Durham University

ISIPTA '09

Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications

Edited by
Thomas Augustin
Frank P. A. Coolen
Serafin Moral
Matthias C. M. Troffaes

ISIPTA '09

Proceedings of the Sixth International Symposium on
Imprecise Probability: Theories and Applications

Durham University, United Kingdom
14–18 July 2009

Edited by
Thomas Augustin
Frank P. A. Coolen
Serafín Moral
Matthias C. M. Troffaes

Published by SIPTA
Society for Imprecise Probability: Theories and Applications
http://www.sipta.org

This book was typeset using LaTeX.

Contents

Preface

The *Sixth International Symposium on Imprecise Probability: Theories and Applications* is held in Durham, United Kingdom, 14–18 July 2009. In addition to an extensive scientific program, the meeting includes visits to Durham Cathedral and Castle, which together are a UNESCO world heritage site.

The ISIPTA meetings are a primary forum for presenting and discussing new advances in imprecise probability, and are held once every two years. The first meeting was held in Gent in 1999, followed by meetings in Ithaca (Cornell University), Lugano, Pittsburgh (Carnegie Mellon University), and Prague. In the decade since the first meeting, imprecise probability has come a long way, which is reflected by the wide range of topics presented at the 2009 meeting, but particularly also in the wider acceptance of imprecise probability in journals and at other conferences.

As with previous ISIPTA meetings, we have avoided parallel sessions. In total, 47 papers are presented by a short talk and poster presentation, which guarantees ample time for discussion of each contribution. The papers are included in these proceedings, and are also available on the SIPTA webpage (`http://www.sipta.org`). Submitted papers have undergone a high quality reviewing process by members of the Program Committee, to whom we are very grateful. The selectivity resulting from the review process, provides trust in the quality of the presented research results.

Nevertheless, it has long been acknowledged that, at the ISIPTA meetings, some good quality papers could not be accepted due to the limited number of papers that can be presented at the meeting. To provide a platform for novel ideas and challenging applications for which the research is not yet completed, poster-only presentations have been introduced at ISIPTA'09. About 25 such contributions will be presented. For each, a short abstract will be distributed at the conference. The abstracts are also available on the SIPTA webpage.

As with previous ISIPTA meetings, a wide variety of theories and applications of imprecise probability will be presented. New application areas and novel ways for dealing with limited information prove the increasing success of imprecise probability. For ISIPTA'09, statistical inference and decision making with imprecise probability has specifically been emphasized, as successful applications in these areas are crucial for wider uptake of imprecise probability. There will be a special discussion session on statistical inference. We are grateful to Kurt Weichselberger who agreed to open this session by reporting on some recent developments and applications of his "Symmetric Theory of Probability". Also the topics of the four tutorials at ISIPTA'09 reflect this emphasis: inference, reliability, decision making, and graphical models. We thank Erik Quaeghebeur, Lev Utkin, Robert Hable, and Cassio de Campos, respectively, for preparing and presenting these tutorials, and for making excellent materials available to the wider community (also on the SIPTA webpage).

Two special session are held at ISIPTA'09. One special session is organized in memory of Henry E. Kyburg Jr. (1928–2007). Henry Kyburg was Gideon Burbank Professor of Moral Philosophy and Professor of Computer Science at the University of Rochester, where he was an active member of their faculty from 1965 until his death, and Chair of their Philosophy Department from 1969–1982. He was among the first researchers to develop a rigorous theory in which probability is interval-valued, and did so by keeping probability an issue of inference and logic, separate from decision making. His original interval-valued theory, *Epistemological Probability*, is found in his 1961 book *Probability and the Logic of Rational Belief*, itself a development of

ideas he presented in his 1956 Columbia University PhD thesis *Probability and Induction in the Cambridge School*, which was written under the supervision of Ernest Nagel. A later version of his theory, with a variety of applications, is given in his 1974 book *The Logical Foundations of Statistical Inference*. Henry's theory can be seen as a generalization of R.A. Fisher's fiducial probability, where the generalization allows interval-valued probability when precise frequency information is lacking about pivotal quantities. Also, Henry's *Epistemological Probability* is an inductive logic: it carries forward foundational ideas found in J.M. Keynes' *Treatise on Probability*, including Keynes' idea that not all probabilities are comparable.

Henry was a member of the Program Committees for the ISIPTA meetings and gave a memorable after-dinner address to the Society at the 2nd Symposium, ISIPTA'01, held at Cornell University. His calm, caring, and deliberate manner, tempered by a dry wit, will be sorely missed.

The organisers are pleased that Isaac Levi has agreed to make opening remarks at the special session in memory of Henry Kyburg. Isaac followed Henry at Columbia University as a student of Ernest Nagel. They were dear friends, and fierce competitors for more than 45 years over how interval-valued probability should be developed.

A second special session is organized in memory of Pauline Coolen-Schrijner (1968–2008). Pauline was Reader in Probability and Statistics at Durham University, and involved in the early plans of organising ISIPTA'09 in Durham. Her PhD research was on quasi-stationarity of discrete-time Markov chains, a topic generalized to imprecise probability by her PhD student Richard Crossman, who presents this work at ISIPTA'09. Pauline was intensively engaged, together with her husband and colleague Frank Coolen, in the development of nonparametric predictive inference (NPI), which is a new exciting methodology for predictive inference under low structure assumptions leading to interval-valued probabilities.

Two PhD students who Pauline supervised present results on NPI at ISIPTA'09: Tahani Maturi presents NPI for multiple comparisons and Rebecca Baker presents the application of NPI to category selection. Pauline published over 40 papers in a wide range of journals in Stochastics, Statistics, Operational Research and Reliability. She particularly developed NPI for replacement problems, which offers great adaptivity to process data. Research results which Pauline achieved with her PhD students, as well as further results with Frank Coolen, will lead to a substantial number of further papers to be published in the near future.

During the last two decades, Pauline suffered from increasingly devastating diseases, which forced her to gradually give up many of the things most precious to her. However, she was always full of optimism, mental strength and energy, with a special sense for the small things in life, being well aware that, quoting the last line of her web page, "breathing is not something we can take for granted."

We believe that, in the 10 years since ISIPTA'99, imprecise probability has found a solid place in research on uncertainty quantification and related fields. Because applications are increasing, both in number and success, we are optimistic about the future impact of imprecise probability. However, with widening acceptance of the theories, new challenges for the ISIPTA meetings arise: emphasis is likely to shift from raising awareness of developments and opportunities, to stimulating discussion between people with a wider range of interests. We believe that the current format of ISIPTA is successful, and we hope that all participants will find the meeting pleasant, informative, and beneficial. We hope that ISIPTA'09 provides a good platform to present and discuss work, and hopefully also leads to new ideas and collaborations. Whether or not this format will remain suitable in the future, in anticipation of more and more participants at these meetings, is an interesting problem which we happily leave for the organisers of ISIPTA'11 and beyond to contemplate.

Finally, we wish to thank several people for their support. Marco Zaffalon, the SIPTA President, regularly provided us with useful information, and ensured that this conference benefits from previous experiences. Gert de Cooman and Teddy Seidenfeld did similarly, and as members of the Steering Committee they provided useful input throughout the preparations for this conference. We also thank Teddy for suggesting the special sessions in honour of Henry Kyburg and Pauline Coolen-Schrijner. We are grateful to the Durham Events staff, in particular Judith Aird, for their professional support in the organisation of the conference. We also thank all

who have contributed to the success of ISIPTA'09, be it by submitting their research results, presenting them at the conference, reviewing papers, or by attending sessions and participating in discussions. Finally, we thank *you* for picking up these proceedings and reading the excellent papers. We are confident that you will enjoy them! May these papers be an everlasting proof of the success of ISIPTA'09.

Thomas Augustin
Frank P. A. Coolen
Serafín Moral
Matthias C. M. Troffaes

July 2009

Organisation

Steering Committee

Thomas Augustin, Germany
Frank P. A. Coolen, UK
Gert de Cooman, Belgium
Serafín Moral, Spain
Teddy Seidenfeld, US
Matthias C. M. Troffaes, UK

Sponsors

ELSEVIER

http://www.elsevier.com/

EPSRC Engineering and Physical Sciences Research Council

http://www.epsrc.ac.uk/

The London Mathematical Society

http://www.lms.ac.uk/

Program Committee Board

Thomas Augustin, Germany
Frank P. A. Coolen, UK
Serafín Moral, Spain
Matthias C. M. Troffaes, UK

Program Committee Members

Joaquín Abellán, Spain
Alessandro Antonucci, Switzerland
Horacio Arló-Costa, USA
Yakov Ben-Haim, Israel
Salem Benferhat, France
Dan Berleant, USA
Mikelis Bickis, Canada
Sudip Bose, USA
Cassio Campos, Switzerland
Andrea Capotorti, Italy
Marco Cattaneo, Germany
Giorgio Corani, Switzerland
Inés Couso, Spain
Fabio Cozman, Brazil
Richard Crossman, UK
Fabio Cuzzolin, UK
Gert de Cooman, Belgium
Thierry Denœux, France
Sébastien Destercke, France
Serena Doria, Italy
Love Ekenberg, Sweden
Malcolm Farrow, UK
Scott Ferson, USA
Thomas Fetz, Austria
Pablo Fierens, Argentina
Terrence Fine, USA
Angelo Gilio, Italy
Robert Hable, Germany
Jim Hall, UK
Peter Harremoës, The Netherlands
Hermann Held, Germany
Filip Hermans, Belgium
Aparna V. Huzurbazar, USA
Manfred Jaeger, Denmark
Radim Jiroušek, Czech Republic
George J. Klir, USA
Igor Kozine, Denmark
Vladik Kreinovich, USA
Jonathan Lawry, UK
Isaac Levi, USA
Thomas Lukasiewicz, UK
Enrique Miranda, Spain
Michael Oberguggenberger, Austria
Endre Pap, Serbia and Montenegro
Renato Pelessoni, Italy
Erik Quaeghebeur, Belgium
Peter Reichert, Switzerland
David Ríos Insúa, Spain
Fabrizio Ruggeri, Italy
Damjan Škulj, Slovenia
Michael Smithson, Australia
Wynn Stirling, USA
Jörg Stoye, USA
Carolin Strobl, Germany
Choh M. Teng, USA
Lev Utkin, Russia
Barbara Vantaggi, Italy
Jiřina Vejnarová, Czech Republic
Paolo Vicig, Italy
Frans Voorbraak, The Netherlands
Kurt Weichselberger, Germany
Alyson Wilson, USA
Nic Wilson, UK
Marco Zaffalon, Switzerland

Iterated Random Selection as Intermediate Between Risk and Uncertainty

Horacio Arló Costa
Carnegie Mellon University
hcosta@andrew.cmu.edu

Jeffrey Helzner
Columbia University
jh2239@columbia.edu

Abstract

In [7] Hertwig et al. draw a distinction between *decisions from experience* and *decisions from description*. In a decision from experience an agent does not have a summary description of the possible outcomes or their likelihoods. A career choice, deciding whether to back up a computer hard drive, cross a busy street, etc., are typical examples of decisions from experience. In such decisions agents can rely only of their encounters with the corresponding prospects. By contrast, an agent furnished with information sources such as drug-package inserts or mutual-fund brochures—all of which describe risky prospects—will often make decisions from description.

In [7] it is shown (empirically) that decisions from experience and decisions from description can lead to dramatically different choice behavior. Most of these results (summarized and analyzed in [6]) are concerned with the role of risk in decision making. This article presents some preliminary results concerning the role of *uncertainty* in decision making. We focus on Ellsberg's two-color problem and consider a chance setup based on double sampling. We report empirical results which indicate that decisions from description where subjects select between a clear urn, the chance setup based on *double sampling* and Ellsberg's *vague* urn, are such that subjects perceive the chance setup at least as an intermediate option between clear and vague choices (and there is evidence indicating that the double sampling chance setup is seen as operationally indistinguishable from the vague urn). We then suggest how the iterated chance setup can be used in order to study decisions from experience in the case of uncertainty.

Keywords. decisions from description, decisions from experience, random selection, uncertainty

1 Introduction

Consider a scenario in which a well-educated couple must decide whether or not their child should receive a particular vaccination. To assist in their decision making, the couple is provided with statistics concerning the frequency of serious, adverse reactions associated with the vaccination in question. Though the frequency of such adverse reactions is quite low, the couple is reluctant to have their child vaccinated. Concerned, the child's pediatrician provides reasons in favor of vaccination, noting that she herself in fact had chosen to vaccinate her own children. What might explain the difference between the judgement of the child's parents and that of the child's pediatrician? One plausible explanation that is of general, theoretical interest focuses on the way in which the relevant parties are acquainted with the chances associated with an adverse reaction. While the child's parents are provided with frequencies, and presumably the child's pediatrician is privy to this information, the pediatrician can also draw upon her own clinical experience.

Hertwig et al. propose to analyze cases of this type in terms of a distinction between *decisions from description* and *decisions from experience*. As suggested by the scenario in the previous paragraph, which is among the scenarios that provide motivation for the work reported in [7], decisions from description are made with the benefit of a description of the relevant chance mechanism, e.g., associated probabilities, while decisions from experience are informed by repeated encounters with the chance mechanism itself, i.e., sampling. As noted in [7], the vast majority of experimental work on decision making has focused on decisions from description. The following example, taken from Kahneman and Tversky's 1979 classic on prospect theory [8], illustrates the methodology typical among work in this area:

Example 1. *Which of the following do you prefer?*

Alternative 1 pays \$5000 with probability .001 and \$0 with probability .999. Alternative 2 pays \$5 with probability 1.

Is there any reason to think that this focus on decisions from description has been significant with respect to the results gathered through numerous studies which employ the sort of methodology illustrated in Example 1? In [7], Hertwig et al. answer this question in the affirmative. Specifically, they present evidence indicating that people tend to "underweight the probability of rare events" when making decisions from experience. This is in stark contrast to the well-known results of Kahneman and Tversky, based on items such as Example 1, which indicate that people tend to overweight the probability of rare events when making decisions from description. Thus, it seems that Hertwig et al. have isolated an important psychological effect. The opening scenario clearly illustrates the effect Hertwig et al. have isolated. The child's parents are presented with frequencies which they interpret as a description of the relevant chance mechanism. As predicted by Hertwig et al., the child's parents overweight the probability of an adverse reaction and decide not to vaccinate. By contrast, the pediatrician's decision making is informed by her clinical experience. As predicted by Hertwig et al., the child's physician underweights the probability of an adverse reaction and recommends vaccination.

Hertwig et al. restrict their attention to decision making under risk. In particular, the descriptions that they employ include a numerically precise probability distribution. The main purpose of this paper is to begin an investigation into the possibility of a *description-versus-experience* effect in the context of decision making under uncertainty, i.e., in contexts where the description of the relevant chance setup does not determine a numerically precise probability distribution. On the basis of experimental evidence reported in this paper we conjecture that the gap between *vague* and *clear* is less pronounced in the case of decisions from experience than in the case of decisions from description.

2 From Risk to Uncertainty

Consider the design of the first study reported by Hertwig et al. in [7]. The subjects were divided into two groups: Description and Experience. Those in Description were presented with several choice problems, each of which consisted of a pair of risky alternatives. For example, one such choice problem consisted of the alternatives $(4, .8)$ and $(3, .1)$, where (m, p) denotes the risky alternative that pays amount m with probability p and pays amount 0 with probability $1 - p$.

Let (m, p) be a risky alternative of the indicated sort. One can construct a chance setup that satisfies description (m, p). For example, a chance setup for the alternative $(4, .8)$ could use random draws (with replacement) from an urn consisting of 80 black balls and 20 white balls. The implementation of the second group, Experience, is less familiar. Consider a decision-from-experience counterpart to the choice problem consisting of $(4, .8)$ and $(3, .1)$. One could, for example, present the subject with two buttons, say A and B, where pressing A (B) results in a trial on a particular chance setup corresponding to $(4, .8)$ ($(3, .1)$). The subject, who sees the result of each trial (e.g., in the case of A, whether the payoff would have been 4 or 0), is permitted to sample the two chance setups as many times as they wish before they are required to make a choice and play one of the two setups for real payoffs. Essentially, this is the way in which Hertwig et al. study decision making from experience.

What happens when we move from risk to uncertainty (where probabilities are imprecise)? The obvious candidate on the description side is familiar through the presentation of Ellsberg problems such as following:

Example 2 (Ellsberg's two-color problem [4])**.** *Consider the following two cases:*

Urn A *contains exactly 100 balls. 50 of these balls are solid black and the remaining 50 are solid white.*

Urn B *contains exactly 100 balls. Each of these balls is either solid black or solid white, although the ratio of black balls to white balls is unknown.*

Consider now the following questions: How much would you be willing to pay for a ticket that pays \$55 (\$0) if the next random selection from Urn A results in black (white) ball? Repeat then the same question for Urn B.

Following the above presentation, an uncertain alternative over a pair of prizes (only one of which is nonzero) can be specified by providing the amount of the nonzero prize and the *set* of probabilities that are associated with that prize. Thus, for example, $(55, \{\frac{i}{100} \mid 0 \leq i \leq 100\})$ is the uncertain alternative in which the probability of winning \$55 is known to be in $\{\frac{i}{100} \mid 0 \leq i \leq 100\}$.

Presenting alternatives in this way has the virtue of generalizing the risky alternatives that are employed by Hertwig et al., since these risky alternatives are simply those of the form $(m, \{p\})$. However, once the probabilities are allowed to be indeterminate, it is not clear how to complete the analogy in a way that would support decision from *experience* under uncertainty. Recall the desired relationship in the case of

risk. Given a risky alternative, e.g., $(m, \{p\})$, one can construct a corresponding chance setup, i.e., one that satisfies description $(m, \{p\})$. Doing the analogous thing in the case of uncertainty would seem to require chance setups that implement indeterminate probabilities. Are there such things in any interesting sense? After all, the uncertainty described in Ellsberg-type examples is purely epistemic, e.g., the ratio between black ball and white balls in Urn B *is* determinate even though it is not known to the decision maker. On the other hand, consider the following description of a chance setup:

B^***:** First, select an integer between 0 and 100 at random, and let n be the result of this selection. Second, make a random selection from an urn consisting of exactly 100 balls, where n of these balls are solid black and $100 - n$ are solid white.

As in the case of Urn B, the outcome of a trial on chance setup B^* depends on a random selection from an urn such that the ratio of black balls to white balls is not known to the subject. However, unlike the case of Urn B, the subject knows that the urn that is sampled in the second stage of B^* is determined by a random selection in the first stage of B^*. According to at least one familiar line of reasoning, this second consideration suggests that a play on B^* is equivalent to a play on Urn A, rather than Urn B. The indicated line of reasoning is roughly as follows: The random selection in the first stage entails that, for each integer i, where $0 \leq i \leq 100$, there is a probability of $\frac{1}{101}$ that the urn sampled in the second stage consist of i black balls and $100 - i$ white balls. Moreover, according to this line of reasoning the random selection in the second stage entails that if i is selected in the first stage, then the probability of selecting a black ball in the second stage is $\frac{i}{100}$. This line of reasoning then continues by combining the first and second stage probabilities to conclude that the probability of getting a black ball on a trial of B^* is $\frac{1}{101}(\sum_{i=0}^{100} \frac{i}{100}) = \frac{1}{2}$, as in the case of Urn A. There are, of course, well-known responses to this line, the most obvious being one that questions the relevance of a chance setup's long-run behavior when it comes to assigning probabilities for a single trial of the setup; here we are assuming that the relevant probabilities are based on frequencies rather than something like propensities.

A less familiar response maintains that one has complete uncertainty over the collection of chance setups that satisfy description B^* and that, since some of the setups will select a black ball on their next trial while others will select a white ball, one has complete uncertainty with respect to the outcome of the next trial. For example, even if one maintains that there are truly nondeterministic chance setups, deterministic chance setups are common and are of the sort that Hertwig et al. employ in [7]. If random selection is understood to mean selection by a mechanism such that (1) future behavior of the mechanism cannot be predicted from a mere knowledge of its past behavior and (2) the various possible outcomes are distributed evenly in the long run – and these are important matters that will be considered in the sections that follow – then the description of B^* is compatible with the use of such deterministic mechanisms.

For all the subject knows, the first stage selection in B^* can be made according to a deterministic process that will select 33 on its next run, while the second stage mechanism will be made according to a deterministic mechanism that will select a black ball on its next draw from the 33:77 urn. Similarly, it is compatible with the information that is presented to the subject that the first stage selection in B^* will be made according to a deterministic process that will select 61 on its next run, while the second stage mechanism will be made according to a deterministic mechanism that will select a white ball on its next draw from the 61:39 urn. The subject has complete uncertainty over these selection mechanisms and, more generally, over the collection of all chance setups that might be used to carry out the selections in B^*. Since the final outcome, i.e., the selection of a black or white ball, is a function of the chance setups which are employed, at least where deterministic mechanisms are used, complete uncertainty over the collection of these mechanisms suggests complete uncertainty with respect to the final outcome. Note that this line is rather extreme, since it suggests complete uncertainty even in situations where the second stage urn is fixed, e.g., as in a chance setup that makes selections with replacement from Urn A. Rather than trying to achieve consensus with respect to such *a priori* considerations, we now turn our attention to psychological matters.

3 Study

What is the psychological relationship between the description of Urn B and B^*? To investigate this question we asked subjects to state their maximum buying prices with respect to hypothetical situations involving the descriptions at issue. Our study included 89 undergraduates from Carnegie Mellon. At the time of the study the participating students were enrolled in 80-100, an introductory philosophy course at Carnegie Mellon. Each subject was presented with a questionnaire, the contents of which will now be described.

After instructing the subjects that they would be presented with questions involving hypothetical scenar-

ios, the questionnaire continued with the following tutorial on chance setups:

> Think of a roulette wheel of the sort that one would find in any American casino. The casino employee spins the wheel in one direction and then sends a ball in the other direction along a track that goes around the circumference of the wheel. Eventually the ball comes to rest in one of the wheel's 38 pockets. Players expect that this setup is fair in the sense that the following conditions are satisfied: (1) In the long run the number of times that the ball lands in a particular pocket is equal to the number of times that the ball lands in any other particular pocket and (2) One cannot predict where the ball will land on the next spin simply by knowing where the ball landed on previous spins.
>
> Roulette wheels are a special case of a more general class of systems. More generally, a *chance setup* is a system that includes a finite set $\{o_1, ..., o_n\}$ of possible outcomes and that outputs one of these outcomes each time that it is run. Such a chance setup is *fair* just in case the following conditions are satisfied: (1) If the system were run repeatedly, then in the long run the number of trials that would result in outcome o_i would be equal to the number of trials that would result in outcome o_j and (2) One cannot predict which outcome will result from the next run of the system simply by knowing the outcome of each of the previous runs of the system.

The questionnaire then continued by instructing the subjects that "random selection" in the context of the questionnaire is to be understood as selection via a fair chance setup. This instruction was followed by three questions:

1. An urn has been filled with exactly 100 balls. 50 of the balls are black and the remaining 50 are white. A random selection from the urn will be made. **What is the most that you would be willing to spend on a ticket that pays \$55 if the random selection results in a black ball and pays \$0 if the random selection results in a white ball?**

2. Consider the following two-stage process: (1) A random selection is made from the collection $\{0, 1, ..., 100\}$ and (2) A second random selection is made from an urn that contains exactly n black balls and $100 - n$ white balls, where n is the result of the random selection from the first stage. Thus, for example, if 23 was the result of the random selection in the first stage, then the random selection in the second stage would be from an urn containing exactly 23 black balls and 77 white balls. **What is the most that you would be willing to spend on a ticket that pays \$55 if the random selection in the second stage results in a black ball and pays \$0 if the random selection in the second stage results in a white ball?**

3. An urn has been filled with exactly 100 balls. Each ball in the urn is either black or white. However, the ratio of black balls to white balls in the urn is unknown. A random selection from the urn will be made. **What is the most that you would be willing to spend on a ticket that pays \$55 if the random selection results in a black ball and pays \$0 if the random selection results in a white ball?**

Note that the ticket is played against the "clear" chance setup in the first question, against the "double sampling" chance setup in the second question, and against the "vague" chance setup in the third question. Thus, referring back to Example 2, the first of these questions asks the subject to price the ticket on Urn A, while the second asks the subject to price the ticket with respect to B^*, and the third asks the subject to price the ticket on Urn B.

Recognizing that the order in which the questions appeared might affect the responses, we created three different versions of the questionnaire: CDV, VDC, and DVC. Version CDV was administered to 41 subjects and presented the questions in the order given above, i.e., the question concerning the clear setup followed by the question concerning the double-sampling setup followed by the question concerning the vague setup. Version VDC, which was administered to 32 subjects, reversed this order. Version DVC was given to 16 subjects and had the question about double-sampling occurring first, the question about the vague setup occurring second and the question about the clear chance setup occurring third. In each version, subjects were instructed to answer the questions in the order that they were presented.

For each of the three groups, Table 1 shows the mean maximum buying price for the three questions. Thus, for example, the first row of the second column indicates that, in the case of the ticket on the clear chance setup, \$22.68 was the mean maximum buying price for the group of subjects that received the VDC version of the questionnaire.

Question	CDV mean	VDC mean	DVC mean
Clear	22.89	22.68	19.02
Double	14.68	9.77	7.70
Vague	5.82	7.10	3.25

Table 1

While the order of the questions appears to have had some bearing, the basic pattern of Vague < Double < Clear for the mean maximum buying prices seems robust across the three versions of the questionnaire. The mean maximum buying prices over all subjects are shown in Table 2.

Question	Mean
Clear	22.12
Double	11.66
Vague	5.82

Table 2

If we turn our attention to the level of individual subjects, then the above pattern of strict inequalities is less pronounced since approximately $\frac{1}{3}$ of the subjects gave the same maximum buying price for Double and Vague; moreover, these subjects were not distributed evenly across the three groups. However, as shown in Table 3, a clear pattern emerges if we weaken the first inequality.

Group	$\# V \leq D < C$	$\% V \leq D < C$
CDV	29	71%
VDC	24	75%
DVC	12	75%
All	65	73%

Table 3

The first column of Table 3 shows the number of subjects in each group (and overall) that satisfied the Vague ≤ Double < Clear pattern, while the second column shows the associate percentages. As Table 3 shows, these percentages are quite stable across the three groups individually and their union. It is also worth noting that, as shown in Table 4, relatively few of the subjects gave the same maximum buying price for Clear and Double.

Group	$\# C = D$	$\% C = D$
CDV	12	29%
VDC	6	19%
DVC	4	25%
All	22	25%

Table 4

The first column of Table 4 shows the number of subjects in each group (and overall) that gave the same maximum buying price for Clear and Double, while the second column shows the associate percentages.

The data also reveal as well an interesting result if we consider the values for **V**ague, **D**ouble, and **C**lear given by the mean maximum buying prices for **V**DC, **D**VC and **C**DV, respectively. These values are important because these are the results for the three cases without considering comparisons – subjects were instructed to answer the questions sequentially and not to return to previous questions. For example, since Vague occurs first in VDC, it seems reasonable to assume that it is evaluated in a non-comparative context. Table 5 shows the mean maximum buying prices for these three non-comparative cases:

Question	Mean
Clear	22.89
Double	7.70
Vague	7.10

Table 5

These results suggest an operational identification of Vague and Double, which have almost identical values; this is of interest to us since Double, unlike Vague, seems to be implementable in a way that could support decisions from experience, and this is something that we will revisit in the following section. Furthermore, these values for Vague and Double are clearly separated from the mean maximum buying price for Clear as reported in Table 5. Perhaps the only concern here is that the number of subjects who received VDC is low in comparison to the number of subjects in the other two groups. We expect to run a larger experiment including decisions from experience. In this case it would be interesting to see whether the pattern verified in the pilot is robustly maintained.

Another issue that we intend to address in future experiments concerns the worry, expressed by one anonymous referee, that the complexity of B^* rather than its status with respect to uncertainty is responsible for its lower price. This raises two interesting issues. First, how might we control for this in future experiments? Perhaps one way to do this would be to have the subjects explain the reasoning behind their responses. We can certainly ask the subjects why they priced one alternative lower than another. The second issue concerns the relationship between complexity and uncertainty. In the present context this seems to line up with the familiar distinction between imprecision and indeterminacy. One might claim that the credal probabilities in the case of B^* are imprecise

but not indeterminate, e.g. that the rational agent is committed to a particular credal probability but is unable identify that particular distribution. This sort of situation is not unlike the imprecision that arises in connection with the measurement of physical concepts, e.g. length or weight. In contrast, one might claim that in the case of Urn B the rational agent's credal probability itself – rather than just its estimation of that credal probability – ought to be indeterminate. Surely the distinction is not mere stipulation, but what is wrong with maintaining that the rational agent's credal probabilities with respect to B^* should also be indeterminate (i.e. that the rational agent is not committed to a determinate credal probability in such a case)? It seems that considerations of this sort lead in the direction of bounded rationality, in particular the tenability of capacity independent notions of rationality.

4 Discussion

The results of our study suggest that double sampling is perceived as something in between risk and uncertainty when comparative contexts are allowed, but is there a way to make this intermediate position more transparent? Perhaps one way to do this would be to describe a mixed chance setup in which the subject is told that the selection will be made from Urn A with probability p and from Urn B with probability $1-p$. One could then attempt to identify the value of p at which the subject's maximum buying price is equal to the maximum buying price that the subject stated with respect to B^*. The value of p so identified could be taken as a representation of the intermediate position that B^* occupies between risk (Urn A) and uncertainty (Urn B).

One drawback to using descriptions of mixed chanced setups in the manner suggested above is that such descriptions do not appear to fit, at least psychologically, into the sets-of-probabilities approach to representing credal states. While the set of all distributions p such that $p(\text{Black}) = \lambda p_1(\text{Black}) + (1-\lambda)p_2(\text{Black})$, where $p_1 \in X$ and $p_2 \in Y$ might seem like a natural representation of a mixture with probability λ on X and probability $1-\lambda$ on Y, some preliminary results reported in [2] suggest that this is not the case.

If descriptions of double sampling are perceived as something distinct from risk, what might an implementation of such a description look like in a study of decision making from experience? Recall the study by Hertwig et al. in [7]. They implemented the description of a risky alternative, such as (m, p), as an appropriate chance setup. A trial on this chance setup is activated by a button on a computer screen. After pushing this button, the subject sees the outcome of the trial on the computer screen. If we attempt to implement B^* in such a way that the subject sees only the final result of the two-stage process after pushing the appropriate button, then we run into problems. The issue is that there is nothing to guarantee that such an implementation of B^* could not just as well serve as an implementation of Urn A. By assumption, a chance setup that satisfies B^* will, in the long run, draw j in the first stage approximately $\frac{1}{101}$th of the time. Moreover, in the long run, trials in which j is drawn in the first stage result in the selection of a black ball approximately $\frac{j}{100}$th of the time. Hence, in the long run, $(\frac{1}{101})(\frac{j}{100})$th of the trials result in a black ball drawn from the urn having j black balls and $100-j$ white balls. Thus, in the long run, approximately

$$(\frac{1}{101})(\frac{1}{100})\sum_{j=0}^{100} j = \frac{5050}{10100} = \frac{1}{2}$$

of the trials result in the draw of a black ball, which agrees with the limiting frequencies for an implementation of Urn A. If we assume that such an implementation of B^* would also satisfy the condition that future behavior cannot be predicted solely from a knowledge of past behavior, and this seems to be a psychological matter, then it appears that there is nothing to prevent such an implementation of B^* from serving as an implementation of Urn A. Clearly the implementations of B^* and Urn A must be distinct in a meaningful way if one is to conduct the desired study of decision making from experience. We now consider other proposals for implementing B^*.

One way to avoid the sort of problematic collapse discussed at the end of the previous paragraph would be to make both stages of the double sampling visible to the subject. Thus, for example, pressing the appropriate button on a computer screen for the first time would run the the first-stage selection, and the result of that selection would be shown to the subject, e.g., that the urn with 35 black balls and 65 white balls had been selected. Pressing the button for the second time would initiate a draw from the urn that had been selected, and the result of that second-stage selection would then be shown to the subject, e.g., that a white ball had been drawn. Making both stages of double sampling visible to the subject avoids the problematic collapse since such an implementation of B^* no longer qualifies an an implementation of Urn A.

However, there is a possible objection to this design, based on the fact that it implements a sort of hybrid experimental condition that does not correspond purely to decisions from experience or decisions from description. Typically in decisions from experience

the subjects do not have access to the probabilities of the option considered. In the previous design one makes at least intermediate (i.e., first stage) probabilities explicit by revealing the composition of the selected urn.

There is a remedy to the previous objection via the implementation of the following experimental design. This design assumes that the following four buttons are available to the subject: PLAY, SELECT GAME, V, and C. The subject's initial choice concerns V and C. Button C implements the "clear" scenario that was presented in the questionnaire . If C is selected, then the agent can press PLAY repeatedly. Pressing PLAY samples from the implementation of the clear urn. While the urn structure is hidden from the subject, the subject sees the associated payoffs, \$55 if black and \$0 if white, after each pressing of PLAY. If V is selected, then the subject is instructed to press SELECT GAME. Unbeknownst to the subject, pressing SELECT GAME selects an implementation based on one of the 101 possible urns considered above (i.e., an urn consisting of n black balls and $100-n$ white balls for some $n \leq 100$).

The following algorithm is used to select a game: consider the space of all possible ordered sequences of 101 urns. Then a sequence in this space is selected at random and fixed. When the game starts and the agent presses SELECT GAME for the first time the first urn in the sequence is selected. Say that the subject has pressed SELECT GAME n times. Then when he presses SELECT GAME once more the selection mechanism picks the urn in the $n+1$ position in the sequence and samples it every time that PLAY is selected. Now at each point the probability of white or black will depend on the previous actions of the subject playing the game. Since probabilities should not be attributed to acts these probabilities remain indeterminate.When the sequence terminates the algorithm starts again at the initial point of the selected sequence.

It is important to remark that what *does not* have a determinate probability is the color of the first ball prior to selecting or not a game (i.e. prior to choosing to play). Likewise for the probability of the nth ball prior at the moment of the choice whether or not to select a game for the nth time. After the agent selects a game (i.e. after the agent decides to play the game) we have a uniform precise probability over the 101 configurations of the urn.

In addition after selecting games a few times and sampling them the calculation of the probabilities of the color of the ball in the next trial grows ever more complicated after conditioning on what has been done and on what has been seen from past plays of the game. Even for an ideal agent these probabilities will be imprecise. The agent will have bounds of the values of the probabilities or he can form qualitative judgments comparing probabilities but the agent will not have precise probabilities at his disposal. So, under a normative point of view there is indeterminacy at the moment contemporary to the selection of the game, and after a few trials there will be imprecision in the corresponding probabilities.

When we consider bounded agents the situation is even worse. The agent might not remember well what he did in the past and what he saw in the past and we can have recency effects as well. So, in the real situation we have to deal with imprecise probabilities and choices under uncertainty.

After SELECT GAME is pressed the subject has a choice: explore the game that was selected by pressing PLAY or select a (possibly) new game by pressing SELECT GAME. Pressing PLAY samples from the current game. While the urn structure of the game is hidden from the subject, the subject sees the associated payoffs, \$55 if black and \$0 if white, after each pressing of PLAY. The agent can interact with the two buttons as long as he wants. Notice that it is perfectly possible that an agent selects a game and then presses PLAY repeatedly without selecting any other game. Notice that in this case the argument in terms of frequencies fails (the agent does not see data from all urns, he just considers a single (or a few) urns). So, the shift to the design where the two stages are visible is essential in order to avoid the argument for the collapse into urn A.

Example 3. ***A possible session involving V:***

The subject first presses V.

On screen: *Select a game by pressing SELECT GAME. OR EXIT*

The subject presses SELECT GAME.

On screen: *You have been awarded a game. You can play this game by pressing PLAY.*

The subject presses PLAY and a payoff appears, for example:

On screen: *You won \$55.*

The subject then agent faces the choice of pressing PLAY again or pressing SELECT GAME or EXIT (in which case he faces the election of V and C again).

If the agent chooses instead button C, he will have the option of pressing PLAY as many times as he wants. A payoff will appear each time that PLAY is pressed. At any time he can STOP and choose between V and

C.

Example 4. ***A possible session involving C:***

The subject first presses C.

On screen: *Press PLAY.*

The subject presses PLAY and a payoff appears, for example:

On screen: *You won $0.*

The subject will can press PLAY as many times as he wants. A payoff will appear after each time that PLAY is pressed.

After receiving feedback from these two buttons the agent has to select V or C and in this case he will play for real money. Of course, if he selects V a new game will be selected by pressing SELECT GAME and he will receive the payoff determined by the next pressing of PLAY, i.e., by sampling the urn corresponding to the current game.

This design makes visible the two-stage nature of the V button but, as in the case of decisions from experience for risk, the agent does not receive any information about intermediate probabilities. Notice that the algorithm used in the proposed implementation of decisions from experience is a particular instance of a selection via a fair chance set up (as described to subjects in the tutorial). Since it is clear that in this case there is no collapse of the implemented mechanism with the clear urn, it follows that in general there is no reason to expect a collapse of B^* and C. This shows that the operational identification of the V and B^* conditions might not just be attributable to a statistical error on the part of the subjects.

We conjecture that this design of decisions from experience will also avoid an identification of the V and C conditions. We also conjecture nevertheless that the gap between the V and C conditions (buttons in decisions from experience) will not be as severe between the gap between the corresponding "vague" condition and the "clear" conditions in the case of decisions from description. This is because it is unlikely that the subject encounters rare events while obtaining feedback by interacting with button V, and this suggests that the subject will remain ignorant of their existence (examples of extreme values or rare events will be the case where either Black or White are zero or very low in the sampled urns).

We conjecture therefore that this proposal will show a significant difference between decisions from description and decisions from experience, demonstrating that the distinction between these two types of decisions is robust and applicable not only to risk but also to the case of uncertainty.

5 Further Considerations

We conclude by mentioning another issue that is raised by our study, an issue that seems to have general significance for experimental work on decision making. An unusual aspect of the questionnaire that we used in the study that is reported in Section 3 is the fact that it is explicit about what is meant by a random selection. While references to random selection are common in experimental work on decision making, these references are seldom accompanied by something like the tutorial on fair chance setups that was part of our study. It is natural to wonder if this makes a difference. While we have yet to conduct a study of this particular question, we do have data from an earlier study that seems to suggest that it does make a difference if one is explicit about what is meant by a random selection.

As part of the study that we reported in [1], we used a questionnaire that asked subjects to state their maximum buying price for what were essentially questions Clear and Vague as presented in Section 3. It is important to note that the questionnaire that was used in [1] did not include any tutorial on fair chance setups, nor for that matter did it include any elaboration regarding the nature of random selection. Finally, it should be noted that the subjects in this earlier study were, like those of the study reported in Section 3, undergraduates at Carnegie Mellon who, at the time of the study, were enrolled in 80-100, which as noted in Section 3 is an introductory philosophy course. Table 6 shows the mean maximum buying prices for the two groups in the earlier study that received Clear as the first question on their questionnaire. [1]

Group	Mean for Clear (2005)
I	15.33
II	13.65

Table 6

These values seem significantly less than the mean maximum buying prices for Clear that are reported in Section 3. These differences seem striking when one considers the mean maximum buying prices that were obtained for Vague in the earlier study. Table 7 shows the mean maximum buying prices for the two groups in the earlier study that received Vague as the

[1] The two groups were distinguished by the fact that they were given slightly different questionnaires. Both groups received a questionnaire that had Clear as the first question, but there were some differences between the two questionnaires in their later sections. We do not think that these differences are significant in the present context, but the interested reader can consult [1] for a detailed description of the questionnaires that were involved.

first question on their questionnaire. [2]

Group	Mean for Vague (2005)
I	5.42
II	6.4

Table 7

These values seem more in line with the mean maximum buying prices for Vague that are reported in Section 3. As a measure of this effect, Table 8 shows the ratio of the mean maximum buying price for Vague to that of Clear.

Group	Vague/Clear
2005	.41
2008	.31

Table 8

Column 2 of the first row in Table 8 shows the average of the two values reported in Table 7 divided by the average of the two values reported in Table 6. The second row of Table 8 shows the mean maximum buying price for Vague as reported in Table 5 divided by the mean maximum buying price for Clear as reported in Table 5. Taken together, these further considerations would seem to raise an important question as to how subjects are interpreting references to random selection in those studies that do not elaborate on what is meant by such a thing.

ACKNOWLEDGMENTS:

A first version of this paper was read at the Seminar of Games and Decisions at CMU. We received numerous comments then. We want to thank in particular Teddy Seidenfeld who proposed a variant of the design of decisions from experience that we are adopting here. Isaac Levi and Paul Pedersen also provided useful comments and criticism. Finally we want to thank specially Ralph Hertwig who provided invaluable feedback and criticism. Various insightful conversations with Ralph during his recent visit to Columbia University motivated us to write this paper.

References

[1] Arló-Costa, H. & Helzner, J. (2005). Comparative Ignornace and the Ellsberg Phenomenon. *Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*, eds. Cozman, F, Nau, R. & Seidenfeld, T.

[2] Arló-Costa, H. & Helzner, J. (2007). On the Explanatory Value of Indeterminate Probabilities. *Proceedings of the Fifth International Symposium on Imprecise Probabilities and Their Applications*, eds. De Cooman, G, Vejnarova, J. & Zaffalon, M.

[3] Chow, C.C. & Sarin, R.K. (2001). Comparative Ignorance and the Ellsberg paradox. *Journal of Risk and Uncertainty*, Vol. 22, No. 2, 129-139.

[4] Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, Vol. 75, No. 4, 643-669.

[5] Fox, C.R., & Tversky, A. (1991). Ambiguity Aversion and Comparative Ignorance. *The Quarterly Journal of Economics*, Vol. 110, No. 3, 585-603.

[6] Hertwig, R. (2009) The Psychology and Rationality of Decisions from Experience, *Synthese*, forthcoming.

[7] Hertwig, R., Barron, G, Weber, E.U., & Erev, I. (2003). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, Vol. 15, No. 8, 534-539.

[8] Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Rconometrica*, Vol. 47, 263-291.

[2] The distinction between the two groups in this case is completely analogous to the distinction between the two groups in Table 6.

Closure of independencies under graphoid properties: some experimental results

Marco Baioletti
Dip. Matematica e Informatica,
Università di Perugia, Italy,
baioletti@dipmat.unipg.it

Giuseppe Busanello **Barbara Vantaggi**
Dip. Metodi e Modelli Matematici,
Università "La Sapienza" Roma, Italy,
e-mail: {busanello, vantaggi}@dmmm.uniroma1.it

Abstract

In this paper we describe an algorithm for computing the closure with respect to graphoid properties of a set of independencies. Since the computation of the complete closure is infeasible, we provide a procedure, called FC1, which is based on a unique inference rule and on the elimination of redundant independencies. FC1 is able to compute a reduced form of the closure, called *fast closure*, which is equivalent to the complete closure, but whose size is much smaller. Some experimental tests have been performed with an implementation of the procedure in order to show the computational behavior of the algorithm. We have also compared the computational cost and the size of the fast closure with the corresponding data for the complete closure.

Keywords. Conditional independence models, Graphoid properties, Inferential rules.

1 Introduction

Conditional independence structures arise in different frameworks, in particular, in probability and in multivariate statistics [11, 14, 15, 18, 20, 23, 31]. It is well known [14] that for any probability measure P the associated independence model $\mathcal{M}$, under the classical definition of independence, is a semi–graphoid (i.e. it satisfies symmetry, decomposition, weak union, contraction) and if P is strictly positive, then $\mathcal{M}$ is a graphoid (also intersection property holds). On the other hand, other independence notions have been introduced in a probabilistic setting [7, 8, 12, 21, 26] and under them graphoid properties have been tested. Moreover, it is well known that graphoid properties are met also by other relations (see [15]) like separation property in graph.

The significance of independence models and graphoid structures is not limited to probabilistic models: in fact many independence models arising from different uncertainty measures are tested on the basis of graphoid properties (see e.g. [1, 9, 10, 13, 15, 16, 17, 19, 23, 27, 30]) and obviously not all the properties among those of graphoid hold.

A significant problem is when a field expert provides an uncertainty measure φ (or better a partial uncertainty assessment, e.g. a coherent conditional probability assessment) and a set J of conditional independence statements, in such case it is necessary to check whether the set J is induced or compatible with φ [29] and then to find all the set of independencies deducible from J.

Then, the aim of this paper is to consider a set J of conditional independence statements, compatible with an uncertainty assessment, and to build in an efficient way the closure through graphoid properties of J.

The computation of the closure is infeasible since its size is exponentially larger than the size of the initial set J of independence statements (see [23, 24]). Then, our aim in [3, 4] (as that in [23, 24] essentially for the case of semi–graphoids) is to build a suitable reduced set of independence statements (obviously included in the closure of J with respect to graphoids), which is as small as possible and it represents the same independence structure. From this reduced set all the relations in the closure should be easily deducible.

In other words, this small set of independence statements, which is called "fast closure", can be considered a basis for the closure.

The computation of the fast closure is relevant also for the selection problem (based essentially on statistical tests) of a model on the basis of data for building, for example, the relevant Bayesian network.

In this paper we describe an algorithm to compute the reduced set. This algorithm is based on a unique inference rule introduced in [4]. In the quoted paper we have also compared this algorithm with another

based on two inferential rules, which are deduced from [24] and studied in our previous paper.

An empirical evaluation of the performance of the introduced algorithm is provided by showing computation times and number of iterations, as well as a comparison between the needed time to compute the fast closure and the time for computing the complete closure (the size of both closures is compared).

The paper is organized as follows: in Section 2 some preliminaries concepts about graphoids, closure and implications for independence relations are recalled. In Section 3 we describe the generalized inference rules and the concept of fast closure; while in Section 4 a system based on a unique inference rule and its corresponding algorithm FC1 are introduced. In Section 5 we describe and comment some experimental results.

2 Graphoid structures

Throughout the paper the symbol $\tilde{S} = \{Y_1, \ldots, Y_n\}$ denotes a finite not empty set of variables. Given an uncertainty measure φ, a conditional independence statement $Y_A \perp\!\!\!\perp Y_B | Y_C$ (compatible with φ), where A, B, C are disjoint subsets of the set of indices $S = \{1, \ldots, n\}$, is denoted simply also as an ordered triple (A, B, C).

Let $S^{(3)}$ be the set of triples (A, B, C) of disjoint sets of S such that A and B are not empty, then a conditional independence model, related to an uncertainty measure φ, is a subset of $S^{(3)}$.

In particular, we deal with independence models closed under graphoid properties. We recall that a graphoid is a couple $(S, \mathcal{I})$, where $\mathcal{I}$ is a ternary relation on the set S, which satisfies the following properties:

G1 if $(A, B, C) \in \mathcal{I}$, then $(B, A, C) \in \mathcal{I}$ (Symmetry);

G2 if $(A, B, C) \in \mathcal{I}$, then $(A, B', C) \in \mathcal{I}$ for any nonempty subset B' of B (Decomposition);

G3 if $(A, B_1 \cup B_2, C) \in \mathcal{I}$ with B_1 and B_2 disjoint, then $(A, B_1, C \cup B_2) \in \mathcal{I}$ (Weak Union);

G4 if $(A, B, C \cup D) \in \mathcal{I}$ and $(A, C, D) \in \mathcal{I}$, then $(A, B \cup C, D) \in \mathcal{I}$ (Contraction);

G5 if $(A, B, C \cup D) \in \mathcal{I}$ and $(A, C, B \cup D) \in \mathcal{I}$, then $(A, B \cup C, D) \in \mathcal{I}$ (Intersection).

$(S, \mathcal{I})$ is a semi–graphoid if it satisfies only the properties G1–G4.
The symmetric versions of rules G2 and G3 are denoted by

G2s if $(A, B, C) \in \mathcal{I}$, then $(A', B, C) \in \mathcal{I}$ for any nonempty subset A' of A;

G3s if $(A_1 \cup A_2, B, C) \in \mathcal{I}$, then $(A_1, B, C \cup A_2) \in \mathcal{I}$.

Let θ, $\theta' \in S^{(3)}$, we denote by

$$\theta \vdash_R \theta'$$

the fact that θ' is obtained by applying once the property R to θ, where in this context R can be G1, G2 or G3.
Moreover, let θ_1, θ_2, $\theta \in S^{(3)}$;

$$\theta_1, \theta_2 \vdash_R \theta$$

denotes that θ is obtained by applying once R to the pair θ_1, θ_2 of triples. In this case R can be either G4 or G5.

Now, we start from a set $J \subset S^{(3)}$ of triples, compatible with an uncertainty measure, and we are interested to establish whether a triple $\theta \in S^{(3)}$ can be derived from J, in symbols

$$J \vdash^* \theta\,.$$

This means that θ can be obtained by applying a finite number of times the rules G1–G5 starting from the set of triples J. This problem is called "implication problem" and has been already studied, for instance, in [32].
A strictly related problem is to compute the closure of a set J, defined as

$$\bar{J} = \{\theta \in S^{(3)} : J \vdash^* \theta\}\,.$$

It is clear that the implication problem can be easily solved once the closure of J has been computed. But the computation of the closure is infeasible because its size is exponentially larger than the size of J.

Then, in the following sections we describe how it is possible to compute a smaller set of triples having the same information as the closure.

This problem has been already faced in [24], with particular attention to semi–graphoid structures.

3 Generalized inference rules

In the following subsections we recall some notions introduced in [2, 4] useful to compute the closure in a more efficient way.

In particular, in Subsection 3.1 a notion of generalized inclusion, that is related to the notion of dominance given in [23] is studied.

In Subsection 3.2 we study some properties of intersection and contraction, which lead to suitable inferential rules. Moreover, we provide a procedure to

compute a "small" set that can be considered a sort of basis for the closure, with respect to graphoid, of a given set of conditional independence statements.

3.1 Generalized inclusion

Let us focus our attention, first of all, to the first three graphoid rules. Given a triple $\theta_2 \in S^{(3)}$, it is possible to compute all the triples θ_1 which can be obtained from θ_2 with a finite number of applications of G1, G2 and G3. We say (see [2, 3, 4]) that, for any such pair of triples, θ_1 is *generalized–included* in θ_2 (briefly g–included), in symbol $\theta_1 \sqsubseteq \theta_2$.

In order to simplify the notation in the following, given a triple $\theta_i = (A_i, B_i, C_i)$, X_i stands for $(A_i \cup B_i \cup C_i)$.

Now, some properties of g–inclusion are recalled.

Proposition 1 *Given $\theta_1 = (A_1, B_1, C_1)$ and $\theta_2 = (A_2, B_2, C_2)$, then $\theta_1 \sqsubseteq \theta_2$ if and only if the following conditions hold*

(i) $C_2 \subseteq C_1 \subseteq X_2$;

(ii) *either* $A_1 \subseteq A_2$ *and* $B_1 \subseteq B_2$ *or* $A_1 \subseteq B_2$ *and* $B_1 \subseteq A_2$.

Generalized inclusion is strictly related to the *partial order relation* $\sqsubseteq_a$ on $S^{(3)}$, defined in [23] and called dominance: the triple $\theta = (A, B, C)$ is said to dominate $\theta' = (A', B', C')$ (in symbol $\theta' \sqsubseteq_a \theta$) if θ' can be derived from θ by means of decomposition, weak union and their symmetric properties (i.e. G2, G3, G2s and G3s).

The relation between $\sqsubseteq$ and $\sqsubseteq_a$ is simple: $\theta' \sqsubseteq \theta$ if and only if

$$\text{either } \theta' \sqsubseteq_a \theta \text{ or } \theta' \sqsubseteq_a \theta^T,$$

where θ^T is the transpose of θ (i.e. if $\theta = (A, B, C)$, then $\theta^T = (B, A, C)$).

The g–inclusion verifies almost all the properties of a partial order relation on $S^{(3)}$ [4], in fact it is reflexive and transitive, but it is not anti–symmetric. However, it satisfies a weak form of anti–symmetry, and denoted by $(AS)^*$:

$\theta_1 \sqsubseteq \theta_2$ and $\theta_2 \sqsubseteq \theta_1$ implies either $\theta_1 = \theta_2$ or $\theta_1 = \theta_2^T$.

The definition of g–inclusion between triples can be extended as follows to the case of sets of triples.

Definition 1 *Let H, J be subsets of $S^{(3)}$. J is a covering of H (in symbol $H \sqsubseteq J$) if and only if for any triple $\theta \in H$ there exists a triple $\theta' \in J$ such that $\theta \sqsubseteq \theta'$.*

The g–inclusion between sets of triples verifies reflexivity and transitivity, while as the following example shows it does not satisfy the anti–symmetry neither in its weak form.

Example 1 *Given $S = \{1, 2, 3, 4\}$, consider the triples $\theta = (\{1\}, \{2\}, \{3\})$, $\theta' = (\{1, 4\}, \{2\}, \{3\}) \in S^{(3)}$ and the subsets $H = \{\theta, \theta'\}$ and $J = \{\theta'\}$ of $S^{(3)}$. It is easy to check that $H \sqsubseteq J$ and $J \sqsubseteq H$, but $\theta \in H$ is such that $\theta \notin J$ and $\theta^T \notin J$.*

However, in [3] we show that weak anti–symmetry holds for particular sets.

3.2 Closure through the generalization of G4 and G5

Now, we recall the two inference rules introduced in [2, 3].

Given $\theta_1, \theta_2 \in S^{(3)}$, $W_C(\theta_1, \theta_2)$ is the set

$$\{\tau : \theta_1', \theta_2' \vdash_{G4} \tau, \text{ with } \theta_1' \sqsubseteq_a \theta_1, \theta_2' \sqsubseteq_a \theta_2\}.$$

Concerning $W_C(\theta_1, \theta_2)$ the following result holds (see [3, 4]).

Proposition 2 *Let $\theta_1 = (A_1, B_1, C_1)$, $\theta_2 = (A_2, B_2, C_2)$ be a pair of triples belonging to $S^{(3)}$, then*

1. *$W_C(\theta_1, \theta_2)$ is not empty if and only if all the following five conditions hold:*
   - *(a)* $A_1 \cap A_2 \neq \emptyset$;
   - *(b)* $C_1 \subseteq X_2$ *and* $C_2 \subseteq X_1$;
   - *(c)* $B_1 \setminus C_2 \neq \emptyset$;
   - *(d)* $B_2 \cap X_1 \neq \emptyset$;
   - *(e)* $|(B_1 \setminus C_2) \cup (B_2 \cap X_1)| \geq 2$.
2. *If $W_C(\theta_1, \theta_2)$ is not empty the triple $gc(\theta_1, \theta_2) =$*

   $$(A_1 \cap A_2, (B_1 \setminus C_2) \cup (B_2 \cap X_1), C_2 \cup (A_2 \cap C_1)),$$

   is in $W_C(\theta_1, \theta_2)$ and dominates any triple belonging to $W_C(\theta_1, \theta_2)$.

When $W_C(\theta_1, \theta_2)$ is empty, we set $gc(\theta_1, \theta_2) = \perp$.

The function $gc(\cdot, \cdot)$ has already been introduced in [24] in an essentially equivalent form.

The conditions *(a)–(e)*, which assure that $W_C(\theta_1, \theta_2)$ is not empty, are however stronger than those given in [24]: in fact, we are looking for the triple dominating all the triples obtained, through G4, from θ_1 and θ_2 or from some of their dominated triples. This is clarified in the next example.

Example 2 *Consider the triples*

$$\theta_1 = (\{1,4\},\{2\},\{3\})$$

and

$$\theta_2 = (\{1,3\},\{2\},\{4\}).$$

The condition (e) fails, since $(B_1 \setminus C_2) = (B_2 \cap X_1)$ *and it contains just the element* 2*.*

Then, in this case $W_C(\theta_1,\theta_2) = \emptyset$*, however it could be noted that by applying G3 to one of the two triples we get* $\theta = (\{1\},\{2\},\{3,4\}) \sqsubseteq_a \theta_i$ *(for* $i = 1, 2$*) and so* θ *adds no further information.*

We denote with $GC(\theta_1,\theta_2)$ the set formed by the possible (i.e. belonging to $S^{(3)}$) triples among $gc(\theta_1,\theta_2)$, $gc(\theta_1,\theta_2^T)$, $gc(\theta_1^T,\theta_2)$ and $gc(\theta_1^T,\theta_2^T)$.

Obviously, $GC(\theta_1,\theta_2)$ is in general different from $GC(\theta_2,\theta_1)$.

Note if $\theta_1,\theta_2 \vdash_{G4} \tau$, then $\tau = gc(\theta_1,\theta_2)$.

A result similar to Proposition 2, related to intersection property, holds (see [3]) by considering the set

$$W_I(\theta_1,\theta_2) = \{\tau : \theta_1',\theta_2' \vdash_{G5} \tau, \text{ with } \theta_1' \sqsubseteq_a \theta_1, \theta_2' \sqsubseteq_a \theta_2\}.$$

Proposition 3 *Let* $\theta_1 = (A_1,B_1,C_1)$*,* $\theta_2 = (A_2,B_2,C_2)$ *be a pair of triples belonging to* $S^{(3)}$*, then*

1. $W_I(\theta_1,\theta_2)$ *is not empty if and only if all the following five conditions hold:*
   (a) $A_1 \cap A_2 \neq \emptyset$*;*
   (b) $C_1 \subseteq X_2$ *and* $C_2 \subseteq X_1$*;*
   (c) $B_1 \cap X_2 \neq \emptyset$*;*
   (d) $B_2 \cap X_1 \neq \emptyset$*;*
   (e) $|(B_1 \cap X_2) \cup (B_2 \cap X_1)| \geq 2$*.*
2. *If* $W_I(\theta_1,\theta_2)$ *is not empty, then the triple* $gi(\theta_1,\theta_2) = (A_{gi},B_{gi},C_{gi})$ *with*
   - $A_{gi} = A_1 \cap A_2$*;*
   - $B_{gi} = (B_1 \cap X_2) \cup (B_2 \cap X_1)$*;*
   - $C_{gi} = (C_1 \cap A_2) \cup (C_2 \cap A_1) \cup (C_2 \cap C_1)$*;*

   is in $W_I(\theta_1,\theta_2)$ *and dominates any triple belonging to* $W_I(\theta_1,\theta_2)$*.*

Given two triples θ_1, θ_2, Proposition 3 gives rise to the dominant triple generated, through G5, by θ_1, θ_2 or by some dominated triples, respectively, by θ_1 and θ_2.

The set $GI(\theta_1,\theta_2)$ is formed by the possible (i.e. belonging to $S^{(3)}$) triples among $gi(\theta_1,\theta_2)$, $gi(\theta_1,\theta_2^T)$, $gi(\theta_1^T,\theta_2)$ and $gi(\theta_1^T,\theta_2^T)$.

Then, $GI(\theta_1,\theta_2) = GI(\theta_2,\theta_1)$.

Also in this case, if $\theta_1,\theta_2 \vdash_{G5} \tau$, then $\tau = gi(\theta_1,\theta_2)$.

The previous sets GC and GI are used to introduce two new inference rules

G4* "generalized contraction": from θ_1,θ_2 deduce any triple $\tau \in GC(\theta_1,\theta_2)$;

G5* "generalized intersection": from θ_1,θ_2 deduce any triple $\tau \in GI(\theta_1,\theta_2)$;

which, as explained above, generalize the two classical inference rules. These rules are useful to compute the closure of a set J of triples in $S^{(3)}$, that is

$$J^* = \{\tau : J \vdash_G^* \tau\} \tag{1}$$

where $J \vdash_G^* \tau$ means that τ is obtained by applying a finite number of times the rules G4* and G5*.

In [3, 4] the relationship between the two closures J^* and $\bar{J}$ is studied, in particular, we prove that any triple obtained through G1–G5 is g–included in a triple deduced from G4* and G5*. This implies that $J^* \subseteq \bar{J}$ and moreover

$$\bar{J} \sqsubseteq J^*.$$

Note that J^* is a subset of $\bar{J}$, so even if J^* has the same information of $\bar{J}$, is smaller than $\bar{J}$. Actually, J^* contains some "redundant" triples, that means that are g–included in some of the other ones. In fact, (see (1)) each application G4* and G5* can generate a triple which is g–included in a triple of J or in an already generated triple.

3.3 Fast closure

In [2, 3, 4] we introduced the concept of "maximal"(with respect to g–inclusion) triple: given a set J of triples, a triple τ is maximal in J if there exists no $\bar{\tau} \in J$ with $\bar{\tau} \neq \tau, \tau^T$ such that $\tau \sqsubseteq \bar{\tau}$.

We denote with $J_{/\sqsubseteq}$ the subset of J composed only by its maximal triples and we call FindMaximal the function which computes $J_{/\sqsubseteq}$ from J.

There is no loss of information by using $J_{/\sqsubseteq}$ instead of J [3], in fact

$$J \sqsubseteq J_{/\sqsubseteq}.$$

Then, given a set J of triples in $S^{(3)}$, we compute J^* (see equation (1)) and then we take only its maximal triples, i.e. $J^*_{/\sqsubseteq}$.

We call the set $J^*_{/\sqsubseteq}$ "fast closure" and we denote it, for simplicity, with J_*.

Note that we have also the following relationship: $J_* \subseteq \bar{J}$ and

$$\bar{J} \sqsubseteq J_*.$$

It is interesting to observe $\bar{J}_{/\sqsubseteq}$ and J_* essentially coincide [3], in fact

$$\bar{J}_{/\sqsubseteq} \sqsubseteq J_* \qquad \text{and} \qquad J_* \sqsubseteq \bar{J}_{/\sqsubseteq}.$$

4 Unique inference rule

In [3, 4] we describe a procedure to compute efficiently the closure of a set of conditional independence statements, which is based on the two above inferential rules (generalized contraction and intersection). In order to improve such procedure, in we look for a unique inferential rule with the aim of simplifying the procedure.

In particular, by taking into account Proposition 2 and Proposition 3, which provide necessary and sufficient conditions for the application of generalized contraction and intersection, respectively, the notion of almost complete pair of triples is introduced in [4] in order to characterize the couples of triples which lead to the largest fast closure.

We recall first of all that the fast closure $\{\theta_1, \theta_2\}_*$ of a couple $\theta_1, \theta_2 \in S^{(3)}$ is composed by a maximum of nine extra triples, no matter how many variables occur in θ_1 and θ_2.

In particular, any pair of triples (θ_1, θ_2) can be re–written, in a general form, as

$$\begin{aligned} \theta_1 &= ([A_A, A_B, A_C, A_N], [B_A, B_B, B_C, B_N], \\ &\quad [C_A, C_B, C_C, C_N]) \\ \theta_2 &= ([A_A, B_A, C_A, A'_N], [A_B, B_B, C_B, B'_N], \\ &\quad [A_C, B_C, C_C, C'_N]) \end{aligned}$$

where some sets can be empty and with the notation that $[A, B, C]$ stands for $A \cup B \cup C$.

Each triple of the fast closure of (θ_1, θ_2) is g–included in the set of possible (i.e. belonging to $S^{(3)}$) triples

$$K(\theta_1, \theta_2) = \{\theta_1, \theta_2, \theta_a, \theta_b, \theta_c, \theta_d, \theta_e, \theta_f, \theta_g, \theta_h, \theta_{ad}\}$$

where

$\theta_a = (A_A, [A_B, B_A, B_B, B_C, C_B, B_N], [A_C, C_A, C_C])$;

$\theta_b = (A_B, [A_A, B_A, B_B, B_C, C_A, B_N], [A_C, C_B, C_C])$;

$\theta_c = (B_A, [A_A, A_B, A_C, B_B, C_B, A_N], [B_C, C_A, C_C])$;

$\theta_d = (B_B, [A_A, A_B, A_C, B_A, C_A, A_N], [B_C, C_B, C_C])$;

$\theta_e = (A_A, [A_B, B_A, B_B, B_C, C_B, B'_N], [A_C, C_A, C_C])$;

$\theta_f = (A_B, [A_A, B_A, B_B, B_C, C_A, A'_N], [A_C, C_B, C_C])$;

$\theta_g = (B_A, [A_A, A_B, A_C, B_B, C_B, B'_N], [B_C, C_A, C_C])$;

$\theta_h = (B_B, [A_A, A_B, A_C, B_A, C_A, A'_N], [B_C, C_B, C_C])$;

$\theta_{ad} = ([A_B, B_A], [A_A, B_B], [A_C, B_C, C_A, C_B, C_C])$.

Therefore,

$$\{\theta_1, \theta_2\}_* \sqsubseteq K(\theta_1, \theta_2).$$

Moreover, in [3, 4] it is also proved that

$$K(\theta_1, \theta_2) \sqsubseteq \{\theta_1, \theta_2\}_*.$$

Note that in general $K(\theta_1, \theta_2)$ may not coincide with $\{\theta_1, \theta_2\}_*$ because it could contain some redundant triple or the transpose triple of one belonging to $\{\theta_1, \theta_2\}_*$.

However, it is easy to see that

$$K(\theta_1, \theta_2)_{/\sqsubseteq} \sqsubseteq \{\theta_1, \theta_2\}_*$$

and

$$\{\theta_1, \theta_2\}_* \sqsubseteq K(\theta_1, \theta_2)_{/\sqsubseteq},$$

since both sets are maximal.

Therefore the set $K(\theta_1, \theta_2)$ allows to compute $\{\theta_1, \theta_2\}_*$: in fact, it is possible to build up such a set and apply the function FindMaximal to it.

All this computation requires a constant number of steps with respect to the size of θ_1, θ_2.

By using $\{\theta_1, \theta_2\}_*$, it is possible to provide a new inference rule

U : from θ_1, θ_2 deduce any triple $\tau \in \{\theta_1, \theta_2\}_*$.

4.1 Algorithm FC1

By using the unique inference rule U, we provided the Algorithm 1.

Concerning the above algorithm we have the following result:

Theorem 1 *Let J be a nonempty subset of $S^{(3)}$, then*

1. *FC1(J) $\sqsubseteq J_*$;*
2. *$J_* \sqsubseteq$ FC1(J).*

Both theoretical and empirical comparisons between FC1 and an algorithm based on two inferential rules in [4] are carried out, hereby showing the better performances of FC1.

Algorithm 1 Fast closure by U

1: **function** FC1(J)
2: $J_0 \leftarrow J$
3: $N_0 \leftarrow J$
4: $k \leftarrow 0$
5: **repeat**
6: $k \leftarrow k + 1$
7: $N_k := \bigcup_{\theta_1 \in J_{k-1}, \theta_2 \in N_{k-1}} \{\theta_1, \theta_2\}_*$
8: $J_k \leftarrow \text{FindMaximal}(J_{k-1} \cup N_k)$
9: **until** $J_k = J_{k-1}$
10: **return** J_k
11: **end function**

Note that FC1 can be optimized by observing that if θ'_1 and θ'_2 belong to $\{\theta_1, \theta_2\}_*$, then $\{\theta'_1, \theta'_2\}_*$ is g–included to $\{\theta_1, \theta_2\}_*$. The validity of this observation follows easily since

$$\{\theta'_1, \theta'_2\}_* \sqsubseteq \{\theta'_1, \theta'_2\}^* \sqsubseteq \{\theta_1, \theta_2\}^* \sqsubseteq \{\theta_1, \theta_2\}_*.$$

Therefore, it is not necessary to apply the inference rule U to a pair of triples θ'_1 and θ'_2, generated by U from the same two triples θ_1 and θ_2, since from θ'_1 and θ'_2 we would obtain only redundant triples, which would be discarded by the function FindMaximal.

Note that for the same reasons, we do not need to apply the rule U between a triple θ and another one θ' generated from θ (by combining θ with another triple θ''): in fact if $\theta' \in \{\theta, \theta''\}_*$, then $\{\theta, \theta'\} \subseteq \{\theta, \theta''\}^*$ and so

$$\{\theta, \theta'\}_* \sqsubseteq \{\theta, \theta''\}_*,$$

which implies that no maximal triple can be obtained.

Then, the use of the inference rule U in FC1 can be enhanced by keeping track of the "parents" of each triple and by neglecting the pairs which satisfies the two previously described situations ("sibling" triples and "father–child").

In our implementation, we use this optimization, but we consider $K(\theta_1, \theta_2)$ instead of $\{\theta_1, \theta_2\}_*$, because in any case in each cycle of FC1 a call to function FindMaximal is however performed.

5 Experimental results

In this section we describe some experimental results obtained with an implementation in C++ of the algorithm FC1, as well as an implementation of an algorithm to compute the complete closure (with respect to G1–G5). The main purpose of these experiments is to prove the viability of the fast closure computation.

The first aspect, that these experiments can clarify, is to show how difficult it is, from the computational point of view, to compute the fast closure. It is clear that this problem is a computationally hard problem, for which no efficient (i.e. polynomial time) solution can exist as already noted in [23, 24].

Therefore an empirical evaluation is necessary in order to establish whether the computation of the fast closure is reasonably fast and uses an acceptable amount of memory.

The other question is which is the quantitative difference in size and in computation time of the fast closure with respect to the complete closure. The fast closure is clearly smaller than the complete closure (each triple $\theta \in J_*$ corresponds to several triple in $\bar{J}$), but we have not been able to find any theoretical bounds for the size of J_* with respect to the size of $\bar{J}$.

The experiments were performed on an AMD Dual Core Opteron running at 1.8 GHz with 2 GByte main memory. We applied a cut–off of 5,000,000 triples that can be stored (to avoid problems with memory) and a time–out of 3600 seconds. Some preliminary results, with different experimental parameters, have already been given in [6, 2].

In the first set of experiments, we have generated 200 random sets of triples having nv variables and nr triples, for $nr = 10, 15, 20, 25, 30$ and $nv = \lfloor 0.5 \cdot nr \rfloor, nr, \lfloor 1.5 \cdot nr \rfloor, 2nr$. and we have computed the fast closure by means of (see Table 1).

In the Table 1, the value *perc* is the percentage of the sets for which FC1 has been able to compute the fast closure, within the limits of time and memory, *time* is the average computation times in seconds, *size* is the average size of the fast closure, *iter* is the average number of iterations needed to find the closure, and *gen* is the average number (rounded to the nearest integer) of the overall generated triples.

The behavior of FC1, as explained in the following, is influenced by many factors, which can have contradictory and not well understandable effects. However it is possible to observe that as nr grows, instances with a small value for $\frac{nv}{nr}$ become more and more difficult: with $nr = 30$ and $nv = 15$ FC1 has not been able to solve any instance. The same happens with $nr = 35$ and 40 (nv being $0.5 \cdot nr$), in experimental tests not described here.

On the other hand, when the ratio $\frac{nv}{nr}$ is large, instances get easier and easier to solve.

The first behavior can be explained with the fact that generating at random an instance with fewer variables, with respect to the number of relations, can produce many triples to which it is possible to repeatedly apply the generalized inference rules. In these

Table 1: Fast Closure FC1

nr	nv	perc	time	size	iter.	gen.
10	5	100	0	10.83	3.99	202
10	10	100	1.06	95.93	6.42	27524
10	15	99	44.43	226.08	6.263	241219
10	20	98.5	22.16	153.54	4.81	115006
15	7	100	9.11E-02	46.84	5.50	5841
15	15	63	500.42	982.68	10.03	1926990
15	22	80.5	111.49	365.29	6.63	359213
15	30	98	9.77	72.14	3.25	32615
20	10	100	79.19	433.835	7.41	652608
20	20	27.5	376.43	921.47	10.2	1105693
20	30	93.5	84.64	305.21	5.58	240052
20	40	98.5	3.64	54.95	2.20	16514
25	12	49.5	1383.23	1354.33	8.3	5231558
25	25	35	254.46	719.69	9.04	720993
25	37	97.5	14.25	124.42	3.8	62761
25	50	100	1.1E-03	29.685	1.445	84
30	15	0	–	–	–	–
30	30	51.28	118.59	514.58	7.65	3631898
30	45	100	0.03	48.38	2.41	1063
30	60	100	8.55E-05	31.06	1.12	7

cases, the computation of the fast closure requires several iterations during which a large number of triples are generated (most of them are discarded). These kinds of instances seem to be the hardest to solve, if compared to the other kinds.

At the same time, if the number of variables is too large, the chance of application of the inference rules becomes very low, as proved by the average size of the fast closure (which is roughly similar to nr) and the number of generated triples (which is rather small). In these cases, the closure often coincides or is similar to the initial set of triples and therefore can be computed with a little computational effort.

In the second set of experiments we compare the computation time needed for finding the complete closure and its size with respect to the time and size of the fast closure. The complete closure is obtained by using an algorithm similar to FC1, which uses all the inference rules G1–G5, without calling FindMaximal. Furthermore, we did not apply for it any cut–off with respect to number of triples.

Since we expect that the complete closure is much larger than its fast version, we have performed these new experiments with smaller instances, instead of using the previous one. In particular, we generate 20 sets of nr triples and nv variables, for $nr = 4, 7, 10$ and $nv = nr, \lfloor 1.5 \cdot nr \rfloor$.

In Table 2 the results for the fast closure are reported, with the average values calculated with respect to the solved instances by FC1, the average computation time is negligible, except that in the last row, where we obtain results similar in magnitude order, as those displayed in Table 1. The algorithm FC1 has been able to build the closure for each instance.

Table 2: Fast Closure with FC1

nr	nv	time	size	iter.	gen.
4	4	0	3.95	2.75	12.1
4	6	0	5.85	2.95	29.2
7	7	2E-03	18.65	4.95	559.25
7	10	1.8E-02	32.05	4.7	1756.15
10	10	0.6755	86.9	5.95	18415
10	15	42.7225	320.45	6.7	335910.5

In Table 3 we report the results obtained in the computation of the complete closure. The last column contains the number of instances for which the algorithm has been able to compute the complete closure within an hour of computation. Note that with $nr = 10$ and $nv = 15$ we could solve only one instance, which almost reached the time limit, while the fast closure of this instance has only 27 triples and has been found in a negligible amount of time. The values in the last column are used to compute the average values showed in Table 2.

The comparison of the size between fast and complete closure is impressive, as it is possible to see in the graph of Figure 1 (the last rows of both tables have been ignored).

Table 3: Complete Closure

nr	nv	time	size	iter.	gen.	res.
4	4	0	64	7	57	20
4	6	0.05	527	8.9	899	20
7	7	1.75	3282	13.15	9526	20
7	10	248	28808	13.89	147249	19
10	10	603	50760	16.67	268381	15
10	15	3513	159164	14	683991	1

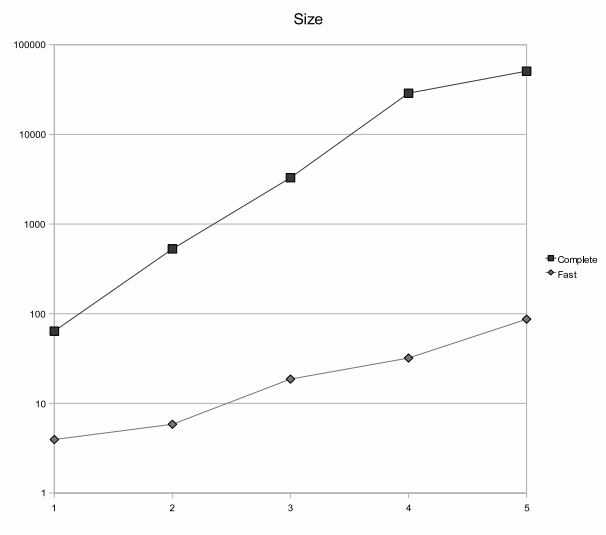


Figure 1: Sizes of the closure

Clearly also the computation times for computing the complete closure are much higher than the time needed to compute the fast closure, as displayed in the Figure 2.

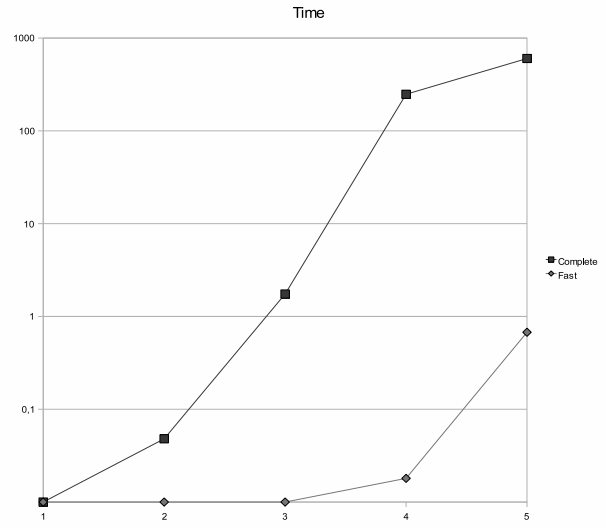


Figure 2: Computation times

6 Conclusions

We study some properties of graphoid structures with the aim to compute efficiently the closure of a set of conditional independence statements. It is well known that the size of the closure of a set is exponentially greater than the size of the given set.

In particular, we give an algorithm FC1, which is able to compute the closure of a set of triples by looking for a suitable subset of the closure, that has the same information, but it is smaller than the closure as shown by experimental results. Actually, FC1 computes just the maximal (with respect to g–inclusion) triples, then it also allows to improve the computational time.

By means of this set also the well known implication problem can be solved in an efficient way: in fact, to verify whether a triple belongs to the closure it is enough to look for a triple in the set, obtained through FC1, which g–includes the given triple. Moreover, to check the g–inclusion relation requires constant time, therefore the computational time is linear with respect to the size of the set.

A straightforward extension of this work is to adapt this framework for computing the closure by using semi-graphoid axioms and compare it with that proposed in [24].

A further open problem, partially studied in [5], consists into using this set for building in an efficient way an acyclic directed graph representing the independence statements in the closure.

References

[1] B.N. Amor, S. Benferhat. Graphoid properties of qualitative possibilistic independence relations. *Inter. Jour. Uncertainty Fuzziness Knowledge-Based Systems*, 13(1):59–96, 2005.

[2] M. Baioletti, G. Busanello, B. Vantaggi. Algorithms for the closure of graphoid structures *Proc. of 12th Inter. Conf. IPMU 2008*, Malaga, 899-906, 2008.

[3] M. Baioletti, G. Busanello, B. Vantaggi. Conditional independence structure and its closure: inferential rules and algorithms. In *Technical Report*, 5/2009 of University of Perugia.

[4] M. Baioletti, G. Busanello, B. Vantaggi. Conditional independence structure and its closure: inferential rules and algorithms. *International Journal of Approximate Reasoning*, in press doi: 10.1016/j.ijar.2009.05.002.

[5] M. Baioletti, G. Busanello, B. Vantaggi. Acyclic Directed Graphs to Represent Conditional Independence Models. *Proc. of ECSQARU 2009. Lecture Notes* LNAI 5590, 530541, 2009.

[6] G. Busanello (2008). Probabilistic conditional independence models: closure and construction of acyclic directed graphs. PhD Thesis.

[7] G. Coletti, R. Scozzafava. Zero probabilities in stochastical independence. In *Information, Uncertainty, Fusion, Kluwer Academic Publishers, Dordrecht*, B. Bouchon- Meunier, R.R. Yager, L.A. Zadeh (Eds.), 185–196, 2000.

[8] G. Coletti, R. Scozzafava. *Probabilistic logic in a coherent setting.* Dordrecht/Boston/London: Kluwer (Trends in logic n.15), 2002.

[9] G. Coletti, B. Vantaggi. Possibility theory: conditional independence. *Fuzzy Sets and Systems*, 157:1491–1513, 2006.

[10] I. Couso, S. Moral, P.Walley. A survey of concepts of independence for imprecise probabilities. *Risk, Decision and Policy*, 5:165–181, 2000.

[11] R.G. Cowell, A.P. Dawid, S.L. Lauritzen, D.J. Spiegelhalter. Probabilistic Networks and Expert Systems, Springer-Verlag, New York, 1999.

[12] F.G. Cozman, T. Seidenfeld. Independence for full conditional measures, graphoids and Bayesian networks, *Boletim BT/PMR/0711 Escola Politecnica da Universidade de Sao Paulo*, Sao Paulo, Brazil (To appear at Benedikt Lwe, Eric Pacuit, Jan-Willem Romeijn (eds.), Foundations of the Formal Sciences VI - Reasoning about Probabilities and Probabilistic Reasoning), 2007.

[13] F. G. Cozman, P. Walley. Graphoid properties of epistemic irrelevance and independence. *Annals of Mathematics and Artificial Intelligence*, 45: 173–195, 2005.

[14] A.P. Dawid. Conditional independence in statistical theory. *J. Roy. Stat. Soc. B*, 41:15–31, 1979.

[15] A.P. Dawid. Separoids: a mathematical framework for conditional independence and irrelevance. Representations of uncertainty. *Annals of Mathematics and Artificial Intelligence*, 32:335–372, 2001.

[16] L.M. de Campos, J.F. Huete. Independence concepts in possibility theory. I. *Fuzzy Sets and Systems*, 103:127–152, 1999.

[17] L. de Campos, S. Moral. Independence Concepts for Convex Sets of Probabilities", *Proc. XI Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco, CA, 108–115, 1995.

[18] S.L. Lauritzen. *Graphical models.* Clarendon Press, Oxford, 1996.

[19] S. Moral, A. Cano. Strong conditional independence for credal sets. *Annals of Mathematics and Artificial Intelligence*, 35:295–321, 2002.

[20] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann, Los Altos, CA, 1988.

[21] E. San Martìn, M. Mouchart, J. Rolin (2005). Ignorable common information, null sets and Basu's first theorem. *Sankhya*, 67:674–698, 2005.

[22] P.P. Shenoy. Conditional independence in valuation-based systems. *International Journal of Approximate Reasoning*, 10:203-234, 1994.

[23] M. Studený. Semigraphoids and structures of probabilistic conditional independence. *Ann. Math. Artif. Intell.*, 21:71–98, 1997.

[24] M. Studený. Complexity of structural models. *Proc. Prague Stochastics '98*, Prague, 521–528, 1998.

[25] M. Studený, R.R. Bouckaert (1998). On chain graph models for description of conditional independence structures. *Ann. Statist.*, 26(4):1434–1495, 1998.

[26] B. Vantaggi. Conditional independence in a coherent setting. *Ann. Math. Artif. Intell.*, 32:287–313, 2001.

[27] B. Vantaggi. Conditional independence structures and graphical models. *Int. J. Uncertain. Fuzziness Knowledge-Based Systems*, 11(5): 545–571, 2003.

[28] B. Vantaggi. Qualitative Bayesian networks with logical constraints. *Lecture Notes in Computer Sciences*, 2711, Springer, New York, 100–112, 2003.

[29] B. Vantaggi. The role of coherence for handling probabilistic evaluations and independence. *Soft Computing*, Springer, Berlin / Heidelberg, 69(8):617–628, 2005.

[30] J. Veinarova. Conditional Independence Relations in Possibility Theory. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 8(3):253–269, 2000.

[31] J. Witthaker. *Graphical models in applied multivariate statistic*, Wiley, New York, 1990.

[32] S.K.M. Wong, C.J. Butz, D. Wu. On the Implication Problem for Probabilistic Conditional Independency. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 30(6):785–805, 2000.

Category Selection for Multinomial Data

Rebecca M. Baker
Department of Mathematical Sciences
University of Durham, England
r.m.baker@durham.ac.uk

Frank P.A. Coolen
Department of Mathematical Sciences
University of Durham, England
frank.coolen@durham.ac.uk

Abstract

A new method is presented for selecting a single category or the smallest subset of categories, based on observations from a multinomial data set, where the selection criterion is a minimally required lower probability that (at least) a specific number of future observations will belong to that category or subset of categories. The inferences about the future observations are made using an extension of Coolen and Augustin's nonparametric predictive inference (NPI) model to a situation with multiple future observations.

Keywords. imprecise probability, predictive inference, categorical data, selection

1 Introduction

Selection is a wide-ranging topic in statistics for choosing the optimal member(s) of some group. This group may be, for example, a set of data categories or a range of data sources. With regard to multinomial data, interest may be in choosing the category that has the largest probability of occurrence. Existing methods for this type of selection [2] are all non-predictive, i.e. the selection of the optimal category is based solely on hypothesis testing and does not use any type of predictive inference.

NPI for learning from multinomial data in the absence of prior knowledge has been developed by Coolen and Augustin [1, 6, 7]. The model gives predictive inferences about a single future observation in the form of probability intervals $P = [\underline{P}, \overline{P}]$. Throughout this paper, P denotes interval probability, which we often just call 'probability'. When an explicitly precise probability is used, it is denoted by p. NPI is based on a probability wheel representation of the data, where each category is represented by a segment of the wheel.

Selection methods based on NPI have been developed by Coolen and van der Laan [3] and Coolen and Coolen-Schrijner [4, 5]. These methods use predictive inferences which are based on past observations, and make use of Hill's assumption A_n [9].

Coolen and van der Laan [3] developed an NPI selection method for real-valued data from k different sources. Their objective was to select the source which would provide the largest next observation. Probabilities were determined for the event that the next observation from one source would exceed the next observation from all other sources. They also considered two ways of selecting a subset of sources: first, they determined the interval probability that some subset would contain the source providing the largest next observation, and second, they found the interval probability that the next observations from every source in some subset would all exceed the next observations from the remaining sources.

Coolen and Coolen-Schrijner [4, 5] developed an NPI selection method for Bernoulli data from k different groups. Their objective was to select the group which would have the highest number of future successes. Here, inferences were made about m future observations rather than just the next observation. Subsets of the groups were also considered [4], and probabilities were presented for the event that some subset contains the group which has the most future successes and for the event that all groups in some subset will have more future successes than every other group.

In this paper, we discuss the use of NPI for selection from a multinomial data set. We consider selection of a single optimal category, and selection of an optimal subset of categories, where we define the optimal subset to be the subset which satisfies the required probability criterion, is of minimal size and has the largest lower probability amongst all subsets of the same size.

2 Predictive category selection

We develop NPI for category selection from a multinomial data set. We have K possible categories, labelled $c_1, ..., c_K$, and our aim is to select the category with the largest probability of occurrence. Suppose that we have a data set consisting of n observations, and let $n_1, ..., n_K$ denote the number of observations in categories $c_1, ..., c_K$ respectively. We consider m future observations, and select a category based on predictive inferences about these m observations. These inferences will be made by using and adapting the general theory of nonparametric predictive inference for multinomial data [1, 6, 7], discussed previously. Let the vector of random quantities $(M_1, ..., M_K)$ denote the number of the m future observations that belong to categories $c_1, ..., c_K$, such that $\sum_{j=1}^{K} M_j = m$.

2.1 One future observation

The simplest case is where $m = 1$, so inference is about one future observation. We may want to select a single category with the largest probability of occurrence. According to the NPI model [7], the lower and upper probabilities that the future observation will belong to category c_j are

$$\underline{P}(M_j = 1) = (\frac{n_j - 1}{n})^+,$$

where $(x)^+$ denotes $\max\{x, 0\}$, and

$$\overline{P}(M_j = 1) = \min\{\frac{n_j + 1}{n}, 1\}.$$

The above formulae are derived through the use of the probability wheel model [6], as illustrated in the example below. We can evaluate these probabilities for each of the possible categories and then select the category with the highest probability.

Example 2.1. *Suppose that our possible categories are blue (B), red (R), yellow (Y) and green (G). Our data set consists of 8 observations: 3 B, 2 G, 2 Y and 1 R. We want to select a single category with the highest probability that the next observation will be in that category. First, we find the probability that the next observation will be blue. Let n_B denote the number of B observations in the data set, and let M_B denote the number of future B observations. The minimum number of slices of the wheel that we can assign to B is equal to $n_B - 1 = 3 - 1 = 2$. This leads to the lower probability $\underline{P}(M_B = 1) = \frac{n_j - 1}{n} = \frac{2}{8}$.*

The maximum number of slices of the wheel that we can assign to B is equal to $n_B + 1 = 3 + 1 = 4$. This leads to the upper probability $\overline{P}(M_B = 1) = \frac{n_j + 1}{n} = \frac{4}{8}$.

We then carry out the same process for the other categories, and we find that $P(M_Y = 1) = P(M_G = 1) = [\frac{1}{8}, \frac{3}{8}]$, and $P(M_R = 1) = [0, \frac{2}{8}]$. So we select the blue category.

Theorem 2.1. *When $m = 1$, and we want to select a single category with the largest probability of occurrence, it is always optimal to choose the category which has the greatest number of observations in the data set.*

Proof. We select the category with the highest probability $P(M_j = 1)$, where $P(M_j = 1) = [\frac{n_j - 1}{n}, \frac{n_j + 1}{n}]$, so it is optimal to select the category with the largest value of n_j. □

2.2 Multiple future observations

Whereas Coolen and Augustin [6, 7] only considered one future observation, we now consider inferences about multiple future observations, so $m > 1$. Suppose that our data set is represented on a probability wheel, and the n slices on the wheel are numbered 1 to n. Each of our m future observations must fall into one of these n slices. Let the vector $(S_1, ..., S_n)$ denote the number of future observations which fall into slices 1 to n, respectively. The total number of different arrangements of these m observations is $\binom{n+m-1}{m}$ [8], which leads to the precise probability for a particular arrangement

$$p(\bigcap_{j=1}^{n} \{S_j = s_j\}) = \binom{n + m - 1}{m}^{-1}$$

where $s_j \geq 0$ and $\sum_{j=1}^{n} s_j = m$.

More generally, the total number of different arrangements of f future observations within a segment made up of $S + 1$ observations is equal to

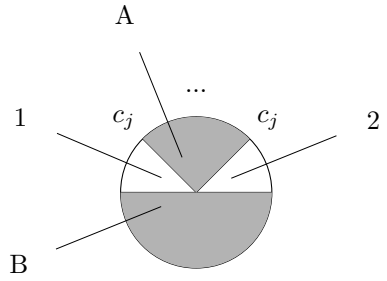
$$\binom{(S - 1) + f}{f}. \tag{1}$$

This is because there are $S - 1$ existing observations within the interior of such a segment, and so we are considering the number of arrangements of f future observations amongst a total of $(S - 1) + f$ observations.

Consider the general case where m may take any value. We want to find the probability that a certain proportion of these m future observations is in some category c_j. We may wish to specify a particular number of observations, in which case the event of interest will be $M_j = m_j$ for some $m_j \leq m$. We may also wish to specify a threshold for M_j, corresponding to the event $M_j \geq m_j$ for some $m_j \leq m$.

2.2.1 Deriving $\mathbf{P(M_j = m_j)}$

We can use NPI to find the probabilities that precisely m_j of the m future observations will belong to category c_j. The bounds derived here are the most conservative bounds achievable within the NPI framework, due to the way in which the slices of the wheel are assigned to categories. This is explained below. The diagram illustrates the relevant segments of the wheel.



It is assumed throughout this section that $1 < n_j < n - 1$. In the case $n_j \leq 1$, we are not forced to assign any slices of the wheel to c_j, leading to $\underline{P}(M_j = m_j) = 0$.

The shaded segment A represents all slices which must be assigned to c_j. There are $n_j - 1$ such slices. By (1), the number of different arrangements of m_j future observations within this segment is $\binom{n_j - 2 + m_j}{m_j}$.

The shaded segment B represents all slices which must be assigned to a category other than c_j. There are $n - n_j - 1$ such slices. By (1), the number of different arrangements of $m - m_j$ future observations within this segment is $\binom{n - n_j - 2 + (m - m_j)}{m - m_j}$.

Multiplying these two binomial coefficients gives us the minimum number of arrangements in which m_j future observations are in c_j, showing that the lower probability is equal to

$$\underline{P}(M_j = m_j) = \binom{n + m - 1}{m}^{-1} \binom{n_j - 2 + m_j}{m_j} \times \binom{n - n_j - 2 + (m - m_j)}{m - m_j}. \quad (2)$$

This general formula is applicable to any positive integers m and m_j such that $m_j \leq m$.

We can also find the equivalent upper probability. We now want to maximise the number of arrangements of the m future observations in which m_j future observations are in c_j. There are $n_j + 1$ slices of the wheel which we can allocate to category c_j, including two slices which we may or may not assign to c_j, which we will term 'optional slices' (labelled 1 and 2 in the diagram above).

As in the case of lower probability, we count all arrangements where m_j observations fall in segment A and $m - m_j$ observations fall in segment B. We showed previously that there are $\binom{n_j - 2 + m_j}{m_j}\binom{n - n_j - 2 + (m - m_j)}{m - m_j}$ such arrangements. However, we now also consider the two optional slices on the wheel. Any observations which fall in one of the optional slices may be counted either as belonging to c_j or as not belonging to c_j. This means that to find the upper probability we need to count any arrangement with one or more observations in the optional slices.

Let T denote the total number of future observations in the optional slices, where T ranges from 1 to m. For $T = 1$, there are two possible arrangements, as the observation could fall either in slice 1 or in slice 2. By similar reasoning, for $T = 2$, there are three possible arrangements. In general, there are $T + 1$ possible arrangements for each value of T.

However, there are a number of different orderings that give T observations in the optional slices. Let X be a non-negative integer such that $X \leq m_j$ and $T - X \leq m - m_j$. Then, we may have $m_j - X$ observations in segment A, $(m - m_j) - (T - X)$ observations in segment B, and T observations in the optional slices, where X ranges from $T - (m - m_j)$ to m_j. Therefore, the total number of arrangements with one or more observations in the optional slices is equal to

$$\sum_{T=1}^{m} \sum_{X=\{T-(m-m_j)\}^+}^{\min\{m_j,T\}} (T+1)\binom{n_j - 2 + (m_j - X)}{m_j - X} \times \binom{n - n_j - 2 + (m - m_j) - (T - X)}{m - m_j - (T - X)}.$$

This enables us to find the maximum number of different arrangements of the m future observations in which m_j observations are in c_j, leading to the upper probability

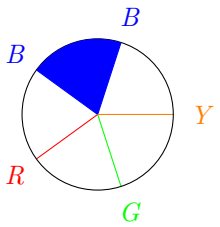
$$\overline{P}(M_j = m_j) = \binom{n + m - 1}{m}^{-1} [\binom{n_j - 2 + m_j}{m_j} \times \binom{n - n_j - 2 + (m - m_j)}{m - m_j} + \sum_{T=1}^{m} \sum_{X=\{T-(m-m_j)\}^+}^{\min\{m_j,T\}} \times (T+1)\binom{n_j - 2 + (m_j - X)}{m_j - X} \times \binom{n - n_j - 2 + (m - m_j) - (T - X)}{m - m_j - (T - X)}]. \quad (3)$$

Again, this formula holds for any positive integers m and m_j such that $m_j \leq m$. As before, it is assumed here that $n_j \geq 2$. An unobserved category can be assigned at most one slice of the wheel, leading to $\overline{P}(M_j = m_j) = \binom{n+m-1}{m}^{-1}\binom{n-n_j-2+m-m_j}{m-m_j}$. In the case $n_j = 1$, the formula reduces to

$$\overline{P}(M_j = m_j) = \binom{n+m-1}{m}^{-1}(m_j+1) \times \binom{n-n_j-2+m-m_j}{m-m_j}.$$

In the case $n_j \geq n-1$, every slice on the wheel may be assigned to category j and furthermore there is only one optional slice.

Example 2.2. *Suppose that our possible categories are blue (B), red (R), yellow (Y) and green (G). Our data set consists of 5 observations as shown on the probability wheel below.*



We want to make inferences about 3 future observations, and we want to find the probability that precisely two of these are blue. To find the lower probability, we use (2) with $m_B = 2$. Using the values $n = 5$, $m = 3$ and $n_j = 2$, this gives

$$\underline{P}(M_B = 2) = \frac{1}{35}\binom{2}{2}\binom{2}{1} = \frac{2}{35}.$$

To find the upper probability, we use (3) with $m_B = 2$. This gives

$$\overline{P}(M_B = 2) = \frac{1}{35}[2+2+4+3+6+4] = \frac{21}{35}.$$

So we see that $P(M_B = 2) = [\frac{2}{35}, \frac{21}{35}]$.

Theorem 2.2. *For general m, when selecting the category which has the largest lower or upper probability of containing all of the future observations, it is optimal to select the category with the greatest number of observations.*

Proof. The general formulae for the lower probability (2) and upper probability (3) can be simplified in the case $M_j = m$, because in this case $m - m_j = 0$ and also the only possible value of X in the summation is T, leading to $T - X = 0$. We find that

$$\underline{P}(M_j = m) = \binom{n+m-1}{m}^{-1}\binom{n_j-2+m}{m}$$

and

$$\overline{P}(M_j = m) = \binom{n+m-1}{m}^{-1}[\binom{n_j-2+m}{m} + \sum_{T=1}^{m}\binom{n_j-2+(m-T)}{m-T}].$$

The values of n, m and T do not depend on the category selected, and since these lower and upper probability formulae are both increasing in n_j, it is always optimal to select the category with the largest value of n_j, ie. the greatest number of data observations. □

It is also of interest to investigate which value of n_j will maximise the lower probability $\underline{P}(M_j = m_j)$. We will henceforth call this value n_j^*. Plotting $\underline{P}(M_j = m_j)$ against values of n_j ranging from 1 to n shows the graph to be monomodal with a smooth line of best fit. Intuitively, we expect that the peak will occur near to $n_j = \frac{nm_j}{m}$, because it seems natural that the proportion of the future observations which are in c_j should be similar to the proportion of the data observations that are in c_j. We will now formally assess which value of n_j gives the maximal lower probability.

Theorem 2.3. *For general m, the value of n_j which will maximise $\underline{P}(M_j = m_j)$ is the integer which lies in the interval $[1 + \frac{m_j}{m}(n-3), 2 + \frac{m_j}{m}(n-3)]$.*

Proof. The proof follows from considering the two ratios

$$\frac{\underline{P}(M_j = m_j|n_j)}{\underline{P}(M_j = m_j|n_j+1)}$$

and

$$\frac{\underline{P}(M_j = m_j|n_j)}{\underline{P}(M_j = m_j|n_j-1)}.$$

□

To see whether this result corresponds to our initial prediction, we check whether $\frac{nm_j}{m}$ lies in this interval, as shown below.

$$1 + \frac{m_j}{m}(n-3) \leq \frac{nm_j}{m} \iff m_j \geq \frac{1}{3}m$$

$$\frac{nm_j}{m} \leq 1 + \frac{m_j}{m}(n-3) \iff m_j \leq \frac{2}{3}m$$

We see that if $\frac{1}{3}m \leq m_j \leq \frac{2}{3}m$, then $\frac{nm_j}{m}$ will indeed be within the interval. We can also show that if $m_j < \frac{1}{3}m$, then $\frac{nm_j}{m} + 1$ is within the interval, meaning that $\frac{nm_j}{m}$ is just to the left of the interval. Similarly, if $m_j > \frac{2}{3}m$, then $\frac{nm_j}{m} - 1$ is within the interval, meaning that $\frac{nm_j}{m}$ is just to the right of the interval. So in all cases, the optimal value n_j^* is close to $\frac{nm_j}{m}$, as intuitively expected.

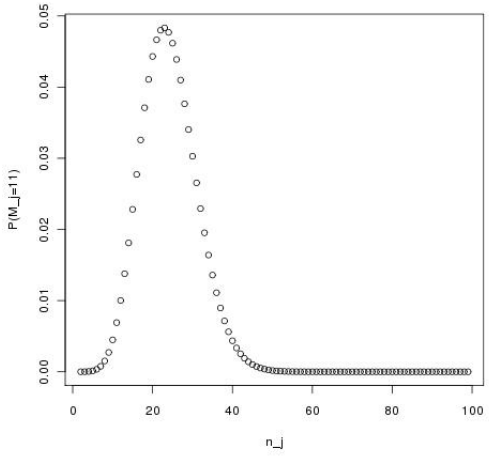
Corollary 2.1. *For general m, when selecting a category which maximises $\underline{P}(M_j = m_j)$, the optimal category is selected as follows:*

1. *If there exists c_j such that $n_j \in [1+\frac{m_j}{m}(n-3), 2+\frac{m_j}{m}(n-3)]$, then this category is optimal.*
2. *If there is no c_j such that $n_j \in [1+\frac{m_j}{m}(n-3), 2+\frac{m_j}{m}(n-3)]$, then find the value of n_j which is closest to the interval on each side. Compare the values of $\underline{P}(M_j = m_j)$ for the two corresponding categories. The category which gives the largest lower probability is optimal.*

We also notice that if we have a lot of observations and if both m_j and m are very large, then $\frac{m_j}{m}$ will tend to some limit l and therefore the interval $[1+\frac{m_j}{m}(n-3), 2+\frac{m_j}{m}(n-3)]$ will shrink to the point value nl. This means that the optimal value of the ratio will tend to the same limit l, as is to be expected.

Example 2.3. *Suppose we have a categorical data set consisting of 100 observations. There are 4 possible categories: blue (B), red (R), yellow (Y) and green (G). We have observed 20 B, 25 R, 28 Y and 27 G. We are making inferences about the next 50 observations, and we wish to select the category that maximises the lower probability $\underline{P}(M_j = 11)$.*

The plot of $\underline{P}(M_j = 11)$ against all possible values of n_j is shown below. From this graph, we expect that n_j^* will be between 20 and 25, as this is where the peak occurs.



By Theorem 2.3, the optimal value n_j^* lies in the interval $[1+\frac{11}{50}(97), 2+\frac{11}{50}(97)] = [22.34, 23.34]$, so the ideal choice of n_j would be $n_j = 23$. However, there is no c_j in the data set with this value of n_j, and so by Corollary 2.1 we must look at either side of the interval.

To the left of the interval, we have $n_j = 20$ corresponding to the blue category. By (2), the relevant lower probability here is $\underline{P}(M_B = 11) = 0.0443$. To the right of the interval, we have $n_j = 25$ corresponding to the red category. The lower probability here is $\underline{P}(M_R = 11) = 0.0462$. As the second probability is largest, we see that $n_j = 25$ is the optimal choice, and so we select red as our optimal category.

2.2.2 Deriving $\mathbf{P(M_j \geq m_j)}$

The other event of interest here is that *at least* m_j of the m future observations will belong to category c_j. For the lower probability, we again count the minimum number of relevant arrangements of the future observations. However, we are now interested in all arrangements which have R future observations which fall in the shaded segment A, where $m_j \leq R \leq m$. We consider each possible value of R separately in order to avoid counting any arrangements more than once. For a given value of R, there are $\binom{n_j-2+R}{R}$ different arrangements within this segment. We must also consider the remaining $m - R$ observations. Contrary to our lower probability formula above (2), arrangements with one or more observations in an optional slice will now be counted. We did not count these when finding the lower probability $\underline{P}(M_j = m_j)$, because for example an arrangement with m_j observations in segment A and 1 in an optional slice could be allocated to the event $M_j = m_j + 1$ when deriving $\underline{P}(M_j = m_j)$. However, such arrangements are now relevant because we are simultaneously considering all events $M_j \in \{m_j, m_j + 1, ..., m\}$.

By (1), the number of different arrangements of $m - R$ future observations within the shaded segment B plus the two optional slices is equal to $\binom{n-n_j+(m-R)}{m-R}$.

Multiplying the two binomial coefficients above leads to the minimum number of arrangements in which R future observations are in c_j. We now sum over R from m_j to m, which gives the lower probability

$$\underline{P}(M_j \geq m_j) = \binom{n+m-1}{m}^{-1} \sum_{R=m_j}^{m} \binom{n_j-2+R}{R} \times \binom{n-n_j+(m-R)}{m-R}. \quad (4)$$

It is assumed here that $n_j \geq 2$, because otherwise the lower probability will be zero. We also assume $m_j > 0$.

To find the corresponding upper probability, we have to maximise the number of arrangements which have at least m_j of the m future observations in category c_j. We still need to count all the arrangements described above, so all of the $\binom{n_j-2+R}{R}\binom{n-n_j+(m-R)}{m-R}$ arrangements will be included in our total, where $m_j \leq R \leq m$. However, we also want to include any arrangements where there are fewer than m_j observations in segment A but where observations in the optional slices can be counted as belonging to c_j.

Suppose we have Y observations in segment A, where $0 \leq Y \leq m_j - 1$. We need to count any arrangement which has $m_j - Y$ or more observations in an optional slice. Let T denote the total number of future observations in the optional slices. T may range from $m_j - Y$ to $m - Y$ for a given value of Y. As explained above, there are $T + 1$ possible arrangements of these observations for each value of T. Therefore, by (1), the number of different arrangements is equal to

$$\sum_{Y=0}^{m_j-1}\sum_{T=m_j-Y}^{m-Y}(T+1)\binom{n_j-2+Y}{Y} \times \binom{n-n_j-2+(m-Y-T)}{m-Y-T}.$$

Summing together both of the above numbers gives the total number of relevant arrangements, leading to the upper probability

$$\begin{aligned}\overline{P}(M_j \geq m_j) = \binom{n+m-1}{m}^{-1}\Big[&\sum_{R=m_j}^{m}\binom{n_j-2+R}{R} \times \binom{n-n_j+(m-R)}{m-R} \\ &+ \sum_{Y=0}^{m_j-1}\sum_{T=m_j-Y}^{m-Y}(T+1)\binom{n_j-2+Y}{Y} \times \binom{n-n_j-2+(m-Y-T)}{m-Y-T}\Big].\end{aligned} \tag{5}$$

As before, we assume $n_j \geq 2$ and $m_j > 0$. In the cases $n_j = 1$ and $n_j = 0$, the formula reduces to

$$\overline{P}(M_j \geq m_j) = \binom{n+m-1}{m}^{-1}\sum_{T=m_j}^{m}(T+1) \times \binom{n-n_j-2+(m-T)}{m-T}$$

and

$$\overline{P}(M_j \geq m_j) = \binom{n+m-1}{m}^{-1}\sum_{T=m_j}^{m} \times \binom{n-n_j-2+(m-T)}{m-T}$$

respectively.

These formulae can be used in a number of different ways. For example, suppose we wanted to select a category for which there was at least a 75% lower probability that two or more of the future observations would be in that category. We would use the above formulae to find all c_j such that $\underline{P}(m_j \geq 2) \geq 0.75$. Alternatively, suppose we wanted to select the category which was most likely to contain 10% or more of the future observations. We would evaluate $P(m_j \geq \frac{m}{10})$ for each of the possible categories, and then select the category according to these values.

This method of selection is illustrated in the example below.

Example 2.4. *Consider Example 2.2, where our possible categories are blue (B), red (R), yellow (Y) and green (G) and our data set consists of 5 observations as shown on the probability wheel in Example 2.2.*

We are making inferences about 3 future observations, and we want to select the category with the highest probability of containing at least one third of the future observations. To find the lower probability of the event $M_j \geq \frac{m}{3}$, we use (4) with $m_j = 1$. We first consider the blue category. Using the values $n = 5$, $m = 3$ and $n_j = 2$, we find that

$$\underline{P}(M_B \geq 1) = \frac{1}{35}[\binom{5}{2} + \binom{4}{1} + \binom{3}{0}] = \frac{15}{35}.$$

To find the upper probability, we use (5) with $m_j = 1$. For blue, this gives

$$\begin{aligned}\overline{P}(M_B \geq 1) &= \frac{1}{35}[15 + \sum_{T=1}^{3}(T+1)\binom{0}{0}\binom{4-T}{3-T}] \\ &= \frac{1}{35}[15 + 2\binom{3}{2} + 3\binom{2}{1} + 4\binom{1}{0}] = \frac{31}{35}.\end{aligned}$$

So we see that $P(M_B \geq 1) = [\frac{15}{35}, \frac{31}{35}]$. We investigate the three remaining categories in the same way, and we find that $P(M_j \geq 1) = [0, \frac{25}{35}]$ for all three categories. So the category we select here is blue.

3 Predictive subset selection

We now consider the use of predictive methods to select a subset of categories, rather than a single category, from a multinomial data set. As before, we have K possible categories, and we have a data set consisting of n observations where $n_1, ..., n_K$ denote the number of times we have observed categories $c_1, ..., c_K$ respectively. Recall that k represents the total number of categories that have been observed. We will select our subset based on inferences about m future observations. Our inferences use the general theory of nonparametric predictive inference [7].

3.1 One future observation

In this case, our aim will be to select a subset in order to maximise the NPI lower probability that the next observation, Y_{n+1}, belongs to a category within that subset.

Let S denote our selected subset of categories. Let OS denote the index set for already-observed categories in S, and let US denote the index set for unobserved categories in S. The sizes of these sets are denoted r and l respectively. Then, according to the NPI model [7], the formula for the lower probability $\underline{P}(Y_{n+1} \in S)$ is

$$\underline{P}(Y_{n+1} \in S) = \sum_{j \in OS} \frac{n_j - 1}{n} + \frac{(2r + l - K)^+}{n} \quad (6)$$

and the formula for the upper probability $\overline{P}(Y_{n+1} \in S)$ is

$$\overline{P}(Y_{n+1} \in S) = \sum_{j \in OS} \frac{n_j - 1}{n} + \frac{\min\{2r + l, k\}}{n}. \quad (7)$$

Our objective is to find some S such that

$$\underline{P}(Y_{n+1} \in S) \geq p^*$$

for some specified threshold probability p^*. We also want S to be of minimal size. If several such subsets exist, we select the one with maximum lower probability.

Example 3.1. *Consider Example 2.1, where our possible categories are blue (B), red (R), yellow (Y) and green (G), and our data set consists of 8 observations including 3 B, 2 G, 2 Y and 1 R. Now, we want to find a subset of categories S of minimal size which satisfies the criterion* $\underline{P}(Y_{n+1} \in S) \geq \frac{3}{8}$. *As shown in Example 2.1, B is the optimal choice when we are selecting a single category, and* $\underline{P}(m_B = 1) = \frac{2}{8}$. *So a subset of size 1 will not satisfy our requirements. We instead look for a subset of size 2. Consider the subset* $S = \{B, G\}$. *Here,* $r = 2$ *and* $l = 0$. *The formula (6) gives*

$$\underline{P}(Y_{n+1} \in \{B, G\}) = \frac{3-1}{8} + \frac{2-1}{8} + (4-4) = \frac{3}{8}.$$

This satisfies the selection criterion. Applying the same formula to other possible subsets of size 2 shows that $\frac{3}{8}$ *is the highest lower probability that we can achieve with a subset of size 2. So the subset we select is* $S = \{B, G\}$.

Theorem 3.1. *When* $m = 1$, *and we want to select a subset of categories according to our aforementioned definition of the optimal subset, it is always optimal to add categories to the subset in decreasing order of number of observations in the data set.*

Proof. We select a subset according to which gives the highest lower probability $\underline{P}(Y_{n+1} \in S)$. The addition of an already-observed category to S will add $\frac{n_j - 1}{n}$ to the first term in the lower probability formula and will add 2 to the second term. The addition of an unobserved category to S will add 0 to the first term and 1 to the second term. So we should always add observed categories before unobserved categories. Furthermore, the observed categories which will give the largest increase to the lower probability when added to S are those with the largest values of n_j. So it is always optimal to include categories in S in decreasing order of n_j, ie. in decreasing order of the number of observations. □

3.2 m future observations

We now consider inferences about multiple future observations. This requires some new notation: let M_S represent the number of future observations that are in S. In terms of the probability wheel, the event $M_S = m_s$ means that precisely m_s future observations fall in a slice allocated to S. Based on the NPI model [7], there are

$$L = \sum_{j \in OS} (n_j - 1) + (2r + l - K)^+ \quad (8)$$

slices of the wheel which must be assigned to a category in S.

In this section, we will consider the general case where m may take any value. We will focus on the event that M_S reaches a certain threshold value, ie. the event $M_S \geq m_S$, because for selection purposes, this is a more natural and useful event to consider than the event that M_S takes one specific value. As before, we derive the most conservative bounds possible within the NPI framework.

First we consider the lower probability. We need to find the minimum number of arrangements of the m future observations such that at least m_S are in the subset S. This involves counting all arrangements such that R observations fall in a slice which must be assigned to S, where $m_S \leq R \leq m$. It is important that we do not count any arrangement multiple times, and so we consider each value of R separately and then sum over R to avoid this.

There are L slices which must be assigned to S, so for a certain value of R, there are $\binom{L-1+R}{R}$ arrangements of the R observations within the slices which must be assigned to S.

We must also account for the other $m-R$ observations. The remainder of the wheel consists of $n-L$ slices, and by (1) there are $\binom{n-L-1+(m-R)}{m-R}$ different arrangements of the $m-R$ observations within these slices.

Multiplying the above binomial coefficients tells us the minimum number of arrangements for which $M_S = R$. We can now sum over all relevant values of R, leading to the lower probability

$$\underline{P}(M_S \geq m_s) = \binom{n+m-1}{m}^{-1} \sum_{R=m_S}^{m} \binom{L-1+R}{R} \times \binom{n-L-1+(m-R)}{m-R}. \tag{9}$$

We assume $0 < L < n$, because $L = 0$ leads to lower probability zero. We also assume $m_s > 0$.

Now we consider the upper probability, which means we need to maximise the number of arrangements which have at least m_S of the m future observations in the subset S. We must still count all of the arrangements described above, i.e. those where at least m_S of the future observations are in a slice which must be assigned to S. As explained above, there are a total of $\binom{L-1+R}{R}\binom{n-L-1+(m-R)}{m-R}$ arrangements such as this.

However, there are other arrangements which must now be included. We can now make use of the optional slices, ie. those slices which we can choose to assign either to S or to its complement. By considering the difference between the lower and upper probabilities given by the NPI model [7], we see that there are

$$Q = \min\{2r+l, k\} - (2r+l-K)^+$$

optional slices. If we have fewer than m_S observations in slices which must be assigned to S, but we have observations which fall in the Q optional slices, then we can count these observations as belonging to S.

Suppose we have Y observations which fall in a slice that must be assigned to the subset S, where $0 \leq Y \leq m_S - 1$. Any arrangement which has $m_S - Y$ or more observations in one of the optional slices must be counted when calculating the upper probability. Let T denote the total number of future observations in the optional slices. T can take values from $m_S - Y$ to $m - Y$ for a particular value of Y. For a certain Y, there are $\binom{L-1+Y}{Y}$ different arrangements of the Y observations within the slices which must be assigned to S. Also, there are $\binom{Q-1+T}{T}$ different arrangements of the T observations within the optional slices. Finally, there are $\binom{n-L-Q-1+(m-Y-T)}{m-Y-T}$ different arrangements of the other observations within the remaining slices of the wheel.

Combining these three binomial coefficents gives us the following upper probability:

$$\overline{P}(M_S \geq m_S) = \binom{n+m-1}{m}^{-1} [\sum_{R=m_S}^{m} \binom{L-1+R}{R} \times \binom{n-L-1+(m-R)}{m-R} + \sum_{Y=0}^{m_S-1} \sum_{T=m_S-Y}^{m-Y} \binom{L-1+Y}{Y}\binom{Q-1+T}{T} \times \binom{n-L-Q-1+(m-Y-T)}{m-Y-T}]. \tag{10}$$

As before, we assume $L > 0$ and $m_s > 0$. It is also assumed here that $L + Q < n$. This is because in the situation $L + Q = n$, every slice on the wheel may be assigned to the subset S, leading to the upper probability $\overline{P}(M_S \geq m_S) = 1$. In the case $L = 0$, the formula reduces to

$$\overline{P}(M_S \geq m_S) = \binom{n+m-1}{m}^{-1} [\sum_{Y=0}^{m_S-1} \sum_{T=m_S-Y}^{m-Y} \binom{Q-1+T}{T}\binom{n-Q-1+(m-Y-T)}{m-Y-T}].$$

Example 3.2. *Consider the data set in Example 2.2, where our possible categories are blue (B), red (R), yellow (Y) and green (G) and we have seen 5 observations including 2 B, 1 G, 1 Y and 1 R.*

We use inferences about three future observations, and we want to find the probability that at least one of these is in the subset $S = \{B, G\}$. To find the lower probability of this event, we use (9) with $m_S = 1$. We find that

$$L = \sum_{j \in OS} (\frac{n_j - 1}{n}) + (2r + l - K)^+ = 1$$

and

$$Q = \min\{2r + l, k\} - (2r + l - K)^+ = 4$$

for this example, and we also know that $n = 5$ and $m = 3$. Using these values we find that

$$\underline{P}(M_S \geq 1) = \frac{1}{35}[\binom{1}{1}\binom{5}{2} + \binom{2}{2}\binom{4}{1} + \binom{3}{3}] = \frac{15}{35}.$$

When finding the upper probability, we observe that $L + Q = n$, and this leads to $\overline{P}(M_S \geq 1) = 1$ because we may assign every slice on the wheel to S.

Now suppose that we want to find the probability that at least two of the three future observations are in S. We now apply (9) with $m_S = 2$, and we find that

$$\underline{P}(M_S \geq 2) = \frac{1}{35}[\binom{2}{2}\binom{4}{1} + \binom{3}{3}\binom{3}{0}] = \frac{5}{35}.$$

As before, every slice on the wheel can be assigned to S, and so $\overline{P}(M_S \geq 2) = 1$.

So we see that $P(M_S \geq 1) = [\frac{15}{35}, 1]$ and $P(M_S \geq 2) = [\frac{5}{35}, 1]$.

Theorem 3.2. *For general m, when selecting an optimal subset of categories (see Introduction for our optimality criteria), categories should always be added to the subset in decreasing order of number of observations in the data set.*

Proof. Our aim is to select the subset which has the highest lower probability $\underline{P}(M_S \geq m_s)$ for some given value m_s. L is the only variable in this formula which changes according to which categories are included in S. We therefore wish to determine the behaviour of $\underline{P}(M_S \geq m_s)$ as L increases. To do this, we will consider two consecutive values of L. Consider the ratio

$$\frac{\underline{P}(M_S \geq m_s|L)}{\underline{P}(M_S \geq m_s|L+1)}. \tag{11}$$

If $\underline{P}(M_S \geq m_s)$ were increasing in L, we would expect this ratio to be always less than 1. Now consider the term within the summation in the formula for this lower probability. If

$$\frac{\binom{L-1+R}{R}\binom{n-L-1+(m-R)}{m-R}}{\binom{L+R}{R}\binom{n-L+(m-R)}{m-R}} \tag{12}$$

is less than 1 for every possible value of R, then (11) must always be less than 1. Using the identities of the binomial coefficients, we can rewrite (12) as

$$\frac{L(n-L)}{(L+R)(n-L+m-R)}.$$

Then, $L(n-L) < (L+R)(n-L+m-R) \Leftrightarrow 0 < (L+R)(m-R) + R(n-L)$. The term $(L+R)$ is clearly always positive, $(m-R)$ must always be positive regardless of the value of R since m is the maximum value of R, and $(n-L)$ must always be positive since L will always be less than n. Therefore $\underline{P}(M_S \geq m_s)$ is increasing in L, and our initial aim translates to making L as large as possible.

We now consider how the composition of the subset S affects the value of L. By (8), the inclusion of an unobserved category in S will add 0 to the first term in L and 1 to the second term in L. The inclusion of an observed category in S will add $\frac{n_j - 1}{n}$ to the first term in L and 2 to the second term in L. So we see that it is always optimal to include observed categories in S before unobserved ones. Additionally, we see that the observed categories which will increase L by the greatest amount are those with the largest values of n_j. It is therefore always optimal to add categories to S in decreasing order of n_j. □

The following example illustrates how Theorem 3.2 can be implemented when selecting subsets.

Example 3.3. *Suppose that we have 8 possible categories, which we label A to H. We have made 100 observations. The table below shows how many of these observations were in each category.*

Category	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Observations	25	20	18	13	10	9	5	0

We want to investigate subsets of these 8 categories, and we will do this by making inferences about 2 future observations. There are two events of interest here: first, the event that at least one of the two future observations is in some subset S, and second, the event that both of the two future observations are in S.

Consider an increasing sequence of subsets $S_1, ..., S_8$, where we begin with a subset of size 1 and add one category at a time. By Theorem 3.2, we know that the categories will be added in decreasing order of number of observations. The table below shows the composition of each of the subsets.

i	S_i	$P(M_{S_i} \geq 1)$	$P(M_{S_i} \geq 2)$
1	A	$[0.4206, 0.4505]$	$[0.0594, 0.0695]$
2	A,B	$[0.6727, 0.7166]$	$[0.1873, 0.2234]$
3	A-C	$[0.8376, 0.8822]$	$[0.3624, 0.4378]$
4	A-D	$[0.9196, 0.9543]$	$[0.5204, 0.6257]$
5	A-E	$[0.9697, 0.9846]$	$[0.6903, 0.7754]$
6	A-F	$[0.9945, 0.9980]$	$[0.8655, 0.9220]$
7	A-G	$[0.9998, 1.0000]$	$[0.9802, 1.0000]$
8	A-H	$[1.0000, 1.0000]$	$[1.0000, 1.0000]$

Using (9) and (10) with $m_S = 1$, we can find the lower and upper probabilities that at least one of the two future observations will be in S_i for $i = 1, ..., 8$. Similarly, we can use (9) and (10) with $m_S = 2$ to find the lower and upper probabilities that both of the two future observations will be in S_i for $i = 1, ..., 8$. The above table shows these probabilities.

Suppose that we want to select a subset of minimal size such that there is at least a 50% lower probability that one or more of the future observations will belong to a category in that subset. Looking at the above table of probabilities for the event $(M_{S_i} \geq 1)$, we see that the first row which satisfies $\underline{P}(M_{S_i} \geq 1) \geq 0.5$ is the row corresponding to $i = 2$. We therefore select the subset $S_2 = \{A, B\}$.

However, now suppose that we want to select the smallest possible subset of categories such that there is at least a 50% lower probability that both of the future observations will belong to a category in that subset. We will now need to select a larger subset in order to achieve the minimally required probability. Looking at the above table for the event $(M_{S_i} \geq 2)$, we see that the first row which satisfies $\underline{P}(M_{S_i} \geq 2) \geq 0.5$ is the row corresponding to $i = 4$. We therefore select the subset $S_4 = \{A, B, C, D\}$.

4 Concluding remarks

Coolen and Augustin [7] proved strong consistency properties for NPI, including F-probability in Weichselberger's theory of interval probability [10], but only for inferences involving a single future observation. For the case with multiple future observations, considered in this paper, these properties have not yet been proved, as we have thus far only derived the lower and upper probabilities of specific events. We would need to derive general formulae in order to investigate such properties. This is an interesting and important topic for future research. Further related research topics include other applications of NPI for multinomial data, where for example applications to classification are being investigated. Detailed comparisons of the NPI methods to more established alternatives may provide further insight into their practical value.

Acknowledgements

The authors thank the referee for helpful comments that led to the improvement of this paper.

References

[1] Augustin, T. and Coolen, F.P.A. (2004) Nonparametric predictive inference and interval probability *Journal of Statistical Planning and Inference*, **124**, 251-272.

[2] Bechhofer, R., Santner, T. and Goldsman, D. (1995) *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons.* Wiley.

[3] Coolen, F.P.A. and van der Laan, P. (2001) Imprecise predictive selection based on low structure assumptions *Journal of Statistical Planning and Inference*, **98**, 259-277.

[4] Coolen, F.P.A. and Coolen-Schrijner, P. (2006) Nonparametric predictive subset selection for proportions *Statistics and Probability Letters*, **76**, 1675-1684.

[5] Coolen, F.P.A. and Coolen-Schrijner, P. (2007) Nonparametric predictive comparison of proportions *Journal of Statistical Planning and Inference*, **137**, 23-33.

[6] Coolen, F.P.A. and Augustin, T. (2005) Learning from multinomial data: a nonparametric predictive alternative to the Imprecise Dirichlet Model *ISIPTA '05*, 125-134.

[7] Coolen, F.P.A. and Augustin, T. (2009) A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories *International Journal of Approximate Reasoning*, **50**, 217-230.

[8] Coolen, F.P.A. (2006) On nonparametric predictive inference and objective Bayesianism *Journal of Logic, Language and Information*, **15**, 21-47.

[9] Hill, B.M. (1968) Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, **63**, 677-691.

[10] Weichselberger, K. (2000) The theory of interval probability as a unifying concept for uncertainty *International Journal of Approximate Reasoning*, **24**, 149-170.

Aggregating Imprecise Probabilistic Knowledge

Alessio Benavoli
IDSIA
Lugano, Switzerland
alessio@idsia.ch

Alessandro Antonucci
IDSIA
Lugano, Switzerland
alessandro@idsia.ch

Abstract

The problem of aggregating two or more sources of information containing knowledge about a same domain is considered. We propose an aggregation rule for the case where the available information is modeled by *coherent lower previsions*, corresponding to convex sets of probability mass functions. The consistency between aggregated beliefs and sources of information is discussed. A closed formula, which specializes our rule to a particular class of models, is also derived. Finally, an alternative explanation of Zadeh's paradox is provided.

Keywords. Information fusion, coherent lower previsions, independent natural extension, generalized Bayes rule.

1 Introduction

In practical problems where modeling and handling knowledge is required, information often comes piecewise from different sources. The modeler usually wants to aggregate these pieces of information into a global model, that serves as a basis for various kinds of inference, like decision making, estimation and many others. If the available information is characterized by uncertainty, Bayesian theories can offer a suitable approach to problems of this kind. Yet, there are situations where the level of uncertainty characterizing the sources is so high that single probability measures cannot properly model the available information. This goes beyond the standard Bayesian theory, and leads to alternative models of uncertainty, like for example Choquet capacities [3], belief functions [7], possibility measures [6], and fuzzy measures [15]. As shown in [14], all these models represent uncertainty through sets instead of single probability measures, and can be all regarded as special cases of Walley's *coherent lower previsions* [13]. This theory, which is usually referred to as *imprecise probability*, provides a very general model of uncertain knowledge, for which also some rationality criteria, that can be used to identify conflicts among the different sources and determine whether the model is self-consistent, are provided. All these features seem to be particularly suited for the aggregation of different sources of information, that might be not only uncertain and vague when considered singularly, but also conflictual or contradictory when considered jointly.

In this paper, we apply Walley's theory of coherent lower previsions to develop a method of aggregation for uncertain information coming from different sources. In order to describe this aggregation task, let us us first formalize the problem in the Bayesian framework.

Consider n sources of information, all reporting knowledge about a variable X, whose generic value x varies in a finite set $\mathscr{X}$.[1] For each $j = 1,\ldots,n$, the knowledge associated to the j-th source is modeled by a conditional probability mass function $p_j(X|A_j = a_j)$. In this formalism, the conditioning event $A_j = a_j$ describes the actual *internal state* of each source, which is in fact modeled by a variable A_j, whose possible realizations take values a_j in a finite set $\mathscr{A}_j$. Examples of internal states of the sources can be the two states of a binary variable denoting the fact that a source is reliable or not, or a collection of measurements collected for the phenomenon under study.

The information associated to the different sources is collected by a single *information fusion center* (IFC), which aims at aggregating this information together with its prior knowledge about X, modeled as a probability mass function $p_0(X)$. This is achieved by identifying the sources' beliefs about A_j given that $X = x$ with those of the IFC:

$$p_0(a_j|x) := p_j(a_j|x) = \frac{p_j(x|a_j)p_j(a_j)}{\sum_{a_j\in\mathscr{A}_j} p_j(x|a_j)p_j(a_j)}, \qquad (1)$$

for each $x \in \mathscr{X}$, where $p_j(A_j)$ is the prior over the internal states of the j-th source. Thus, assuming conditional independence between the variables in $(A_1,\ldots,A_n)$ given X, we can aggregate those beliefs into the following joint:

$$p_0(x,a_1,\ldots,a_n) \quad = \quad \prod_{j=1}^{n} \frac{p_j(x|a_j)p_j(a_j)}{p_j(x)} p_0(x) \ , \qquad (2)$$

[1] Variables are denoted in this paper by uppercase letters; the corresponding calligraphic and lowercase letters denote respectively their sets of possible values and the generic values of these sets.

with $p_j(x) = \sum_{a_j \in \mathscr{A}_j} p_j(x|a_j)p_j(a_j)$ prior of the j-th source. Finally, from (2), the *aggregated* posterior is:

$$p_0(x|\tilde{a}_1,\ldots,\tilde{a}_n) \quad \propto \quad \prod_{j=1}^{n} \frac{p_j(x|\tilde{a}_j)}{p_j(x)} p_0(x) \ , \qquad (3)$$

where $\tilde{a}_j$ denotes the element of $\mathscr{A}_j$ corresponding to the observed internal state of the source. According to (3), $p_0(x|\tilde{a}_1,\ldots,\tilde{a}_n)$ is only a function of the IFC's prior $p_0(X)$, and of the sources' conditional $p_j(X|\tilde{a}_j)$ and prior $p_j(X)$, where the latter two are the only pieces of information to be shared between the sources and the IFC. Note also that the prior over the internal states $p_j(A_j)$ has been dropped from (3) because of normalization.

Figure 1 depicts the sequential steps involved in the above derivation. The idea there is that each source should be regarded as an independent subject, that has inferred its conditional beliefs about X given the actual internal state of the source. As formalized in (1), each source induces a *model revision* into the IFC's beliefs. This means that, regarding the state of the source conditional on X, the IFC identifies its own beliefs with those of the source. Finally, the IFC defines a global model over all the variables by exploiting the independence among the sources as in (2).

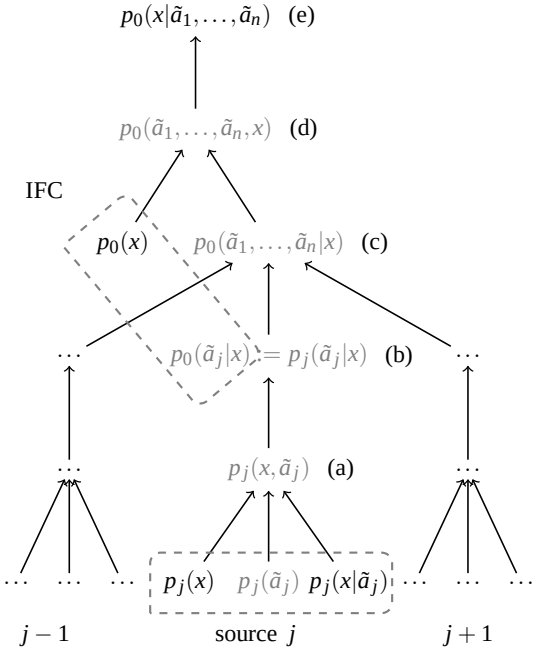


Figure 1: Aggregation of the sources of information in the Bayesian framework. The black-highlighted text describes the information used by the IFC to compute the final posterior density (still in black). The gray-highlighted text denotes the intermediate steps needed to aggregate the information. The dashed boxes are used to group the beliefs whose coherence will be checked in Section 4.

In this architecture it has been assumed that each source processes its own information in order to compute the posterior probability $p_j(x|a_j)$, which can be regarded as a *sufficient statistical descriptor*, to be shared with the IFC together with $p_j(x)$. This is a high-level form of aggregation, since the IFC aggregates pieces of information which have already been elaborated from the sources. This is one of the most common architectures for data fusion (see for example [2, Chapter 8]).

In this paper we aim at generalizing this approach to Walley's theory of imprecise probability in the general case where, instead of probability mass functions, the uncertainty about a variable is described by *coherent lower previsions*. To this end, in Section 2 we first recall the basics of the theory of coherent lower previsions. In Section 3, we detail the different steps of our derivation leading to a combination rule for the general case of coherent lower previsions. The consistency between the obtained results and the original assessment is discussed in Section 4. The rule is indeed specialized in Section 5 for a special class of coherent lower previsions, called *linear-vacuous mixtures*. Finally, in Section 6, we show how this rule can be applied in practice for a possible explanation of Zadeh's paradox [16]. Conclusions and outlooks for future developments are in Section 7.

2 Coherent Lower Previsions

The *imprecise probability* theory [13] is an extension of the Bayesian theory of subjective probability. The goal is to model a subject's uncertainty by looking at his dispositions toward taking certain actions, and imposing requirements of rationality, or consistency, on these dispositions. In order to do that, let us first recall the fundamental notion of *coherent lower prevision*.

Given a variable X taking values in a set $\mathscr{X}$, we use *gambles*, i.e. bounded functions $f : \mathscr{X} \to \mathbb{R}$, in order to test a subject's uncertainty about X. For each $x \in \mathscr{X}$, the real number $f(x)$ is regarded as the (possibly negative) reward, expressed in some linear utility units, that the subject receives by accepting the gamble if $X = x$. Uncertainty about the actual value of X can be modeled by the willingness to accept certain gambles and to reject others. Bayesian theory assumes that subjects are always able to provide a fair price $P(f)$ for f, whatever information is available about X. This assumption is relaxed in the imprecise probability framework, where subjects can express two different prices, called respectively lower and upper previsions and denoted by $\underline{P}(f)$ and $\overline{P}(f)$, that correspond to the highest (lowest) buying (selling) price for the gamble f. Since selling a gamble f for a given price r is the same as buying $-f$ for the price $-r$, the conjugacy relation $\overline{P}(f) = -\underline{P}(-f)$ holds and we can therefore focus on lower previsions only. If $\mathscr{L}(\mathscr{X})$ denotes the set of all the

bounded[2] gambles on $\mathscr{X}$, a lower prevision $\underline{P}$ can be regarded as a real-valued functional on $\mathscr{L}(\mathscr{X})$.

Indicator functions[3] are clearly a special class of gambles. Given a set $\mathscr{X}' \subseteq \mathscr{X}$, we can consider the lower prevision for the corresponding indicator function $I_{\mathscr{X}'}$. The behavioural interpretation of $\underline{P}(I_{\mathscr{X}'})$ is the supremum rate for which the subject is disposed to bet on the event $x \in \mathscr{X}'$, which is the subject's *lower probability* for this event, similarly $\overline{P}(I_{\mathscr{X}'}) = 1 - \underline{P}(I_{\mathscr{X}\setminus\mathscr{X}'})$ is the *upper probability*.

Since lower previsions represent a subject's dispositions to act in certain ways, some criteria ensuring that these dispositions do not lead to irrational behaviours should be imposed. *Coherence* is the strongest requirement considered in the theory of imprecise probability. A lower prevision $\underline{P}$ is *coherent* if and only if it satisfies the following properties:

(P1) $\min_{x\in\mathscr{X}} f(x) \leq \underline{P}(f)$ [accepting sure gains],

(P2) $\underline{P}(f+g) \geq \underline{P}(f) + \underline{P}(g)$ [super-additivity],

(P3) $\underline{P}(\lambda f) = \lambda \underline{P}(f)$ [positive homogeneity],

for all $f, g \in \mathscr{L}(\mathscr{X})$ and non-negative real numbers λ. We point the reader to [13, Chapter 2] for a deep explanation of the irrational consequences of modeling beliefs by lower previsions that are not coherent. Here, we regard a *coherent lower prevision* (CLP) as the more general model of a subject's (rational) beliefs about a variable.

Let us present some examples of CLP. A *linear prevision* P on $\mathscr{L}(\mathscr{X})$ is a CLP which is also self-conjugate, i.e., $P(-f) = -P(f)$ for each $f \in \mathscr{L}(\mathscr{X})$. This property makes the prevision a linear functional, i.e., $P(\lambda(f+g)) = \lambda P(f) + \lambda P(g)$ for all $f, g \in \mathscr{L}(\mathscr{X})$ and real λ. Any linear prevision P is completely determined by its *mass function* $p(x) := P(I_{\{x\}})$, since it follows from the previous properties that for any gamble f, $P(f) = \sum_{x\in\mathscr{X}} p(x) f(x)$. A CLP $\underline{P}$ on $\mathscr{L}(\mathscr{X})$ such that $\underline{P}(f) = \min_{x\in\mathscr{X}} f(x)$ can be easily identified as the most conservative (i.e., less informative) CLP and is therefore called *vacuous*. As both linear and vacuous previsions are coherent, we can construct new coherent lower previsions by convex combination of the two [13, Chapter 2]. If P is a linear prevision, for each $0 \leq \varepsilon \leq 1$, $\underline{P}(f) := \varepsilon P(f) + (1-\varepsilon)\min_{x\in\mathscr{X}} f(x)$ defines a new CLP which is called *linear-vacuous mixture*. Walley proved that a CLP can be equivalently specified by a convex set of linear previsions, and hence a convex set of probability distributions [13].

Now consider also a second variable A with values in $\mathscr{A}$. Given a CLP $\underline{P}$ on $\mathscr{L}(\mathscr{X}\times\mathscr{A})$, we can easily compute its *marginal* prevision on $\mathscr{A}$ for each $f \in \mathscr{L}(\mathscr{A})$ by noting that f can be equivalently regarded as a gamble in $\mathscr{L}(\mathscr{X}\times\mathscr{A})$ which is constant with respect to X, and set

$$\underline{P}^A(f) := \underline{P}(f), \tag{4}$$

where the superscript A emphasizes the fact that the marginal prevision is defined on $\mathscr{L}(\mathscr{A})$.

For each $h \in \mathscr{L}(\mathscr{X}\times\mathscr{A})$ and $a \in \mathscr{A}$, a subject's *conditional lower prevision* $\underline{P}^{X|A}(h|A=a)$, denoted also as $\underline{P}^{X|A}(h|a)$, is the highest real number r for which the subject would buy the gamble h for any price strictly lower than r, if he knew in addition that the variable A assumes the value a. We denote by $\underline{P}^{X|A}(h|A)$ the gamble on A that assumes the value $\underline{P}^{X|A}(h|A=a)$ for each $a \in \mathscr{A}$. Overall, $\underline{P}^{X|A}(h|A)$ is a gamble on $\mathscr{A}$ for each $h \in \mathscr{L}(\mathscr{X}\times\mathscr{A})$ and $\underline{P}^{X|A}(\cdot|A)$ is a map between $\mathscr{L}(\mathscr{X}\times\mathscr{A})$ and $\mathscr{L}(\mathscr{A})$.

A conditional lower prevision $\underline{P}^{X|A}(\cdot|A)$ is said to be *separately coherent* if $\underline{P}^{X|A}(\cdot|a)$ is a CLP on $\mathscr{L}(\mathscr{X}\times\mathscr{A})$ and $\underline{P}^{X|A}(I_{\mathscr{X}\times\{a\}}|a) = 1$, for each $a \in \mathscr{A}$. The last condition means that if the subject knew that $A = a$, he would be disposed to bet at all non-trivial odds on the event that $A = a$.

If, besides the separately coherent conditional lower prevision $\underline{P}^{X|A}(\cdot|A)$ on $\mathscr{L}(\mathscr{X}\times\mathscr{A})$, the subject has also specified an unconditional CLP $\underline{P}$ on $\mathscr{L}(\mathscr{X}\times\mathscr{A})$, then $\underline{P}$ and $\underline{P}^{X|A}(\cdot|A)$ should in addition satisfy the criterion of *joint coherence*, that requires

$$\underline{P}\left(I_{\mathscr{X}\times\{a\}}\left[h - \underline{P}^{X|A}(h|a)\right]\right) = 0, \tag{5}$$

for each $a \in \mathscr{A}$ and $h \in \mathscr{L}(\mathscr{X}\times\mathscr{A})$. It can be proved [13, Chapter 6] that, if $\underline{P}(I_{\mathscr{X}\times\{a\}}) > 0$, $\underline{P}^{X|A}(h|a)$ is the only solution of (5). Thus, given a joint CLP on $\mathscr{L}(\mathscr{X}\times\mathscr{A})$, a (separately coherent) conditional lower prevision can be obtained from (5). For this reason, this equation is also called *generalized Bayes rule* (GBR). GBR cannot be applied if $\underline{P}(I_{\mathscr{X}\times\{a\}}) = 0$. Nevertheless, if $\overline{P}(I_{\mathscr{X}\times\{a\}}) > 0$, a conditional prevision $\underline{P}^{X|A}(\cdot|a)$ can be computed by the following *regular extension*

$$\underline{P}^{X|A}(h|a) = \max\{\mu : \underline{P}\left(I_{\mathscr{X}\times\{a\}}[h-\mu]\right) \geq 0\}. \tag{6}$$

On the other side, given a (separately coherent) conditional lower prevision $\underline{P}^{X|A}(\cdot|A)$ and a coherent marginal prevision $\underline{P}^A$ on $\mathscr{A}$, a joint CLP on $\mathscr{L}(\mathscr{X}\times\mathscr{A})$ can be obtained by *marginal extension*:

$$\underline{P}(h) = \underline{P}^A\left(\underline{P}^{X|A}(h|A)\right). \tag{7}$$

The marginal extension $\underline{P}$ in (7) can be proved to be jointly coherent with $\underline{P}^{X|A}$ as in (5), and its marginal on A is still $\underline{P}^A$ [13, Chapter 6].

The standard notion of conditional independence considered in the Bayesian theory, requires a more general formulation in the framework of CLPs. Given a joint CLP $\underline{P}$

[2] Although Walley's theory has been developed for bounded gambles only, an extension to the unbounded case can be found in [12].

[3] A real-valued function on a domain is called the *indicator function* of a given subset of this domain if it takes the value one inside the subset and zero otherwise.

on $\mathscr{L}(\mathscr{X}\times\mathscr{A}_i\times\mathscr{A}_j)$, we say that, according to $\underline{P}$, A_j is *epistemically irrelevant* to A_i given X, if:

$$\underline{P}^{A_i|X,A_j}(h|x,a_j)=\underline{P}^{A_i|X}(h|x), \tag{8}$$

for each $h\in\mathscr{L}(\mathscr{A}_i)$, $x\in\mathscr{X}$ and $a_j\in\mathscr{A}_j$, where both $\underline{P}^{A_i|X,A_j}$ and $\underline{P}^{A_i|X}$ are obtained from $\underline{P}$ through GBR. If A_j is epistemically irrelevant to A_i given X, and A_i is epistemically irrelevant to A_j given X, then A_i and A_j are said to be epistemically independent (given X).

Let us adopt, for sake of compactness, the notation $A^n:=(A_1,\ldots,A_n)$ and $\mathscr{A}^n:=\times_{j=1}^n\mathscr{A}_j$. Given a collection of separately coherent conditional lower previsions $\underline{P}_j^{A_j|X}$ on $\mathscr{L}(\mathscr{A}_j)$, for each $j=1,\ldots,n$, the most conservative separately coherent conditional lower prevision $\underline{P}^{A^n|X}$ which is coherent with each $\underline{P}_j^{A_j|X}$, under the assumption that, for each $i,j=1,\ldots,n$ with $i\neq j$, A_i and A_j are epistemically independent given X, is defined as follows:[4]

$$\underline{P}(g|x)=\sup_{\substack{g_j\in\mathscr{L}(\mathscr{A}_j)\\ j=1,\ldots,n}}\ \inf_{\substack{a_j\in\mathscr{A}_j\\ j=1,\ldots,n}}\Big\{g(a_1,\ldots,a_n)-\sum_{j=1}^n\Big[g_j(a_1,\ldots,a_n)-\underline{P}_j(g_j(a_1,\ldots,a_{j-1},\cdot,a_{j+1}\ldots,a_n)|x)\Big]\Big\} \tag{9}$$

This is the *independent natural extension* [5][5]. The notion of joint coherence between a separately coherent conditional lower prevision and a joint CLP in (5) reflects the fact that our assessments should be consistent not only separately, but also with each other. For this case, joint coherence can be characterized by the following theorem.

Theorem 1. *The separately coherent conditional lower previsions $\underline{P}_j^{A_j|X}$, with $j=1,\ldots,n$, are jointly coherent if there is a CLP $\underline{P}$ on $\mathscr{L}(\mathscr{X}\times\mathscr{A}^n)$ such that: (i) its marginal $\underline{P}^X$ assigns positive probability to the elements of $\mathscr{X}$; (ii) its marginals $\underline{P}^{A_j,X}$ are jointly coherent with $\underline{P}_j^{X|A_j}$, for each $j=1,\ldots,n$, in the sense of (5).*

A more general formulation of Theorem 1 and its proof can be found in [9].

3 Aggregating Coherent Lower Previsions

The theoretical results reviewed in Section 2 can be employed for a generalization to imprecise probabilities of the aggregation rule presented in Section 1. Accordingly, we suppose that the j-th source of information, for each $j=1,\ldots,n$, makes assessments about the value that X assumes in $\mathscr{X}$ conditionally on its internal states $\tilde{a}_j\in\mathscr{A}_j$. Such assessments are expressed through separately coherent conditional lower previsions $\underline{P}_j^{X|A_j}$. Furthermore, also extra assessments about the internal states of the sources are available and again expressed in terms of CLPs $\underline{P}_j^{A_j}$ on $\mathscr{L}(\mathscr{A}_j)$ for $j=1,\ldots,n$. The IFC should therefore gather this information and aggregate it with its prior about X, which is expressed as a CLP $\underline{P}_0^X$ on $\mathscr{L}(\mathscr{X})$.

Our goal is to compute the IFC's joint CLP $\underline{P}_0$ on $\mathscr{L}(\mathscr{X}\times\mathscr{A}^n)$ from which the beliefs about X conditional on the actual internal states of the sources $(\tilde{a}_1,\ldots,\tilde{a}_n)$ could be computed. By analogy with the derivation in Section 1, this task is achieved by the following sequential steps:

(a) As outlined in (7), a CLP $\underline{P}_j$ on $\mathscr{L}(\mathscr{X}\times\mathscr{A}_j)$ can be derived from $\underline{P}_j^{X|A_j}$ and $\underline{P}_j^{A_j}$ by marginal extension

$$\underline{P}_j(f_j):=\underline{P}_j^{A_j}\left(\underline{P}_j^{X|A_j}(f_j|A_j)\right), \tag{10}$$

for each $f_j\in\mathscr{L}(\mathscr{X}\times\mathscr{A}_j)$ and $j=1,\ldots,n$.

(b) GBR is used to compute, given $\underline{P}_j$, the conditional CLP $\underline{P}_j^{A_j|X}$ on $\mathscr{L}(\mathscr{X}\times\mathscr{A}_j)$.[6] Accordingly, by computing the solution μ of the equation

$$\underline{P}_j\left(I_{\{\tilde{x}\}}\cdot[f_j-\mu]\right)=0, \tag{11}$$

we have $\underline{P}_j^{A_j|X}(f_j|\tilde{x}):=\mu$, for each $f_j\in\mathscr{L}(\mathscr{A}_j)$, $\tilde{x}\in\mathscr{X}$, and $j=1,\ldots,n$.

The so-obtained separately coherent conditional lower previsions associated to the sources are assumed to induce a *model revision* into the corresponding beliefs of the IFC, i.e.,

$$P_0^{A_j|X}(f_j|x):=P_j^{A_j|X}(f_j|x), \tag{12}$$

for each $f_j\in\mathscr{L}(\mathscr{A}_j)$ and $x\in\mathscr{X}$.

(c) A conditional CLP $\underline{P}_0^{A^n|X}$ is obtained from $\underline{P}_0^{A_j|X}$ by independent natural extension (9):

$$\underline{P}_0^{A^n|X}(g|x)=\sup_{\substack{g_j\in\mathscr{L}(\mathscr{A}_j)\\ j=1,\ldots,n}}\ \inf_{\substack{a_j\in\mathscr{A}_j\\ j=1,\ldots,n}}\Big\{g(a_1,\ldots,a_n)-\sum_{j=1}^n\Big[g_j(a_1,\ldots,a_n)-\underline{P}_0^{A_j|X}(g_j(a_1,\ldots,a_{j-1},\cdot,a_{j+1}\ldots,a_n)|x)\Big]\Big\}. \tag{13}$$

(d) Then, the joint CLP $\underline{P}_0$ on $\mathscr{L}(\mathscr{X}\times\mathscr{A}^n)$ is derived by marginal extension (7):

$$\underline{P}_0(g):=\underline{P}_0^X\left(\underline{P}_0^{A^n|X}(g|X)\right), \tag{14}$$

for each $g\in\mathscr{L}(\mathscr{X}\times\mathscr{A}^n)$.

[4] A more general formula for non-linear spaces can be found in [10].

[5] This paper includes a survey of different aggregation rules for CLPs. Yet, our approach differs in aggregating knowledge referred to the same domain.

[6] We noted that GBR requires $\underline{P}_j^X(I_{\{\tilde{x}\}})>0$. If only $\overline{P}_j^X(I_{\{\tilde{x}\}})>0$ holds, regular extension (6) should be employed instead. An example of the calculations required in this latter case is in Section 6.

(e) Finally, assuming that $\underline{P}_0^{A^n}(\tilde{a}_1,\ldots,\tilde{a}_n) > 0$, where $(\tilde{a}_1,\ldots,\tilde{a}_n) \in \mathscr{A}^n$ are the observed internal states of the sources, we again apply GBR,

$$\underline{P}_0\left(I_{\{\tilde{a}_1,\ldots,\tilde{a}_n\}} \cdot [g-\mu]\right) = 0, \qquad (15)$$

to compute the separately coherent conditional lower prevision $\underline{P}_0^{X|A^n}(\cdot|A^n)$ on $\mathscr{L}(\mathscr{X})$.[7]

The above derivation has been achieved by complete analogy with that in Section 1, but in the more general framework of CLPs. Notice that, if the sources directly provide the CLPs $\underline{P}_j^{A_j|X}$, we could still apply our procedure by considering only the steps from (c) to (e). In this case, the posterior probabilities coincide with those returned by a *naive credal classifier* (e.g., compare the equation in Table 2 with the results in [17]). This holds in spite of the different notion of independence assumed in [17], and can be verified by means of the algorithm in [4].

The coherence between the joint CLP obtained at the step (d) and the initial assessments will be investigated in the next section.

4 Checking Coherence

The subjects involved in the derivation formalized in the previous section (i.e., the sources and the IFC) should be regarded as autonomous and distinct individuals. Nevertheless, we have assumed that the uncertain information associated to a subject can induce in another subject a *model revision*, i.e., the second agent can replace his own CLPs (even in the conditional case) with those of the first agent. More specifically, in our architecture, we allow for an *asymmetrical* model revision, as we assume that each source revises the IFC's beliefs as in (1) or in (12), while the contrary cannot take place because of the way the sources and the IFC share the information. In this section we discuss the coherence between the different beliefs specified in our model. According to the previous argument, this will be done separately for each subject, by considering also the beliefs induced by other subjects via model revision.

Let us start from the coherence of the IFC's beliefs. In order to do that, we first consider the derivation in the precise case as in Section 1. As outlined in Figure 1, the mass functions to be considered are the conditionals $p_0(A_j|x)$, for each $j = 1,\ldots,n$, which are obtained through model revision from the sources, and the marginal $p_0(X)$. The consistency between these assessments when considered jointly follows from the existence of a joint probability mass function, which is clearly the one in (2), from which these mass functions can be obtained. Concerning the IFC, we should also verify that this joint probability mass function preserves the assumption of independence between the sources given X. This holds since, after marginalization and Bayes rule, the joint probability mass function p_0 in (2) is such that $p_0(a_i|x,a_j) = p_0(a_i|x)$ for each $i,j = 1,\ldots,n$, $a_i \in \mathscr{A}_i$, $a_j \in \mathscr{A}_j$ and $x \in \mathscr{X}$. Analogous results, in the more general framework of imprecise probability, can be obtained by considering the joint CLP $\underline{P}_0$ in (14), which is the basis to prove the following result.

Theorem 2. *The separately coherent conditional lower previsions $\underline{P}_0^{A_j|X}$ in (12) and $\underline{P}_0^X$ are jointly coherent.*

Proof. The joint coherence of the assessments $\underline{P}_0^X$ and $\underline{P}_j^{A_j|X}(\cdot|x)$, considered for each $j = 1,\ldots,n$, can be proved by considering the joint CLP $\underline{P}^{X,A^n}$ in (14). As a consequence of marginal extension, $\underline{P}^{X,A^n}$ is jointly coherent with both $\underline{P}_0^X$ and $\underline{P}^{A^n|X}(\cdot|x)$. Furthermore, as a consequence of independent natural extension, $\underline{P}^{A^n|X}(\cdot|x)$ is jointly coherent with all the $\underline{P}_j^{A_j|X}(\cdot|x)$ for $j = 1,\ldots,n$, because of the epistemic independence between the variables in $(A_1,\ldots,A_n)$ given X. Finally, assuming $\underline{P}_j^{A^n}(I_{\{a_1,\ldots,a_n\}}) > 0$ because of GBR, the coherence of $\underline{P}^{X|A^n}(\cdot|a_1,\ldots,a_n)$ follows from Theorem 1. □

On the other side, checking the coherence of the beliefs associated to a particular source is trivial, as $\underline{P}_j^{X|A_j}$ and $\underline{P}_j^{A_j}$ are jointly coherent because of (5), for each $j = 1,\ldots,n$. We have argued that the IFC's beliefs are not required to be coherent with those of the sources, as they refer to separate subjects. Nevertheless, let us consider what can be said about the consistency between different subjects in the Bayesian (i.e., precise) formulation. By exploiting the independencies between the sources, (2) rewrites as:

$$p_0(x,a_1,\ldots,a_n) \;=\; \prod_{j=1}^{n} \frac{p_0(x|a_j)p_0(a_j)}{p_0(x)} p_0(x) \;. \qquad (16)$$

By comparing (16) with (2), it can be noticed that the joint coherence between the IFC's beliefs and those of the j-th source cannot be guaranteed in general. In fact, we can always impose $p_0(x|a_j) := p_j(x|a_j)$ and $p_0(a_j) := p_j(a_j)$, but, at least in general, it is not possible to have at the same time $p_0(X) = p_j(X)$, for each $j = 1,\ldots,n$. In fact, since each source and the IFC are considered autonomous subjects and the information flows from the sources to the IFC, we cannot require that the sources agree on their marginals over $\mathscr{X}$, i.e., $p_i(X) = p_j(X)$ for each $i,j = 1,\ldots,n$. Thus, the IFC can define a single global probabilistic model over all the variables that reproduces all the inputs from the sources only if the IFC and all the sources have the same prior over X.

[7] Note that, also in this case, if we only have that $\overline{P}_0^{A^n}(\tilde{a}_1,\ldots,\tilde{a}_n) > 0$, the regular extension (6) can be used instead.

5 Mathematical Derivation for Linear-Vacuous Mixtures

Let us detail the derivation described in Section 3 in the special case where the marginal associated to the IFC and the separately coherent conditional lower previsions specified for the sources are linear-vacuous mixtures, while the marginals over $\mathscr{A}_j$ are linear.[8] This corresponds to the following settings:

$$\begin{aligned}
&\underline{P}_0^X(h) := \varepsilon_0 \sum_{x\in\mathscr{X}} p_0(x)h(x) + (1-\varepsilon_0)\min_{x\in\mathscr{X}} h(x),\\
&\underline{P}_j^{X|A_j}(f_j|a_j) := \varepsilon_j^{a_j} \sum_{x\in\mathscr{X}} p_j(x|a_j) f_j(x,a_j) \qquad (17)\\
&+(1-\varepsilon_j^{a_j})\min_{x\in\mathscr{X}} f_j(x,a_j), \quad \forall a_j\in\mathscr{A}_j\\
&P_j^{A_j}(g_j) := \sum_{a_j\in\mathscr{A}_j} p_j(a_j)g_j(a_j), \qquad (18)
\end{aligned}$$

where $p_j(X|a_j)$, $p_j(A_j)$ and $p_0(X)$ are probability mass functions, $f_j \in \mathscr{L}(\mathscr{X}\times\mathscr{A}_j)$, $g_j \in \mathscr{L}(\mathscr{A}_j)$, and $h \in \mathscr{L}(\mathscr{X})$, for all $j = 1,\dots,n$. The derivation is as follows.

(a) In this particular case, (10) rewrites as

$$\underline{P}_j(f_j) = \sum_{a_j\in\mathscr{A}_j} p_j(a_j)\cdot\Big(\varepsilon_j^{a_j}\sum_{x\in\mathscr{X}} p_j(x|a_j)\cdot f_j(x,a_j) + (1-\varepsilon_j^{a_j})\min_{x\in\mathscr{X}} f_j(x,a_j)\Big), \quad (19)$$

for each $f_j \in \mathscr{L}(\mathscr{X}\times\mathscr{A}_j)$ and $j = 1,\dots,n$.

(b) Thus, for each $\tilde{x}\in\mathscr{X}$, (11) becomes:

$$\sum_{a_j\in\mathscr{A}_j} p_j(a_j)\cdot\Big(\varepsilon_j^{a_j}[f_j(\tilde{x},a_j)-\mu]p_j(\tilde{x}|a_j) + (1-\varepsilon_j^{a_j})\min\{0, f_j(\tilde{x},a_j)-\mu\}\Big) = 0. \quad (20)$$

Define the subset $\mathscr{A}_j^*(\mu)$ of $\mathscr{A}_j$ as follows:

$$\mathscr{A}_j^*(\mu) := \{a_j\in\mathscr{A}_j :\ f_j(\tilde{x},a_j)-\mu<0\}, \quad (21)$$

where f_j, $\tilde{x}$ are omitted from the arguments of $\mathscr{A}_j^*$ for sake of simpler notation. Equation (20) rewrites as:

$$\begin{aligned}
&\sum_{a_j\in\mathscr{A}_j} p_j(a_j)[\varepsilon_j^{a_j}p_j(\tilde{x}|a_j)+(1-\varepsilon_j^{a_j})I_{\mathscr{A}_j^*(\mu)}(a_j)]f_j(\tilde{x},a_j)\\
&-\mu\sum_{a_j\in\mathscr{A}_j} p_j(a_j)[\varepsilon_j^{a_j}p_j(\tilde{x}|a_j)+(1-\varepsilon_j^{a_j})I_{\mathscr{A}_j^*(\mu)}(a_j)] = 0.
\end{aligned} \quad (22)$$

The solution of (20) is non-trivial because $\mathscr{A}_j^*$ is a function of μ. Yet, we can compute $\mathscr{A}_j^*(\mu)$ for the particular value $\tilde{\mu}$ of μ that solves (20), without explicitly solving this equation. Accordingly, we set $\mathscr{A}_j^* := \mathscr{A}_j^*(\tilde{\mu})$, and the solution $\underline{P}_j^{A_j|X}(f_j|\tilde{x})$ of (22) is:[9]

$$\frac{\sum_{a_j\in\mathscr{A}_j} p_j(a_j)[\varepsilon_j^{a_j}p_j(\tilde{x}|a_j)+(1-\varepsilon_j^{a_j})I_{\mathscr{A}_j^*}(a_j)]f_j(\tilde{x},a_j)}{\sum_{a_j\in\mathscr{A}_j} p_j(a_j)[\varepsilon_j^{a_j}p_j(\tilde{x}|a_j)+(1-\varepsilon_j^{a_j})I_{\mathscr{A}_j^*}(a_j)]}. \quad (23)$$

(c) The (separately coherent) conditional lower previsions associated to the sources and defined as in (23) induce the following *model revision* into the IFC's beliefs,

$$P_0^{A_j|X}(f_j|x) := P_j^{A_j|X}(f_j|x), \quad (24)$$

for each $f_j\in\mathscr{L}(\mathscr{A}_j)$, $j = 1,\dots,n$ and $x\in\mathscr{X}$. Their independent natural extension to $\mathscr{A}^n$ can be therefore considered:

$$\begin{aligned}
\underline{P}_0^{A^n|X}(g|\tilde{x}) = \sup_{\substack{g_j\in\mathscr{L}(\mathscr{X}\times\mathscr{A}_j)\\ j=1,\dots,n}} \inf_{\substack{a_j\in\mathscr{A}_j\\ j=1,\dots,n}} \Big\{ & g(\tilde{x},a_1,\dots,a_n)\\
&-\sum_{j=1}^n\Big[g_j(\tilde{x},a_1,\dots,a_n)\\
&-\underline{P}_0^{A_j|X}(g_j(\tilde{x},a_1,\dots,a_{j-1},\cdot,a_{j+1}\dots,a_n)|\tilde{x})\Big]\Big\}, \quad (25)
\end{aligned}$$

for each $\tilde{x}\in\mathscr{X}$. Notice that the gamble $g_j(\tilde{x},a_1,\dots,a_{j-1},\cdot,a_{j+1}\dots,a_n)$ is in $\mathscr{L}(\mathscr{X}\times\mathscr{A}_j)$. Let us consider, in (25), only gambles $g\in\mathscr{L}(\mathscr{X}\times\mathscr{A}^n)$ such that, for $X=\tilde{x}$ and each $(a_1,\dots,a_n)\in\mathscr{A}^n$, factorize as follows:

$$g(\tilde{x},a_1,\dots,a_n) = \prod_{j=1}^n g'_j(\tilde{x},a_j), \quad (26)$$

with $g'_j\in\mathscr{L}(\mathscr{X}\times\mathscr{A}_j)$ for each $j = 1,\dots,n$. Assume also that the gamble $g'_j(\tilde{x},\cdot)\in\mathscr{L}(\mathscr{A}_j)$ has a constant sign in $\mathscr{A}_j$, and denote its sign by $\sigma_j = \sigma_j(\tilde{x})$[10]. Under these assumptions, if we intend, for fixed $\tilde{x}$, g as a gamble on $\mathscr{A}^n$, we have that g has constant sign and (25) reduces to:

$$\underline{P}_0^{A^n|X}(g|\tilde{x}) = \begin{cases} \prod_{j=1}^n \underline{P}_0^{A_j|X}(g'_j|\tilde{x}) & \text{if } g\geq 0\\ -\ \prod_{j=1}^n \overline{P}_0^{A_j|X}(\sigma_j g'_j|\tilde{x}) & \text{if } g<0 \end{cases} \quad (27)$$

where g'_j is the g_j defined in (25), for each $j = 1,\dots,n$. The proof is in [10]. The gambles we consider in the following factorize as in (26), and we can therefore use (27) instead of (25).

[8] The last assumption will be relaxed at the end of this section.

[9] This is possible unless $\underline{P}_j(I_{\{\tilde{x}\}\times\mathscr{A}_j}) = \sum_{a_j\in\mathscr{A}_j} p_j(a_j)\varepsilon_j^{a_j}p_j(\tilde{x}|a_j) > 0$.

[10] Set $\sigma_j = +1$ if $g'_j(\tilde{x},\cdot) > 0$, $\sigma_j = -1$ if $g'_j(\tilde{x},\cdot) < 0$ and $\sigma_j = 0$ otherwise.

(d) By marginal extension (14), the following joint CLP can be calculated:

$$\underline{P}_0(h) = \underline{P}_0^X\left(\underline{P}_0^{A^n|X}(h|x)\right) = \varepsilon_0 \sum_{x\in\mathscr{X}} \underline{P}_0^{A^n|X}(h|x)p_0(x) + (1-\varepsilon_0)\min_{x\in\mathscr{X}} \underline{P}_0^{A^n|X}(h|x). \quad (28)$$

(e) Thus, by GBR, given $\{\tilde{a}_1,\ldots,\tilde{a}_n\} \in \mathscr{A}^n$, the conditional CLP $\underline{P}_0^{X|A^n}(g|\tilde{a}_1,\ldots,\tilde{a}_n)$ is the solution of:

$$\underline{P}_0(I_{\{\tilde{a}_1\}\times\cdots\times\{\tilde{a}_n\}}(g-\mu)) = 0, \quad (29)$$

where we assume $\underline{P}_0(I_{\{\tilde{a}_1\}\times\cdots\times\{\tilde{a}_n\}}) > 0$. Note also that the only values of the gamble g that should be considered for the solution of (29) are those such that $A^n \neq \tilde{a}^n$, because otherwise the argument of $\underline{P}_0$ is zero. Furthermore, for fixed x, $g(x,\tilde{a}_1,\ldots,\tilde{a}_n)-\mu$ is constant. Thus, the gamble factorizes as in (26), with $g_i'(\tilde{x},a_i) = I_{\{\tilde{a}_i\}}\ \forall i<n$ and $g_i'(\tilde{x},a_n) = I_{\{\tilde{a}_n\}}(g(\cdot)-\mu)$. Therefore, notice that $\sigma_i = 1\ \forall i<n$ and $\sigma_n = \operatorname{sgn}(g(\cdot)-\mu)$. Thus, (27) holds and we can write:[11]

$$\begin{aligned}\underline{P}^{A^n|X}(h|x) = {} & \underline{P}_1^{A_1|X}(I_{\{\tilde{a}_1\}}|x)\cdots\underline{P}_n^{A_n|X}(I_{\{\tilde{a}_n\}}|x) \\ & [g(x,\tilde{a}_1,\ldots,\tilde{a}_n)-\mu]I_{\{g(x,\tilde{a}_1,\ldots,\tilde{a}_n)-\mu\geq 0\}} \\ & + \overline{P}_1^{A_1|X}(I_{\{\tilde{a}_1\}}|x)\cdots\overline{P}_n^{A_n|X}(I_{\{\tilde{a}_n\}}|x) \\ & [g(x,\tilde{a}_1,\ldots,\tilde{a}_n)-\mu]I_{\{g(x,\tilde{a}_1,\ldots,\tilde{a}_n)-\mu<0\}} \quad (30)\end{aligned}$$

According to (30), (29) can be written as in Table 1, where from (23) it can be derived that:

$$\underline{P}_j^{A_j|X}(I_{\{\tilde{a}_j\}}|\tilde{x}) = \frac{p_j(\tilde{a}_j)\varepsilon_j^{\tilde{a}_j}p_j(x|\tilde{a}_j)}{\sum_{a_j\in\mathscr{A}_j}p_j(a_j)[\varepsilon_j^{a_j}p_j(x|a_j)+(1-\varepsilon_j^{a_j})I_{\mathscr{A}_j\setminus\{\tilde{a}_j\}}(a_j)]}. \quad (31)$$

It can be easily verified that $\mathscr{A}_j^* = \mathscr{A}_j\setminus\{\tilde{a}_j\}$ in this case. Again from (23) it follows that:

$$\underline{P}_j^{A_j|X}\left(I_{\{\mathscr{A}_j\setminus\tilde{a}_j\}}|x\right) = \frac{\sum_{a_j\in\mathscr{A}_j, a_j\neq\tilde{a}_j}p_j(a_i)\varepsilon_j^{a_j}p_j(x|a_j)}{\sum_{a_j\in\mathscr{A}_j}p_j(a_j)[\varepsilon_j^{a_j}p_j(x|a_j)+(1-\varepsilon_j^{a_j})I_{\{\tilde{a}_j\}}(a_j)]}, \quad (32)$$

where, in this case, $\mathscr{A}_j^* = \{\tilde{a}_j\}$. According to the duality relation reviewed in Section 2, the corresponding upper probability is one minus the lower probability in (32), and hence:

$$\overline{P}_j^{A_j|X}\left(I_{\{\tilde{a}_j\}}|x\right) = \frac{p_j(\tilde{a}_j)[\varepsilon_j^{\tilde{a}_j}p_j^{\tilde{a}_j}(x)+(1-\varepsilon_j^{\tilde{a}_j})]}{\sum_{a_j\in\mathscr{A}_j}p_j(a_i)[\varepsilon_j^{a_j}p_j(x|a_j)+(1-\varepsilon_j^{a_j})I_{\{\tilde{a}_j\}}(a_j)]}. \quad (33)$$

Finally, by solving the equation in Table 1 with respect to μ, the conditional CLPs $\underline{P}_0^{X|A^n}(g|\tilde{a}_1,\ldots,\tilde{a}_n)$ can be calculated for each $\{\tilde{a}_1,\ldots,\tilde{a}_n\} \in \mathscr{A}^n$.

[11] Note that the indicator functions in (30) refer to sets that are implicitly defined through inequalities over gambles. This kind of specification will be employed also in the followings.

The assumption of linearity for the prior beliefs over the sources can be relaxed to the case where the previsions $\underline{P}_j^{A_j}$ are CLPs generated by the lower envelope of a finite set of linear previsions [13, Chapter 3]. In this case, we solve the equation in Table 1 for each element of this set, and the minimum over these values is the solution in the general case. The following results can be easily verified to follow from our derivation.

1. If $\underline{P}_0^X$ is vacuous (i.e., $\varepsilon_0 = 0$), then also $\underline{P}_0^{X|A^n}$ is vacuous. This is consistent with the results in [11].
2. If $\underline{P}_j^{X|A_j}$ is vacuous (i.e., $\varepsilon_j^{\tilde{a}_j} = 0$) for each $j = 1,\ldots,n$, then $\underline{P}_j(I_{\{\tilde{x}\}\times\mathscr{A}_j}) = 0$ and, (20) cannot be solved by (23). In this case, from (20) it is straightforward to verify that $\underline{P}_j^{A_j|X}(f_j|\tilde{x})$ is vacuous (if $p_j(a_i) > 0$ for each i), that $\underline{P}^{A^n|X}(g|\tilde{x})$ is also vacuous and that $\underline{P}_0^{X|A^n}(g|\tilde{a}_1,\ldots,\tilde{a}_n)$ is equal to $\underline{P}_0^X(g)$.
3. In (3), it is shown that, since the posterior probability distribution $p_0(x|a_1,\ldots,a_n)$ does not depend on $p(a_j)$, the only pieces of information to be shared between sources and IFC are $p_j(x)$ and $p_j(x|a_j)$. In the imprecise case, additional information must be shared between sources and IFC. In fact, from Table 1 and from (31) and (33), it can be seen that $\underline{P}_0^{X|A^n}(g|\tilde{a}_1,\ldots,\tilde{a}_n)$ depends on the sources' priors $\underline{P}_j^X$ and on $(1-\varepsilon_j^{\tilde{a}_j})p(\tilde{a}_j)$. Notice, in fact, that the denominator in (31) is just equal to $\overline{P}_j^X(I_{\{x\}}) - (1-\varepsilon_j^{\tilde{a}_j})p(\tilde{a}_j) = \underline{P}_j^X(I_{\mathscr{X}\setminus\{x\}}) - (1-\varepsilon_j^{\tilde{a}_j})p(\tilde{a}_j)$, while the denominator in (33) is $\underline{P}_j^X(I_{\{x\}}) + (1-\varepsilon_j^{\tilde{a}_j})p(\tilde{a}_j)$. Conversely, the dependency on $p(\tilde{a}_j)$ in the numerators of (31) and (33) is dropped in Table 1, since the sum and the minimum are over x and, thus, the $p(\tilde{a}_j)$ can be simplified. Summarizing, the pieces of information to be shared between sources and IFC are: the marginal CLP $\underline{P}_j^X$, which corresponds to the prior CLP of the sources; the quantity $(1-\varepsilon_j^{\tilde{a}_j})p(\tilde{a}_j)$, which is equal to the probability that the j-th source is in the state $p(\tilde{a}_j)$ multiplied by the *degree of uncertainty* $\overline{P}_j^{X|A_j}(I_{\{x\}}) - \underline{P}_j^{X|A_j}(I_{\{x\}}) = 1-\varepsilon_j^{\tilde{a}_j}$.

6 Zadeh's Paradox

The problem of aggregating beliefs over the same variable has been already considered in other uncertainty theories. In the case of Dempster-Shafer (DS) theory [7], Dempster's combination rule allows for the following aggregation of two belief functions m_1 and m_2:[12]

$$m_{12}(X) \propto \sum_{X_1,X_2:X_1\cap X_2=X} m_1(X_1)\cdot m_2(X_2). \quad (34)$$

[12] We point to [7] for details about DS theory.

Table 1: The unique solution μ of GBR corresponding to the conditional CLP $\underline{P}_0^{X|A^n}(g|\tilde{a}_1,\ldots,\tilde{a}_n)$

$$
\begin{aligned}
0 \;=\;& \varepsilon_0 \sum_{x\in\mathscr{X}} \Big\{ \Big[\underline{P}_0^{A_1|X}(I_{\{\tilde{a}_1\}}|x)\cdots\underline{P}_0^{A_n|X}(I_{\{\tilde{a}_n\}}|x) I_{\{g(x,\tilde{a}_1,\ldots,\tilde{a}_n)-\mu\geq 0\}} \\
&+ \overline{P}_0^{A_1|X}(I_{\{\tilde{a}_1\}}|x)\cdots\overline{P}_0^{A_n|X}(I_{\{\tilde{a}_n\}}|x) I_{\{g(x,\tilde{a}_1,\ldots,\tilde{a}_n)-\mu<0\}} \Big] (g(x,\tilde{a}_1,\ldots,\tilde{a}_n)-\mu) p_0(x) \Big\} \\
&+ (1-\varepsilon_0)\min_{x\in\mathscr{X}} \Big\{ \Big[\underline{P}_0^{A_1|X}(I_{\{\tilde{a}_1\}}|x)\cdots\underline{P}_0^{A_n|X}(I_{\{\tilde{a}_n\}}|x) I_{\{g(x,\tilde{a}_1,\ldots,\tilde{a}_n)-\mu\geq 0\}} \\
&+ \overline{P}_0^{A_1|X}(I_{\{\tilde{a}_1\}}|x)\cdots\overline{P}_0^{A_n|X}(I_{\{\tilde{a}_n\}}|x) I_{\{g(x,\tilde{a}_1,\ldots,\tilde{a}_n)-\mu<0\}} \Big] (g(x,\tilde{a}_1,\ldots,\tilde{a}_n)-\mu) \Big\}
\end{aligned}
$$

Yet, in the 1980s, DS theory suffered a serious blow when Zadeh proposed his "paradox", an example for which the Dempster's rule of combination gave an apparently counter-intuitive result [16].

Zadeh's example is as follows. Two doctors examine a patient and agree that he suffers from either meningitis (x_1), contusion (x_2) or brain tumor (x_3). Thus, $\mathscr{X}=\{x_1,x_2,x_3\}$ is the frame of the variable of interest. The doctors agree in considering a tumor quite unlikely, but disagree in the likely cause, thus providing the following diagnosis:

$$
\begin{array}{ll} \text{Doctor 1} \rightarrow & m_1(x_1)=0.99, \quad m_1(x_3)=0.01, \\ \text{Doctor 2} \rightarrow & m_2(x_2)=0.99, \quad m_2(x_3)=0.01, \end{array} \tag{35}
$$

while the basic belief masses of the other elements of the power set of $\mathscr{X}$ are null. By (34) one gets

$$
m_{12}(x_1)=0, \quad m_{12}(x_2)=0, \quad m_{12}(x_3)=1. \tag{36}
$$

Hence, from direct application of the DS theory, it turns out that the patient suffers from brain tumor with certainty. This result arises from the fact that the two doctors agree that the patient most likely does not suffer from tumor but are in almost full contradiction for the other causes of the disease. Since doctors' diagnoses are modeled by precise probability mass functions, also Bayesian approaches like the one in Section 1 might be applied to Zadeh's example; yet the same result is obtained.

Haenni has shown that the controversy of Zadeh's example can be overcome by assuming that the doctors are not fully reliable [8]. To take this into account, one has to build a model that includes two more variables, modeling the reliabilities of the doctors. Let $A_1=a_1$ correspond to the statement "Doctor 1 is reliable", and $A_1=\neg a_1$ to "Doctor 1 is unreliable", $p_1(a_1)$ can be therefore interpreted as the probability that the first source is reliable, $p_1(\neg a_1)=1-p_1(a_1)$ that is unreliable, and similarly for Doctor 2. By following this idea, our aggregation rule can be applied to Zadeh's example. The doctors' diagnoses (35) can be formalized as in (17) by setting $\varepsilon_1^{a_1}=1$, $p_1(x_1|a_1)=0.99$, $p_1(x_2|a_1)=0$, $p_1(x_3|a_1)=0.01$ and $\varepsilon_1^{\neg a_1}=0$ for Doctor 1, and similarly but with $p_2(x_1|a_2)=0$ and $p_2(x_2|a_2)=0.99$ for Doctor 2. Notice that, by setting $\varepsilon_1^{\neg a_1}=\varepsilon_2^{\neg a_2}=0$, it has been assumed that $\underline{P}_1^{X|\neg a_1}$ and $\underline{P}_1^{X|\neg a_2}$ are vacuous, i.e., when the doctors are unreliable they do not provide any useful information. Furthermore, we assume that $p_1(a_1)=p_2(a_2)=\delta$ with $\delta\in(0,1)$ and $\varepsilon_0=1$, $p_0(x_1)=p_0(x_2)=p_0(x_3)=1/3$. The goal is the evaluate the posterior belief $\underline{P}_0^{X|A_1,A_2}(I_{\{\tilde{x}\}}|\tilde{a}_1,\tilde{a}_2)$, which represents the lower probability of the diagnosis $\tilde{x}\in\mathscr{X}$ conditional on the fact that the sources are in a particular state $(\tilde{a}_1,\tilde{a}_2)$. In this case, we can compute the lower probability $\underline{P}_0^{X|A_1,A_2}(I_{\{\tilde{x}\}}|\tilde{a}_1,\tilde{a}_2)$ by simply putting $g(x,\tilde{a}_1,\tilde{a}_2)=I_{\{\tilde{x}\}}$ in the equation in Table 1. The final conditional are shown in Table 2. For Doctor 1, the CLPs $\underline{P}_1^{A_j|X}$ for $X=x_1$ or $X=x_3$ can be derived by applying equations (32)-(33). Conversely, for $X=x_2$, since $\underline{P}_1(I_{\{x_2\}\times\mathscr{A}_1})=0$, the GBR cannot be applied to get $\underline{P}_1^{A_j|x_2}$ and, thus, (32)-(33) are not valid anymore. However, since

$$
\begin{aligned}
&\overline{P}_1(I_{\{x_2\}\times\mathscr{A}_1}) = \sum_{\tilde{a}_j\in\mathscr{A}_1} p_1(\tilde{a}_j)\cdot\Big(\varepsilon_j^{\tilde{a}_j}\sum_{x\in\mathscr{X}} p_1(x|\tilde{a}_j) \\
&\cdot I_{\{x_2\}\times\mathscr{A}_1}(x,\tilde{a}_j)+(1-\varepsilon_j^{\tilde{a}_j})\max_{x\in\mathscr{X}} I_{\{x_2\}\times\mathscr{A}_1}(x,\tilde{a}_j)\Big)\;, \\
&= p_1(\neg a_1) > 0
\end{aligned}
$$

the regular extension (6) can be used to derive

$$
\underline{P}_1^{A_j|x_2}(g|x_2)=\max_{\mu}\underline{P}\left(I_{\{x_2\}\times\mathscr{A}_1}[g-\mu]\right)\geq 0
$$

where the gambles we are interested in are only $I_{\{a_1\}}$ and $I_{\{\neg a_1\}}$. From (22), $\underline{P}_1^{A_j|x_2}(g|x_2)$ can be calculated by finding the maximum value of μ for which

$$
\begin{aligned}
&\sum_{\tilde{a}_j\in\mathscr{A}_1} p_j(\tilde{a}_j)[\varepsilon_j^{\tilde{a}_j}p_j(x_2|\tilde{a}_j)+(1-\varepsilon_j^{\tilde{a}_j})I_{\mathscr{A}_1^*(\mu)}(\tilde{a}_j)]g(\tilde{a}_j) \\
&-\mu\sum_{\tilde{a}_j\in\mathscr{A}_1} p_j(\tilde{a}_j)[\varepsilon_j^{\tilde{a}_j}p_j(x_2|\tilde{a}_j)+(1-\varepsilon_j^{\tilde{a}_j})I_{\mathscr{A}_1^*(\mu)}(\tilde{a}_j)]\geq 0.
\end{aligned} \tag{37}
$$

The values of μ which satisfy (37) in the cases $g=I_{\{a_1\}}$ and $g=I_{\{\neg a_1\}}$ are $\mu=0$ and, respectively, $\mu=1$. Hence, it follows that $\underline{P}_1^{A_j|x_2}(I_{\{a_1\}}|x_2)=\overline{P}_1^{A_j|x_2}(I_{\{a_1\}}|x_2)=$

Table 2: Upper and lower conditional probability for the Zadeh's example for $i, j, k = 1, 2, 3$ and $i \neq j \neq k$

$$\underline{P}_0^{X|A_1,A_2}(I_{\{x_i\}}|\tilde{a}_1,\tilde{a}_2) = \frac{\underline{P}_1^{A_1|X}(I_{\{\tilde{a}_1\}}|x_i)\underline{P}_2^{A_2|X}(I_{\{\tilde{a}_2\}}|x_i)}{\underline{P}_1^{A_1|X}(I_{\{\tilde{a}_1\}}|x_i)\underline{P}_2^{A_2|X}(I_{\{\tilde{a}_2\}}|x_i)+\overline{P}_1^{A_1|X}(I_{\{\tilde{a}_1\}}|x_j)\overline{P}_2^{A_2|X}(I_{\{\tilde{a}_2\}}|x_j)+\overline{P}_1^{A_1|X}(I_{\{\tilde{a}_1\}}|x_k)\overline{P}_2^{A_2|X}(I_{\{\tilde{a}_2\}}|x_k)}$$

$$\overline{P}_0^{X|A_1,A_2}(I_{\{x_i\}}|\tilde{a}_1,\tilde{a}_2) = \frac{\overline{P}_a^{A_1|X}(I_{\{\tilde{a}_1\}}|x_i)\overline{P}_2^{A_2|X}(I_{\{\tilde{a}_2\}}|x_i)}{\overline{P}_1^{A_1|X}(I_{\{\tilde{a}_1\}}|x_i)\overline{P}_2^{A_2|X}(I_{\{\tilde{a}_2\}}|x_i)+\underline{P}_1^{A_1|X}(I_{\{\tilde{a}_1\}}|x_j)\underline{P}_2^{A_2|X}(I_{\{\tilde{a}_2\}}|x_j)+\underline{P}_1^{A_1|X}(I_{\{\tilde{a}_1\}}|x_k)\underline{P}_2^{A_2|X}(I_{\{\tilde{a}_2\}}|x_k)}$$

0 and $\underline{P}_1^{A_j|x_2}(I_{\{\neg a_1\}}|x_2) = \overline{P}_1^{A_j|x_2}(I_{\{\neg a_1\}}|x_2) = 1$. A similar derivation can be clearly achieved for Doctor 2. The posterior lower and upper probabilities calculated for the reliability value $\delta = 0.8$ are shown in Table 3. The values of the conditionals which depend on δ are highlighted in bold-face. It can be noticed that, in the case the sources are in the states $\tilde{a}_1 = a_1$ and $\tilde{a}_2 = a_2$, i.e., both sources are reliable, one gets the following precise conditional probability $\underline{P}_0^{X|A_1,A_2}(I_{\{x_1\}}|a_1,a_2) = \overline{P}_0^{X|A_1,A_2}(I_{\{x_1\}}|a_1,a_2) = 0$, $\underline{P}_0^{X|A_1,A_2}(I_{\{x_2\}}|a_1,a_2) = \overline{P}_0^{X|A_1,A_2}(I_{\{x_2\}}|a_1,a_2) = 0$, and $\underline{P}_0^{X|A_1,A_2}(I_{\{x_3\}}|a_1,a_2) = \overline{P}_0^{X|A_1,A_2}(I_{\{x_3\}}|a_1,a_2) = 1$. This result holds for each value of δ and shows that, when both the sources are reliable, the answer provided in (36) by both DS and Bayesian theory is coherent with the initial assessments. In fact, since Doctor 1 says implicitly that x_2 is wrong (with almost absolute certainty), and Doctor 2 says that x_1 is wrong, it follows then that x_3 must be the true diagnosis when both doctors are reliable.

According to Table 3 it can also be noticed that when both doctors are unreliable the conditionals are vacuous for all the diseases. Conversely, in the case only one doctor is reliable, e.g., Doctor 1 in Table 3, the disease that he believes wrong has precisely zero probability. For $\delta > 0.9$, it can be verified that $\underline{P}_0^{X|A_1,A_2}(I_{\{x_1\}}|a_1,\neg a_2) > \overline{P}_0^{X|A_1,A_2}(I_{\{x_3\}}|a_1,\neg a_2)$ and, thus, the lower probability of x_1 dominates the upper probability of the other element. In this case, the IFC can decide, without doubts, that the patient suffers from the disease x_1.

In general, in this kind of reliability problems, the sources of information do not provide their reliability status $\{\tilde{a}_1,\tilde{a}_2\}$ and, thus, the IFC cannot know it. However, since the doctors' diagnoses are almost in full contradiction, the IFC can infer that at least one of the doctors must be unreliable and, thus, apply the aggregation rule by computing the following lower conditional probability $\underline{P}_0^{X|A_1,A_2}(\cdot|\mathscr{A}^2\backslash\{a_1,a_2\})$. In practice, the conditioning event is the complementary event of $\{a_1,a_2\}$, which means that at least one doctor is unreliable.

Since $I_{\mathscr{A}^2\backslash\{a_1,a_2\}}$ do not factorize as in (26), we cannot apply (30) to compute $\underline{P}^{A^2|X}(\cdot|x)$. However, since $\underline{P}^{A^2|X}(\cdot|x)$ is a CLP, we can exploit the following property: $\underline{P}^{A^2|X}(I_{\mathscr{A}^2\backslash\{a_1,a_2\}}|x) = 1-\overline{P}^{A^2|X}(I_{\{a_1,a_2\}}|x) = 1-\overline{P}^{A_1|X}(I_{\{a_1\}}|x)\overline{P}^{A_2|X}(I_{\{a_2\}}|x)$ and $\overline{P}^{A^2|X}(I_{\mathscr{A}^2\backslash\{a_1,a_2\}}|x) = 1-\underline{P}^{A^2|X}(I_{\{a_1,a_2\}}|x) = 1-\underline{P}^{A_1|X}(I_{\{a_1\}}|x)\underline{P}^{A_2|X}(I_{\{a_2\}}|x)$.

Since $\overline{P}^{A_1|X}(I_{\{a_1\}}|x_i)\overline{P}^{A_2|X}(I_{\{a_2\}}|x_i) = 0$ and $\underline{P}^{A_1|X}(I_{\{a_1\}}|x_i)\underline{P}^{A_2|X}(I_{\{a_2\}}|x_i) = 0$ for $i = 1,2$, and $\overline{P}^{A_1|X}(I_{\{a_1\}}|x_3)\overline{P}^{A_2|X}(I_{\{a_2\}}|x_3) = 1$, the lower and upper probabilities are those in Table 4. Because of $\underline{P}_0^{X|A_1,A_2}(I_{\{x_1\}}|I_{\mathscr{A}^2\backslash\{a_1,a_2\}}) = \underline{P}_0^{X|A_1,A_2}(I_{\{x_2\}}|I_{\mathscr{A}^2\backslash\{a_1,a_2\}}) \geq \overline{P}_0^{X|A_1,A_2}(I_{\{x_3\}}|I_{\mathscr{A}^2\backslash\{a_1,a_2\}})$, the IFC can infer that the patient suffers from x_1 or x_2 but not from x_3. It can be noticed that when the reliability δ approaches one, the lower and upper probabilities converge to the following precise probability mass function: $p_0^{X|A_1,A_2}(I_{\{x_1\}}|I_{\mathscr{A}^2\backslash\{a_1,a_2\}}) = p_0^{X|A_1,A_2}(I_{\{x_2\}}|I_{\mathscr{A}^2\backslash\{a_1,a_2\}}) = 1/2$.

Summarizing, the results of this section generalize those in [8, 1] to CLPs by showing that: (i) if both the doctors are reliable the result obtained by the Bayes' and Dempster's rule in (36) is correct and coherent with the initial assessments; (ii) if we assume that at least one of the doctors is unreliable, we obtain that the patient must suffer from either x_1 or x_2.

7 Conclusions and Outlooks

A general aggregation rule for coherent lower previsions defined on the same domain has been proposed. This is achieved by a simultaneous *model revision* of beliefs associated to different sources of information. The coherence of the aggregated beliefs is also discussed. Furthermore, in the particular case of linear-vacuous mixtures, a closed formula for the aggregated beliefs has been derived. As an example of applications of this approach, Zadeh's paradox is treated and an alternative explanation is concluded.

As a future work, we aim to generalize our formula for linear-vacuous mixtures to the more general case of 2-monotone capacities. That would be the basis for a recursive application of our approach. Furthermore, although the size of the possibility space of the variable of interest has been assumed finite, it seems possible to extend our results to the infinite case. Yet, further investigations

Table 3: Posterior lower and upper probabilities in the case $\delta = 0.8$

	$\underline{P}_0^{X\|A_1,A_2}(\cdot\|a_1,a_2)$	$\overline{P}_0^{X\|A_1,A_2}(\cdot\|a_1,a_2)$	$\underline{P}_0^{X\|A_1,A_2}(\cdot\|a_1,\neg a_2)$	$\overline{P}_0^{X\|A_1,A_2}(\cdot\|a_1,\neg a_2)$	$\underline{P}_0^{X\|A_1,A_2}(\cdot\|\neg a_1,\neg a_2)$	$\overline{P}_0^{X\|A_1,A_2}(\cdot\|\neg a_1,\neg a_2)$
x_1	0	0	**0.45**	1	0	1
x_2	0	0	0	0	0	1
x_3	1	1	0	**0.54**	0	1

Table 4: Upper and lower conditional probabilities conditioned on $I_{\mathscr{A}^2\setminus\{a_1,a_2\}}$ for $i = 1,2$

$$\underline{P}_0^{X|A_1,A_2}(I_{\{x_i\}}|I_{\mathscr{A}^2\setminus\{a_1,a_2\}}) = \frac{1}{3-\underline{P}^{A_1|X}(I_{\{a_1\}}|x_3)\underline{P}^{A_2|X}(I_{\{a_2\}}|x_3)}, \quad \overline{P}_0^{X|A_1,A_2}(I_{\{x_i\}}|I_{\mathscr{A}^2\setminus\{a_1,a_2\}}) = \frac{1}{2}$$

$$\underline{P}_0^{X|A_1,A_2}(I_{\{x_3\}}|I_{\mathscr{A}^2\setminus\{a_1,a_2\}}) = 0, \quad \overline{P}_0^{X|A_1,A_2}(I_{\{x_3\}}|I_{\mathscr{A}^2\setminus\{a_1,a_2\}}) = \frac{1-\underline{P}^{A_1|X}(I_{\{a_1\}}|x_3)\underline{P}^{A_2|X}(I_{\{a_2\}}|x_3)}{3-\underline{P}^{A_1|X}(I_{\{a_1\}}|x_3)\underline{P}^{A_2|X}(I_{\{a_2\}}|x_3)}$$

about the coherence of the corresponding model should be considered. We also want to investigate the relationships between our approach in the case of a single source and Jeffrey's updating. Finally, we intend to apply our rule to practical problems of information fusion in signal and data processing and communications.

Acknowledgments

This research has been partially supported by the Swiss NSF grants n. 200020-121785/1. The authors are grateful to Marco Zaffalon and Enrique Miranda for their useful suggestions.

References

[1] S. Arnborg. Robust Bayesianism: Imprecise and paradoxical reasoning. In *Proceedings of the Seventh International Conference on Information Fusion*, volume I, pages 407–414. International Society of Information Fusion, 2004.

[2] Y. Bar-Shalom. *Multitarget-Multisensor Tracking: Advanced Applications.* Artech-House, Norwood, MA, 1990.

[3] G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, pages 131–295, 1953–1954.

[4] G. de Cooman, F. Hermans, A. Antonucci, and M. Zaffalon. Epistemic irrelevance in credal networks: the case of Markov trees. (under preparation).

[5] G. de Cooman and M. Troffaes. Coherent lower previsions in systems modelling: products and aggregation rules. *Reliability Engineering and System Safety*, 85:113–134, 2004.

[6] D. Dubois and H. Prade. *Possibility Theory*. Plenum Press, New York, 1988.

[7] Shafer G. *A mathematical theory of evidence*. Princeton University Press, 1976.

[8] R. Haenni. Shedding new light on Zadeh's criticism of Dempster's rule of combination. In *Proc. 8th Int. Conf. Information Fusion*, contribution No. C8-1, Philadelphia, USA, 2005.

[9] E Miranda. Updating coherent previsions on finite spaces. Technical reports of statistics and decision sciences, Rey Juan Carlos University, 2008.

[10] E. Miranda and G. de Cooman. Coherence and independence in non-linear spaces. Technical reports of statistics and decision sciences, Rey Juan Carlos University, 2005.

[11] A. Piatti, M. Zaffalon, F. Trojani, and M. Hutter. Limits of learning about a categorical latent variable under prior near-ignorance. *International Journal of Approximate Reasoning*. (accepted for pubblication).

[12] M. Troffaes and G. de Cooman. Extension of coherent lower previsions to unbounded random variables. In *Ninth International Conference IPMU 2002*, pages 735–742, Anney, France, 2002.

[13] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.

[14] P. Walley. Measures of uncertainty in expert systems. *Artificial Intelligence*, 83(1):1–58, 1996.

[15] L. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.

[16] L. A. Zadeh. On the validity of Dempster rule of combination. In *Memo M 79/24*, pages 3–28. Univ. of California, Berkeley, 1979.

[17] M. Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105(1):5–21, 2002.

Tests of the Mean with Distributional Uncertainty: An Info-Gap Approach

Yakov Ben-Haim
Yitzhak Moda'i Chair in Technology and Economics
Technion — Israel Institute of Technology
Haifa 32000 Israel

Abstract

Statistical tests of the mean are quite common. Sometimes the analyst cannot validate the assumptions underlying the test, such as normality, symmetry, independence of measurements, etc. This causes unknown deviation of the actual sampling distribution from the distribution assumed by the test, and thus unknown size and power of the test. This distributional uncertainty makes it difficult to reliably choose the decision threshold (critical value) and sample size. We present a method for evaluating the robustness of a test to an unknown degree of distributional uncertainty, based on info-gap decision theory. Analysis of robustness is useful in evaluating effective size and power, and for selecting the decision threshold and sample-size. We study binary simple-hypothesis tests of the mean and consider both type I and type II errors. We show quantitatively that robustness to distributional uncertainty improves, at fixed nominal level of significance, as the effective level of significance deteriorates. Likewise, robustness improves as the effective power of the test deteriorates. Furthermore, we show how to choose the decision threshold and sample size in light of distributional uncertainty. We illustrate our results by application to the t test and to a test of false nulls in epidemiology.

Keywords: binary hypothesis tests, distributional uncertainty, info-gaps, robustness, tests of the mean, t test, chronic wasting disease, false nulls.

1 Introduction

Statistical tests of the mean value of a population property are exceedingly common, and numerous tests are available. These tests depend on various assumptions about the data and the population. Sometimes normality is assumed, and almost invariably random sampling is posited: the measurements are made independently but with the same instrument and from the same population which is unaffected by the sample. However, in many situations the data generating process is not normal, or the sample is not random: the measurement instrument is not constant, or the sample is biased, or the measurements influence one another to some extent, or the statistical character of the population which is sampled is not constant. Determination of the level of significance and power of the test, and selection of the sample size, depend on the test which is used and its underlying assumptions. In some situations the analyst is unable to characterize the violation of test assumptions and is thus unable to adjust the test accordingly, and unable to reliably evaluate the level of significance and power or choose a sample size. We present a method for dealing with such situations.

Violation of the test assumptions can result in deviation of the actual sampling distribution from the distribution upon which the test is based. In situations where the violations are poorly known, the distributional deviations are similarly uncertain. We will refer to this as *distributional uncertainty.*

Considerable effort has been devoted to deriving methods which are robust to distributional uncertainty. Careful test design is a major antidote, though not always adequate. In some cases the distributional uncertainty can be characterized as a mixture of several (or many) distributions of known structure. Given adequate data, methods exist for estimating the parameters of the distributions and their weights in the mixture (Titterington *et al.* 1985). In other situations Monte Carlo methods are used to construct a sampling distribution based on prior knowledge of the distributional complexity (Robert, 2004). In these situations one can evaluate the robustness of a test as the extent of difference between simulated type I and type II error rates and the theoretical error rates in the absence of distributional uncertainty. Non-parametric methods exist which avoid

or weaken some assumptions about the sampled distribution. These tests do posit random sampling, or identity of two distinct distributions, or other assumptions (Johnson, 1995), and some are strictly valid only asymptotically. Numerical methods are available for evaluating the robustness of non-parametric statistics to specific violations, such as small-sample applications. However, non-parametric statistics can be very sensitive to a small number of outlying measurements. This focusses attention on the problem of long tails of the sampled distribution. The jacknife technique (Mooney and Duval, 1993), or trimmed means (DeGroot, 1986), attempt to rectify the effects of outliers. More generally, robustness can be evaluated as insensitivity to small deviations from the distributional assumptions (Huber, 1981), leading to M estimates and other techniques.

The distributional uncertainty on which we focus here is more unstructured than that for which these methods are explicitly designed. We illustrate our concept of distributional uncertainty, and its origin in ecological assessment and epidemiology, in section 2. Briefly, however, we consider situations in which the sampling distribution is highly uncertain and may be skewed, heavy tailed, multi-modal or non-random in ways which are unknown to the analyst. Distributional uncertainty, in the sense which concerns us here, arises for example in the use of historical data from diverse and unknown sources, taken with a variety of protocols (or lack of protocols in any professional sense), sampled from different and varying populations whose identity is imperfectly known. In such situations one must account for enormous and highly unstructured variability of the sampling distribution.

This sort of distributional uncertainty cannot be handled by the analysis of compound hypotheses. Distributional uncertainty presents us with an unbounded infinity of possible distributions—hypotheses—so it would seem impossible to formulate a compound hypothesis, or to identify a mixture of distributions.

We study two sets of problems. First, in the face of severe distributional uncertainty, what level of significance and power can one reliably ascribe to a binary simple-hypothesis test of the mean? We develop a method for quantitatively evaluating the reduction in level of significance and power, as distributional uncertainty increases. This analysis supports judgments about the effective level of significance and power, as expressed by their robustness to distributional uncertainty. Second, we show how to choose the decision threshold (critical value) and sample size when facing distributional uncertainty.

Our analysis employs info-gap decision theory for evaluating the robustness to large and highly unstructured uncertainty in the sampling distribution. We illustrate our results with simple t tests of the mean, but the methodology is applicable to a broad range of statistical tests.

We begin, in section 2, with an intuitive discussion of the origin and nature of distributional uncertainty. Section 3 formulates the binary hypothesis test which we study. Section 4 presents an info-gap model for distributional uncertainty. Section 5 formulates the info-gap robustness functions for type I and type II errors. A numerical example illustrating the decisions and judgments which the analyst must make is presented in section 6. A concluding discussion appears in section 8.

2 Origins of Distributional Uncertainty in Ecology and Epidemiology

Recall that by 'distributional uncertainty' we mean uncertainty in the form of the sampling distribution which results from unknown violations of assumptions underlying a statistical test. Distributional uncertainty is not uncommon in ecological assessment, arising from violations of test assumptions which the analyst is unable to characterize.

The main antidote to violation of test-assumptions is of course careful test design. This typically requires good basic understanding of the processes which are studied. However, measurements are sometimes made for the very purpose of augmenting our (sometimes quite deficient) understanding of these processes. For instance, Boone and Krohn (1999) show that the accuracy of model-based predictions of occurrence of avian species is a function of the frequency of species occurrences; not surprisingly, rare species are more difficult to model accurately than common species. Similarly, Craft *et al.* (1999) study the rate of restoration of ecological attributes in artificially constructed marshes as compared to natural marshes, noting that there are no long-term comparative studies. If the factors which influence long-term restoration and growth are incompletely understood, it may be difficult to characterize the relevant statistical properties of the control and test sites and to verify that they are equivalent. Finally, it is sometimes necessary to use very small samples, such as when data are based on large-scale natural experiments (Carpenter, 1989). Tests based on phenomena which are rare and poorly understood, or newly identified and unstudied, are vulnerable to distributional uncertainty.

There are also other potential causes of distributional uncertainty. Franklin (1999) uses a range of obser-

vational data from many different sources over the past 150 years—of varying accuracy and reliability—to evaluate change in bird assemblages in northern Australia. Some of these sources were trained biologists, though professional protocols changed over the sampling period. Some observers were casual or untrained observers who may exert less effort, and thus miss the rare events, or who are enthusiastic in the search for rare occurrences and may systematically over-report extreme observations. While historical observational data are an important and valuable source, it is difficult to verify that test-assumptions are not violated.

McCarthy (1998) uses museum collections to evaluate trends in marsupials and monotremes, recognizing that variable collection efforts introduce uncertainties. Similarly, Burgman *et al.* (1995) recognize that "collection frequencies will reflect changing trends in museum and herbarium collections", which introduces uncertainties in evaluating extinction threats based on historical development of collections. Stewart-Oaten *et al.* (1992) study tests of changes of a mean population property, before and after an impact, where the impact cannot be replicated (e.g., construction of a power plant). They note that data from such measurements "do not necessarily satisfy" the assumptions of standard tests. They state that "there is no panacea" for violation of test assumptions, and if the assumptions "are seriously wrong, alternative analyses are needed. This will often require a long time series of data." These authors discuss many sources of violation of test assumptions, stressing the importance of unknown skewness of distributions or correlations among measurements.

Evidence for violation of test assumptions is not uncommon in epidemiological studies. Bausch *et al.* (2003) report non-normal distributions of large samples, and non-random selection of participants, with disproportionate participation of particular subpopulations, due perhaps to the fear of stigma.

In short, analysts not infrequently face considerable uncertainty about the actual sampling distribution of their data. There surely is a true sampling distribution from which the data were obtained, but this distribution is unknown, and unknowable on the basis of available information. On the other hand, there is undoubtedly a population property—such as a mean—which is reflected in some way in the data. It is the aim of the statistical test to discriminate something about this population property, and to assess the confidence of this discrimination. A conventional statistical approach would be to transform the pdf, or modify the test, for formulate a compound or mix-distribution hypothesis, to accommodate violations of specific assumptions. We cannot do that because we don't know what specific violations have occurred. That's precisely the distributional uncertainty which we are studying.

3 Binary-Hypothesis Test

We have a set of measurements $X = \{x_1, \ldots, x_n\}$ which do not necessarily constitute a random sample of any known distribution, as discussed in sections 1 and 2. These data reflect a population mean, μ, but they suffer from an unknown degree of distributional uncertainty. We wish to use this data to decide between two simple hypotheses:

$$H_0: \quad \mu = T_0 \tag{1}$$
$$H_1: \quad \mu = T_1 \tag{2}$$

where each T_i is a specified number, and $T_1 > T_0$.

Let y be a statistic, for instance the t statistic, and let $F_i(y)$ denote the cumulative distribution function (cdf) of y under H_i. For any distribution $F(y)$, let $q_\alpha(F)$ denote the $(1-\alpha)$th quantile of $F(y)$:

$$q_\alpha(F) = \inf\{y: \ F(y) \geq 1-\alpha\} \tag{3}$$

We reject H_0 with significance α if:

$$y \geq q_\alpha(F_0) \tag{4}$$

The size, α, is the probability of falsely rejecting the null hypothesis, H_0, and the power, $1-\beta$, is the probability of correctly rejecting H_0. β is the probability of falsely rejecting H_1. If the cdf's are continuous at $q_\alpha(F)$ then the size α, and power, $1-\beta$, are:

$$1-\alpha = F_0[q_\alpha(F_0)] \tag{5}$$
$$\beta = F_1[q_\alpha(F_0)] \tag{6}$$

4 Info-Gap Models for Distributional Uncertainty

Suppose that the data X are not believed to be a random sample, or other assumptions underlying the test which is to be used are violated, but the nature of the violation is not known. In other words, suppose that the data are subject to distributional uncertainty. Let y be the test statistic (perhaps the t statistic, but not necessarily), and let $\widetilde{F}_i(y)$ denote the best (or perhaps only) guess of the distribution of the test statistic y, under hypothesis H_i. For instance, our best guess might be that $\widetilde{F}_0(y)$ is the t distribution with $n-1$ degrees of freedom for the regular t statistic $y = (\overline{x} - T_0)/(s/\sqrt{n})$ with sample mean and variance

$\overline{x}$ and s^2, while $\widetilde{F}_1(y) = \widetilde{F}_0(y-\delta)$ where $\delta = (T_1 - T_0)/(s/\sqrt{n})$.

$\widetilde{F}_i(y)$ is our best guess of the pdf of y but we don't know how wrong this guess is, and we have no "worst case" estimate. This distributional uncertainty in y, under hypothesis H_i, is represented by an info-gap model, $\mathcal{U}_i(h)$, which is an unbounded family of cdf's centered on $\widetilde{F}_i(y)$. For example, the uniform-bound info-gap model for uncertainty in the cdf of y is:

$$\mathcal{U}_i(h) = \left\{ F(y): \quad F(y) \in \mathcal{P},\ |F(y) - \widetilde{F}_i(y)| \le h, \quad \forall\ y \right\}, \quad h \ge 0 \tag{7}$$

where $\mathcal{P}$ is the set of all normalized non-negative cdf's. The info-gap model is an unbounded family of nested sets, $\mathcal{U}_i(h)$, of cdf's. In the absence of uncertainty, that is, when $h = 0$, the set is a singleton containing only the estimated cdf:

$$\mathcal{U}_i(0) = \{\widetilde{F}_i\} \tag{8}$$

The sets become more inclusive as the horizon of uncertainty increases:

$$h < h' \quad \text{implies} \quad \mathcal{U}_i(h) \subseteq \mathcal{U}_i(h') \tag{9}$$

The horizon of uncertainty, h, is unknown, so there is no known worst case or largest set of cdf's other than the set of all mathematically allowed cdf's (which occurs for $h \ge 1$). Eqs.(8) and (9) are the "contraction" and "nesting" axioms, respectively.

The uniform-bound info-gap model of eq.(7) entails enormous uncertainty in the cdf's. For sufficiently large h, the set $\mathcal{U}_i(h)$ contains densities which are highly asymmetric, multi-modal, with heavy or light tails, and with bumps, dimples, or "atoms" (infinite probability density at a single value of y) arbitrarily far from the mean resulting in arbitrarily large moments. Most importantly, the uncertainty in the cdf's which is represented by an info-gap model such as eq.(7) is different from estimation error or deviation from an asymptotic form. The info-gap model represents distributional uncertainty arising from unknown and possibly serious violation of fundamental assumptions underlying the hypothesis test. We do not motivate the structure of the info-gap model from consideration of estimation analytics or convergence (as in the Berry-Esseen inequality, Feller, 1971). Rather, the family of sets in eq.(7) reflects distributional uncertainty.

Other forms of info-gap model can be used if further information is available to constrain the relevant cdf's (Ben-Haim, 2006). For instance, one might have information indicating that the error of the estimated cdf is localized, e.g. on the tails, so the inequality in eq.(7) is modified in the envelope-bound info-gap model:

$$\mathcal{U}_i(h) = \left\{ F(y): \ F(y) \in \mathcal{P},\ |F(y) - \widetilde{F}_i(y)| \le h\psi(y), \quad \forall\ y \right\}, \quad h \ge 0 \tag{10}$$

where $\psi(y)$ is a known function. A related class of info-gap models is treated by Fox *et al.* (2007).

Alternatively one might make the judgment that probability atoms do not occur, but that the distribution may have bumps or dimples, or the tails may be heavy or light in unknown ways. An info-gap model which represents this is the fractional-error model applied to the probability density function (pdf) rather than to the cdf:

$$\mathcal{U}_i(h) = \left\{ f(y): \ f(y) \in \mathcal{D},\ |f(y) - \widetilde{f}_i(y)| \le h f_i^\star, \quad \forall\ y \right\}, \quad h \ge 0 \tag{11}$$

where $\mathcal{D}$ is the set of all normalized non-negative pdf's and $f_i^\star$ is a normalization constant with units of probability density. For instance, one might choose $f_i^\star = \max_y \widetilde{f}_i(y)$, which is the value of the pdf at its mode.

A much more restrictive info-gap model than eq.(11) is:

$$\mathcal{U}_i(h) = \left\{ f(y): \ f(y) \in \mathcal{D},\ |f(y) - \widetilde{f}_i(y)| \le h \widetilde{f}_i(y), \quad \forall\ y \right\}, \quad h \ge 0 \tag{12}$$

To understand the difference between the uncertainty models in eqs.(11) and (12), consider the case where y varies from $-\infty$ to $+\infty$ and the estimated pdf, $\widetilde{f}_i(y)$, has tails which diminish asymptotically to zero. The uncertainty set $\mathcal{U}_i(h)$ in eq.(11) allows bumps as large as $hf_i^\star$ arbitrarily far out on the tail. This is not the case for the set $\mathcal{U}_i(h)$ in eq.(12) for which a bump cannot be larger than $h\widetilde{f}_i(y)$ which will become very small for large y. The info-gap model of eq.(11) allows much more deviant tails than the info-gap model of eq.(12).

5 Robustnesses for Type I and Type II Errors: Formulation

Consider a test of size $\alpha^\star$, namely, a test which rejects H_0 when:

$$y \ge q_{\alpha^\star}(\widetilde{F}_0) \tag{13}$$

$\alpha^\star$ is the "nominal" size of the test since it is based on the best-estimate of the cdf under H_0, $\widetilde{F}_0$.

We now define the robustness of this test with respect to distributional uncertainty in $\widetilde{F}_0$, for falsely rejecting H_0. The robustness is the maximum horizon of uncertainty, h, up to which the test at nominal size $\alpha^\star$ falsely rejects H_0 with probability no greater than α:

$$\widehat{h}_0(\alpha^\star, \alpha) = \max\left\{h : \left(\min_{F\in\mathcal{U}_0(h)} F[q_{\alpha^\star}(\widetilde{F}_0)]\right) \geq 1-\alpha\right\} \tag{14}$$

We use the quantile $q_{\alpha^\star}(\widetilde{F}_0)$ because the test is implemented with the quantile of the best-guess distribution under H_0, $\widetilde{F}_0(y)$, and is of nominal size $\alpha^\star$, while the actual size (probability of falsely rejecting H_0) is then determined by the unknown true distribution under H_0, $F(y)$, which is info-gap-uncertain.

$\widehat{h}_0(\alpha^\star, \alpha)$ is related to type I error (falsely rejecting H_0). Specifically, $\widehat{h}_0(\alpha^\star, \alpha)$ is the greatest horizon of uncertainty up to which the probability of type I error is no greater than α. The following expression for $\widehat{h}_0(\alpha^\star, \alpha)$, for the info-gap model in eq.(7), is derived in appendix A:

$$\widehat{h}_0(\alpha^\star, \alpha) = \alpha - \alpha^\star \tag{15}$$

or zero if this is negative. We refer to α as the effective size, while $\alpha^\star$ is the nominal size. Section 6 explains how the analyst evaluates and chooses the effective size.

Note that, for any choice of $\alpha^\star$, the robustness curve for type-I error, $\widehat{h}_0(\alpha^\star, \alpha)$ vs. α, is entirely independent of the form of the test statistic. The implementation of the test, eq.(13), does depend on the type of test, through the value of the quantile $q_{\alpha^\star}(\widetilde{F}_0)$.

We now define a different robustness, related to type II error (falsely accepting H_0). $\widehat{h}_1(\alpha^\star, \beta)$ is the greatest horizon of uncertainty up to which the probability of falsely accepting H_0, with a test of nominal size $\alpha^\star$, is no greater than β:

$$\widehat{h}_1(\alpha^\star, \beta) = \max\left\{h : \left(\max_{F\in\mathcal{U}_1(h)} F[q_{\alpha^\star}(\widetilde{F}_0)]\right) \leq \beta\right\} \tag{16}$$

Let $1-\beta^\star$ be the nominal power:

$$1-\beta^\star = 1 - \widetilde{F}_1[q_{\alpha^\star}(\widetilde{F}_0)] \tag{17}$$

The following expression for $\widehat{h}_1(\alpha^\star, \beta)$, for the info-gap model in eq.(7), is derived in appendix B:

$$\widehat{h}_1(\alpha^\star, \beta) = 1 - \beta^\star - (1-\beta) \tag{18}$$

or zero if this is negative. We refer to $1-\beta$ as the effective power, while $1-\beta^\star$ is the nominal power. Section 6 explains how the analyst evaluates and chooses the effective power.

Note that, for any choice of $\alpha^\star$, the robustness curve for type-II error, $\widehat{h}_1(\alpha^\star, \beta)$ vs. β, depends on the form of the test, unlike for the type-I robustness. This is because the value of $\beta^\star$ depends on $\alpha^\star$ through the cdf's of the test statistic, $\widetilde{F}_0$ and $\widetilde{F}_1$.

6 Decisions and Judgments

The analyst must make two decisions and two judgments. The analyst must *decide* on the nominal test size $\alpha^\star$ and the sample size n. Together these decisions determine the decision threshold $q_{\alpha^\star}(\widetilde{F}_0)$ in eq.(13). Also, the analyst must *judge* what are reliable and acceptable values of the effective size α and effective power $1-\beta$ by considering the robustness functions $\widehat{h}_0(\alpha^\star, \alpha)$ and $\widehat{h}_1(\alpha^\star, \beta)$. (Recall that α is the probability of falsely rejecting H_0, while $1-\beta$ is the probability of correctly rejecting H_0.)

We will illustrate these decisions and judgments with an example employing the t test. The test statistic, y, is $(\overline{x} - T_0)(s/\sqrt{n})$ where $\overline{x}$ is the sample mean, s^2 is the sample variance, and n is the sample size. The estimated distribution under H_0, $\widetilde{F}_0(y)$, is the cdf of the t statistic with $n-1$ degrees of freedom. The estimated distribution under H_1 is $\widetilde{F}_1(y) = \widetilde{F}_0(y-\delta)$ where $\delta = (T_1 - T_0)/(s/\sqrt{n})$. The true distributions under H_0 and H_1 are unknown and the uncertainty in each cdf is represented by the info-gap model in eq.(7).

$\alpha^\star = 0.01$		$\alpha^\star = 0.05$	
n	$1-\beta^\star$	n	$1-\beta^\star$
5	0.1027	3	0.1784
7	0.3185	4	0.3736
9	0.5400	5	0.5390
12	0.7644	7	0.7457
31	0.9980	31	0.9997

Table 1: Size and power in the absence of distributional uncertainty.

The need for these judgments disappears in the absence of distributional uncertainty, since $\alpha^\star$ is the actual size and the actual power, $1-\beta^\star$, is entirely determined by $\alpha^\star$ and n. Values of $\alpha^\star$ and $1-\beta^\star$ are shown in table 1 for several sample sizes. Power, $1-\beta^\star$, improves (gets larger) as level of significance $\alpha^\star$ gets worse (gets larger) at fixed sample size n. Likewise, $1-\beta^\star$ improves as n increases at fixed $\alpha^\star$.

However, the presence of distributional uncertainty makes it necessary to form judgments on effective size α and power $1-\beta$. These judgments are based on the robustness functions, plots of which appear in figs. 1 and 2: $\widehat{h}_0(\alpha^\star, \alpha)$ vs. α (positive slope) and $\widehat{h}_1(\alpha^\star, \beta)$

vs. $1-\beta$ (negative slope).

Consider first the robustness curve for type-I error, $\widehat{h}_0(\alpha^\star,\alpha)$. The horizontal intercept of $\widehat{h}_0(\alpha^\star,\alpha)$ is the nominal size, $\alpha^\star$, because $\widehat{h}_0(\alpha^\star,\alpha^\star)=0$. This means that a test designed for size $\alpha^\star$ has no robustness to distributional uncertainty if one requires that the effective size actually equal $\alpha^\star$. The positive slope of $\widehat{h}_0(\alpha^\star,\alpha)$ vs. α means that positive robustness is obtained only for effective size, α, greater (worse) than the nominal size $\alpha^\star$. Stated differently, the positive slope of $\widehat{h}_0(\alpha^\star,\alpha)$ expresses a trade-off: the robustness against distributional uncertainty improves as the effective level of significance, α, get worse: robustness is exchanged for significance.

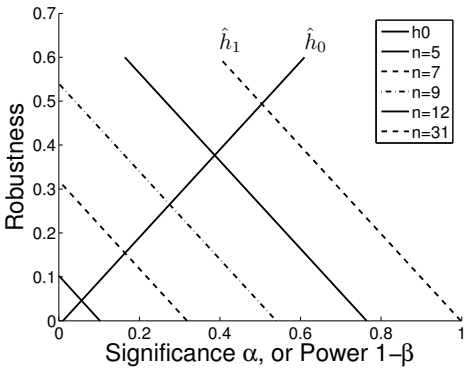


Figure 1: Robustness curves for the t test, $\widehat{h}_0(\alpha^\star,\alpha)$ for falsely rejecting H_0, and $\widehat{h}_1(\alpha^\star,\alpha)$ for falsely rejecting H_1. Nominal size is $\alpha^\star=0.01$. $\widehat{h}_1(\alpha^\star,\alpha)$ calculated at 5 different sample sizes: $n=5$, 7, 9, 12 and 31. $\delta=1$.

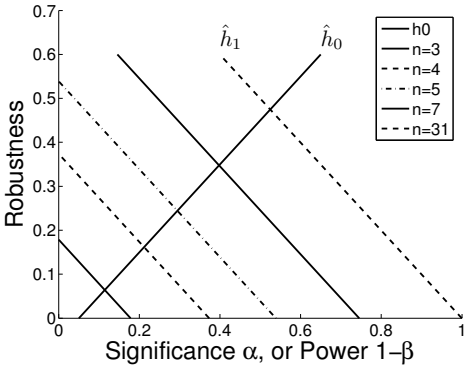


Figure 2: Robustness curves for the t test, $\widehat{h}_0(\alpha^\star,\alpha)$ for falsely rejecting H_0, and $\widehat{h}_1(\alpha^\star,\alpha)$ for falsely rejecting H_1. Nominal size is $\alpha^\star=0.05$. $\widehat{h}_1(\alpha^\star,\alpha)$ calculated at 5 different sample sizes: $n=3$, 4, 5, 7 and 31. $\delta=1$.

We now can see how one makes judgments of reliable effective size, α. A test designed for size $\alpha^\star=0.01$, as in fig. 1, has no robustness for size 0.01. However, consider an effective size $\alpha=0.05$ and refer to eq.(15). The test designed for $\alpha^\star=0.01$ will falsely reject H_0 with probability no greater than 0.05 if the actual cdf, $F(y)$, differs from the estimated cdf, $\widetilde{F}_0(y)$, by no more than 0.04 in cumulative probability. For instance, type I error will have probability no larger than 0.05 if the tails of the true distribution are too heavy or too light by no more than 4% of the total probability weight. The distributional uncertainty may arise from the presence of an outlying sub-population. The probability of type I error will not exceed 0.05 provided the sub-population is no larger than 4% of the total, regardless of how it is distributed. Similarly, at effective size $\alpha=0.1$, a test designed for size $\alpha^\star=0.01$ is robust to distributional uncertainty up to 0.09 in cumulative probability.

Now consider the robustness curves for type-II error, $\widehat{h}_1(\alpha^\star,\beta)$, eq.(18). The horizontal intercept of $\widehat{h}_1(\alpha^\star,\beta)$ is the nominal power, $1-\beta^\star$, because $\widehat{h}_1(\alpha^\star,\beta^\star)=0$. This means that a test designed for size $\alpha^\star$ has no robustness to distributional uncertainty if one requires that the effective power actually equal $1-\beta^\star$. The negative slope of $\widehat{h}_1(\alpha^\star,\beta)$ vs. $1-\beta$ means that positive robustness is obtained only for effective power, $1-\beta$, lower (worse) than the nominal power $1-\beta^\star$. Stated differently, the negative slope of $\widehat{h}_1(\alpha^\star,\beta)$ expresses a trade-off: the robustness against distributional uncertainty improves as the effective power, $1-\beta$, get worse: robustness is exchanged for power.

We can now see how one makes judgments of reliable effective power, $1-\beta$. A test designed for size $\alpha^\star=0.01$ with sample size $n=9$ (dot-dash in fig. 1), has no robustness for power 0.54 (the horizontal intercept and nominal power). However, consider an effective power $1-\beta=0.44$ and refer to eq.(18). This test will falsely accept H_0 with probability of 0.44 if the actual cdf differs from the estimated cdf by no more than 0.1. At effective size $1-\beta=0.44$, this test is robust to distributional uncertainty up to 0.1 in cumulative probability. For instance, if the tails err by as much as 10% of the total probability, or if a sub-population with unknown distribution has no more than 10% weight, then the probability of type II error will be no more than 0.44. Similarly, at effective size $1-\beta=0.34$, this test is robust to distributional uncertainty up to 0.2 in cumulative probability.

Finally, let us consider the choice of the sample size. Only the type-II robustness is influenced by the sample size, as we see from eqs.(15) and (18) and from figs. 1 and 2. The nominal and effective power both

increase with increasing sample size, and are also substantially influenced by the nominal size $\alpha^\star$ as we see by comparing the two figures. The analyst decides on the sample size in light of the effective power and robustness which are needed. We illustrate the decisions and judgments with the aid of fig. 3, which is expanded from fig. 1.

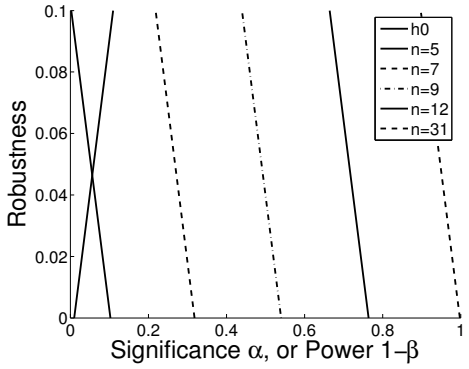


Figure 3: Expanded from fig. 1.

In fig. 3 we are contemplating the choice of nominal size $\alpha^\star = 0.01$. Consider the judgment that effective size $\alpha = 0.05$ is adequate and reliable because the robustness is $\widehat{h}_0(0.01, 0.05) = 0.04$, eq.(15). This judgment considers the robustness and the effective size together since they are linked through the trade-off between them. For instance, the judgment is that the tails are unlikely to err by more than 4%, and the 5% risk of type I error is acceptable. Now apply this robustness to type II error by requiring $\widehat{h}_1(\alpha^\star, \beta) = 0.04$. From fig. 3 we find effective powers of 0.50, 0.72 and 0.96 for sample sizes 9, 12 and 31. Judging that power of 0.50 is too small, we require a sample larger than $n = 9$. If power of 0.72 is adequate then we adopt a sample of size 12. Choosing a sample of size 31 would result in power of 0.96.

Let us continue our consideration of the judgment in the previous paragraph that effective size $\alpha = 0.05$ is adequate and reliable. Judgment is subjective, and this is a two-fold judgment since size and robustness are linked through the trade-off between them. Size, α, is subjectively judged in terms of the risk of type I error. Robustness in this case can also be subjectively judged in terms of probability. For instance one might make the judgment that the distribution is distorted by an outlying sub-population whose weight is no more than a few percent of the main population. This robustness judgment can be cast in terms of risk: by accepting a robustness of 0.04 we are accepting the risk that the parent population is contaminated by an outlying population whose weight is no more than 4%.

It may be convenient and familiar for some analysts to judge robustness in this example in terms of probability and risk as just described, However, this is not necessary. Info-gap models of uncertainty are inherently non-probabilistic, and value judgments about robustness can be formed non-probabilistically. Judgments of acceptable risk are based on experience and context. In the same way, analysts can acquire subjective feel for fractional error, or other non-probabilistic quantities, which leads to judgments of acceptable robustness. The concept of analogical inference has been employed to form non-probabilistic value judgments of robustness (Ben-Haim, 2006, chap. 4).

Let us now return to our discussion of choosing the sample size, three paragraphs before, and remove a simplification which we made: applying the same robustness to both type I and type II errors. Having accepted robustness of 0.04 for type I error, $\widehat{h}_0(0.01, 0.05) = 0.04$, we then evaluated the sample size in terms of the same robustness for type II error, $\widehat{h}_1(\alpha^\star, \beta) = 0.04$. This is justified if one faces the same severity of distributional uncertainty for both hypotheses. However, one might well image situations in which the distributional uncertainty is different for the two hypotheses. For instance, one hypothesis may represent a "healthy" state which is more thoroughly studied than the "unhealthy" state represented by the other hypothesis. In such a situation one makes separate judgments of robustness and its trade-off partner (either size or power) for each hypothesis. The judgment of effective size, α, is linked to a judgment of $\widehat{h}_0$-robustness. Then one chooses the sample size to yield what is judged to be acceptable type-II robustness, $\widehat{h}_1$, at acceptable power.

7 Example: Chronic Wasting Disease

Verbal description. Chronic wasting disease (CWD) in deer can be detected by inoculating a particular strain of mice with an extract from the antler velvet of the infected deer. The prion protein (PrP) which is characteristic of this disease is expressed in the mice after a time t which is randomly distributed. This distribution is highly uncertain, and it has been observed that PrP expression with antler velvet from diseased deer frequently does not occur even anomalously long after the mean time (Angers *et al.* 2009). The expression of PrP is much more reliable if the injections are made from the brains of the deer. However, brains may not be available. For instance, antler velvet is used in various traditional Asian medicines which may be the only source for testing.

Suppose that we have inoculated n mice and after incubation times $t_1, \ldots, t_n$, no expression of the PrP is observed in any of the mice. How confident are we

that CWD is not present in the deer?

System model. Let $p(t)$ denote the probability density function (pdf) of the incubation time, with cumulative distribution function (cdf) $P(t)$. We assume that the incubation times are statistically independent, so the probability of a false null—true presence with no observed expression of the PrP—is:

$$P_{\rm fn}(t_1,\,\ldots,\,t_n) = \prod_{i=1}^{n}[1-P(t_i)] \tag{19}$$

Uncertainty model. Let $\widetilde{p}(t)$ and $\widetilde{P}(t)$ denote the estimated pdf and cdf. Let $t_{\rm s}$ denote a point on the upper tail beyond which the estimated pdf is quite uncertain. For instance we might choose $t_{\rm s}$ to be 2 standard deviations from the mean. We will define an info-gap model in which there are functions whose upper tail, beyond $t_{\rm s}$, decays as $1/t^2$, much slower than the decay of exponential or normal distributions.

Let $\mathcal{P}$ denote the set of non-negative normalized pdf's. The info-gap model, for $h \ge 0$, is:

$$\mathcal{U}(h) = \left\{p:\ p\in\mathcal{P}, p(t) \le \widetilde{p}(t) + \frac{t_{\rm s}h}{t^2}\forall t \ge t_{\rm s}\ \right\} \tag{20}$$

The first condition assures that the functions are mathematically legitimate pdf's. The second condition allows the upper tail, beyond $t_{\rm s}$, to exceed the exponential by as much as $t_{\rm s}h/t^2$, conditional on the rest of the distribution being able to adjust to assure non-negativity and normalization.

Note that $\int_{t_{\rm s}}^{\infty} t_{\rm s}h/t^2\,{\rm d}t = h$. Thus the horizon of uncertainty, h, represents the fraction of the entire statistical weight which is uncertain. For instance, if the uncertainty of the pdf is thought of as an uncertain mixture of populations, then h is the fraction of the non-$\widetilde{p}$ population.

Performance requirement. The probability of a false null must be less than a critical value:

$$P_{\rm fn}(t_1,\,\ldots,\,t_n) \le P_{\rm fnc} \tag{21}$$

Robustness function. The robustness is defined as:

$$\widehat{h}(n,P_{\rm fnc}) = \max\left\{h:\ \left(\max_{p\in\mathcal{U}(h)} P_{\rm fn}\right) \le P_{\rm fnc}\right\} \tag{22}$$

We will evaluate the inverse of $\widehat{h}(n,P_{\rm fnc})$.

Let us denote the inner maximum in eq.(22) by $m(h)$, which is the inverse of $\widehat{h}(n,P_{\rm fnc})$. We will assume that all the observed times, $t_1,\,\ldots,\,t_n$, exceed $t_{\rm s}$, so they fall in the domain of the uncertain tail. In this case, $m(h)$ is evaluated with the upper envelope at horizon of uncertainty h, provided that this distribution can be normalized. For each individual observation:

$$\begin{aligned}\max_{p\in\mathcal{U}(h)}[1-P(t_i)] &= \min\left[1,\ \int_{t_i}^{\infty}\left(\widetilde{p}(t)+\frac{t_{\rm s}h}{t^2}\right)\,{\rm d}t\right]\\ &= \min\left[1,\ 1-\widetilde{P}(t_i)+\frac{t_{\rm s}h}{t_i}\right] \end{aligned}\tag{23}$$

Since the n observations are independent we find the inner maximum in eq.(22) to be:

$$m(h) = \prod_{i=1}^{n}\min\left[1,\ 1-\widetilde{P}(t_i)+\frac{t_{\rm s}h}{t_i}\right] \tag{24}$$

Plotting $m(h)$ vs h is equivalent to $P_{\rm fnc}$ vs $\widehat{h}(n,P_{\rm fnc})$.

Eq.(24) can be simplified when the observations, t_i, are large, so that $1-\widetilde{P}(t_i)$ is nearly zero. For $h \le 1$:

$$m(h) \approx \frac{t_{\rm s}^n h^n}{\prod_{i=1}^{n} t_i} \tag{25}$$

Equating this to $P_{\rm fnc}$ and solving for h yields an approximate expression for the robustness which is valid when the observations are large:

$$\widehat{h}(n,P_{\rm fnc}) \approx \frac{1}{t_{\rm s}}\left(P_{\rm fnc}\prod_{i=1}^{n} t_i\right)^{1/n} \tag{26}$$

Denoting the geometric mean of the n observations by $\overline{t}_{\rm gm}$, this becomes:

$$\widehat{h}(n,P_{\rm fnc}) \approx \frac{\overline{t}_{\rm gm}}{t_{\rm s}}P_{\rm fnc}^{1/n} \tag{27}$$

The geometric mean observation, $\overline{t}_{\rm gm}$, will change as the sample grows, but the dominant effect of sample size is in the term $P_{\rm fnc}^{1/n}$ which grows rapidly as n increases when n and $P_{\rm fnc}$ are small. Furthermore, when $P_{\rm fnc}$ is very small, the slope of $\widehat{h}$ vs $P_{\rm fnc}$ increases as n increases. This means that, when $P_{\rm fnc}$ is small, the cost of robustness, in units of increased $P_{\rm fnc}$, is small when n is large.

Example. Fig. 4 shows robustness curves, based on eq.(24), for 5 sample sizes with the following data t_i = 500, 530, 510, 520, 505 days. The bottom curve (n = 1) uses only the first datum; the next curve uses the first 2 data; etc. The estimated distribution is normal with mean and standard deviation of 450 and 20 days. $t_{\rm s}$ = 490.

The positive slopes of the curves express the trade-off between robustness, $\widehat{h}$, and critical probability of false null, $P_{\rm fnc}$. Large robustness is obtained only by accepting large $P_{\rm fnc}$. The robustness is zero at the estimated value of $P_{\rm fnc}$.

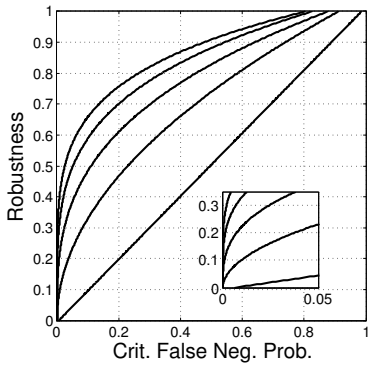


Figure 4: $\widehat{h}(n, P_{\text{fnc}})$ vs P_{fnc}, $n = 1$ to 5 (bottom to top).

The robustness increases substantially as the sample size increases from $n = 1$ to 2. The marginal increase in robustness decreases with increasing n. From the insert in the figure we see that the slope of the robustness curve increases dramatically as the sample size increases. A high slope means that the robustness can be increased without significantly increasing the critical probability of false null, P_{fnc}.

8 Methodological Conclusion

This paper concentrates on binary simple-hypothesis statistical tests, subject to distributional uncertainty, by which we mean uncertainty in the sampling distribution resulting from unknown violations of the test assumptions. We have focussed on two decisions and two judgments which the analyst must make. How can one *decide upon* the decision threshold and the sample size, and how does one *judge* the effective size and power of a test? We have developed a generic approach to these questions based on info-gap decision theory, and illustrated the method with the t test and with a test for false nulls. The method can be applied to other tests as well.

Consider a test which is designed to have nominal level of significance $\alpha^\star$. The robustness of this test with respect to distributional uncertainty, for falsely rejecting H_0 in eq.(1), is denoted $\widehat{h}_0(\alpha^\star, \alpha)$ and defined in eq.(14). $\widehat{h}_0(\alpha^\star, \alpha)$ is the greatest horizon of distributional uncertainty up to which the test, with nominal size $\alpha^\star$, falsely rejects H_0 with probability no greater than α. That is, $\widehat{h}_0(\alpha^\star, \alpha)$ is the greatest horizon of uncertainty up to which the probability of type I error (false rejection of H_0) is no greater than α, when using a test with nominal size $\alpha^\star$.

$\widehat{h}_0(\alpha^\star, \alpha)$ is necessarily zero when $\alpha = \alpha^\star$, implying that the test has no robustness to distributional uncertainty at its nominal size, $\alpha^\star$. The robustness is positive for $\alpha > \alpha^\star$, and the robustness increases as α gets larger. This expresses the trade-off between robustness to distributional uncertainty on the one hand, and effective level of significance on the other hand, as illustrated by eq.(15) and the lines of positive slope in figs. 1–3.

The robustness function $\widehat{h}_0(\alpha^\star, \alpha)$ is the basic tool for choosing the decision threshold, $q_{\alpha^\star}(\widetilde{F}_0)$ in eq.(13), and for evaluating the effective size, α, of the test. If $\widehat{h}_0(\alpha^\star, \alpha)$ is large then one has confidence that the probability of falsely rejecting H_0 is no greater than α. What constitutes a 'large' robustness, and 'how large is large enough' are delicate value judgments, somewhat like the choice of level of significance. We discussed this in section 6, though there is no absolute answer.

We have also considered the robustness to distributional uncertainty in evaluating the effective power. For any test designed for size $\alpha^\star$, the robustness to distributional uncertainty, for falsely accepting H_0 (type II error), is denoted $\widehat{h}_1(\alpha^\star, \beta)$, defined in eq.(16). The power, $1 - \beta$, is the probability of correctly rejecting H_0. $\widehat{h}_1(\alpha^\star, \beta)$ is the greatest horizon of distributional uncertainty up to which the test, with nominal size $\alpha^\star$, will falsely accept H_0 with probability no greater than β. The robustness is zero when β is the value obtained, at size $\alpha^\star$, in the absence of distributional uncertainty. That is, there is no robustness for the nominal power. The robustness increases as the power decreases, as illustrated by eq.(18) and the lines of negative slope in figs. 1–3.

The robustness functions $\widehat{h}_1(\alpha^\star, \beta)$ and $\widehat{h}_0(\alpha^\star, \alpha)$ are the basic tools for choosing the sample size and for evaluating the effective power of a test. If $\widehat{h}_1(\alpha^\star, \beta)$ is large then one has confidence that the probability of correctly rejecting H_0 is no less than $1 - \beta$ with the chosen sample size. Once again, judgments of adequate power and large robustness are subjective.

We have concentrated on tests of the mean with binary simple hypotheses, both because such tests are exceedingly common in practice, and because the main aim was to demonstrate the methodology of info-gap theory for evaluating effective size and power and for selecting the decision threshold and sample-size. The methodology developed in this paper can be extended to other test structures, and to tests of quantities other than the mean. Furthermore, the close relation between hypothesis tests and confidence intervals enables the application of the methodology to evaluating and selecting confidence intervals.

Acknowledgements

The author is indebted to Mark A. Burgman, Ayala Cohen, David Fox, Malka Gorfine, Mick McCarthy, Andrew Robinson and Miriam Zacksenhouse for valuable comments. Funding for this research was

provided by the U.S. Department of Agriculture under USDA/ERS/PREISM Cooperative Agreement No.58-7000-8-0095.

9 References

Angers, Rachel C., *et al.*, 2009, Chronic Wasting Disease Prions in Elk Antler Velvet, *Emerging Infectious Diseases,* Vol. 15, No. 5, pp.696–703.

Bausch, Daniel G. *et al.* 2003, Risk Factors for Marburg Hemorrhagic Fever, Democratic Republic of the Congo, *Emerging Infectious Diseases,* Vol. 9, No. 12, pp.1531–1537.

Ben-Haim, Yakov (2006). *Info-Gap Decision Theory: Decisions Under Severe Uncertainty,* 2nd edition, London: Academic Press.

Boone, Randall B. and William B. Krohn (1999). Modeling the occurrence of bird species: are the errors predictable? *Ecological Applications* **9**, 835–848.

Burgman, Mark A., Roger C. Grimson and Scott Ferson (1995). Inferring threat from scientific collections, *Conservation Biology* **9**, 923–928.

Carpenter, Stephen R. (1989). Replication and treatment strength in whole-lake experiments (1989). *Ecology* **70**, 453–463.

Craft, Christopher, Judy Reader, John N. Sacco and Stephen W. Broome (1999). Twenty-five years of ecosystem development of constructed *Spartina alterniflora* (Loisel) marshes, *Ecological Applications* **9**, 1405–1419.

DeGroot, Morris H. (1986). *Probability and Statistics,* 2nd ed., Reading, MA: Addison-Wesley.

Feller, William (1971). *An Introduction to Probability Theory and Its Applications,* vol. 2, 2nd ed. New York: Wiley.

Fox, David R., Yakov Ben-Haim, Keith R. Hayes, Michael McCarthy, Brendan Wintle, Piers Dunstan (2007). An info-gap approach to power and sample size calculations, *Environmetrics* **18**, 189–203.

Franklin, Donald C. (1999). Evidence of disarray amongst granivorous bird assemblages in the savannas of northern Australia, a region of sparse human settlement, *Biological Conservation* **90**, 53–68.

Huber, Peter J. (1981). *Robust Statistics,* New York: Wiley.

Johnson, Douglas H. (1995). Statistical sirens: The allure of nonparametrics, *Ecology* **76**, 1998–2000.

McCarthy, Michael A. (1998). Identifying declining and threatened species with museum data, *Biological Conservation* **83**, 9–17.

Mooney, Christopher Z. and Robert D. Duval (1993). *Bootstrapping: A Nonparametric Approach to Statistic Inference,* London: Sage Publications.

Robert, Christian P. (2004). *Monte Carlo Statistical Methods,* 2nd ed., New York: Springer.

Stewart-Oaten, Allan, James R. Bence, and Craig W. Osenberg (1992). Assessing effects of unreplicated perturbations: No simple solutions, *Ecology* **73**, 1396–1404.

Titterington, D.M., A.F.M. Smith, U.E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions,* Chichester: Wiley.

A Evaluating the Robustness $\widehat{h}_0(\alpha^\star, \alpha)$ for Falsely Rejecting H_0

In this appendix we derive $\widehat{h}_0(\alpha^\star, \alpha)$ based on the info-gap model in eq.(7).

First $V(x) = 0$ if $x < 0$, $V(x) = x$ if $0 \le x \le 1$, $V(x) = 1$ if $x > 1$.

Let $m_0(h)$ denote the inner minimum in the definition of the robustness in eq.(14). The robustness, $\widehat{h}_0(\alpha^\star, \alpha)$, is the greatest horizon of uncertainty, h, at which $m_0(h) \ge 1 - \alpha$. $m_0(h)$ decreases with increasing h because the sets $\mathcal{U}_0(h)$ of the info-gap model become more inclusive as h increases (the nesting axiom). Hence the robustness is the greatest non-negative value of h for which $m_0(h) = 1 - \alpha$. If there is no such value of h, then the robustness is zero.

The inner minimum in eq.(14) is obtained when $F(y)$ is as small as possible at $q_{\alpha^\star}(\widetilde{F}_0)$, subject to membership in $\mathcal{U}_0(h)$. From the info-gap model in eq.(7) we find:

$$m_0(h) = V\left(\widetilde{F}_0[q_{\alpha^\star}(\widetilde{F}_0)] - h\right) = V(1 - \alpha^\star - h) \quad (28)$$

where we recall that $\widetilde{F}_0[q_{\alpha^\star}(\widetilde{F}_0)] = 1 - \alpha^\star$. The greatest value of h at which $m_0(h) = 1 - \alpha$ is the robustness, eq.(15).

B Evaluating the Robustness $\widehat{h}_1(\alpha^\star, \beta)$ for Correctly Rejecting H_0

In this appendix we derive $\widehat{h}_1(\alpha^\star, \beta)$ based on the info-gap model in eq.(7).

Let $m_1(h)$ denote the inner maximum in the definition of the robustness in eq.(16). The nesting axiom implies that $m_1(h)$ increases monotonically as h increases. Consequently the robustness, $\widehat{h}_1(\alpha^\star, \beta)$, is the greatest horizon of uncertainty, h, at which $m_1(h) = \beta$.

From the info-gap model in eq.(7), and using the step function $V(x)$ defined earlier, we find:

$$m_1(h) = V\left(\widetilde{F}_1[q_{\alpha^\star}(\widetilde{F}_0)] + h\right) \quad (29)$$

Equating this to β and solving for h we find the robustness in eq.(18) with the aid of the expression for the nominal power in eq.(17).

On general conditional random quantities

Veronica Biazzo
Dip. Mat. Inf.
Viale A. Doria, 6
95125 Catania (Italy)
vbiazzo@dmi.unict.it

Angelo Gilio
Dip. Me. Mo. Mat.
Via A. Scarpa, 16
00161 Roma (Italy)
gilio@dmmm.uniroma1.it

Giuseppe Sanfilippo
Dip. Sc. Stat. Mat.
Viale delle Scienze, ed. 13
90128 Palermo (Italy)
sanfilippo@unipa.it

Abstract

In the first part of this paper, recalling a general discussion on iterated conditioning given by de Finetti in the appendix of his book, vol. 2, we give a representation of a conditional random quantity $X|HK$ as $(X|H)|K$. In this way, we obtain the classical formula $\mathbb{P}(XH|K) = \mathbb{P}(X|HK)P(H|K)$, by simply using linearity of prevision. Then, we consider the notion of general conditional prevision $\mathbb{P}(X|Y)$, where X and Y are two random quantities, introduced in 1990 in a paper by Lad and Dickey. After recalling the case where Y is an event, we consider the case of discrete finite random quantities and we make some critical comments and examples. We give a notion of coherence for such more general conditional prevision assessments; then, we obtain a strong generalized compound prevision theorem. We study the coherence of a general conditional prevision assessment $\mathbb{P}(X|Y)$ when Y has no negative values and when Y has no positive values. Finally, we give some results on coherence of $\mathbb{P}(X|Y)$ when Y assumes both positive and negative values. In order to illustrate critical aspects and remarks we examine several examples.

Keywords. conditional events, general conditional random quantities, general conditional prevision assessments, generalized compound prevision theorem, iterated conditioning, strong generalized compound prevision theorem.

1 Introduction

This paper takes as its starting point the definition of *general conditional prevision* introduced by Lad and Dickey in [16] and also considered by Lad in his book [17]. In these works, the authors propose a general theory of conditional prevision specifying its operational meaning. This theory, which considers conditional prevision of the form $\mathbb{P}(X|Y)$ where both X and Y are random quantities, generalizes the de Finetti's definition of a conditional prevision assertion $\mathbb{P}(X|H)$, where H is an event. We observe that, denoting the indicator of H by the same symbol, to assume "H true" amounts to assuming $(H = 1)$ true, that is $(H \neq 0)$ true. Then, in the approach of Lad and Dickey, $X|H$ can be looked at as $X|Y$, where Y is the indicator of H; hence, $\mathbb{P}(X|H) = \mathbb{P}[X|(H = 1)]$. Notice that we discard the case where Y is the constant 0, as it reduces to the case $X|H$ where $(H \neq 0)$ is impossible. We recall that, concerning (precise or imprecise) conditional probability or prevision assessments like $P(E|H)$ or $\mathbb{P}(X|H)$, where E and H are events and X is a random quantity, theoretical results and algorithms in the framework of coherence have been given by many authors (see, for instance, [2, 3, 4, 5, 6, 8, 9, 10, 19, 20, 21, 22]) The checking of coherence and the extension of precise conditional prevision assessments have been studied in [7].

In [16, 17] the general conditional prevision $\mathbb{P}(X|Y)$ is defined as a number that you specify asserting your willingness to engage any transaction yielding a suitable random net gain and it is shown that such a generalization answers to questions of decision problems involving "state dependent preferences". In his book ([17]), Lad introduces the notion of general conditional random quantity $X|Y$ from the definition of conditional prevision $\mathbb{P}(X|Y)$. Obviously, as usual in a subjective setting, engaging a transaction requires a coherency of your assertion. In [16, 17], the coherency of $\mathbb{P}(X|Y)$ requires that a generalized compound prevision theorem is satisfied, that is the quantities $\mathbb{P}(XY)$, $\mathbb{P}(Y)$ and $\mathbb{P}(X|Y)$ must be such that $\mathbb{P}(XY) = \mathbb{P}(X|Y)\mathbb{P}(Y)$. But, the general case is different from the case where Y is the indicator of an event H. In fact, $\mathbb{P}(H) = 0$ implies $\mathbb{P}(XH) = 0$, and using coherence ([15, 18]) we can directly assess $\mathbb{P}(X|H)$. On the contrary, $\mathbb{P}(Y) = 0$ doesn't imply that $\mathbb{P}(XY) = 0$ and it could happen that it doesn't exist a finite value of $\mathbb{P}(X|Y)$ which satisfies the generalized compound prevision theorem. Thus, in this paper we propose a notion of coherence in order to handle the case $\mathbb{P}(Y) = 0$, integrating the Lad's defi-

nition of $\mathbb{P}(X|Y)$. Then, we give a strong generalized compound prevision theorem which follows from our definition of coherence. The random quantities, like X and Y, considered in this paper are finite discrete. The paper is organized as follows. In section 2 we recall some preliminary concepts and results. In section 3 we deepen, in the setting of coherence, the operational meaning of the assessments $\mathbb{P}(X|H)$ and $\mathbb{P}(X|HK)$, where H and K are events and X is a random quantity; then, based on a general discussion on *iterated conditioning* given by de Finetti in ([12], Vol. 2, Appendix, section 13), we look at $B|AH$ and $X|HK$, respectively, as $(B|A)|H$ and $(X|H)|K$); then, we give a representation for $B|AH$ and $X|HK$ which allows to obtain the classical results $\mathbb{P}(AB|H) = \mathbb{P}(B|AH)P(A|H)$ and $\mathbb{P}(XH|K) = \mathbb{P}(X|HK)P(H|K)$, by simply applying the linearity of prevision. In section 4, we recall the definitions of conditional prevision $\mathbb{P}(X|Y)$ and conditional random quantity $X|Y$; then, we examine a critical example. In section 5, after some critical comments, we propose an explicit definition of coherence for the conditional prevision $\mathbb{P}(X|Y)$; then, we give a strong generalized compound prevision theorem; we also examine many examples to illustrate some further aspects. In section 6, we study the coherence of a conditional prevision assessment $\mathbb{P}(X|Y) = \mu$, when Y has no negative values, or Y has no positive values. In section 7, we give some results concerning the coherence of the assessment $\mathbb{P}(X|Y) = \mu$, where Y assumes both positive and negative values. In section 8, we show some results concerning the set of coherent prevision assessments on $X|Y'$, where Y' is a linear transformation of Y. Finally, in section 9 we give some conclusions and an outlook on future research, which should concern more in general the case of imprecise conditional prevision assessments on families of conditional random quantities.

2 Some preliminary notions

We assume that each random quantity has a finite set of possible values. We denote by Ω (resp., $\emptyset$) the sure (resp., impossible) event; moreover, we denote by A^c the negation of A and by $A \vee B$ (resp., AB) the disjunction (resp., the conjunction) of A and B. We use the same symbol to denote an event and its indicator. We recall that in the subjective approach to probability, your assessment $P(E|H) = p$ means that You accept a bet on the conditional event $E|H$ in which You pay an amount ps, with $s \neq 0$, by receiving the random quantity $sHE + psH^c$, so that your net random gain is

$$G = sHE + psH^c - ps = sH(E - p)\,.$$

By excluding trivial cases, the value of G is, respectively, $s(1-p)$, or $-ps$, or 0, according to whether EH is true, or E^cH is true, or H^c is true.
We recall that, considering the restricted random gain $G|H = s(E-p) \in \{s(1-p), -ps\}$, it is $\min G|H \cdot \max G|H = -s^2p(1-p)$. Then, the coherence of p is defined by the condition ([15, 18]): $\min G|H \cdot \max G|H \leq 0$; that is $p(1-p) \geq 0$, which amounts to: $0 \leq p \leq 1$.
We observe that, to determine the coherent values of p, we don't consider all the values of G, but only those of $G|H$; in other words the value 0 of G associated with the case "H false" is "discarded".
We also observe that, denoting by the same symbol the (conditional) events and their indicators, by choosing $s = 1$ we obtain

$$E|H = EH + pH^c = EH + (1 - H)p\,,$$

where the indicator, or truth-value function, $E|H$ represents the quantity we receive when we pay the amount $p = P(E|H)$. Then, by the linearity of prevision, we obtain:

$$P(E|H) = P(EH) + [1 - P(H)]p\,,$$

that is: $P(EH) = P(H)P(E|H)$ (*compound probability theorem*). We recall that, starting with a pioneering work of de Finetti ([11]), the notion of conditional event as a three-valued (logical and/or numerical) entity has been proposed by many authors (see, e.g., [1], [13], [14]). Based on the betting scheme, the notions of conditional prevision and conditional random quantity are defined and widely exploited in [17]. Truth-values of conditional events and their extension to decomposable conditional measures of uncertainty, with the aim of finding reasonable axioms for a general theory, have been discussed in many papers by Coletti and Scozzafava, see e.g. [9].

3 Representation of conditional random quantities

We remark that the general formula $P(AB|H) = P(A|H)P(B|AH)$ can be obtained by using the general coherence condition for conditional probability assessments. The same formula can be obtained, based on the linearity of prevision, by the following refined reasoning. Let $\mathcal{P} = (x, y, z)$ a probability assessment on $\mathcal{F} = \{A|H, B|AH, AB|H\}$. We observe that representing the indicator $B|AH$ as

$$B|AH = ABH + (1 - AH)y\,,$$

we obtain

$$P(B|AH) = y = P(ABH) + [1 - P(AH)]y\,,$$

from which it follows: $P(ABH) = P(AH)y$, i.e. $zP(H) = xyP(H)$; hence, to reach the conclusion we need to assume $P(H) > 0$. To bypass this obstacle, based on the general discussion on *iterated conditioning* given by de Finetti in ([12], Appendix of Vol. 2, section 13), we can look at $B|AH$ as $(B|A)|H$. Moreover, defining $p = P(B|A)$, we have $B|A = AB + (1 - A)p$. Of course, when we pass from $B|A$ to $B|AH$, we must replace p by y. Then

$$B|AH = (B|A)|H = AB|H + [(1 - A)|H]y = \\ = AB|H + (A^c|H)y = (AB + yA^c)|H\,. \tag{1}$$

The representation above is not surprising, as shown by the following remarks:
(i) with the family $\mathcal{F}$ we can associate the partition $\{ABH, AB^cH, A^cH, H^c\}$;
(ii) under the hypothesis "H true", the random quantities $B|AH$ and $(AB + yA^c)|H$ coincide, as they always assume the same value, that is 1, or 0, or y, according to whether ABH is true, or AB^cH is true, or A^cH is true.
Hence, it must be: $\mathbb{P}(B|AH) = \mathbb{P}[(AB + yA^c)|H]$, with $\mathbb{P}(B|AH) = P(B|AH) = y$ and

$$\mathbb{P}(AB + yA^c)|H = \mathbb{P}(AB|H) + \mathbb{P}(yA^c|H) = \\ = P(AB|H) + yP(A^c|H) = z + y(1 - x)\,.$$

Then, we obtain: $y = z + y(1 - x)$, i.e. $z = xy$.
Notice that, based on this result, we have that $B|AH$ and $(AB + yA^c)|H$ coincide also when H^c is true. In fact, the value of $B|AH$ (resp., $(AB+yA^c)|H$) associated with H^c is y (resp., $z+y(1-x) = y+z-xy = y$).
Now, by generalizing the previous reasoning, given an event H and a discrete finite random quantity $X \in \{x_1, x_2, \ldots, x_n\}$, in the subjective approach the conditional prevision assessment $\mu = \mathbb{P}(X|H)$ is the amount to be payed in order to receive the random quantity $X|H = XH + (1 - H)\mu$. The random gain is $G = X|H - \mu = XH - \mu H$ and, as before, the coherence condition for μ is: $\min G|H \cdot \max G|H \leq 0$, which amounts to: $\min X|H \leq \mu \leq \max X|H$.
Of course, we have

$$\mathbb{P}(X|H) = \mu = \mathbb{P}[XH + (1 - H)\mu] = \\ = \mathbb{P}(XH) + \mathbb{P}(1 - H)\mu = \mathbb{P}(XH) + \mu - P(H)\mu\,,$$

from which it follows the well known formula: $\mathbb{P}(XH) = P(H)\mu = P(H)\mathbb{P}(X|H)$.
More in general, given two events H, K and a random quantity X, let $\mathcal{M} = (x, y, z)$ a conditional prevision assessment on $\mathcal{F} = \{H|K, X|HK, XH|K\}$.
By the same kind of reasoning, we have

$$X|HK = (X|H)|K = [XH + (1 - H)y]|K = \\ = XH|K + yH^c|K\,. \tag{2}$$

In fact, as for the case of conditional events, we can show that the conditional random quantities $X|HK$ and $[XH + (1 - H)y]|K$ coincide by the following remarks:
(i) we denote by $\{x_1, \ldots, x_n\}$ the set of possible values of X and, for the sake of simplicity by $\{x_1, \ldots, x_r\}$ (resp., $\{x_1, \ldots, x_r, \ldots, x_t\}$) the set of values of X compatible with HK (resp., with K), where $r \leq t \leq n$; moreover, we set $E_i = (X = x_i)$ and with the family $\mathcal{F}$ we associate the partition (of the sure event Ω) $\{E_1HK, \ldots, E_rHK, H^cK, K^c\}$;
(ii) we have $X = \sum_{i=1}^n x_iE_i$ and $XH = \sum_{i=1}^n x_iE_iH$; then

$$X|HK = \sum_{i=1}^{r} x_iE_iHK + (1 - HK)y\,;$$

$$XH|K + yH^c|K = \sum_{i=1}^{r} x_iE_iHK + (1 - K)z + \\ + yH^cK + (1 - K)y(1 - x)\,;$$

(iii) assuming "K true", if H is true, then $X = x_i$ for some $i \leq r$ and $X|HK = [XH + (1 - H)y]|K = x_i$; if H is false, then $X = x_i$ for some i, with $r < i \leq t$, and $X|HK = XH|K + yH^c|K = y$; hence, under the hypothesis "K true", $X|HK$ and $[XH + (1-H)y]|K$ coincide. Then

$$\mathbb{P}(X|HK) = y = \mathbb{P}([XH + (1 - H)y]|K) = \\ = \mathbb{P}(XH|K) + yP(H^c|K) = z + y(1 - x)\,,$$

from which it follows: $z = xy$, that is:

$$\mathbb{P}(XH|K) = \mathbb{P}(X|H)P(H|K)\,.$$

Notice that, by the previous formula, if K is false we have $X|HK = y$ and

$$XH|K + yH^c|K = z + y(1 - x) = y + z - xy = y\,.$$

Therefore, the conditional random quantities $X|HK$ and $XH|K+yH^c|K = (XH+yH^c)|K$ coincide in all cases.

4 General conditional random quantities

Let be given two random quantities X and Y. In [17] it is proposed the notion of general conditional random quantity $X|Y$ based on the following definition for the prevision of $X|Y$, introduced in [16].

Definition 1. The conditional prevision for X given Y, denoted $\mathbb{P}(X|Y)$, is a number you specify with the understanding that you accept to engage any transaction yielding a random net gain $G = sY[X - \mathbb{P}(X|Y)]$.

The following definition is given for the conditional random quantity $X|Y$.

Definition 2. Having asserted your conditional prevision $\mathbb{P}(X|Y) = \mu$, the conditional random quantity $X|Y$ is defined as

$$X|Y = XY + (1-Y)\mu = \mu + Y(X-\mu)\,. \quad (3)$$

Notice that, if Y assumes only the value 0, that is $Y \equiv 0$, you can pay every real number $\mu = \mathbb{P}(X|Y)$, as you always receive the same amount μ; in fact, the net gain is always 0. To avoid this trivial case we will assume that $(Y = 0) \neq \Omega$.

We remark that such a general notion of conditional random quantity reduces to the classical one $X|H = XH + (1-H)\mu$ when Y coincides with an event H. Lad remarks that the direction of the net gain (or loss) depends on the difference $(X-\mu)$, while the scale depends on the numerical value of Y. Lad also remarks that for $Y = 0$ (resp., $Y = 1$) the net gain is 0 (resp., $s(X-\mu)$), i.e. the possible net gains obtained when Y is an event. Then, by computing the prevision on both sides of (3), Lad obtains

$$\mu = \mu + \mathbb{P}[Y(X-\mu)] = \mu + \mathbb{P}(XY) - \mu\mathbb{P}(Y)\,,$$

so that $\mathbb{P}(XY) = \mathbb{P}(X|Y)\mathbb{P}(Y)$, which becomes $\sum_j p_j y_j \mathbb{P}[X\,|\,(Y = y_j)] = \mathbb{P}(X|Y)\sum_j p_j y_j$, where $p_j = P(Y = y_j)$. This condition, which we call "generalized compound prevision theorem", generalizes the classical one $\mathbb{P}(XH) = \mathbb{P}(X|H)P(H)$, where H is an event. Then, when $\mathbb{P}(Y) \neq 0$ it immediately follows $\mathbb{P}(X|Y) = \frac{\mathbb{P}(XY)}{\mathbb{P}(Y)}$ (actually, we will see that the generalized compound prevision theorem holds in a *stronger* sense). Several properties are obtained by Lad, under the condition $\mathbb{P}(Y) \neq 0$. We also notice that, when X and Y are uncorrelated, i.e. $Cov(X,Y) = 0$, it is $\mathbb{P}(XY) = \mathbb{P}(X)\mathbb{P}(Y)$; then, under the hypothesis $\mathbb{P}(Y) \neq 0$, it follows $\mathbb{P}(X|Y) = \mathbb{P}(X)$. We can say that, *under the condition $\mathbb{P}(Y) \neq 0$, X and Y are uncorrelated if and only if the prevision of 'X given Y' coincides with the prevision of X.*

We examine below an example, in which Y is not an event, to illustrate a critical aspect.

Example 1. We recall that by the formula $\mathbb{P}(XH) = P(H)\mathbb{P}(X|H)$, when $P(H) > 0$ it follows $\mathbb{P}(X|H) = \frac{\mathbb{P}(XH)}{P(H)}$. Moreover, if $P(H) = 0$, then $\mathbb{P}(XH) = 0$; in this case, based on coherence principle ([15, 18]) and assuming $\emptyset \neq H \neq \Omega$, it can be proved that the assessment $(0, 0, \mu)$ on $\{H, XH, X|H\}$ is coherent if and only if: $\min X|H \leq \mu \leq \max X|H$. But, replacing H by a random quantity Y, we are in a *very different situation*, as $\mathbb{P}(Y) = 0$ *doesn't imply* $\mathbb{P}(XY) = 0$. To illustrate this aspect, let us consider a random vector

$$(X,Y) \in \mathcal{C} = \{(0,-1),(0,1),(1,-1),(1,1)\}\,,$$

with

$$p(0,-1) = \frac{1}{3},\ p(0,1) = \frac{1}{6},\ p(1,-1) = \frac{1}{6},\ p(1,1) = \frac{1}{3},$$

where $p(x,y) = P(X = x, Y = y)$. We denote the joint distribution of (X,Y) by the vector $(\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{3})$. We have

$$Y \in \mathcal{C}_Y = \{-1, 1\}\,,\ XY \in \mathcal{C}_{XY} = \{-1, 0, 1\}\,,$$

with $P(Y = -1) = P(Y = 1) = \frac{1}{2}$, and with $P(XY = -1) = \frac{1}{6}$, $P(XY = 0) = \frac{1}{2}$, $P(XY = 1) = \frac{1}{3}$, so that $\mathbb{P}(Y) = 0$ and $\mathbb{P}(XY) = \frac{1}{6}$. In this case, it doesn't exist any finite value $\mathbb{P}(X|Y)$ which satisfies the equality $\mathbb{P}(XY) = \mathbb{P}(X|Y)\mathbb{P}(Y)$. In fact, given any assessment $\mathbb{P}(X|Y) = \mu$, the values of $Y(X-\mu)$ associated with that of (X,Y) are, respectively, $\mu, -\mu, -1+\mu, 1-\mu$; then, assuming (for the sake of simplicity) $s = 1$, one has

$$\mathbb{P}(G) = \mathbb{P}[Y(X-\mu)] =$$

$$= \frac{1}{3}\mu + \frac{1}{6}(-\mu) + \frac{1}{6}(-1+\mu) + \frac{1}{3}(1-\mu) = \frac{1}{6} \neq 0\,,\ \ \forall\,\mu\,.$$

Hence, by starting with a joint probability distribution on (X,Y), it may happen that the equation $\mathbb{P}(XY) = \mathbb{P}(X|Y)\mathbb{P}(Y)$ has no finite solutions in the unknown $\mathbb{P}(X|Y)$.

If we assign the joint distribution $(\frac{1}{3} - \varepsilon, \frac{1}{6} + \varepsilon, \frac{1}{6}, \frac{1}{3})$ on (X,Y), with $\varepsilon \in [-\frac{1}{6}, 0) \cup (0, \frac{1}{3}]$, we obtains

$$P(Y = -1) = \frac{1}{2} - \varepsilon\,,\ P(Y = 1) = \frac{1}{2} + \varepsilon\,,\ \ \mathbb{P}(Y) = 2\varepsilon\,,$$

$$P(Y = -1) = \frac{1}{2} - \varepsilon\,,\ P(Y = 1) = \frac{1}{2} + \varepsilon\,,\ \ \mathbb{P}(Y) = 2\varepsilon\,,$$

while the distribution of XY doesn't change; moreover,

$$\mathbb{P}(G) = (\frac{1}{3} - \varepsilon)\mu + (\frac{1}{6} + \varepsilon)(-\mu) + \frac{1}{6}(-1+\mu) + \frac{1}{3}(1-\mu) =$$

$$= \frac{1}{6} - 2\varepsilon\mu = \mathbb{P}(XY) - \mathbb{P}(Y)\mathbb{P}(X|Y)\,,$$

and imposing $\mathbb{P}(G) = 0$, it follows

$$\mu = \mathbb{P}(X|Y) = \frac{1}{12\,\varepsilon}\,,\quad \varepsilon \in [-\frac{1}{6}, 0) \cup (0, \frac{1}{3}]\,.$$

In particular, for $\varepsilon \in [-\frac{1}{6}, 0)$ it is $\mu \in (-\infty, -\frac{1}{2}]$, while for $\varepsilon \in (0, \frac{1}{3}]$ it is $\mu \in [\frac{1}{4}, +\infty)$.

Finally, if we assign a uniform distribution on (X,Y), that is

$$p(0,-1) = p(0,1) = p(1,-1) = p(1,1) = \frac{1}{4}\,,$$

it follows $\mathbb{P}(Y) = \mathbb{P}(XY) = 0$; then, the equality $\mathbb{P}(XY) = \mathbb{P}(Y)\mathbb{P}(X|Y)$ becomes $0 = 0 \cdot \mathbb{P}(X|Y)$. In this case, we need a *direct* assessment of $\mathbb{P}(X|Y)$ and the problem of *coherence* arises. This basic problem will be addressed in the next section.

5 Coherence of general conditional prevision assessments

A crucial problem arises when $\mathbb{P}(Y) = 0$; what can be said about coherence of a given assessment $\mathbb{P}(X|Y) = \mu$? We remark that this case has not been examined in the book of Lad. We also observe that when Y equals 0 Lad notices that the net gain is 0 without further comments. But, concerning the classical case of a conditional random quantity $X|H$, in order to check the coherence of the assessment $\mathbb{P}(X|H) = \mu$, as is well known the value 0 of the net gain associated with the case $H = 0$ is discarded by the set of values of the net gain G, i.e. coherence checking is based on the values of $G|H$. Hence, in order to integrate the analysis of Lad by properly managing the case $\mathbb{P}(Y) = 0$, we propose:
(i) to give an explicit definition of coherence for a given assessment $\mathbb{P}(X|Y) = \mu$;
(ii) to discard, in the definition of coherence, the value 0 of the net gain associated with the case $Y = 0$.
Then, based on [15, 18], we give the following

Definition 3. Given two random quantities X, Y and a conditional prevision assessment $\mathbb{P}(X|Y) = \mu$, let $G = s(X|Y - \mu) = sY(X - \mu)$ be the net random gain, where s is an arbitrary real quantity, with $s \neq 0$. Defining the event $H = (Y \neq 0)$, the assessment $\mathbb{P}(X|Y) = \mu$ is coherent if and only if: $\inf G|H \cdot \sup G|H \leq 0$, for every s.

In what follows, without loss of generality, we will set $s = 1$.

5.1 A strong generalized compound prevision theorem

Based on Definition 3, we will obtain a stronger version of the generalized compound prevision theorem. We recall that H is the event $(Y \neq 0)$; then, we make the following reasoning (where we assume that $\mu, \mathbb{P}(Y|H)$, and $\mathbb{P}(XY|H)$ are finite):
(i) by Definition 3, μ is the quantity to be payed, in order to receive $X|Y$, under the hypothesis H true; hence, *operatively* μ is the prevision of $X|Y$, *conditional on* H; (ii) hence, a more appropriate representation of $X|Y$ is given by: $X|Y = [\mu + Y(X - \mu)]|H$; (iii) then, by computing the prevision on both sides of the previous equality, we have:

$$\mu = \mathbb{P}(X|Y) = \mathbb{P}[\mu + Y(X-\mu)|H] = \mu + \mathbb{P}[Y(X-\mu)|H]\,,$$

so that $\mathbb{P}[Y(X - \mu)|H] = \mathbb{P}[(XY - \mu Y)|H] = 0$; then, by the linearity of prevision, it follows

$$\mathbb{P}(XY|H) = \mathbb{P}(X|Y)\mathbb{P}(Y|H)\,. \tag{4}$$

Notice that, if Y is a finite discrete random quantity, with $Y \geq 0$, or $Y \leq 0$, surely it is $\mathbb{P}(Y|H) \neq 0$; then, by (4) it follows $\mathbb{P}(X|Y) = \frac{\mathbb{P}(XY|H)}{\mathbb{P}(Y|H)}$.
We recall that H^c is the event $(Y = 0)$; moreover, we observe that $\mathbb{P}(Y|H^c) = \mathbb{P}(XY|H^c) = 0$; hence,

$$\begin{aligned} &\mathbb{P}(Y) = \mathbb{P}(Y|H)P(H) + \mathbb{P}(Y|H^c)P(H^c) = \\ &= \mathbb{P}(Y|H)P(H) = \mathbb{P}(YH)\,, \end{aligned} \tag{5}$$

$$\begin{aligned} &\mathbb{P}(XY) = \mathbb{P}(XY|H)P(H) + \mathbb{P}(XY|H^c)P(H^c) = \\ &= \mathbb{P}(XY|H)P(H) = \mathbb{P}(XYH)\,. \end{aligned} \tag{6}$$

Then, by (4), (5), and (6), one has

$$\mathbb{P}(XY|H)P(H) = \mathbb{P}(X|Y)\mathbb{P}(Y|H)P(H)\,,$$

that is, the formula $\mathbb{P}(XY) = \mathbb{P}(X|Y)\mathbb{P}(Y)$, given in [16] and [17], which we call *weak* generalized compound prevision theorem.

5.2 Some examples and remarks

In the finite case, denoting respectively by $\mathcal{C}_X, \mathcal{C}_Y$ and $\mathcal{C}$ the sets of possible values of X, Y and (X, Y), with each $(x_h, y_k) \in \mathcal{C}$ it is associated for the net gain G the value $g_{hk} = y_k(x_h - \mu)$. We set $\mathcal{C}_0 = \{(x_h, y_k) \in \mathcal{C} : y_k \neq 0\}$; of course $\mathcal{C}_0 \subseteq \mathcal{C}$. Then, by Definition 3, the assessment μ is coherent if and only if: $m \leq 0 \leq M$, where

$$m = \min_{(x_h,y_k)\in\mathcal{C}_0} y_k(x_h - \mu)\,, \quad M = \max_{(x_h,y_k)\in\mathcal{C}_0} y_k(x_h - \mu)\,.$$

We denote by Π the set of coherent assessments μ; then, we remark that, assuming $\mathcal{C}_0 \neq \emptyset$, the assessment $\mu = x_h$ is coherent, as it trivially satisfies the condition of coherence (it is $g_{hk} = 0$, $\forall\,(x_h, y_k) \in \mathcal{C}_0$). Hence, $\mathcal{C}_X \subseteq \Pi$.

Example 2. Given a random vector $(X, Y) \in \mathcal{C} = \{(-1, 0), (1, 1)\}$, consider the assessment $\mathbb{P}(Y|X) = \mu$ on the conditional random quantity $X|Y$. We have $H = (Y \neq 0)$; hence $\mathcal{C}_0 = \{(1, 1)\}$. Moreover, one has $G = Y(X - \mu) \in \{0, 1 - \mu\}$, with $G|H = 1 - \mu$. We observe that Y coincides with the indicator of H, so that $X|Y = X|H$. Then, by Definition 3, μ is coherent if and only if $1 - \mu = 0$, that is $\mu = 1$. Notice that this result is consistent with the usual approach to the notion of conditional prevision.

Remark 1. Notice that in Example 2, while the coherence condition $\inf G|H \cdot \sup G|H \leq 0$ is satisfied uniquely with $\mu = 1$, the condition $\inf G \cdot \sup G \leq 0$ is satisfied for every μ. Then, if the condition $\inf G|H \cdot \sup G|H \leq 0$ were replaced by $\inf G \cdot \sup G \leq 0$, it would follow that every assessment $\mathbb{P}(X|Y) = \mu$ would be coherent, which is clearly unreasonable (however, as we will show by other examples, still applying the condition $\inf G|H \cdot \sup G|H \leq$

0, it may be $\Pi = \mathbb{R}$). Example 2 confirms that, in order to look at $X|Y$ as $X|H$ in the usual sense, when checking coherence we must discard the value 0 of the random gain G associated with the case $Y = 0$. In this way, we can look at the family of conditional random quantities like $X|H$, where H is an event, as a sub-family of the family of general conditional random quantities like $X|Y$, where Y is a random quantity.

We recall that, given any event $H \neq \emptyset$, if X is a constant, say $X = c$, then $\mathbb{P}(X|H) = c$. The following example shows that, if $X = c$ and Y is a random quantity, with $\min Y < 0 < \max Y$, then the assessment $\mathbb{P}(X|Y) = \mu$ is coherent for every $\mu \in \mathbb{R}$.

Example 3. Given $(X, Y) \in \mathcal{C} = \{(c, -y_1), (c, y_2)\}$, with $c \in \mathbb{R}$ and $y_1, y_2 > 0$, consider the coherence of any assessment $\mathbb{P}(X|Y) = \mu$. We have $\mathcal{C}_0 = \mathcal{C}$, so that $H = (Y \neq 0) = \Omega$ and $G|H = G = Y(c - \mu)$. The values of $G|H$ are: $-y_1(c-\mu), y_2(c-\mu)$, and the coherence condition $\inf G|H \cdot \sup G|H \leq 0$ is satisfied for every $\mu \in \mathbb{R}$. Moreover, given a joint distribution on (X, Y), say $(p, 1-p)$, where

$$p = P(X = c, Y = -y_1)\,, \quad 1-p = P(X = c, Y = y_2),\,,$$

with $0 \leq p \leq 1$, we have $\mathbb{P}(Y) = y_2 - p(y_1 + y_2)$ and

$$\mathbb{P}(XY) = c\mathbb{P}(Y) = c[y_2 - p(y_1 + y_2)]\,.$$

Then, if $p \neq \frac{y_2}{y_1+y_2}$, one has $\mathbb{P}(Y) \neq 0$ and c is the unique coherent value of μ associated with the distribution $(p, 1-p)$. Whereas, if $p = \frac{y_2}{y_1+y_2}$, then $\mathbb{P}(Y) = \mathbb{P}(XY) = 0$, and the assessment $\mathbb{P}(X|Y) = \mu$, associated with the distribution $(\frac{y_2}{y_1+y_2}, \frac{y_1}{y_1+y_2})$, is coherent for every $\mu \in \mathbb{R}$.

Example 4. We continue the study of Example 1, by examining the coherence of a given assessment $\mathbb{P}(X|Y) = \mu$. We recall that $(X, Y) \in \mathcal{C} = \{(0, -1), (0, 1), (1, -1), (1, 1)\}$; moreover, we observe that $\mathcal{C}_0 = \mathcal{C}$, as $H = (Y \neq 0) = \Omega$ and hence $G|H = G = Y(X - \mu)$. With the values of (X, Y) are associated respectively the following values of $G|H$: $\mu, -\mu, -1 + \mu, 1 - \mu$; hence, the coherence condition $\inf G|H \cdot \sup G|H \leq 0$ is satisfied for every μ.

Example 5. We assume that $(X, Y) \in \mathcal{C} = \{(0, -1), (1, 1)\}$, by examining the coherence of a given assessment $\mathbb{P}(X|Y) = \mu$. We have $\mathcal{C}_0 = \mathcal{C}$; so that $H = (Y \neq 0) = \Omega$ and we have $G|H = G = Y(X - \mu)$. The values of $G|H$ are: $\mu, 1 - \mu$ and, as it can be verified, the coherence condition $\inf G|H \cdot \sup G|H \leq 0$ is satisfied if and only if $\mu \notin (0, 1)$, that is μ is coherent if and only if $\mu \in (-\infty, 0] \cup [1, +\infty)$. In this example with each coherent assessment μ it is associated a unique joint distribution on (X, Y), say $(p, 1-p)$, where

$$p = P(X = 0, Y = -1)\,,$$

$$1 - p = P(X = 1, Y = 1)\,, \quad 0 \leq p \leq 1\,.$$

The parameter p is determined by requiring that the prevision of the random gain be 0, that is

$$p\mu + (1 - p)(1 - \mu) = 0\,. \tag{7}$$

As it can be verified, one has

$$p = f(\mu) = \frac{1-\mu}{1-2\mu}\,;$$

moreover, when $\mu \leq 0$ it is $\frac{1}{2} < p \leq 1$; when $\mu \geq 1$ it is $0 \leq p \leq \frac{1}{2}$. Notice that

$$\mu = f^{-1}(p) = \frac{1-p}{1-2p}\,;$$

that is: $f^{-1} = f$. This result depends on the symmetry of the equation (7) with respect to p and μ.

As shown by Example 5, the set Π *of the coherent assessments* μ *may be not convex.*
To better analyze this aspect, in what follows we examine separately two cases:
(*i*) $Y \geq 0$, or $Y \leq 0$; (*ii*) $\min Y < 0 < \max Y$.

6 The case $Y \geq 0$, or $Y \leq 0$.

We assume $X \in \mathcal{C}_X = \{x_1, \ldots, x_n\}$ and $Y \in \mathcal{C}_Y = \{y_1, \ldots, y_r\}$, with $y_k \geq 0\,,\ \forall\, k$. Moreover, we denote by X^0 the subset of $\mathcal{C}_X$ such that for each $x_h \in X^0$ there exists $(x_h, y_k) \in \mathcal{C}_0$. Then, we set

$$x_0 = \min X^0\,, \quad x^0 = \max X^0\,. \tag{8}$$

We first consider the case $Y \geq 0$; we have

Theorem 1. Given two finite random quantities X, Y, with $Y \geq 0$, the prevision assessment $\mathbb{P}(X|Y) = \mu$ is coherent if and only if $x_0 \leq \mu \leq x^0$.

Proof. Given any μ, with each pair $(x_h, y_k) \in \mathcal{C}_0$ we associate the inequality $y_k(x_h - \mu) \geq 0$. Under the hypothesis $Y \neq 0$ it is $y_k > 0$; then the inequality is satisfied if and only if $\mu \leq x_h$. We observe that, for each $x_h \in X^0$, there exists (at least) a value $y_k > 0$ such that $(x_h, y_k) \in \mathcal{C}_0$. Then, we distinguish three cases: (i) $\mu < x_0$; (ii) $\mu > x^0$; (iii) $x_0 \leq \mu \leq x^0$. In the first case it is $y_k(x_h - \mu) > 0$ for every $(x_h, y_k) \in \mathcal{C}_0$, so that $\inf G|H \cdot \sup G|H > 0$ and hence μ is not coherent. In the second case it is $y_k(x_h - \mu) < 0$ for every $(x_h, y_k) \in \mathcal{C}_0$, so that $\inf G|H \cdot \sup G|H > 0$ and hence μ is not coherent. In the third case, denoting by y_k and y_s two positive values of Y such that $(x_0, y_k) \in \mathcal{C}_0$, $(x^0, y_s) \in \mathcal{C}_0$, it is $y_k(x_0 - \mu) \leq 0\,,\ y_s(x^0 - \mu) \geq 0$, so that $\inf G|H \leq 0\,,\ \sup G|H \geq 0$ and hence $\inf G|H \cdot \sup G|H \leq 0$. Therefore, for every $\mu \in [x_0, x^0]$, μ is coherent. □

We illustrate the previous result by the following

Example 6. Given a random vector $(X,Y) \in \mathcal{C} = \{(0,1),(1,0),(1,1),(2,2)\}$, let us determine the set Π of coherent prevision assessment $\mathbb{P}(X|Y) = \mu$ on $X|Y$. We observe that $X^0 = X$, so that $x_0 = \min \mathcal{C}_X = 0$, $x^0 = \max \mathcal{C}_X = 2$; moreover, it is $\mathcal{C}_0 = \{(0,1),(1,1),(2,2)\}$ and the values of $Y(X-\mu)$, under the restriction $(X,Y) \in \mathcal{C}_0$ are, respectively, $-\mu, 1-\mu, 2(2-\mu)$; such values are all positive (resp., all negative) when $\mu < 0$ (resp., $\mu > 2$); hence each $\mu \notin [0,2]$ is not coherent. Finally, when $\mu \in [0,2]$ one has $-\mu(2-\mu) \leq 0$, so that the condition $\inf G|H \cdot \sup G|H \leq 0$ is satisfied. Hence, we have $\Pi = [x_0, x^0] = [0,2]$.

We now consider the case $Y \leq 0$; we have

Theorem 2. Given two finite random quantities X, Y, with $Y \leq 0$, the conditional prevision assessment $\mathbb{P}(X|Y) = \mu$ is coherent if and only if $x_0 \leq \mu \leq x^0$.

Proof. We observe that, as $-Y \geq 0$, by Theorem 1 the assessment $\mathbb{P}(X|-Y) = \mu$ is coherent if and only if $x_0 \leq \mu \leq x^0$. On the other hand, defining $G'|H = -Y(X-\mu)|H$, we have $G|H = Y(X-\mu) = -G'|H$. Then

$$\inf G|H = -\sup G'|H\,, \quad \sup G|H = -\inf G'|H\,,$$

and hence: $\inf G|H \cdot \sup G|H = \inf G'|H \cdot \sup G'|H$; thus, the assessment $\mathbb{P}(X|H) = \mu$ is coherent if and only if $x_0 \leq \mu \leq x^0$. □

7 The case $\min Y < 0 < \max Y$.

We now examine the general case in which there exist positive and negative values of Y. We set

$$X^- = \{x_h \in \mathcal{C}_X : \exists (x_h, y_k) \in \mathcal{C}_0, y_k < 0\}\,,$$
$$X^+ = \{x_h \in \mathcal{C}_X : \exists (x_h, y_k) \in \mathcal{C}_0, y_k > 0\}\,;$$
$$\mathcal{C}^- = \{(x_h, y_k) \in \mathcal{C}_0 : y_k < 0\}\,,$$
$$\mathcal{C}^+ = \{(x_h, y_k) \in \mathcal{C}_0 : y_k > 0\}\,.$$

Of course, $\mathcal{C}^- \cap \mathcal{C}^+ = \emptyset$ and $\mathcal{C}^- \cup \mathcal{C}^+ = \mathcal{C}_0$. We have

Theorem 3. Let be given two random quantities X, Y, with $\min Y < 0 < \max Y$. If $X^- \cap X^+ \neq \emptyset$, then the conditional prevision assessment $\mathbb{P}(X|Y) = \mu$ is coherent, for every real number μ.

Proof. Let be given $x_h \in X^- \cap X^+, y_k \in \mathcal{C}_Y, y_t \in \mathcal{C}_Y$ such that $(x_h, y_k) \in \mathcal{C}^-$ and $(x_h, y_t) \in \mathcal{C}^+$; moreover, let μ be any real number. It is $g_{hk} g_{ht} = y_k(x_h - \mu) \cdot y_t(x_h - \mu) = y_k y_t (x_h - \mu)^2 \leq 0$, so that $\inf G|H \cdot \sup G|H \leq 0$. Therefore, for every $\mu \in \Re$, μ is coherent. □

We illustrate the previous result by the following

Example 7. We determine the set Π of coherent prevision assessment $\mathbb{P}(X|Y) = \mu$ on $X|Y$, where $(X,Y) \in \mathcal{C} = \{(0,1),(0,-1),(1,-1),(1,1)\}$, as in Example 1. We have $X^- = X^+ = \{0,1\}$, so that $X^- \cap X^+ \neq \emptyset$; hence, by Theorem 3, $\Pi = \Re$.

In what follows, we examine the cases

$$\min X^- = \max X^+\,, \quad \max X^- = \min X^+\,;$$

then, we study in depth the case $X^- \cap X^+ = \emptyset$. Given any $(x_h, y_s) \in \mathcal{C}^-$, $(x_k, y_t) \in \mathcal{C}^+$, we set

$$m_{hk} = \min\{x_h, x_k\}\,, \quad M_{hk} = \max\{x_h, x_k\}\,;$$

moreover, we denote by I_{hk} the open interval (m_{hk}, M_{hk}). Then, we set

$$I = \bigcap_{x_h \in X^-,\, x_k \in X^+} I_{hk}\,. \tag{9}$$

Notice that, defining

$$\begin{aligned} \mu_0 &= \max\nolimits_{x_h \in X^-,\, x_k \in X^+} m_{hk}\,, \\ \mu^0 &= \min\nolimits_{x_h \in X^-,\, x_k \in X^+} M_{hk}\,, \end{aligned} \tag{10}$$

one has $I \neq \emptyset$ if and only if $\mu_0 < \mu^0$ and, in this case, $I = (\mu_0, \mu^0)$. We have

Theorem 4. Let the quantities μ_0, μ^0 be defined as in (10); then $\mu_0 = \min(\max X^-, \max X^+)$ and $\mu^0 = \max(\min X^-, \min X^+)$.

Proof. We first prove that μ_0 coincides with $\min(\max X^-, \max X^+)$. Let be $x_k = \max X^+$, $x_h = \max X^-$. Then $x_r \leq x_k, \forall x_r \in X^+$ and $x_t \leq x_h, \forall x_t \in X^-$. Let be $\min(\max X^-, \max X^+) = x_h$. Then, there exists $x_r \in X^+$ such that $x_r \geq x_h$, i.e. there exist $(x_h, y_s) \in C^-$ and $(x_r, y_t) \in C^+$, such that $m_{hr} = x_h$. Suppose that $\mu_0 \neq x_h$, i.e. $\mu_0 \neq \min(\max X^-, \max X^+)$; then $\mu_0 > x_h$ and, as $x_h = \max X^-$, it must be $\mu_0 = x_t$ for some $x_t \in X^+$. Then, there exist $(x_v, y_r) \in C^-$, $(x_t, y_s) \in C^+$ such that $x_t \leq x_v$. From $x_v \leq x_h$, it is $x_t \leq x_v \leq x_h$, i.e. $\mu_0 \leq x_h$, which is absurd; hence $\mu_0 = \min(\max X^-, \max X^+)$. The proof is similar if $\min(\max X^-, \max X^+) = x_k$, where $x_k = \max X^+$. We now prove that $\mu^0 = \max(\min X^-, \min X^+)$. Let be $x_k = \min X^+$, $x_h = \min X^-$. Then $x_r \geq x_k, \forall x_r \in X^+$ and $x_t \geq x_h, \forall x_t \in X^-$. Let be $\max(\min X^-, \min X^+) = x_h$. Then, there exists $x_r \in X^+$ such that $x_r \leq x_h$, i.e. there exist $(x_h, y_s) \in C^-$ and $(x_r, y_t) \in C^+$, such that $M_{hr} = x_h$. Suppose that $\mu^0 \neq x_h$, i.e. $\mu^0 \neq \max(\min X^-, \min X^+)$; then $\mu^0 < x_h$ and, as $x_h = \min X^-$, it must be $\mu^0 = x_t$ for some $x_t \in X^+$. Then, there exist $(x_v, y_r) \in C^-$,

$(x_t, y_s) \in C^+$ such that $x_t \geq x_v$. From $x_v \geq x_h$, it is $x_t \geq x_v \geq x_h$, i.e. $\mu^0 \geq x_h$, which is absurd; hence $\mu^0 = \max(\min X^-, \min X^+)$.
The proof is similar if $\max(\min X^-, \min X^+) = x_k$, where $x_k = \min X^+$. □

Thus, if $\mu_0 < \mu^0$, it is $I = (\mu_0, \mu^0) = (\min(\max X^-, \max X^+),\ \max(\min X^-, \min X^+))$. We set $X^- < X^+$ (resp., $X^- > X^+$) if and only if $\max X^- < \min X^+$ (resp., $\min X^- > \max X^+$), otherwise we set $X^- \nsim X^+$. We have

Theorem 5. $I \neq \emptyset$ if and only if $X^- < X^+$, or $X^- > X^+$.

Proof. Obviously, $I \neq \emptyset$ if and only if $\mu_0 < \mu^0$. We prove that $\mu_0 \geq \mu^0$ if and only if $X^- \nsim X^+$. Such a situation happens if and only if $\mu_0 \in X^-$ and $\mu^0 \in X^-$ or $\mu_0 \in X^+$ and $\mu^0 \in X^+$. Suppose that $\mu_0 = x_h \in X^+$ and $\mu^0 = x_k \in X^+$. It is $\mu_0 = \max X^+$ and $\mu^0 = \min X^+$. From $\mu_0 = \min(\max X^-, \max X^+)$, there exists $x_s \in X^-$ such that $x_s \geq x_h$ and, from $\mu^0 = \max(\min X^-, \min X^+)]$, there exists $x_t \in X^-$ such that $x_t \leq x_k$, that is $X^- \nsim X^+$. Moreover, from $\mu_0 = \max X^+$, $\mu^0 = \min X^+$, it is $\mu_0 \geq \mu^0$ and $I = \emptyset$. If we suppose that $\mu_0 = x_h \in X^-$ and $\mu^0 = x_k \in X^-$, by a similar reasoning, we have that $X^- \nsim X^+$ and $\mu_0 > \mu^0$ so that $I = \emptyset$.
Suppose that $I \neq \emptyset$ that is $\mu_0 < \mu^0$. Thus, $\mu_0 = x_k \in X^+$ and $\mu^0 = x_h \in X^-$ or $\mu_0 = x_h \in X^-$ and $\mu^0 = x_k \in X^+$. In the first case it is $X^+ < X^-$, in the other case it is $X^+ < X^-$. Conversely, if $X^+ < X^-$, it is $\max X^+ < \max X^-$ and $\mu_0 = \max X^+$. Moreover, it is $\min X^+ < \min X^-$ and $\mu^0 = \min X^-$, with $\mu_0 < \mu^0$. If, $X^+ > X^-$ it is $\max X^+ > \max X^-$ and $\mu_0 = \max X^-$. Moreover, it is $\min X^+ > \min X^-$ and $\mu^0 = \min X^+$, with $\mu_0 < \mu^0$. □

Based on the previous result, we have the following three cases

1. $X^+ < X^- \Leftrightarrow I \neq \emptyset$ and $I = (\mu_0, \mu^0)$, with $\mu_0 = \max X^+$, $\mu^0 = \min X^-$.
2. $X^+ > X^- \Leftrightarrow I \neq \emptyset$ and $I = (\mu_0, \mu^0)$, with $\mu_0 = \max X^-$, $\mu^0 = \min X^+$.
3. $X^- \nsim X^+ \Leftrightarrow I = \emptyset$.

We have

Theorem 6. Let be given two random quantities X, Y, with $\min Y < 0 < \max Y$. If case 1, or case 2, holds, then $X^- \cap X^+ = \emptyset$ and the conditional prevision assessment $\mathbb{P}(X|Y) = \mu$ is coherent if and only if $\mu \notin I$. In the case 3, the assessment $\mathbb{P}(X|Y) = \mu$ is coherent for every real number μ.

Proof. Case 1. Suppose $\mu \leq \mu_0$. We prove that μ is coherent. It is $\mu \leq \mu_0 = \max X^+ < \min X^- = \mu^0$. Let $x_h \in X^+$, that is there exist $(x_h, y_s) \in C^+$ and $y_s > 0$. It is $x_h - \mu \geq 0$, then $g_{hs} = y_s(x_h - \mu) \geq 0$. Let $x_k \in X^-$, that is there exist $(x_k, y_t) \in C^-$ and $y_s < 0$. It is $x_k - \mu > 0$, then $g_{kt} = y_t(x_k - \mu) < 0$. It follows $\inf G|H \cdot \sup G|H \leq 0$, that is μ is coherent. By a similar reasoning, if $\mu \geq \mu^0$ it follows that μ is coherent.
Conversely, we prove that, if $\mu_0 = \max X^+ < \mu < \min X^- = \mu^0$, μ is not coherent. From $X^+ < X^-$, it is $x_k \leq \mu_0 < \mu < \mu^0 \leq x_h$ for each $x_h \in X^-$,$x_k \in X^+$. Hence, we have that for each $(x_h, y_s) \in \mathcal{C}^-$ one has $g_{hs} = y_s(x_h - \mu) < 0$, as $y_s < 0$ and $x_h - \mu \geq \mu_0 - \mu > 0$; moreover, for each $(x_k, y_t) \in \mathcal{C}^+$ one has $g_{kt} = y_t(x_k - \mu) < 0$, as $y_t > 0$ and $x_k - \mu \leq \mu^0 - \mu < 0$. Hence, for every $(x_h, y_k) \in \mathcal{C}$, it is $g_{hk} = y_k(x_h - \mu) < 0$. Then $\inf G|H \cdot \sup G|H > 0$, that is μ is not coherent.
Case 2. The proof is formally identical to the case 1.
Case 3. There exist $(x_h, y_t) \in C^-$, $(x_k, y_s) \in C^+$, $(x_u, y_r) \in C^-$, $(x_v, y_z) \in C^+$, such that $x_h < x_k$ and $x_u > x_v$. Let μ be a real number. Suppose that $g_{ht} = y_t(x_h - \mu) < 0$. Then, $(x_h - \mu) > 0$ and $(x_k - \mu) > 0$, hence $g_{ks} = y_s(x_h - \mu) > 0$ and μ is coherent.
Suppose that $g_{ht} = y_t(x_h - \mu) > 0$. Then, $(x_h - \mu) < 0$. Thus, suppose that $(x_k - \mu) > 0$. It is $x_h < \mu < h_k$. By absurd, suppose that $g_{ur} = y_r(x_u - \mu) > 0$ and $g_{vz} = y_z(x_v - \mu) > 0$. Thus, it is $x_u - \mu < 0$ and $x_v - \mu > 0$, that is $x_u < \mu < x_v$ and $x_u < x_v$, which is absurd, as $x_u > x_v$. Then, μ is coherent. □

Remark 2. We observe that Theorem 3 is a particular case of Theorem 6, as $X^- \cap X^+ \neq \emptyset$ implies $X^- \nsim X^+$.

We say that $X^+ \leq X^-$ if $\max X^+ = \min X^-$, and $X^+ \geq X^-$ if $\min X^+ = \max X^-$.
From the previous results, we can summarize the case $\min Y < 0 < \max Y > 0$ in the following way

- $X^+ < X^- \Leftrightarrow \mu_0 = \max X^+ < \min X^- = \mu^0$. Then μ is coherent if and only if $\mu \leq \mu_0$ or $\mu \geq \mu^0$.
- $X^+ > X^- \Leftrightarrow \mu_0 = \max X^- < \min X^+ = \mu^0$. Then μ is coherent if and only if $\mu \leq \mu_0$ or $\mu \geq \mu^0$.
- $X^- \nsim X^+$. If $X^+ \leq X^-$ or $X^+ \geq X^-$, then $\mu_0 = \mu^0$, otherwise $\mu_0 > \mu^0$ and in all such cases every real number μ is coherent.

We illustrate the previous result by the example below.

Example 8. We determine the set Π of coherent prevision assessment $\mathbb{P}(X|Y) = \mu$ on $X|Y$, where $(X,Y) \in \mathcal{C} = \{(0,1),(0,2),(1,-1),(1,-2)\}$. We have $X^- = \{1\}$, $X^+ = \{0\}$, so that $X^- \cap X^+ = \emptyset$; we have to consider a unique case: $x_h = 1$, $x_k = 0$, with the associated open interval $I_{hk} = (0,1)$. Then, $I = I_{hk} = (0,1)$ and, by Theorem 6, $\Pi = \Re \setminus (0,1)$; that is, μ is coherent if and only if $\mu \notin (0,1)$. The same result follows directly, by observing that: (i) $\mathcal{C}_0 = \mathcal{C}$, so that $G|H = G$;
(ii) given any μ, the values of G are: $-\mu, -2\mu, -1+\mu, -2+2\mu$; (iii) if $\mu \in (0,1)$, the values of G are all negative; if $\mu \notin (0,1)$, it is: $\min G < 0$, $\max G > 0$.

8 Linear transformations of Y.

In this section we examine the effect produced on the set Π (of coherent conditional prevision assessments on $X|Y$) by a linear transformation on the *conditioning* random quantity Y. Given two random quantities X, Y and two constants c, d, with $(c,d) \neq (0,0)$, we set $y_0 = min\, Y$, $y^0 = max\, Y$, $Y' = cY + d$ and, if $c \neq 0$, $Y^* = Y + \frac{d}{c}$,; moreover, we denote by Π' (resp., Π^*) the set of coherent prevision assessments on $X|Y' = X|(cY+d)$ (resp., $X|Y^* = X|(Y + \frac{d}{c})$). We show below, among other things, that: (a) for $d \neq 0$ both cases $\Pi^* = \Pi$, or $\Pi^* \neq \Pi$, are possible; (b) $\Pi' = \Pi^*$.

Theorem 7. Given two finite random quantities X, Y and two constants c, d, with $(c,d) \neq (0,0)$, we have:
1. if $c = 0, d \neq 0$, then $\mathbb{P}(X|Y') = \mathbb{P}(X|d) = \mathbb{P}(X)$ and $\Pi' = [min\, X, max\, X]$;
2. if $c \neq 0$, $\frac{d}{c} \notin (-y^0, -y_0)$, then $\Pi^* = [x_0, x^0]$, where the values x_0, x^0 are defined as in (8) with Y replaced by Y^*;
3. if $c \neq 0$, $\frac{d}{c} \in (-y^0, -y_0)$, then $\Pi^* = \mathbb{R} \setminus I$, where the (possibly empty) interval I is defined as in (9) with Y replaced by Y^*;
4. $\Pi' = \Pi^*$.

Proof. In case 1 it is $G = d(X - \mu)$; then, under coherence of $\mathbb{P}(X)$, from $\mathbb{P}(G) = 0$ it follows $\mu = \mathbb{P}(X) \in [min\, X, max\, X]$. In case 2, it is $Y^* \geq 0$, when $\frac{d}{c} \geq -y_0$, and $Y^* \leq 0$, when $\frac{d}{c} \leq -y^0$; then, by Theorems 1 and 2, it follows $\Pi^* = [x_0, x^0]$. In case 3, as $-y^0 \leq \frac{d}{c} \leq -y_0$, it is $min\, Y^* < 0 < max\, Y^*$; then, by Theorem 6, one has $\Pi^* = \mathbb{R} \setminus I$, with the interval I possibly empty.
In case 4 it is $Y' = cY^*$ and, denoting by G' (resp., G^*) the random gain associated with $X|Y'$ (resp., $X|Y^*$), we have $G' = cY^*(X - \mu) = cG^*$. Then

$$\inf G'|H \cdot \sup G'|H = c^2 \inf G^*|H \cdot \sup G^*|H\,,$$

and, being $c^2 \neq 0$, the assessment $\mathbb{P}(X|Y') = \mathbb{P}(X|cY^*) = \mu$ is coherent if and only if $\mathbb{P}(X|Y^*) = \mu$ is coherent; thus $\Pi' = \Pi^*$. □

We give below an example where $\Pi^* \neq \Pi$.

Example 9. As in Example 6, we consider the random vector $(X,Y) \in \mathcal{C} = \{(0,1),(1,0),(1,1),(2,2)\}$. We recall that $\Pi = [0,2]$. Given $Y' = 2Y - 2 = 2Y^*$, where $Y^* = Y - 1$, let us determine the set $\Pi' = \Pi^*$. It is

$$(X,Y^*) \in \mathcal{C}^* = \{(0,0),(1,-1),(1,0),(2,1)\}$$

$$X^{*-} = \{1\}\,,\ \ X^{*+} = \{2\}\,,\ \ X^{*-} \cap X^{*+} = \emptyset\,.$$

Then: $X^{*-} < X^{*+}$, $\mu_0 = 1$, $\mu^0 = 2$, and we have $I = (1,2)$; moreover

$$y_0 = 0\,,\ \ y^0 = 2\,,\ \ \frac{d}{c} = -1 \in (-2,0) = (-y^0, -y_0)\,.$$

Then, by Theorem 7, case 3, we obtain

$$\Pi' = \Pi^* = (-\infty, 1) \cup (2, +\infty) = \mathbb{R} \setminus (1,2) \neq \Pi\,.$$

9 Conclusions

In this paper, recalling a general discussion on iterated conditioning given by de Finetti in his book, vol. 2, Appendix, section 13, we have given a representation of a conditional random quantity $X|HK$ as $(X|H)|K$. In this way, we have obtained the classical formula $\mathbb{P}(XH|K) = \mathbb{P}(X|H)P(H|K)$, by simply using linearity of prevision. Then, we have considered the notion of general conditional prevision $\mathbb{P}(X|Y)$, where X and Y are two random quantities, introduced in 1990 in a paper by Lad and Dickey, also discussed by Lad in his book published in 1996. After recalling the case where Y is an event, we have considered the case of discrete finite random quantities and we made some critical comments and examples. We have given a notion of coherence for such more general conditional prevision assessments; then, we have obtained a strong generalized compound prevision theorem. We have studied the coherence of a general conditional prevision assessment $\mathbb{P}(X|Y)$ when Y has no negative values and when Y has no positive values. We gave some results concerning the set of coherent conditional prevision assessments of $X|Y'$, where Y' is a linear transformation of Y. Finally, we have given some results on coherence of $\mathbb{P}(X|Y)$ when Y assumes both positive and negative values. To better illustrate some critical points and remarks we have also examined several examples. Future research more in general should concern: (i) the coherence of a conditional prevision assessment $\mathcal{A} = (\mu_1, \ldots, \mu_n)$ on a family of n conditional

random quantities $\mathcal{F} = \{X_1|Y_1, \ldots, X_n|Y_n\}$; (ii) the generalized coherence of imprecise conditional prevision assessments, for instance an interval-valued assessment $\mathcal{A} = ([l_1, u_1], \ldots, [l_n, u_n])$, on $\mathcal{F}$.

Acknowledgements
We are grateful to the anonymous referees for their very useful criticisms and suggestions.

References

[1] Baioletti M., and Capotorti A., A comparison between classical logic and three-valued logic for conditional events, *Proc. of IPMU'96*, Granada, July 1-5, 1996, Vol. 3, pp. 1217-1221.

[2] Biazzo V., and Gilio A., A generalization of the fundamental theorem of de Finetti for imprecise conditional probability assessments, *International Journal of Approximate Reasoning,* 24, 251-272, 2000.

[3] Biazzo V., and Gilio A., On the linear structure of betting criterion and the checking of coherence, *Annals of Mathematics and Artificial Intelligence,* 35, 83-106, 2002.

[4] Biazzo V., and Gilio A., Some results on imprecise conditional prevision assessments, *Proc. of ISIPTA'07*, Prague, Czech Republic, July 16-19, 2007, pp. 31-40.

[5] Biazzo V., Gilio A., and Sanfilippo G., Coherence Checking and Propagation of Lower Probability Bounds, *Soft Computing,* 7, 310-320, 2003.

[6] Biazzo V., Gilio A., and Sanfilippo G., Generalized coherence and connection property of imprecise conditional previsions, *Proc. of IPMU'08*, Malaga, Spain, June 22 - 27, 2008, pp. 907-914.

[7] Capotorti A., and Paneni T., An operational view of coherent conditional previsions, *Proc. of EC-SQARU'01*, Toulouse, France, September 19-21, 2001, pp. 132-143.

[8] Capotorti A., and Vantaggi B., Locally strong coherence in inference processes, *Annals of Mathematics and Artificial Intelligence,* 35, 125-149, 2002.

[9] Coletti G., and Scozzafava R., *Probabilistic logic in a coherent setting*, Kluwer Academic Publishers, 2002.

[10] de Cooman G., and Miranda E., Marginal extension in the theory of coherent lower previsions, *International Journal of Approximate Reasoning*, 46, 188-225, 2007.

[11] de Finetti B. (1935), La logique de la probabilité, *Actes du Congrès International de Philosophie Scientifique*, Paris 1935, Hermann: IV, 1-9, 1936.

[12] de Finetti B. (1970), *Teoria delle probabilità*, voll. 1-2, Einaudi, Torino, 1970 (Engl. transl.: *Theory of Probability*, voll. 1-2, Wiley, Chichester, 1974).

[13] Gilio A., Criterio di penalizzazione e condizioni di coerenza nella valutazione soggettiva della probabilità, *Bollettino Un. Matem. Ital.*, [7a] 4-B(3): 645-660, 1990.

[14] Gilio A., Incomplete probability assessments in decision analysis, *J. It. Statist. Soc.*, Vol. 1, N. 1, 67-76, 1992.

[15] Holzer S., On coherence and conditional prevision, *Boll. Un. Mat. Ital.* 4 (6), 441-460, 1985.

[16] Lad F., and Dickey J. M., A general theory of conditional prevision, $P(X|Y)$, and the problem of state-dependent preferences, *Economic Decision-Making: Games, Econometrics and Optimization, Essays in Honor of Jaques Dreze*, J.J. Gabsewicz, J.F. Richard, and L.A. Wolsey (eds.), Amsterdam: North Holland, 369-383, 1990.

[17] Lad F., *Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction.* New York, Wiley, 1996.

[18] Regazzini E., Finitely additive conditional probabilities, *Rend. Sem. Mat. Fis. Milano* 55, 69-89, 1985.

[19] Vicig P., Zaffalon M., and Cozman F.G., Notes on "Notes on conditional previsions", *International Journal of Approximate Reasoning*, 44, 358-365, 2007.

[20] Walley P., *Statistical reasoning with imprecise probabilities*, Chapman and Hall, London, 1991.

[21] Walley P., Pelessoni R., and Vicig P., Direct Algorithms for Checking Coherence and Making Inferences from Conditional Probability Assessments, *Journal of Statistical Planning and Inference*, 126(1), 119-151, 2004.

[22] Williams P.M., Notes on conditional previsions, Technical report, University of Sussex, 1975. Reprinted in a revised form in: *International Journal of Approximate Reasoning*, 44(3):366-383, 2007.

Approximation of coherent lower probabilities by 2-monotone measures

Andrey G. Bronevich
Technological Institute of Southern Federal University, Taganrog, RUSSIA
brone@mail.ru

Thomas Augustin
Department of Statistics, Ludwig-Maximilians University (LMU), Munich, GERMANY
thomas@stat.uni-muenchen.de

Abstract

The paper investigates outer approximations of coherent lower probabilities by 2-monotone measures. We characterize the set of (Pareto)-optimal outer approximations and provide powerful iterative algorithms to calculate such measures.

Keywords. Pareto optimal 2-monotone measure, additivity on lattices, simplex method, imprecision indices.

1 Introduction

Walley [21, p. 51] is often cited in saying that he does not "…know any 'rationality' argument for two-monotonicity, beyond its computational convenience." Of course, in particular in problems of larger scale, computational convenience, and even computational tractability, is still an issue, and so the problem of finding a suitable approximation of a coherent lower probability by 2-monotone measures arises naturally in many applications of imprecise probabilities (see also Section 3).

As analysis shows, the optimal choice of a 2-monotone measure can not be made uniquely, which may be understood from the fact that the minimum of two 2-monotone measures is not again a 2-monotone measure in general, and so we will characterize and derive Pareto optimal solutions to that problem.

The main idea of this paper consists of the following. For any coherent probability μ, we define a convex set $M_{2-mon\le\mu}$ of 2-monotone measures that are dominated by μ. Then any possible optimal choice of a 2-monotone measure in $M_{2-mon\le\mu}$ is produced by finding extreme points of $M_{2-mon\le\mu}$, which are not dominated by other measures in $M_{2-mon\le\mu}$, and any optimal measure is represented as a linear convex combination of such points. After some technical preliminaries (section 2) and a slightly more detailed look at the convenience of 2-monotonicity, we give in section 4 a necessary and sufficient condition for a 2-monotone measure to be an extreme point through lattices on which a 2-monotone measure is additive. In Section 5, we provide iterative algorithms for searching optimal extreme points, which then are illustrated by two examples. In the Appendix the reader can find some results on canonical sequences of monotone measures [5], which are used in the proofs.

2. Technical preliminaries

Let X be a measurable space and $\mathfrak{A}$ be a σ-algebra of its subsets. A set function $\mu:\mathfrak{A}\to[0,1]$ is called a *monotone measure* [14] if 1) $\mu(\varnothing)=0$, $\mu(X)=1$; and 2) $A,B\in\mathfrak{A}$, $A\subseteq B$ implies $\mu(A)\le\mu(B)$. We write $\mu_1\le\mu_2$ for monotone measures μ_1,μ_2 on $\mathfrak{A}$ if $\mu_1(A)\le\mu_2(A)$ for all $A\in\mathfrak{A}$. In this paper we consider the following families of monotone measures:

1) M_{mon} is the set of all monotone measures on $\mathfrak{A}$;

2) M_{pr} is the set of all finite additive probability measures on $\mathfrak{A}$, i.e. $M_{pr}\subseteq M_{mon}$ and additionally $\mu(A\cup B)=\mu(A)+\mu(B)$ for disjoint sets $A,B\in\mathfrak{A}$;

3) M_{low} is the set of all *lower probabilities* [22] on $\mathfrak{A}$, i.e. $M_{low}\subseteq M_{mon}$ and for any $\mu\in M_{low}$ there exists $P\in M_{pr}$ such that $\mu\le P$ (, and so $\mu\in M_{low}$ iff it satisfies the avoiding sure loss property [22]);

4) M_{coh} is the set of all *coherent lower probabilities* [22] on $\mathfrak{A}$, i.e. for any $\mu\in M_{coh}$ and $B\in\mathfrak{A}$ there exists $P\in M_{pr}$ such that $\mu\le P$ and $\mu(B)=P(B)$;

5) M_{2-mon} is the set of all 2-monotone measures [11] on $\mathfrak{A}$, i.e. $M_{2-mon}\subseteq M_{mon}$ and $\mu(A)+\mu(B)\le\mu(A\cup B)+\mu(A\cap B)$ for any $A,B\in\mathfrak{A}$.

6) M_{chain} is the set of all chain measures [14] on $\mathfrak{A}$, i.e. if $\mu\in M_{chain}$, then there is a chain $\Gamma\subseteq\mathfrak{A}$ such that $\varnothing\in\Gamma$, $X\in\Gamma$ and, for all B, $\mu(B)=\sup_{A\in\Gamma|A\subseteq B}\mu(A)$.

3. On the convenience of 2-monotonicity

As also discussed below, 2-monotone measures have some regular properties compared to coherent lower probabilities, which are very convenient from the computational point of view. Of particular importance is the property recalled in Remark 1 below, ensuring that for any chain of events there is a single classical probability

in the core simultaneously attaining the lower probability for all elements of the chain. As a consequence, the enveloping lower and upper distribution functions define probabilities in the core, and so, for instance, a closed form for natural extension (calculating expectation of random variables) is available (repeated, e.g, in. [22, p. 30ff], where also some direct applications are given). By similar arguments a convenient closed form for calculating lower and upper conditional probabilities (in Walley's sense) can be derived (see, e.g., [22, p. 301, including the corresponding footnote]). Moreover, other common forms of conditioning, like Dempster's rule of conditioning ([13]), also called maximum likelihood updating ([15]), are then guaranteed to lead to a coherent, and indeed again 2-monotone, solution.

Our main motivation for the present study, however, has been the case of hypothesis testing, where one has to distinguish between two hypotheses described by imprecise probabilities, and decide which one is more likely to have produced the data. Similarly as in the case of calculating the conditional distribution or the natural extension, the testing problem can be expressed in terms of a single linear optimization problem (see [1, chapter 4]), but, even with the considerable improvement along the lines developed for decision problems in [20, section 3.2], the problem still increases exponentially in the sample size, and so still is, for the sample sizes usually common in statistics, simply computationally intractable.

A powerful way out is offered by Huber-Strassen theory ([18], and the work following it, see also [17, 3, 4] for reviews from different perspectives). The famous Huber-Strassen theorem (in [18, cf. also the finally obtained extension in [9]) ensures that 2-monotonicity is sufficient for the existence of a globally least favorable pair, i.e. a pair of classical probability distributions that

i) allow to represent the whole testing problem in determining the optimal test and

ii) can be calculated by considering sample size 1 only.

While i) can be alleviated by a concept of local least favorability ([1, chapter 3), [2], [16]), property ii) can not be generalized appropriately (see the analysis of the proof in [1, p. 223ff.]). As a consequence, statistical models described by coherent, but not 2-monotone measures, often have to be approximated appropriately to be able to determine appropriate statistical testing procedures.

4 Approximation by 2-monotone measures (finite case)

In this case, we assume that X is a finite set and $\mathfrak{A}$ is the power set of X, i.e. $\mathfrak{A}=2^X$. Let $\mu\in M_{low}$, then $\nu\in M_{mon}$ is defined as a Pareto optimal approximation of μ if $\nu\le\mu$ and $\nu\le\nu'\le\mu$ for $\nu'\in M_{mon}$ implies that $\nu'=\nu$. For any $\mu\in M_{low}$, we denote $M_{2-mon\le\mu}=\{\nu\in M_{2-mon}\mid\nu\le\mu\}$.

Lemma 1. *Any Pareto optimal 2-monotone measure for a $\mu\in M_{coh}$ can be represented as a convex linear combination of Pareto optimal extreme points of $M_{2-mon\le\mu}$.*

Proof. It is clear that the set $M_{2-mon\le\mu}$ has a finite set of extreme points $\{\mu_i\}$, because it can be described by a finite number of inequalities. Therefore any $\nu\in M_{2-mon\le\mu}$ can be represented as a linear convex combination of these points, i.e. $\nu=\sum_i a_i\mu_i$, where $a_i\ge 0$, $\sum_i a_i=1$. Assume that in the above representation there is an extreme point $\mu_{i'}$ such that $a_{i'}>0$ and $\mu_{i'}$ is not Pareto optimal, i.e. there is $\mu'\in M_{2-mon\le\mu}$ such that $\mu_{i'}<\mu'$ (i.e., $\mu_{i'}\le\mu'$ and $\mu_{i'}\ne\mu'$). Then we define $\nu'=\sum_{i\|\ne i'} a_i\mu_i\ +a_{i'}\mu'$. It is clear that $\nu'\in M_{2-mon\le\mu}$ and $\nu<\nu'$, therefore, ν is not Pareto optimal, which means that the coefficient a_i has to be equal to zero if the corresponding extreme measure μ_i is not Pareto optimal. This fact proves the lemma.■

The previous lemma says that the full description of Pareto optimal 2-monotone measures for $\mu\in M_{coh}$ can be given by knowing only its Pareto optimal extreme 2-monotone measures. Therefore, we have to answer the following question: what characteristics define extreme points uniquely? For this reason, we further involve some results concerning additivity properties of 2-monotone measures. We will consider lattices of the algebra $\mathfrak{A}$. A lattice is a subset of $\mathfrak{A}$ closed with respect to intersection and union. We say that $\mu\in M_{2-mon}$ is additive on a lattice $\mathfrak{L}\subseteq\mathfrak{A}$ if $\mu(A)+\mu(B)=\mu(A\cup B)+\mu(A\cap B)$ for any $A,B\in\mathfrak{L}$. Next straightforward result shows the way how we can describe additivity of 2-monotone measures.

Lemma 2. *Let $\mathfrak{S}$ be the set of all possible lattices in $\mathfrak{A}$, on which $\mu\in M_{2-mon}$ is additive. Then $\mathfrak{S}$ is a covering[1] of $\mathfrak{A}$.*

Proof. Let $X=\{x_1,...,x_n\}$. Consider maximal chains in $\mathfrak{A}=2^X$ of the type $\Gamma=\{B_0,B_1,...,B_n\}$, $\varnothing=B_0\subset B_1\subset...\subset B_n=X$, $|B_i\setminus B_{i-1}|=1$, $i=1,...,n$. It is clear that such chains are lattices and every monotone measure

[1] An arbitrary covering $\mathfrak{C}$ of $\mathfrak{A}$ is a family of non-empty subsets of $\mathfrak{A}$ such that $\bigcup_{\mathfrak{a}\in\mathfrak{C}}\mathfrak{a}=\mathfrak{A}$.

is additive on them, i.e., we get the required covering that consists of all these lattices.■

We denote by $\mathfrak{S}_\mu$ the covering of $\mathfrak{A}$ that consists of all maximal lattices, on which $\mu \in M_{2-mon}$ is additive. For example, if μ is a probability measure, then the covering is a singleton, which contains only one element $\mathfrak{A}$. If a $\mu \in M_{2-mon}$ is such that $\mu(A)+\mu(B) < \mu(A \cup B) + \mu(A \cap B)$ for any $A, B \in \mathfrak{A}$ with $A \not\subseteq B$ and $B \not\subseteq A$ then $\mathfrak{S}_\mu$ obviously consists of all maximal chains in $\mathfrak{A}$. It is important to emphasize that any $\Lambda \in \mathfrak{S}_\mu$ has to contain $\varnothing$ and X, since these sets are additive elements for any $\mu \in M_{2-mon}$.

Another convenient characterization of 2-monotone measures is recalled in

Remark 1. For any $\mu \in M_{2-mon}$, define the convex set $core(\mu)$ of probability measures, defined by $core(\mu) = \{P \in M_{pr} \mid P \geq \mu\}$. It is well-known that this set is non-empty and usually called the core of μ. Moreover, it is possible to describe all extreme points of this set [10]. To do this, we should consider all maximal chains of the algebra 2^X on $X = \{x_1, x_2, ..., x_n\}$. Then any extreme point P_γ is generated by a maximal chain $\gamma = \{B_0, B_1, ..., B_n\}$, where $\varnothing = B_0 \subset B_1 \subset ... \subset B_n = X$ and $B_k = \{x_{i_1}, ..., x_{i_k}\}$, $k = 1, ..., n$, as $P_\gamma(\{x_{i_k}\}) = P_\gamma(B_k \setminus B_{k-1}) = \mu(B_k) - \mu(B_{k-1})$, i.e. P_γ is chosen such that $P_\gamma(B) = \mu(B)$ for all $B \in \gamma$.

Lemma 3. *Any lattice in $\mathfrak{S}_\mu$ contains a maximal chain.*

Proof. Consider an arbitrary lattice $\Lambda \subseteq 2^X$, on which μ is additive. Let Γ be a sequence of sets with the following properties: 1) a minimal algebra that contains Γ coincides with 2^X; 2) first elements of Γ are all elements of Λ. Then the limit measure[2] μ_Γ, in the canonical sequence constructed by Γ is a probability measure, and also $\mu_\Gamma(A) = \mu(A)$ for all $A \in \Lambda$. Since any such sequence Γ is equivalent to some maximal chain $\gamma \subseteq 2^X$, we get $\mu_\Gamma(A) = \mu(A)$ for all $A \in \gamma$. Consider a lattice, on which μ and μ_Γ have the same values. It is clear that this lattice contains Λ and γ, and also μ is additive on it. It means that any lattice in $\mathfrak{S}_\mu$ contains a maximal chain. ■

[2] The explanation of terms: "limit measure", "canonical sequence of monotone measures", … are given in Appendix.

Proposition 1. *There is the one-to-one correspondence between maximal lattices in $\mathfrak{S}_\mu$ and extreme points of $core(\mu)$ for every $P \in core(\mu)$ defined by $\Lambda = \{A \in \mathfrak{A} \mid P(A) = \mu(A)\}$, where $\Lambda \in \mathfrak{S}_\mu$.*

Proof. Because any lattice $\Lambda \in \mathfrak{S}_\mu$ contains a maximal chain $\gamma \subseteq \Lambda$, we can define that P_γ corresponds to Λ. Using canonical sequences of 2-monotone measures, it easy to prove that $P_\gamma(B) = \mu(B)$ for all $B \in \Lambda$. This proves that if Λ contains two different maximal chains, then they generate the same probability measure, i.e. we show that such a construction generates the unique probability measure P_γ, where $\gamma \subseteq \Lambda$, with $P_\gamma(B) = \mu(B)$ for all $B \in \Lambda$. We finish the proof of the proposition by showing that for any maximal chain γ the set $\{B \in 2^X \mid P_\gamma(B) = \mu(B)\} \in \mathfrak{S}_\mu$. It is easy to check that this set is a lattice. Let $P_\gamma(A) = \mu(A)$ and $P_\gamma(B) = \mu(B)$ for some $A, B \in 2^X$. Then we have to prove that also $P_\gamma(A \cap B) = \mu(A \cap B)$ and $P_\gamma(A \cup B) = \mu(A \cup B)$. The above condition implies that

$$\mu(A) + \mu(B) \leq \mu(A \cap B) + \mu(A \cup B) \leq$$
$$P_\gamma(A \cap B) + P_\gamma(A \cup B) = P_\gamma(A) + P_\gamma(B) = \mu(A) + \mu(B),$$

i.e. $\mu(A) + \mu(B) = \mu(A \cap B) + \mu(A \cup B)$, $P_\gamma(A \cap B) = \mu(A \cap B)$ and $P_\gamma(A \cup B) = \mu(A \cup B)$. Using again canonical sequences of 2-monotone measures, it is easy to prove that such a lattice is maximal, i.e. we have the one-to-one correspondence between maximal lattices in $\mathfrak{S}_\mu$ and extreme points in $core(\mu)$.■

Proposition 2. *Let $\mu \in M_{coh}$, $\nu \in M_{2-mon \leq \mu}$, $S_{\nu=\mu} = \{A \in \mathfrak{A} \mid \nu(A) = \mu(A)\}$, $S_{\nu=0} = \{A \in \mathfrak{A} \mid \nu(A) = 0\}$. Then ν is an extreme point of $M_{2-mon \leq \mu}$ iff its values are defined by the sets $S_{\nu=\mu}$, $S_{\nu=0}$, $\mathfrak{S}_\nu$ uniquely.*

Proof. A set function ν is in $M_{2-mon \leq \mu}$ iff it satisfies the following conditions:

1) $\nu(\varnothing) = 0$, $\nu(X) = 1$;
2) $\nu(A) \geq 0$ for all $A \in A$;
3) $\nu(A) \leq \nu(B)$ if $A \subseteq B$;
4) $\nu(A) + \nu(B) \leq \nu(A \cap B) + \nu(A \cup B)$ for all $A, B \in A$;
5) $\nu(A) \leq \mu(A)$ for all $A \in 2^X$.

These conditions can be considered as a system of linear inequalities on values $\nu(A)$, $A \in 2^X$. From the theory of linear inequalities, we know that any extreme point can be calculated by solving linear equalities, obtained by the subset of inequalities if we change "$\leq$" to "$=$". Show that we can confine ourselves to using equalities

that are generated by 2), 4), and 5). It is not necessary to use 1) because $\nu(\varnothing)=\mu(\varnothing)=0$, $\nu(X)=\mu(X)=1$. We show further that any equality $\nu(C)=\nu(D)$ for $C\subset D$ ($C\neq D$), generated by 3), can be derived from the additivity of ν. In this case, we take $A=C$, $B=D\setminus C$. Then $A\cap B=\varnothing$, $\nu(B)=0$, $\nu(A)+\nu(B)=\nu(A\cap B)+\nu(A\cup B)$, and the last equality, $\nu(B)=0$, $\nu(A\cap B)=0$ implies that $\nu(C)=\nu(D)$. Therefore, we conclude that the proposition is true. ■

Consider some corollaries from Propositions 1 and 2:

Corollary 1. *Let the notation of Proposition 2 be used. Then ν is an extreme point of $M_{2-mon\le\mu}$ if for any $\Lambda\in\mathfrak{S}_\nu$ a probability measure P_Λ with $P_\Lambda(A)=\mu(A)$ for all $A\in S_{\nu=\mu}\cap\Lambda$ and $P_\Lambda(A)=0$ for all $A\in S_{\nu=0}\cap\Lambda$ is defined uniquely.*

Proof. It is easy to see that Corollary 1 is a direct consequence of Propositions 1 and 2. ■

Corollary 2. *Let notations of Proposition 2 be used. Then ν is an extreme point of $M_{2-mon\le\mu}$ if for any $\Lambda\in\mathfrak{S}_\nu$ the set $\Lambda\cap\left(S_{\nu=\mu}\cup S_{\nu=0}\right)$ contains a maximal chain. In addition, $\nu=\underset{\gamma}{\wedge}P_\gamma$, where the minimum in the right side of the last formula is taken over all possible probability measures P_γ, defined for each maximal chain $\gamma\subseteq S_{\nu=\mu}\cup S_{\nu=0}$ by $P_\gamma(A)=\mu(A)$ for $A\in\gamma\cap S_{\nu=\mu}$ and $P_\gamma(A)=0$ for $A\in\gamma\cap S_{\nu=0}$. Moreover, if $S_{\nu=0}\setminus S_{\nu=\mu}=\varnothing$, then such a ν is Pareto optimal.*

Proof. Because P_Λ is defined uniquely if $\Lambda\cap\left(S_{\nu=\mu}\cup S_{\nu=0}\right)$ contains a maximal chain, we conclude that ν is an extreme point by Corollary 1. The formula $\nu=\underset{\gamma}{\wedge}P_\gamma$ is also true, since 2-monotonicity of ν implies that $P_\gamma\ge\nu$ for any $\gamma\subseteq S_{\nu=\mu}\cup S_{\nu=0}$. Observe also that, for any $\nu'\in M_{2-mon\le\mu}$ with $S_{\nu'=\mu}=S_{\nu=\mu}$ and $S_{\nu'=0}=S_{\nu=0}$, we have $\nu'\le\nu$, i.e. ν have the largest values for the fixed $S_{\nu=\mu}$ and $S_{\nu=0}$. Show that ν is Pareto optimal if $S_{\nu=0}\setminus S_{\nu=\mu}=\varnothing$. Suppose on the contrary that there is another $\nu'\in M_{2-mon\le\mu}$ such that $\nu'>\nu$. Then we should conclude that $S_{\nu=\mu}\subseteq S_{\nu'=\mu}$ and $S_{\nu=\mu}\neq S_{\nu'=\mu}$. We see that $\nu'\le\underset{\gamma\subseteq S_{\nu'=\mu}}{\wedge}P_\gamma\le\underset{\gamma\subseteq S_{\nu=\mu}}{\wedge}P=\nu$, and such a ν' does not exist, i.e. the corollary is proved in the whole. ■

Pareto optimal extreme points, described in Corollary 2, have desirable properties. They are uniquely defined by a chosen set $S_{\nu=\mu}$ and their values can be easily computed using explicit formulas. Therefore, it is desirable to study the conditions of existence of these extreme points, and to construct the algorithm for finding such sets $S_{\nu=\mu}$.

We see from Proposition 1 that any extreme point of $M_{2-mon\le\mu}$ is characterized by $S_{\nu=\mu}$, $S_{\nu=0}$, $\mathfrak{S}_\nu$. But we know that an arbitrary extreme point is not necessarily Pareto optimal. To investigate this situation, introduce so called elementary lattices in 2^X of two types. An elementary lattice Λ of the first type is given by $\Lambda=\left\{A,A\cup\{x_i\}\right\}$, where $A\in 2^X$ and $x_i\notin A$, and an elementary lattice of the second type by $\Lambda=\left\{A,A\cup\{x_i\},A\cup\{x_j\},A\cup\{x_i\}\cup\{x_j\}\right\}$, where $A\in 2^X$ and $x_i,x_j\notin A$. Using the above definition we can formulate the following necessary and sufficient feature of 2-monotonicity [7, 12].

Proposition 3. *A set function $\mu:2^X\to[0,1]$ is a 2-monotone measure iff*

1) $\mu(\varnothing)=0$, $\mu(X)=1$;

2) μ is monotone on all possible lattices in 2^X of the first type;

3) μ is 2-monotone on all possible lattices in 2^X of the second type.

Remark 2. Proposition 3 can be reformulated in the following simple way:

A set function $\mu:2^X\to[0,1]$ is a 2-monotone measure iff

1) $\mu(\varnothing)=0$, $\mu(X)=1$;

2) $\mu(A)\le\mu\left(A\cup\{x_i\}\right)$ for all possible $A\in 2^X$ and $x_i\notin A$;

3) $\mu\left(A\cup\{x_i\}\right)+\mu\left(A\cup\{x_j\}\right)\le\mu\left(A\right)+\mu\left(A\cup\{x_i\}\cup\{x_j\}\right)$ for all possible $A\in 2^X$ and $x_i,x_j\notin A$.

However, the consideration of elementary lattices is useful for characterizing Pareto optimal 2-monotone measures.

Proposition 4. *Let $\nu\in M_{2-mon\le\mu}$, $\mathcal{L}_1$ be the set of all elementary lattices of the first type on which ν is constant, and $\mathcal{L}_2$ be the set of all elementary lattices of the second type, on which ν is additive. Then ν is not Pareto optimal iff there is a non-identical zero, non-negative set function $\Delta\nu:2^X\to\mathbb{R}_+$ such that*

1) $\Delta\nu(A)=0$ if $A\in S_{\nu=\mu}$;

2) $\Delta\nu$ is monotone on all lattices in $\mathcal{L}_1$;

3) $\Delta\nu$ is 2-monotone on all lattices in $\mathcal{L}_2$.

Proof. *Necessity.* Let ν be not Pareto optimal. Then there is a $\nu' \in M_{2-mon \le \mu}$ such that $\nu' > \nu$. It is easy to check that $\Delta\nu = \nu' - \nu$ obeys all required properties.

Sufficiency. Let such a set function $\Delta\nu$ exist. Consider the following positive numbers:

$$\varepsilon_1 = \max\left\{h\left(\mu(A) - \nu(A)\right) \middle| A \in 2^X\right\},$$

$$\varepsilon_2 = \max\left\{h\left(\nu\left(A \cup \{x_i\}\right) - \nu(A)\right) \middle| A \in 2^X, x_i \notin A\right\},$$

$$\varepsilon_3 = \max\left\{w\left(A, x_i, x_j\right) \middle| A \in 2^X, x_i, x_j \notin A\right\},$$

where $h(t) = t$ if $t > 0$ and $t = 1$ else; $w\left(A, x_i, x_j\right) = h\left(\nu(A) + \nu\left(A \cup \{x_i\} \cup \{x_j\}\right) - \nu\left(A \cup \{x_i\}\right) - \nu\left(A \cup \{x_j\}\right)\right)$. Then choosing $\Delta\nu$ such that $\max\left\{\Delta\nu(A) \mid A \in 2^X\right\} \le \varepsilon$, where $\varepsilon = \min\left\{\varepsilon_1, \varepsilon_2, 0.5\varepsilon_3,\right\}$, we get that the set function $\nu' = \nu + \Delta\nu$ is in $M_{2-mon \le \mu}$ and obviously $\nu' > \nu$, i.e. ν is not Pareto optimal. ■

5. Algorithms for searching Pareto optimal 2-monotone measures

In this section we present two algorithms. The first one improves a given approximation (two-monotone probability) to a Pareto-optimal approximation, the second one places the choice of a certain Pareto-optimal approximation on a certain linear imprecision index as an objective function.

Algorithm I

Input data: coherent lower probability μ on 2^X.

First step. Finding a 2-monotone measure ν_0 with $\nu_0 \le \mu$.

Second step. Finding a Pareto optimal 2-monotone measure ν with $\nu_0 \le \nu \le \mu$.

The first step can be based on different approaches. For example, we can choose as ν_0 an arbitrary chain measure, generated by some maximal chain Γ of algebra 2^X. Then $\nu_0(B) = \sup_{A \in \Gamma \mid A \subseteq B} \mu(A)$ for all $B \in 2^X$. However, it is clear that the realization of the second step of the algorithm can be produced more effectively if the values ν_0 are close to the values of μ. In this sense, the following procedure is better than the first one.

1) Compute an auxiliary 2-monotone set function g on 2^X using the following formulas:

a) $g(A) = \mu(A)$ for all $A \in 2^X$ with $|A| \le 1$;

b) Let us compute all values of g on sets with cardinality less or equal to k. Then values of g on sets A with cardinality that is equal to $k+1$ are computed by

$$g(A) = \max\left\{\mu(A), \max_{x_i, x_j \in A} g\left(A \setminus \{x_i\}\right) + g\left(A \setminus \{x_j\}\right) - g\left(A \setminus \{x_i, x_j\}\right)\right\}.$$

Observe that in the last formula $g\left(A \setminus \{x_i\}\right) + g\left(A \setminus \{x_j\}\right) - g\left(A \setminus \{x_i, x_j\}\right) = g\left(A \setminus \{x_i\}\right)$ for $i = j$. Therefore, g is 2-monotone by Proposition 3. It is easy to see that $g \ge \mu$ and $g = \mu$ iff μ is 2-monotone, and also it is not necessarily $g(X) = 1$.

2) A 2-monotone measure $\nu_0 = \varphi \circ g$ is computed using a convex distortion function $\varphi : [0, g(X)] \to [0,1]$ that has to obey the following properties:

(i) $\varphi(0) = 0$, $\varphi(g(X)) = 1$;

(ii) $\varphi(g(A)) \le \mu(A)$ for all $A \in 2^X$.

According to, e.g., [14] ν_0 has to be also 2-monotone, i.e. $\nu_0 \in M_{2-mon} \le \mu$. The search of the mapping φ is also connected with solving the system of linear inequalities. It is clear that it is sufficient to know the values of φ only in the points in the set $Y = \left\{g(A) \mid A \in 2^X\right\}$. Let $Y = \{y_i\}_{i=0}^m$, where $0 = y_0 < y_1 < \ldots < y_m = g(X)$. Then the condition (ii) is transformed to $\varphi\left(y_i\right) \le \psi\left(y_i\right)$, where $\psi\left(y_i\right)$, $i = 1, \ldots, m-1$, are corresponding values of μ in (ii), and convexity of φ means that $\varphi\left(y_i\right) \le \varphi\left(y_{i+1}\right)$, $i = 0, \ldots, m-1$, and

$$\frac{\varphi\left(y_{i+1}\right) - \varphi\left(y_i\right)}{y_{i+1} - y_i} \le \frac{\varphi\left(y_i\right) - \varphi\left(y_{i-1}\right)}{y_i - y_{i-1}}, \; i = 1, \ldots, m.$$

Clearly the problem of searching φ is simpler than the initial problem, and we should try to choose φ with the largest values.

The second step can be performed iteratively by using procedures that are similar to the usual simplex method. Consider an algorithm that seems to be easily realizable and computationally effective. Let $\nu_k \in M_{2-mon \le \mu}$, and the following values

$$\Delta_1 = \mu(A) - \nu_k(A),$$

$$\Delta_2 = \min_{x_i \in X \setminus A} \left(\nu_k\left(A \cup \{x_i\}\right) - \nu_k(A)\right),$$

$$\Delta_3 = \min_{x_i \in X \setminus A, x_j \in A} \left(\nu_k\left(A \cup \{x_i\}\right) - \nu_k(A) - \nu_k\left(\left(A \setminus \{x_j\}\right) \cup \{x_i\}\right) + \nu_k\left(A \setminus \{x_j\}\right)\right)$$

are positive for a given $A \in 2^X$. Then, by Proposition 3, we can increase values of ν_k on the set A without any changes on other sets, and get a measure $\nu_{k+1} \in M_{2-mon \le \mu}$

by the rule $\nu_{k+1}(B)=\nu_k(B)+d$ if $B=A$ and $\nu_{k+1}(B)=\nu_k(B)$ otherwise, where $d=\min\{\Delta_1,\Delta_2,\Delta_3\}$. Thus, we can increase values by this rule until $d=0$ for any $A\in 2^X$. It easy to show that this procedure converges to a Pareto optimal 2-monotone measure after a finite number of iterations due to simplex method. Show that a measure ν_k is Pareto optimal if $d=0$ for any $A\in 2^X$. In this case, we have to show that a convex set $M=\{\nu\in M_{2-mon}\,|\,\nu_k\le\nu\le\mu\}$ is a singleton, i.e. $M=\{\nu_k\}$. Observe that values of ν_k can be considered as basic variables and the above condition ($d=0$ for any $A\in 2^X$) means that we cannot change them, i.e. the convex set M contains the only one extreme point ν_k, i.e. ν_k is Pareto optimal. Analogously, any iteration of the proposed procedure can be considered as an iterative step of the simplex method. This means that this procedure converges by a finite number of iterations.

Algorithm II. It is based on the usual application of simplex method. As a criterion a linear imprecision index can be used. By definition [8], a linear imprecision index f is a non-negative functional on M_{low} that satisfies the following properties:

1) $f(P)=0$ for any $P\in M_{pr}$;

2) $f\left(\eta_{\langle X\rangle}\right)=1$, where $\eta_{\langle X\rangle}$ describes the situation of complete ignorance, i.e. $\eta_{\langle X\rangle}(A)=1$ if $A=X$, $\eta_{\langle X\rangle}(A)=0$ otherwise;

3) $f(\nu_1)\le f(\nu_2)$ for any $\nu_1,\nu_2\in M_{low}$ such that $\nu_1\ge\nu_2$;

4) $f(a\nu_1+(1-a)\nu_2)=af(\nu_1)+(1-a)f(\nu_2)$ for arbitrary $a\in[0,1]$ and $\nu_1,\nu_2\in M_{low}$.

The notable examples of such imprecision indices are the generalized Hartley measure [19] defined by

$$GH(\nu)=\frac{1}{\ln|X|}\sum_{A\in 2^X}m(A)\ln|A|,$$

where m is the Möbius transform [10] of the given $\nu\in M_{low}$, and an index f_{L_1} based on L_1 distance defined by

$$f_{L_1}(\nu)=\frac{1}{2^{|X|}-2}\sum_{A\in 2^X}\left|\overline{\nu}(A)-\nu(A)\right|,$$

where $\overline{\nu}$ is the dual of ν, i.e. $\overline{\nu}(A)=1-\nu(A^c)$. Notice that linear imprecision indices are linear functions w.r.t. values of a given $\nu\in M_{low}$. In particular, since $\overline{\nu}\ge\nu$ for any $\nu\in M_{low}$, we get

$$f_{L_1}(\nu)=\frac{1}{2^{|X|}-2}\sum_{A\in 2^X}\left(1-\nu(A)-\nu(A^c)\right)=$$
$$1-\frac{1}{2^{|X|-1}-1}\sum_{A\in 2^X\setminus\{\varnothing,X\}}\nu(A).$$

Notice that we can use also as a linear functional the L_1 distance between μ and its approximation ν, i.e. in this case

$$f(\nu)=\sum_{A\in 2^X}\left|\mu(A)-\nu(A)\right|.$$

Because $\nu\le\mu$, we obtain

$$f(\nu)=\sum_{A\in 2^X\setminus\{\varnothing,X\}}\mu(A)-\sum_{A\in 2^X\setminus\{\varnothing,X\}}\nu(A),$$

i.e. the criterion based on this metric is equivalent to the criterion f_{L_1}.

Therefore, the choice of Pareto optimal 2-monotone measure, based on a linear inclusion index, can be conceived as a linear programming problem, where we have a system of inequalities that describe a convex set $M_{2-mon\le\mu}$ and a linear criterion f.

6. Examples of the proposed algorithms working

To illustrate our method, we use examples of coherent lower probabilities from [6].

Example 1. Let $X=\{x_1,x_2,x_3,x_4\}$ and let $\mu\in M_{coh}$ be defined on 2^X by $\mu(A)=\min\{P_1(A),P_2(A)\}$, where $A\in 2^X$ and $P_1,P_2\in M_{pr}$ are defined through their values on singletons by $P_1(\{x_1\})=1/4$; $P_1(\{x_2\})=0$, $P_1(\{x_3\})=3/4$; $P_1(\{x_4\})=0$; $P_2(\{x_1\})=0$; $P_2(\{x_1\})=1/2$, $P_2(\{x_3\})=0$; $P_2(\{x_4\})=1/2$. The values of μ are given in Table 1. It is clear that $\mu\notin M_{2-mon}$, because, for example, $\mu(A)+\mu(B)>\mu(A\cup B)$ for $A=\{x_1,x_2\}$, $B=\{x_2,x_3\}$. Following the first step of Algorithm 1, we get an auxiliary 2-monotone set function g on 2^X with values also shown in Table 1. Then we need to find a convex distortion function φ, that is lower than function ψ (see Fig. 1). The found distortion function is also shown in Fig. 1 and can be given by the formula

$$\varphi(x)=\begin{cases}0.5x, & x\in[0,0.5],\\ 0.75x-0.125, & x\in(0.5,1.5].\end{cases}$$

It easy to check that ν_0 is not Pareto optimal in this case, because, for example, $d=1/8$ for the set $A=\{x_1,x_2,x_3\}$ and according to Algorithm 1, we obtain the next approximation $\nu_1\in M_{2-mon\le\mu}$ by the rule

$\nu_1(B) = \nu_0(B) + d$ if $B = A$ and $\nu_1(B) = \nu_0(B)$ otherwise. Producing in such a way iterations for sets $\{x_1, x_3, x_4\}$, $\{x_2, x_3, x_4\}$, $\{x_2, x_3\}$, $\{x_3, x_4\}$, we obtain a Pareto optimal measure $\nu \in M_{2-mon \le \mu}$ with values given in Table 1.

x_1	x_2	x_3	x_4	μ	g	ν_0	ν
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0
1	1	0	0	1/4	1/4	1/8	1/8
0	0	1	0	0	0	0	0
1	0	1	0	0	0	0	0
0	1	1	0	1/2	1/2	1/4	3/8
1	1	1	0	1/2	3/4	7/16	½
0	0	0	1	0	0	0	0
1	0	0	1	1/4	1/4	1/8	1/8
0	1	0	1	0	0	0	0
1	1	0	1	1/4	1/2	1/4	¼
0	0	1	1	1/2	1/2	1/4	3/8
1	0	1	1	1/2	3/4	7/16	½
0	1	1	1	3/4	1	5/8	¾
1	1	1	1	1	3/2	1	1

Table 1. Results for Example 1.

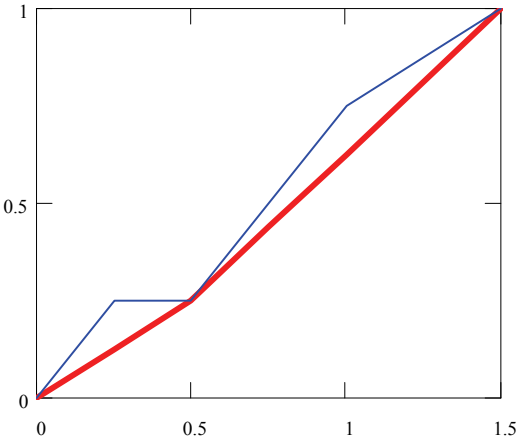


Figure 1: The distortion function for Example 1: φ - red line; ψ - blue line.

Example 2. Let $X = \{x_1, x_2, x_3, x_4\}$ and let $\mu \in M_{coh}$ have the values given in Table 2. We see that $\mu \notin M_{2-mon}$, since $\mu(A) + \mu(B) > \mu(A \cap B) + \mu(A \cup B)$ for $A = \{x_1, x_4\}$, $B = \{x_2, x_4\}$. Then, following the steps of Algorithm 1, we can get results that are shown in Table 2 and Fig. 2. The distortion function for this case can be defined by the formula

$$\varphi(x) = \begin{cases} 0.5x, & x \in [0, 2/3], \\ x - 1/3, & x \in (2/3, 4/3]. \end{cases}$$

x_1	x_2	x_3	x_4	μ	g	ν_0	ν_1
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0
1	0	1	0	0	0	0	0
0	1	1	0	0	0	0	0
1	1	1	0	2/3	2/3	1/3	2/3
0	0	0	1	0	0	0	0
1	0	0	1	1/3	1/3	1/6	1/6
0	1	0	1	1/3	1/3	1/6	1/6
1	1	0	1	1/3	2/3	1/3	1/3
0	0	1	1	1/3	1/3	1/6	1/6
1	0	1	1	1/3	2/3	1/3	1/3
0	1	1	1	1/3	2/3	1/3	1/3
1	1	1	1	1	4/3	1	1

Table 2. Results for Example 2.

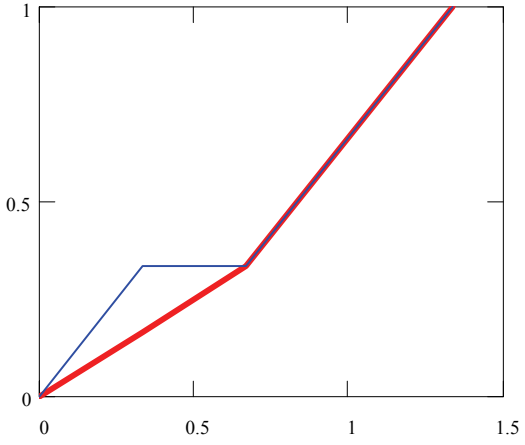


Figure 2: The distortion function for Example 2: φ - red line; ψ - blue line.

It is easy to check that ν_0 is not Pareto optimal in this case, because $d = 1/3$ for set $A = \{x_1, x_2, x_3\}$, and according to Algorithm 1, we obtain a Pareto optimal measure $\nu_1 \in M_{2-mon \le \mu}$ by the rule $\nu_1(B) = \nu_0(B) + d$ if $B = A$ and $\nu_1(B) = \nu_0(B)$ otherwise.

Notice that we can indeed apply the proposed algorithms to any monotone measure, i.e. μ need not be a coherent

lower probability. This case is considered in the next example.

Example 3. Let $X=\{x_1,x_2,x_3,x_4\}$ and let $\mu\in M_{mon}$ be defined by $\mu(A)=1$ if $|A|\geq 1$ and $\mu(A)=0$. In this case the set of all Pareto optimal 2-monotone measures coincides with the set of all probability measures on 2^X, and, by Algorithm 1, we obtain a probability measure $\nu=\nu_0$ defined by $\nu(\{x_i\})=1/4$, where $i=1,...,4$.

7. Concluding remarks

We have characterized and computed Pareto optimal outer approximations of coherent lower probabilities by 2-monotone measures. Further research includes obviously the study of the sensitivity of the results with respect to the choice of the approximation.

Also a closer investigation of some modifications of the algorithms is certainly rewarding, in particular in the following directions.

Because in principle the solution of the optimization problem is computationally very hard for large $n=|X|$, it is possible to solve it for some subalgebra $\mathfrak{B}\subseteq 2^X$. Let ν be Pareto optimal on $\mathfrak{B}$ for some μ on 2^X, then its inner extension $\underline{\nu}$ on 2^X defined by $\underline{\nu}(B)=\sup_{A\in\mathfrak{B}|A\subseteq B}\nu(A)$, $B\in 2^X$, is 2-monotone [14], and can be considered as an approximation of a Pareto optimal measure. The same approach can be used for a general infinite algebra $\mathfrak{A}$.

In light of the intended application to statistical hypotheses testing, it will also be interesting to replace the linear imprecision index in the objective function by the Kullback-Leibler distance, which has some close relation to the likelihood ratio underlying optimal hypotheses testing.

Notice that a Pareto optimal measure is not uniquely defined even in a case when we use a linear imprecision index in the linear programming problem. To get uniqueness, it seems to be possible to use the following approach: Let $\mathfrak{A}=2^X$, where $|X|=n$, we have a linear order on $\mathfrak{A}$ defined by indexing its elements, i.e. $\mathfrak{A}=\{B_i\}_{i=1}^{2^n}$ and B_i is more preferable than B_j if $i<j$. Then we say that $\nu_1\in M_{2-mon\leq\mu}$ is more preferable than $\nu_2\in M_{2-mon\leq\mu}$ if there is an index k such that $\nu_1(B_i)=\nu_2(B_i)$ for $i=1,...,k-1$, and $\nu_1(B_k)>\nu_2(B_k)$.

Another rewarding issue has been raised by one of the referees, looking at the so-to-say inverse problem: can every Pareto-optimal solution be obtained from a certain imprecision index? Irrespective of whether the answer is affirmative or not, in any way that would give a vivid natural characterization and classification of the Pareto optimal solutions.

Appendix: Canonical sequences of monotone measures: main results

Here we give a brief overview on results concerning canonical sequence of monotone measures. The detailed description with proofs can be found in [5].

Let μ_0 be a monotone measure on $\mathfrak{A}$, $\Gamma=\{B_k\}_{k=1}^{\infty}$ a sequence of sets in $\mathfrak{A}$. Then a sequence of monotone measures $\{\mu_k\}_{k=0}^{\infty}$, defined as

$$\mu_k(A)=\mu_{k-1}(A\cup B_k)-\mu_{k-1}(B_k)+\mu_{k-1}(A\cap B_k),$$

is called a *canonical sequence* of monotone measures, generated by Γ. It is easy to see that if μ_0 is 2-monotone, then the sequence $\{\mu_k\}_{k=0}^{\infty}$ is increasing, i.e. $\mu_0\leq\mu_1\leq...$, and there is a limit $\mu_\Gamma(A)=\lim_{k\to\infty}\mu_k(A)$ for all $A\in\mathfrak{A}$, and $\mu_\Gamma\in M_{2-mon}$. If μ_0 is 2-alternating (submodular), the sequence $\{\mu_k\}_{k=0}^{\infty}$ is decreasing, i.e. $\mu_0\geq\mu_1\geq...$, and the limit measure $\mu_\Gamma(A)=\lim_{k\to\infty}\mu_k(A)$, $A\in\mathfrak{A}$, is also 2-alternating. For our purpose, it is sufficient to consider the finite case where $\mathfrak{A}=2^X$, $\Gamma=\{B_k\}_{k=1}^{m}$, and $\mu_\Gamma=\mu_m$.

Two sequences $\Gamma_1=\{B_k\}_{k=1}^{n}$ and $\Gamma_2=\{C_k\}_{k=1}^{m}$ in $\mathfrak{A}$ are called to be *equivalent* ($\Gamma_1\sim\Gamma_2$) iff $\mu_{\Gamma_1}=\mu_{\Gamma_2}$ for any generating monotone measure μ_0.

Theorem 1. *Let $\Gamma_A=\{A_k\}_{k=1}^{n}\subseteq\mathfrak{A}$. Then there is a increasing sequence of sets $\Gamma_B=\{B_k\}_{k=1}^{m}\subseteq\mathfrak{A}$, $B_1\subseteq B_2\subseteq...\subseteq B_m$, such that $\Gamma_A\sim\Gamma_B$. Minimal algebras $\mathfrak{A}_A$ and $\mathfrak{A}_B$, generated by Γ_A and Γ_B respectively, coincide, i.e. $\mathfrak{A}_A=\mathfrak{A}_B$.*

Let $\mu\in M_{mon}$ be a monotone measure on $\mathfrak{A}$. We call a set $B\in\mathfrak{A}$ an *additive element* w.r.t. μ iff $\mu(A)=\mu(A\cup B)-\mu(B)+\mu(A\cap B)$ for all $A\in\mathfrak{A}$. It is easy to check that $\varnothing, X$ are additive elements w.r.t. any $\mu\in M_{mon}$ and the set of all additive elements w.r.t. a monotone measure μ is an algebra.

Theorem 2. *Let $\{\mu_n\}_{n=0}^{\infty}$ be a canonical sequence of monotone measures, generated by $\{B_n\}_{n=1}^{\infty}\subseteq\mathfrak{A}$. Denote by $\mathfrak{M}_n$ the algebra, consisting of all additive elements w.r.t. μ_n. Then*

1) $\mathfrak{M}_0\subseteq\mathfrak{M}_1\subseteq...\subseteq\mathfrak{M}_n\subseteq....$;

2) μ_n is additive on $\mathfrak{M}_n$;

3) if $C \in \mathfrak{M}_n$, *then* $\mu_n(C) = \mu_k(C)$ *for* $k \geq n$;

4) $\{B_1, B_2, .., B_n\} \subseteq \mathfrak{M}_n$.

Notice that the above results imply several important consequences, which are used in this paper. In particular, let $X = \{x_1, ..., x_n\}$. $\mathfrak{A} = 2^X$, $\mu_0 \in M_{2-mon}$. Consider a canonical sequence of 2-monotone measures, generated by $\Gamma_A = \{A_k\}_{k=1}^m \subseteq \mathfrak{A}$, assuming that the minimal algebra containing Γ_A coincides with $\mathfrak{A}$. Then, by Theorem 2, $\mu_\Gamma \geq \mu$, μ_Γ is additive on $\mathfrak{A}$, i.e. μ_Γ is a probability measure, and by Theorem 1, there is a maximal chain $\Gamma_B = \{B_k\}_{k=0}^n \subseteq \mathfrak{A}$ such that $\varnothing = B_0 \subset B_1 \subset ... \subset B_n = X$, $|B_k \setminus B_{k-1}| = 1$, $k = 1, .., n$; μ_Γ is uniquely defined by $\mu_\Gamma(B_k) = \mu_0(B_k)$, $k = 1, .., n$.

Acknowledgement. We are very grateful to the referees for many very helpful and detailed remarks.

Andrey Bronevich from his side expresses his sincere thanks to the German Academic Exchange Service (DAAD), Russian Ministry of Education, Ludwig-Maximilians University and Prof. Dr. Thomas Augustin for the research opportunity provided.

References

[1] T. Augustin. *Optimale Tests bei Intervallwahrscheinlichkeit.* Vandenhoeck & Ruprecht, Göttingen, 1998.

[2] T. Augustin. On data-based checking of hypotheses in the presence of uncertain knowledge. In: Gaul, W., Locarek-Junge, H. (eds.). Classification in the Information Age. Springer. Heidelberg, 1999, pp. 127 – 135.

[3] T. Augustin. Neyman–Pearson testing under interval probability by globally least favorable pairs. Reviewing Huber–Strassen theory and extending it to general interval probability. *Journal of Statistical Planning and Inference* 105: 149 – 173, 2002.

[4] T. Augustin and R. Hable. On the impact of robust statistics on imprecise probability models: a review. To appear in: *Proc. of the the 10th International Conference on Structural Safety and Reliability*, Osaka, 2009.

[5] A.G. Bronevich. Canonical sequences of fuzzy measures. In Proc. of International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU-2004), Perugia-Italy, 2004, 8 pp.

[6] A.G. Bronevich. An investigation of ideals in the set of fuzzy measures. *Fuzzy Sets and Systems* 152: 271 – 288, 2005.

[7] A.G. Bronevich. On the closure of families of fuzzy measures under eventwise aggregations. *Fuzzy Sets and Systems* 153: 45 – 70, 2005.

[8] A.G. Bronevich and A.E. Lepskiy. Measuring uncertainty with imprecision indices. In de Cooman, G., Vejnarova. J., Zaffalon, M. (eds.). Proc. of the Fifth International Symposium on imprecise probability: Theory and Applications, Prague, Czech Republic, 2007, pp. 47 – 56.

[9] A. Buja. On the Huber-Strassen theorem. *Probability Theory and Related Fields* 73: 149 – 152, 1986.

[10] A. Chateauneuf and J.Y. Jaffray. Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. *Mathematical Social Sciences* 17: 263 – 283, 1989.

[11] G. Choquet. Theory of capacities. *Ann. Inst. Fourier*, 5: 131 – 295, 1954.

[12] V.I. Danilov. *Lectures on Game Theory*. Russian Economical School, Moscow, 2002. (In Russian)

[13] A.P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics* 37: 325 – 339, 1967.

[14] Denneberg D. *Non-additive Measure and Integral*. Dordrecht, Kluwer, 1997.

[15] I. Gilboa and D. Schmeidler. Updating ambiguous beliefs. *Journal of Economic Theory* 59: 33 – 49, 1993.

[16] R. Hable. Data-based decisions under imprecise probability and least favorable models. *International Journal of Approximate Reasoning* 50: 642 – 654, 2009.

[17] R. Hafner. Konstruktion robuster Teststatistiken. In: Schach, S., Trenkler, G. (eds.). Data Analysis and Statistical Inference. Festschrift in Honour of Prof. Dr. Friedhelm Eicker. Eul. Bergisch Gladbach, 1992, pp. 145-160.

[18] P.J. Huber, V. Strassen. Minimax tests and the Neyman–Pearson lemma for capacities. *Ann. Statist.* 1: 251–263, 1973.

[19] G. J. Klir. *Uncertainty and Information: Foundations of Generalized Information Theory*, Hoboken, NJ: Wiley-Interscience, 2006.

[20] L.V. Utkin and T. Augustin. Powerful algorithms for decision making under partial prior information and general ambiguity attitudes. In Cozman, F.G., Nau, R., Seidenfeld, T. (eds.). Proc. of the Fourth International Symposium on Imprecise Probabilities and their Applications, Pittsburgh (Carnegie Mellon), SIPTA, Manno, 2005, pp. 349-358.

[21] P. Walley. *Coherent lower (and upper) probabilities.* Research Report, University of Warwick, Department of Statistics, Warwick, 1981.

[22] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, London, 1991.

On the use of a new discrepancy measure to correct incoherent assessments and to aggregate conflicting opinions based on imprecise conditional probabilities

Andrea Capotorti, Giuliana Regoli, Francesca Vattari
Dipartimento di Matematica e Informatica, Perugia, Italy
{capot,regoli,francesca.vattari}@dipmat.unipg.it

Abstract

We give a preliminary study of a new procedure to correct incoherent imprecise conditional probability assessments. The procedure is based on parametric optimization problems which have as objective function a new discrepancy measure. We show through simple examples how the procedure of correcting incoherent assessments can be properly extended to aggregate conflicting opinions, and can be generalized to embed importance weights of each assessment.

Keywords. Imprecise conditional probabilities, inconsistency handling, aggregation opinions, divergence measures.

1 Introduction

In this paper we illustrate a preliminary study for the adoption of a new procedure to correct inconsistent imprecise conditional probability assessments. The procedure is based on parametric optimization problems the objective function of which is a discrepancy measure recently introduced in [4] for a similar purpose with respect to precise assessments. Such discrepancy originates from a peculiar choice of a scoring rule, and it behaves like ordinary divergences among probability distributions.

Care must be taken for the notion of incoherence. In fact, for imprecise conditional probability assessments, different coherence requirements are possible (see e.g. the comparison among them done in [20, 21]). We choose to proceed along the line of de Finetti [12, 13], adopting the most stringent generalization of his coherence notion for precise assessments to imprecise ones, as proposed by Coletti and Scozzafava (see e.g. [7]).

Assessments inconsistency can naturally arise whenever there is the need to merge different sources of uncertainty information. The extension of our correction procedure to aggregation of opinions comes quite naturally. It is in fact sufficient to formally duplicate the elements in common among the assessments to have a joint one, and treat it as generated by a unique source.

Aggregation of different opinions is actually a subject which has been studied in depth, both in precise (see e.g. [10, 14, 22, 25]) and imprecise (see e.g. [11, 16, 19, 23, 24]) evaluation frameworks. Some aggregation rules are based solely on the assessed values, others rely on auxiliary over structures, like for example second order assessments or risk neutral probabilities. Our choice lies in between: once a specific scoring rule is chosen, the aggregation proceeds "alone" by working only on the assessed values.

The procedure we propose reveals its efficacy especially whenever opinions are given on different domains and the envelope of opinions union turns out to be incoherent "per se".

While theoretical details will be the object of a future contribution, we present here some simple examples to show peculiarities and potentialities of our procedure.

The paper continues with Section 2, where the notation and basic notions are introduced. In particular, the discrepancy mentioned above is described and its justification and properties are reported. After that, in Section 3 we illustrate how to use such discrepancy as objective function of parametric optimization problems, so that, by an iteration, it is possible to select a set of coherent precise assessments whose lower-upper envelopes induce the correction of an initially incoherent assessment. In Section 4 we extend the procedure to the aim of aggregating different opinions. This generalization comes quite naturally by a simple rewriting of the joint assessment. After that we generalize the discrepancy measure by introducing a weighted version. In fact, it is possible to differentiate the importance of the single opinions, and inside them of the single assessed values. Finally, we end by Section 6, where a short conclusion is reported.

2 Basic notions

We formalize the domain of the evaluation through a finite family of conditional events of the type $\mathcal{E} = [E_1|H_1, \ldots, E_n|H_n]$.

Events E_i-s usually represent the situations under consideration, while the H_i-s usually represent the different contexts, or scenarios, under which the E_i-s are evaluated.

The basic events $E_1, \ldots, E_n, H_1, \ldots, H_n$ can be endowed with logical constraints, that represent dependencies among particular configurations of them (e.g. incompatibilities, implications, partial or total coincidences, etc.).

In the following E_iH_i will denote the logical connection "E_i and H_i", $\neg E_i$ will indicate "not E_i" and the event $H^0 = \bigvee_{i=1}^{n} H_i$ will represent the whole set of contexts considered.

By the basic events $E_1, \ldots, E_n, H_1, \ldots, H_n$ it is possible to span a sample space $\Omega = \{\omega_1, \ldots, \omega_k\}$, where ω_j represents generic atoms, in some context named "possible worlds ". Note that the sample space Ω and H^0 are not part of the assessment but only auxiliary tools.

The numerical part of the assessment is elicited through interval values

$$\mathbf{lub} = ([lb_1, ub_1], \ldots, [lb_n, ub_n]) \tag{1}$$

thought as honest ranges for the probabilities $p_i = P(E_i|H_i)$, $i = 1, \ldots, n$. Of course, some of the intervals $[lb_i, ub_i]$-s could degenerate to precise values p_i-s.

For assessments like $(\mathcal{E}, \mathbf{lub})$, although defined on finite spaces, there could be different kinds of consistency requirements (for a detailed exposition, among others, refer to [20]). In this paper we focus on the most stringent one: (strong) coherence. By adopting a Bayesian sensitivity analysis interpretation, coherent lower-upper conditional probability assessments $(\mathcal{E}, \mathbf{lub})$ are those the numerical part **lub** of which can be obtained as lower-upper envelopes of sets of coherent precise, i.e. linear, conditional probability assessments on $\mathcal{E}$; coherence for precise assessments is thought in the most general sense of restrictions on $\mathcal{E}$ of full finitely additive conditional probability distributions. For a complete and rigorous description see the exhaustive treatise [9].

It follows that to have a coherent assessment on $\mathcal{E}$, there should exist a set of probability distributions over Ω such that, on one hand it induces probabilities for the $E_i|H_i$-s inside the ranges $[lb_i, ub_i]$, and on the other hand it is such that each lower (lbi_s) and upper (ubi_s) bound of the ranges is attained through at least one distribution in the set.

We denote by $\mathcal{M}$ such set of coherent precise conditional assessments compatible with $(\mathcal{E}, \mathbf{lub})$

$$\begin{aligned} \mathcal{M} \quad := \quad & \{P \text{ coherent} \,|\, lb_i \leq P(E_i|H_i) \leq ub_i, \\ & i = 1, \ldots, n\}. \end{aligned} \tag{2}$$

We shall focus on the situations with an empty $\mathcal{M}$ that characterize incoherent assessments $(\mathcal{E}, \mathbf{lub})$. Such kind of incoherence is usually denoted as "incurring in uniform loss" (see [27]) or as "not g-coherent" (see [1]).

In literature it is commonly faced an other kind of incoherence: $\mathcal{M}$ is not empty but there exist at least one index $i \in \{1, \ldots, n\}$ such that

$$lb_i < \inf_{P \in \mathcal{M}} P(E_i|H_i) \quad \text{or} \quad \sup_{P \in \mathcal{M}} P(E_i|H_i) < ub_i \quad . \tag{3}$$

In this cases $(\mathcal{E}, \mathbf{lub})$ is said to "avoid uniform loss but not strong coherent" (see [26, 28]), or simply "incoherent" (see [7]). This second kind of incoherence can be directly solved by computing the "natural extension" of $(\mathcal{E}, \mathbf{lub})$ (see again [1, 21, 27], among others).

Actually, there is a third type of incoherence: when the assessment is not "weak coherent" (see again [26]). For finite domains, this subtle "weak incoherence" derives, as well illustrated in [20, 21], by the exclusion of conditioning on events with zero probability. Since, on the contrary, we believe that it is important to include such assessments in $\mathcal{M}$ (see e.g. [8]), we do not tackle this further type of inconsistency.

Whenever $(\mathcal{E}, \mathbf{lub})$ incurs in a uniform loss, there is no unique way to adjust it. In this paper we propose to find out the "closest" correction, with a specific choice for the "distance" notion. In [4] we already did this for precise assessments taking advantage of the aforementioned discrepancy measure. We propose now to extend such method to imprecise assessments by generalizing the discrepancy among sets of assessments.

Before introducing the discrepancy measure, we need some further auxiliary notions.

Every probability distribution $\alpha : \mathcal{P}(\Omega) \to \mathbb{R}$ corresponds to a non-negative vector $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k]$, with $\alpha_j = \alpha(\omega_j)$; then for every event E it will be $\alpha(E) = \sum_{\omega_j \subseteq E} \alpha_j$. We will refer to a nested hierarchy of probability distributions over Ω. This to properly separate inner from boundary situations:

- let $\mathcal{A} := \{\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k] \,|\, \sum_1^k \alpha_i = 1, \alpha_j \geq 0, j = 1, \ldots, k\}$ represents the whole set of probability distributions on Ω;

- let $\mathcal{A}_0 := \left\{\boldsymbol{\alpha} \in \mathcal{A} | \alpha(H^0) = \alpha(\bigvee H_i) = 1\right\}$ be the subset of probability distributions on Ω that concentrate all the probability mass on the contemplated scenarios[1];
- let $\mathcal{A}_1 := \{\boldsymbol{\alpha} \in \mathcal{A}_0 | \alpha(H_i) = \sum_{\omega_j \in H_i} \alpha_j > 0, i = 1, \ldots, n\}$ be the subset of probability distributions on Ω that give positive probability to every scenario;
- let $\mathcal{A}_2 = \{\boldsymbol{\alpha} \in \mathcal{A}_1 | 0 < \frac{\sum_{j:\, \omega_j \subset E_i H_i} \alpha_j}{\sum_{j:\, \omega_j \subset H_i} \alpha_j} < 1, i = 1, \ldots, n\}$ be the subset of probability distributions that avoid boundary values $\{0, 1\}$ for the conditional probabilities.

Any $\boldsymbol{\alpha} \in \mathcal{A}_1$ induces a coherent precise conditional assessment on $\mathcal{E}$

$$\mathbf{q}_{\boldsymbol{\alpha}} := [q_i = \frac{\sum_{j:\, \omega_j \subset E_i H_i} \alpha_j}{\sum_{j:\, \omega_j \subset H_i} \alpha_j}, i = 1, \ldots, n]. \quad (4)$$

Associated to any (coherent or not) precise assessment $\mathbf{p} = [p_1, \ldots, p_n] \in (0, 1)^n$ over $\mathcal{E} = [E_1|H_1, \ldots, E_n|H_n]$ we can introduce a scoring rule

$$S(\mathbf{p}) := \sum_{i=1}^{n} |E_i H_i| \ln p_i + \sum_{i=1}^{n} |\neg E_i H_i| \ln(1 - p_i) \quad (5)$$

with $|\cdot|$ indicator function of unconditional events.

Note that such scoring rule is not defined for boundary values 0 or 1 of the assessed probabilities. This is of course a limitation of our approach, but all the same we believe it is significant. In fact if the assessor had so strong a belief in assessing such extreme values, it could mean that the component did not want to be the object of a settlement. Hence if any of the lbi_s or of the ubi_s turn out to be 0 or 1, they are maintained fixed in their values, if this of course will not induce any evident contradiction, otherwise they must be treated outside our procedure.

This score $S(\mathbf{p})$ is an "adaptation" of the "proper scoring rule" for probability distributions proposed by Lad in [18](pag. 355). We have extended it to partial and conditional probability assessments.

Such a score is motivated by a conditional event $E_i|H_i$ being a three-valued logical entity, partitioning Ω in three parts (omnia Gallia divisa est in partes tres): the atoms satisfying $E_i H_i$ and therefore verifying the conditional, those satisfying $\neg E_i H_i$, thus falsifying the conditional, and those not fulfilling the context H_i, to which the conditional may not be applied at all. Hence the assessor of $\mathbf{p}$ "loses less" the higher are the probabilities assessed for events that are verified, and at the same time, the lower are the probabilities assessed for those that are not verified. The values assessed on events that turn out to be undetermined do not influence the score. In fact the realization of the random value $S(\mathbf{p})$ when the atom ω_j occurs is

$$S_j(\mathbf{p}) = \sum_{i:\, E_i H_i \supset \omega_j} \ln p_i + \sum_{i:\, \neg E_i H_i \supset \omega_j} \ln(1 - p_i). \quad (6)$$

The simultaneous involvement in this score of events that turn out to be true and of those that turn out to be false modifies the peculiar property of the usual logarithmic scoring rule to depend only on the true ones.

We now have all the elements to introduce the "discrepancy" between a precise assessment $\mathbf{p}$ over $\mathcal{E}$ and a distribution $\boldsymbol{\alpha} \in \mathcal{A}_2$, with respect to its induced conditional coherent assessment $\mathbf{q}_{\boldsymbol{\alpha}}$, as

$$\Delta(\mathbf{p}, \boldsymbol{\alpha}) := E_{\boldsymbol{\alpha}}(S(\mathbf{q}_{\boldsymbol{\alpha}}) - S(\mathbf{p})) \quad (7)$$

$$= \sum_{j=1}^{k} \alpha_j [S_j(\mathbf{q}_{\boldsymbol{\alpha}}) - S_j(\mathbf{p})]. \quad (8)$$

The distributions $\boldsymbol{\alpha}$ are restricted to be in $\mathcal{A}_2$ because only there the scoring rule $S(\mathbf{q}_{\boldsymbol{\alpha}})$ is properly defined. It is however possible to extend by continuity the previous definition of $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ to any distribution $\boldsymbol{\alpha}$ in $\mathcal{A}_0$ through the expression

$$\Delta(\mathbf{p}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \ln(\frac{q_i}{p_i}) \alpha(E_i H_i) + \ln(\frac{1 - q_i}{1 - p_i}) \alpha(\neg E_i H_i) \quad (9)$$

$$= \sum_{i=1}^{n} \alpha(H_i) \left(q_i \ln(\frac{q_i}{p_i}) + (1 - q_i) \ln(\frac{1 - q_i}{1 - p_i}) \right). \quad (10)$$

This discrepancy $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ behaves in a way that is analogous to other usual Bregman divergences[2] (see [2]). In fact in[5] we formally proved that the following properties hold:

- $\Delta(\mathbf{p}, \boldsymbol{\alpha}) \geq 0 \quad \forall \boldsymbol{\alpha} \in \mathcal{A}$;
- $\Delta(\mathbf{p}, \boldsymbol{\alpha}) = 0$ iff $\mathbf{p} \equiv \mathbf{q}_{\boldsymbol{\alpha}}$;
- $\Delta(\mathbf{p}, \cdot)$ is convex on $\mathcal{A}_2$;
- $\Delta(\mathbf{p}, \cdot)$ always admits a minimum on $\mathcal{A}_0$;

[1] This is commonly done in conditional frameworks to avoid unpleasant consequences. See Walley[26] about Avoiding Uniform Loss assessments or Holzer[17] about the Principle of Conditional Coherence

[2] Actually $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ turns out to be a generalization of the sum of two different "Bregman divergences".

- If $\Delta(\mathbf{p},\cdot)$ attains its minimum value on $\mathcal{A}_1$; then there is a unique coherent assessment $\mathbf{q}_{\underline{\alpha}}$ on $\mathcal{E}$ such that $\Delta(\mathbf{p},\underline{\boldsymbol{\alpha}})$ is minimum;
- If $\Delta(\mathbf{p},\cdot)$ attains its minimum value on $\mathcal{A}_0 \setminus \mathcal{A}_1$, then any distribution $\boldsymbol{\alpha} \in \mathcal{A}_0$ that minimizes $\Delta(\mathbf{p},\cdot)$ induces the same significant conditional probabilities $(\mathbf{q}_\alpha)_j$ on the conditional events $E_j|H_j$ such that $\alpha(H_j) > 0$.

The last two items are the crucial ones: for precise numerical evaluations $\mathbf{p}$, they always guarantee the existence of a coherent assessment $(\mathcal{E}, \mathbf{q}_{\widetilde{\boldsymbol{\alpha}}})$ "close as much as possible" to $(\mathcal{E},\mathbf{p})$. And this also with respect to the most general notion of conditional coherence that contemplates the hierarchy of the so called "zero layers" (see again [9] for details about this delicate and crucial notion).

3 Correcting incoherent assessments

Let us see how the properties of $\Delta(\mathbf{p},\boldsymbol{\alpha})$ could help us in the correction of an incoherent assessment.

The starting point is that incoherence of $(\mathcal{E},\mathbf{lub})$ is equivalent to the incoherence of any precise assessment $\mathbf{v} = (v_1,\ldots,v_n)$ with $lb_i \le v_i \le ub_i$. On the other hand, the assessor elicitates the bounds lb_i-s and ub_i-s as effectively attainable. For this reason, we iteratively fix a specific bound lb_f (or ub_f) and we find the precise coherent assessment $\tilde{\mathbf{q}}$ that is the closest to the subset of precise assessments $\mathbf{v}$-s that reach lb_f (or ub_f), while remaining inside the ranges $[lb_i, ub_i]$-s for the others elements.

More precisely, by fixing an index $f \in \{1,\ldots,n\}$, we can find two coherent assessments $\underline{\mathbf{q}}_f$ and $\overline{\mathbf{q}}_f$ on $\mathcal{E}$, induced respectively, by the solutions of the following two parametric optimization problems, with parameter $\mathbf{v}$:

$$\text{minimize } \Delta(\mathbf{v},\boldsymbol{\alpha}) \tag{11}$$

under the constraints

$$v_f = lb_f \quad \mathbf{or} \quad v_f = ub_f \tag{12}$$

$$\forall i \neq f \quad lb_i \le v_i \le ub_i \quad , \quad i \in \{1,\ldots,n\} \tag{13}$$

$$\sum_{j:\,\omega_j \subset E_k H_k} \alpha_j = q_k \sum_{j:\,\omega_j \subset H_k} \alpha_j \quad , \quad k = 1,\ldots,n \tag{14}$$

$$\boldsymbol{\alpha} \in \mathcal{A}_0 \quad . \tag{15}$$

The choice in (12) of whether to fix the lower or upper bound distinguishes one problem from the other.

The $n-1$ constraints (13) reflect the compatibility of $\mathbf{v}$ with the other intervals in $\mathbf{lub}$, while the n constraints (14) impose the coherence of the assessment $\mathbf{q}_{\boldsymbol{\alpha}}$ induced by a solution $\boldsymbol{\alpha}$.

Note that if any optimal solution $\widetilde{\boldsymbol{\alpha}}$ of (11) is in $\mathcal{A}_0 \setminus \mathcal{A}_1$, the associated conditional assessment $\mathbf{q}_{\widetilde{\boldsymbol{\alpha}}}$ is properly defined only for those conditional events $E_k|H_k$ with $\widetilde{\alpha}(H_k) > 0$, some component of $\underline{\mathbf{q}}_f$ (or of $\overline{\mathbf{q}}_f$) remaining unspecified. Hence, in these cases, we need to explore other "zero layers". This can be simply done by reiterating the optimization problem over the part of $\mathcal{E}$ with probability of the conditioning events induced by $\widetilde{\boldsymbol{\alpha}}$ equal to 0. The new optimal solutions are distributions defined on sample spaces spanned by the sub-domain, so that they significantly induce some of the unspecified component of $\underline{\mathbf{q}}_f$ (or $\overline{\mathbf{q}}_f$). Since for each iteration there will be at least one conditioning event H_k with strictly positive induced probability, at worst in $n-1$ steps the assessments $\underline{\mathbf{q}}_f$ (or $\overline{\mathbf{q}}_f$) are fully determined.

By letting the index f vary over the full range $1,\ldots,n$ we obtain a set of $2n$ coherent assessments

$$\mathcal{Q} = \{\underline{\mathbf{q}}_f, \overline{\mathbf{q}}_f, f = 1,\ldots,n\}. \tag{16}$$

By definition, the imprecise assessment on $\mathcal{E}$

$$\mathbf{luc} = ([lc_1, uc_1],\ldots,[lc_n, uc_n]), \tag{17}$$

which is bounded by the lower and upper envelope of $\mathcal{Q}$, i.e.

$$lc_i := \min_{\tilde{\mathbf{q}}\in\mathcal{Q}} \tilde{\mathbf{q}}(E_i|H_i) \qquad uc_i := \max_{\tilde{\mathbf{q}}\in\mathcal{Q}} \tilde{\mathbf{q}}(E_i|H_i), \tag{18}$$

is coherent and can be adopted as correction of $\mathbf{lub}$.

Note moreover that we have no guarantees about the uniqueness of $\underline{\mathbf{q}}_f$, or of $\overline{\mathbf{q}}_f$, because the set of optimal solutions

$$\mathcal{O}_f = \{\widetilde{\boldsymbol{\alpha}} \in \mathcal{A}_0 | \widetilde{\boldsymbol{\alpha}} \text{ optimal solution of } (11-15)\} \tag{19}$$

could induce different coherent precise assessments over $\mathcal{E}$. At the moment, numerical experiments support uniqueness, but further theoretical investigations are needed. In any case, if there were different assessments induced by (19), we could take the whole set of them instead of the single $\underline{\mathbf{q}}_f$ (or $\overline{\mathbf{q}}_f$) to determine the envelope (18).

Let us see how our correction procedure works with a simple example.

Example 1 *By borrowing the framework from [15], we consider the domain* $\mathcal{E} = [C|A, C|B, C|A \vee B]$ *built by three basic unconditional logically independent events* A, B, C. *Hence the whole sample space would be of 8 atoms, but only 6 are inside* $H^0 \equiv A \vee B$. *The set of coherent assessments on* $\mathcal{E}$ *is made by the triples*

$[q_1, q_2, q_3] \in [0,1]^3$*, with the last component forced to be in the range*

$$q_3 \in \left[\frac{q_1\, q_2}{q_1 + q_2 - q_1\, q_2}, \frac{q_1 + q_2 - 2q_1\, q_2}{1 - q_1\, q_2} \right] \qquad (20)$$

(see Fig.1). Note the evident non-convexity of this coherent set.

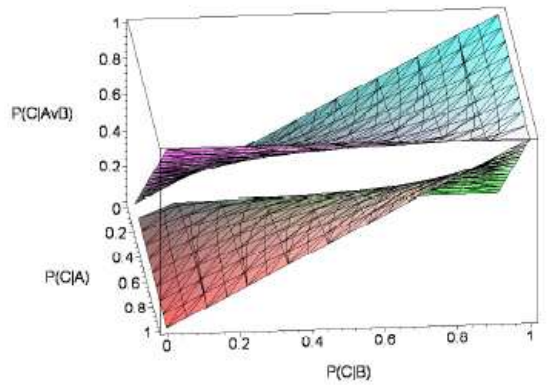


Figure 1: Lower and upper bounds for coherent assessments on $\mathcal{E} = [C|A, C|B, C|A \vee B]$

Let us firstly consider an assessments **lub** *that incurs in a uniform loss*

$\mathcal{E}$	$C\|A$	$C\|B$	$C\|A \vee B$
lb_i	.1	.2	.6
ub_i	.3	.4	.8

. (21)

Incoherence can be highlighted by taking as good the first two components of **lub***, so that the coherent (natural) extension to $C|A \vee B$ should be, by (20), the interval* $[.0714, .5227]$*, that does not overlap the assessed range* $[.6, .8]$*. Hence we have that the set $\mathcal{M}$ of precise assessments compatible with* **lub** *is empty.*
By performing [3] *the 6 optimization problems of type (11), we obtain the following coherent precise assessments:*

$\mathcal{E}$	$C\|A$	$C\|B$	$C\|A \vee B$
$\underline{\mathbf{q}}_1$	.1196	.4797	.5140
$\overline{\mathbf{q}}_1$	.3263	.4344	.5560
$\underline{\mathbf{q}}_2$	.3830	.2558	.4910
$\overline{\mathbf{q}}_2$	.3263	.4344	.5560
$\underline{\mathbf{q}}_3$	.3263	.4344	.5560
$\overline{\mathbf{q}}_3$	.4078	.5440	.6530

, (22)

whose lower-upper envelope results the following coherent imprecise assessment:

$\mathcal{E}$	$C\|A$	$C\|B$	$C\|A \vee B$
lc_i	.1196	.2558	.4910
uc_i	.4078	.5440	.6530

. (23)

[3] Numerical results obtained with the nonlinear optimization software CONOPT of the GAMS package [3]

4 Aggregating conflicting opinions

The merging, or aggregation, of different opinions has a considerable importance for both theoretical and practical aspects. This subject has been widely treated in several scientific fields, and even restricting attention to probabilistic models, there is a great number of proposals. An interesting feature occurs when the different opinions are in conflict, i.e. the whole assessment results incoherent (see e.g. [6, 25] for precise assessments and [16, 19] for imprecise ones).

In our approach, conflict among opinions can be expressed through disjoint intervals associated to the same conditional events, and/or through incoherence of the joint assessment.

Here we propose to adopt the previous procedure, which we have seen to correct incoherent imprecise assessments, also for aggregation purposes.

First of all, if we have evaluations assessed on $\mathcal{E}^s = [E_{1.s}|H_{1.s}, \ldots, E_{n.s}|H_{n.s}]$, with the index $s \in S$ expressing the different sources, we denote the joint domain by $\mathcal{E} = \bigvee_{s \in S} \mathcal{E}^s$.

Secondly, we can replace the possible multiple ranges assigned to single elements of $\mathcal{E}$ duplicating such elements and adding coincidence constraints in list of the logical relationships. For example, if we have two different ranges $[lb'_i, ub'_i]$ and $[lb''_i, ub''_i]$ associated to the same $E_i|H_i \in \mathcal{E}$, we can actually associate the second interval $[lb''_i, ub''_i]$ to a new conditional event $E''_i|H''_i$ added to $\mathcal{E}$, and increase the logical relationships with the constraints

$$E_i H_i \equiv E''_i H''_i \quad ; \qquad (24)$$

$$H_i \equiv H''_i \quad . \qquad (25)$$

In this way, we will have the different opinions joined in a single imprecise (and incoherent) assessment of the type $(\mathcal{E}, \mathbf{lub})$, so that its correction $(\mathcal{E}, \mathbf{luc})$ will represent an aggregation result. Of course, since **luc** is a coherent imprecise assessment, equal intervals $([lc'_i, uc'_i] = [lc''_i, uc''_i])$ will be associated to coincident elements of $\mathcal{E}$ ($E_i|H_i$ and $E''_i|H''_i$).

Let us see how this works with an example.

Example 2 *Let us consider again the framework of the previous Example 1, but now with two different opinions given on separate, but overlapping, subdomains:*

	$C\|A$	$C\|B$	$C\|A \vee B$
$\mathbf{lub}'$	$[.1, .3]$	$[.2, .4]$	–
$\mathbf{lub}''$	–	$[.5, .7]$	$[.6, .8]$

, (26)

by duplicating $C|B$, we obtain a unique whole assessment:

$\mathcal{E}$	$C\|A$	$C\|B$	$C''\|B''$	$C\|A\vee B$
lub	$[.1,.3]$	$[.2,.4]$	$[.5,.7]$	$[.6,.8]$

(27)

with the logical constraints

$$C''B'' \equiv CB \quad , \quad B'' \equiv B. \tag{28}$$

The 8 iterations of the optimization problem of type (11) give the following set $\mathcal{Q}$ of coherent precise assessments:

$\mathcal{E}$	$C\|A$	$C\|B$	$C''\|B''$	$C\|A\vee B$
$\underline{\mathbf{q}}_1$	.1193	.4872	.4872	.5205
$\overline{\mathbf{q}}_1$	.3196	.4612	.4612	.5700
$\underline{\mathbf{q}}_2$	.4053	.3678	.3678	.4855
$\overline{\mathbf{q}}_2$	.3196	.4612	.4612	.5700
$\underline{\mathbf{q}}_3$	.3196	.4612	.4612	.5700
$\overline{\mathbf{q}}_3$	.3547	.5749	.5749	.5380
$\underline{\mathbf{q}}_4$	.3196	.4612	.4612	.5700
$\overline{\mathbf{q}}_4$	.4078	.5440	.5440	.6530

(29)

lower-upper envelope of which gives us the coherent aggregation

$\mathcal{E}$	$C\|A$	$C\|B$	$C\|A\vee B$
luc	$[.1193,.4078]$	$[.3678,.5749]$	$[.4855,.6530]$

(30)

Of course, the approach doesn't change if more than two assessments are given to the same conditional event $E_i|H_i$. We simply have as many coincidence constraints (24,25) as assessed intervals for $E_i|H_i$.

Note that the aggregation (30) we obtained in the previous example deforms all the original opinions (26). This is because the two assessments are strongly in conflict. In fact, apart from the obvious inconsistence due to the two disjoint intervals given on $C|B$, the range $[.6,.8]$, in $\mathbf{lub}''$ associated to $C|A\vee B$, does not overlap the natural extension of $\mathbf{lub}'$

$$[lb'_{C|A\vee B}, ub'_{C|A\vee B}] = [.0714,.5227] \quad . \tag{31}$$

But there are cases in which our procedure gives an aggregation result that reconciles, without misshaping, the original assessments. We can see this in the next example.

Example 3 *If we modify the two separate opinions (26) of the previous example to*

	$C\|A$	$C\|B$	$C\|A\vee B$
$\mathbf{lub}'$	$[.1,.3]$	$[.35,.6]$	–
$\mathbf{lub}''$	–	$[.3,.55]$	$[.1,.6]$

, (32)

our procedure (we skip here the detailed computations of $\mathcal{Q}$) gives as lower-upper envelope the assessment

$\mathcal{E}$	$C\|A$	$C\|B$	$C\|A\vee B$
luc	$[.1,.3]$	$[.3,.6]$	$[.1,.6]$

(33)

that coincide with the least commitment aggregation of (32), that is the lower-upper envelope of the union of the single intervals.

More generally, we can emphasize that our aggregation procedure is particularly significant when joining the different opinions gives an incoherent result, so that each assessed interval influences the result of the merging. On the other hand, note that if all the original opinions are coherent and given on the same domain $\mathcal{E}$, our aggregation result coincides with the least commitment aggregation mentioned above, also named "unanimity rule". Although a "weak" result, this coincidence allows us to compare the behavior of our procedure with many properties for aggregation rules suggested by various authors (see for example [10, 14, 19, 24] among others). In fact, in such situations we trivially have that

- *Unanimity Preservation*: if all the experts agree and give the same assessments for the same events, then the aggregate agrees with all the experts;
- *Symmetry*: for any permutation in the set of the experts, we have the same aggregate;
- *Invariance with respect to noninformative opinion*: the aggregated assessment of N experts yields the same result as the aggregation of the N opinions with a further noninformative opinion, i.e. with one already implied by the natural extension of the aggregation of the first N;
- *Generalized External Bayesianity*: the aggregation of the original assessments, followed by the coherent extension to a new event, gives the same result as the aggregation of the coherent extensions of the initial assessments.

Moreover, we leave to a future investigation some further basic properties, like for example those proposed by Moral and del Sagrado [23].

5 Weighted aggregation

It is possible to associate different weights to the elements of the joined assessment $(\mathcal{E}, \mathbf{lub})$, as we have already done for precise assessments, reflecting either possible repetitions of the values or different trust on

the various sources of information. If we denote by $\mathbf{w} = [w_1, \ldots, w_n]$ such weights, we can adjust the expression of $\Delta(\mathbf{v}, \boldsymbol{\alpha})$ in the optimization problems (11) as

$$\Delta^{\mathbf{w}}(\mathbf{v}, \boldsymbol{\alpha}) := \sum_{i=1}^{n} w_i \alpha(Hi) \left(q_i \ln(\frac{q_i}{v_i}) + (1 - q_i) \ln(\frac{(1 - q_i)}{(1 - v_i)}) \right). \quad (34)$$

We can see directly the effects of this adjustment by a slight modification of Example 2

Example 4 *Let us modify the two original assessments (26) by adding exact overlapping to the previously missing intervals:*

	$C\|A$	$C\|B$	$C\|A \vee B$
$\mathbf{lub}'$	[.1, .3]	[.2, .4]	[.6, .8]
$\mathbf{lub}''$	[.1, .3]	[.5, .7]	[.6, .8]

. (35)

We can now avoid duplicating identical conditional events with identical intervals in the joint assessment **lub***, and yet maintain the information of their multiplicity by using the following frequencies weights:*

$\mathcal{E}$	$C\|A$	$C\|B$	$C''\|B''$	$C\|A \vee B$
lub	[.1, .3]	[.2, .4]	[.5, .7]	[.6, .8]
w	2	1	1	2

. (36)

Performing the 8 optimizations with the new objective function (34) under the same constraints (12-15), we obtain as lower-upper envelope of $\mathcal{Q}$

$\mathcal{E}$	$C\|A$	$C\|B$	$C\|A \vee B$
luc	[.1139, .3747]	[.3750, .6242]	[.5355, .6933]

. (37)

Note how the highest weights have "attracted" the aggregation ranges to the corresponding initial assessments.

Weights w_i could be given in an "imprecise" fashion through intervals $[\underline{w_i}, \overline{w_i}]$, especially when they represent trust levels on the sources of information. This does not change the method, but increases the procedure complexity. In fact, in such cases, we can think of the w_i in (34) as further variables in the optimization problems (11-15), with additional constraints $\underline{w_i} \leq w_i \leq \overline{w_i}$. This will affect the numerical expression of the elements inside $\mathcal{Q}$ in (16), but all other considerations will remain the same.

6 Conclusion

The core of our proposal is in the parametric optimization problems (11), based on the discrepancy measure $\Delta(\cdot, \boldsymbol{\alpha})$. Such discrepancy was originally proposed in [4] by a generalization of the logarithmic scoring rule to partial conditional assessments, and has been used to adjust precise evaluations. In [6] we have extended its use to merge different sources of information, and now to correct incoherent imprecise conditional probability assessments.

We have seen through examples (1-4) that the procedure to correct incoherent assessments can be properly extended to aggregate different opinions and generalized to embed importance weights of each assessment. Effectiveness changes if the joint assessment has a coherent least commitment aggregation or not. In fact, if the lower-upper envelope of the union of the opinions turned out to be coherent, our procedure weakens its peculiarity and reduces to the so called "unanimity rule". Anyhow, our proposal is meaningful in the situations when most of the known rules do not apply. In fact, our procedure applies also when the domains of the opinions do not coincide and the numerical parts are strongly inconsistent, so that the aggregation turns out to be a reasonable compromise between the elicited values and the consistency requirement.

This paper reflects just a preliminary study, because, as already mentioned, theoretical aspects will have to be fixed. To begin with, we need to investigate the presumed uniqueness of the assessments induced by the optimal solutions of the parametric optimization problems (11). In fact, the same method applies also when the solution is not unique, but some operational troubles could appear.

Another open problem is about complexity. The check of coherence is already a NP-complete problem "per se". As a consequence, our parametric nonlinear optimization problems (11 - 15) are even harder. Modern optimization tools like GAMS make medium-size problems treatable with some tens of events. One would need heuristic procedures for larger domain problems.

Yet another important further investigation would be to study the relationships with other aggregation rules, in particular comparing properties, and characterizing possible coincidences.

References

[1] V. Biazzo and A. Gilio: A generalization of the fundamental theorem of de Finetti for imprecise conditional probability assessments.*International Journal of Approximate Reasoning*, 24, 251-272, 2000.

[2] L. M. Bregman. The relaxation method of find-

ing the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Physics*, 7:200–217, 1967.

[3] A. Brooke , D. Kendrick , A. Meeraus and R. Raman. GAMS: a Users Guide , Washington, D.C.: GAMS Development Corp, 2003.

[4] A. Capotorti, G. Regoli. Coherent correction of inconsistent conditional probability assessments. in *Proc. of IPMU'08 - Malaga (Es)*, 2008.

[5] A. Capotorti, G. Regoli, F. Vattari. Theoretical properties of a discrepancy measure among partial conditional probability assessments. *To appear*

[6] A. Capotorti, G. Regoli, F. Vattari. Merging different probabilistic information sources through a new discrepancy measure. Submitted for the acceptance in the *Proceedings of WUPES'09*, to appear.

[7] G. Coletti and R. Scozzafava: Conditional measures: old and new. In *New trends in Fuzzy Systems*, World Scientific: 107-120, 1998.

[8] G. Coletti, R. Scozzafava. The role of coherence in eliciting and handling imprecise probabilities and its application to medical diagnosis. *Information Science*, 13:41–65, 2000.

[9] G. Coletti, R. Scozzafava. *Probabilistic Logic in a Coherent Setting*, Dordrecht: Kluwer, Series "Trends in Logic", 2002.

[10] R.M. Cooke. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. New York: Oxford University Press.

[11] G. de Cooman, M. C. M. Troffaes. Coherent lower previsions in systems modelling: products and aggregation rules. *Reliability Engineering and System Safety*, 85(1-3):113-134, 2004.

[12] B. de Finetti: Sull' impostazione assiomatica delle probabilità. *Annali Univ. Trieste*, **19**, 3-55 1949. (Engl. transl.: Ch. 5 in *Probability, Induction, Statistics* Wiley, London, 1972).

[13] B. de Finetti: *Teoria della probabilità*, Einaudi, Torino, 1970. (Engl. transl.: *Theory of Probability*, Vol.1 and 2. Wiley, Chichester, 1974).

[14] C. Genest, JV Zidek. Combining probability distributions: A critique and an annotated bibliography, *Statistical Science*, 1:114–148, 1986.

[15] A. Gilio. Probabilistic Relations Among Logically Dependent Conditional Events, *Soft Computing*,3:154–161, 1999.

[16] M. Ha-Duong, Hierarchical fusion of expert opinions in the Transferable Belief Model, application to climate sensitivity, *International Journal of Approximate Reasoning*, 49, 555-574, 2008.

[17] S. Holzer On coherence and conditional prevision. *Bull. Unione Matematica Italiana, Analisi funzionale e applicazioni*. 6(4): 441-460, 1985.

[18] F. Lad, *Operational Subjective Statistical Methods: a mathematical, philosophical, and historical introduction*, New York: John Wiley, 1996.

[19] R. F. Nau. The aggregation of imprecise probabilities, *Journal of Statistical Planning and Inference* 105(1):265–282, 2002.

[20] E. Miranda, Updating coherent previsions on finite spaces, *Fuzzy Sets and Systems*(2008) to appear,doi: 10.1016/j.fss.2008.10.005.

[21] E. Miranda, M. Zaffalon. Coherence graph, *Artificial Intelligence*, 173:104-144, 2009.

[22] S. Rässler, *Statistical Matching: a frequentist theory, practical applications and alternative Bayesian applications*, Springer, 2002.

[23] S. Moral, J. del Sagrado. Aggregation of imprecise probabilities. In: Bouchon-Meunier, *Aggregation and Fusion of Imperfect Information*, Physica-Verlag, New York, 162–188, 1998.

[24] M.C.M. Troffaes Generalizing the Conjunction Rule for aggregating Conflict Expert Opinions *International Journal of Intelligent Systems*, 21: 361–380, 2006.

[25] B. Vantaggi. Statistical matching of multiple sources: A look through coherence, *International Journal of Approximate Reasoning*, 49(3):701–711, 2008.

[26] P. Walley. *Statistical reasoning with Imprecise Probabilities*, Chapman and Hall, London, 1991.

[27] Walley P. Pelessoni R., Vicig P.: Direct Algorithms for Checking Coherence and Making Inferences from Conditional Probability Assessments, *Journal of Statistical Planning and Inference*, 126 (1), 119–151, 2004.

[28] P.M. Williams: Note on conditional previsions. *School of Mathematical and Physical Sciences*, The University of Sussex, working paper, 1975.

A Generalization of Credal Networks

Marco E. G. V. Cattaneo
Department of Statistics, LMU Munich
cattaneo@stat.uni-muenchen.de

Abstract

The likelihood approach to statistics can be interpreted as a theory of fuzzy probability. This paper presents a generalization of credal networks obtained by generalizing imprecise probabilities to fuzzy probabilities; that is, by additionally considering the relative plausibility of different values in the probability intervals.

Keywords. Bayesian networks, credal networks, graphical models, d-separation, imprecise probabilities, updating, likelihood function, hierarchical model, fuzzy probabilities.

1 Introduction

A common interpretation of membership functions of fuzzy sets is as statistical likelihood functions. With this interpretation, the well-established likelihood approach to statistics appears as a theory of fuzzy probabilities. These generalize imprecise probabilities by additionally considering the relative plausibility of different values in the probability or expectation intervals. Besides the increased expressive power, the fundamental advantage of the likelihood-based fuzzy probabilities with respect to imprecise probabilities is the ability of using all the information provided by the data. In fact, the resulting hierarchical model exploits the outstanding statistical properties of the likelihood function, which makes it an ideal basis for inference and decision making (see Cattaneo, 2005, 2007).

In the present paper, the hierarchical model is used in the framework of belief networks, to describe the uncertain knowledge about the values of the involved variables. This leads to a generalization of Bayesian networks and credal networks, combining the possibility of imprecision in the probability values with the ability of using all the information provided by the data.

In Section 2 the hierarchical model is briefly introduced (see Cattaneo, 2008*a*, for a more detailed description), while in Section 3 some aspects of the model of great practical importance are presented. Finally, in Section 4 the hierarchical networks are defined and compared with credal networks.

2 Hierarchical Model

In most theories of imprecise probability, the model corresponds to a set $\mathcal{P}$ of probability measures on a measurable space $(\Omega, \mathcal{A})$. The set $\mathcal{P}$ is often assumed to be convex, and when an event $A \in \mathcal{A}$ is observed, $\mathcal{P}$ is usually updated to

$$\mathcal{P}' = \{P(\cdot \mid A) : P \in \mathcal{P},\, P(A) > 0\} \qquad (1)$$

(that is, each $P \in \mathcal{P}$ is conditioned on A). The conditional probability measure $P(\cdot \mid A)$ is obtained by normalizing the "restricted" measure $P(\cdot \cap A)$, but the normalization step deletes the information about the value $P(A)$. The values $P_1(A), P_2(A)$ describe the relative ability of the probability measures $P_1, P_2 \in \mathcal{P}$ to forecast the observed event A (before observing it): the larger the probability value, the better the forecast. These values are combined in the *likelihood* function lik' on $\mathcal{P}'$ defined (up to a positive multiplicative constant) by

$$lik'(P') \propto \sup_{P \in \mathcal{P} \,:\, P(\cdot \mid A) = P'} lik(P)\, P(A) \qquad (2)$$

for all $P' \in \mathcal{P}'$, where lik was the likelihood function on $\mathcal{P}$ before observing A. The likelihood function is a central concept in statistical inference: it is usually interpreted as a measure of the *relative* plausibility of the probability measures as models of the reality under consideration (proportional likelihood functions are considered equivalent). When A is the first observed event, the *prior* likelihood function $lik : \mathcal{P} \to (0, \infty)$ can be interpreted as a (subjective) measure of the relative plausibility of the elements of $\mathcal{P}$ according to the prior information (see also Dahl,

2005). In particular, prior ignorance is described by a constant prior likelihood function lik; in this case, (2) corresponds to the usual definition of statistical likelihood function induced by the observation of the event A (apart from the fact that lik' is defined on $\mathcal{P}'$ instead of $\mathcal{P}$). In general, the prior likelihood function is interpreted and used as if it were the statistical likelihood function induced by (hypothetical) past data.

In the likelihood approach to statistics (see for example Pawitan, 2001), the likelihood of a set of probability measures is usually defined as the supremum of the likelihood of its elements (this idea is used also in (2), if there are several $P \in \mathcal{P}$ such that $P(\cdot \mid A) = P'$). When lik is a likelihood function on $\mathcal{P}$, the set function LR on $2^{\mathcal{P}}$ defined by

$$LR(\mathcal{H}) = \frac{\sup_{P\in\mathcal{H}} lik(P)}{\sup_{P\in\mathcal{P}} lik(P)}$$

for all $\mathcal{H} \subseteq \mathcal{P}$ (in this paper, $\sup \varnothing = 0$) is a normalized *possibility* measure with possibility distribution proportional to lik. A possibility distribution can also be considered as the membership function of a *fuzzy* set (see Zadeh, 1978). In the present paper, possibility distributions and membership functions are interpreted as proportional to likelihood functions: this is a common interpretation (see in particular Hisdal, 1988, and Dubois, 2006). Hence, it suffices to consider normalized fuzzy sets and normalized possibility measures, but grades of membership and degrees of possibility have only a relative meaning.

The set $\mathcal{P}$ of probability measures and the likelihood function lik on $\mathcal{P}$ can be considered as the two levels of a *hierarchical* model: these two levels describe different kinds of uncertainty (probabilistic and possibilistic, respectively). When an event $A \in \mathcal{A}$ is observed, the two levels $\mathcal{P}$ and lik of the hierarchical model are updated to $\mathcal{P}'$ and lik' according to (1) and (2), respectively. The uncertain knowledge about the value $g(P)$ of a function $g : \mathcal{P} \to \mathcal{G}$ is described by the induced possibility measure $LR \circ g^{-1}$ on $\mathcal{G}$ (in this paper, g^{-1} denotes the set function associating to each subset of $\mathcal{G}$ its inverse image under g); that is, by the normalized fuzzy subset of $\mathcal{G}$ with membership function proportional to the *profile* likelihood function lik_g on $\mathcal{G}$ defined (up to a positive multiplicative constant) by

$$lik_g(\gamma) \propto \sup_{P\in\mathcal{P}\,:\,g(P)=\gamma} lik(P)$$

for all $\gamma \in \mathcal{G}$. In particular, if g associates to each probability measure $P \in \mathcal{P}$ the expectation $g(P) = E_P(X)$ of a random variable X, or the probability $g(P) = P(B)$ of an event $B \in \mathcal{A}$, then the normalized fuzzy subset of $\mathbb{R}$ with membership function proportional to lik_g can be interpreted as the fuzzy expectation of X, or the fuzzy probability of B, respectively. In this sense, the likelihood approach to statistics can be interpreted as a theory of fuzzy probability. The discussion on how to evaluate by one or more real numbers the normalized fuzzy subset of $\mathbb{R}$ with membership function proportional to lik_g goes beyond the scope of the present paper (but see Cattaneo, 2007, for some interesting results): only the α-cut

$$\left\{x \in \mathbb{R} : lik_g(x) \geq \alpha \sup_{y\in\mathbb{R}} lik_g(y)\right\}$$

with $\alpha \in (0, 1)$ will be considered here. This is a likelihood-based confidence region for $g(P)$, whose coverage probability can often be approximated thanks to the result of Wilks (1938): in particular, 95% coverage probability corresponds to $\alpha = 0.1465$.

Example 1 *Consider an urn containing* 3 *balls: one ball is white, another is black, while the third one could be white or black. We have no idea about the color (white or black) of the third ball, but we can perform a sequence of random draws with replacement from the urn, and observe the colors of the balls drawn. Conditional on the composition of the urn, these observations can be described as a sequence of independent Bernoulli trials with constant probability* $\frac{1}{3}$ *or* $\frac{2}{3}$ *of observing a black ball (depending on the color of the third ball: white or black, respectively). We shall never be able to determine with absolute certainty the composition of the urn, but if in a long sequence of draws the proportion of black balls is approximately* $\frac{2}{3}$*, then it is much more plausible that the color of the third ball is black than it is white.*

Let $\mathcal{P}$ *be the (convex) set of probability measures resulting from the only imprecise prior probability measure about the composition of the urn (that is, about the color of the third ball) such that the probability of observing a black ball in the first draw is described by the interval* $[\frac{1}{3}, \frac{2}{3}]$*. This is the vacuous imprecise prior, and therefore, if* $\mathcal{P}$ *is updated according to (1), then the (posterior) imprecise probability of observing a black ball in the next draw remains* $[\frac{1}{3}, \frac{2}{3}]$*, independently of the number and colors of the balls drawn. By contrast, the (posterior) fuzzy probability of observing a black ball in the next draw (resulting from the hierarchical model with constant prior likelihood function on* $\mathcal{P}$*) evolves as expected: it tends to concentrate on the value* $\frac{2}{3}$*, when in a sequence of draws of increasing length the proportion of black balls remains approximately* $\frac{2}{3}$*. Figure 1 shows the graphs of the membership functions of the fuzzy probability p of observing a black ball in the next draw: prior to any draw (dotted line), after drawing* 2 *white balls and* 5 *black balls (dashed line), and after drawing* 8 *white balls and* 15 *black balls (solid line); in particular, the*

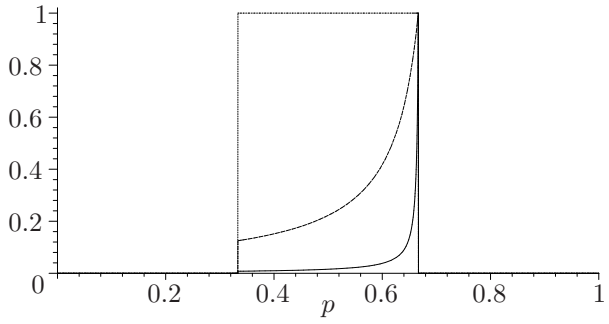


Figure 1: Membership functions of the fuzzy probability p of observing a black ball in the next draw (in the situation of Example 1): prior to any draw (dotted line), after drawing 2 white balls and 5 black balls (dashed line), and after drawing 8 white balls and 15 black balls (solid line).

corresponding α-cuts with $\alpha = 0.1465$ are the intervals $[0.333, 0.667]$, $[0.389, 0.667]$, *and* $[0.651, 0.667]$, *respectively.*

These membership functions can be easily obtained by applying the results of Section 3; the detailed calculations will be presented in Example 4.

The hierarchical model with levels $\mathcal{P}$ and lik generalizes the imprecise probability model $\mathcal{P}$, since the probabilistic level is updated in the same way (1) as the imprecise probability model, while the possibilistic level carries additional information. In particular, the fuzzy expectation of a random variable X is a fuzzy subset of the imprecise expectation

$$[\underline{E}(X), \overline{E}(X)] = [\inf_{P\in\mathcal{P}} E_P(X), \sup_{P\in\mathcal{P}} E_P(X)]\,,$$

and the fuzzy probability of an event $B \in \mathcal{A}$ is a fuzzy subset of the imprecise probability

$$[\underline{P}(B), \overline{P}(B)] = [\inf_{P\in\mathcal{P}} P(B), \sup_{P\in\mathcal{P}} P(B)]\,,$$

since their membership functions are constant equal to 0 outside these intervals (for example, the fuzzy probabilities of Figure 1 are fuzzy subsets of the imprecise probability $[\frac{1}{3}, \frac{2}{3}]$). That is, fuzzy probabilities generalize imprecise probabilities by additionally considering the relative plausibility of different values in the probability intervals (imprecise probabilities correspond to the crisp case of fuzzy probabilities). This additional information allows us in particular to get out of the state of complete ignorance; that is, to reach nontrivial conclusions also when starting with the vacuous prior, as in Example 1. Alternative updating rules for the imprecise probability model $\mathcal{P}$, making use of some information contained in the possibilistic level lik, have been proposed in particular by Moral (1992), Wilson (2001), and Held et al. (2008): these updating rules discard some of the less plausible probability measures in $\mathcal{P}$, but this can lead to important problems, since any discarded probability measure can become the most plausible one in the light of new data. To avoid these problems, it is necessary to store more information than it is possible in an imprecise probability model: the hierarchical model provides a simple solution.

When the probabilistic level of the hierarchical model is a singleton $\mathcal{P} = \{P\}$, the possibilistic level contains no information, since the likelihood function is defined only up to a positive multiplicative constant. In this case, the membership function of the fuzzy expectation of a random variable X, or of the fuzzy probability of an event $B \in \mathcal{A}$, is the indicator function of $\{E_P(X)\}$, or of $\{P(B)\}$, respectively; and when an event $A \in \mathcal{A}$ is observed, the probabilistic level is updated according to (1) by conditioning P on A. Hence, the purely probabilistic description of uncertain knowledge about $\omega \in \Omega$ (that is, the Bayesian model) is a special case of the hierarchical model. The same is true also for the purely possibilistic description of uncertain knowledge about $\omega \in \Omega$: a normalized possibility measure Π on Ω with possibility distribution π can be described by the hierarchical model with as probabilistic level the set $\mathcal{P} = \{\delta_\omega : \omega \in \Omega, \pi(\omega) > 0\}$ (where δ_ω is the Dirac measure on Ω concentrated on ω), and as possibilistic level the likelihood function lik on $\mathcal{P}$ defined (up to a positive multiplicative constant) by $lik(\delta_\omega) \propto \pi(\omega)$ for all $\delta_\omega \in \mathcal{P}$. In this case, $\Pi = LR \circ t^{-1}$ is the possibility measure on Ω induced by the identification of each Dirac measure $\delta_\omega \in \mathcal{P}$ with the corresponding $\omega \in \Omega$, described by the function $t : \mathcal{P} \to \Omega$ with $t(\delta_\omega) = \omega$ for all $\delta_\omega \in \mathcal{P}$. The fuzzy expectation of a random variable X corresponds then to the possibility measure $\Pi \circ X^{-1}$ on $\mathbb{R}$ induced by $X : \Omega \to \mathbb{R}$; and when an event $A \in \mathcal{A}$ is observed, the hierarchical model is updated according to (1) and (2) to the hierarchical model with levels $\mathcal{P}' = t^{-1}(A)$ and $lik' = lik|_{\mathcal{P}'}$ (the restriction of lik to $\mathcal{P}'$). That is, when A is observed, Π is updated to the normalized possibility measure Π' on Ω with possibility distribution proportional to the pointwise product of π and the indicator function of A.

The hierarchical model offers a unified approach to the combination of probabilistic and possibilistic uncertainty (for instance, fuzzy data can be used without problem). Since membership functions and possibility distributions are interpreted as proportional to likelihood functions, the rules for manipulating fuzzy probabilities are implied by the well-established theories of probability and likelihood. It is important to underline that other interpretations of mem-

bership functions and possibility distributions would lead to other rules for manipulating fuzzy probabilities; in particular, the updating rule would be different. For example, Walley (1997) and De Cooman (2005) interpret possibility measures as upper probability measures: the resulting fuzzy probability model is a special case of the imprecise probability model (at least from the mathematical standpoint); in particular, constant possibility distributions remain constant independently of the data observed (that is, we cannot get out of the state of complete ignorance).

3 Convex Hierarchical Models

Let $\mathcal{M}_0$ be the set of all finite measures μ on the measurable space $(\Omega, \mathcal{A})$, and let $\mathcal{P}_0 \subset \mathcal{M}_0$ be the set of all probability measures P on $(\Omega, \mathcal{A})$. Hence, $\mathcal{M}_0$ and $\mathcal{P}_0$ are subsets of the real vector space of all finite signed measures on $(\Omega, \mathcal{A})$. Let $\mu_0 \in \mathcal{M}_0 \setminus \mathcal{P}_0$ be the measure on $(\Omega, \mathcal{A})$ with $\mu_0(\Omega) = 0$ (that is, μ_0 has constant value 0). The *normalization* function $n : \mathcal{M}_0 \setminus \{\mu_0\} \to \mathcal{P}_0$ is defined by $n(\mu) = [\mu(\Omega)]^{-1}\,\mu$ for all $\mu \in \mathcal{M}_0 \setminus \{\mu_0\}$, where the multiplication of μ with the normalization constant $[\mu(\Omega)]^{-1}$ is to be interpreted pointwise. The restriction $n|_{\mathcal{P}_0}$ of n to $\mathcal{P}_0$ is the identity function on $\mathcal{P}_0$, since $P(\Omega) = 1$ for all $P \in \mathcal{P}_0$. A set $\mathcal{M} \subset \mathcal{M}_0$ is said to be *bounded* if $\sup_{\mu\in\mathcal{M}} \mu(\Omega)$ is finite.

Each bounded set $\mathcal{M} \subset \mathcal{M}_0$ such that $\mathcal{M} \setminus \{\mu_0\}$ is not empty describes a hierarchical model: the probabilistic level

$$\mathcal{P} = \{n(\mu) : \mu \in \mathcal{M} \setminus \{\mu_0\}\}$$

is the image of $\mathcal{M}\setminus\{\mu_0\}$ under n, and the possibilistic level is the likelihood function lik on $\mathcal{P}$ defined (up to a positive multiplicative constant) by

$$lik(P) \propto \sup_{\substack{\mu\in\mathcal{M}\setminus\{\mu_0\}:\\ n(\mu)=P}} \mu(\Omega)$$

for all $P \in \mathcal{P}$. Each hierarchical model can be described in this way by a subset of $\mathcal{M}_0$: for example the hierarchical model with levels $\mathcal{P}$ and lik is described by

$$\mathcal{M} = \{lik(P)\,P : P \in \mathcal{P}\}$$

(where the multiplication of P with the constant $lik(P)$ is to be interpreted pointwise), but such a description is not unique: for instance the sets $\mathcal{M}\cup\{\mu_0\}$ and $\mathcal{M} \setminus \{\mu_0\}$ describe the same hierarchical model. The advantage of the description by a subset of $\mathcal{M}_0$ is that the updating is particularly simple: when an event $A \in \mathcal{A}$ is observed, the set $\mathcal{M}$ is updated to

$$\mathcal{M}' = \{\mu(\cdot \cap A) : \mu \in \mathcal{M}\}. \tag{3}$$

That is, the updated description $\mathcal{M}'$ is the image of $\mathcal{M}$ under r_A, where r_A is the function on $\mathcal{M}_0$ defined by $r_A(\mu) = \mu(\cdot \cap A)$. It can be easily proved that the update of $\mathcal{M}$ according to (3) corresponds to the update of the hierarchical model according to (1) and (2), because if $n(\mu) = P$ and $P(A) > 0$, then $(n \circ r_A)(\mu) = P(\cdot \,|\, A)$. In particular, when applied to the probability measures $P \in \mathcal{P}_0$ with $P(A) > 0$, the function $n \circ r_A$ describes the conditioning on A; hence, the updating (3) of the set $\mathcal{M}$ of measures corresponds to the updating (1) of the imprecise probability model, but without the normalization step (which deletes the information about the relative ability of the probability measures to forecast the observed event A). For the hierarchical model described by $\mathcal{M}$, the uncertain knowledge about the value $g(P)$ of a function $g : \mathcal{P} \to \mathcal{G}$ is described by the normalized fuzzy subset of $\mathcal{G}$ with membership function proportional to the profile likelihood function lik_g on $\mathcal{G}$, which satisfies

$$lik_g(\gamma) \propto \sup_{\substack{\mu\in\mathcal{M}\setminus\{\mu_0\}:\\ (g\circ n)(\mu)=\gamma}} \mu(\Omega)$$

for all $\gamma \in \mathcal{G}$.

The imprecise probability model $\mathcal{P}$ corresponds to the hierarchical model with as probabilistic level the set $\mathcal{P}$, and as possibilistic level a constant likelihood function lik on $\mathcal{P}$; this hierarchical model is described by the set $\mathcal{M} = \mathcal{P} \subset \mathcal{M}_0$. The imprecise probability model $\mathcal{P}$ is often assumed to be convex; it can be easily proved that a set $\mathcal{M}'$ can be obtained by updating a convex set $\mathcal{M} = \mathcal{P}$ according to (3) if and only if $\mathcal{M}'$ is convex. A hierarchical model is said to be *convex* if it can be described by a convex subset of $\mathcal{M}_0$. Hence, the convex hierarchical models are the hierarchical models that can be interpreted as the result of updating (with real or hypothetical data) a convex imprecise probability model; that is, the convex hierarchical models are the direct generalizations of the convex imprecise probability models.

Let $\mathcal{L}_1, \mathcal{L}_2$ be real vector spaces, and let $\mathcal{C} \subseteq \mathcal{L}_1$ be convex. A function $f : \mathcal{C} \to \mathcal{L}_2$ is said to *maintain segments* if for all $x, y \in \mathcal{C}$, the image of the set

$$\{\lambda\, x + (1 - \lambda)\, y : \lambda \in [0, 1]\}$$

under f is the set

$$\{\lambda\, f(x) + (1 - \lambda)\, f(y) : \lambda \in [0, 1]\}\,.$$

The *convex hull* of a set $\mathcal{S} \subseteq \mathcal{L}_1$ is denoted by $\mathrm{ch}(\mathcal{S})$. The following result can be easily proved.

Theorem 2 *Let $\mathcal{L}_1, \mathcal{L}_2$ be real vector spaces, and let $\mathcal{C} \subseteq \mathcal{L}_1$ be convex. If the function $f : \mathcal{C} \to \mathcal{L}_2$ main-*

tains segments, and $\mathcal{S} \subseteq \mathcal{C}$, then the image of the convex hull of $\mathcal{S}$ under f is the convex hull of the image of $\mathcal{S}$ under f; that is,

$$\{f(x) : x \in \mathrm{ch}(\mathcal{S})\} = \mathrm{ch}\left(\{f(y) : y \in \mathcal{S}\}\right).$$

The *convexification* of a hierarchical model described by the set $\mathcal{M} \subset \mathcal{M}_0$ is the convex hierarchical model described by the set $\mathrm{ch}(\mathcal{M}) \subset \mathcal{M}_0$. The function r_A on $\mathcal{M}_0$ maintains segments, since it is the restriction to $\mathcal{M}_0$ of a linear map; hence, Theorem 2 implies that if $\mathcal{M}$ is updated to $\mathcal{M}'$ according to (3), then $\mathrm{ch}(\mathcal{M})$ is updated to $\mathrm{ch}(\mathcal{M}')$ according to (3). This result is particularly useful for updating the convexification of a hierarchical model described by a finite set $\mathcal{M} \subset \mathcal{M}_0$ (such models are very important in the framework of belief networks, studied in Section 4). Since the normalization function n on $\mathcal{M}_0 \setminus \{\mu_0\}$ maintains segments, Theorem 2 can be used to prove also the well-known result that if a set $\mathcal{P}$ of probability measures is updated to $\mathcal{P}'$ according to (1), then $\mathrm{ch}(\mathcal{P})$ is updated to $\mathrm{ch}(\mathcal{P}')$ according to (1).

Let $\rho : [0,\infty] \to [0,\infty]$ be the function defined by $\rho(0) = \infty$, $\rho(\infty) = 0$, and $\rho(x) = \frac{1}{x}$ for all $x \in (0,\infty)$. The function ρ is an involution; that is, $\rho \circ \rho$ is the identity function on $[0,\infty]$. The *convex hull* of a function $\phi : \mathcal{C} \to [0,\infty]$ is denoted by $\mathrm{ch}(\phi)$; that is, $\mathrm{ch}(\phi)$ is the (pointwise) largest convex function $\gamma : \mathcal{C} \to [0,\infty]$ such that $\gamma(x) \leq \phi(x)$ for all $x \in \mathcal{C}$. The following theorem is useful because for example the functions g associating to each probability measure $P \in \mathcal{P}_0$ the expectation $g(P) = E_P(X)$ of a bounded random variable X, or the probability $g(P) = P(B)$ of an event $B \in \mathcal{A}$, are the restrictions to $\mathcal{P}_0$ of linear maps. It is a consequence of Theorem 2, since if $g : \mathcal{P}_0 \to \mathcal{G}$ is the restriction to $\mathcal{P}_0$ of a linear map, then the function $f : \mathcal{M}_0 \setminus \{\mu_0\} \to \mathcal{G} \times \mathbb{R}$ defined by

$$f(\mu) = \left((g \circ n)(\mu), [\mu(\Omega)]^{-1}\right)$$

for all $\mu \in \mathcal{M}_0 \setminus \{\mu_0\}$ maintains segments.

Theorem 3 *Let $\mathcal{G}$ be a real vector space, and let $g : \mathcal{P}_0 \to \mathcal{G}$ be the restriction to $\mathcal{P}_0$ of a linear map. If π and π_{ch} are the membership functions of the normalized fuzzy subsets of $\mathcal{G}$ describing the uncertain knowledge about the value $g(P)$ of g for a hierarchical model and its convexification, respectively, then*

$$\pi_{\mathrm{ch}} = \rho \circ \mathrm{ch}(\rho \circ \pi).$$

Theorem 3 implies in particular that for a convex hierarchical model, the membership function π of the fuzzy expectation of X, or of the fuzzy probability of B, is "reciprocally convex", in the sense that $\rho \circ \pi$ is convex (since $\pi = \pi_{\mathrm{ch}}$). Moreover, Theorem 3 implies that for the convexification of a hierarchical model described by a finite set $\mathcal{M} \subset \mathcal{M}_0$ (such models are very important in the framework of belief networks, studied in Section 4), the membership function π_{ch} of the fuzzy expectation of X, or of the fuzzy probability of B, is *piecewise hyperbolic*, in the sense that $\rho \circ \pi_{\mathrm{ch}}$ is piecewise linear; in this case, the construction of π_{ch} is particularly simple, as shown in the following example.

Example 4 *Consider the situation of Example 1. Conditional on the composition of the urn (that is, conditional on the color of the third ball: white or black), the observations about the colors of the balls drawn are modeled as a sequence of independent Bernoulli trials with constant probability $\frac{1}{3}$ or $\frac{2}{3}$ of observing a black ball, described by the probability measures $P_{\frac{1}{3}}$ and $P_{\frac{2}{3}}$, respectively. The imprecise probability model $\mathcal{P}$ resulting from the vacuous imprecise prior about the composition of the urn (that is, about the color of the third ball) is the convex hull of the finite set $\mathcal{P}_B = \{P_{\frac{1}{3}}, P_{\frac{2}{3}}\}$ of probability measures. The hierarchical model with constant prior likelihood function on $\mathcal{P}$ is described by the set $\mathcal{P} = \mathrm{ch}(\mathcal{P}_B) \subset \mathcal{M}_0$; hence, it is the convexification of the hierarchical model described by the finite set $\mathcal{M} = \mathcal{P}_B \subset \mathcal{M}_0$.*

When the colors of the balls drawn are observed, the updating to $\mathcal{M}'$ according to (3) of the hierarchical model described by the finite set $\mathcal{M} = \mathcal{P}_B$ is very simple. In fact, the updating (1) of the probabilistic level $\mathcal{P}_B = \{P_{\frac{1}{3}}, P_{\frac{2}{3}}\}$ is unimportant for the probability of observing a black ball in the next draw, because the Bernoulli trials are independent under both probability measures $P_{\frac{1}{3}}$ and $P_{\frac{2}{3}}$. The constant prior likelihood function lik on $\mathcal{P}_B$ is updated to lik' according to (2): since $\mathcal{P}_B = \{P_{\frac{1}{3}}, P_{\frac{2}{3}}\}$ has only two elements, lik' is determined (up to a positive multiplicative constant) by the likelihood ratio

$$\frac{lik'(P_{\frac{2}{3}})}{lik'(P_{\frac{1}{3}})} = \frac{(\frac{1}{3})^w\,(\frac{2}{3})^b}{(\frac{2}{3})^w\,(\frac{1}{3})^b} = 2^{b-w}$$

of $P_{\frac{2}{3}}$ and $P_{\frac{1}{3}}$, where w and b are the numbers of white and black balls observed, respectively. Assume that $b \geq w$: the hierarchical model described by the finite set $\mathcal{M}'$ simply tells us that the probability of observing a black ball in the next draw is either $\frac{1}{3}$ or $\frac{2}{3}$, with a likelihood ratio of 2^{b-w} in favor of the second value. This uncertain knowledge is described by the fuzzy probability p of observing a black ball in the next draw, whose membership function π on $[0,1]$

satisfies

$$\pi(p) = \begin{cases} (\frac{1}{2})^{b-w} & \text{if } p = \frac{1}{3}, \\ 0 & \text{if } p \in [0,1] \setminus \{\frac{1}{3}, \frac{2}{3}\}, \\ 1 & \text{if } p = \frac{2}{3}. \end{cases}$$

Theorem 3 allows us to easily obtain the membership function π_{ch} *of the fuzzy probability of observing a black ball in the next draw for the convexification of the hierarchical model described by* $\mathcal{M} = \mathcal{P}_B$*; that is, for the hierarchical model with constant prior likelihood function on* $\mathcal{P}$*, which was considered in Example 1. Since the function* $\rho \circ \pi$ *on* $[0,1]$ *satisfies*

$$(\rho \circ \pi)(p) = \begin{cases} 2^{b-w} & \text{if } p = \frac{1}{3}, \\ \infty & \text{if } p \in [0,1] \setminus \{\frac{1}{3}, \frac{2}{3}\}, \\ 1 & \text{if } p = \frac{2}{3}, \end{cases}$$

its convex hull $\mathrm{ch}(\rho \circ \pi)$ *is the piecewise linear function on* $[0,1]$*, whose values in* $(\frac{1}{3}, \frac{2}{3})$ *are obtained by linear interpolation of the values of* $\rho \circ \pi$ *in* $\frac{1}{3}$ *and* $\frac{2}{3}$*; that is,*

$$(\mathrm{ch}(\rho \circ \pi))(p) = \begin{cases} 2^{b-w} - 3\,(2^{b-w}-1)\,(p-\frac{1}{3}) & \text{if } p \in [\frac{1}{3}, \frac{2}{3}], \\ \infty & \text{if } p \in [0,1] \setminus [\frac{1}{3}, \frac{2}{3}]. \end{cases}$$

Hence, for the hierarchical model with constant prior likelihood function on $\mathcal{P}$ *(which was considered in Example 1), the fuzzy probability* p *of observing a black ball in the next draw, after having observed* w *white balls and* b *black balls (with* $b \geq w$*), has membership function* π_{ch} *on* $[0,1]$ *satisfying*

$$\pi_{\mathrm{ch}}(p) = \begin{cases} [2^{b-w} - 3\,(2^{b-w}-1)\,(p-\frac{1}{3})]^{-1} & \text{if } p \in [\frac{1}{3}, \frac{2}{3}], \\ 0 & \text{if } p \in [0,1] \setminus [\frac{1}{3}, \frac{2}{3}]. \end{cases}$$

Figure 1 shows the graphs of the piecewise hyperbolic function π_{ch} *when* $b - w = 0$ *(dotted line),* $b - w = 3$ *(dashed line), and* $b - w = 7$ *(solid line).*

4 Hierarchical Networks

Let $X_1, \ldots, X_k$ be some variables taking value in the finite sets $\mathcal{X}_1, \ldots, \mathcal{X}_k$, respectively. An elegant and useful way of constructing a probability measure P on $\Omega = \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$ (that is, a purely probabilistic description of uncertain knowledge about the values of the variables $X_1, \ldots, X_k$) is through a *Bayesian network* (see for example Pearl, 1988, or Jensen, 2001). This consists of a directed acyclic graph with nodes $X_1, \ldots, X_k$, such that to each node X_i is associated a stochastic kernel P_i from $\mathcal{PA}_i$ to $\mathcal{X}_i$, where $\mathcal{PA}_i$ is the image of Ω under PA_i, and PA_i is the function on Ω assigning to each $\omega = (x_1, \ldots, x_k) \in \Omega$ the vector $(x_{j_1}, \ldots, x_{j_l})$ of the values of the *parents* $X_{j_1}, \ldots, X_{j_l}$ of X_i (that is, the nodes from which start the edges pointing to X_i). The stochastic kernel P_i associates to each vector $pa_i \in \mathcal{PA}_i$ a probability measure $P_i(\cdot \,|\, pa_i)$ on $\mathcal{X}_i$; in particular, if X_i is a *root* (that is, it has no parents), then PA_i assigns the "empty vector" $()$ to all $\omega \in \Omega$, and therefore $\mathcal{PA}_i = \{()\}$ is a singleton and the stochastic kernel P_i reduces to a probability measure $P_i(\cdot \,|\, ())$ on $\mathcal{X}_i$. The probability measure $P_{P_1,\ldots,P_k}$ on Ω associated to the Bayesian network is defined by

$$P_{P_1,\ldots,P_k}\{\omega\} = \prod_{i=1}^{k} P_i(\{x_i\} \,|\, PA_i(\omega))$$

for all $\omega = (x_1, \ldots, x_k) \in \Omega$. A key property of Bayesian networks is that the graph encodes conditional independences between the variables $X_1, \ldots, X_k$: these conditional independences can be determined by the graphical criterion of *d-separation*.

Bayesian networks can be generalized to *credal networks* by associating to each node X_i a set $\mathcal{P}_i$ of stochastic kernels P_i from $\mathcal{PA}_i$ to $\mathcal{X}_i$, instead of a single stochastic kernel (see for example Cozman, 2005, or Antonucci and Zaffalon, 2008). The set $\mathcal{P}_i$ associated to a node X_i is said to be *separately specified* if for each $pa_i \in \mathcal{PA}_i$ we can specify a set $\mathcal{P}_{i,pa_i}$ of probability measures on $\mathcal{X}_i$, and obtain $\mathcal{P}_i$ as the set of all stochastic kernels P_i from $\mathcal{PA}_i$ to $\mathcal{X}_i$ such that $P_i(\cdot \,|\, pa_i) \in \mathcal{P}_{i,pa_i}$ for each $pa_i \in \mathcal{PA}_i$ (that is, $\mathcal{P}_i$ can be identified with the Cartesian product of the sets $\mathcal{P}_{i,pa_i}$). The imprecise probability model usually associated to the credal network (called *strong extension* of the credal network) is the convex hull of the set

$$\mathcal{P}_{\mathcal{P}_1,\ldots,\mathcal{P}_k} = \{P_{P_1,\ldots,P_k} : P_1 \in \mathcal{P}_1, \ldots, P_k \in \mathcal{P}_k\}.$$

In practical applications of credal networks the sets $\mathcal{P}_i$ of stochastic kernels are often finite, and thus the set $\mathcal{P}_{\mathcal{P}_1,\ldots,\mathcal{P}_k}$ of probability measures is finite too.

Credal networks can be generalized to *hierarchical networks* by associating to each node X_i also a (prior) likelihood function lik_i on the set $\mathcal{P}_i$ of stochastic kernels associated to X_i. When the set $\mathcal{P}_i$ associated to a node X_i is separately specified by the sets $\mathcal{P}_{i,pa_i}$ of probability measures on $\mathcal{X}_i$ (where $pa_i \in \mathcal{PA}_i$), the likelihood function lik_i on $\mathcal{P}_i$ associated to X_i is said to be *separately specified* if for each $pa_i \in \mathcal{PA}_i$ we can specify a likelihood function lik_{i,pa_i} on $\mathcal{P}_{i,pa_i}$, and obtain lik_i as the function on $\mathcal{P}_i$ defined by

$$lik_i(P_i) = \prod_{pa_i \in \mathcal{PA}_i} lik_{i,pa_i}(P_i(\cdot \,|\, pa_i))$$

for all $P_i \in \mathcal{P}_i$ (that is, lik_i can be interpreted as the independent combination of the marginals lik_{i,pa_i}). A node X_i is said to be *Bayesian* if the set $\mathcal{P}_i$ of stochastic kernels associated to X_i is a singleton;

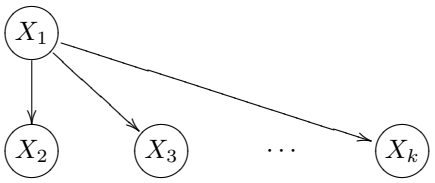


Figure 2: Directed acyclic graph of the hierarchical network of Example 5.

that is, the uncertain knowledge about the value of a Bayesian node conditional on the values of its parents is purely probabilistic. A node X_i is said to be *fuzzy* if $P_i(\cdot \,|\, pa_i)$ is a Dirac measure on $\mathcal{X}_i$ for all $pa_i \in \mathcal{PA}_i$ and all stochastic kernels P_i in the set $\mathcal{P}_i$ associated to X_i; that is, the uncertain knowledge about the value of a fuzzy node conditional on the values of its parents is purely possibilistic. The hierarchical model associated to the hierarchical network has as probabilistic level the set $\mathcal{P}_{\mathcal{P}_1,\ldots,\mathcal{P}_k}$, and as possibilistic level the likelihood function lik on $\mathcal{P}_{\mathcal{P}_1,\ldots,\mathcal{P}_k}$ defined (up to a positive multiplicative constant) by

$$lik(P) \propto \sup_{\substack{P_1 \in \mathcal{P}_1,\ldots,P_k \in \mathcal{P}_k : \\ P_{P_1,\ldots,P_k} = P}} \prod_{i=1}^{k} lik_i(P_i)$$

for all $P \in \mathcal{P}_{\mathcal{P}_1,\ldots,\mathcal{P}_k}$. Hence, the hierarchical model associated to the hierarchical network is described by the set $\mathcal{M} \subset \mathcal{M}_0$ consisting of all measures $\mu_{P_1,\ldots,P_k}$ on Ω with $P_1 \in \mathcal{P}_1, \ldots, P_k \in \mathcal{P}_k$, where

$$\mu_{P_1,\ldots,P_k}\{\omega\} = \prod_{i=1}^{k} [lik_i(P_i)\, P_i(\{x_i\} \,|\, PA_i(\omega))]$$

for all $\omega = (x_1, \ldots, x_k) \in \Omega$. If only convexifications of hierarchical models are considered, then credal networks correspond to the hierarchical networks with constant likelihood functions lik_i, and it often suffices to use finite sets $\mathcal{P}_i$ of stochastic kernels, so that the set $\mathcal{M}$ of measures is finite and the results of Section 3 can be exploited, as in the following examples.

Example 5 *Consider a hierarchical network about the value of the binary variables $X_1, \ldots, X_k \in \{0,1\}$. The directed acyclic graph is plotted in Figure 2. The root X_1 is Bayesian with uniform probability; that is, $\mathcal{P}_1 = \{P_1\}$ with $P_1(\{0\} \,|\, ()) = P_1(\{1\} \,|\, ()) = \frac{1}{2}$. For each $i \geq 2$ the set $\mathcal{P}_i$ associated to the node X_i consists of all stochastic kernels P_i from $\mathcal{PA}_i = \{0,1\}$ to $\mathcal{X}_i = \{0,1\}$ such that $P_i(\{x\} \,|\, (x)) \geq 0.9$ for both $x \in \{0,1\}$. All (prior) likelihood functions lik_i on the sets $\mathcal{P}_i$ are constant. Hence, the hierarchical network corresponds to a credal network with separately specified sets $\mathcal{P}_i$. It can be interpreted as follows: X_1 is the unobservable variable of interest, and for each $i \geq 2$ the variable X_i describes the observation returned by a sensor with a probability of being correct of at least 90%. We want to describe the uncertain knowledge about the value of X_1 that we gain from the observations returned by the $k-1$ sensors, which are assumed to be independent conditional on X_1.*

The case with $k = 3$ (interpreted as a credal network) was studied by Antonucci et al. (2007, Example 1): they showed that if the observations x_2, x_3 returned by the two sensors are equal, then the posterior imprecise probability that X_1 has value $x_2 = x_3$ is $[0.988, 1]$, while if the observations x_2, x_3 are different, then the posterior imprecise probability about the value of X_1 is vacuous. This can be reasonable, but the problem is that the model behaves in the same way in the cases with $k > 3$: it suffices that one of the observations $x_2, \ldots, x_k$ returned by the $k-1$ sensors is different from the others, in order for the posterior imprecise probability about the value of X_1 to be vacuous, independently of the number of sensors. The reason is that for each $i \geq 2$ it is considered possible that the sensor returning the observation X_i is perfect (that is, always correct) while all others are not (that is, they can be wrong), and in this case the posterior probability that X_1 has value x_i is 1, even when all observations returned by the other sensors are different from x_i. However, even if the sensor returning the observation X_i is always correct while all others can be wrong, it is extremely improbable that all others are wrong at the same time. Hence, when the observation returned by a sensor is different from all others, it is extremely implausible that this sensor is perfect. This information about plausibility is described by the likelihood function, and in fact the problem disappears when the network is interpreted as a hierarchical network instead of a credal network.

The convexification of the hierarchical model associated to the hierarchical network can be easily updated thanks to the results of Section 3: for instance, in the case with $k = 5$, when 3 of the observations $x_2, \ldots, x_5$ returned by the 4 sensors are equal x and one is different from x, the membership function of the posterior fuzzy probability p that X_1 has value x is plotted in Figure 3; in particular, the α-cut with $\alpha = 0.1465$ is the interval $[0.932, 1]$. As expected, this fuzzy probability is very high, although no probability value in the interval $[0,1]$ is completely excluded.

To solve the above problem in the framework of credal networks, we should exclude the possibility of perfect sensors by bounding from above the probability that sensors are correct. That is, we should choose a small $\varepsilon > 0$, and for each $i \geq 2$ replace the set $\mathcal{P}_i$ by the set of all stochastic kernels P_i from $\mathcal{PA}_i = \{0,1\}$ to $\mathcal{X}_i = \{0,1\}$ such that $P_i(\{x\} \,|\, (x)) \in [0.9, 1-\varepsilon]$

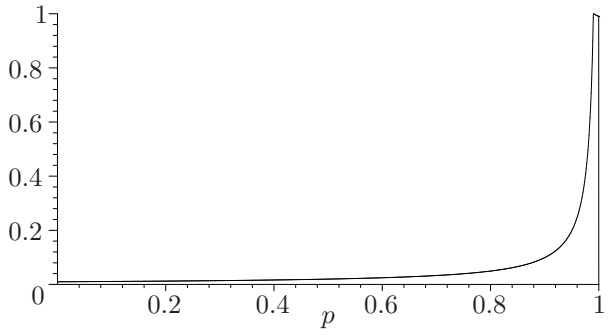


Figure 3: Membership function of the posterior fuzzy probability p that X_1 has value x, when 3 of the observations $x_2, \dots, x_5$ returned by the 4 sensors are equal x and one is different from x (for the hierarchical network of Example 5 with $k = 5$).

for both $x \in \{0, 1\}$. *However, the resulting posterior imprecise probabilities can depend strongly on the choice of* ε*: for instance, in the case with* $k = 5$*, when* 3 *of the observations* $x_2, \dots, x_5$ *returned by the* 4 *sensors are equal* x *and one is different from* x*, the posterior imprecise probability that* X_1 *has value* x *is* $[0.422, 1.000]$ *when* $\varepsilon = 0.001$*,* $[0.786, 1.000]$ *when* $\varepsilon = 0.005$*, and* $[0.880, 1.000]$ *when* $\varepsilon = 0.01$*. By contrast, in these cases the membership functions of the posterior fuzzy probability that* X_1 *has value* x *are (almost) equal to the pointwise product of the indicator function of the corresponding posterior imprecise probability and the membership function for the case with* $\varepsilon = 0$ *(plotted in Figure 3). Hence, this fuzzy probability does not change much when* ε *is varied, since only rather implausible probability values are excluded; in particular, the* α*-cuts with* $\alpha = 0.1465$ *for the cases with* $\varepsilon = 0.001$*,* $\varepsilon = 0.005$*, or* $\varepsilon = 0.01$ *are practically equal to the* α*-cut* $[0.932, 1]$ *for the case with* $\varepsilon = 0$.

The possibilistic level of the hierarchical model associated to the hierarchical network of Example 5 contains no information before the updating, because the (prior) likelihood functions lik_i on the sets $\mathcal{P}_i$ of stochastic kernels associated to the nodes X_i are constant. But also hierarchical networks such that the possibilistic levels of the associated hierarchical models contain some prior information (that is, some of the likelihood functions lik_i are not constant) can be useful. In particular, when the stochastic kernels of the network are learned from training data, it is not necessary to reduce the likelihood function to the maximum likelihood estimates (and thus discard the information about the uncertainty of these estimates): the whole likelihood function induced by the training data can be maintained as the possibilistic level of the hierarchical model associated to the hierarchical network. This is a very interesting topic, but goes beyond the scope of the present paper.

Another useful application of hierarchical networks with nonconstant (prior) likelihood functions lik_i is the contamination of a Bayesian (or credal) network: for each node X_i we can give high relative plausibility to the original stochastic kernels P_i associated to X_i, and low relative plausibility to all (or a subset of) other stochastic kernels P_i from $\mathcal{PA}_i$ to $\mathcal{X}_i$. A similar contamination would be possible also in the framework of credal networks (by considering neighborhoods of the original stochastic kernels), but we could not include all possible stochastic kernels (since otherwise the resulting imprecise probability model would be useless), and the final considerations of Example 5 suggest that the resulting posterior imprecise probabilities would be much more sensitive than the posterior fuzzy probabilities to the exact choice of the contamination. In a certain sense, in the framework of hierarchical networks the contamination can be at the possibilistic level, while in the framework of credal networks it must be at the probabilistic level, and this can lead to instability.

Example 6 *Consider the Bayesian network obtained from the hierarchical network of Example 5 by selecting, for each* $i \geq 2$*, the stochastic kernel* P_i *from* $\mathcal{PA}_i = \{0, 1\}$ *to* $\mathcal{X}_i = \{0, 1\}$ *such that* $P_i(\{x\} \,|\, (x)) = 0.95$ *for both* $x \in \{0, 1\}$*. We can contaminate this Bayesian network by choosing a small* $\gamma > 0$ *and associating to each node* X_i *the (separately specified) set* $\mathcal{P}_i$ *of all stochastic kernels* P_i *from* $\mathcal{PA}_i = \{0, 1\}$ *to* $\mathcal{X}_i = \{0, 1\}$ *and the (prior) likelihood function* lik_i *on* $\mathcal{P}_i$ *separately specified by the likelihood functions* $lik_{i,(x)}$ *on the set of all probability measures on* $\{0, 1\}$ *such that* $lik_{i,(x)}(P_i(\cdot \,|\, (x))) = 1$ *if* $P_i(\cdot \,|\, (x))$ *is the corresponding conditional probability in the Bayesian network, and* $lik_{i,(x)}(P_i(\cdot \,|\, (x))) = \gamma$ *otherwise, for both* $x \in \{0, 1\}$*. The resulting hierarchical network describes the situation in which there is some uncertainty about the conditional probabilities of the Bayesian network; it is useful because it tells us how robust against modifications of the conditional probabilities are the conclusions of the Bayesian network.*

The convexification of the hierarchical model associated to the hierarchical network can be easily updated thanks to the results of Section 3: for instance, Figure 4 shows the graphs of the membership functions of the fuzzy probability p *of* $X_1 = 1$ *in the case with* $k = 3$ *and* $\gamma = 0.05$*: prior to any observation (dashed line), after observing* $X_2 = X_3 = 0$ *(solid line with maximum near* 0*), after observing* $X_2 = 1$ *and* $X_3 = 0$ *or vice versa (dotted line), and after observing* $X_2 = X_3 = 1$ *(solid line with maximum near* 1*); in particular, the corresponding* α*-*

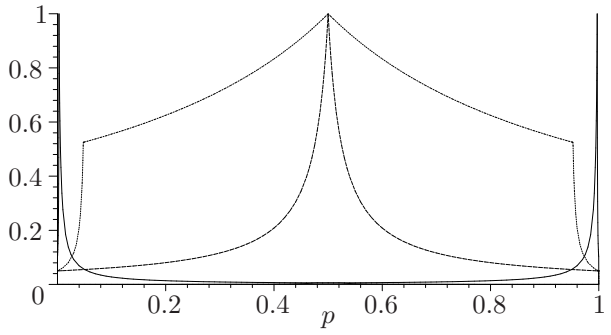


Figure 4: Membership functions of the fuzzy probability p of $X_1 = 1$ (for the hierarchical network of Example 6 with $k = 3$ and $\gamma = 0.05$): prior to any observation (dashed line), after observing $X_2 = X_3 = 0$ (solid line with maximum near 0), after observing $X_2 = 1$ and $X_3 = 0$ or vice versa (dotted line), and after observing $X_2 = X_3 = 1$ (solid line with maximum near 1).

cuts with $\alpha = 0.1465$ *are the intervals* $[0.347, 0.653]$, $[0.001, 0.019]$, $[0.035, 0.965]$, *and* $[0.981, 0.999]$, *respectively. Hence, the conclusions of the Bayesian network are pretty robust when the two sensors agree (the uncertainty about the probability of* $X_1 = 1$ *decreases), while they are not robust at all when the two sensors disagree (the uncertainty about the probability of* $X_1 = 1$ *increases).*

When $X, Y, Z \subseteq \{X_1, \ldots, X_k\}$ are three disjoint sets of variables, Y is said to be *irrelevant* to X given Z (with respect to a hierarchical model on Ω) if the fuzzy probability distribution for the variables in X conditional on any realization of the variables in Z does not change when also something about the variables in Y is observed. This definition of conditional irrelevance is stronger than the corresponding one for imprecise probability models, since the invariance of both levels of the hierarchical model is required. However, when the hierarchical model is constructed through a hierarchical network, the following fundamental result holds (for a sketch of the proof see Cattaneo, 2008*b*, Subsection 3.1).

Theorem 7 *Let* $X, Y, Z \subseteq \{X_1, \ldots, X_k\}$ *be three disjoint sets of variables. If* X *and* Y *are d-separated by* Z *in the directed acyclic graph of a hierarchical network, then* Y *is irrelevant to* X *given* Z*, with respect to the hierarchical model associated to the hierarchical network.*

Theorem 7 is of crucial importance for the meaning and usefulness of hierarchical networks: conditional irrelevances between the variables $X_1, \ldots, X_k$ are encoded in the graph and can be determined by the graphical criterion of d-separation. Together with the results of Section 3, Theorem 7 allows the calculation of exact inferences in simple hierarchical networks.

Any probability measure on Ω can be constructed through a Bayesian network with nodes $X_1, \ldots, X_k$. By contrast, not all hierarchical models on Ω can be constructed through hierarchical networks with nodes $X_1, \ldots, X_k$. However, any hierarchical model describing the uncertain knowledge about the values of the variables $X_1, \ldots, X_k$ can be constructed through a hierarchical network with nodes $X_1, \ldots, X_{k+1}$: it suffices to add a root X_{k+1}, which in general is a parent of all other nodes, and which indexes the probability measures in the probabilistic level $\mathcal{P}$ of the hierarchical model. Hence, the variable X_{k+1} takes values in the set $\mathcal{P}$, which can be infinite, but this is unimportant, since the root X_{k+1} is fuzzy (with likelihood function lik_{k+1} corresponding to the possibilistic level lik of the hierarchical model); by contrast, the nodes $X_1, \ldots, X_k$ are Bayesian.

More generally, we can easily transform any hierarchical network with nodes $X_1, \ldots, X_k$ into a larger hierarchical network which describes the same uncertain knowledge about the values of the variables $X_1, \ldots, X_k$, but such that each node is either Bayesian or fuzzy (and we can also require that only roots can be fuzzy). In fact, when a node X_i is neither Bayesian nor fuzzy (or it is fuzzy but not a root), we can simply add a root which is a parent of X_i only, and which indexes the set $\mathcal{P}_i$ of stochastic kernels associated to X_i. This additional root is fuzzy (with likelihood function corresponding to the likelihood function lik_i on $\mathcal{P}_i$ associated to X_i), while the node X_i becomes Bayesian. In particular, we can always obtain a hierarchical network such that each node X_i is either Bayesian or fuzzy and both the set $\mathcal{P}_i$ of stochastic kernels and the likelihood function lik_i on $\mathcal{P}_i$ associated to X_i are separately specified (since this is always the case for roots and Bayesian nodes).

From the above considerations it follows easily the result (showed by Antonucci and Zaffalon, 2008) that we can transform any credal network with nodes $X_1, \ldots, X_k$ into a larger credal network which describes the same uncertain knowledge about the values of the variables $X_1, \ldots, X_k$, but such that each node X_i is either Bayesian or the set $\mathcal{P}_i$ of stochastic kernels associated to X_i is separately specified by vacuous imprecise probability models. More specifically, we can always obtain a credal network such that each node X_i is either Bayesian or it is a root and the set $\mathcal{P}_i$ of probability measures on $\mathcal{X}_i$ is the vacuous imprecise probability model. The difference between the hierarchical model and the imprecise probability model is in the way in which such roots X_i are updated when data

are observed (since the Bayesian nodes are updated in the same way in both models): in the framework of credal networks we remain in the state of complete ignorance about the value of X_i (apart from when we get deterministic information about it), while in the framework of hierarchical networks the possibilistic level allows us to get out of the state of complete ignorance about the value of X_i.

This shows in particular that hierarchical networks cannot be described by possibly larger credal networks (for instance by interpreting possibility measures as upper probability measures), because these could not display the same behavior when data are observed, not even with an alternative updating rule.

5 Conclusion

In the present paper, the use of fuzzy probabilities to describe the uncertain knowledge about the values of the nodes of belief networks has been studied. The increased expressive power, the ability of using all the information provided by the data, and the increased robustness of the conclusions are important advantages over credal networks. The possibility of using the whole likelihood function induced by training data (and not only the maximum likelihood estimates) seems very promising and deserves further study. The description of convex hierarchical models by finite sets of measures and the validity of the criterion of d-separation allow the calculation of the desired inferences in simple hierarchical networks. However, approximation algorithms are necessary for the calculation of inferences in more complex networks: some algorithm for credal networks can probably be adapted to hierarchical networks, thanks to the strong similarity between the descriptions of the hierarchical model and of the imprecise probability model as convex sets of measures.

References

Antonucci, A., Brühlmann, R., Piatti, A., and Zaffalon, M. (2007). Credal networks for military identification problems. In *ISIPTA '07*. SIPTA, 1–10.

Antonucci, A., and Zaffalon, M. (2008). Decision-theoretic specification of credal networks: A unified language for uncertain modeling with sets of Bayesian networks. *Int. J. Approx. Reasoning* 49, 345 – 361.

Cattaneo, M. (2005). Likelihood-based statistical decisions. In *ISIPTA '05*. SIPTA, 107–116.

Cattaneo, M. (2007). *Statistical Decisions Based Directly on the Likelihood Function*. PhD thesis, ETH Zurich.

Cattaneo, M. (2008*a*). Fuzzy probabilities based on the likelihood function. In *Soft Methods for Handling Variability and Imprecision*. Springer, 43–50.

Cattaneo, M. (2008*b*). Probabilistic-possibilistic belief networks. Technical Report 32. Department of Statistics, LMU Munich.

Cozman, F. G. (2005). Graphical models for imprecise probabilities. *Int. J. Approx. Reasoning* 39, 167–184.

Dahl, F. A. (2005). Representing human uncertainty by subjective likelihood estimates. *Int. J. Approx. Reasoning* 39, 85–95.

De Cooman, G. (2005). A behavioural model for vague probability assessments. *Fuzzy Sets Syst.* 154, 305–358.

Dubois, D. (2006). Possibility theory and statistical reasoning. *Comput. Stat. Data Anal.* 51, 47–69.

Held, H., Augustin, T., and Kriegler, E. (2008). Bayesian learning for a class of priors with prescribed marginals. *Int. J. Approx. Reasoning* 49, 212 – 233.

Hisdal, E. (1988). Are grades of membership probabilities? *Fuzzy Sets Syst.* 25, 325–348.

Jensen, F. V. (2001). *Bayesian Networks and Decision Graphs*. Springer.

Moral, S. (1992). Calculating uncertainty intervals from conditional convex sets of probabilities. In *UAI '92*. Morgan Kaufmann, 199–206.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

Pearl, J. (1988). *Probabilistic Inference in Intelligent Systems*. Morgan Kaufmann.

Walley, P. (1997). Statistical inferences based on a second-order possibility distribution. *Int. J. Gen. Syst.* 26, 337–383.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* 9, 60–62.

Wilson, N. (2001). Modified upper and lower probabilities based on imprecise likelihoods. In *ISIPTA '01*. Shaker, 370–378.

Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets Syst.* 1, 3–28.

A tree augmented classifier based on Extreme Imprecise Dirichlet Model

Giorgio Corani
IDSIA
Manno, Switzerland
giorgio@idsia.ch

Cassio P. De Campos
IDSIA
Manno, Switzerland
cassio@idsia.ch

Sun Yi
IDSIA
Manno, Switzerland
yi@idsia.ch

Abstract

In this paper we present TANC, i.e., a tree-augmented naive credal classifier based on imprecise probabilities; it models prior near-ignorance via the Extreme Imprecise Dirichlet Model (EDM) [1] and deals conservatively with missing data in the training set, without assuming them to be missing-at-random. The EDM is an approximation of the global Imprecise Dirichlet Model (IDM), which considerably simplifies the computation of upper and lower probabilities; yet, having been only recently introduced, the quality of the provided approximation needs still to be verified. As first contribution, we extensively compare the output of the naive credal classifier (one of the few cases in which the global IDM can be exactly implemented) when learned with the EDM and the global IDM; the output of the classifier appears to be identical in the vast majority of cases, thus supporting the adoption of the EDM in real classification problems. Then, by experiments we show that TANC is more reliable than the precise TAN (learned with uniform prior), and also that it provides better performance compared to a previous [13] TAN model based on imprecise probabilities. TANC treats missing data by considering all possible completions of the training set, but avoiding an exponential increase of the computational times; eventually, we present some preliminary results with missing data.

Keywords. Imprecise Dirichlet Model, Extreme Imprecise Dirichlet Model, Classification, TANC, Naive Credal Classifier.

1 Introduction

Classifiers based on imprecise probabilities are progressively becoming known and appreciated also outside the area of imprecise probabilities [2]; typically, they are based on the Imprecise Dirichlet Model (IDM) to model a condition of prior *near-ignorance*. When faced with an instance whose classification is prior-dependent, they preserve reliability by returning a set of classes (*indeterminate classifications*) instead of a single class. Thanks to the IDM, credal classifiers robustly deal with cases where the evidence arising from the data is not strong enough to smooth the effect of the prior choice.

Two IDM variants have been adopted in credal classifiers: the global IDM or the local IDM; the local lacks some constraints present in the global. The global IDM can make it very difficult to solve the optimization problem to determine lower and upper probabilities. So far, the naive credal classifier (NCC) of [10] is the only case in which it has been possible to develop a credal classifier based on the global IDM. On the contrary, the local IDM allows for an easier solution of the optimization problem; yet, it can return probability intervals that can be unnecessarily wide, compared to the global IDM.

Recently, the EDM (Extreme Dirichlet Model) [1] has been introduced; it restricts the credal set of the global IDM only to its extreme distributions. The intervals returned by the EDM are hence included in the intervals returned by the global IDM; however, the EDM can considerably simplify the solution of the optimization problem. So far, the EDM has been used only in very preliminary experiments; as recognized also in [1], it is still necessary to test the EDM in real classification problems and to study the difference with the global IDM. A first contribution of this paper is that we have implemented NCC with EDM and we have compared it (using 23 data sets) against NCC with global IDM; results show that the two models returns the same set of classes in the large majority of cases.

However, besides prior-ignorance, there is another kind of ignorance involved in the process of learning from data, i.e., ignorance about the missingness process. Usually, classifiers ignore missing data; this entails the idea that the missingness process (MP) is non-selective in producing missing data, i.e., it is MAR (*missing at random* [6]). However, assuming MAR cannot be regarded as an objective-minded approach, if one is ignorant about the MP. According to the Conservative Updating Rule [11, 12], in order to deal conservatively with nonMAR[1]

[1] The term nonMAR is used to indicate that MAR is not assumed.

missing data in the training set, it is necessary to compute a likelihood for each possible completion of the data set. The naive credal classifier of [10] implements such an approach for data that are missing in the training set.

However, naive classifiers can become inadequate on certain data sets, as they assume the statistical independence of the features given the class. Tree augmented naive classifiers [5] have been shown to often outperform naive Bayes, as they can model more realistically complex data sets. An attempt to extend TAN to imprecise probabilities has been proposed in [13]; in the following, this algorithm is referred to as TANC*. TANC* is based on the local IDM, to keep the computation affordable; yet, this choice is likely to make TANC* much more indeterminate than if the global IDM was used. In fact, TANC* returns a considerable number of indeterminate classifications [13]. A further characteristic of TANC* is that it assumes missing data to be MAR, which also contributes for its efficiency.

In this paper we present TANC, i.e., a tree-augmented naive credal classifier based on imprecise probabilities, which (a) models prior near-ignorance via the EDM and (b) treats missing data in the training set[2] without assuming MAR, thus computing a set of likelihoods. Although the number of possible likelihoods is in principle exponential with respect to the number of missing values, we show that the computational complexity of TANC does not necessarily increase exponentially with the total number of missing data in the training set.

We thoroughly evaluate TANC by experiments. Firstly, we evaluate TANC against the precise TAN (i.e., learned with uniform prior) on several data sets; we show that TANC is effective at detecting hard-to-classify instances, over which TAN becomes unreliable; instead, TANC preserve its reliability thanks to indeterminate classifications. In a second series of experiments, we compare TANC and TANC*; we show that TANC is less indeterminate than TANC*; the results suggest moreover that TANC returns determinate and correct answers on instances over which TANC* is unnecessarily indeterminate. Since the difference between TAN, TANC and TANC* lies in the model of prior ignorance, the differences between them decreases with the size of the data set: large amount of data reduce the role of prior densities.

Eventually, we present some preliminary results with non-MAR missing data, comparing TANC against the naive credal classifier (which is also able to treat missing data as nonMAR). Under this setting, TANC appears to be much more indeterminate than the naive credal classifier, because of the more complex graph.

The paper is structured as follows: Section 2 introduces the notation and the basic definitions; Section 3 describes the Imprecise Dirichlet Model in its local, global and extreme specifications; in Section 5 we experimentally show that using the naive credal classifier with global IDM or with the EDM leads to equivalent classifications in most cases. Section 6 presents the TANC algorithm and proves its correctness; Section 7 shows the experimental results, including the comparison against TAN, TANC* and some preliminary results with missing data. Finally, Section 8 contains the conclusions.

[2]The extension to nonMAR missing data in the testing set is left for future development.

2 Notation and Basic Definitions

This section presents the notation used later in the paper, the definition of a credal network and the specification of the data that is employed for learning the parameters of the network. To simplify, we use a definition of credal network were the factorization is enforced in a set of joint probability distributions.

Definition 1 *A credal network is a triple $(\mathcal{G}, \mathcal{X}, \mathcal{K})$, where $\mathcal{G}$ is a directed acyclic graph with nodes associated to discrete random variables $\mathcal{X} = \{X_1, \ldots, X_m\}$ and $\mathcal{K}$ is a set of multinomial probability distributions on $\mathcal{X}$ such that each $p \in \mathcal{K}$ factorizes as $p(\mathcal{X}) = \prod_i p(X_i|\Pi_i)$ (which can be read as* every variable is conditionally independent of its non-descendants given its parents*), where Π_i denotes the parents of X_i in $\mathcal{G}$ (when $\Pi_i = \emptyset$, $p(X_i|\Pi_i)$ is in fact the marginal $p(X_i)$).*

The state space of a variable X_i is denoted by Ω_{X_i}, and the joint space on a set of variables $\mathcal{Y}$ by $\Omega_{\mathcal{Y}} = \times_{X \in \mathcal{Y}} \Omega_X$. Lowercase letters are used to specify assignments to variables: $x_i \in \Omega_{X_i}$ is a category of X_i, and $\pi_i \in \Omega_{\Pi_i}$ is an assignment to all parents of X_i. Parents and children of variables are denoted always with respect to the graph $\mathcal{G}$ of the network. A variable X_i with $\Pi_i = \emptyset$ is called a *root* variable. We further denote by Λ_i the set of children of X_i.

We assume the training data set D to contain n instances of type $\mathbf{x} = \{x_1, \ldots, x_m\}$. With reference to the subset of variables $\mathcal{Y} \subseteq \mathcal{X}$, we define $n_{\mathbf{y}}$ as the number of instances for which the set of variables $\mathcal{Y}$ is set to $\mathbf{y}$.

We allow the training data set to contain missing values, that is, for each instance $\mathbf{x}$ some of its elements may be absent. A *completion* of $\mathbf{x}$ is an assignment to the missing values such that $\mathbf{x}$ becomes complete. A completion of the data set is a completion for all its instances. We denote by $\mathbf{d}_{\mathcal{Y}}$ a possible realization of the training data set (i.e., the observed values plus a possible realization for missing data, if any) restricted to the variables $\mathcal{Y} \subseteq \mathcal{X}$.

3 Variants of the Imprecise Dirichlet Model

The *Imprecise Dirichlet Model*(IDM) [8] is a tool for inference from categorical data, based on a *set* of prior Dirichlet densities. In the following, we illustrate the different variants of the IDM, considering as an example the simple credal network $X_1 \rightarrow X_2$.

As for the marginal distribution $p(X_1)$, the Dirichlet density is proportional to $\prod_{x_1 \in \Omega_{X_1}} \theta_{x_1}^{\alpha_{x_1}-1}$, where $\alpha_{x_1} > 0$ and $\sum_{x_1 \in \Omega_{X_1}} \alpha_{x_1} = s$, where s represents the *equivalent sample size* (or *hidden instances*), which determines the weight of the prior compared to the total number of instances in the training set. By letting the hyper-parameters α_{x_1} take all the possible values in their domain of definition, the IDM produces an interval posterior estimate of the chance, which for each $x_1 \in \Omega_{X_1}$ is:

$$\left[\frac{n_{x_1}}{n+s}, \frac{n_{x_1}+s}{n+s}\right]. \tag{1}$$

When we move to the estimation of $p(x_2|x_1)$, the IDM can be applied locally or globally. By the *local* IDM, we repeat the estimation of formula (1), thus obtaining:

$$\left[\frac{n_{x_1x_2}}{n_{x_1}+s}, \frac{n_{x_1x_2}+s}{n_{x_1}+s}\right]. \tag{2}$$

In other terms, the hyper-parameters $\alpha_{x_1x_2}$ can vary between 0 and s ($0 < \alpha_{x_1x_2} < s$). In this way, we obtain a local credal set for each variable and each assignment of its parents; the global credal set is eventually obtained by the multiplication of the local credal sets.

Alternatively, one can use the *global* IDM; in this case the hyper-parameter $\alpha_{x_1x_2}$ is constrained by $\sum_{x_2} \alpha_{x_1x_2} = \alpha_{x_1}$, where α_{x_1} is the hyper-parameter of the marginal distribution of the parent. The intervals computed by the global IDM are:

$$\left[\frac{n_{x_1x_2}}{n_{x_1}+\alpha_{x_1}}, \frac{n_{x_1x_2}+\alpha_{x_1}}{n_{x_1}+\alpha_{x_1}}\right]. \tag{3}$$

The global IDM estimates narrower posterior intervals than the local IDM because of these additional constraints. In fact, the intervals computed by the local IDM can be very wide when we analyze the corresponding set of joint distributions. On the other hand, under the global IDM, it is usually hard to solve inferences, because the parameters of the network become all correlated in some way. One of the few cases in which this computation is tractable is the naive credal classifier [10].

The EDM is a modification of the global IDM which restricts the IDM to its extreme distributions. Let us consider X_1 again; the EDM allows α_{x_1} to assume two values: 0 or s; hence, it does *not* consider all the Dirichlet distributions defined by the constraint $\sum_{x_1 \in \Omega_{X_1}} \alpha_{x_1} = s$, which are infinite. Analogously, $\alpha_{x_1x_2}$ can assume only two values: 0 or s, but still depends on α_{x_1}. In fact, the EDM treats the s hidden instances as s rows of missing data; the rows are assumed to be identical, but there is ignorance about the value assumed by each variable; such an ignorance determines the credal set.

When applied to a single variable, EDM returns the same interval of the global IDM; however, when applied to a credal network, it returns intervals that are included (or at most equivalent) in the intervals computed by the global IDM [1].

4 Credal Classification

We denote the class variable as C, assuming values in Ω_C; while the set of remaining variables $\mathcal{Y} = \mathcal{X} \setminus C$ are called *features*. The goal of classification is to build a classifier on a training set, and then to predict the unknown class of new instances, given the values $\mathbf{y}$ of the features.

According to [7], the optimality criterion for classification based on imprecise probabilities is to return the *non-dominated* classes. In particular, given the values $\mathbf{y}$ of the features, class *c'* dominates (or *credal-dominates*) class *c"* if and only if:

$$\min_{p \in \mathcal{K}} (p(c'|\mathbf{y}) - p(c''|\mathbf{y})) > 0$$

The set of non-dominated classes can be detected by performing repeated pairwise comparisons, as shown in Figure 1.

IDENTIFICATION OF NON-DOMINATED CLASSES

1. set NonDominatedClasses := Ω_C;
2. for class $c' \in \Omega_C$
   - for class $c'' \in \Omega_C$, $c'' \neq c'$
     - if c'' is dominated by c', drop c'' from NonDominatedClasses and break the internal loop;
3. return NonDominatedClasses.

Figure 1: Identification of non-dominated classes via pairwise comparisons.

A key point is that there can be several non-dominated classes and that these classes are incomparable; in this case, the classifier returns an indeterminate (or set-valued) classification. Classifiers that issue set-valued classifications are called *credal classifiers* . Intuitively, credal classifiers will return determinate classifications (i.e., a single class) on easy-to-classify instances, and more classes on hard-to-classify instances.

5 IDM vs. EDM: empirical comparison on Naive Credal Classifier

Before describing TANC, we experimentally evaluate the naive credal classifier with adoption of the EDM (NCC-EDM) against the traditional naive credal classifier based on the global IDM (NCC). When checking credal-dominance between c' and c'', NCC searches the minimum of $p(c')/p(c'')$ over $(0, s)$, while NCC-EDM evaluates the ratio $p(c')/p(c'')$ only in 0 and s. We have implemented NCC with EDM by reworking the code of JNCC2 [3], an open source implementation of NCC.

The answers returned by NCC and NCC-EDM might be different: when checking whether c' credal-dominates c'', it can happen that NCC-EDM detects credal-dominance while NCC, using a larger credal set (which by the way contains the former), does not detect credal-dominance (in other terms: NCC can find a lower minimum, implying non-dominance, than NCC-EDM).

Some of this different dominance tests do not affect the final set of non-dominated classes, because several pairwise comparisons are run, but some do. Therefore, NCC and NCC-EDM may return distinct sets of non-dominated classes.

To empirically evaluate the difference between NCC and NCC-EDM we have worked on 23 data sets from the UCI repository [4]. Each data set has been used as training and then as testing set; in fact, the goal here is to compare the answers of the two classifiers and not to provide an assessment of their accuracy.

On 22 data sets out of 23, the percentage of credal-dominance tests which receive a different answer from NCC-EDM and NCC is far smaller than 1%; the percentage of instances over which the two models return a different set of dominated classes is very low: 0.01% on average. The number of performed pairwise comparison overall is in the order of 10^6, while the total number of instances classified by NCC and NCC-EDM is around 10^5.

There is however a single data set over which NCC and NCC-EDM lead to different results: the *audiology*. It has 226 instances, 24 classes and 69 features. Remarkably, most binary features have *very* skewed distributions, such as 224 versus 2, or 225 versus 1. Because of the many classes and of the unevenly distributed features, the differences on the model of prior ignorance can lead to a different set of non-dominated classes. This happens on 51/226 instances, i.e., about 22% of the instances.

We conclude that NCC and NCC-EDM are practically equivalent on most cases; however, differences between the two models can arise on data sets with many classes and unevenly distributed features. Still, such indications support the introduction of EDM in classification.

[3] http://www.idsia.ch/~giorgio/jncc2.html

[4] http://archive.ics.uci.edu/ml/

6 Tree Augmented Naive Credal Classifier

The Tree-Augmented Naive (TAN) structure has the characteristic that each feature has at least C as parent and at most one other parent constituted by another feature. By Tree Augmented Naive Credal Classifier (TANC), we mean a credal network over a TAN graph.

As described in Section 4, TANC performs pairwise comparison to detect credal-dominance; for every comparison between two classes, the minimization is performed over (a) all possible completions of the training data (because missing data of the training set are nonMAR) and (b) over the prior densities belonging to the EDM. The credal dominance condition can be rewritten as:

$$\min_{\mathbf{d}_{\mathcal{X}},\boldsymbol{\alpha}} (p(c'|\mathbf{y}) - p(c''|\mathbf{y})) > 0,$$

because the distributions $p \in \mathcal{K}$ are completely defined by $\mathbf{d}_{\mathcal{X}}$ and $\boldsymbol{\alpha}$.

We assume further that there is no missing values in the class and that the hyper-parameters $\boldsymbol{\alpha}_C$ are fixed (we may solve at each time a given extreme configuration of $\boldsymbol{\alpha}_C$). Hence, the credal dominance problem is equivalent to

$$\min_{\mathbf{d}_{\mathcal{X}},\boldsymbol{\alpha}} (p(\mathbf{y}|c')p(c') - p(\mathbf{y}|c'')p(c'')) > 0$$

because $p(\mathbf{y})$ is positive and so does not affect the sign of the formula. Then we can separately solve each optimization as follows:

$$p(c') \cdot \min_{\mathbf{d}_{\mathcal{X}},\boldsymbol{\alpha}\setminus\boldsymbol{\alpha}_C} p(\mathbf{y}|c') - p(c'') \cdot \max_{\mathbf{d}_{\mathcal{X}},\boldsymbol{\alpha}\setminus\boldsymbol{\alpha}_C} p(\mathbf{y}|c'') \quad (4)$$

because $p(\mathbf{y}|c')$ only depends on $\alpha_{c'}$ and on the data of instances with $C = c'$, while $p(\mathbf{y}|c'')$ depends on $\alpha_{c''}$ and counts from instances with $C = c''$ (data with $C = c'$ and $C = c''$ are obviously disjoint).

Because we take the Extreme IDM as model for the priors, $\boldsymbol{\alpha}$ only assumes extreme values. Hence, it is possible to tackle the problem by introducing s new instances to the training set that are completely missing. As this new *fake* instance of missing values has also missing classes, it could introduce a dependence between the minimization and the maximization of Equation (4). However, it is possible to solve the optimization for every possible completion of the missing data of the class in this additional instance (which are just two extremes). Thus we have

$$p(c') \cdot \min_{\mathbf{d}_{\mathcal{X}}} p(\mathbf{y}|c') - p(c'') \cdot \max_{\mathbf{d}_{\mathcal{X}}} p(\mathbf{y}|c''), \quad (5)$$

which is solved for every possible completion of the data (including the fake instance).

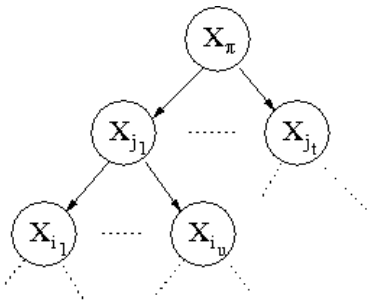


Figure 2: Part of the computation tree of the TANC algorithm.

The idea of the algorithm to evaluate Equation (5) is to combine the computations that are performed separately in the children of each variable and then to propagate the best possible solution to their sole parent. We ignore the arcs from C because we look for $\underline{p}(\mathbf{y}|c') = \min_{\mathbf{d}_\mathcal{X}} p(\mathbf{y}|c')$ and $\overline{p}(\mathbf{y}|c'') = \max_{\mathbf{d}_\mathcal{X}} p(\mathbf{y}|c'')$, that is, the actual root variable C is observed. The computation starts on the leaves and follows in a bottom-up idea. At each variable X_i, the goal is to obtain the joint probability $p(\mathbf{y}_{\mathbf{\Lambda_i}}|y_i, c)$ of its children conditional on y_i[5] (c equals c' or c'' depending whether it is the minimization or the maximization). This evaluation is done for all possible completions d_{X_i} and it is optimized over the completions of the children. The result is stored in a *cache* $\phi_i(d_{x_i})$. Figure 2 shows part of a network. At X_{j_1}, the joint probabilities $p(\mathbf{y}_{\mathbf{\Lambda_{i_k}}}|y_{i_k}, c)$ of every child $X_{i_k} \in \Lambda_{j_1}$ (for every possible completion of that sub-tree) are already computed. So, they are combined to obtain $p(y_{i_1}, \ldots, y_{i_u}|y_{j_1}, c)$, for every possible completion of X_{j_1}. These new probabilities $p(\mathbf{y}_{\mathbf{\Lambda_{j_1}}}|y_{j_1}, c)$ are then made available to the parent X_π, where the computations are analogous but using the information obtained from X_{j_1} and its siblings. The process goes through the tree structure until reaching the root.

Denote by $\mathbf{y}_{\sigma(\mathbf{i})}$ the assignment for all the variables in the sub-tree rooted at X_i, that is, $\mathbf{y}_{\sigma(\mathbf{i})} \in \Omega_{\mathbf{X}_{\sigma(\mathbf{i})}}$ is the queried assignment over $\mathbf{X}_{\sigma(\mathbf{i})} \subseteq \mathcal{X}$, the set of variables in the sub-tree rooted at X_i. Suppose that the root variables (if C is not considered) are $X_1, \ldots, X_r$. So,

$$\begin{aligned} p(\mathbf{y}|c') &= \prod_{j=1}^{r} p(\mathbf{y}_{\sigma(\mathbf{j})}|c') \\ &= \prod_{j=1}^{r} p(y_j|c') \cdot \prod_{X_i \in \Lambda_j} p(\mathbf{y}_{\sigma(\mathbf{i})}|y_j, c'), \end{aligned}$$

and, in general, $\min_{\mathbf{d}_\mathcal{X}} p(\mathbf{y}_{\sigma(\mathbf{j})}|\pi_j^y, c') =$

$$= \min_{\mathbf{d}_\mathcal{X}} \left(p(y_j|\pi_j^y, c') \cdot \prod_{X_i \in \Lambda_j} p(\mathbf{y}_{\sigma(\mathbf{i})}|y_j, c') \right),$$

where $\pi_j^y \in \Omega_{\Pi_j}$ is the assignment of Π_j that is being queried. (the maximization is analogous). Now, when you complete the variable X_j, the children Λ_j have separable computations. They are separable because the counts n that appear in the children of X_j are independent of each other as they concern disjoint subsets of variables (the structure is a tree, so $\mathbf{X}_{\sigma(\mathbf{i})} \cap \mathbf{X}_{\sigma(\mathbf{i}')} = \emptyset$ for $X_i, X_{i'} \in \Lambda_j$, with $i \neq i'$ and $X_j = \Pi_i = \Pi_{i'}$.). The only dependent value is n_{y_j}, as it appears in the denominators of distinct children of X_j. However, n_{y_j} is fixed as the problem is solved for every possible completion of X_j. Besides that, note that the terms $\boldsymbol{\alpha}$ are not present because we treat them using the *fake* missing instance. Hence, the overall computation can be decomposed as

$$= \min_{d_{X_j}} \left(p(y_j|\pi_j^y, c') \cdot \prod_{X_i \in \Lambda_j} \min_{d_{\mathbf{X}_{\sigma(\mathbf{i})}}} p(\mathbf{y}_{\sigma(\mathbf{i})}|y_j, c') \right).$$

To prove that this idea is correct, we rewrite it as a function of completions: $\forall d_{\mathbf{X}_{\sigma(\mathbf{j})}}$, we have

$$\phi_j(d_{\mathbf{X}_{\sigma(\mathbf{j})}}) = \prod_{X_i \in \Lambda_j} \min_{d_{\mathbf{X}_{\sigma(\mathbf{i})}}} \left(\frac{n_{y_i y_j}}{n_{y_j}} \phi_i(d_{\mathbf{X}_{\sigma(\mathbf{i})}}) \right), \quad (6)$$

where the product is assumed to be 1 when Λ_j is empty. The maximization version is analogous. We prove by induction on the tree the following property:

$$\phi_j(d_{\mathbf{X}_{\sigma(\mathbf{j})}}) = \begin{cases} 1, & \text{if } X_j \text{ is a leaf,} \\ \underline{p}(\mathbf{y}_{\sigma(\mathbf{j})} \setminus \{y_j\}|y_j, c'), & \text{otherwise.} \end{cases} \quad (7)$$

The base of induction holds by definition. Now assume that Equation (7) holds for every $X_i \in \Lambda_j$. By applying this hypothesis on Equation (6), we have

$$\phi_j(d_{\mathbf{X}_{\sigma(\mathbf{j})}}) = \prod_{X_i \in \Lambda_j} \min_{d_{\mathbf{X}_{\sigma(\mathbf{i})}}} \left(\frac{n_{y_i y_j}}{n_{y_j}} \underline{p}(\mathbf{y}_{\sigma(\mathbf{i})} \setminus \{y_i\}|y_i, c') \right), \quad (8)$$

where n_{y_j} is fixed and $n_{y_i y_j}$ depends on the completion d_{X_i}, which belongs to $d_{\mathbf{X}_{\sigma(\mathbf{i})}}$. Thus, it is possible to minimize the factor of each child separately and we obtain $\phi_j(d_{\mathbf{X}_{\sigma(\mathbf{j})}}) = \underline{p}(\mathbf{y}_{\sigma(\mathbf{j})} \setminus \{y_j\}|y_j, c')$.

The derivation so far requires exponential time over all missing values. Nevertheless, an important fact in Equation (6) is that $\phi_i(d_{\mathbf{X}_{\sigma(\mathbf{i})}}) = \phi_i(d_{X_i})$, for d_{X_i} compatible with $d_{\mathbf{X}_{\sigma(\mathbf{i})}}$, that is, it is enough to keep the best possible solution for every completion of a variable without having to record all the completions of its descendants. This is valid because $n_{y_i y_j}$ is known when the completion d_{X_i} is given, so completions of variables in $\mathbf{X}_{\sigma(\mathbf{i})} \setminus \{X_i\}$ are irrelevant for the minimization in Equation (6), and it is enough to have the best possible solution for each d_{X_i}. This leads us only to compute:

$$\forall_{d_{X_j}} \ \phi_j(d_{X_j}) = \prod_{X_i \in \Lambda_j} \min_{d_{X_i}} \left(\frac{n_{y_i y_j}}{n_{y_j}} \phi_i(d_{X_i}) \right), \quad (9)$$

[5] $y_i \in \Omega_{X_i}$ is used as the notation for the queried state of X_i.

and equivalently in the maximization case. Now the algorithm can be implemented in a bottom-up manner so as the ϕ's of children are available when a given variable is treated, which reduces the complexity of the method to be exponential in the number of missing values of only two variables (a variable and its parent) instead of all missing values.

The described formulation obtains $\overline{p}(\mathbf{y} \setminus \{y_i\}|c'')$ and $\underline{p}(\mathbf{y} \setminus \{y_i\}|c')$, for each root variable X_i, $i \leq r$. Those values still need to be multiplied by the corresponding $p(y_i|c)$ (using the proper c). We leave this last step intentionally apart to show how to deal with the forest of trees. The probability of the variables that have only C as parent are multiplied all together, just as if we had computed $\phi_C(\cdot)$ according to Equation (9):

$$\underline{p}(\mathbf{y}|c') = \phi_C(\cdot) = \prod_{X_i \in \Lambda_C} \min_{d_{X_i}} \left(\frac{n_{y_i c'}}{n_{c'}} \phi_i(d_{X_i}) \right), \quad (10)$$

and similarly for the maximization. In case $r = 1$ (single root), the outer product of Equation (10) disappears. This final step returns the desired values $\overline{p}(\mathbf{y}|c'')$ and $\underline{p}(\mathbf{y}|c')$, which are later multiplied by $p(c'')$ and $p(c')$, respectively, to evaluate Equation (5).

We point out that, if the data set is complete, the only missing data that must be processed by the algorithm are those introduced by the fake instance (for the treatment of the EDM). In such case, the complexity of the method is clearly linear in the input size, as there is a constant number of computations by variable (there are only two ways of completing the data by variable and the algorithm is locally exponential). In fact there are other ideas that might be employed to solve the problem of selecting the hyperparameters $\boldsymbol{\alpha}$ of the EDM, but we use the idea of *fake* instance because it fits straightforward into the framework of the proposed algorithm. In the presence of missing data, the idea spends exponential time in the number of missing data of two linked variables, which is already much better than an overall exponential but still slow for data sets with many missing values. Using dynamic programming, it might be possible to further reduce this complexity to exponential in the missing of a single variable.

7 Experiments on TAN

We have performed experiments on several data sets retrieved from the UCI repository. The data sets cover a wide spectrum in terms of number of instances (min: 101; max: 12960) and classes (min: 2; max: 11). On each data set, we have performed 10-folds cross-validation to the performance of the classifiers. Numerical features have been discretized using supervised discretization [4]; the features discretized into a single bin have been removed from the computation. TANC requires the features to be discrete; however, supervised discretization of the features is a good practice in general, as it has been shown to improve the accuracy of several classifiers [3].

The instances over which TANC return a single class are referred to as *determinately* classified, while those over which TANC returns more classes are referred to as *indeterminately* classified.

For some data sets, we report results before and after having performed feature selection. To perform feature selection, we have cross-checked the suggestions of two feature selectors implemented in WEKA [9]: correlation-based and wrapper. Both approaches are multivariate, i.e., they are designed to identify an optimal subset of feature, by also considering interaction between features.

7.1 TANC vs TAN

In this section, we compare TANC against TAN on *complete* data sets, i.e., with no missing data. On each cross-validation run, we first learn the structure of the graph using WEKA [9]; later, we run TAN and TANC, using the same network structure for both.

We adopt a set of indicators already known in literature [10] for comparing a credal and a Bayesian classifier; in particular:

- *determinacy (D%)*: the percentage of instances classified determinately by TANC;
- *TAN-D* and *TAN-I*: the accuracy of TAN on the instances which are classified determinately and indeterminately by TANC. As TANC is designed to separate hard-to-classify instances (that are prior dependent, and hence indeterminately classified) and easy-to-classify instances (those determinately classified), we shall observe TAN-D>TAN-I, because TAN-D is in fact the accuracy achieved both by TAN and TANC on the determinately classified instances. Actually, if TANC is determinate, TAN and TANC return the same classification (although the uniform prior adopted for TAN is not included in the credal set of the EDM, it empirically appears that if both the extreme priors of the EDM indicate the same class as the most probable one, TAN will lead to the same conclusion too).
- *set-accuracy (S-acc%)*: the ratio of the number indeterminate classifications which contain the actual class to the total number of indeterminate classifications;
- *indeterminate output size (ind. sz.)*: the average number of classes returned on the instances indeterminately classified.

Note that set-accuracy and indeterminate output size are meaningful only if the data set has more than two classes.

data set	#inst.	#cl.	D	TAN-D	TAN-I	S-acc	Ind. Sz.
zoo	101	7	77%	100%	84%	100%	5.9/7
iris	150	3	93%	97%	44%	92%	2.8/3
diabetes	767	2	98%	80%	9%	-	-
segment	810	7	17%	99%	93%	98%	4.1/7
vehicle	846	7	67%	82%	60%	88%	2.4/7
vowel	990	11	66%	98%	77%	99%	7.6/11
credit	1000	2	87%	78%	55%	-	-
splice	3190	3	90%	97%	79%	99%	2.1/3
kr-kp	3196	2	99%	92%	40%	-	-
waveform	5000	3	88%	85%	73%	100%	2.1/3
nursery	12960	5	90%	97%	79%	99%	3.5/5
average			**79%**	**91%**	**63%**	**97%**	**72%**

Table 1: The data sets (sorted according to the number of instances) and the indicators. The meaning of the header is as follows: #inst denotes the number of instances of the data set and #cl the number of classes; D is the determinacy of TANC; TAN-D and TAN-I the accuracy achieved by TAN on the instances classified determinately and indeterminately by TANC; S-acc is the set-accuracy of TANC while Ind. Sz. is the average number of instances returned by TANC on the instances indeterminately classified.

The results of Table 1 show that the determinacy of TANC is quite high: 79% on average, and often above 90%. In general, the determinacy of TANC increases with the number of instances in the data set (as large the data set as reduced the importance of the prior) and decreases, for similarly-sized data sets, with the number of classes. The major exception to this is segment (17 features, 7 classes, 810 instances); however, in this case feature selection can be helpful. It turns out that 10 out of the 17 features in segment are irrelevant; removing them from the data set and re-running the experiment increases the determinacy from 17% to 57%, with only a minor drop of accuracy on the instances determinately classified (TAN-D decreases from 99% on 17% of instances to 95% on 57% of instances).

Most importantly, TANC is quite effective in separating hard-to-classify from easy-to-classify instances. There is a sharp drop of accuracy of TAN when we move from determinate to indeterminate instances; on the average, the drop is about 28 percentage points. On data sets with two classes, the accuracy of TAN on the instances indeterminately classified is comparable to random guessing or even worse (diabetes: 9%; credit: 55%, kr-kp: 40%); however, as the number of classes increases, TAN performs better on the instances indeterminately classified (see for instance segment and zoo: TAN-I is 84% and 93% respectively). This might show that as the number of classes increases, TANC becomes indeterminate also on some instances that could be successfully classified. However, studies suggest that even this kind of problem can be significantly mitigated by feature selection [2]. As an example, let us consider the zoo data set, which has 15 features. By running feature selection, we find that there are 4 irrelevant features out of 15. Re-running the experiment on the pruned data, determinacy rises from 77% to 79%, TAN-D remains close to 100%, while TAN-I drops from 84% to 72%. Hence, in some particular data sets, feature selection can be helpful to improve the determinacy and/or the detection of hard-to-classify instances.

On the hard-to-classify instances, TANC preserves its reliability thanks to indeterminate classifications, providing set-accuracy close to 100%, while returning on the average about 70% of the total classes. All these findings are in good agreement with previous comparisons of Bayesian classifiers against their imprecise probability counterparts [2, 13].

7.2 TANC vs. TANC*

Two main differences exist between TANC and TANC* regarding the model of prior ignorance (TANC adopts the EDM, while TANC* adopts the local IDM) and the treatment of missing data in the training set (TANC* assumes MAR, while TANC does not). In this section, we focus on the impact of the models of prior ignorance on the two classifiers; in order to remove the effect of the treatment of missing data, we consider *complete* data sets. We did not implement TANC* in our code; rather, we have compared our results with those published in [13]. For this reason, the analysis tries to draw general conclusions rather than punctual ones. We consider here all the 6 complete data sets analyzed in [13].

On the basis of previous explanation, we can expect TANC to be more determinate that TANC*. However we have also to verify that it becomes determinate on instances that can be safely classified with a single class.

The comparison between the determinacy of TANC and TANC* is shown in the upper plot of Figure 3. On the

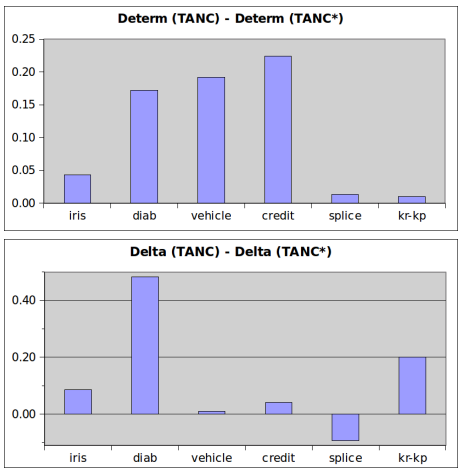


Figure 3: Comparison between TANC and TANC*.

average, TANC is 11 percentage points higher than that of TANC (89% vs. 78%). However, the determinacy of the two classifiers is almost equivalent on both splice and kr-kp; this might be due to the large size of the two data sets (around 3200 instances each), which reduces the role of the prior distributions.

In order to compare the ability of isolating hard-to-classify instances, we introduce the indicator Delta $=$ (TAN-D $-$ TAN-I), which evaluates the difference in accuracy achieved by TAN between the instances classified determinately and indeterminately by TANC [resp. TANC*]. The results are displayed in the lower plot of Figure 3; they suggests that the increased determinacy of TANC corresponds also to a better ability in isolating hard-to-classify instances, thus supporting the hypothesis that TANC is returning determinate answers on instances over which TANC* is unnecessarily indeterminate. However, these results should be taken with some cautiousness, as it has not been possible to actually run side-by-side the two classifiers.

7.3 Preliminary results with missing data

In this section we focus on comparing the determinacy of the classifier in the presence of missing data. The effect of the treatment of missing data is also important so as to verify the consequences of nonMAR in terms of accuracy, but a deeper analysis is left for future work. We note that the term nonMAR is employed to indicate the ignorance about the MP, that is, MAR is not assumed. In particular, we consider the crx data set, which has 16 features; the structure of the network has 14 links among features (besides those which connect the class to all the features). We consider the complete data set and then artificially generate 30 missing values, distributed among 6 different features. Even such a small quantity of missing data decreases the determinacy from 87% to 77%. On the very same data sets, we run the naive credal classifier 2 [2] which can be seen as NCC enabled for NonMAR treatment of missing data; the determinacy of NCC2 (assuming NonMAR) remains stable around 95% on both cases. Hence, it seems that the TAN structure can lead to much larger indeterminacy than the naive one, if MAR is not assumed. This result is somehow expected, as TAN introduces the possibility of having linked features with missing values, while a naive structure does not.

8 Conclusions

TANC is a new credal classifier based on a Tree-Augmented Naive structure; it treats missing data conservatively by considering all possible completions of the training set, but avoiding an exponential increase of the computational time. TANC adopt the EDM as a model of prior ignorance; we have shown that EDM is a reliable and computationally affordable model of prior near-ignorance for credal classifiers. We have shown that TANC is more reliable than precise TAN (learned with uniform prior) and that it obtains better performance compared to a previous TAN model based on imprecise probabilities, but learned with a local IDM approach; the adoption of EDM overcomes the problem of the unnecessary imprecision induced by the local IDM, while keeping the computation affordable.

The TANC classifier has room for many improvements. The treatment of MAR and nonMAR missing data all together, appearing both in the training and the testing set are the main topics for future work. In order to make TANC less indeterminate on incomplete data sets, a solution could be to allow for mixed configurations, in which some features are treated as MAR and some others are not. This would allow both for a decrease of indeterminacy and for a finer-grained tuning of the way that missing data are dealt with. Besides that, the computational performance of TANC can also be further improved, for example, with the use of dynamic programming. Extensions beyond trees are also of interest, but they fall into the need of fast and accurate inference methods for general credal networks.

Acknowledgments

Work partially supported by the Swiss NSF grant n. 200021-118071/1 and 200020-116674/1 and from the project 'Ticino in rete'.

References

[1] A. Cano, M. Gómez-Olmedo, and S. Moral. Credal nets with probabilities estimated with an extreme imprecise Dirichlet model. In *Proceedings of the Fifth International Symposium on Imprecise Probability:*

Theories and Applications (ISIPTA'07), Action M Agency, Prague, pages 57–66, 2007.

[2] G. Corani and M. Zaffalon. Learning Reliable Classifiers from Small or Incomplete Data Sets: the Naive Credal Classifier 2. *Journal of Machine Learning Research*, 9:581–621, 2008.

[3] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In A. Prieditis and S. Russell, editors, *Proceedings of the 12th conference on machine learning*, pages 194–202, San Francisco, CA, 1995. Morgan Kaufmann.

[4] U. M. Fayyad and K. B. Irani. Multi-interval Discretization of Continuous-valued Attributes for Classification Learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1022–1027, San Francisco, CA, 1993. Morgan Kaufmann.

[5] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian Network Classifiers. *Machine Learning*, 29(2):131–163, 1997.

[6] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, 1987.

[7] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.

[8] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *J. R. Statist. Soc. B*, 58(1):3–57, 1996.

[9] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann, 2005.

[10] M. Zaffalon. Statistical inference of the naive credal classifier. In G. de Cooman, T. L. Fine, and T. Seidenfeld, editors, *ISIPTA '01: Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, pages 384–393, The Netherlands, 2001. Shaker.

[11] M. Zaffalon. Exact credal treatment of missing data. *Journal of Statistical Planning and Inference*, 105(1):105–122, 2002.

[12] M. Zaffalon. Conservative rules for predictive inference with incomplete data. In F. G. Cozman, R. Nau, and T. Seidenfeld, editors, *ISIPTA '05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*, pages 406–415, Manno, Switzerland, 2005. SIPTA.

[13] M. Zaffalon and E. Fagiuoli. Tree-Based Credal Networks for Classification. *Reliable Computing*, 9(6):487–509, 2003.

Sets of Desirable Gambles and Credal Sets

Inés Couso
Dpto. Estadística e IO
EUIT Industrial, Univ. de Oviedo, Spain
couso@uniovi.es

Serafín Moral
Dpto. de Ciencias de la Computación e IA
ETSI Informática, Univ. de Granada, Spain
smc@decsai.ugr.es

Abstract

Sets of desirable gambles were proposed by Walley [7] as a general theory of imprecise probability. The main reasons for this are: it is a very general model, including as particular cases most of the existing theories for imprecise probability; it has a deep and simple axiomatic justification; and mathematical definitions are natural and intuitive. However, there is still a lot of work to be done until the theory of desirable gambles is operative for its use in general reasoning tasks. This paper gives an overview of some of the fundamental concepts expressed in terms of desirable gambles in the finite case, gives a characterization of regular extension, and studies the nature of maximally coherent sets of gambles.

Keywords. Desirable gambles, regular extension, zero probabilities, sets of probability measures.

1 Introduction

Sets of desirable gambles are a powerful and simple model for representing and reasoning with imprecise probabilities. For these reasons, they were proposed by Walley [7] as a general model for imprecise probability after studying the limitations of other models.

The axioms for desirable gambles were introduced by Williams [9] and Walley studied them in Appendix F of his book [6]. They were also considered in [5] as a basic for a logical approach to probability. They are mathematically equivalent to partial probability orderings [1, 3], but they are simpler [7]. Because of this, desirable gambles are a more suitable theory of uncertainty. Even though, their use in the literature is very scarce. In many cases, it is possible to find papers based on other representations, as for example lower and upper previsions, in which the rules for inference are deduced making arguments which are based on desirability. This makes desirability a more primitive notion.

Moral [4] recently studied the concept of epistemic irrelevance in terms of desirable gambles which resulted in a very natural approach to this notion, as it was possible to show a number of properties in a simple form.

In this paper, we give an overview of some of the main concepts of desirable gambles in the finite case, showing the difference between desirable gambles and almost desirable gambles (Section 2). Then, we study the concept of conditioning, showing how the rules of conditioning for lower previsions can be obtained from the simple definition of conditioning for sets of desirable gambles and giving an axiomatic justification of regular extension (Section 3). One of the problems associated to the use of desirable gambles is the lack of effective methods of representing information and algorithms to make inference from available information. Section 4 discusses this issue and shows that there are algorithms in the literature which can be directly applied in this theory. Finally Section 5 studies the case of maximally coherent sets of desirable gambles. These sets have always an associated precise probability measure. But, as sets of desirable gambles contain more information than probability measures, we prove that we can associate a more complex structure to the maximal coherent sets: a sequence of probability measures, each one of them defined in the set in which the previous measure in the sequence assigns a zero probability, similar to the sequences defined in [2]. We also show that a general coherent set can be expressed in terms of maximal (precise) coherent sets.

2 Sets of Desirable Gambles

Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ denote the (finite) set of outcomes. We assume that there is an unkown true value belonging to Ω. A *gamble* on Ω is a bounded mapping from Ω to $\mathbb{R}$, i.e., $X : \Omega \to \mathbb{R}$. Gambles are used to represent an agent's beliefs and information. If an agent accept a gamble X, then the value $X(\omega)$ rep-

resents the reward she would obtain if ω is the true unknown value (this value can be negative and then it represents a loss).

Let $\mathcal{L}$ denote the set of all gambles defined on Ω. For $X, Y \in \mathcal{L}$, let $X \geq Y$ mean that $X(\omega) \geq Y(\omega)$ for all $\omega \in \Omega$, and let $X > Y$ mean that $X \geq Y$ and $X(\omega) > Y(\omega)$ for some $\omega \in \Omega$.

A subset $\mathcal{D}$ of $\mathcal{L}$ is said to be a *coherent set of desirable gambles* relative to $\mathcal{L}$ [7] when it satisfies the following four axioms:

D1. $0 \notin \mathcal{D}$,

D2. if $X \in \mathcal{L}$ and $X > 0$ then $X \in \mathcal{D}$,

D3. if $X \in \mathcal{D}$ and $c \in \mathbb{R}^+$ then $cX \in \mathcal{D}$,

D4. if $X \in \mathcal{D}$ and $Y \in \mathcal{D}$ then $X + Y \in \mathcal{D}$.

In what follows, $\mathcal{D}$ is assumed to be a coherent set of gambles. We assume that information is represented by means of a coherent set of gambles. These rules represent the consistency conditions for the gambles that are considered desirable. For example, Axiom D4 says that if we consider as desirable X and Y, then we should consider as desirable the gamble resulting from adding the rewards of both gambles. Axiom D2 says that a positive gamble (we can win but never lose) is always desirable.

The null gamble is neutral and then it is not included in the set of really desirable gambles, but this is not an important fact. In some cases, as in [4, 6], the null gamble has been considered desirable. The real important condition for coherence is that if $X < 0$, then $X \notin \mathcal{D}$ (avoiding *partial loss*). In our approach, this condition is a consequence of D1 and the other axioms (D2 and D4). But both options are completely equivalent, in the sense that the only difference is the inclusion of the null gamble in the set of desirable gambles and this does not have any difference in practice. The only consequence of taking one of the two possible options is that some mathematical definitions have to be changed (for example, conditioning is different if we accept the null gamble). Walley first considered the null gamble desirable in [6], but then he changed to consider it non desirable in [4]. In this moment, we also consider that this last option is simpler and more intuitive.

The *lower prevision induced by* $\mathcal{D}$ is the function $\underline{P} : \mathcal{L} \to \mathbb{R}$ defined as follows: $\underline{P}(X) = \sup\{c : X - c \in \mathcal{D}\}$.

The *upper prevision induced by* $\mathcal{D}$ is the function $\overline{P} : \mathcal{L} \to \mathbb{R}$ defined as follows: $\overline{P}(X) = \inf\{c : c - X \in \mathcal{D}\}$.

The set of linear previsions induced by $\mathcal{D}$ is defined as:

$$\mathcal{P}_{\mathcal{D}} = \{P : P(X) \geq 0 \text{ for all } X \in \mathcal{D}\}.$$

$\mathcal{P}_{\mathcal{D}}$ is always a *credal set* (a closed and convex set of probability measures). $\underline{P}$ and $\overline{P}$ are dual and they respectively coincide with the pointwise infimum and the pointwise supremum of $P \in \mathcal{P}_{\mathcal{D}}$. There can be two different sets of desirable gambles $\mathcal{D} \neq \mathcal{D}'$ inducing the same class of linear previsions $\mathcal{P}_{\mathcal{D}} = \mathcal{P}_{\mathcal{D}'}$.

Conversely, given a set of linear previsions $\mathcal{P}$, define

$$\mathcal{D}_{\mathcal{P}} = \{X \in \mathcal{L} : P(X) > 0, \forall P \in \mathcal{P}\} \cup \{X : X > 0\}.$$

$\mathcal{D}_{\mathcal{P}}$ is called the set of *strictly desirable gambles* associated to $\mathcal{P}$ [6].

$\mathcal{D}_{\mathcal{P}}$ is coherent and, if $\mathcal{P}$ has been induced by a set of desirable gambles $\mathcal{D}$, then $\mathcal{D}_{\mathcal{P}}$ is a subset of it. In other words, the following inclusion holds:

$$\mathcal{D}_{\mathcal{P}_{\mathcal{D}}} \subseteq \mathcal{D}$$

$\mathcal{D}_{\mathcal{P}}$ is the smallest set of gambles associated to a credal set $\mathcal{P}$.

$\mathcal{P}$ can be recovered from $\mathcal{D}_{\mathcal{P}}$ by

$$\mathcal{P} = \mathcal{P}_{\mathcal{D}_{\mathcal{P}}}. \tag{1}$$

Another possible set of desirable gambles associated to $\mathcal{P}$, but with more gambles in it is:

$$\mathcal{D}'_{\mathcal{P}} = \{X \in \mathcal{L} : P(X) \geq 0, \forall P \in \mathcal{P} \text{ and } \exists P \in \mathcal{P}, \text{ with } P(X) > 0\} \cup \{X : X > 0\}$$

A coherent set $\mathcal{D}$ of *almost desirable gambles* is a set of gambles which satisfies axioms D2, D3, and D4 and the following two axioms (the first one is a modification of the corresponding axiom for desirable gambles. The new version is called *avoiding sure loss*):

D1'. $-1 \notin \mathcal{D}$,

D5. if $X + \epsilon \in \mathcal{D}, \forall \epsilon > 0$, then $X \in \mathcal{D}$

A set of almost desirable gambles $\mathcal{D}$ can define a lower prevision, an upper prevision, and a credal set, by means of expressions completely analogous to the case of desirable gambles. But now, from a credal set $\mathcal{P}$, the associated set of almost desirable gambles $\mathcal{D}$ is given by:

$$\mathcal{D}^*_{\mathcal{P}} = \{X \in \mathcal{L} : P(X) \geq 0, \forall P \in \mathcal{P}\} \tag{2}$$

Intuitively, the set of desirable gambles contains all the gambles that are really desirable, i.e. the agent has reasons to accept them as desirable. The set of almost desirable gambles also includes all the gambles that are limit of desirable gambles, though some of them as the null gamble is not really desirable. $\mathcal{D}_{\mathcal{P}}$, the set of striclty desirable gambles associated to $\mathcal{P}$ is the interior of $\mathcal{D}^*_{\mathcal{P}}$ in the supremum norm topology [6].

If $\mathcal{D}$ is a coherent set of desirable gambles, then $\mathcal{D}^*$ will be the coherent set of almost desirable gambles obtained by adding to it the gambles resulting of the application of Axiom D5 (closure). Both $\mathcal{D}$ and $\mathcal{D}^*$ always define the same credal set. If the credal set is $\mathcal{P}$, then $\mathcal{D}_{\mathcal{P}} \subseteq \mathcal{D} \subseteq \mathcal{D}^*$. $\mathcal{D}_{\mathcal{P}}$ contains the strictly desirable gambles. If a gamble X is in $\mathcal{D}_{\mathcal{P}}$, then there is a $\delta > 0$ such that $X - \delta \in \mathcal{D}_{\mathcal{P}}$, i.e. even paying a quantity δ, the gamble continues being desirable. $\mathcal{D}^*$ contains more gambles, all the gambles such that for any $\epsilon > 0$, $X + \epsilon$ is desirable, i.e., if we receive any positive quantity, this is enough to make the gamble desirable (but the gamble alone may not be desirable). $\mathcal{D}$ is the set of gambles that are considered desirable by an agent without any additional consideration in the limit.

Coherent sets of almost desirable gambles and credal sets are equivalent, in the sense that there is a one-to-one correspondence between these two families. If $\mathcal{D}$ is a set of almost desirable gambles: $\mathcal{D}^*_{\mathcal{P}_{\mathcal{D}}} = \mathcal{D}$. A credal set is a convex and closed set of probabilities and an almost desirable gamble can be interpreted as a linear restriction on the credal set by means of expression $P(X) \geq 0$. The difference between desirable and almost desirable gambles is that a set of almost desirable gambles is always closed, and a set of desirable gambles is never closed (the null is the limit of desirable gambles and is never desirable) but not necessarily open either. The set of strictly desirable gambles is always open. Axioms can be also defined for strict desirable gambles [6] and it is possible to show the equivalence between sets of strict desirable gambles and credal sets.

Example 1 *Consider the credal set $\mathcal{P}$ represented in Figure 2 for a frame with three elements $\{\omega_1, \omega_2, \omega_3\}$, where each point is a probability mass function with values determined by the distances to the triangle edges. Imagine that $\mathcal{D}$ and $\mathcal{D}^*$ are a set of desirable gambles and the set of almost desirable gambles associated to it. A gamble can be associated to a linear restriction about the probabilities through the inequality $P(X) \geq 0$. If this inequality is not trivial in the set of probabilities (X is trivial if $X \geq 0$), then in the triangle we will see the inequality as a segment dividing the triangle in two parts and a direction determining in which of the two parts the inequality is verified. So, a non trivial gamble X can be associated with a segment and a direction. A gamble is almost desirable if all the probabilities in the credal set verify the restriction. In the figure, X_1 and X_3 are almost desirable and X_2 is not as there is a probability in $\mathcal{P}$ not verifying the inequality associated to X_2. X_1 is also strictly desirable. For desirability we have a necessary condition: if X is desirable then $P(X) \geq 0$ for any $P \in \mathcal{P}$. So, as X_2 does not verify it, it can not be desirable. We also have a sufficient condition: if $P(X) > 0$, for any P, then X has to be strictly desirable and desirable. So X_1 is desirable and strictly desirable. The difference is in those gambles X, for which $P(X) \geq 0$ for any P, but $P(X) = 0$ for some P. This gamble is almost desirable and can not be strictly desirable, but it can be desirable or not desirable. So, it is not determined whether gamble X_3 (touching the border of the credal set) is or is not desirable. These gambles in the border determine the difference between desirability, almost desirability, and strict desirability. They have behabioural consequences, in particular after conditioning to events of probability 0.*

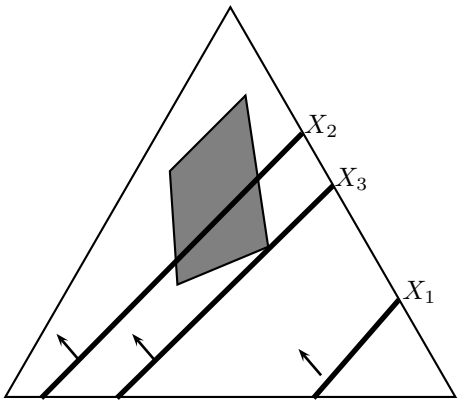


Figure 1: Desirable and almost desirable gambles

If $\mathcal{G}$ is an arbitrary set of gambles, then the set of all gambles obtained by applying axioms D2, D3, and D4 is called the *set of gambles generated* by $\mathcal{G}$ and it is denoted by $\overline{\mathcal{G}}$. If this set is coherent ($0 \notin \overline{\mathcal{G}}$) then it will be called its *natural extension* (the minimum coherent set containing $\mathcal{G}$). If $0 \in \overline{\mathcal{G}}$ we will say that $\mathcal{G}$ is incoherent. If $X < 0$ and $X \in \overline{\mathcal{G}}$ we will say that $\mathcal{G}$ does not *avoid partial loss*.

It is an immediate result that

$$\overline{\mathcal{G}} = \{\sum_{i=1}^{n} \lambda_i X_i \,:\, \lambda_i > 0,\ [X_i \in \mathcal{G} \text{ or } X_i > 0] \quad i \leq n \in \mathbb{N},\ n \geq 1\}.$$

Walley [6] considers the gambles that dominate (are greater or equal) than the positive linear combination

of gambles in $\mathcal{G}$. Our expression with equality is equivalent as we allow to combine positive gambles, except that we avoid to add the 0 gamble.

3 Conditioning

Let us consider a set of desirable gambles $\mathcal{D}$ on Ω. Let B denote (the indicator function of) an arbitrary subset of Ω. The set of *B-desirable gambles* ([6], Section 6.1.6) can be defined as follows:

$$\mathcal{D}_B = \{X \in \mathcal{L} : BX \in \mathcal{D}\} \cup \{X : X > 0\}.$$

This set will be also called the set of *conditional desirable gambles given B*. This set is determined by those gambles Y that are desirable and that outside of B are null, i.e. nothing happens if B does not occur. A gamble X belongs to $\mathcal{D}_B$ if BX is equal to one of these gambles or is positive.

The following results relate this definition with the usual concept in the associated credal set, consisting in computing the conditional probability given B of all the probability measures in the credal set (when $P(B) > 0$ for all the probabilities). In all of them, $\mathcal{D}^*$ is the set of almost desirable gambles associated to the set of desirable gambles $\mathcal{D}$.

Lemma 1 *Let $\mathcal{D} \subset \mathcal{L}$ be a coherent set of desirable gambles and B a subset of Ω such that $\underline{P}(B) > 0$. Then:*

$$X \in \mathcal{D}^* \Rightarrow X + \epsilon\, B \in \mathcal{D},\ \forall \epsilon > 0.$$

Proof: According to the above hypotheses, $\underline{P}(B) > 0$ and thus, there exists some $c > 0$ such that $B - c \in \mathcal{D}$. Furthermore, the gamble $X + \epsilon c$ is assumed to belong to $\mathcal{D}$, for all $\epsilon > 0$. By the coherence of $\mathcal{D}$, the gambles $\epsilon(B - c) = \epsilon B - \epsilon c$ and $X + \epsilon\, B = (X + \epsilon c) + (\epsilon B - \epsilon c)$ belong to it, for each $\epsilon > 0$, and thus the thesis of the lemma is checked. □

Lemma 2 *Let $\mathcal{D} \subset \mathcal{L}$ be a coherent set of desirable gambles satisfying the condition:*

$$X \in \mathcal{D}^* \text{ and } -X \notin \mathcal{D}^* \Rightarrow X \in \mathcal{D}. \qquad (3)$$

Then, for any B subset of Ω such that $\overline{P}(B) > 0$., the following condition is also verified:

$$X \in \mathcal{D}^* \Rightarrow X + \epsilon\, B \in \mathcal{D},\ \forall \epsilon > 0.$$

Proof: Let us assume that $X + \epsilon \in \mathcal{D},\ \forall \epsilon > 0$. Then, by the coherence of $\mathcal{D}$, $(X + \epsilon B) + \epsilon' = (X + \epsilon') + \epsilon\, B \in \mathcal{D},\ \forall \epsilon, \epsilon' > 0$. So $X + \epsilon B \in \mathcal{D}^*$. To prove that this gamble is also in $\mathcal{D}$, we only have to prove that $-X - \epsilon B \notin \mathcal{D}^*$, i.e., there exists some $\epsilon'' > 0$ such that $-(X + \epsilon B) + \epsilon'' \notin \mathcal{D}$. Let us check it by contradiction. Let us suppose that $\epsilon'' - (X + \epsilon B) \in \mathcal{D},\ \forall \epsilon'' > 0$. Then the gamble $\epsilon''' + \epsilon'' - \epsilon B = (X + \epsilon''') + (\epsilon'' - (X + \epsilon B))$ belongs to $\mathcal{D}$, for all $\epsilon'', \epsilon''' > 0$ by the coherence of $\mathcal{D}$. But the last assertion contradicts the assumption $\overline{P}(B) > 0$. □

Theorem 3 *Let $\mathcal{D} \subset \mathcal{L}$ be a coherent set of desirable gambles and let B be a subset of the universe Ω. Let us assume that the following condition holds:*

$$(X \in \mathcal{D}^* \Rightarrow X + \epsilon\, B \in \mathcal{D},\ \forall \epsilon > 0). \qquad (4)$$

Then,

$$\mathcal{P}_{\mathcal{D}_B} = (\mathcal{P}_{\mathcal{D}})_{|B},$$

where $(\mathcal{P}_{\mathcal{D}})_{|B}$ denotes the set of linear previsions

$$(\mathcal{P}_{\mathcal{D}})_{|B} = \{P(\cdot|B) : P \in \mathcal{P}_{\mathcal{D}} \text{ and } P(B) > 0\},$$

and, for each P with $P(B) > 0$, $P(\cdot|B)$ is defined as follows:

$$P(X|B) = \frac{P(BX)}{P(B)},\ \forall X \in \mathcal{L}.$$

Proof:

First, let us prove that $(\mathcal{P}_{\mathcal{D}})_{|B} \subseteq \mathcal{P}_{\mathcal{D}_B}$. If $Q \in (\mathcal{P}_{\mathcal{D}})_{|B}$, then $Q = P(.|B)$, where $P \in \mathcal{P}_{\mathcal{D}}$ and $P(B) > 0$.

If $X \in \mathcal{D}_B$, then either $X > 0$, and then it is verified $Q(X) \geq 0$, or $XB \in \mathcal{D}$. In the last case, as $P \in \mathcal{P}_{\mathcal{D}}$, we have that $P(XB) \geq 0$, and as $Q = P(.|B)$, then $Q(X) = Q(XB) = \frac{P(XB)}{P(B)} \geq 0$. Being $Q(X) \geq 0$ for any $X \in \mathcal{D}_B$, we can conclude that $Q \in \mathcal{P}_{\mathcal{D}_B}$.

To prove the other inclusion $\mathcal{P}_{\mathcal{D}_B} \subseteq (\mathcal{P}_{\mathcal{D}})_{|B}$, first consider that both are credal sets with probabilities which are 0 outside of B, then the inclusion can be obtained if we show that any linear restriction $P(X) \geq 0$ verified by probabilities in $(\mathcal{P}_{\mathcal{D}})_{|B}$ with $X(\omega) = 0, \forall \omega \in \Omega - B$, it is also verified by probabilities in $\mathcal{P}_{\mathcal{D}_B}$.

Assume that $X(\omega) = 0, \forall \omega \in \Omega - B$ and that $P(X) \geq 0, \quad \forall P \in (\mathcal{P}_{\mathcal{D}})_{|B}$. Then, we have that $Q(X|B) \geq 0, \quad \forall Q \in \mathcal{P}_{\mathcal{D}}$, with $Q(B) > 0$. As, $X(\omega) = 0, \forall \omega \in \Omega - B$, then $Q(X|B) = Q(XB)/Q(B) \geq 0, \quad \forall Q \in \mathcal{P}_{\mathcal{D}}, Q(B) > 0$. As, the inequality is trivially verified if $Q(B) = 0$, then we have that $Q(XB) \geq 0, \quad \forall Q \in \mathcal{P}_{\mathcal{D}}$.

If we add an amount ϵ to the gamble we obtain a desirable gamble: $XB + \epsilon \in \mathcal{D}, \quad \forall \epsilon > 0$, and by condition (4) we have that $XB + \epsilon B \in \mathcal{D}, \quad \forall \epsilon > 0$. By the definition of $\mathcal{D}_B$, we obtain that $XB + \epsilon B \in \mathcal{D}_B, \quad \forall \epsilon > 0$. This implies that $P(XB + \epsilon B) \geq 0, \quad \forall \epsilon > 0, \forall P \in \mathcal{P}_{\mathcal{D}_B}$ and therefore $P(XB) \geq 0, \forall P \in \mathcal{P}_{\mathcal{D}_B}$. As

$XB = X$, then the inequality $P(X) \geq 0$ is also verified by probabilities $P \in \mathcal{P}_{\mathcal{D}_B}$. $\square$

According to Lemmas 1 and 2 and Theorem 3, we derive the following corollary:

Corollary 4 *Let $\mathcal{D} \subset \mathcal{L}$ be a coherent set of desirable gambles and let B be an arbitrary subset of the universe Ω. Let us assume that one of the following conditions holds:*

1. $\underline{P}(B) > 0$.
2. $\overline{P}(B) > 0$ *and* $\mathcal{D} \subset \mathcal{L}$ *satisfies the restriction considered in Equation (3).*

Then:

$$\mathcal{P}_{\mathcal{D}_B} = (\mathcal{P}_{\mathcal{D}})_{|B}.$$

Remark 3.1 *When $\underline{P}(B) > 0$ the set of linear previsions $(\mathcal{P}_{\mathcal{D}})_{|B}$ can be written as follows:*

$$(\mathcal{P}_{\mathcal{D}})_{|B} = \{P(\cdot|B) : P \in \mathcal{P}_{\mathcal{D}}\},$$

since then condition $P(B) > 0$ is redundant.

This corollary represents the main result in this paper. First it shows the known fact that when $\underline{P}(B) > 0$, conditioning (in terms of credal sets) can be done by conditioning all the probability measures. The second thing is relative to conditioning when $\underline{P}(B) = 0$, but $\overline{P}(B) > 0$, in this case conditioning is not determined when we look at the associated credal set, but if we assume condition (3), then conditioning can be obtained in the associated credal set by conditioning all the probabilities with $P(B) > 0$, this conditioning was called *regular extension*. Condition (3) can be seen as a weaker version of Axiom D5, as here an almost desirable gamble X is also desirable when $-X$ is not almost desirable. If X and $-X$ are both almost desirable and we were accepting both of them as desirable, then we would obtain that the null gamble is desirable, and then the associated set would not be coherent. But, if X is almost desirable, but $-X$ is not, then it could be considered that we have some reasons to assume that X is desirable. Here, we have shown that this implies regular conditioning.

Example 2 *Assume that we have the credal set of Figure 2 and that we want to compute its conditional credal set to $B = \{\omega_1, \omega_2\}$ and that the points P with $P(B) = 1$ are the triangle base. In this case conditional gambles are those gambles X such that $X(\omega_3) = 0$ and the associated linear restrictions pass through the vertex opposite to the triangle base. When the credal set does not contain this vertex ($\underline{P}(B) > 0$),*

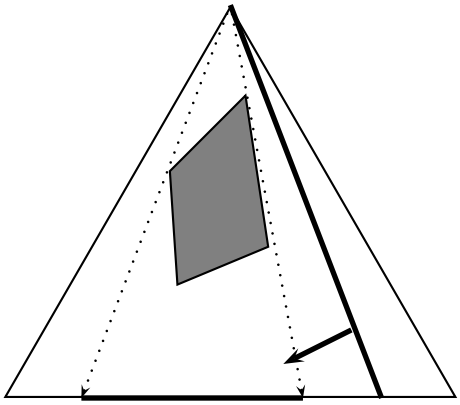


Figure 2: Conditional Gambles

then there are desirable conditional gambles that determine that the conditional credal set is the thick segment represented in the basis and that is equal to the projection of all the probabilities in the credal set from the upper vertex (the projection of a probability P is its conditional probability $P(.|B)$). In other words, the set linear restrictions associated to the conditional gambles (passing through the upper vertex) that are strictly desirable (all the probabilities verify them and are not touching the credal set) as the one in the figure are enough to restrict the set on conditional probabilities to the segment in the figure.

However, when $\underline{P}(B) = 0$, then the upper vertex is in the credal set, as in Figure 3, and all the conditional gambles as the one depicted in the figure are touching the border of the credal set, and therefore their desirability is not determined by the credal set. The set of conditional desirable gambles could contain only the trivial gambles and then the conditional credal set is the full base (the natural extension *of the* generalized Bayes rule *[6]) or it could be a more restrictive one and include all the gambles with linear restrictions verified by the probabilities in the segment $\overline{AC}$ (the smallest possible conditional credal set: the regular extension).*

4 Introduction to Representation and Algorithms

A very important issue to make desirable gambles useful in practice is to determine an effective method to represent information and to develop algorithms able of working with this representation. In particular we would like to have procedures that have as input a set of gambles $\mathcal{F}$ and are able of carrying out the following basic reasoning tasks:

1. to determine whether the natural extension $\overline{\mathcal{F}}$ is coherent (i.e. $0 \notin \overline{\mathcal{F}}$),

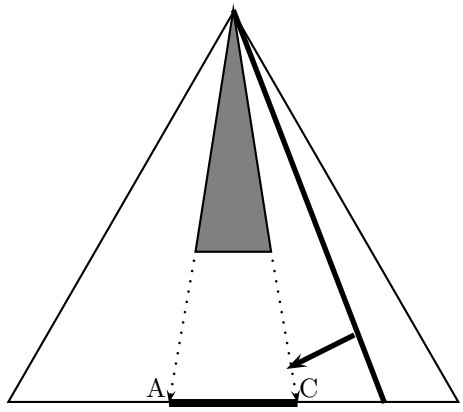


Figure 3: Conditional Gambles, $\underline{P}(B) = 0$

2. given X, to determine whether $X \in \overline{\mathcal{F}}$,
3. given X and $B \subset \Omega$, to compute $\underline{P}(X|B)$ and $\overline{P}(X|B)$ under $\overline{\mathcal{F}}$ when this set is coherent.

The second question is immediate to answer if we can solve the first one, as the following theorem shows.

Theorem 5 *If $\mathcal{F}$ is an arbitrary set of gambles such that $\overline{\mathcal{F}}$ is coherent, then $X \in \overline{\mathcal{F}}$ if and only if $\overline{\mathcal{F} \cup \{-X\}}$ is not coherent.*

Proof: If $X \in \overline{\mathcal{F}}$, then $X, -X \in \overline{\mathcal{F} \cup \{-X\}}$, and $X - X = 0 \in \overline{\mathcal{F} \cup \{-X\}}$. So this set is not coherent.

On the other hand, if $\overline{\mathcal{F} \cup \{-X\}}$ is not coherent, then $0 \in \overline{\mathcal{F} \cup \{-X\}}$. This set is equal to all the gambles $Y = \alpha.Z - \beta.X$, where $Z \in \overline{\mathcal{F}}$ and $\alpha, \beta \geq 0, \alpha > 0$ or $\beta > 0$. In particular, there must be α, β such that $0 = \alpha.Z - \beta.X$, where $Z \in \overline{\mathcal{F}}$. As $\overline{\mathcal{F}}$ is coherent, $\beta \neq 0$, and we obtain $X = \frac{\alpha}{\beta} Z$, and by Axiom D3, $X \in \overline{\mathcal{F}}$. □

A coherent set of gambles $\mathcal{D}$ contains infinite gambles. If we want to represent them in a computer in order to manipulate them by means of algorithms, we need to determine a procedure to represent a coherent set of gambles $\mathcal{D}$ by means of a set $\mathcal{F}$ such that $\mathcal{D} = \overline{\mathcal{F}}$, and for representing the set $\mathcal{F}$ in some formal language. A basic issue is: to determine the type of sets $\mathcal{F}$ we are going to consider and the representation we are going to use. For sets of almost desirable gambles, we can start with a finite set of gambles $\mathcal{F}$ (which can be represented by enumerating the gambles in the set $\mathcal{F}$). This could also be done with sets of desirable gambles, but the capabilities of representation would be too limited, as the following example shows.

Example 3 *Assume that we know that $\overline{P}(B) = 0$, then the only possible set of desirable gambles representing this fact, should include all the gambles $\epsilon - B$ for any $\epsilon > 0$, but not the gamble in the limit $-B$. If $B \neq \Omega$, then $-B$ can be almost desirable without giving rise to an incoherent set. So this fact can be represented with a finite set of almost desirable gambles, but not with a finite set of desirable gambles.*

If we start with a finite set of gambles and compute its natural extension then some of the basic pieces of information can not be represented. In this paper, we want to point out a representation scheme which is not general enough for all the sets of desirable gambles, but which is enough for some of the most usual types of information and for which there are efficient algorithms in the literature.

Definition 1 *A* basic set of gambles *is a set of gambles $\mathcal{F}_{X,B} = \{X + \epsilon B : \epsilon > 0\}$, where X is an arbitrary gamble and $B \subseteq \Omega$. This set of gambles will be denoted as (X, B).*

When $B = \emptyset$, we have a single gamble, X. Otherwise, (X, B) is an infinite set with X in the limit.

The representation we propose is based in considering sets $\mathcal{F}$ given by the union of a finite family of basic sets of gambles: $(X_1, B_1), \ldots, (X_k, B_k)$.

With this system, $\underline{P}(X|B) = c$ is represented by means of $((X - c)B, B)$, i.e. in frame B, we are ready to pay $c - \epsilon$ to get reward $X(\omega)$, for any $\epsilon > 0$. $\overline{P}(X|B) = c$ is represented by means of $((c - X)B, B)$.

Coherence of the set of gambles generated by a finite set of basic gambles, $(X_1, B_1), \ldots, (X_k, B_k)$[1], is equivalent to the fact that the 0 gamble is not in the set of gambles generated by these gambles, which can be checked by showing that the following system in λ_i and ϵ has no solution:

$$\begin{array}{l} \sum_{i=1}^{k} \lambda_i (X_i + \epsilon B_i) \leq 0 \\ \lambda_i \geq 0, \quad \epsilon > 0 \end{array}$$

This is due to the fact that the set of gambles $\sum_{i=1}^{k} \lambda_i (X_i + \epsilon B_i)$ where $\lambda_i \geq 0, \quad \epsilon > 0$ is the set of gambles generated by the finite set of basic gambles by applying Axioms D3 and D4. So, we are checking whether the null gamble is contained in the natural extension.

An algorithm to solve this system is given by Walley, Pelessoni, and Vicig [8]. They start with a set of lower previsions of events, but they finally arrive to a system of this form, and propose an efficient algorithm to solve it, based on the resolution of a sequence of linear programming problems.

[1] We are considering coherence of the generated set of gambles and not the usual notion of coherence for conditional previsions which implies that none of the initial statements is strictly redundant.

To compute $\underline{P}(X|B)$ it is necessary to solve the following optimization problem (we are computing the supremum value of α such that $(X-\alpha)B$ is desirable in the natural extension of the basic gambles:

$$\begin{array}{l}\sup \alpha \\ s.t. \\ \sum_{i=1}^{k} \lambda_i (X_i + \epsilon B_i) \leq (X-\alpha)B \\ \epsilon > 0, \lambda_i \geq 0\end{array}$$

This paper [8] also proposes algorithms to solve an optimization problem completely analogous to this one than can be easily adapted.

A basic question is whether there are simple sets of gambles which can not be covered with this representation. The following example shows a simple case in which there is no obvious solution by using this representation.

Example 4 *Consider $\Omega = \{\omega_1, \omega_2\}$ and the two gambles X, Y given by $X(\omega_1) = 1$, $X(\omega_2) = -1$ and $Y(\omega_1) = -1$, $Y(\omega_2) = 1$. Consider the set of gambles $\mathcal{F}$ given by $\epsilon_1 X + \epsilon_2 Y$, where $\epsilon_1, \epsilon_2 > 0$. This set of gambles is not coherent, as $X + Y = 0$ belongs to it. However, if we start with any representation $(X, B_1), (Y, B_2)$, then either $B_1 = B_2 = \emptyset$ with which we are adding X and Y to the set $\mathcal{F}$ (they were not initially in $\mathcal{F}$ as they can not be expressed as $\epsilon_1 X + \epsilon_2 Y$ with $\epsilon_1, \epsilon_2 > 0$) or if one of them, B_1 or B_2, is not empty, then the set generated by $(X, B_1), (Y, B_2)$ is coherent.*

The solution could be to start with more complex representations as $(X_1, \ldots, X_k)$ representing all the gambles $Z = \sum_{i=1}^{k} \epsilon_i X_i$ where $\epsilon_i > 0$, that we will call an *open set of gambles*. And then to work with sets of gambles which are generated by a finite family of open sets of gambles. However, the development of algorithms for coherence and inference is something to be done in the future, though it does not seem to be simple task.

5 Maximal Sets of Gambles

In this section we will investigate maximal coherent sets of gambles. These sets of gambles represent a *complete* uncertain knowledge: adding a single more gamble will give rise to an incoherent set. The associated credal sets will be linear previsions (probability measures). But, we will also be able of associating finite sequences of probability measures similar to the ones considered by Coletti and Scozzafava [2].

Definition 2 *We will say that a set of gambles $\mathcal{D}$ is* maximal *if it is coherent and there does not exist any $X \notin \mathcal{D}$ such that $\overline{\mathcal{D} \cup \{X\}}$ is coherent.*

Lemma 6 *If $\mathcal{D}$ is coherent and $-X \notin \mathcal{D}$, $X \neq 0$, then $\overline{\mathcal{D} \cup \{X\}}$ is coherent.*

Proof: Let us check it by contradiction. Let us suppose that $\overline{D \cup \{X\}}$ is not coherent. Then there exists a collection of non-negative numbers $c_1, \ldots, c_n, c_{n+1}$ such that $\sum_{i=1}^{n} c_i X_i + c_{n+1} X = 0$, where some of the c_i's is non zero. Furthermore, according to the coherence of $\mathcal{D}$, $c_{n+1} \neq 0$. And, as $X \neq 0$, some of the $c_i, i = 1, \ldots, n$ is also different from 0. Thus, $-X$ can be written as follows: $-X = \sum_{i=1}^{n} \frac{c_i}{c_{n+1}} X_i$. Then, by the coherence of $\mathcal{D}$, $-X$ belongs to it, and we get a contradiction. □

Theorem 7 *A coherent set of gambles $\mathcal{D}$ is maximal if and only if $X \in \mathcal{D}$ xor $-X \in \mathcal{D}$, for all $X \in \mathcal{L}$, $X \neq 0$.*

Proof: Let us suppose that $\mathcal{D}$ is maximal and $X \notin \mathcal{D}$. Then, by definition, $\overline{(\mathcal{D} \cup \{X\})}$ is not coherent. Thus, according to Lemma 6, $-X$ must belong to $\mathcal{D}$. On the other hand, if for any $X \in \mathcal{L}$, $X \in \mathcal{D}$ xor $-X \in \mathcal{D}$, then $\mathcal{D}$ is maximal, as if $X \notin \mathcal{D}$, then $-X \in \mathcal{D}$ and $\overline{\mathcal{D} \cup \{X\}}$ can not be coherent. □

Lemma 8 *If $\mathcal{D}$ is maximal then $\mathcal{D}_B$ is maximal for all $B \subseteq \Omega$, $B \neq \emptyset$.*

Proof: It is trivially derived from Theorem 7. □

Lemma 9 *Let $\mathcal{D}$ be a maximal set of gambles and let $\underline{P}$ and $\overline{P}$ be respectively the lower and the upper previsions associated to it. Then $\underline{P}(B) = \overline{P}(B)$, $\forall\, B \subseteq \Omega$.*

Proof: Let us prove it by contradiction. Let us suppose that there exists some $B \subseteq \Omega$ such that $\underline{P}(B) < \overline{P}(B)$. Then, for all $p \in (\underline{P}(B), \overline{P}(B))$, $B - p \notin \mathcal{D}$ and $-(B-p) = p - B \notin \mathcal{D}$. According to Theorem 7, $\mathcal{D}$ cannot be maximal. □

If we have a sequence of nested sets $\Omega = C_0 \supset C_1 \supset \cdots \supset C_n = \emptyset$, and $B \subseteq \Omega$, then the *layer* of B with respect to this sequence, will be the minimum value of i such that $B \cap (C_i \setminus C_{i+1}) \neq \emptyset$. It will be denoted by layer(B).

Theorem 10 *If $\mathcal{D}$ is maximal then there is a sequence of nested sets $\Omega = C_0 \supset C_1 \supset \cdots \supset C_n = \emptyset$ and a sequence of probability measures $P_0, \ldots, P_{n-1}$ satisfying the following conditions:*

1. *for each probability P_i, $P_i(C_i \setminus C_{i+1}) = 1$, $P_i(\omega) > 0$ for any $\omega \in C_i \setminus C_{i+1}$,*
2. *for each $A \subseteq B \subseteq \Omega$, if $i = \text{layer}(B)$, then $\underline{P}(A|B) = \overline{P}(A|B) = P_i(A|B)$, where $\underline{P}(A|B)$ and $\overline{P}(A|B)$ are the lower and upper probabilities computed from $\mathcal{D}_B$.*

Proof: According to Lemma 9, the lower and the upper probabilities associated to $\mathcal{D}$ do coincide. In other words, the class $\mathcal{P}_{\mathcal{D}}$ is a singleton (it is determined by an additive probability measure P on Ω). Let $C_1 \subsetneq \Omega$ be the subset of elements of probability 0 $C_1 = \{\omega \in \Omega : P(\{\omega\}) = 0\}$. If $C_1 \neq \emptyset$, according to Lemma 8, $\mathcal{D}_{C_1}$ is maximal. Based again on Lemma 9, it induces a probability measure on C_1, P_1. We can repeat the same process again and get a strictly decreasing finite sequence of nonempty sets C_i and an associated finite family of probability measures P_i. (Note that, after a finite sequence of n steps, the set C_n will be the empty set and the process is finished.)

On the other hand, if $A \subseteq B \subseteq \Omega$ and i is the layer of B, then we have that $B \subseteq C_i$ and $P_i(B) > 0$ (remember that C_{i+1} is the subset of C_i given by the $\omega \in C_i$ such that $P_i(\omega) = 0$). Probability P_i is defined in C_i and as it is associated to $\mathcal{D}_{C_i}$ and this set is maximal, we have that $\underline{P}(E|C_i) = \overline{P}(E|C_i) = P_i(E)$ for any $E \subseteq C_i$. As $P_i(B) > 0$, then its lower conditional probability is greater than 0, and by Corollary 4, the conditional probability can be computed by conditioning in the associated credal set, obtaining the desired result:

$$P_i(A|B) = \underline{P}(E|C_i \cap B) = \overline{P}(A|C_i \cap B) = \underline{P}(A|B) = \overline{P}(A|B)$$

□

This theorem shows the great similarity between maximal coherent gambles and the sequence of probabilities associated to a coherent set of conditional assessments given by Coletti and Scozzafava [2]. The layer of B is also the minimum value of i for which $P_i(B) > 0$, and therefore is the equivalent concept to the *zero layer* of B proposed by these authors. However, there are some differences between the two models as we will show later: we can have the same sequence of probabilities associated to different maximal coherent sets of desirable gambles.

In the following we show that any coherent set of gambles is the intersection of a family of coherent maximal sets of gambles. First, we need a technical result.

Lemma 11 *If $\mathcal{D}$ is coherent and $-X, X \notin \mathcal{D}$ and $X \neq 0$, then $\mathcal{D}^{+X} = \overline{(\mathcal{D} \cup \{X\} \cup \{-X + Y : Y \in \mathcal{D})}$ is coherent.*

Proof: If this set is not coherent, then we have that there are $\alpha_1, \alpha_2, \alpha_3 \geq 0$, and $Y_1, Y_2 \in \mathcal{D}$ such that $\alpha_1 Y_1 + \alpha_2 X + \alpha_3(-X + Y_2) \leq 0$, and at least one of the α_i is not equal to 0.

From this inequality we have that: $\alpha_1 Y_1 + \alpha_3 Y_2 \leq (\alpha_3 - \alpha_2)X$.

First, notice that α_1, α_3 can not be both equal to 0, because otherwise, $0 \leq (-\alpha_2)X$, and as $X \neq 0$ and $\alpha_2 \neq 0$, we have that $-X \in \mathcal{D}$, which is in contradiction with the fact that $-X, X \notin \mathcal{D}$.

Then, at least one of the values α_1, α_3 is different from 0, and thus $\alpha_1 Y_1 + \alpha_3 Y_2 \in \mathcal{D}$.

Three situations are now possible:

- $\alpha_3 = \alpha_2$, which is in contradiction with the fact that $\mathcal{D}$ is coherent.
- $(\alpha_3 - \alpha_2) > 0$, which is in contradiction with the fact that $\mathcal{D}$ is coherent and $X \notin \mathcal{D}$.
- $(\alpha_2 - \alpha_3) > 0$, which is in contradiction with the fact that $\mathcal{D}$ is coherent and $-X \notin \mathcal{D}$.

In any case, we arrive to a contradiction, so $\mathcal{D}^{+X}$ must be coherent. □

Theorem 12 *Let $\mathcal{D}$ be a coherent set of gambles. Then, there exists at least one maximal set of gambles containing it.*

Proof:

Let us start with a coherent set and then, repeat the following process:

1. If for any gamble X ($X \neq 0$), we have that $X \in \mathcal{D}$ or $-X \in \mathcal{D}$, then $\mathcal{D}$ is maximal and the procedure stops.
2. Select a gamble X such that $-X, X \notin \mathcal{D}$ and $X \neq 0$.
3. Transform $\mathcal{D}$ by making it equal to $\mathcal{D}^{+X}$. By Lemma 11, this new set is coherent and contains to the old $\mathcal{D}$.
4. Go to step 1.

The main point of this procedure is that it arrives to a maximal coherent set after a finite number of steps. This result is based on the fact that if in the first $k+1$ loops of this process we select gambles $X_1, X_2, \ldots, X_{k+1}$, then these gambles are linearly independent. This fact is obtained by proving that after having added $X_1, \ldots, X_k$, then for any linear combination of these gambles $Y = \sum_{i=1}^{k} \alpha_i X_i$, and $Y \neq 0$, we have that either $Y \in \mathcal{D}$ or $-Y \in \mathcal{D}$. So, in step 2, we have to select a gamble which is linearly independent of the previously selected ones.

This is going to be proved by induction in k. For $k = 1$, $Y = \alpha_1 X_1$. Then if $\alpha_1 > 0$, $Y \in \mathcal{D}$, and if $\alpha_1 < 0$, then $-Y \in \mathcal{D}$. α_1 can not be equal to 0 because $Y \neq 0$.

Now, assume that it is true for the first k gambles $X_1, \ldots, X_k$, and let us prove it for $X_1, X_2, \ldots, X_{k+1}$.

Assume, $Y = \sum_{i=1}^{k+1} \alpha_i X_i$. Let us denote by $Z = \sum_{i=1}^{k} \alpha_i X_i$.

If $\alpha_i = 0$ for all $1 \leq i \leq k$, then $Y = \alpha_{k+1} X_{k+1}$ and we are in a situation similar to the case $k = 1$.

If some α_i with $i \leq k$ is different from 0, then by induction, we have that either Z or $-Z$ is in $\mathcal{D}$, after adding $X_1, \ldots, X_k$.

We have that $Y = Z + \alpha_{k+1} X_{k+1}$. The following situations are possible:

- $\alpha_{k+1} = 0$, then $Y = Z$ and we have that either Y or $-Y$ is in $\mathcal{D}$.
- $\alpha_{k+1} > 0$ and $Z \in \mathcal{D}$, then by coherence $Y \in \mathcal{D}$.
- $\alpha_{k+1} > 0$ and $-Z \in \mathcal{D}$, then $-Y = -Z - \alpha_{k+1} X_{k+1}$, and by the way we compute $\mathcal{D}^{+X_{k+1}}$ in which we add any gamble $-X_{k+1} + U$, and therefore any gamble $-\alpha_{k+1} X_{k+1} + U$ where $U \in \mathcal{D}$, we have that $-Y \in \mathcal{D}$.
- $\alpha_{k+1} < 0$ and $-Z \in \mathcal{D}$, then by coherence $-Y \in \mathcal{D}$.
- $\alpha_{k+1} < 0$ and $Z \in \mathcal{D}$, then $Y = \alpha_{k+1} X_{k+1} + Z \in \mathcal{D}$ after replacing $\mathcal{D}$ by $\mathcal{D}^{+X_{k+1}}$.

As, we always choose a gamble that is linearly independent of the previous one, and Ω being finite, the dimension of $\mathcal{L}$ as a linear space is finite, and so the process has to stop after a finite number of steps. □

Theorem 13 *Let $\mathcal{D}$ be a coherent set of gambles. Then $\mathcal{D} = \cap_{i \in I} \mathcal{D}_i$, where $\{\mathcal{D}_i : i \in I\}$ is the class of maximal sets of gambles containing $\mathcal{D}$.*

Proof: We only have to check the inclusion $\cap_{i \in I} \mathcal{D}_i \subseteq \mathcal{D}$. We will prove it by contradiction. Let us suppose that $X \in \cap_{i \in I} \mathcal{D}_i \setminus \mathcal{D}$. Then, by Lemma 6, $\overline{\mathcal{D} \cup \{-X\}}$ is coherent. By Theorem 12 there exists at least one maximal set of gambles containing $\overline{\mathcal{D} \cup \{-X\}}$. This maximal set coincides with one of the $\mathcal{D}_i$, for some $i \in I$. Then there exists some $i \in I$ such that $-X \in \mathcal{D}_i$. It contradicts the assumption of coherence of $\mathcal{D}_i$. □

This theorem can be the basis to obtain a representation of gambles analogous to the credal sets for sets of almost desirable gambles. Now, a coherent set of gambles can be expressed as a family of maximally coherent sets of gambles, each one of them has an associated sequence of probability measures. There are important problems to be solved. One of them is that a maximally consistent coherent set of gambles is not exactly equivalent to a sequence of probability measures as the following example shows.

Example 5 *Assume that $\Omega = \{\omega_1, \omega_2\}$ and the probability given by $P_0(\omega_1) = P_0(\omega_2) = 0.5$ (only one probability in the sequence). It is clear that any gamble with $X(\omega_1) + X(\omega_2) > 0$ should be desirable. But, this probability does not determine whether the gamble $Y(\omega_1) = 1, Y(\omega_2) = -1$ is desirable. We can have a coherent set in which neither Y nor $-Y$ is desirable, another coherent set in which Y is desirable, and other one in which $-Y$ is desirable. Only the two last ones are maximal.*

An alternative model that allows to establish a correspondence between maximally coherent sets of gambles and sequences of probability measures is obtained by making the consistency Axiom D1 stronger, modifying it to the following version:

D1". If $X \in \mathcal{D}$, then there is $\epsilon > 0$, such that $-X + \epsilon\,\mathrm{supp}(X) \notin \mathcal{D}$.

where $\mathrm{supp}(X)$, the support of X, is the set of $\omega \in \Omega$ such that $X(\omega) \neq 0$.

This consistency condition is stronger than Axiom D1, as this axiom was assuming that we can not have X and $-X$ as desirable. D1" implies that the null gamble is not desirable. This axiom says that we can not have X as desirable if $-X$ is the limit of desirable gambles with the same support. This is a kind of a minimum of separation between X and $-Y$, if both X and Y are desirable and have the same support. The support is necessary in the condition, as if we had only considered $-X + \epsilon$ (as in strict desirability axioms [6, Section 3.7.8]), then it would have become a too strong condition. Imagine that $P_0(B) = 0$, then as $P_0(B) = 0$ we have that $-B + \epsilon$ is also desirable for any $\epsilon > 0$. But we have that $B \in \mathcal{D}$. So, the separation condition without considering the support would not be fulfilled.

The following theorem shows that a sequence of probability measures as the one generated in Theorem 10 can always be represented by means of a maximally coherent set of gambles among those satisfying D1" and that this set is unique.

Theorem 14 *If we have a sequence of nested sets $\Omega = C_0 \supset C_1 \supset \cdots \supset C_n = \emptyset$ and a sequence of probability measures $P_0, \ldots, P_{n-1}$ satisfying condition:*

- *for any i, $P_i(C_i \setminus C_{i+1}) = 1$, and $P_i(\omega) > 0, \quad \forall \omega \in C_i \setminus C_{i+1}$,*

then the set of gambles $\mathcal{D} = \{X : P_i(X) > 0,$ where $i = \mathrm{layer}(\mathrm{supp}(X))\}$ is the only maximally

coherent set of desirable gambles among those satisfying Axiom D1" and that for any i the credal set associated to $\mathcal{D}_{C_i}$ contains only probability measure P_i.

Proof: First, it is easy to prove that this set of gambles satisfies all the axioms for coherence including Axiom D1". Considering $P_i(X) > 0$, we have that $P_i(\text{supp}(X)) > 0$, $P_i(-X + \epsilon\,\text{supp}(X)) = -P_i(X) + \epsilon < 0$ if we choose $\epsilon > 0$ small enough. Therefore, there is an $\epsilon > 0$ such that $-X + \epsilon\,\text{supp}(X) \notin \mathcal{D}$.

It is immediate to prove that P_i is the credal set associated to $\mathcal{D}_{C_i}$, as there can not be a probability Q different to P_i defined on C_i for which $Q(X) > 0$ for all $X \in \mathcal{D}_{C_i}$.

On the other hand, this set is unique: if $\mathcal{D}'$ is such that for any i the credal set associated to $\mathcal{D}'_{C_i}$ is the probability P_i, then, if X is a gamble and $i = \text{layer}(\text{supp}(X))$, then $\text{supp}(X) \subseteq C_i$ and:

- if X is such that $P_i(X) > 0$, then $X \in \mathcal{D}'$,
- if X is such that $P_i(X) < 0$, then $X \notin \mathcal{D}'$,
- if X is such that $P_i(X) = 0$, then as there is $\omega \in C_i \setminus C_{i+1}$ such that $X(\omega) > 0$ and $P_i(\omega) > 0$, we have that for any $\epsilon > 0$, $P_i(-X+\epsilon\,\text{supp}(X)) = P_i(-X) + \epsilon = \epsilon > 0$ then we have that $-X + \epsilon\,\text{supp}(X) \in \mathcal{D}'_{C_i} \subseteq \mathcal{D}'$. By Axiom D1", $X \notin \mathcal{D}'$.

As a consequence, $\mathcal{D}'$ obeys the same criteria than $\mathcal{D}$ to determine whether a gamble belongs to it ($P_i(X) > 0$) and thus $\mathcal{D} = \mathcal{D}'$.

The fact that $\mathcal{D}$ is maximal is a consequence of the uniqueness. □

6 Conclusions

In this paper we have presented some basic concepts under the light of desirability. We have tried to show that this approach can shed light on some important notions in imprecise probability, such as conditioning. It can be useful for showing the relationships with other approaches, such as as probabilistic coherence [2]. There is important work to do, mainly on the algorithmic side. Here we have shown some existing algorithms which can be directly applied to some restricted forms of coherent sets, but it is necessary to determine whether this restriction is too severe to leave out some important real situations. Also, it would be interesting to determine some extra axioms under which the representation based on what we have called basic sets of gambles is enough to cover any possible set of desirable gambles.

As we have mentioned, Moral [4] studied the concept of epistemic irrelevance and independence taking desirability as basis. An important problem for the future is how to use graphical models to represent and use epistemic independence assessments in the computation of conditional sets of desirable gambles.

Acknowledgments

This work has been jointly supported by the Spanish Ministry of Education and Science under project TIN2007-67418-C03-03, by European Regional Development Fund (FEDER), and by the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018). We are grateful to the two anonymous reviewers of this paper for ISIPTA' 09 conference who helped us to correct and improve it.

References

[1] R.J. Buheler, Coherence Preferences, *The Annals of Statistics* 4, 1976, 1051–1064.

[2] G. Coletti and R. Scozzafava, *Probabilistic Logic in a Coherent Setting*, Kluwer Academic Publishers (Dordrecht, 2002).

[3] F.J. Girón and S. Ríos, Quasi-Bayesian behaviour: A more realistic approach to decision making? In: *Bayesian Statistics* (J.M. Bernardo, J.H. DeGroot, D.V. Lindley, and A.F.M. Smith, eds.) University Press (Valencia, 1980) 17–38.

[4] S. Moral, Epistemic Irrelevance on Sets of Desirable Gambles, *Annals of Mathematics and Artificial Intelligence* 45, 2005, 197–214.

[5] S. Moral, N. Wilson, A logical view of probability, *ECAI-94 Proceedings* (A. Cohn, ed.) John Wiley (Chichester, 1994) 386–390.

[6] P. Walley, *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall (London, 1991).

[7] P. Walley, Towards a unified theory of imprecise probability, *International Journal of Approximate Reasoning*, 24, 2000, 125–148.

[8] P. Walley, R. Pelessoni, P. Vicig, Direct algorithms for checking consistency and making inferences from conditional probability assessments, *Journal of Statistical Planning and Inference* 126, 2004, 119–151.

[9] P.M. Williams, Coherence, strict coherence and zero probabilities, *Proceedings of the Fifth International Congress of Logic, Methodology and Philosophy of Science, VI*, 1975, 29–33.

Concentration Inequalities and Laws of Large Numbers under Epistemic Irrelevance

Fabio Gagliardi Cozman
Escola Politécnica, Universidade de São Paulo - São Paulo, SP - Brazil
fgcozman@usp.br

Abstract

This paper presents concentration inequalities and laws of large numbers under weak assumptions of irrelevance, expressed through lower and upper expectations. The results are variants and extensions of De Cooman and Miranda's recent inequalities and laws of large numbers. The proofs indicate connections between concepts of irrelevance for lower/upper expectations and the standard theory of martingales.

1 Introduction

This paper investigates concentration inequalities and laws of large numbers under weak assumptions of "irrelevance" that are expressed using lower and upper expectations. The starting point is the assumption that, given bounded variables $X_1, \ldots, X_n$, we have:

$$\begin{array}{c}\text{for each } i \in [2,n], \text{ variables } X_1, \ldots, X_{i-1} \\ \text{are epistemically irrelevant to } X_i.\end{array} \quad (1)$$

Epistemic irrelevance of variables $X_1, \ldots, X_{i-1}$ to X_i obtains when [26, Def. 9.2.1]

$$\overline{E}[f(X_i)|A(X_{1:i-1})] = \overline{E}[f(X_i)] \quad (2)$$

for any bounded function f of X_i and any nonempty event $A(X_{1:i-1})$ defined by variables $X_{1:i-1}$, where the functional $\overline{E}$ is an *upper expectation* (Section 2). Here and in the remainder of the paper we simplify notation by using $X_{1:i}$ for $X_1, \ldots, X_i$.

A judgement of epistemic irrelevance can be interpreted as a relaxed judgement of stochastic independence, perhaps motivated by a robustness analysis or by disagreements amongst a set of decision makers. Alternatively, one might consider epistemic irrelevance as *the* appropriate concept of independence when expectations are not known precisely.

De Cooman and Miranda have recently proven a number of inequalities and laws of large numbers that also deal with judgements of irrelevance expressed through lower/upper expectations [5]. De Cooman and Miranda's weak law of large numbers implies that, given Assumption (1), for any $\epsilon > 0$,

$$\underline{P}\left(\underline{\mu}_n - \epsilon \leq \frac{\sum_{i=1}^n X_i}{n} \leq \overline{\mu}_n + \epsilon\right) \geq 1 - 2e^{-\frac{n\epsilon^2/4}{(\max_i B_i)^2}},$$

where B_i is such that $|X_i| \leq B_i$, and

$$\underline{\mu}_n \doteq \frac{\sum_{i=1}^n \underline{E}[X_i]}{n}, \qquad \overline{\mu}_n \doteq \frac{\sum_{i=1}^n \overline{E}[X_i]}{n}.$$

Moreover, De Cooman and Miranda's results and Assumption (1) imply a two-part strong law of large numbers: for any $\epsilon > 0$, there is $N \in \mathbf{N}_+$ such that for any $N' \in \mathbf{N}_+$,

$$\overline{P}\left(\exists n \in [N, N+N'] : \frac{\sum_{i=1}^n X_i}{n} \geq \overline{\mu} + \epsilon\right) < \epsilon,$$

$$\overline{P}\left(\exists n \in [N, N+N'] : \frac{\sum_{i=1}^n X_i}{n} \leq \underline{\mu} - \epsilon\right) < \epsilon.$$

This law of large numbers corresponds to a finitary version of the usual strong law of large numbers [9]; the focus on a finitary law is justified by the fact that De Cooman and Miranda do not assume countable additivity. If countable additivity holds, the finitary strong law of large numbers implies convergence of empirical means with probability one [5, Sec. 5.3].

To obtain their results, De Cooman and Miranda assume, following Walley's theory of lower previsions, that all variables are bounded, and that conglomerability (and consequently disintegrability) holds. These assumptions are discussed in more detail later.

The present paper derives laws of large numbers by exploiting concentration and martingale inequalities that are adapted to the setting of lower/upper expectations. These results use either Assumption (1) or the weaker assumption that, for each $i \in [2,n]$ and any nonempty event $A(X_{1:i-1})$,

$$\begin{array}{c}\underline{E}[X_i|A(X_{1:i-1})] = \underline{E}[X_i] \\ \text{and} \\ \overline{E}[X_i|A(X_{1:i-1})] = \overline{E}[X_i].\end{array} \quad (3)$$

Several results for *bounded* variables presented in this paper are basically implied by De Cooman and Miranda's work. Regarding bounded variables our contribution lies in offering tighter inequalities and alternative proof techniques that are more closely related to established methods in standard probability theory (in particular, close to Hoeffding's and Azuma's inequalities). In Section 4 we offer more significant contributions as we lift the assumption of boundedness for variables, and use martingale theory to prove laws of large numbers under elementwise disintegrability.

2 Expectations, disintegrability, and zero probabilities

In this section we present notation and terminology. Throughout the paper we assume that an expectation functional E maps bounded variables into real numbers, and satisfies:
(1) if $\alpha \leq X \leq \beta$, then $\alpha \leq E[X] \leq \beta$;
(2) $E[X+Y] = E[X] + E[Y]$;
where X, Y are bounded variables and α, β are real numbers (inequalities are understood pointwise).

From such an expectation functional, a *finitely additive* probability measure P is induced by $P(A) \doteq E[A]$ for any event A; note that A denotes both the event and its indicator function.[1]

Given a set of expectation functionals, the lower and upper expectations of variable X are respectively

$$\underline{E}[X] = \inf E[X]\,, \qquad \overline{E}[X] = \sup E[X]\,.$$

Lower and upper probabilities are defined similarly using indicator functions. Given an event A, a conditional expectation functional is constrained by $E[X|A]\,P(A) = E[XA]$. If we have a set of expectation functionals, then a set of conditional expectation functionals given an event A is produced by elementwise conditioning on event A (that is, each expectation functional is conditioned on A).

2.1 Disintegrability and factorization

We will employ an assumption of *disintegrability* in our proofs; namely,

$$\overline{E}[W] \leq \overline{E}\big[\overline{E}[W|Z]\big] \tag{4}$$

for any $W \geq 0$, $Z \geq 0$ of interest, where W and Z may stand for sets of (non-negative) variables. Note that disintegrability can fail for a single finitely additive probability measure over an infinite space [6, 10]; that is, there is a finitely additive probability measure P such that

$$E_P[W] > E_P[E_P[W|Z]]\,.$$

One way to obtain disintegrability is to restrict attention to simple variables; that is, variables that take on finitely many distinct values. In particular, indicator functions are simple variables; hence simple variables suffice to express convergence of relative frequencies, and our results apply then.

Another way to obtain disintegrability for every probability measure P is to adopt countable additivity [1]. That is, assume that if

$$A_1 \supset A_2 \supset \ldots$$

is a countable sequence of events, then

$$\cap_i A_i = \emptyset \quad \text{implies} \quad \lim_{n\to\infty} \overline{P}(A_n) = 0. \tag{5}$$

This assumption says that if $\cap_i A_i = \emptyset$, then $\lim_{n\to\infty} P(A_n) = 0$ for every possible probability measure.

A third way to obtain disintegrability is simply to impose it. One may consider disintegrability a "rationality" requirement.

- The theories of coherent behavior by Heath and Sudderty [14] and by Lane and Sudderth [19] follow this path by axiomatizing the *strategic* measures of Dubins and Savage [11], and thus prescribing probability measures that disintegrate appropriately along some predefined partitions. This would be sufficient for our purposes, but there are limitations in the approach as summarized by Kadane et al [16]. The disintegrability of strategic measures has actually been used to prove various laws of large numbers in a finitely additive setting [17].
- Another scheme that imposes disintegrability is Walley's theory of lower previsions; in that theory, Expression (4) is a consequence of axioms for "coherent" behavior. This is the path adopted by De Cooman and Miranda, who consequently have Expression (4) at their disposal.

When disintegrability holds, recursive application of Expression (4) yields: if $f_i(X_i) \geq 0$ for $i \in \{1, \ldots, n\}$, then

$$\overline{E}\left[\prod_{i=1}^{n} f_i(X_i)\right] \leq \overline{E}\left[\ldots \overline{E}\left[\overline{E}\left[\prod_{i=1}^{n} f_i(X_i)|X_{1:n-1}\right]|X_{1:n-2}\right]\ldots\right];$$

[1] A probability measure defined on a field completely characterizes an expectation functional on bounded functions that are measurable with respect to the field and vice-versa [26, Theorem 3.2.2].

Assumption (1) then implies an inequality we use later: for bounded and nonnegative functions,

$$\overline{E}\left[\prod_{i=1}^{n} f_i(X_i)\right] \leq \prod_{i=1}^{n} \overline{E}[f_i(X_i)]\,. \tag{6}$$

2.2 Zero probabilities, full conditional measures and weak irrelevance

It should be noted that the definition of epistemic irrelevance (Expression (2)) does not contain any clause concerning zero probabilities. Indeed, Walley's theory of lower previsions follows de Finetti in adopting *full conditional measures*, and in this setting Expression (2) can be imposed without concerns about zero probabilities. Recall that a full conditional measure $P : \mathcal{B} \times (\mathcal{B}\backslash\emptyset) \to \Re$, where $\mathcal{B}$ is a Boolean algebra, is a set-function that for every nonempty event C satisfies [10, 18]:
(1) $P(C|C) = 1$;
(2) $P(A|C) \geq 0$ for all A;
(3) $P(A \cup B|C) = P(A|C) + P(B|C)$ for all disjoint A and B;
(4) $P(A \cap B|C) = P(A|B \cap C)\, P(B|C)$ for all A and B such that $B \cap C \neq \emptyset$.

Full conditional measures are not adopted in the usual Kolmogorovian theory, and if countable additivity is adopted and conditioning is defined through Radon-Nykodym derivatives, it may be impossible to satisfy the axioms for full conditional measures [23, 24]. Thus there are are some differences between epistemic irrelevance (at least as defined by Walley) and the usual Kolmogorovian set-up, besides the obvious set-valued/point-valued distinction.

Suppose that one wishes to deal with sets of probability measures and associated lower/upper expectations, but chooses to adopt the Kolmogorovian set-up for each measure. That is, each measure satisfies countable additivity and thus disintegrability, and conditioning is left undefined when the conditioning event has probability zero. It might seem reasonable to amend Expression (2) as follows:

$$\begin{aligned} \overline{E}[f(X_i)|A(X_{1:i-1})] &= \overline{E}[f(X_i)] \\ &\quad \text{if } \underline{P}(A(X_{1:i-1})) > 0. \end{aligned} \tag{7}$$

This condition is a natural for theories that do not define conditioning on events of lower probability zero, such as Giron and Rios' theory [13]. Alas, this weaker condition is really too weak to produce laws of large numbers, as the following example shows.

Example 1 *Suppose* $X_1, X_2, \ldots$ *assume values in* $\{0, 1, 2\}$*, and*

$$\underline{P}(X_i = x) = 0,$$

$$\overline{P}(X_i = x) = 1/2$$

for $x \in \{0, 1, 2\}$*. Consequently,* $E[X_i] \in [1/2, 3/2]$*. Suppose additionally that*

$$P(X_i = x | X_{i-1} = x, X_{1:i-2}) = 1$$

for $x \in \{0, 1, 2\}$*; that is, the* i*th variable reproduces the value of the* $(i-1)$*th variable. They are obviously dependent variables. However, all events have lower probability zero, so variables* $X_{1:i-1}$ *would be irrelevant to* X_i *by Expression (7).*

In this example, Expression (6) fails. For instance,

$$\overline{E}[X_1 X_2] = 5/2 > 9/4 = (3/2)(3/2) = \overline{E}[X_1]\,\overline{E}[X_2]\,.$$

Moreover the example illustrates a failure of any sensible law of large numbers, as for any $\epsilon > 0$*,*

$$P\left(1/2 - \epsilon \leq \frac{\sum_{i=1}^{n} X_i}{n} \leq 3/2 + \epsilon\right) \in [0, 1/2],$$

because the inequality inside the probability is only satisfied when $\{X_1 = 1\}$ *obtains.*

We might thus consider an alternative to Expression (7):

$$\begin{aligned} \overline{E}[f(X_i)|A(X_{1:i-1})] &= \overline{E}[f(X_i)] \\ &\quad \text{if } \overline{P}(A(X_{1:i-1})) > 0. \end{aligned} \tag{8}$$

The concept of irrelevance conveyed by Expression (8) does lead to Expression (6). To see this, note that for nonnegative X and Y, we have

$$\begin{aligned} \overline{E}[XY] &\leq \sup_P E_P\left[\overline{E}[XY|Y]\right] \\ &= \sup_P E_P\left[A\overline{E}[XY|Y] + A^c\overline{E}[XY|Y]\right], \end{aligned}$$

using disintegrability and defining A as the set of all values of Y such that $\overline{P}(A^c) = 0$. Hence $P(A^c) = 0$ for every P and using Expression (8):

$$\begin{aligned} \overline{E}[XY] &\leq \sup_P E_P\left[AY\overline{E}[X|Y]\right] \\ &= \sup_P E_P\left[AY\overline{E}[X]\right] \\ &= \sup_P E_P[AY]\,\overline{E}[X] \\ &= \overline{E}[X] \sup_P E_P[Y] \\ &= \overline{E}[X]\,\overline{E}[Y]\,. \end{aligned}$$

[As a digression, note that one might *define* conditional expectations as $\underline{E}[X|A] = \inf_{P:P(A)>0} E_P[X|A]$ and $\overline{E}[X|A] = \sup_{P:P(A)>0} E_P[X|A]$. This form of conditioning has been advocated by several authors [27, 28], and it is quite similar to Walley's concept of regular

extension [26, Ap. J]. For such a form of conditioning, Expression (8) seems to be the natural definition of irrelevance.]

In short, more than one combination of definitions and assumptions lead to the results presented in the remainder of this paper. For instance, Expression (6) obtains when Assumption (1) holds *and* disintegrability holds (because all variables are simple, *or* because countable additivity is assumed, *or* because disintegrability is imposed). Alternatively, Expression (6) obtains when Expression (8) holds for any $i \in [2, n]$, any bounded function f of X_i, and any event $A(X_{1:i-1})$, and additionally disintegrability holds.

Similar remarks concerning zero probabilities can be directed at Assumption (3). We say that *weak irrelevance* obtains when either:

- For any $i \in [2, n]$ and any nonempty event $A(X_{1:i-1})$,

$$\underline{E}[X_i|A(X_{1:i-1})] = \underline{E}[X_i] \quad \text{and} \quad \overline{E}[X_i|A(X_{1:i-1})] = \overline{E}[X_i]$$

  [this is Assumption (3), and it requires full conditional measures].

or:

- For any $i \in [2, n]$ and any event $A(X_{1:i-1})$,

$$\underline{E}[X_i|A(X_{1:i-1})] = \underline{E}[X_i] \text{ if } \overline{P}(A(X_{1:i-1})) > 0 \quad \text{and} \quad \overline{E}[X_i|A(X_{1:i-1})] = \overline{E}[X_i] \text{ if } \overline{P}(A(X_{1:i-1})) > 0.$$

3 Bounded variables

Take variables $X_1, \ldots, X_n$ such that $|X_i| \leq B_i$ and define

$$\gamma_n \doteq \sum_{i=1}^{n} B_i^2 > 0.$$

We start by deriving two concentration inequalities.

3.1 Concentration inequalities

The following inequality is a counterpart of Hoeffding inequality [8, 15] in the context of lower/upper expectations; it is slightly tighter than similar inequalities by De Cooman and Miranda [5]. It is interesting to note that the proof is remarkably similar to the proof of the original Hoeffding inequality.

Theorem 1 *If bounded variables $X_1, \ldots, X_n$ satisfy Expression (6), then if $\gamma_n > 0$,*

$$\overline{P}\left(\sum_{i=1}^{n}(X_i - \overline{E}[X_i]) \geq \epsilon\right) \leq e^{-2\epsilon^2/\gamma_n},$$

$$\overline{P}\left(\sum_{i=1}^{n}(X_i - \underline{E}[X_i]) \leq -\epsilon\right) \leq e^{-2\epsilon^2/\gamma_n}.$$

Proof. By Markov inequality, if $X \geq 0$, then for any $\epsilon > 0$ we have $P(X \geq \epsilon) \leq E[X]/\epsilon$. Consequently, for $s > 0$, any variable X satisfies

$$\overline{P}(X \geq \epsilon) = \overline{P}\left(e^{sX} \geq e^{s\epsilon}\right) \leq e^{-s\epsilon}\overline{E}[\exp(sX)].$$

Using this inequality and Expression (6):

$$\begin{aligned}\overline{P}&\left(\sum_{i=1}^{n}(X_i - \overline{E}[X_i]) \geq \epsilon\right) \\ &\leq e^{-s\epsilon}\overline{E}\left[\exp\left(\sum_{i=1}^{n} s(X_i - \overline{E}[X_i])\right)\right] \\ &\leq e^{-s\epsilon}\prod_{i=1}^{n}\overline{E}\left[\exp\left(s(X_i - \overline{E}[X_i])\right)\right].\end{aligned}$$

We now use Hoeffding's result (Expression (11)) that if variable X satisfies $a \leq X \leq b$ and $E[X] \leq 0$, then $E[\exp(sX)] \leq \exp(s^2(b-a)^2/8)$ for any $s > 0$. Thus for any P, $E_P\left[\exp(s(X_i - \overline{E}[X_i]))\right] \leq \exp(s^2B_i^2/8)$, and then $\overline{E}\left[\exp\left(s(X_i - \overline{E}[X_i])\right)\right] \leq \exp(s^2B_i^2/8)$. Consequently,

$$\overline{P}\left(\sum_{i=1}^{n}(X_i - \overline{E}[X_i]) \geq \epsilon\right) \leq e^{-s\epsilon}e^{s^2\gamma_n/8} \leq e^{-2\epsilon^2/\gamma_n},$$

where the last inequality is obtained by taking $s = 4\epsilon/\gamma_n$. This proves the first inequality in the theorem; the second inequality is proved by taking $\overline{P}\left(\sum_{i=1}^{n}((-X_i) - \overline{E}[-X_i]) \geq \epsilon\right)$ and noting that $\underline{E}[X_i] = -\overline{E}[-X_i]$. □

We now move to weak irrelevance and obtain an analogue of Azuma's inequality [2, 7]. It is again interesting to note that the proof is remarkably similar to the proof of the original Azuma inequality. De Cooman and Miranda [5, Sec. 4.1] show that their inequalities are valid under weak irrelevance; the next inequality is slightly tighter than theirs.

Theorem 2 *If bounded variables $X_1, \ldots, X_n$ satisfy weak irrelevance and disintegrability (Expression (4)) holds, then if $\gamma_n > 0$,*

$$\overline{P}\left(\sum_{i=1}^{n}(X_i - \overline{E}[X_i]) \geq \epsilon\right) \leq e^{-2\epsilon^2/\gamma_n},$$

$$\overline{P}\left(\sum_{i=1}^{n}(X_i - \underline{E}[X_i]) \leq -\epsilon\right) \leq e^{-2\epsilon^2/\gamma_n}.$$

Proof. Using both Markov's inequality (as in the proof of Theorem 1) and disintegrability, for any $s > 0$ we get

$$\begin{aligned}
&\overline{P}\left(\sum_{i=1}^{n}(X_i - \overline{E}[X_i]) \geq \epsilon\right) \\
&\quad\leq e^{-s\epsilon}\overline{E}\left[\exp\left(\sum_{i=1}^{n} s(X_i - \overline{E}[X_i])\right)\right] \\
&\quad\leq e^{-s\epsilon}\overline{E}\left[\overline{E}\left[\exp\left(\sum_{i=1}^{n} s(X_i - \overline{E}[X_i])\right) \mid X_{1:n-1}\right]\right] \\
&\quad\leq e^{-s\epsilon}\overline{E}\left[\exp\left(\sum_{i=1}^{n-1} s(X_i - \overline{E}[X_i])\right) h(X_{1:n-1})\right],
\end{aligned}$$

where

$$h(X_{1:n-1}) = \overline{E}\left[\exp\big(s(X_n - \overline{E}[X_n])\big) \mid X_{1:n-1}\right].$$

Due to weak irrelevance,

$$E_P[X_n|X_{1:n-1}] \leq \overline{E}[X_n|X_{1:n-1}] = \overline{E}[X_n];$$

consequently, for any P,

$$E_P\left[X_n - \overline{E}[X_n]\,|X_{1:n-1}\right] \leq 0.$$

We now use Hoeffding's result (Expression (11)) that if variable X satisfies $a \leq X \leq b$ and $E[X] \leq 0$, then $E[\exp(sX)] \leq \exp(s^2(b-a)^2/8)$ for any $s > 0$. Thus for any P we have

$$E_P\left[\exp\left(s(X_n - \overline{E}[X_n])\right)|X_{1:n-1}\right] \leq \exp(s^2B_n^2/8)$$

and then $h(X_{1:n-1}) \leq \exp(s^2B_n^2/8)$. Thus

$$\begin{aligned}
&\overline{P}\left(\sum_{i=1}^{n}(X_i - \overline{E}[X_i]) \geq \epsilon\right) \\
&\quad\leq e^{-s\epsilon}\overline{E}\left[\exp\left(\sum_{i=1}^{n} s(X_i - \overline{E}[X_i])\right)\right] \\
&\quad\leq e^{-s\epsilon}\overline{E}\left[\exp\left(\sum_{i=1}^{n-1} s(X_i - \overline{E}[X_i])\right)\exp(s^2B_n^2/8)\right] \\
&\quad\leq e^{-s\epsilon}\exp(s^2B_n^2/8)\overline{E}\left[\exp\left(\sum_{i=1}^{n-1} s(X_i - \overline{E}[X_i])\right)\right].
\end{aligned}$$

These inequalities can be iterated to produce:

$$\overline{P}\left(\sum_{i=1}^{n}(X_i - \overline{E}[X_i]) \geq \epsilon\right) \leq e^{-s\epsilon}\exp\left(s^2\sum_{i=1}^{n}B_i^2/8\right).$$

Finally, by taking $s = 4\epsilon/\gamma_n$,

$$\overline{P}\left(\sum_{i=1}^{n}(X_i - \overline{E}[X_i]) \geq \epsilon\right) \leq e^{-2\epsilon^2/\gamma_n}.$$

The second inequality in the theorem is proved by noting that weak irrelevance of $X_1, \ldots, X_n$ implies weak irrelevance of $-X_1, \ldots, -X_n$ (as $\underline{E}[X_i] = -\overline{E}[-X_i]$), and then by taking $\overline{P}\big(\sum_{i=1}^{n}((-X_i) - \overline{E}[-X_i]) \geq \epsilon\big)$. □

3.2 Laws of large numbers

Theorem 1 leads to simple proofs of laws of large numbers already stated by De Cooman and Miranda [5]. To start, take Assumption (1). Using subadditivity of upper probability and Theorem 1,

$$\overline{P}\left(\left(\sum_{i=1}^{n} X_i \geq n\overline{\mu}_n + \epsilon\right) \cup \left(\sum_{i=1}^{n} X_i \leq n\underline{\mu} - \epsilon\right)\right) \leq 2e^{-\frac{2\epsilon^2}{\gamma_n}},$$

where as before, $\underline{\mu}_n \doteq (1/n)\sum_{i=1}^{n}\underline{E}[X_i]$ and $\overline{\mu}_n \doteq (1/n)\sum_{i=1}^{n}\overline{E}[X_i]$. By noting that $\underline{P}(A) = 1 - \overline{P}(A^c)$ for any event A, by including the endpoints of relevant inequalities, and by using $n\epsilon$ instead of ϵ:

$$\begin{aligned}
&\underline{P}\left(\underline{\mu} - \epsilon \leq \frac{\sum_{i=1}^{n} X_i}{n} \leq \overline{\mu} + \epsilon\right) \quad \geq \\
&\underline{P}\left(\underline{\mu} - \epsilon < \frac{\sum_{i=1}^{n} X_i}{n} < \overline{\mu} + \epsilon\right) \quad \geq \quad 1 - 2e^{-\frac{2n\epsilon^2}{B^2}},
\end{aligned}$$

where we define $B \doteq \max_i B_i$. By taking limits, we obtain a weak law of large numbers:

$$\lim_{n\to\infty}\underline{P}\left(\underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^{n} X_i}{n} < \overline{\mu}_n + \epsilon\right) = 1.$$

An analogue of De Cooman and Miranda's finitary strong law of large numbers can be deduced as well from the previous inequalities, as follows. Here and in the remainder of the paper, n, N and N' denote positive integers. For all $\epsilon > 0$, $N > 0$ and $N' > 0$,

$$\begin{aligned}
&\overline{P}\left(\exists n \in [N, N+N'] : \frac{\sum_{i=1}^{n} X_i}{n} \geq \overline{\mu} + \epsilon\right) \\
&\qquad\leq \sum_{n=N}^{N+N'} \overline{P}\left(\frac{\sum_{i=1}^{n} X_i}{n} \geq \overline{\mu} + \epsilon\right) \\
&\qquad\leq \sum_{n=N}^{N+N'} e^{-2n\epsilon^2/B^2} \\
&\qquad= \left(e^{-2N\epsilon^2/B^2}\right)\sum_{n=0}^{N'} e^{-2n\epsilon^2/B^2} \\
&\qquad= \left(e^{-2N\epsilon^2/B^2}\right)\frac{1 - e^{2(N'+1)\epsilon^2/B^2}}{1 - e^{-2\epsilon^2/B^2}} \\
&\qquad< \frac{e^{-2N\epsilon^2/B^2}}{1 - e^{-2\epsilon^2/B^2}}.
\end{aligned}$$

Consequently,

$$\overline{P}\left(\exists n \in [N, N+N'] : \frac{\sum_{i=1}^n X_i}{n} \geq \overline{\mu} + \epsilon\right) < \epsilon,$$

provided that N is a positive integer such that

$$N > -(B^2/(2\epsilon^2)) \ln \epsilon (1 - e^{-2\epsilon^2/B^2}).$$

An analogous argument leads to

$$\overline{P}\left(\exists n \in [N, N+N'] : \frac{\sum_{i=1}^n X_i}{n} \leq \underline{\mu} - \epsilon\right) < \epsilon.$$

By superadditivity of upper probability, we obtain a perhaps more intuitive statement of the strong law of large numbers: for all $\epsilon > 0$, there is N such that for any N',

$$\underline{P}\left(\forall n \in [N, N+N'] : \underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) > 1 - 2\epsilon,$$

thus reproducing De Cooman and Miranda's strong laws.

We now present a pair of weak/strong laws of large numbers under weak irrelevance. De Cooman and Miranda prove a similar pair of laws by resorting to their previous results on *forward irrelevant natural extensions* [5, Sec. 4.1]. The proof offered now is perhaps more direct, using our analogue of Azuma's inequality.

Theorem 3 *If bounded variables $X_1, \ldots, X_n$ satisfy weak irrelevance and Expression (4) holds, then for any $\epsilon > 0$,*

$$\underline{P}\left(\underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} \overline{\mu}_n + \epsilon\right) \geq 1 - 2e^{-2n\epsilon^2/B^2},$$

and there is N such that for any N',

$$\underline{P}\left(\forall n \in [N, N+N'] : \underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) > 1 - 2\epsilon.$$

Proof. Using subadditivity of upper probability and Theorem 2, and defining again $B \doteq \max_i B_i$,

$$\overline{P}\left(\left(\sum_{i=1}^n X_i \geq n\overline{\mu}_n + \epsilon\right) \cup \left(\sum_{i=1}^n X_i \leq n\underline{\mu} - \epsilon\right)\right) \leq 2e^{-\frac{2n\epsilon^2}{B^2}},$$

and we obtain the first expression in the theorem. To produce the second inequality (strong law), note:

$$\begin{aligned}
&\overline{P}\left(\exists n \in [N, N+N'] : \frac{\sum_{i=1}^n X_i}{n} \geq \overline{\mu} + \epsilon\right) \\
&\qquad \leq \sum_{n=N}^{N+N'} \overline{P}\left(\frac{\sum_{i=1}^n X_i}{n} \geq \overline{\mu} + \epsilon\right) \\
&\qquad \leq \sum_{n=N}^{N+N'} e^{-2n\epsilon^2/B^2} \\
&\qquad < \frac{e^{-2N\epsilon^2/B^2}}{1 - e^{-2\epsilon^2/B^2}}.
\end{aligned}$$

Again,

$$\overline{P}\left(\exists n \in [N, N+N'] : \frac{\sum_{i=1}^n X_i}{n} \geq \overline{\mu} + \epsilon\right) < \epsilon$$

provided that N is a positive integer such that

$$N > -(B^2/(2\epsilon^2)) \ln \epsilon (1 - e^{-2\epsilon^2/B^2}).$$

This is "half" of the second expression in the theorem; the other "half" is proved analogously. □

The theorem easily implies the following concise weak law of large numbers, by taking limits:

$$\lim_{n\to\infty} \underline{P}\left(\underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) = 1.$$

4 Laws of large numbers without boundedness

We now consider variables without bounds in their ranges under the assumption of weak irrelevance; the resulting laws of large numbers are the main contribution of the paper. We will assume in this section that countable additivity holds (Expression (5)). This assumption of countable addivity implies disintegrability; that is, $E_P[W] = E_P[E_P[W|Z]]$ for any P, W and Z. Thus our setup is close to the standard (Kolmogorovian) one, where any expectation functional is a linear monotone and monotonically convergent functional that can be expressed through Lebesgue integration. We only depart from the Kolmogorovian tradition in explicitly letting a *set* of such functionals to be permissible given a set of assessments.

We will use a sequence of variables $\{Y_n\}$ defined as follows:

$$Y_n \doteq \sum_{i=1}^n X_i - E_P[X_i|X_{1:i-1}].$$

The key observation is that Y_n is a function of all variables $X_{1:n}$ such that

$$\begin{aligned}
E_P[Y_n|X_{1:n-1}] &= \left(\sum_{i=1}^{n-1} X_i - E_P[X_i|X_{1:i-1}]\right) + \\
&\quad E_P[X_n - E_P[X_n|X_{1:n-1}]\,|X_{1:n-1}] \\
&= Y_{n-1} + \\
&\quad E_P[X_n|X_{1:n-1}] - E_P[X_n|X_{1:n-1}] \\
&= Y_{n-1};
\end{aligned}$$

so, $\{Y_n\}$ is a *martingale* with respect to P. Thus,

$$\begin{aligned}
&E_P\left[(Y_n - Y_{n-1})^2|X_{1:n-1}\right] \\
&= E_P\left[Y_n^2|X_{1:n-1}\right] - 2E_P[Y_{n-1}Y_n|X_{1:n-1}] + Y_{n-1}^2 \\
&= E_P\left[Y_n^2|X_{1:n-1}\right] - 2Y_{n-1}E_P[Y_n|X_{1:n-1}] + Y_{n-1}^2 \\
&= E_P\left[Y_n^2|X_{1:n-1}\right] - 2Y_{n-1}Y_{n-1} + Y_{n-1}^2 \\
&= E_P\left[Y_n^2|X_{1:n-1}\right] - Y_{n-1}^2.
\end{aligned}$$

And by taking expectations on both sides and noting that $Y_i - Y_{i-1} = X_i - E_P[X_i|X_{1:i-1}]$, we get

$$E_P\left[Y_n^2\right] = E_P\left[(X_n - E_P[X_n|X_{1:n-1}])^2\right] + E_P\left[Y_{n-1}^2\right].$$

Iterating this expression, we obtain:

$$E_P\left[Y_n^2\right] = \sum_{i=1}^{n} E_P\left[(X_i - E_P[X_i|X_{1:i-1}])^2\right]. \quad (9)$$

With these preliminaries, we have:

Theorem 4 *Assume countable additivity. If variables $X_1, \ldots, X_n$ satisfy weak irrelevance, and $\underline{E}[X_i]$ and $\overline{E}[X_i]$ are finite quantities such that $\overline{E}[X_i] - \underline{E}[X_i] \leq \delta$, and the variance of any X_i is no larger than a finite quantity σ^2, then for any $\epsilon > 0$,*

$$\underline{P}\left(\underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) \geq 1 - \frac{\sigma^2 + \delta^2}{\epsilon^2 n},$$

and there is $N > 0$ such that for any $N' > 0$,

$$\underline{P}\left(\forall n \in [N, N+N'] : \underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) > 1 - 2\epsilon.$$

Consequently,

$$\forall \epsilon > 0 : \lim_{n\to\infty} \underline{P}\left(\underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) = 1,$$

$$\underline{P}\left(\limsup_{n\to\infty} \left(\frac{\sum_{i=1}^n X_i}{n} - \overline{\mu}_n\right) \leq 0\right) = 1,$$

$$\underline{P}\left(\liminf_{n\to\infty} \left(\frac{\sum_{i=1}^n X_i}{n} - \underline{\mu}_n\right) \geq 0\right) = 1.$$

Proof. For a fixed P and for all $\epsilon > 0$,

$$\begin{aligned}
&P\left(\underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) \\
&= P\left(\sum_{i=1}^n \underline{E}[X_i] - \epsilon n < \sum_{i=1}^n X_i < \sum_{i=1}^n \overline{E}[X_i] + \epsilon n\right) \\
&\geq P\left(\sum_{i=1}^n E_P[X_i|X_{1:i-1}] - \epsilon n < \sum_{i=1}^n X_i \right. \\
&\qquad\qquad \left. < \sum_{i=1}^n E_P[X_i|X_{1:i-1}] + \epsilon n\right) \\
&\qquad \text{(using weak irrelevance)} \\
&= P\left(-\epsilon < \frac{\sum_{i=1}^n X_i - E_P[X_i|X_{1:i-1}]}{n} < \epsilon\right) \\
&= P(-\epsilon < Y_n/n < \epsilon) \\
&= P(|Y_n/n| < \epsilon).
\end{aligned}$$

Applying Chebyshev's inequality and Expression (9),

$$\begin{aligned}
P(|Y_n/n| \geq \epsilon) &\leq \frac{E_P\left[Y_n^2\right]}{\epsilon^2 n^2} \\
&= \frac{\sum_{i=1}^n E_P\left[(X_i - E_P[X_i|X_{1:i-1}])^2\right]}{\epsilon^2 n^2}.
\end{aligned}$$

Now write $(X_i - E_P[X_i|X_{1:i-1}])^2$ as

$$\left((X_i - E_P[X_i]) + (E_P[X_i] - E_P[X_i|X_{1:i-1}])\right)^2,$$

and then:

$$\begin{aligned}
&\sum_{i=1}^n E_P\left[(X_i - E_P[X_i|X_{1:i-1}])^2\right] \\
&= \sum_{i=1}^n E_P\left[(X_i - E_P[X_i])^2\right] \\
&\quad + 2E_P[(X_i - E_P[X_i])(E_P[X_i] - E_P[X_i|X_{1:i-1}])] \\
&\quad + E_P\left[(E_P[X_i] - E_P[X_i|X_{1:i-1}])^2\right] \\
&\leq \sum_{i=1}^n \sigma^2 + \delta^2 \\
&\quad + 2(E_P[X_i] - E_P[X_i|X_{1:i-1}])E_P[X_i - E_P[X_i]] \\
&= \sum_{i=1}^n \sigma^2 + \delta^2.
\end{aligned}$$

Hence

$$\sum_{i=1}^n E_P\left[(X_i - E_P[X_i|X_{1:i-1}])^2\right] \leq n(\sigma^2 + \delta^2), \quad (10)$$

and combining these inequalities, we obtain:

$$P(|Y_n/n| \geq \epsilon) \leq \frac{\sigma^2 + \delta^2}{\epsilon^2 n},$$

and then

$$P\left(\underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) \geq 1 - \frac{\sigma^2 + \delta^2}{\epsilon^2 n}$$

for any P, as desired. By taking the limit as n grows without bound, we obtain

$$\lim_{n\to\infty} \underline{P}\left(\underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) = 1.$$

The proof of the strong law of large numbers uses the same strategy, but replaces the appeal to Chebyshev's inequality by an appeal to the Kolmogorov-Hajek-Renyi inequality (described in the Appendix), following the proof of the strong law of large numbers by Whittle [29, Thm. 14.2.3].So, for a fixed P and for all $\epsilon > 0$, we proceed as previously to obtain:

$$\begin{aligned}
&P\left(\forall n \in [N, N+N'] : \underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) \\
&\qquad \geq P\left(\forall n \in [N, N+N'] : -\epsilon < \frac{Y_n}{n} < \epsilon\right) \\
&\qquad = P(\forall n \in [N, N+N'] : |Y_n/n| < \epsilon).
\end{aligned}$$

As $\{Y_N, Y_{N+1}, \ldots, Y_{N+N'}\}$ forms a martingale, we use the Kolmogorov-Hajek-Renyi inequality to produce:

$$\begin{aligned}
&P(\forall n \in [N, N+N'] : |Y_n/n| < \epsilon) \\
&\quad \geq 1 - \frac{\sum_{i=1}^{N} E_P\left[(X_i - E_P[X_i|X_{1:i-1}])^2\right]}{\epsilon^2 N^2} \\
&\qquad - \sum_{i=N+1}^{N+N'} \frac{E_P\left[(X_i - E_P[X_i|X_{1:i-1}])^2\right]}{\epsilon^2 i^2} \\
&\quad \geq 1 - \frac{\sigma^2+\delta^2}{\epsilon^2 N} - \sum_{i=N+1}^{N+N'} \frac{\sigma^2+\delta^2}{\epsilon^2 i^2} \\
&\qquad \text{(using Expression (10))} \\
&\quad \geq 1 - \frac{\sigma^2+\delta^2}{\epsilon^2 N} - \sum_{i=N+1}^{\infty} \frac{\sigma^2+\delta^2}{\epsilon^2 i^2} \\
&\quad \geq 1 - \frac{\sigma^2+\delta^2}{\epsilon^2} \left(\frac{1}{N} + \int_N^{\infty} 1/i^2 di\right) \\
&\quad = 1 - \frac{\sigma^2+\delta^2}{\epsilon^2} \left(\frac{1}{N} + \frac{1}{N}\right) \\
&\quad = 1 - 2\frac{\sigma^2+\delta^2}{\epsilon^2 N}.
\end{aligned}$$

Consequently, for integer $N > (\sigma^2+\delta^2)/\epsilon^3$, we obtain the desired inequality

$$\underline{P}\left(\forall n \in [N, N+N'] : \underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^{n} X_i}{n} < \overline{\mu}_n + \epsilon\right) > 1 - 2\epsilon.$$

As we assume countable additivity for every P, the proof of the Kolmogorov-Hajek-Renyi can be extended to an infinite intersection of (decreasing) events expressed as $\{\forall j \geq 1 : |X_j| < \epsilon_j\}$; thus

$$\begin{aligned}
&\forall \epsilon > 0 : \forall \delta > 0 : \exists N > 0 : \\
&\quad P\left(\forall m \geq N : \frac{\sum_{i=1}^{m} X_i - \overline{E}[X_i]}{m} > \epsilon\right) \geq 1 - \delta,
\end{aligned}$$

and this is equivalent to:

$$\forall \epsilon > 0 : \lim_{N \to \infty} P\left(\forall m \geq N : \frac{\sum_{i=1}^{m} X_i - \overline{E}[X_i]}{m} > \epsilon\right) = 1.$$

As the events in these probability values form an increasing sequence, we have, for all $\epsilon > 0$,

$$P\left(\exists N > 0 : \forall m \geq N : \frac{\sum_{i=1}^{m} X_i - \overline{E}[X_i]}{m} > \epsilon\right) = 1.$$

Now this is equivalent to $\forall k > 0 : P(A_k) = 1$, where $A_k = \{\exists N > 0 : \forall m \geq N : (1/m)\sum_{i=1}^{m} X_i - \overline{E}[X_i] > 1/k\}$, and because $P(\cup_{k>0} \neg A_k) \leq \sum_{k>0} P(\neg A_k) = 0$, we have $P(\forall k > 0 : A_k) = 1$, so

$$P\left(\forall k > 0 : \exists N > 0 : \forall m \geq N : \frac{\sum_{i=1}^{m} X_i - \overline{E}[X_i]}{m} > \epsilon\right) = 1.$$

This is exactly the desired expression

$$\underline{P}\left(\limsup_{n \to \infty} \left(\frac{\sum_{i=1}^{n} X_i}{n} - \overline{\mu}_n\right) \leq 0\right) = 1.$$

A similar argument proves the last inequality in the theorem, starting from:

$$\begin{aligned}
&\forall \epsilon > 0 : \forall \delta > 0 : \exists N > 0 : \\
&\quad P\left(\forall m \geq N : \frac{\sum_{i=1}^{m} X_i - \underline{E}[X_i]}{m} < -\epsilon\right) \geq 1 - \delta.
\end{aligned}$$

□

5 Discussion

The concentration inequalities and laws of large numbers proved in this paper assume rather weak conditions of epistemic irrelevance. When compared to usual laws of large numbers, both premises and consequences are weaker: expectations are not assumed precisely known, and convergence is interval-valued.

Theorems 1 and 2 and their ensuing laws of large numbers are implied by De Cooman and Miranda's seminal work [5] (and their results generalize several previous efforts [12]). Actually, De Cooman and Miranda start from a weaker condition of *forward factorization* that implies both Assumption (1) and weak irrelevance. The possible advantage of our proof techniques for these two theorems is that they are rather close to well-known methods in standard probability theory, such as Hoeffding's inequality (it should be noted that De Cooman and Miranda already indicate the similarity between their inequalities and Hoeffding's).

The most significant results of the paper employ weak irrelevance to produce concentration inequalities (Theorem 2) and laws of large numbers (Theorems 3 and 4). The latter theorem is possibly the most valuable contribution. The strategy for most proofs is to translate assumptions of weak irrelevance into facts regarding martingales, and to adapt results for martingales to this setting. This strategy keeps the proof relatively short and close to well-known results in probability theory. The connection between lower/upper expectations and the theory of martingales seems rather natural [4, 25], but the relationship between epistemic irrelevance and martingales does not appear to have been explored in depth so far. We note that the basic constraint defining martingales (that is, $E[Y_n|X_{1:n-1}] = Y_{n-1}$) is preserved by convex combination of mixtures; therefore, the study of martingales seems appropriate when one deals with convex sets of probability measures — certainly it seems less contorted than the analysis through stochastic independence, as stochastic independence is *not* preserved by convex combination.

The proofs presented in this paper need assumptions of disintegrability that can be easily satisfied if countable additivity is adopted. It is an open question whether similar results can be proven without disintegrability, particularly when one deals with unbounded variables.

Acknowledgements

Thanks to Gert de Cooman and Enrique Miranda for their generous suggestions regarding content and presentation. Thanks to a reviewer who indicated the work by Sadrolhefazi and Fine on ergodic properties [21], a topic to be pursued in the future.

The author is partially supported by CNPq.

A Two auxiliary inequalities

The following inequality is a simple extension of a basic result by Hoeffding [8, 15]: If variable X satisfies $a \leq X \leq b$ and $E[X] \leq 0$, then for any $s > 0$,

$$E[\exp(sX)] \leq \exp(s^2(b-a)^2/8). \tag{11}$$

First, the inequality is clearly valid if $a = b$, or if $a = 0$, or if $b < 0$. From now on, suppose $b \geq 0 > a$. By convexity of the exponential function,

$$\exp(sx) \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa} \quad \text{for } x \in [a,b].$$

Given monotonicity of expectations and $E[X] \leq 0$,

$$E[\exp(sX)] \leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \doteq \exp(\phi(s(b-a)))$$

for $\phi(u) = -pu + \log(1-p+pe^u)$ with $p = -a/(b-a)$ (and note that $p \in (0,1]$ in the situation under consideration). Given that $\phi(0) = \phi'(0) = 0$ and $\phi''(u) \leq 1/4$ for $u > 0$ (as the maximum of $\phi''(u)$ is $1/4$, attained at $e^u = (1-p)/p$), we can use Taylor's theorem as follows. For some $v \in (0,u)$, $\phi(u) = \phi(0) + u\phi'(0) + (u^2/2)\phi''(v) \leq (u^2/8)$ and consequently $\phi(s(b-a)) \leq s^2(b-a)^2/8$. By putting together these inequalities, we obtain Expression (11).

We now review the Kolmogorov-Hajek-Renyi inequality, almost exactly as proved by Whittle [29]; this is presented just to indicate the role of (elementwise) disintegrability in the derivation. Let $\{X_i\}$ be a martingale with $X_0 = 0$, and let $\{\epsilon_i\}$ be a sequence $0 = \epsilon_0 \leq \epsilon_1 \leq \ldots$; the inequality is

$$P(\forall j \in [1,n] : |X_j| < \epsilon_j) \geq 1 - \sum_{i=1}^{n} \frac{E\big[(X_i - X_{i-1})^2\big]}{\epsilon_i^2}.$$

To prove this inequality, define

$$A_n \doteq \{\forall j \in [1,n] : |X_j| < \epsilon_j\}.$$

Using $\xi_i = X_i - X_{i-1}$, and again denoting an event and its indicator function by the same symbol, we have

$$\begin{aligned}
P(A_n) &= E_P[A_n] \\
&= E_P[A_{n-1}\{|X_n| < \epsilon_n\}] \\
&\geq E_P\big[A_{n-1}(1 - X_n^2/\epsilon_n^2)\big] \\
&\quad \text{(as } \{|X| < \epsilon\} \geq 1 - X^2/\epsilon^2) \\
&= E_P\big[A_{n-1}(1 - (X_{n-1}^2 + \xi_n^2)/\epsilon_n^2)\big] \\
&\quad \text{(by the martingale property)} \\
&\geq E_P\big[A_{n-2}(1 - X_{n-1}^2/\epsilon_{n-1}^2)\big] - E_P\big[\xi_n^2/\epsilon_n^2\big] \\
&\quad \text{(as } \epsilon_{n-1} \leq \epsilon_n \text{ and} \\
&\quad \{|X| < \epsilon\}(1 - X^2/\epsilon^2) \geq (1 - X^2/\epsilon^2)).
\end{aligned}$$

Iteration of the last inequality yields the result. Note that it was necessary to apply disintegrability of P when applying the martingale property (that is, elementwise disintegrability is used).

References

[1] R. B. Ash and C. A. Dóleans-Dade. *Probability and Measure Theory (2nd ed.).* Academic Press, 1999.

[2] F. Chung and L. Lu. Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, 3(1):79–127, 2006.

[3] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems.* Springer-Verlag, New York, 1999.

[4] G. de Cooman and F. Hermans. Imprecise probability trees: Bridging two theories of imprecise probability *Artificial Intelligence*, 172:1400–1427, 2008.

[5] G. de Cooman and E. Miranda. Weak and strong laws of large numbers for coherent lower previsions. *Journal of Statistical Planning and Inference*, 138(8):2409–2432, August 2008.

[6] B. de Finetti. *Theory of probability, vol. 1-2.* Wiley, New York, 1974.

[7] L. Devroye. Exponential inequalities in nonparametric estimation. In G.Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, NATO ASI Series, pages 31–44. Kluwer Academic Publishers, Dordrecht, 1991.

[8] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer Verlag, New York, 1996.

[9] L. E. Dubins. On Lebesgue-like extensions of finitely additive measures. *Annals of Probability*, 2(3):456–463, 1974.

[10] L. E. Dubins. Finitely additive conditional probability, conglomerability and disintegrations. *Annals of Statistics*, 3(1):89–99, 1975.

[11] L. E. Dubins and L. J. Savage. *How to Gamble If You Must: Inequalities for Stochastic Processes*. McGraw-Hill, New York, 1965.

[12] L. G. Epstein and M. Schneider. IID: independently and indistinguishably distributed. *Journal of Economic Theory*, 113(1):32–50, 2003.

[13] F. J. Giron and S. Rios. Quasi-Bayesian behaviour: A more realistic approach to decision making? In J. M. Bernardo, J. H. DeGroot, D. V. Lindley and A. F. M. Smith, *Bayesian Statistics*, pages 17–38, University Press, Valencia, Spain, 1980.

[14] D. Heath and W. Sudderth. On finitely additive priors, coherence, and extended admissibility. *Annals of Mathematical Statistics*, 43:2072–2077, 1978.

[15] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[16] J. B. Kadane, M. J. Schervish, and T. Seidenfeld. Reasoning to a foregone conclusion. *Journal of the American Statistical Association*, 91(435):1228–1235, 1996.

[17] R. L. Karandikar. A general principle for limit theorems in finitely additive probability. *Transactions of the American Mathematical Society*, 273(2):541–550, 1982.

[18] P. Krauss. Representation of conditional probability measures on Boolean algebras. *Acta Mathematica Academiae Scientiarum Hungaricae*, 19(3-4):229–241, 1968.

[19] D. A. Lane and W. D. Sudderth. Coherent predictions are strategic. *Annals of Statistics*, 13(3):1244–1248, 1985.

[20] D. Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, 2002.

[21] A. Sadrolhefazi and T. L. Fine. Finite-dimensional distributions and tail behavior in stationary interval-valued probability models. *Annals of Statistics*, 22(4):1840–1870, 1994.

[22] M. Schervish, T. Seidenfeld, and J. B. Kadane. The extent of non-conglomerability of finitely additive measures. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwardte Gebiete*, 66:205–226, 1984.

[23] T. Seidenfeld. Remarks on the theory of conditional probability: some issues of finite versus countable additivity. In *Statistics – Philosophy, Recent History, and Relations to Science*, pages 167–177. Kluwer Academic, 2001.

[24] T. Seidenfeld, M. J. Schervish, and J. B. Kadane. Improper regular conditional distributions. *Annals of Probability*, 29(4):1612–1624, 2001.

[25] G. Shafer and V. Vovk. *Probability and Finance: It's Only a Game!*. Wiley, 2001.

[26] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[27] K. Weichselberger. The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24(2-3):149–170, 2000.

[28] K. Weichselberger, T. Augustin (assistant), and A. Wallner (assistant). *Elementare Grundbegriffe einer Allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als Umfassendes Konzept*. Physica-Verlag Heidelberg, 2001.

[29] P. Whittle. *Probability via Expectation (3rd ed.)*. Springer-Verlag, 1992.

6th International Symposium on Imprecise Probability: Theories and Applications, Durham, United Kingdom, 2009

Imprecise Markov Chains with an Absorbing State

Richard J. Crossman
Durham University
r.j.crossman@durham.ac.uk

Pauline Coolen-Schrijner
Durham University

Damjan Škulj
University of Ljubljana
damjan.skulj@fdv.uni-lj.si

Frank P.A. Coolen
Durham University
frank.coolen@durham.ac.uk

Abstract

Several authors have presented methods for considering the behaviour of Markov chains in the generalised setting of imprecise probability. Some assume a constant transition matrix which is not known precisely, instead bounds are given for each element. Others consider a transition matrix which is neither known precisely nor assumed to be constant, though each element is known to exist within intervals that are constant over time. In both cases results have been published regarding the long-term behaviour of such chains. When a finite Markov chain is considered with a single absorbing state, however, eventual absorption is generally certain in both cases. Thus it is of interest to consider the long-term behaviour of the chain, conditioned on non-absorption, within the setting of imprecise probability. Methods have previously been presented for the case of a constant transition matrix, and submitted for the case of a non-constant transition matrix. In this paper the methods for the two cases are compared.

Keywords. Absorbing state, imprecise probability, Markov chains, time-inhomogeneity

1 Introduction

There are several papers in which the theory of interval probability has been applied to the consideration of Markov chains. Kozine and Utkin [10] consider the situation in which the individual elements of the transition matrix are assumed to be constant over time, but may not be known precisely (thus that paper can be thought of as generalising the time-homogeneous case). Instead, all that is known are the intervals in which each individual matrix element is contained. This property can be relaxed, as it was by Škulj [13, 14], by only requiring that the intervals to which those elements belong remain constant over time, and allowing the elements to vary with time (thus those papers can be thought of as generalising the time-inhomogeneous case). In those same papers the concept of the initial distribution is also generalised, so that rather than assume a specific initial distribution, an entire set of possible initial distributions is defined. The papers then considered the long-term behaviour of such chains, and proved that, subject to certain conditions, the possible distributions as time approaches infinity form a set that is independent of the set of initial distributions. An alternate method for considering the situation found in [13, 14] was offered by de Cooman *et al.* in [3]; we explain in Section 3 why we have not adopted their method in this paper.

It can be proved that for a finite Markov chain with one absorbing state eventual absorption is certain both in the case given by Kozine and Utkin [10], and also the case found in [14], assuming the conditions required in that paper (the respective proofs for these results can be found in Crossman *et al.* [4, 5]). In this situation, then, it is of more interest to consider the long-term behaviour of the chain when conditioning on non-absorption at each step.

What follows can be thought of as a generalisation of the *limiting conditional distribution* in the precise case. The limiting conditional distribution, if used as the initial distribution, is referred to as the *quasi-stationary distribution* (QSD). The QSD has many applications. For example, it is used by Pakes [11] to better understand population sizes, which are modelled in that paper as birth-death processes with catastrophes. In this case the QSD represents the long-term behaviour of a stable population, before the point at which it becomes extinct. Further, Parsons and Pollet [12] apply QSDs to describe the long-term behaviour of certain catalytic chemical reactions.

Crossman *et al.* [4] considered the long-term behaviour conditioned on non-absorption for the model given in [10], and the consideration of sets of initial distributions is introduced as in [13]. Crossman and Škulj [5] then applied this consideration to the model

given in [13], though the restrictions upon each row of the transition matrix is given as a closed probability set, rather than a group of intervals. In this paper the method found in [4] is similarly expanded to using closed probability sets (more on this can be found in Crossman [6]), and the two different approaches are compared.

1.1 Markov chains with imprecision

The following model is given in a slightly different form by Škulj [13]. Let $\mathcal{X} = \{X(n), n = 0, \ldots\}$ be a discrete-time Markov chain on the state space $S = \{-1\} \cup C$ with $C = \{0, \ldots, s\}$ where -1 is an absorbing state and C is a set of transient states. Imprecision is introduced by the assumption that the transition matrix for any given time step is not known precisely. Instead, limitations are imposed upon the possible values of each transition probability at each step.

Define $s + 2$ closed sets of probability distributions, $\mathcal{P}^{(i)},\ i = -1, 0, \ldots, s$.

Definition 1.1 All potential transition matrices for a given time step belong to the set

$$\mathcal{M}(P) := \left\{ \begin{pmatrix} \boldsymbol{p}^{(-1)} \\ \vdots \\ \boldsymbol{p}^{(s)} \end{pmatrix} \mid \boldsymbol{p}^{(i)} \in \mathcal{P}^{(i)},\ \forall i \in C \right\}$$

where the choice of the element from $\mathcal{P}^{(i)}$ has no effect on the choice of the element $\mathcal{P}^{(j)}$ if $i \neq j$.

Thus, each row of the transition matrix for a given time step is chosen from a set of probability distributions, and each choice is made independently.

Further conditions are now given. First, as -1 is an absorbing state $\mathcal{P}^{(-1)} = \{(1, 0, \ldots, 0)\}$ is required. Further, each of the possible transition matrices must guarantee that C is a single communicating class with each set in C aperiodic.[1]

Definition 1.2 The set of all possible initial distributions over S is denoted by

$$\mathcal{M}_0 := \{\boldsymbol{v} = (v_{-1}, v_0, \ldots, v_s) | v_i \geq 1\ \forall i, \sum_{i=-1}^{s} v_i = 1\}.$$

Furthermore, $\mathcal{D}_0$ is used to denote a strict subset of $\mathcal{M}_0$.

Thus, $\mathcal{D}_0$ can be thought of as the set of initial distributions deemed possible for a given process, where this conclusion is arrived at by some unspecified method. $\mathcal{M}_0$ would be used only when nothing whatsoever is known about the initial distribution.

[1] Note that this is a more general formulation than can be found in [4], the justification for this change can be found in [6].

2 Imprecise Markov chains with constant transition matrix

In this section it is assumed that there is a single element of $\mathcal{M}(P)$ that describes the transition probabilities at every time step, i.e. the transition matrix is unknown, but constant.

As mentioned, $\mathcal{D}_0$ represents the set of initial distributions over S that have been judged possible. Thus for a matrix $P \in \mathcal{M}(P)$ the set $\tilde{\mathcal{D}}_n(P)$ of all possible distributions over S at time $n \geq 1$ can be defined as follows.

Definition 2.1

$$\tilde{\mathcal{D}}_n(P) := \{\boldsymbol{v}P \mid \boldsymbol{v} \in \tilde{\mathcal{D}}_{n-1}(P)\} = \{\boldsymbol{v}P^n \mid \boldsymbol{v} \in \tilde{\mathcal{D}}_0(P)\}$$

where $\tilde{\mathcal{D}}_0(P) := \mathcal{D}_0$. Should every possible initial distribution be considered possible, the appropriate definition becomes

$$\tilde{\mathcal{M}}_n(P) := \{\boldsymbol{v}P | \boldsymbol{v} \in \tilde{\mathcal{M}}_{n-1}(P)\} = \{\boldsymbol{v}P^n | \boldsymbol{v} \in \tilde{\mathcal{M}}_0(P)\}$$

where $\tilde{\mathcal{M}}_0(P) := \mathcal{M}_0$.

However, since it is unknown which element of the set $\mathcal{M}(P)$ actually describes the behaviour of the chain, it is of more pratical use to introduce the following definition.

Definition 2.2

$$\tilde{\mathcal{M}}_n := \bigcup_{P \in \mathcal{M}(P)} \tilde{\mathcal{M}}_n(P). \tag{2.1}$$

Thus $\tilde{\mathcal{M}}_n$ contains every distribution possible at time n.

Theorem 2.1 For each $P \in \mathcal{M}(P)$ and $n \geq 0$,

$$\tilde{\mathcal{M}}_{n+1}(P) \subseteq \tilde{\mathcal{M}}_n(P).$$

Proof. For each $P \in \mathcal{M}(P)$, it follows from the definition of $\tilde{\mathcal{M}}_0(P)$ and the fact that P is a strictly stochastic matrix that $\tilde{\mathcal{M}}_1(P) = \{\boldsymbol{v}P | \boldsymbol{v} \in \tilde{\mathcal{M}}_0(P)\} \subseteq \tilde{\mathcal{M}}_0(P)$. Now assume that for a certain $n > 1$, $\tilde{\mathcal{M}}_n(P) \subseteq \tilde{\mathcal{M}}_{n-1}(P)$. Then

$$\begin{aligned} \tilde{\mathcal{M}}_{n+1}(P) &= \{\boldsymbol{v}P \mid \boldsymbol{v} \in \tilde{\mathcal{M}}_n(P)\} \\ &\subseteq \{\boldsymbol{v}P \mid \boldsymbol{v} \in \tilde{\mathcal{M}}_{n-1}(P)\} \\ &= \tilde{\mathcal{M}}_n(P). \end{aligned}$$

□

It is therefore appropriate to define

Definition 2.3

$$\tilde{\mathcal{M}}_\infty(P) := \bigcap_{n=0}^{\infty} \tilde{\mathcal{M}}_n(P)$$

This set $\tilde{\mathcal{M}}_\infty(P)$ describes the behaviour of the chain as time approaches infinity. Once again, though, since the correct matrix from $\mathcal{M}(P)$ is unknown, the following definition is of more practical use.

Definition 2.4

$$\tilde{\mathcal{M}}_\infty = \bigcup_{P\in\mathcal{M}(P)} \tilde{\mathcal{M}}_\infty(P).$$

It is proved in [4] that in our current case

$$\tilde{\mathcal{M}}_\infty = \bigcup_{P\in\mathcal{M}(P)} \{(1,0,\ldots,0)\} = \{(1,0,\ldots,0)\}.$$

Thus, since $\tilde{\mathcal{D}}_n(P) \subseteq \tilde{\mathcal{M}}_n(P)$ for all n, eventual absorption is certain irrespective of our choice of $\mathcal{D}_0$ or the actual element of the set $\mathcal{M}(P)$ that correctly describes the chain. We therefore consider the situation in which the chain is conditioned on non-absorption at each step.

We now define the following functions.

Definition 2.5 For

$$\boldsymbol{v} \in \mathcal{M}_0 \setminus \{(1,0,\ldots,0)\} \tag{2.2}$$

we have

$$f(\boldsymbol{v}) = f((v_{-1}, \boldsymbol{v}^*)) = \frac{1}{1-v_{-1}}\boldsymbol{v}^*,$$

and

$$\begin{aligned}\tilde{f}_\alpha(f(\boldsymbol{v})) &= \tilde{f}_\alpha(\frac{1}{1-v_{-1}}(v_0,\ldots,v_s)) \\ &:= (\alpha,(1-\alpha)(v_0,\ldots,v_s))\end{aligned}$$

where $\alpha \in [0,1)$.

Thus $f(\cdot)$ takes a distribution over S (for which absorption is not certain) and conditions it on non-absorption. $\tilde{f}_\alpha(\cdot)$ takes a distribution over C and maps it to a distribution in S for which the relative probabilities for being in any two states in C remain constant.

Lemma 2.1 $f(\tilde{f}_\alpha(\boldsymbol{v})P) = f(\tilde{f}_\beta(\boldsymbol{v})P)$ for any $P \in \mathcal{M}(P)$, independently of the values of α and β.

Proof.

$$\begin{aligned} f(\tilde{f}_\alpha(\boldsymbol{v})P) &= f\left((\alpha,(1-\alpha)\boldsymbol{v})\begin{pmatrix}1 & \boldsymbol{0}\\ \boldsymbol{p} & Q\end{pmatrix}\right) \\ &= \frac{(1-\alpha)\boldsymbol{v}Q}{|(1-\alpha)\boldsymbol{v}Q|} \\ &= \frac{\boldsymbol{v}Q}{|\boldsymbol{v}Q|}.\end{aligned}$$

□

Using $f(\cdot)$ it becomes possible to define the set of all possible distributions over C, conditioned on non-absorption, given $\mathcal{D}_0$, in the following way

Definition 2.6

$$\tilde{\mathcal{M}}_n^C := \{f(\boldsymbol{v}) | \boldsymbol{v} \in \tilde{\mathcal{M}}_n \setminus \{(1,0,\ldots,0)\}\} \tag{2.3}$$

and

$$\tilde{\mathcal{D}}_n^C := \{f(\boldsymbol{v}) | \boldsymbol{v} \in \tilde{\mathcal{D}}_n \setminus \{(1,0,\ldots,0)\}\}.$$

Theorem 2.2 For each $P \in \mathcal{M}(P)$ and $n \geq 0$,

$$\tilde{\mathcal{M}}_{n+1}^C \subseteq \tilde{\mathcal{M}}_n^C.$$

Proof. For each $P \in \mathcal{M}(P)$, and for $\mathcal{M}_0^C(P) = \mathcal{M}_0^C$, we have from (2.3) and Theorem 2.1 that

$$\begin{aligned}\tilde{\mathcal{M}}_{n+1}^C(P) &= \{f(\boldsymbol{v}) \mid \boldsymbol{v} \in \tilde{\mathcal{M}}_{n+1}(P)\backslash\{(1,0,\ldots,0)\}\} \\ &\subseteq \{f(\boldsymbol{v}) \mid \boldsymbol{v} \in \mathcal{M}_n(P)\backslash\{(1,0,\ldots,0)\}\} \\ &= \tilde{\mathcal{M}}_n^C(P).\end{aligned}$$

By taking the union of both sides over all $P \in \mathcal{M}(P)$ the proof is complete. □

The following definition is therefore appropriate.

Definition 2.7

$$\tilde{\mathcal{M}}_\infty^C := \bigcap_{i=0}^{\infty} \tilde{\mathcal{M}}_i^C.$$

Equivalently

$$\tilde{\mathcal{M}}_\infty^C = \bigcup_{P\in\mathcal{M}(P)} \tilde{\mathcal{M}}_\infty^C(P).$$

Thus $\tilde{\mathcal{M}}_\infty^C$ contains the possible distributions, conditioned on non-absorption as time goes to infinity, for all possible matrices from $\mathcal{M}(P)$ assuming nothing is known about the initial distribution.

Associated with each element $P \in \mathcal{M}(P)$ is a unique limiting conditional distribution $\boldsymbol{\alpha}_P$. In [4] it is proved that

$$\tilde{\mathcal{M}}_\infty^C = \bigcup_{P\in\mathcal{M}(P)} \tilde{\mathcal{M}}_\infty^C(P) = \bigcup_{P\in\mathcal{M}(P)} \boldsymbol{\alpha}_P \tag{2.4}$$

That is, although the correct element of $\mathcal{M}(P)$ is unknown, we know that the only possible distributions that can occur as time approaches infinity are the limiting conditional distributions of the elements of $\mathcal{M}(P)$. Since we have from Darroch and Seneta [7] that $\boldsymbol{\alpha}_P$ is reached independently of the initial distribution, we have that the right hand side of (2.4) represents the long-term behaviour of the chain, conditioned on non-absorption, independent of the choice of $\mathcal{D}_0^C$.

3 Imprecise Markov chains with non-constant transition matrix

In the case considered in Section 2, the long-term behaviour conditioned on non-absorption is easy to define, since such behaviour in the time-homogeneous case is well-known. In this section it is no longer assumed that the unknown transition matrix for time step n equals that for time step $m \neq n$. This corresponds in the precise case to the concept of time-inhomogeneous chains, the long-term behaviour of which is far less well understood.

A further condition is required in this case, namely that if $[P]_{ij} = 0$ for any $P \in \mathcal{M}(P)$ then $[Q]_{ij} = 0$ for all $Q \in \mathcal{M}(P)$. Thus a jump from state i to state j is either possible at all time steps, or impossible at all time steps. This is to prevent situations in which two or more transition matrices, each of which has C as a single communicating class, can be chosen from $\mathcal{M}(P)$ which, when multiplied, form a matrix for which C is *not* a single communicating class. It is certainly true that such matrices exist, but it is possible that they may already be disqualified by the conditions in Section 1.1 (most critically the assumption of independence), making this new condition unnecessary. Work is currently being conducted into ascertaining whether or not the new condition is redundant.

Definition 3.1 The set of possible n step transition matrices $\mathcal{M}^n(P)$ is defined as follows:

$$\mathcal{M}^n(P) := \{P_1 P_2 \dots P_n |\ P_i \in \mathcal{M}(P)\}.$$

Definition 3.2 The set $\mathcal{M}(P)$ is referred to as *regular* if for some n every $P \in \mathcal{M}^n(P)$ has only strictly positive elements. Further, the set $\mathcal{M}(P)$ is referred to as *conditionally regular on* C if for some r every $P \in \mathcal{M}^r(P)$ has all elements $[P]_{ij}$ strictly positive, where $i \in C$, $j \in S$.

Lemma 3.1 All matrices which belongs to the set $\mathcal{M}^{s+1}(P)$ are conditionally regular on C.

Proof. Any matrix P_{s+1} contained in $\mathcal{M}^{s+1}(P)$ represents the behaviour of a time-inhomogeneous Markov chain over $s+1$ time steps. By assumption each of the time steps are described by transition matrices for which C is a single communicating class, and each state in C is aperiodic. There must therefore be a path of n states, denoted $\{a_k\}_{k=1,\dots,n}$, strictly between i and j, where $i, j \in C$, and no element of $\{a_k\}_{k=1,\dots,n}$ is equal to either i or j.

Assume $i \neq j$. By assumption a jump from state i to state j is either possible or not at a given time step *independent of that time step*. Therefore if there exists $k_1 \neq k_2$ such that $a_{k_1} = a_{k_2}$, the elements $a_{k_1}, a_{k_1+1}, \dots, a_{k_2-1}$ can be removed from $\{a_k\}_{k=1,\dots,n}$ and the remainder still represents a viable path from i to j. This process can continue until there remains no duplicated value in the path, which forces $n \leq s-1$. Thus j can be reached from i in s jumps, forcing $P(X(s) = j | X(0) = i) > 0$. $P(X(s+1) = j | X(0) = i) > 0$ follows immediately from the fact that each possible transition matrix has C as a single communicating class, and thus cannot contain a column of zeroes.

Now assume $i = j$. The same process as above applies, except that without duplicated values in the path we have $n \leq s$, and hence we can return to i after $s+1$ jumps, and $P(X(s+1) = j | X(0) = i) > 0$. □

$\mathcal{M}_0$ and $\mathcal{D}_0$ are defined just as they were in Section 2. Since a constant transition matrix can no longer be assumed, the following definitions are required.

Definition 3.3

$$\mathcal{M}_n := \{\boldsymbol{v}P | \boldsymbol{v} \in \mathcal{M}_{n-1}, P \in \mathcal{M}(P)\}$$

and

$$\mathcal{D}_n := \{\boldsymbol{v}P | \boldsymbol{v} \in \mathcal{D}_{n-1}, P \in \mathcal{M}(P)\}.$$

Furthermore

$$\mathcal{M}_n^C := \{f(\boldsymbol{v}) | \boldsymbol{v} \in \mathcal{M}_n \setminus \{(1, 0, \dots, 0)\}\}$$

and

$$\mathcal{D}_n^C := \{f(\boldsymbol{v}) | \boldsymbol{v} \in \mathcal{D}_n \setminus \{(1, 0, \dots, 0)\}\}.$$

It should be clear that

$$\tilde{\mathcal{M}}_n^C \subseteq \mathcal{M}_n^C, \forall n > 0 \tag{3.1}$$

and moreover that

$$\tilde{\mathcal{M}}_1^C = \mathcal{M}_1^C$$

where $\tilde{\mathcal{M}}_n^C$ is as defined in (2.3).

Lemma 3.2

$$\mathcal{M}_n^C = \{f(\tilde{f}_\alpha(\boldsymbol{v}) \cdot P) | \boldsymbol{v} \in \mathcal{M}_{n-1}^C, P \in \mathcal{M}(P)\}$$

and

$$\mathcal{D}_n^C = \{f(\tilde{f}_\alpha(\boldsymbol{v}) \cdot P) | \boldsymbol{v} \in \mathcal{D}_{n-1}^C, P \in \mathcal{M}(P)\}.$$

Proof. $\boldsymbol{v} \in \mathcal{D}_{n-1}^C \Rightarrow \tilde{f}_\alpha(\boldsymbol{v}) \in \mathcal{D}_{n-1}$ for some $\alpha \in [0,1)$ by definition. Thus $\tilde{f}_\alpha(\boldsymbol{v})P \in \mathcal{D}_n^C$. By Lemma 2.1, however $f(\tilde{f}_\alpha(\boldsymbol{v})P) = f(\tilde{f}_\beta(\boldsymbol{v})P)$ for any $\beta \in [0,1)$, and so in fact $f(\tilde{f}_\alpha(\boldsymbol{v})P) \in \mathcal{D}_n^C$ independently of our choice of α. □

It is proven in [13] that

$$\mathcal{M}_{n+1} \subseteq \mathcal{M}_n$$

making the following definition appropriate.

Definition 3.4

$$\mathcal{M}_\infty := \bigcap_{n=0}^{\infty} \mathcal{M}_n.$$

It is proven in [5] that

$$\mathcal{M}_\infty = \{(1,0,\ldots,0)\}$$

so absorption is certain even when the transition matrix is unknown and can change between time steps. Once again the long-term behaviour of the chain conditioned on non-absorption is considered.

It is proved in [5] (in an almost identical manner to Theorem 2.2) that

$$\mathcal{M}_{n+1}^C \subseteq \mathcal{M}_n^C \tag{3.2}$$

and hence the following definition is appropriate.

Definition 3.5

$$\mathcal{M}_\infty^C := \bigcap_{n=0}^{\infty} \mathcal{M}_n^C.$$

Definition 3.6 A set of distributions $\mathcal{M}$ is denoted a *conditionally invariant set of distributions*, henceforth known as CISD, if

$$f(\tilde{f}_\alpha(\mathcal{M}) \cdot \mathcal{M}(P)) = \mathcal{M}$$

for some α and therefore for every $\alpha \in [0,1)$, where $\cdot$ represents an element-wise product.

Thus if at any time-step the set of possible distributions over C is a CISD every subsequent time-step will have an identical set of possible distributions over C. Note that $\mathcal{M}_\infty^C$ must be a CISD By Lemma 2.1.

$\mathcal{M}_\infty^C$ describes the behaviour of the chain, conditioned on non-absorption, as time approaches infinity, assuming that there is nothing whatsoever that can be said regarding the initial distribution over C. An important property of the limiting conditional distribution in the precise case, however, is that the behaviour of the chain, conditioned on non-absorption, tends toward it independently of the choice of initial distribution over C. In what follows we outline the method by which the generalisation of this property can be proved.

Definition 3.7 Two sets of distributions over S, $\mathcal{M}$ and $\mathcal{N}$, are described as *conditionally equal* if $f(\mathcal{M}) = f(\mathcal{N})$, where $f(\mathcal{M}) := \{f(\boldsymbol{v}) | \boldsymbol{v} \in \mathcal{M}\}$.

A non-symmetric distance measure $d(\cdot,\cdot)$ between two sets of distributions over S is defined in [5], where $d(\mathcal{M},\mathcal{N}) = 0$ if and only if for every $\boldsymbol{v} \in \mathcal{M}$ there is a $\boldsymbol{w} \in \mathcal{N}$ such that $f(\boldsymbol{v}) = f(\boldsymbol{w})$.

Corollary 3.1 Let $\mathcal{M}$ and $\mathcal{N}$ be closed sets of distributions. Then $f(\mathcal{M}) \subseteq f(\mathcal{N})$ if and only if $d(\mathcal{M},\mathcal{N}) = 0$.

Proof. $f(\mathcal{M}) \subseteq f(\mathcal{N})$ implies that for every $f(\boldsymbol{v}) \in \mathcal{M}$ there exists $\boldsymbol{w} \in \mathcal{N}$ such that $f(\boldsymbol{v}) = f(\boldsymbol{w})$. Thus $d(\mathcal{M},\mathcal{N}) = 0$.

Let $d(\mathcal{M},\mathcal{N}) = 0$. By the above assertion, for every $\boldsymbol{v} \in \mathcal{M}$ there exists $\boldsymbol{w} \in \mathcal{N}$ such that $d(\boldsymbol{v},\boldsymbol{w}) = 0$. Thus $f(\mathcal{M}) \subseteq f(\mathcal{N})$.

□

It is proven in [5] that, under the conditions given in this paper

$$d(\mathcal{M} \cdot \mathcal{M}(P), \mathcal{N} \cdot \mathcal{M}(P)) < d(\mathcal{M},\mathcal{N}) \tag{3.3}$$

and

$$f(\mathcal{M}) = f(\mathcal{M}') \Rightarrow d(\mathcal{M},\mathcal{N}) = d(\mathcal{M}',\mathcal{N}) \tag{3.4}$$

for any set of distributions $\mathcal{N}$.

Definition 3.8 Let $\mathcal{M}$ be a compact set of distributions and $\mathcal{M}(P)$ a set of transition matrices that are conditionally regular on C. Then $\mathcal{M}$ is a *fixed set of $\mathcal{M}(P)$ conditionally on C* if $f(\mathcal{M} \cdot \mathcal{M}(P)) = f(\mathcal{M})$, or equivalently, if $\mathcal{M}$ and $\mathcal{M} \cdot \mathcal{M}(P)$ are conditionally equal on C.

It is important to note that if $\mathcal{M}$ is a fixed set of $\mathcal{M}(P)$ conditionally on C, then $f(\mathcal{M})$ must be a conditionally invariant set of distributions.

Theorem 3.1 Let $\mathcal{M}$ and $\mathcal{N}$ be conditionally fixed sets of $\mathcal{M}(P)$ on C. Then they are conditionally equal on C.

Proof. It follows from Corollary 3.1 that the sets $\mathcal{M}$ and $\mathcal{N}$ are conditionally equal on C if and only if $d(\mathcal{M},\mathcal{N}) = d(\mathcal{N},\mathcal{M}) = 0$. Suppose that one of the distances is greater than 0, say $d(\mathcal{M},\mathcal{N}) > 0$. By the

assertion that both sets are conditionally fixed sets, we have that $f(\mathcal{M}) = f(\mathcal{M} \cdot \mathcal{M}(P))$ and $f(\mathcal{N}) = f(\mathcal{N} \cdot \mathcal{N}(P))$ Then, by Corollary 3.1, (3.3) and (3.4), $d(\mathcal{M},\mathcal{N}) = d(\mathcal{M} \cdot \mathcal{M}(P),\mathcal{N}) = d(\mathcal{M} \cdot \mathcal{M}(P),\mathcal{N} \times \mathcal{M}(P)) < d(\mathcal{M},\mathcal{N})$, which is a contradiction. □

Thus we have that there can be only one conditionally invariant set of distributions for a given imprecise Markov chain. Finally, [5] goes on to prove that convergence to this set is certain, conditioned upon non-absorption, irrespective of the choice of $\mathcal{D}_0^C$.

These results confirm the CISD as the imprecise analog to the QSD. Not only does setting $\mathcal{D}_0^C = \mathcal{M}_\infty^C$ ensure that $\mathcal{D}_n^C = \mathcal{M}_\infty^C$, for all n, but the set of possible distributions tends towards $\mathcal{M}_\infty^C$ no matter what initial distributions are allowed.

We now discuss the method offered in [3], and explain why we do not make use of it here. The results presented in Section 3 are based on the notion of regularity defined in Definition 3.2. In this sense the results on convergence directly generalise those found in [7], where the analogous notion of regularity is used in the precise case.

Two important further insights for the case of unconditional convergence of imprecise Markov chains are found in [3], which suggest that the concept of regularity that we use might be too strong. First, de Cooman *et al.* show that even the concept of regularity itself can be transferred to the imprecise case in a weaker form, which suggests that there might be different types of convergence with different properties. However, their approach is substantially different from ours, where the main difference is that they represent imprecision in terms of lower and upper expectation operators instead of sets of probabilities and moreover, the calculations of the distributions at further time steps are done by the use of so called backwards recursion.

While in the case where sets of probabilities are convex, which certainly is the most important case, the representation with expectation operators coincides with the approach with sets of probabilities, our approach is more general if sets of probabilities are not assumed to be convex. Our stronger notion of regularity seems to be necessary in this case to assure convergence. The second problem with efficiently applying the approach taken in [3] to studying convergence under conditioning on non-absorption is that it is not obvious to us how the conditioning that must take place at every step would be combined with the backward recursion method. This effectively means that we do not see how the step performed in Lemma 2.1, which is shown to be easy using the forward calculations, could be done using the backwards recursion. Of course, while the chain is still finite, conditioning can be done at an arbitrary step n, but when convergence is in question as n tends to infinity, it is not clear how and where conditioning can be done, as it is clearly too late to condition at infinity where absorption takes place with certainty. Despite the above difficulties we believe that combining our results with those of de Cooman *et al.* is possible in some way, which is a possible path of our future research.

The second important insight given in [3] is that, in the case without conditioning and even in the precise case, instead of regularity a weaker condition called "regular absorption" is sufficient to assure unique convergence, which also seems to be possible to apply to the problem of unique convergence under conditioning.

4 Comparison between the models

In this section we consider two examples. In the first, movement from all three transient states exhibits imprecise behaviour, but the bounds on that behaviour are comparatively tight. In the second example, movement from only one transient state exhibits imprecise behaviour, but the bounds on that behaviour are comparatively much wider. In each example we consider the difference between applying the model given in Section 2 and that given in Section 3. Note that throughout this section $\mathcal{M}_0^C$ is used as the set of possible initial distributions over C.

In this section simplex diagrams (see e.g. Walley [16]) are used to graphically represent probability distributions with three elements. A simplex diagram is an equilateral triangle with height one unit in which each vertex represents the probability distribution with all mass in one state of C. The probabilities assigned to the three elements of C are identified with perpendicular distances from the three sides of the triangle. Thus the set $\mathcal{M}_0^C$ is represented by the whole simplex diagram.

Example 1

Consider a time-homogeneous birth-death process $\mathcal{X}$ with state space $\Omega = \{-1\} \cup C$ where $C = \{0, 1, 2\}$. The set of all possible one-step transition matrices $\mathcal{M}(P)$ is given as follows. Each $P \in \mathcal{M}(P)$ takes the form

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ a & 0 & 1-a & 0 \\ 0 & b & 0 & 1-b \\ 0 & 0 & c & 1-c \end{pmatrix}$$

where $a \in [0.1, 0.3]$, $b \in [0.5, 0.6]$, and $c \in [0.67, 0.73]$.

Generating either $\tilde{\mathcal{M}}_n^C$ or $\mathcal{M}_n^C$ in their entirety for this example (or any other) is a non-trivial task. There are

several alternative methods that can be used to gain sensible approximations. For instance, the maximum and minimum values of each element of the vectors contained in $\tilde{\mathcal{M}}_n^C$ and $\mathcal{M}_n^C$ can be calculated. The simplex diagrams in Figure 1 below show such approxiations for $\tilde{\mathcal{M}}_n^C$ for $n = 2, 3, 4$ (left column, from top to bottom), and $\mathcal{M}_n^C$, also for $n = 2, 3, 4$ (right column, from top to bottom). Bounds have also been approximated for the sets $\tilde{\mathcal{M}}_{100}^C$ and $\mathcal{M}_{100}^C$. These were found by randomly generating 1000 100-step transition matrices for each of the two cases, multiplying each one by $(1, 0, 0), (0, 1, 0)$ and $(0, 0, 1)$, and finding the overall maximum and minimum of each element. The 100th time step is an excellent approximation to the case as time approaches infinity.

Recall that it is known that the size of the bounded areas are non-increasing from time step n to $n+1$, from (2.3) and (3.2). Figure 1 demonstrates these properties very well. Note also that, as expected, for each times step the bounded areas on the right are larger than those on the left. This again is exactly what was expected given (3.1), and moreover is consistent with the idea that more can be said about the long term behaviour for the case where the transition matrix is constant than can be said for the case where the transition matrix is potentially non-constant between time steps. One could say that the second case allows for "more imprecision," in that less can be assumed about the underlying process.

It is important to note that the variable a does in fact play a role in the example, despite the fact that by conditioning on non-absorption we implicitly assume that every transition from state 0 must have been to state 1. This can be easily seen by noting that

$$f((0,x,y,z)\begin{pmatrix} 1 & 0 & 0 & 0 \\ a & 0 & 1-a & 0 \\ 0 & b & 0 & 1-b \\ 0 & 0 & c & 1-c \end{pmatrix})$$

$$= (\frac{1}{1-ax})(by, (1-a)x + cz, (1-b)y + (1-c)z)$$

and therefore conditioning on non-absorption does not prevent a from contributing to the distribution over C.

Example 2

Consider a time-homogeneous birth-death process $\mathcal{X}$ with state space $\Omega = \{-1\} \cup C$ where $C = \{0, 1, 2\}$. The set of all possible one-step transition matrices $\mathcal{M}(P)$ is given as follows. Each $P \in \mathcal{M}(P)$ has the following form.

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.6 & 0 & 0.4 & 0 \\ 0 & d & 0 & 1-d \\ 0 & 0 & 0.7 & 0.3 \end{pmatrix}$$

where $d \in [0.37, 0.73]$. The diagrams were created using identical methods to those used in the first example.

The same comments regarding Figure 1 also apply to Figure 2. It should also be noted that in the second example more can be said about the probability of being in state 1, conditioned on non-absorption. as time approaches infinity, but less can be said about the probabilities of being in states 1 or 3. This may be explained as follows. Note that in the method used in Section 2, the bounds upon $\tilde{\mathcal{M}}_\infty^C$ are simply the bounds upon the set $\bigcup_{P \in \mathcal{M}(P)} \boldsymbol{\alpha}_P$ (see (2.4)). Thus the bounds approximated in the bottom-left simplex of Figure 1 relate to the three elements of a vector function with three unknowns, a, b and c, all with comparatively small ranges. In comparison, the bounds approximated in the bottom-left simplex of Figure 2 relate to the three elements of a vector function with one unknown, d, which has a comparatively large range. The elongated, thinner shape in Figure 2 is thus intuitively unsurprising, though the validity of this intuition can be questioned, as d will eventually appear in all transition proabilities given enough jumps.

The final point regarding Figures 1 and 2 is the fact that in both the situation in which little is known about one state's behaviour, and in that where no state's behaviour is completely known, there is much that can be said about the long-term behaviour conditioned on non-absorption. It is *not* the case, as may have been feared, that the imprecision grows with each new iteration until there is nothing to be said about a given time-step. Moreover, this is true even when the transition matrix is not assumed to be constant. This is particularly important because it suggests that the model used in Section 3 can be applied to approximating the long-term behaviour of precise time-inhomogeneous chains with an absorbing state, conditioned upon non-absorption, an area in which comparatively little work has been done.

Note that is would also be possible to compare the two models by creating a set of r initial distributions to approximate $\mathcal{M}_0^C$ and a set of s transition matrices to approximate $\mathcal{M}(P)$. These can then be used to create sets of vectors to approximate $\tilde{\mathcal{M}}_n^C$ and $\mathcal{M}_n^C$. The drawback to this method is that it rapidly becomes computationally heavy. In the example above, allowing $\mathcal{M}_0^C$ to be approximated by the 231 vectors $\{\frac{i}{20}, \frac{j}{20}, \frac{k}{20}\}$, where i, j, k are the set of non-negative integers for which $i + j + k = 20$, and allowing $\mathcal{M}(P)$ to be approximated by the 264 matrices for which $a \in [0.1, 012, \ldots, 0.3]$, $b \in [0.5, 0.52, \ldots, 0.7]$, and $c \in [0.67, 0.69, 0.71, 0.73]$, then by the time $n = 4$ there are over a thousand *billion* vectors to calculate.

5 Concluding remarks

In this paper we have summarised two methods in which imprecision can be applied to the theory of Markov chains, and discussed that in each case, given certain conditions and conditioned on non-absorption, convergence to a unique conditionally invariant set is guaranteed, and that using this set as the set of initial distributions, the possible behaviour of the chain is unchanging over time. It has also been demonstrated that, by considering this extension of the QSD, it is possible to say something regarding the long-term behaviour, conditioned on non-absorption, of finite Markov chains with an absorbing state, in situations in which the transition matrix at each time-step is not known precisely. Moreover, it has been shown that much can be said even in situations where the transition matrix is not assumed to be constant over time, and in which there is no transient state from which the transition probabilities are known precisely. This in turn means that the model presented in Section 3 could be applied when considering the long-term behaviour of certain precise time-inhomogeneous chains.

6 Acknowledgements

We are grateful to two referees for their detailed comments which helped to improve this paper.

References

[1] Beer, G. (1993) *Topologies On Closed And Closed Convex Sets*, Kluwer Academic Publishers, Dordrecht.

[2] Coolen-Schrijner, P. & van Doorn, E. (2006) Quasi-stationary distributions for a class of discrete-time Markov chains, *Methodology and Computing in Applied Probability*, **8**, 449-465.

[3] de Cooman, G., Quaeghebeur, E. & Miranda, E. (2009) Imprecise Markov chains and their limit behaviour. *Pre-print at arXiv.org*, arXiv:0801.0980.

[4] Crossman R.J., Coolen-Schrijner P. & Coolen F.P.A. (2009) Time homogeneous birth-death processes with probability intervals and absorbing state, *Journal of Statistical Theory & Practice* **3**, 103-118.

[5] Crossman, R.J. & Škulj D. (2009) Finite discrete-time Markov chains with absorbing state and imprecise probability, in submission, copy available upon request.

[6] Crossman, R.J. (2009) *Limiting Conditional Distributions: Imprecision and Relation to Hazard Rate*, PhD Thesis, Durham University, in submission, copy available upon request.

[7] Darroch, J.N. & Seneta, E. (1965) On quasi-stationary distributions in absorbing discrete-time finite Markov chains, *Journal of Applied Probability* **2**, 88-100.

[8] Hartfiel, D.J. (1998) *Markov Set-Chains*, Springer-Verlag, Berlin, Heidelberg, New York.

[9] Kijima, M. (1997) *Markov Processes for Stochastic Modeling*. Chapman & Hall, London.

[10] Kozine, I.O. & Utkin, L.V. (2002) Interval-valued finite Markov chains, *Reliable Computing 8*, **2**, 97-113.

[11] Pakes, A. (1987) Limit theorems for the population size of a birth and death process allowing catastrophes, *Journal of Mathematical Biology*, 25(3), 307-325.

[12] Parsons, R. & Pollett P. (1987) Quasistationary distributions for auto-catalytic reactions, *Journal of Statistical Physics*, 46(1-2), 249-254.

[13] Škulj, D. (2006) Finite discrete time Markov chains with interval probabilities, *Soft Methods for Integrated Uncertainty Modelling*, Springer, Berlin.

[14] Škulj, D. (2007) Finite discrete time Markov chains with interval probabilities. In J. Lawry, E.Miranda, A. Bugarin, S.Li, M.A. Gil, P.Grzegorzewski, O.Hryniewicz (esd.), *Soft Methods for Integrated Uncertainty Modelling*, 299-306. Springer, Berlin.

[15] Walley, P. (1981) Coherent lower (and upper) probabilities. Department of Statistics, University of Warwick.

[16] Walley, P. (1991) *Statistical Reasoning With Imprecise Probabilities*, Chapman & Hall, London.

[17] Weichselberger, K. (2000) The theory of interval-probability as a unifying concept for uncertainty, *International Journal of Approximate Reasoning*, **24** 149-170.

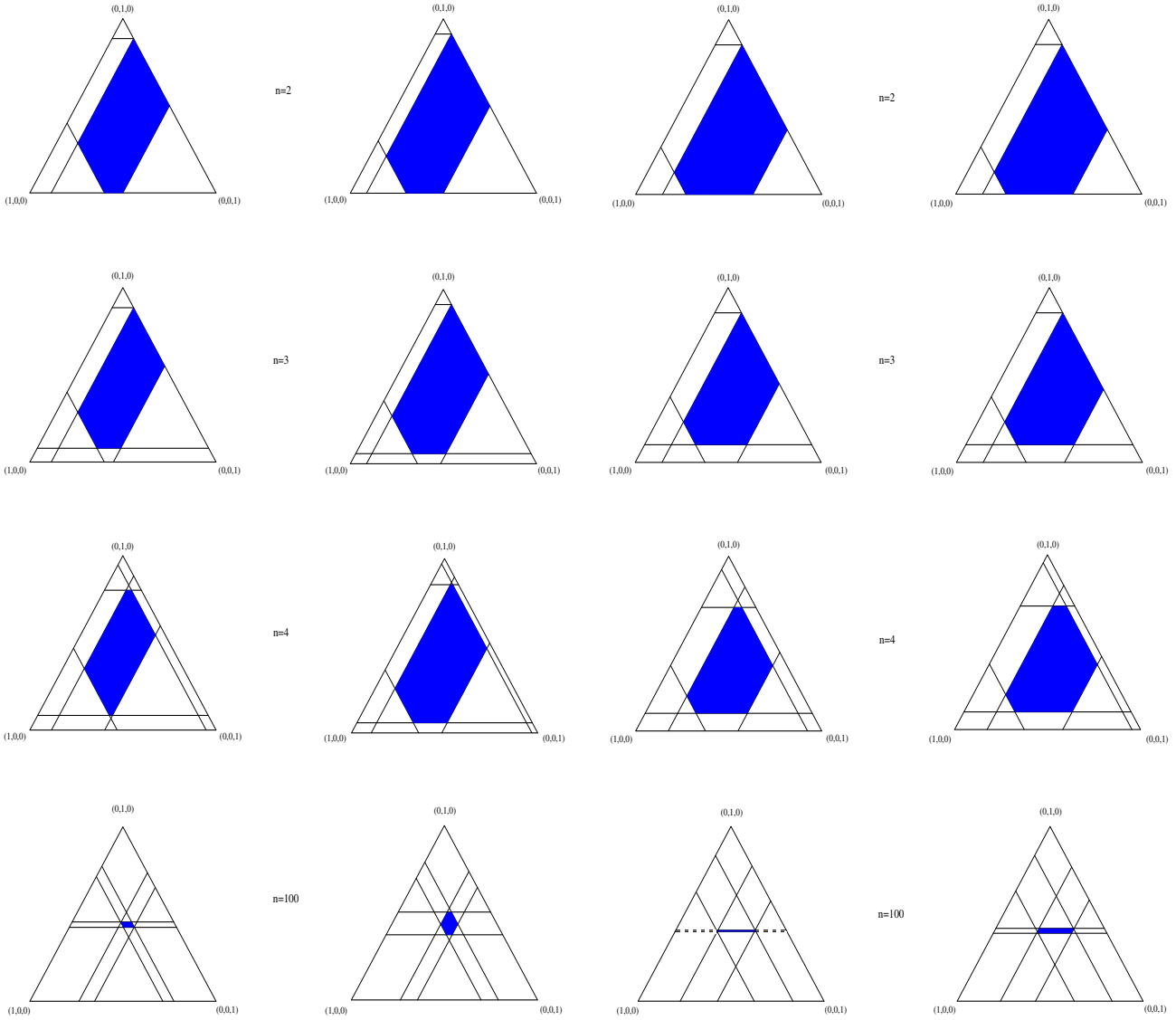


Figure 1: Bounds for the sets $\tilde{\mathcal{M}}_n^C$ and $\mathcal{M}_n^C$, all for $n = 2, 3, 4$ and 100.

Figure 2: Bounds for the sets $\tilde{\mathcal{M}}_n^C$ and $\mathcal{M}_n^C$, all for $n = 2, 3, 4$ and 100.

Credal semantics of Bayesian transformations

Fabio Cuzzolin
Oxford Brookes University, Oxford, UK
fabio.cuzzolin@brookes.ac.uk

Abstract

In this paper we propose a credal representation of the set of interval probabilities associated with a belief function, and show how it relates to several classical Bayesian transformations of belief functions through the notion of "focus" of a pair of simplices. Starting from the interpretation of the pignistic function as the center of mass of the credal set of consistent probabilities, we prove that relative belief and plausibility of singletons and intersection probability can be described as foci of different pairs of simplices in the simplex of all probability measures. Such simplices are associated with the lower and upper probability constraints, respectively. This paves the way for the formulation of frameworks similar to the transferable belief model for lower, upper, and interval constraints.

Keywords. Belief functions, credal sets, Bayesian transformations, upper and lower simplices, focus.

1 Introduction

Consider a given decision or estimation problem Q. We assume that the possible answers to Q form a finite set $\Theta = \{x_1, ..., x_n\}$ called "frame of discernment". Given a certain amount of evidence, we are allowed to describe our belief on the outcome of Q in several possible ways: the classical option is to assume a probability distribution on Θ. However, as we may need to incorporate imprecise measurements and people's opinions in our knowledge state, or cope with missing or scarce information, a more sensible approach is to assume that we have no access to the "correct" probability distribution. The available evidence, though, provides us with some sort of constraint on this true distribution.

Such a constraint is often given in the form of a *credal set*, i.e., the convex set of probability distributions maintained by the agent [14]. A specific class of credal sets is formed by *belief functions* [16]. Even though in their original definition [16] belief functions are defined as set functions $b : 2^\Theta \to [0, 1]$ on the power set 2^Θ of a finite universe Θ, they are equivalent to a set of linear constraints determining a credal set. Belief functions are a popular tool for representing uncertain knowledge under scarce information, as they can naturally cope with ignorance, qualitative judgements, and missing data.

Their credal interpretation is at the core of a widely adopted approach to the theory of evidence, the "Transferable Belief Model" (TBM) [20, 21]. In the TBM, decisions are made by resorting to a probability called "pignistic function". Based on a number of rationality principles, the pignistic function has a nice geometric interpretation as the center of mass of the credal set of probability measures "consistent" with b, i.e. the probabilities that dominate b on all the events A: $\mathcal{P}[b] \doteq \{p \in \mathcal{P} : p(A) \geq b(A)\ \forall A \subseteq \Theta\}$ (here $\mathcal{P}$ denotes the set of all the probability measures on Θ).

The relation between belief and probability measures or "Bayesian belief functions" has been widely studied in the context of the theory of evidence [1, 10, 11, 13, 26, 27], often with different goals. While some authors have looked for efficient implementations of the rule of combination [15, 23], others have argued that Bayesian and belief calculi have the same expressive power as each model can be transformed into the other.

An approach to the Bayesian transformation problem seeks approximations which enjoy commutativity properties with respect to some evidence combination rule, in particular the original Dempster's sum [9]. Voorbraak [24] was probably the first to explore this direction. He proposed to adopt the *relative plausibility of singletons*, i.e., the unique probability that, given a belief function b with plausibility $pl_b : 2^\Theta \to [0, 1]$, $pl_b(A) = 1 - b(A^c)$, assigns to each element $x \in \Theta$ of the domain its normalized plausibility. Cobb and Shenoy later analyzed its properties in detail [3]. More recently, a dual *relative belief of singletons* has been investigated in terms of both its

semantics [7] and its properties with respect to Dempster's rule. The condition under which some of those transformations coincide has been studied in [4].
Unlike the case of the pignistic transformation, a credal semantic is still lacking for most other major Bayesian approximations of belief functions. Moreover, not all such transformations are consistent with the original belief function, i.e., they do not necessarily fall into the corresponding credal set. We address this issue here in the framework of "probability intervals". An admissible constraint on the true probability p which describes the given problem Q can be provided by enforcing lower and upper bounds on its probability values on the elements of the frame Θ. What we get is a set of *probability intervals* [8, 22, 25]:

$$\big\{l(x) \leq p(x) \leq u(x), \forall x \in \Theta\big\}. \tag{1}$$

Probability intervals are themselves a special class of credal sets. Besides, each belief function determines itself such a set of intervals, in which the lower bound to $p(x)$ is the belief value $b(x)$ on x, while its upper bound is the plausibility value $pl_b(x) = 1 - b(\{x\}^c)$:

$$\mathcal{P}[b, pl_b] \doteq \big\{p \in \mathcal{P} : b(x) \leq p(x) \leq pl_b(x), \forall x \in \Theta\big\}. \tag{2}$$

The credal set (2) determined by the set of probability intervals associated with a belief function is strictly related to the credal set of consistent probabilities. More precisely, it is the intersection of two higher-dimensional triangles or "simplices": A "lower simplex" $T^1[b]$ determined by the lower bound constraint $b(x) \leq p(x)$, and an "upper simplex" $T^{n-1}[b]$ determined by the upper bound constraint $p(x) \leq pl_b(x)$.

1.1 Contribution

We can exploit the different credal sets associated with a belief function to provide several important Bayesian transformations with a credal semantic similar to that of the pignistic transformation. In this paper we focus on relative plausibility [24] and belief [7] of singletons, and on the so-called *intersection probability*, a new Bayesian approximation introduced in [5]. We prove that each of the above transformations can be geometrically described in a homogeneous fashion as the *focus* $f(S, T)$ of a pair S, T of simplices, i.e., the unique point which has the same coordinates w.r.t. the two simplices. When the focus of two simplices falls into their intersection, it is the unique intersection of the lines joining corresponding vertices of S and T (see Figure 1).

Here we consider the pairs of simplices $\{\mathcal{P}, T^1[b]\}$, $\{\mathcal{P}, T^{n-1}[b]\}$, $\{T^1[b], T^{n-1}[b]\}$. We prove that, while the relative belief of singletons is the focus of $\{\mathcal{P}, T^1[b]\}$, the relative plausibility of singletons is the focus of $\{\mathcal{P}, T^{n-1}[b]\}$ and the intersection probability that of $\{T^1[b], T^{n-1}[b]\}$.
Their respective focal coordinates encode major features of the underlying belief function: the total mass it assigns to singletons, their total plausibility, and the fraction of the related probability intervals which determines the intersection probability.
This provides a consistent, comprehensive credal semantics for a wide family of Bayesian transformations in terms of geometric loci in the probability simplex. In perspective, this paves the way for TBM-like frameworks based on those same transformations.

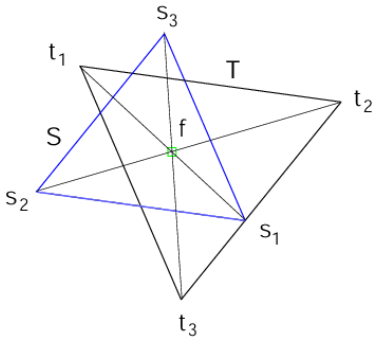


Figure 1: The focus f of a pair of simplices (e.g. two triangles S, T in the 2-D case) is the unique intersection of the lines joining their corresponding vertices.

1.2 Paper outline

We start by recalling the credal interpretation of belief functions and interval probabilities as convex constraints on the value of the unknown probability distribution assumed to describe the problem (Section 2). In particular we focus on the credal sets of probabilities consistent with a belief function and a set of probability intervals, respectively, and introduce what we call the "lower" and "upper" simplices, i.e. the sets of probability measures which meet the lower and upper probability constraints on singletons.
As the pignistic function has a strong credal interpretation in its capacity of center of mass of the polytope of consistent probabilities, we can conjecture the existence of an analogous credal interpretation for other major Bayesian transformations of belief functions (Section 3).
Drawing inspiration from the ternary case, we prove in Section 4 that all the considered probability transformations (relative belief and plausibility of singletons, intersection probability) are geometrically the foci of different pairs of simplices, and discuss the meaning of the map associated with a focus in terms of mass assignment. Finally, in Section 5 we comment on those results, and outline alternative reasoning frameworks

based on the introduced credal interpretations of upper and lower probability constraints and the associated probability transformations.

2 Credal semantics of belief functions and probability intervals

Belief functions and probability intervals are different but related mathematical representations of the bodies of evidence we possess on a given decision or estimation problem Q. They determine different *credal sets* or sets of probability distributions on Θ.

2.1 Credal interpretation of belief functions

A *belief function* (BF) $b : 2^\Theta \to [0,1]$ on a finite set or "frame" Θ has the form

$$b(A) = \sum_{B \subseteq A} m_b(B), \tag{3}$$

where $m_b : 2^\Theta \to [0,1]$ is a set function called "basic probability assignment" (b.p.a.) or "basic belief assignment", and is such that $m_b(A) \geq 0 \ \forall A \subseteq \Theta$ and $\sum_{A \subseteq \Theta} m_b(A) = 1$.
Events $A \subseteq \Theta$ such that $m_b(A) \neq 0$ are called "focal elements". *Bayesian* BFs are belief functions which assign non-zero mass to singletons only: $m_b(A) = 0 \ \forall A : |A| > 1$.
In the following we denote by b_A the unique *categorical* belief function assigning unitary mass to a single event A: $m_{b_A}(A) = 1$, $m_{b_A}(B) = 0 \ \forall B \neq A$. We can then write each belief function b with b.p.a. m_b as [6]

$$b = \sum_{A \subseteq \Theta} m_b(A) b_A. \tag{4}$$

Belief functions have a natural interpretation as constraints on the "true", unknown probability distribution of Q. According to this interpretation the mass assigned to each event $A \subseteq \Theta$ can float freely among its elements $x \in A$. A probability distribution "consistent" with b emerges by redistributing the mass of each focal element to its singletons.

Example. Let us consider a little example, namely a belief function b on a frame of cardinality three $\Theta = \{x, y, z\}$ with focal elements (Figure 2-top): $m_b(\{x,y\}) = \frac{2}{3}$, $m_b(\{y,z\}) = \frac{1}{3}$. One way of obtaining a probability consistent with b is, for instance, to equally share the mass of $\{x,y\}$ among its elements x and y, while attributing the entire mass of $\{y,z\}$ to y (Figure 2-bottom-left). Or, we can assign all the mass of the focal element $\{x,y\}$ to y, and give the mass of $\{y,z\}$ to z only, obtaining the Bayesian belief function of Figure 2-bottom-right.

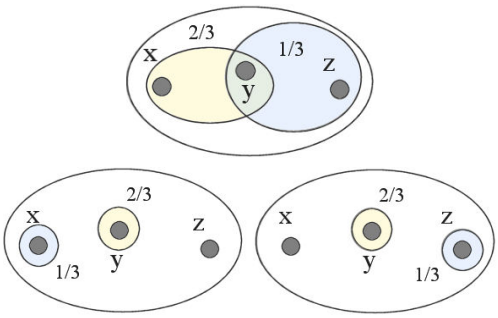


Figure 2: Top: A simple belief function in a frame of size 3. Bottom: two probabilities consistent with it on the same frame.

Belief function as lower bound. The credal set associated with a belief function b (i.e., the set of all the probability distributions consistent with b) is

$$\mathcal{P}[b] = \Big\{ p \in \mathcal{P} : p(A) \geq b(A) \ \forall A \subseteq \Theta \Big\} \tag{5}$$

i.e. the set of distributions whose values dominate that of b on all events A. These are well known to form a polytope in the space $\mathcal{P}$ of all probability measures [2], whose center of mass coincides with the pignistic transformation. Let us denote by Cl the convex closure operator: $Cl(b_1, ..., b_k) = \{b \in \mathcal{B} : b = \alpha_1 b_1 + \cdots + \alpha_k b_k, \sum_i \alpha_i = 1, \ \alpha_i \geq 0 \ \forall i\}$, where $\mathcal{B}$ is the space of all belief functions.

Proposition 1. *The polytope $\mathcal{P}[b]$ of all the probability measures consistent with a belief function b can be expressed as the convex closure $P[b] = Cl(p^\rho[b] \ \forall \rho)$, where ρ is any permutation $(x_{\rho(1)}, ..., x_{\rho(n)})$ of the elements of $\Theta = \{x_1, ..., x_n\}$, and the vertex $p^\rho[b]$ is the unique Bayesian BF such that*

$$p^\rho[b](x_{\rho(i)}) = \sum_{A \ni x_\rho(i), A \not\ni x_\rho(j) \forall j < i} m_b(A). \tag{6}$$

Each probability function (6) assigns to each singletons $x = x_{\rho(i)}$ the mass of all the focal elements of b which contain it, but do not contain the elements which precede x in the ordered list $(x_{\rho(1)}, ..., x_{\rho(n)})$ generated by the permutation ρ.

2.2 Credal interpretation of probability intervals

A *set of probability intervals* provides instead lower and upper bounds for the probability values of the elements of Θ (singletons):

$$\big\{ l(x) \leq p(x) \leq u(x), \forall x \in \Theta \big\}.$$

Any belief function determines itself such a set of intervals, in which the lower bound to $p(x)$ is the belief

value $b(x)$ on x, while its upper bound is the *plausibility value* $pl_b(x)$ of x, $\{b(x) \leq p(x) \leq pl_b(x), \forall x \in \Theta\}$. The plausibility function $pl_b(A) = 1 - b(A^c)$ expresses the evidence not against an event A.
Probability intervals possess themselves a credal representation, which for intervals associated with belief functions is also strictly related to the credal set $\mathcal{P}[b]$ of all consistent probabilities.

Credal form. By definition (5) of $\mathcal{P}[b]$ it follows that the polytope of consistent probabilities can be decomposed into a number of polytopes

$$\mathcal{P}[b] = \bigcap_{i=1}^{n-1} \mathcal{P}^i[b], \tag{7}$$

where $\mathcal{P}^i[b]$ is the set of probabilities meeting the lower probability constraint *for size i events*:

$$\mathcal{P}^i[b] \doteq \{p \in \mathcal{P} : p(A) \geq b(A), \forall A : |A| = i\}.$$

Note that for $i = n$ the constraint is trivially met by all the probability distributions: $\mathcal{P}^n[b] = \mathcal{P}$.
In fact, a simple and elegant geometric description can be given if we consider the credal sets

$$T^i[b] \doteq \{p \in \mathcal{P}' : p(A) \geq b(A), \forall A : |A| = i\}$$

where $\mathcal{P}'$ denotes the set of all *pseudo*-probabilities[1] on Θ, the functions $p : \Theta \to \mathbb{R}$ which meet the normalization constraint $\sum_{x \in \Theta} p(x) = 1$ but not necessarily the non-negativity one: it may exist an element x such that $p(x) < 0$.
In particular we focus here on the set of pseudo-probability measures which meet the lower constraint *on singletons*

$$T^1[b] \doteq \{p \in \mathcal{P}' : p(x) \geq b(x) \quad \forall x \in \Theta\}, \tag{8}$$

and the set $T^{n-1}[b]$ of pseudo-probabilities which meet the lower constraint on events of size $n-1$: $T^{n-1}[b] \doteq$

$$\begin{array}{rl} \doteq & \{p \in \mathcal{P}' : p(A) \geq b(A) \quad \forall A : |A| = n-1\} \\ = & \{p \in \mathcal{P}' : p(\{x\}^c) \geq b(\{x\}^c) \quad \forall x \in \Theta\} \\ = & \{p \in \mathcal{P}' : p(x) \leq pl_b(x) \quad \forall x \in \Theta\}, \end{array} \tag{9}$$

i.e., the set of pseudo-probabilities which meet the *upper bound for the elements x* of Θ.

Simplicial form. The generalization to pseudo-probabilities allows to give the credal sets (8) and (9) the form of *simplices*. A *simplex* is the convex closure of a collection of "affinely independent" points $v_1, ..., v_k$ of a vector space, i.e., points which cannot be expressed as an affine combination of the others:

$$\nexists \left\{\alpha_j, j \neq i : \sum_{j \neq i} \alpha_j = 1\right\} \; s.t. \; v_i = \sum_{j \neq i} \alpha_j v_j.$$

[1] Also called "normalized signed measures" in measure theory.

The notation introduced in Equation (4) is extensively used in the following [4].

Proposition 2. *The credal set $T^1[b]$ or* lower simplex *can be written as the convex closure*

$$T^1[b] = Cl(t^1_x[b], x \in \Theta) \tag{10}$$

of the vertices

$$t^1_x[b] = \sum_{y \neq x} m_b(y) b_y + \left(1 - \sum_{y \neq x} m_b(y)\right) b_x. \tag{11}$$

Dually, the upper simplex $T^{n-1}[b]$ *reads as the convex closure*

$$T^{n-1}[b] = Cl(t^{n-1}_x[b], x \in \Theta) \tag{12}$$

of the vertices

$$t^{n-1}_x[b] = \sum_{y \neq x} pl_b(y) b_y + \left(1 - \sum_{y \neq x} pl_b(y)\right) b_x. \tag{13}$$

To clarify the above results, let us denote by

$$k_b \doteq \sum_{x \in \Theta} m_b(x) \leq 1, \quad k_{pl_b} \doteq \sum_{x \in \Theta} pl_b(x) \geq 1$$

the total mass and plausibility of singletons, respectively. By Equation (11) each vertex $t^1_x[b]$ of the lower simplex is the distribution that adds the mass $1 - k_b$ of non-singletons to the original mass of the element x, leaving all the others unchanged:

$$m_{t^1_x[b]}(x) = m_b(x) + 1 - k_b, \quad m_{t^1_x[b]}(y) = m_b(y) \; \forall y \neq x.$$

As $m_{t^1_x[b]}(z) \geq 0 \; \forall z \in \Theta \; \forall x$ (all the $t^1_x[b]$ are actual probabilities) we have that

$$T^1[b] = \mathcal{P}^1[b] \tag{14}$$

is *completely included* in the probability simplex $\mathcal{P}$.
On the other hand the vertices (13) of the upper simplex are not guaranteed to be valid probabilities, but only *pseudo*-probabilities in the sense that they may assign negative values to some element of Θ. Each vertex $t^{n-1}_x[b]$ assigns to each element of Θ different from x its plausibility $pl_b(y)$, while it subtracts from $pl_b(x)$ the plausibility "in excess" $k_{pl_b} - 1$:

$$\begin{array}{rl} m_{t^{n-1}_x[b]}(x) & = \; pl_b(x) + (1 - k_{pl_b}), \\ m_{t^{n-1}_x[b]}(y) & = \; pl_b(y) \; \forall y \neq x. \end{array}$$

As $1 - k_{pl_b}$ can be a negative quantity, $m_{t^{n-1}_x[b]}(x)$ can be negative too and $t^{n-1}_x[b]$ is not guaranteed to be a "true" probability. We will see this in Section 4.
In conclusion, by Equations (2), (14) and (9) the credal set of probabilities consistent with a probability interval is the intersection[2] $\mathcal{P}[b, pl_b] = T^1[b] \cap T^{n-1}[b]$.

[2] This credal set is an outer approximation [10] of $\mathcal{P}[b]$.

3 Bayesian transformations

The relation between belief and probability measures or "Bayesian belief functions" is central in uncertainty theory [1, 10, 11, 13, 27], and in the theory of evidence [16] in particular.

3.1 Pignistic function as center of mass of consistent probabilities

In Smets' "Transferable Belief Model" [17, 18, 20, 21] beliefs are represented as convex sets of probabilities, while decisions are made by resorting to a Bayesian BF called *pignistic function*:

$$BetP[b](x) = \sum_{A \supseteq \{x\}} \frac{m_b(A)}{|A|}. \tag{15}$$

The rationality principle behind the pignistic function can be explained in terms of the "floating mass" interpretation of focal elements exposed in Section 2.1. If the mass of each focal element is *uniformly* distributed among all its elements, the probability we obtain is (15). The pignistic function $BetP[b]$ has a strong credal interpretation, as it is known [2, 12] to be the center of mass of the set $\mathcal{P}[b]$ of probabilities consistent with b. Many other popular and significant Bayesian functions used to approximate belief functions or to represent them in a decision process, however, *have not* yet a similar credal interpretation. The aim of this paper is indeed to show that relative plausibility [24], relative belief of singletons [7], and intersection probability [5] possess such credal interpretations in terms of the probability intervals associated with a belief function.

3.2 Relative plausibility and belief

The *relative plausibility of singletons* [24] $\tilde{pl}_b$ is the unique probability that, given a belief function b with plausibility pl_b, assigns to each singleton its normalized plausibility:

$$\tilde{pl}_b(x) = \frac{pl_b(x)}{\sum_{y \in \Theta} pl_b(y)} = \frac{pl_b(x)}{k_{pl_b}}. \tag{16}$$

Voorbraak has proven that $\tilde{pl}_b$ is a perfect representative of b when combined with other probabilities through Dempster's orthogonal sum $\oplus$ [9], $\tilde{pl}_b \oplus p = b \oplus p \ \forall p \in \mathcal{P}$. Cobb and Shenoy [3] have later shown that (16) meets a number of other interesting properties with respect to $\oplus$.

Dually, a *relative belief of singletons* $\tilde{b}$ [7] can be defined. This probability function assigns to the elements of Θ their normalized belief values:

$$\tilde{b}(x) \doteq \frac{b(x)}{\sum_{y \in \Theta} b(y)}. \tag{17}$$

Even though the existence of (17) is subject to quite a strong condition

$$k_b = \sum_{x \in \Theta} m_b(x) \neq 0,$$

the case in which $\tilde{b}$ does not exist is indeed pathological, as it excludes a great number of belief and probability measures [7].

While $\tilde{pl}_b$ is associated with the less conservative (but incoherent) scenario in which all the mass that can be assigned to a singleton is actually assigned to it, $\tilde{b}$ reflects the most conservative (but still not coherent) choice of assigning to x only the mass that the BF b (seen as a constraint) assures it belong to x.

It can be proven that relative belief meets a number of dual properties with respect to Dempster's sum which are the dual of those enjoyed by relative plausibility [7]. These two approximations form a strongly linked couple: we will see what this implies in terms of their geometry in the probability simplex.

3.3 Intersection probability

For any set of probability intervals (1) we can define its *intersection probability* as the unique probability of the form $p(x) = l(x) + \alpha(u(x) - l(x))$ for all $x \in \Theta$ for some $\alpha \in [0, 1]$ such that:

$$\sum_{x \in \Theta} p(x) = \sum_{x \in \Theta} \Big[l(x) + \alpha(u(x) - l(x)) \Big] = 1$$

(see Figure 3). This corresponds to the reasonable request that the desired probability, as a candidate to represent the set of intervals (1), should behave homogeneously for each element x of the domain. When

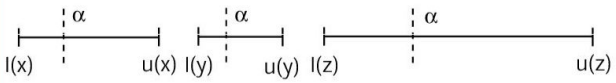


Figure 3: An illustration of the notion of intersection probability for an upper/lower probability system.

the set of intervals is that associated with a belief function, the upper bound to the probability of a singleton is obviously $u(x) = pl_b(x)$, its lower bound $l(x) = b(x) = m_b(x)$. The intersection probability can then be written as [5]

$$p[b](x) = \beta[b] pl_b(x) + (1 - \beta[b]) b(x) \tag{18}$$

as the quantity α of Figure 3 has value

$$\beta[b] = \frac{1 - k_b}{k_{pl_b} - k_b}. \tag{19}$$

Here k_{pl_b}, k_b denote again the total plausibility and belief of singletons, respectively.
The ratio $\beta[b]$ (19) measures the fraction of *each* probability interval which we need to add to the lower bound $b(x)$ to obtain a valid distribution.
Another interpretation of the intersection probability comes from its alternative form

$$p[b](x) = b(x) + \Big(1 - \sum_x b(x)\Big) R[b](x) \tag{20}$$

where

$$R[b](x) \doteq \frac{pl_b(x) - b(x)}{k_{pl_b} - k_b} = \frac{pl_b(x) - b(x)}{\sum_y (pl_b(y) - b(y))}. \tag{21}$$

The quantity $pl_b(x) - b(x)$ measures the width of the probability interval on x, i.e., the uncertainty on the probability value on each element of Θ. Then $R[b](x)$ indicates how much the uncertainty on the probability value on x "weights" on the total uncertainty associated with the set of intervals (1). It is the natural to call it *relative uncertainty* on singletons.
According to (20), $p[b]$ re-distributes to each $x \in \Theta$ a fraction of the mass of non-singletons $(1 - \sum_x b(x))$ in proportion to the relative uncertainty $R[b](x)$ of each singleton in the set of intervals.

4 Credal interpretation of Bayesian approximations

4.1 The ternary case

Taking inspiration from the important case of the pignistic transformation, here we will be able to prove that other Bayesian transformations of belief functions possess a similar credal interpretation.
Let us first analyze the case of a frame of cardinality three: $\Theta = \{x, y, z\}$. Consider the BF

$$\begin{array}{ccc} m_b(x) = 0.2, & m_b(y) = 0.1, & m_b(z) = 0.3, \\ m_b(\{x,y\}) = 0.1, & m_b(\{y,z\}) = 0.2, & m_b(\Theta) = 0.1. \end{array} \tag{22}$$

Figure 4 illustrates the geometry of the related consistent polytope $\mathcal{P}[b]$ in the simplex $Cl(b_x, b_y, b_z)$ of all probability measures on Θ. By Proposition 1 $\mathcal{P}[b]$ has as vertices $\rho^1, \rho^2, \rho^3, \rho^4, \rho^5[b]$

$$\begin{array}{l} \rho^1 = (x,y,z), \\ \quad \rho^1[b](x) = .4, \ \ \rho^1[b](y) = .3, \ \ \rho^1[b](z) = .3; \\ \rho^2 = (x,z,y), \\ \quad \rho^2[b](x) = .4, \ \ \rho^2[b](y) = .1, \ \ \rho^2[b](z) = .5; \\ \rho^3 = (y,x,z), \\ \quad \rho^3[b](x) = .2, \ \ \rho^3[b](y) = .5, \ \ \rho^3[b](z) = .3; \\ \rho^4 = (z,x,y), \\ \quad \rho^4[b](x) = .3, \ \ \rho^4[b](y) = .1, \ \ \rho^4[b](z) = .6; \\ \rho^5 = (z,y,x), \\ \quad \rho^5[b](x) = .2, \ \ \rho^5[b](y) = .2, \ \ \rho^5[b](z) = .6; \end{array} \tag{23}$$

(as the permutations (y,x,z) and (y,z,x) yield the same probability distribution). We can notice that:

1. $\mathcal{P}[b]$ (the polygon delimited by little squares) is the intersection of two triangles (2-dimensional simplices) $T^1[b]$ and $T^2[b]$;
2. the relative belief of singletons

   $$\tilde{b}(x) = \frac{1}{3}, \ \ \tilde{b}(y) = \frac{1}{6}, \ \ \tilde{b}(z) = \frac{1}{2}$$

   is the *intersection of the lines joining the corresponding vertices of the probability simplex $\mathcal{P}$ and the lower simplex $T^1[b]$*;
3. the relative plausibility of singletons

   $$\tilde{pl}_b(x) = \frac{4}{15}, \ \ \tilde{pl}_b(y) = \frac{1}{3}, \ \ \tilde{pl}_b(z) = \frac{2}{5}$$

   is the *intersection of the lines joining the corresponding vertices of $\mathcal{P}$ and upper simplex $T^2[b]$*;
4. finally, the intersection probability

   $$\begin{array}{rcl} p[b](x) & = & m_b(x) + \beta[b](m_b(\{x,y\}) + m_b(\Theta)) \\ & = & .2 + \frac{.4}{1.5 - 0.4} 0.2 = .27, \\ p[b](y) & = & .1 + \frac{.4}{1.1} 0.4 = .245, \ \ p[b](z) = .485 \end{array}$$

   is the unique intersection of the lines joining the corresponding vertices of upper $T^2[b]$ and lower $T^1[b]$ simplices.

Point 1. is easily explained by noticing that in the ternary case, by Equation (7), $\mathcal{P}[b] = T^1[b] \cap T^2[b]$. Figure 4 suggests that $\tilde{b}$, $\tilde{pl}_b$ and $p[b]$ might be consistent with b, i.e. they could lie inside the consistent simplex $\mathcal{P}[b]$. This, though, is not guaranteed to be true in the general case.

Theorem 1. *The relative belief of singletons is not always consistent.*

A counterexample similar to that of the proof of Theorem 1 can be found for $\tilde{pl}_b$. The inconsistency of relative belief and plausibility is due to the fact that those functions only constrain the probabilities of singletons, not considering higher size events as full belief functions do. Indeed these approximations $\tilde{b}$, $\tilde{pl}_b$, $p[b]$ are *consistent with the set of probability intervals associated with b*:

$$\tilde{b}, \tilde{pl}_b, p[b] \in \mathcal{P}[b, pl_b] = T^1[b] \cap T^{n-1}[b].$$

Their geometric behavior, described by points 2., 3. and 4., holds in the general case too.

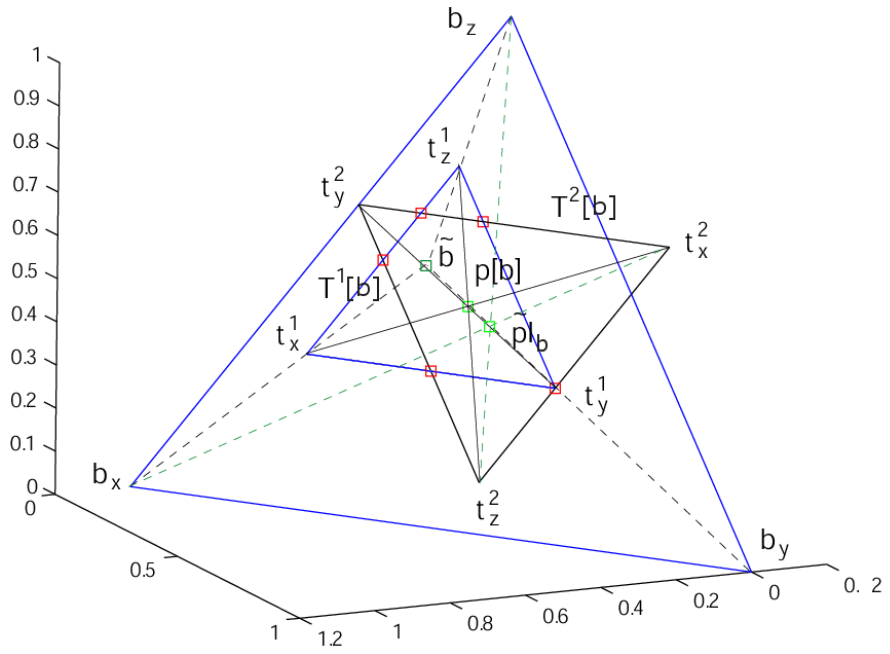


Figure 4: The polytope of all the probabilities consistent with the belief function (22) is shown here in the simplex $\mathcal{P} = Cl(b_x, b_y, b_z)$ of all probability distributions on $\Theta = \{x, y, z\}$. Its vertices (red squares) are given by Equation (23). Intersection probability $p[b]$, relative belief $\tilde{b}$ and plausibility $\tilde{pl}_b$ of singletons are the foci of the pairs of simplices $\{T^1[b], T^2[b]\}$, $\{T^1[b], \mathcal{P}\}$ and $\{\mathcal{P}, T^2[b]\}$ respectively. In the ternary case the lower and upper simplices $T^1[b]$ and $T^2[b]$ are nothing but triangles. Their focus is geometrically the intersection of the lines joining corresponding vertices (dashed lines for $\{\mathcal{P}, T^1[b]\}$ and $\{\mathcal{P}, T^2[b]\}$, solid ones for $\{T^1[b], T^2[b]\}$).

4.2 Focus of a pair of simplices

In the ternary case relative belief, plausibility and intersection probability lie in the intersection of the lines joining corresponding vertices of pairs formed by the upper simplex, the lower simplex, or the probability simplex. This remark can be formalized through the notion of *focus* of a pair of simplices, laying the foundations for a credal interpretation of these three Bayesian transformations.

Definition 1. *Consider a pair of simplices* $S = Cl(s_1, ..., s_n)$, $T = Cl(t_1, ..., t_n)$ *in* $\mathbb{R}^{n-1}$.
We call focus *of the pair* (S, T) *the unique point* $f(S, T)$ *of* $S \cap T$ *which has the same affine coordinates* $\{\alpha_1, ...\alpha_n\}$ *in both simplices:*

$$f(S,T) = \sum_{i=1}^{n} \alpha_i s_i = \sum_{i=1}^{n} \alpha_i t_j, \qquad \sum_{i=1}^{n} \alpha_i = 1. \quad (24)$$

Such point always exists, even though it does not always fall into the intersection of the two simplices. In the latter case, though, the focus coincides with the unique intersection of the lines joining corresponding vertices of S and T (see Figure 1 again).
Suppose indeed that a point p is such that

$$p = \alpha s_i + (1-\alpha) t_i, \qquad \forall\, i = 1, ..., n \quad (25)$$

(i.e. p lies on the line passing through s_i and t_i $\forall i$). Then necessarily $t_i = \frac{1}{1-\alpha}[p - \alpha s_i]$ $\forall\, i = 1, ..., n$. If p has coordinates $\{\alpha_i, i = 1, ..., n\}$ in T, $p = \sum_{i=1}^{n} \alpha_i t_i$, then

$$\begin{aligned} p &= \sum_{i=1}^{n} \alpha_i t_i = \frac{1}{1-\alpha} \sum_i \alpha_i [p - \alpha s_i] \\ &= \tfrac{1}{1-\alpha} [p \textstyle\sum_i \alpha_i - \alpha \sum_i \alpha_i s_i] \\ &= \frac{1}{1-\alpha} [p - \alpha \sum_i \alpha_i s_i] \end{aligned}$$

which implies $p = \sum_i \alpha_i s_i$, i.e. p is the focus of (S, T). Notice that the center of mass itself of a simplex is a special case of focus. Indeed, the center of mass of a d-dimensional simplex S is the intersection of the medians of S, i.e. the lines joining each vertex with the center of mass of the opposite ($d-1$ dimensional) face (see Figure 5). But those centers of mass for all $d-1$ dimensional faces form themselves the vertices of a simplex T. Therefore, the pignistic function itself can be thought of as the focus of two simplices.

4.3 Bayesian transformations as foci

Theorem 2. *The relative belief of singletons is the focus of the pair of simplices* $\{\mathcal{P}, T^1[b]\}$.

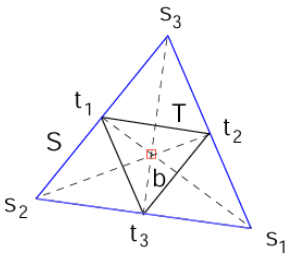


Figure 5: The center of mass b of a simplex S is the focus of the simplex itself and the simplex T formed by the centers of mass of all its $n-1$-dimensional faces. Here a 2-dimensional example is shown.

A dual result can be proven for the relative plausibility of singletons.

Theorem 3. *The relative plausibility of singletons is the focus of the pair of simplices* $\{\mathcal{P}, T^{n-1}[b]\}$.

The notion of focus of upper and lower simplices provides indeed the desired credal semantics for the family of Bayesian transformations linked to Dempster's rule of combination, in terms of the credal set associated with the related set of probability intervals.
The coordinate of the focus on the intersecting lines also has a meaning in terms of degrees of belief.

Theorem 4. *The affine coordinate of* $\tilde{b}$ *as focus of* $\{\mathcal{P}, T^1[b]\}$ *on the corresponding intersecting lines is the inverse of the total mass of singletons.*

A similar result holds for the relative plausibility of singletons.

Theorem 5. *The affine coordinate of* $\tilde{pl}_b$ *as focus of* $\{\mathcal{P}, T^{n-1}[b]\}$ *on the corresponding intersecting lines is the inverse of the total plausibility of singletons.*

An analogous result has recently been proven [4] for the intersection probability (18).

Proposition 3. *The intersection probability is the focus of the pair of simplices* $\{T^{n-1}[b], T^1[b]\}$.

As we could have expected, the line coordinate of the intersection probability as a focus also corresponds to a basic feature of the underlying belief function (or better, the associated set of probability intervals).

Theorem 6. *The coordinate of the intersection probability as focus of* $\{T^1[b], T^{n-1}[b]\}$ *on the corresponding intersecting lines is the ratio* $\beta[b]$ *(19).*

The fraction of the uncertainty of the singletons that generates the intersection probability *can be read in the probability simplex*, as its coordinate on any the lines determining the focus of $\{T^1[b], T^{n-1}[b]\}$.

5 Comments and conclusions

The notion of focus of a pair of simplices provides a unifying geometric framework for a number of different Bayesian transformations of belief functions. In fact, as we pointed out here, it is more correct to think of relative belief, plausibility, and intersection probability as transformations/approximations/representatives of lower, upper, and interval probability systems respectively. While $\tilde{b}$, $\tilde{pl}_b$ and $p[b]$ are potentially inconsistent with the original BF, they are perfectly consistent with the associated lower/upper probability systems (as they fall into the corresponding credal set). Therefore we can argue that simply replacing the pignistic transform with a different transformation when operating on BFs in the TBM would not be semantically correct.

The geometric notion of focus has a simple semantic in terms of probability constraints. Selecting the focus of two simplices representing two different constraints (i.e., the point with the same convex coordinates in the two simplices) means adopting the single probability distribution which meets both constraints *in exactly the same way*. Notice that the second constraint can be empty, like in the case of upper or lower probability systems. If we assume homogeneous behavior in the two sets of constraints as a rationality principle for a probability transformation, then the above Bayesian functions follow as the necessary unique solutions to the corresponding transformation problems. The notion can be easily extended to more than two constraints.

Finally, the credal interpretation of upper, lower, and interval probability constraints on singletons lays in perspective the foundations of the formulation of TBM-like frameworks for such systems.
In the Transferable Belief Model belief functions b are represented by their credal sets, while decisions are made through the corresponding center of mass, the pignistic function $BetP[b]$: $\{\mathcal{P}[b], BetP[b]\}$. We can therefore imagine similar frameworks

$$\left\{\{\mathcal{P}, T^1[b]\}, \tilde{b}\right\}, \quad \left\{\{\mathcal{P}, T^{n-1}[b]\}, \tilde{pl}_b\right\}, \quad \left\{\{T^1[b], T^{n-1}[b]\}, \tilde{pl}_b\right\} \tag{26}$$

in which lower, upper, and interval constraints on a probability distribution on $\mathcal{P}$ are represented by the associated credal sets. This would involve replacing the TBM's disjunctive/conjunctive combination rules [19] by specific evidence elicitation/revision operators for lower, upper, and interval probability systems. Decisions would then be made based on the appropriate focus probability: relative belief, plausibility,

or interval probability respectively.
Notice that, even though in the case of belief functions such systems are simply less informative than the original BF, and their credal sets outer approximations of the credal set of consistent probabilities $\mathcal{P}[b]$, they can be defined independently in their own right, according to the available evidence at hand. In such a case, the use of the appropriate transformation according to the above rationality principle would ensure the consistency of the result. We plan to elaborate on this line of research in the near future.

Appendix: proofs

Proof of Theorem 1. Consider a belief function $b: 2^\Theta \to [0,1]$, $\Theta = \{x_1, x_2, ..., x_n\}$ such that $m_b(x_i) = k_b/n$, $m_b(\{x_1, x_2\}) = 1 - k_b$. Then

$$b(\{x_1,x_2\}) = 2 \cdot \frac{k_b}{n} + 1 - k_b = 1 - k_b\left(\frac{n-2}{n}\right),$$
$$\tilde{b}(x_1) = \tilde{b}(x_2) = \frac{1}{n} \Rightarrow \tilde{b}(\{x_1,x_2\}) = \frac{2}{n}.$$

For $\tilde{b}$ to be consistent with b it is necessary that $\tilde{b}(\{x_1,x_2\}) \geq b(\{x_1,x_2\})$, i.e.

$$\frac{2}{n} \geq 1 - k_b\frac{n-2}{n} \Rightarrow k_b \geq 1$$

i.e. $k_b = 1$. Therefore if $k_b < 1$ (b is not a probability) its relative belief of singletons is not consistent.

Proof of Theorem 2. We need to prove that $\tilde{b}$ has the same simplicial coordinates in $\mathcal{P}$ and $T^1[b]$. By definition (17) $\tilde{b}$ can be expressed in terms of the vertices of the probability simplex $\mathcal{P}$ as

$$\tilde{b} = \sum_{x\in\Theta} \frac{m_b(x)}{k_b} b_x.$$

We then need to prove that $\tilde{b}$ can be written as the same affine combination

$$\tilde{b} = \sum_{x\in\Theta} \frac{m_b(x)}{k_b} t^1_x[b]$$

in terms of the vertices $t^1_x[b]$ of $T^1[b]$. Replacing (11) in the above equation yields $\sum_{x\in\Theta} \frac{m_b(x)}{k_b} t^1_x[b] =$

$$= \sum_{x\in\Theta} \frac{m_b(x)}{k_b}\Big[\sum_{y\neq x} m_b(y) b_y + \Big(1 - \sum_{y\neq x} m_b(y)\Big) b_x\Big] =$$
$$= \sum_{x\in\Theta} b_x\Big(\frac{m_b(x)}{k_b}\sum_{y\neq x} m_b(y)\Big) + \sum_{x\in\Theta}\frac{m_b(x)}{k_b} b_x +$$
$$- \sum_{x\in\Theta} b_x\Big(\frac{m_b(x)}{k_b}\sum_{y\neq x} m_b(y)\Big) = \sum_{x\in\Theta}\frac{m_b(x)}{k_b} b_x = \tilde{b}.$$

Proof of Theorem 3. We just need to replace belief with plausibility values in the proof of Theorem 2.

Proof of Theorem 4. In the case of the pair $\{\mathcal{P}, T^1[b]\}$ we can compute the (affine) line coordinate α of $\tilde{b} = f(\mathcal{P}, T^1[b])$ by imposing condition (25). The latter assumes the following form (being $s_i = b_x$, $t_i = t^1_x[b]$): $\sum_{x\in\Theta} \frac{m_b(x)}{k_b} b_x =$

$$= t^1_x[b] + \alpha(b_x - t^1_x[b]) = (1-\alpha) t^1_x[b] + \alpha b_x$$
$$= (1-\alpha)\Big[\sum_{y\neq x} m_b(y) b_y + \big(1 - k_b + m_b(x)\big) b_x\Big] + \alpha b_x$$
$$= b_x\Big[(1-\alpha)\big(1 - k_b + m_b(x)\big) + \alpha\Big] +$$
$$+ \sum_{y\neq x} m_b(y)(1-\alpha) b_y,$$

and for $1 - \alpha = \frac{1}{k_b}$, $\alpha = \frac{k_b - 1}{k_b}$ the condition is met.

Proof of Theorem 5. Again we can compute the line coordinate α of $\tilde{pl}_b = f(\mathcal{P}, T^{n-1}[b])$ by imposing condition (25). The latter assumes the form (being $s_i = b_x$, $t_i = t^{n-1}_x[b]$): $\sum_{x\in\Theta} \frac{pl_b(x)}{k_{pl_b}} b_x =$

$$= t^{n-1}_x[b] + \alpha(b_x - t^{n-1}_x[b]) = (1-\alpha) t^{n-1}_x[b] + \alpha b_x$$
$$= (1-\alpha)\Big[\sum_{y\neq x} pl_b(y) b_y + \big(1 - k_{pl_b} + pl_b(x)\big) b_x\Big] + \alpha b_x$$
$$= b_x\Big[(1-\alpha)\big(1 - k_{pl_b} + pl_b(x)\big) + \alpha\Big] +$$
$$+ \sum_{y\neq x} pl_b(y)(1-\alpha) b_y,$$

and for $1 - \alpha = \frac{1}{k_{pl_b}}$, $\alpha = \frac{k_{pl_b} - 1}{k_{pl_b}}$ the condition is met.

Proof of Theorem 6. Again, we need to impose condition (25) on the pair $\{T^1[b], T^{n-1}[b]\}$, or

$$p[b] = t^1_x[b] + \alpha(t^{n-1}_x[b] - t^1_x[b]) = (1-\alpha) t^1_x[b] + \alpha t^{n-1}_x[b]$$

for all the elements $x \in \Theta$ of the frame, α being some constant. This is equivalent to (after replacing the expressions (11), (13) of $t^1_x[b]$ and $t^{n-1}_x[b]$)

$$\sum_{x\in\Theta} b_x\big[m_b(x) + \beta[b](pl_b(x) - m_b(x))\big] =$$
$$= (1-\alpha)\Big[\sum_{y\neq x} m_b(y) b_y + (1 - k_b + m_b(x)) b_x\Big] +$$
$$+\alpha\Big[\sum_{y\neq x} pl_b(y) b_y + \Big(1 - \sum_{y\neq x} pl_b(y)\Big) b_x\Big]$$
$$= (1-\alpha)\Big[\sum_{y\in\Theta} m_b(y) b_y + (1 - k_b) b_x\Big] +$$
$$+\alpha\Big[\sum_{y\in\Theta} pl_b(y) b_y + (1 - k_{pl_b}) b_x\Big]$$
$$= b_x\Big[(1-\alpha)(1-k_b) + (1-\alpha) m_b(x) + \alpha pl_b(x) +$$
$$+\alpha(1 - k_{pl_b})\Big] + \sum_{y\neq x} b_y\big[(1-\alpha) m_b(y) + \alpha pl_b(y)\big]$$
$$= b_x\Big\{(1 - k_b) + m_b(x) +$$
$$+\alpha\big[pl_b(x) + (1 - k_{pl_b}) - m_b(x) - (1 - k_b)\big]\Big\} +$$
$$+ \sum_{y\neq x} b_y\big[m_b(y) + \alpha(pl_b(y) - m_b(y))\big].$$

If we set $\alpha = \beta[b] = \frac{1-k_b}{k_{pl_b}-k_b}$ we get for the coefficient of b_x (the probability value of x)

$$\frac{1-k_b}{k_{pl_b}-k_b}\big[pl_b(x) + (1-k_{pl_b}) - m_b(x) - (1-k_b)\big] + \\ +(1-k_b) + m_b(x) = \beta[b]\big[pl_b(x) - m_b(x)\big] + (1-k_b) \\ +m_b(x) - (1-k_b) = p[b](x)$$

while obviously $m_b(y) + \alpha(pl_b(y) - m_b(y)) = m_b(y) + \beta[b](pl_b(y) - m_b(y)) = p[b](y)$ for all $y \neq x$, no matter the choice of x.

References

[1] M. Bauer, *Approximation algorithms and decision making in the Dempster-Shafer theory of evidence–an empirical study*, IJAR **17** (1997), 217–237.

[2] A. Chateauneuf and J.Y. Jaffray, *Some characterization of lower probabilities and other monotone capacities through the use of Moebius inversion*, Math. Soc. Sci. **17** (1989), 263–283.

[3] B. Cobb and P.P. Shenoy, *On the plausibility transformation method for translating belief function models to probability models*, IJAR **41** (2006), no. 3, 314–330.

[4] F. Cuzzolin, *Rationale and properties of the intersection probability*, submitted to AI (2007).

[5] F. Cuzzolin, *Two new Bayesian approximations of belief functions based on convex geometry*, IEEE Trans. SMC-B **37** (2007), no. 4, 993–1008.

[6] F. Cuzzolin, *A geometric approach to the theory of evidence*, IEEE Trans. SMC-C **38** (2008), no. 4, 522–534.

[7] F. Cuzzolin, *Semantics of the relative belief of singletons*, International Workshop on Uncertainty and Logic UNCLOG'08, Kanazawa, Japan, 2008.

[8] L. de Campos, J. Huete, and S. Moral, *Probability intervals: a tool for uncertain reasoning*, IJUFKS **1** (1994), 167–196.

[9] A.P. Dempster, *Upper and lower probabilities generated by a random closed interval*, Annals of Mathematical Statistics **39** (1968), 957–966.

[10] T. Denoeux, *Inner and outer approximation of belief structures using a hierarchical clustering approach*, IJUFKS **9** (2001), no. 4, 437–460.

[11] T. Denoeux and A. Ben Yaghlane, *Approximating the combination of belief functions using the Fast Moebius Transform in a coarsened frame*, IJAR **31** (2002), no. 1-2, 77–101.

[12] D. Dubois, H. Prade, and Ph. Smets, *New semantics for quantitative possibility theory.*, ISIPTA, 2001, pp. 152–161.

[13] R. Haenni and N. Lehmann, *Resource bounded and anytime approximation of belief function computations*, IJAR **31** (2002), no. 1-2, 103–154.

[14] I. Levi, *The enterprise of knowledge*, MIT Press, 1980.

[15] J.D. Lowrance, T.D. Garvey, and T.M. Strat, *A framework for evidential-reasoning systems*, Proceedings of the National Conference on Artificial Intelligence, 1986, pp. 896–903.

[16] G. Shafer, *A mathematical theory of evidence*, Princeton University Press, 1976.

[17] Ph. Smets, *Belief functions versus probability functions*, Uncertainty and Intelligent Systems, Springer Verlag, Berlin, 1988, pp. 17–24.

[18] Ph. Smets, *Constructing the pignistic probability function in a context of uncertainty*, Uncertainty in Artificial Intelligence, 5, Elsevier Science Publishers, 1990, pp. 29–39.

[19] Ph. Smets, *Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem*, IJAR **9** (1993), 1–35.

[20] Ph. Smets, *Decision making in the TBM: the necessity of the pignistic transformation*, IJAR **38** (2005), no. 2, 133–147.

[21] Ph. Smets and R. Kennes, *The transferable belief model*, AI **66** (1994), 191–234.

[22] B. Tessem, *Interval probability propagation*, IJAR **7** (1992), 95–120.

[23] B. Tessem, *Approximations for efficient computation in the theory of evidence*, AI **61** (1993), no. 2, 315–329.

[24] F. Voorbraak, *A computationally efficient approximation of Dempster-Shafer theory*, Int. J. on Man-Machine Studies **30** (1989), 525–536.

[25] K. Weichselberger and S. Pohlmann, *A methodology for uncertainty in knowledge-based systems*, Lecture Notes in Artificial Intelligence, vol. 419, Springer, Berlin, 1990.

[26] T. Weiler, *Approximation of belief functions*, IJUFKS **11** (2003), no. 6, 749–777.

[27] A. Ben Yaghlane, T. Denoeux, and K. Mellouli, *Coarsening approximations of belief functions*, ECSQARU'2001, 2001, pp. 362–373.

Consistent approximations of belief functions

Fabio Cuzzolin
Oxford Brookes University, Oxford, UK
fabio.cuzzolin@brookes.ac.uk

Abstract

Consistent belief functions represent collections of coherent or non-contradictory pieces of evidence. As most operators used to update or elicit evidence do not preserve consistency, the use of consistent transformations $cs[\cdot]$ in a reasoning process to guarantee coherence can be desirable. Such transformations are turn linked to the problem of approximating an arbitrary belief function with a consistent one.
We study here the consistent approximation problem in the case in which distances are measured using classical L_p norms. We show that, for each choice of the element we want them to focus on, the partial approximations determined by the L_1 and L_2 norms coincide, and can be interpreted as classical focused consistent transformations. Global L_1 and L_2 solutions do not in general coincide, however, nor are they associated with the highest plausibility element.

Keywords. Consistent belief function, simplicial complex, approximation, L_p norms.

1 The consistent approximation problem

Belief functions (b.f.s) [19] are complex objects, in which different and sometimes contradictory bodies of evidence may coexist, as they mathematically describe the fusion of possibly conflicting expert opinions and/or imprecise/ corrupted measurements, etcetera. Making decisions based on such objects can then be misleading. This is a well known problem in classical logics, where the application of inference rules to inconsistent sets of assumptions or "knowledge bases" may lead to incompatible conclusions, depending on the subset of assumptions we start our reasoning from.
Consistent belief functions (cs.b.f.s), i.e. belief functions whose non-zero mass events or "focal elements" have non-empty intersection or "core", are then particularly interesting as they represent collections of coherent or non-contradictory pieces of evidence. In some situations it may then be desirable to design a method which, given an arbitrary belief function b, generates a consistent or non-contradictory belief function $cs[b]$: we call this *consistent transformation*. Such a transformation is all the more valuable as several important operators used to update or elicit evidence represented as belief measures, like Dempster's sum [8] and disjunctive combination [21], do not preserve consistency. To guarantee the consistency of a state of belief we may want to seek a scheme in which each time new evidence is combined to yield a new b.f., the consistent transformation $cs[\cdot]$ is applied to reduce it to a coherent knowledge state.
Now, consistent transformations can be built by solving a minimization problem of the form $cs[b] = \arg\min_{cs \in \mathcal{CS}} dist(b, cs)$, where $dist$ is some distance measure between belief functions, and $\mathcal{CS}$ denotes the collection of all consistent b.f.s. We call this the *consistent approximation problem*. By plugging in different distance functions we get different consistent transformations.

In this paper, in particular, we study what happens when using classical L_p norms. Indeed, consistent belief functions correspond to possibility distributions (Section 2), which are in turn inherently related to the L_∞ norm. Besides, the region of all cs.b.f.s is geometrically the set of belief functions for which the L_∞ norm of the plausibility distribution is equal to 1. We can then conjecture that L_p consistent approximations will be meaningful in terms of degrees of belief. This is indeed the case.

From a technical point of view, consistent b.f.s do not live in a single linear space, but in a collection of higher-dimensional triangles or simplices, called "simplicial complex" [11]. A partial solution has then to be found separately for each maximal simplex $\mathcal{CS}_x$ of the consistent complex $\mathcal{CS}$, i.e., the set of cs.b.f.s whose core includes the element x. These partial solu-

tions are later to be compared to determine the global optimal solution.

We will prove here that the partial approximations determined by both the L_1 and the L_2 norms are unique and coincide. We will also prove that the L_1/L_2 consistent approximation onto each component $\mathcal{CS}_x$ of $\mathcal{CS}$ generates indeed the *consistent transformation focused on* x [10, 1], i.e. a new belief function whose focal elements have the form $A' = A \cup \{x\}$, where A is a focal element of the original b.f. b. As we will see, though, the associated global L_1/L_2 solutions do not lie in general on the same component of the consistent complex.

1.1 Paper outline

After recalling the notions of consistent and consonant belief functions, we will recall their semantics and stress why it can be desirable to transform a generic belief function into a consistent one (Section 2). As we pose the approximation problem in a geometric framework, we will briefly recall in Section 3 the geometry of consistent b.f.s. As the latter form a complex, we need to solve the approximation problem separately for each maximal simplicial component of such complex (Section 4). After gaining some insight from the analysis of the binary case (Section 5), we will proceed to solve the L_1 and L_2 consistent approximation problems in the general case in Section 6. We will finally comment and interpret our results.

2 Semantics of consistent belief functions

2.1 Consistent belief functions

We first recall the basic notions of the theory of evidence, and the definition of consistent belief functions in particular, to later discuss their semantics [19].

Definition 1 *A* basic probability assignment *(b.p.a.) on a finite set* (frame of discernment [19]) Θ *is a set function* $m_b : 2^\Theta \to [0,1]$ *on* $2^\Theta \doteq \{A \subseteq \Theta\}$ *s.t.*

$$m_b(\emptyset) = 0, \quad \sum_{A \subseteq \Theta} m_b(A) = 1, \quad m_b(A) \geq 0 \; \forall A \subseteq \Theta.$$

Subsets of Θ associated with non-zero values of m_b are called *focal elements* (f.e.), and their intersection *core*:

$$\mathcal{C}_b \doteq \bigcap_{A \subseteq \Theta : m_b(A) \neq 0} A.$$

Definition 2 *The* belief function *(b.f.)* $b : 2^\Theta \to [0,1]$ *associated with a basic probability assignment* m_b *on* Θ *is defined as:*

$$b(A) = \sum_{B \subseteq A} m_b(B).$$

A dual mathematical representation of the evidence encoded by a belief function b is the *plausibility function* (pl.f.) $pl_b : 2^\Theta \to [0,1]$, $A \mapsto pl_b(A)$ where

$$pl_b(A) \doteq 1 - b(A^c) = 1 - \sum_{B \subseteq A^c} m_b(B)$$

expresses the amount of evidence *not against* A.
In the theory of evidence a probability function is simply a special belief function assigning non-zero masses to singletons only (*Bayesian* b.f.): $m_b(A) = 0 \; |A| > 1$. *Consonant* belief functions are b.f.s whose f.e.s $A_1 \subset \cdots \subset A_m$ are nested. Consonant b.f.s always have a non-empty core, namely their smallest f.e. A_1. However, not all b.f.s whose core is non-empty are consonant.

Definition 3 *A belief function is said to be* consistent *if its core is non-empty.*

2.2 Semantics of consistent belief functions

Consistent belief functions (cs.b.f.s) form a significant class of b.f.s, for several reasons. On one side, they correspond to possibility distributions, and form therefore with consonant b.f.s the link between evidence and possibility theory. More importantly, though, they are the analogues of consistent, non-contradictory sets of propositions ("knowledge bases") in logics. As maintaining coherence along an inference process is highly desirable, the utility of an operator which maps arbitrary belief functions to consistent ones emerges. This is all the more valuable as several evidence combination rules, like Dempster's sum [8] and disjunctive combination [21] do not preserve consistency. To guarantee the consistency of the knowledge state a scheme like the following (where we use $\oplus$ to denote a valid combination rule) can be brought forward

$$\begin{array}{ccccccc} b_1, b_2 & \to & b_1 \oplus b_2 & & & & \\ & & \downarrow & & & & \\ & & cs[b_1 \oplus b_2], & b_3 & \to & cs[b_1 \oplus b_2] \oplus b_3 & \\ & & & & & \downarrow & \\ & & & & & cs[cs[b_1 \oplus b_2] \oplus b_3] & \end{array} \quad (1)$$

in which when new evidence is combined to yield a new belief state, the consistent transformation $cs[\cdot]$ is applied to ensure coherence.

2.3 Consistent b.f.s and possibility distributions

In possibility theory [9, 14], subjective probability is mathematically described by *possibility measures*, i.e. functions $Pos : 2^\Theta \to [0,1]$ such that $Pos(\emptyset) = 0$, $Pos(\Theta) = 1$ and $Pos\big(\bigcup_i A_i\big) = \sup_i Pos(A_i)$, for any family of subsets $\{A_i | A_i \in 2^\Theta, i \in I\}$, where I is an arbitrary set index.
Each measure Pos is uniquely characterized by a *possibility distribution* $\pi : \Theta \to [0,1]$, $\pi(x) \doteq Pos(\{x\})$, via the formula $Pos(A) = \sup_{x \in A} \pi(x)$.
A central role in the connection between possibility and evidence theory [20, 18, 14, 12, 23, 3] is played by consonant and consistent belief functions. On one side,

Proposition 1 *The plausibility function pl_b associated with a b.f. b is a possibility measure iff b is consonant.*

On the other, after calling *plausibility assignment* $\bar{pl}_b$ the restriction of the plausibility function to singletons $\bar{pl}_b(x) = pl_b(\{x\})$ it can be proven that [13, 5]

Proposition 2 *The plausibility assignment $\bar{pl}_b$ associated with a belief function b is the admissible possibility distribution of a possibility measure iff the b.f. b is consistent.*

Consistent b.f.s are then the counterparts of possibility distributions in the theory of evidence.
A different, powerful semantics comes in terms of consistent knowledge bases.

2.4 Consistent b.f.s as collections of coherent pieces of evidence

Belief functions are complex objects, in which sometimes contradictory bodies of evidence may coexist, as they may result from the fusion of possibly conflicting expert opinions and/or imprecise/corrupted measurements. In formal logics, the application of inference rules to inconsistent sets of assumptions or "knowledge bases" may lead to incompatible conclusions, depending on the subset of assumptions we start from. A variety of approaches to solve this problem have been proposed. These include fragmenting the knowledge base into maximally consistent subsets, limiting the power of the formalism, or adopting non-classical semantics [17, 2]. Paris, on his side, tackles the problem by not assuming each proposition in the knowledge base as a fact, but by attributing to it a certain degree of belief [16]. This leads to something similar to a belief function.
A mechanism able to obtain a consistent knowledge base from an inconsistent one is therefore desirable. In the theory of evidence such a mechanism can be described as an operator

$$cs : \mathcal{B} \to \mathcal{CS}, \quad b \mapsto cs[b]$$

where $\mathcal{B}, \mathcal{CS}$ denote respectively the set of all b.f.s, and that of all cs.b.f.s.

2.5 Consistent belief functions and combination rules

Such a transformation acquires even more importance when we notice that most operators used to update/elicit evidence in the theory of evidence *do not preserve* consistency.

Definition 4 *The* orthogonal sum *or* Dempster's sum *of two belief functions b_1, b_2 is a new belief function $b_1 \oplus b_2$ with b.p.a.*

$$m_{b_1 \oplus b_2}(A) = \frac{\sum_{B \cap C = A} m_{b_1}(B)\, m_{b_2}(C)}{\sum_{B \cap C \neq \emptyset} m_{b_1}(B)\, m_{b_2}(C)},$$

where m_{b_i} denotes the b.p.a. associated with b_i.
Their disjunctive *combination is a new belief function $b_1 \cap b_2$ with b.p.a.*

$$m_{b_1 \cap b_2}(A) = \sum_{B \cap C = A} m_{b_1}(B) m_{b_2}(C).$$

Their conjunctive *combination is instead the b.f. $b_1 \cup b_2$ with b.p.a.*

$$m_{b_1 \cup b_2}(A) = \sum_{B \cup C = A} m_{b_1}(B) m_{b_2}(C).$$

Now, it is not difficult to prove that:

Proposition 3 *If b_1, b_2 are consistent then $b_1 \cup b_2$ is also consistent. On the other hand, if b_1, b_2 are consistent and their cores $\mathcal{C}_{b_1}$, $\mathcal{C}_{b_2}$ have non-empty intersection, then both $b_1 \oplus b_2$ and $b_1 \cap b_2$ are consistent with core $\mathcal{C}_{b_1 \cap b_2} = \mathcal{C}_{b_1} \cap \mathcal{C}_{b_2}$. Finally, if $\mathcal{C}_{b_1} \cap \mathcal{C}_{b_2} = \emptyset$ then $b_1 \oplus b_2$, $b_1 \cap b_2$ are* not *consistent.*

In other words, consistency is preserved by the conjunctive rule, the price to pay being increasing uncertainty as new evidence is combined, since the core of the belief state tends to Θ (complete ignorance). On the other side, both Dempster's rule and disjunctive combination preserve consistency only when the collection of focal elements of b_1 *and* b_2 is *already* consistent (i.e. any intersection $A \cap B$ of a f.e. A of b_1 and a f.e. B of b_2 is non-empty). As long as the new evidence is consistent with the existing one uncertainty is reduced. The price to pay is the loss of consistency in most cases.
The use of a consistent transformation in a reasoning process (1) would then guarantee consistency, while allowing the degree of uncertainty affecting our knowledge of the problem to decrease with time.

2.6 Making a belief function consistent

Consistent transformations can be built by solving a minimization problem of the form

$$cs[b] = \arg\min_{cs \in \mathcal{CS}} dist(b, cs) \tag{2}$$

where *dist* is some distance measure between belief functions, and $\mathcal{CS}$ denotes again the collection of all consistent b.f.s.
We call (2) the *consistent approximation problem*.
Plugging in different distance functions in (2) we get different consistent transformations.

In this paper we study what happens when using classical L_p norms in the approximation problem. As possibility measures are inherently related to the L_∞ norm (see above) cs.b.f.s live in a space linked to such a norm (Section 3). This leads to suppose that L_p-based approximations may indeed generate meaningful consistent transformations.

3 The simplicial complex of consistent belief functions

To solve the consistent approximation problem (2) we need to understand the structure of the space in which consistent belief functions live. We can then move forward and find the projection of b onto this space by minimizing the chosen distance.

3.1 The consistent complex

A belief function is determined by its $N-2$, $N = 2^{|\Theta|}$ belief values $\{b(A)\ \emptyset \subsetneq A \subsetneq \Theta\}$ (since $b(\emptyset) = 0$, $b(\Theta) = 1$ for all b.f.s). It can then be thought of as a vector of $\mathbb{R}^{N-2}$. The collection $\mathcal{B}$ of points of $\mathbb{R}^{N-2}$ which are b.f.s is a "simplex" (in rough words a higher-dimensional triangle), which we call *belief space*. $\mathcal{B}$ is the convex closure[1]

$$\mathcal{B} = Cl(b_A, \emptyset \subsetneq A \subseteq \Theta)$$

of the ("categorical") belief functions b_A assigning all the mass to a single event A: $m_b(A) = 1$, $m_b(B) = 0$ $\forall B \neq A$. In the belief space the vector $b \in \mathcal{B}$ which represents a belief function is the convex combination

$$b = \sum_{\emptyset \subsetneq A \subseteq \Theta} m_b(A) b_A \tag{3}$$

of the vectors b_A representing all the categorical belief functions.

[1]Here Cl denotes the convex closure operator: $Cl(b_1, ..., b_k) = \{b \in \mathcal{B} : b = \alpha_1 b_1 + \cdots + \alpha_k b_k, \sum_i \alpha_i = 1,\ \alpha_i \geq 0\ \forall i\}$.

The geometry of consistent belief functions can be described as a structure collection of simplices or *simplicial complex* [7]. More precisely, $\mathcal{CS}$ is the union

$$\mathcal{CS} = \bigcup_{x \in \Theta} Cl(b_A, A \ni x)$$

of the maximal simplices $Cl(b_A, A \ni x)$ formed by all the b.f.s with core containing a given element x of Θ.

3.2 Example: the binary case

As an example let us consider a frame of discernment formed by just two elements, $\Theta_2 = \{x, y\}$. In this very simple case each belief function $b : 2^{\Theta_2} \to [0, 1]$ is completely determined by its belief values $b(x)$, $b(y)$ as $b(\Theta) = 1, b(\emptyset) = 0\ \forall b \in \mathcal{B}$.
We can then represent each b.f. b as the vector

$$[b(x) = m_b(x), b(y) = m_b(y)]'$$

of $\mathbb{R}^{N-2} = \mathbb{R}^2$ (since $N = 2^2 = 4$). Since

$$m_b(x) \geq 0,\ m_b(y) \geq 0,\ m_b(x) + m_b(y) \leq 1$$

the set $\mathcal{B}_2$ of all the possible belief functions on Θ_2 is the triangle of Figure 1, whose vertices are the points $b_\Theta = [0, 0]'$, $b_x = [1, 0]'$, $b_y = [0, 1]'$ which correspond respectively to the vacuous belief function b_Θ ($m_{b_\Theta}(\Theta) = 1$), the Bayesian b.f. b_x with $m_{b_x}(x) = 1$, and the Bayesian b.f. b_y with $m_{b_y}(y) = 1$. The region $\mathcal{P}_2$ of all the Bayesian b.f.s on Θ_2 is the segment $Cl(b_x, b_y)$. In the binary case consistent belief functions can have as list of focal elements either $\{\{x\}, \Theta_2\}$

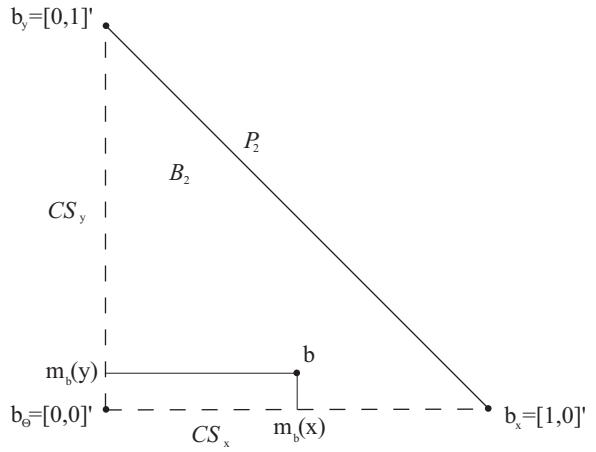


Figure 1: The belief space $\mathcal{B}$ for a binary frame is a triangle of $\mathbb{R}^2$ whose vertices are the categorical b.f.s focused on $\{x\}, \{y\}$ and Θ. The probability region is the segment $Cl(b_x, b_y)$, while all consistent b.f.s live in the union of the two segments $\mathcal{CS}_x = Cl(b_\Theta, b_x)$ and $\mathcal{CS}_y = Cl(b_\Theta, b_y)$.

or $\{\{y\}, \Theta_2\}$. Therefore the space of cs.b.f.s $\mathcal{CS}_2$ is the union of two one-dimensional simplices (line segments):

$$\mathcal{CS}_2 = \mathcal{CS}_x \cup \mathcal{CS}_y = Cl(b_\Theta, b_x) \cup Cl(b_\Theta, b_y).$$

4 The L_p consistent approximation problem

4.1 Using norms of the L_p family

The geometry of the binary case hints to a strict relation between consistent belief functions and L_p norms. As the plausibility of all the elements of their core is

$$pl_b(x) = \sum_{A \supseteq \{x\}} m_b(A) = 1 \ \ \forall x \in \mathcal{C}_b,$$

the region of consistent b.f.s

$$\mathcal{CS} = \Big\{b : \max_{x \in \Theta} pl_b(x) = 1\Big\} = \Big\{b : \|\bar{pl}_b\|_{L_\infty} = 1\Big\}$$

is the set of b.f.s for which the L_∞ norm of the plausibility distribution is equal to 1. This reinforces the observation that cs.b.f.s correspond to possibility distributions (Section 2), which are in turn inherently related to L_∞.
It makes then sense to conjecture that the consistent transformation we obtain by picking as distance function in the approximation problem (2) one of the classical L_p norms

$$\begin{array}{lcl} \|b - b'\|_{L_1} & = & \sum_{A \subseteq \Theta} |b(A) - b'(A)|, \\ \|b - b'\|_{L_2} & = & \sqrt{\sum_{A \subseteq \Theta} (b(A) - b'(A))^2}, \\ \|b - b'\|_{L_\infty} & = & \max_{A \subseteq \Theta} \{|b(A) - b'(A)|\} \end{array}$$

will be meaningful.
When looking for a probabilistic approximation $p[b] = \arg\min_{p \in \mathcal{P}} dist(b, p)$ the use of L_p norms leads indeed to quite interesting results. The L_2 approximation produces the so-called "orthogonal projection" of b onto $\mathcal{P}$ [6], while, at least in the binary case, the set of L_1/L_∞ probabilistic approximations of b coincide with the set of probabilities dominating b:

$$\mathcal{P}[b] \doteq \{p \in \mathcal{P} : p(A) \geq b(A) \ \forall A \subseteq \Theta\}.$$

4.2 Approximation on a complex

As the consistent complex $\mathcal{CS}$ is a *collection* of linear spaces (better, simplices which generate a linear space) solving the problem (2) involves finding a number of partial solutions

$$cs^x_{L_p}[b] = \arg \min_{cs \in \mathcal{CS}_x} \|b - cs\|_{L_p} \qquad (4)$$

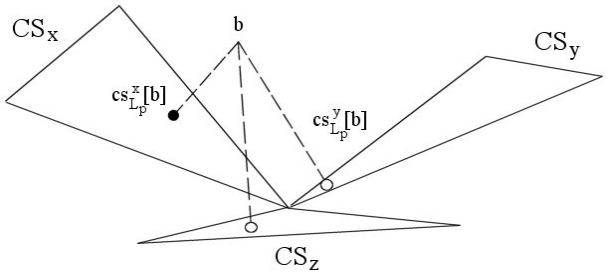


Figure 2: To minimize the distance of a point from a simplicial complex, we need to find all partial solutions (4) for all maximal simplices in the complex (empty circles), and later compare these partial solutions to select the global optimum (black circle).

(see Figure 2). Then, the distance of b from all such partial solutions has to be assessed in order to select a global optimal approximation.
In the rest of the paper we will apply this scheme to both the approximation problems associated with L_1 and L_2, respectively.

5 Approximation in the binary case

To get some insight on how to proceed in the general case, we will first consider the case study of a binary frame (Figure 3), and discuss how to approximate a belief function $b \in \mathcal{B}_2$ with a Bayesian or a consistent b.f. using an L_p norm. We will denote by

$$p_{L_p}[b] \doteq \arg \min_{p \in \mathcal{P}} \|b - p\|_{L_p}$$

the probability which minimizes the L_p distance from b. Analogously, we will use the notation

$$cs_{L_p}[b] \doteq \arg \min_{cs \in \mathcal{CS}} \|b - cs\|_{L_p}$$

for L_p consistent approximations.
In the Bayesian case we get

$$p_{L_2}[b] = \Big[m_b(x) + \frac{m_b(\Theta)}{2}, m_b(y) + \frac{m_b(\Theta)}{2}\Big]';$$

this probability is called *orthogonal projection* $\pi[b]$ of b onto $\mathcal{P}$ [6], and coincides with the pignistic function $BetP[b]$ [22, 4] in the binary case.
The L_1 solution $p_{L_1}[b]$, instead, is the whole set of probabilities "dominating" b [15], i.e.,

$$p_{L_1}[b] = \mathcal{P}[b] \doteq \big\{p \in \mathcal{P} : p(A) \geq b(A) \ \forall A \subseteq \Theta\big\}. \quad (5)$$

Figure 3 illustrates the geometry of all L_p Bayesian and consistent approximations of a belief function b in the binary frame. We can notice that:

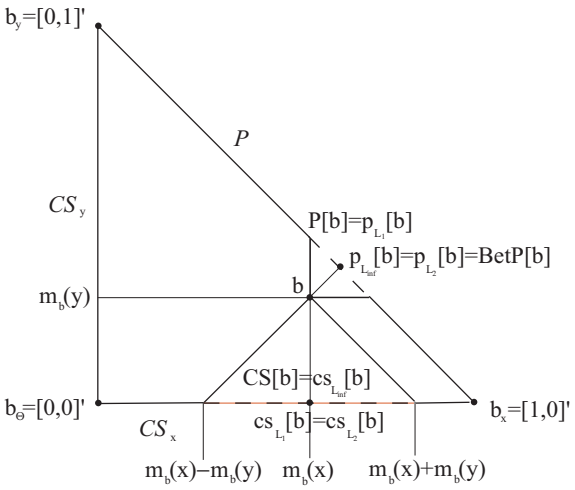


Figure 3: The dual behavior of Bayesian $p_{L_i}[b]$ and consistent $c_{L_i}[b]$ approximations of a b.f. b associated with the norms L_1, L_2, L_∞ is shown in the binary case.

1. the solution of the L_∞ approximation problem determines an entire set $\mathcal{CS}[b]$ of consistent b.f.s;
2. on the other hand, L_1/L_2 approximations on the same component $\mathcal{CS}_x$ of $\mathcal{CS}$ are point-wise and coincide;
3. the corresponding consistent transformation $cs^x_{L_2}[b]$ maps the original belief function b to a new b.f. with a focal element $A \cup \{x\}$ whenever A is a f.e. of b. The resulting b.p.a. is

$$\begin{aligned} m_{cs^x_{L_2}[b]}(x) &= \sum_{A: A\cup\{x\}=\{x\}} m_b(A) = m_b(x), \\ m_{cs^x_{L_2}[b]}(\Theta) &= \sum_{A: A\cup\{x\}=\Theta} m_b(A) \\ &= m_b(y) + m_b(\Theta). \end{aligned}$$

4. finally, the *global* L_1/L_2 consistent transformations also coincide, as they belong to the same component of the consistent complex ($\mathcal{CS}_x$ in the figure).

These facts (except the last point, which turns out to be an artifact of binary frames) are valid in the general case. Here we are going to focus on L_1/L_2 approximations.

6 Consistent L_1/L_2 approximations

6.1 Reducing the approximation problem to a linear system

In the case of an arbitrary frame a cs.b.f. $cs \in \mathcal{CS}_x$ is a solution of the L_2 *partial* approximation problem if $b - cs$ is orthogonal to all the generators $b_B - b_\Theta$ of the simplex $\mathcal{CS}_x = Cl(b_B, B \supseteq \{x\})$:

$$\langle b - cs, b_B - b_\Theta \rangle = \langle b - cs, b_B \rangle = 0 \quad \forall B \supseteq \{x\}$$

(as $b_\Theta = \mathbf{0}$ is the origin of $\mathbb{R}^{N-2}$, see binary example). We denote by $\alpha(A) \doteq m_{cs}(A)$ the b.p.a. of cs so that we can write each consistent belief function whose core contains $\{x\}$ as

$$cs = \sum_{A \supseteq \{x\}} \alpha(A) b_A$$

(by Equation (3)). After introducing the notation

$$\beta(A) \doteq m_b(A) - \alpha(A)$$

we can write $b - cs = \sum_{A \subsetneq \Theta} \beta(A) b_A$ and the orthogonality condition reads as

$$\Big\langle \sum_{A \subsetneq \Theta} \beta(A) b_A, b_B \Big\rangle = 0 \quad \forall B \supseteq \{x\}$$

i.e. (still for $\forall B \supseteq \{x\}$),

$$\Big\{ \sum_{A \supseteq \{x\}} \beta(A) \langle b_A, b_B \rangle + \sum_{A \not\supseteq \{x\}} m_b(A) \langle b_A, b_B \rangle = 0. \tag{6}$$

The L_1 minimization problem reads instead as

$$\begin{aligned} &\arg\min_{\vec{\alpha}} \Big\{ \sum_{A \supseteq \{x\}} \Big| \sum_{B \subseteq A} m_b(B) - \sum_{B \subseteq A, B \supseteq \{x\}} \alpha(B) \Big| \Big\} = \\ &\arg\min_{\vec{\beta}} \Big\{ \sum_{A \supseteq \{x\}} \Big| \sum_{B \subseteq A, B \supseteq \{x\}} \beta(B) + \sum_{B \subseteq A, B \not\supseteq \{x\}} m_b(B) \Big| \Big\} \end{aligned}$$

which is clearly solved by setting all addenda to zero, obtaining the linear system:

$$\Big\{ \sum_{B \subseteq A, B \supseteq \{x\}} \beta(B) + \sum_{B \subseteq A, B \not\supseteq \{x\}} m_b(B) = 0 \; \forall A \supseteq \{x\}. \tag{7}$$

6.2 Linear transformation

We are going to show here that the two minimization problems associated with the linear systems (6) and (7) coincide. The solution is indeed conserved due to the fact that the second linear system is obtained from the first one through a linear transformation.

Lemma 1 $\sum_{B \supseteq A} \langle b_B, b_C \rangle (-1)^{|B \setminus A|} = 1$ *if* $C \subseteq A$, *0 otherwise.*

Corollary 1 *The linear system (6) can be reduced to the system (7) through a linear transformation of rows:*

$$row_A \mapsto \sum_{B \supseteq A} row_B (-1)^{|B \setminus A|}. \tag{8}$$

Proof. If we apply the linear transformation (8) to the system (6) we get

$$\sum_{B\supseteq A}\Big[\sum_{C\supseteq\{x\}}\beta(C)\langle b_B,b_C\rangle+\sum_{C\not\supset\{x\}}m_b(C)\langle b_B,b_C\rangle\Big]\cdot$$
$$\cdot(-1)^{|B\setminus A|}=\sum_{C\supseteq\{x\}}\beta(C)\sum_{B\supseteq A}\langle b_B,b_C\rangle(-1)^{|B\setminus A|}+$$
$$+\sum_{C\not\supset\{x\}}m_b(C)\sum_{B\supseteq A}\langle b_B,b_C\rangle(-1)^{|B\setminus A|}\quad\forall A\supseteq\{x\}.$$

Therefore by Lemma 1 we get

$$\sum_{C\supseteq\{x\},C\subseteq A}\beta_C+\sum_{C\not\supset\{x\},C\subseteq A}m_b(C)=0\qquad\forall A\supseteq\{x\}$$

i.e. the system of equations (7). □

6.3 Form of the solution

To obtain both the L_2 and the L_1 consistent approximations of b it then suffices to solve the system (7) associated with the L_1 norm.

Theorem 1 *The unique solution of the linear system (7) is given by*

$$\beta(A)=-m_b(A\setminus\{x\}).$$

Proof. We can prove it by substitution. System (7) becomes

$$-\sum_{B\subseteq A,B\supseteq\{x\}}m_b(B\setminus\{x\})+\sum_{B\subseteq A,B\not\supset\{x\}}m_b(B)=$$
$$=-\sum_{C\subseteq A\setminus\{x\}}m_b(C)+\sum_{B\subseteq A,B\not\supset\{x\}}m_b(B)=$$
$$=-\sum_{C\subseteq A\setminus\{x\}}m_b(C)+\sum_{C\subseteq A\setminus\{x\}}m_b(C)=0.\qquad\square$$

Therefore, according to what discussed in Section 4, the partial L_1/L_2 consistent approximations of b on the maximal component $\mathcal{CS}_x$ of the consistent complex have b.p.a.

$$\begin{aligned}m_{cs^x_{L_1}}(A)&=m_{cs^x_{L_2}}(A)=\alpha(A)=m_b(A)-\beta(A)\\&=m_b(A)+m_b(A\setminus\{x\})\end{aligned}$$

for all events A such that $\{x\}\subseteq A\subsetneq\Theta$.
The value of $\alpha(\Theta)$ can be obtained by normalization:

$$\begin{aligned}\alpha(\Theta)&=1-\sum_{\{x\}\subseteq A\subsetneq\Theta}\alpha(A)\\&=1-\sum_{\{x\}\subseteq A\subsetneq\Theta}m_b(A)+m_b(A\setminus\{x\})\\&=1-\sum_{\{x\}\subseteq A\subsetneq\Theta}m_b(A)-\sum_{\{x\}\subseteq A\subsetneq\Theta}m_b(A\setminus\{x\})\\&=1-\sum_{A\neq\Theta,\{x\}^c}m_b(A)=m_b(\{x\}^c)+m_b(\Theta)\end{aligned}$$

as $B\not\supset\{x\}$ iff $B=A\setminus\{x\}$ for $A=B\cup\{x\}$.

Corollary 2 *The partial L_1 and L_2 consistent approximations of a belief function b with b.p.a. m_b onto the component $\mathcal{CS}_x$ of the consistent complex coincide. They have b.p.a.*

$$m_{cs^x_{L_1}}(A)=m_{cs^x_{L_2}}(A)=m_b(A)+m_b(A\setminus\{x\})$$

$\forall x\in\Theta$, and for all A s.t. $\{x\}\subseteq A\subseteq\Theta$.

6.4 Partial solutions as focused consistent transformations

The basic probability assignment of the L_1/L_2 consistent approximations of b has an elegant expression. It also has a straightforward interpretation: to get a consistent b.f. focused on a singleton x, the mass contribution of all the events B such that $B\cup\{x\}=A$ coincide is assigned indeed to A. But there are just two such events: A itself, and $A\setminus\{x\}$.
As an example, the partial consistent approximation of a belief function on a frame $\Theta=\{x,y,z,w\}$ with core $\{x\}$ is illustrated in Figure 4. The b.f. with focal

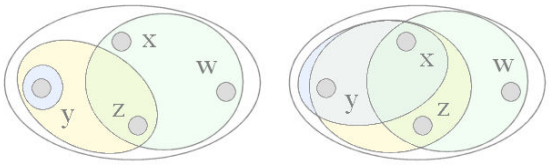


Figure 4: A belief function (left) and its L_1/L_2 consistent approximation with core $\{x\}$ (right).

elements $\{y\}$, $\{y,z\}$, and $\{x,z,w\}$ is transformed by the map

$$\begin{array}{lcl}\{y\}&\mapsto&\{x\}\cup\{y\}=\{x,y\},\\\{y,z\}&\mapsto&\{x\}\cup\{y,z\}=\{x,y,z\},\\\{x,z,w\}&\mapsto&\{x\}\cup\{x,z,w\}=\{x,z,w\}\end{array}$$

into the consistent b.f. with focal elements $\{x,y\}$, $\{x,y,z\}$, and $\{x,z,w\}$ and the same b.p.a.

Partial solutions to the L_1/L_2 consistent approximation problem turn out to be related to classical *inner consonant approximations* of a belief function b, i.e. the set of consonant b.f.s such that $c(A)\geq b(A)$ $\forall A\subseteq\Theta$ (or equivalently $pl_c(A)\leq pl_b(A)$ $\forall A$).
Dubois and Prade [10] proved indeed that such an approximation exists iff b is consistent. However, when b is *not* consistent a "focused consistent transformation" can be applied to get a new belief function b' such that

$$m'(A\cup x_i)=m(A)\qquad\forall A\subseteq\Theta$$

and x_i is the element of Θ with highest plausibility.
Theorem 1 and Corollary 2 state that the L_1/L_2 consistent approximation onto each component $\mathcal{CS}_x$ of $\mathcal{CS}$ generates the consistent transformation focused on x.

6.5 Global optimal solution for L_1

To find the *global* consistent approximation of b we need to work out which of the partial approximations $cs^x_{L_{1/2}}[b]$ has minimal distance from b. To do so we need to find

$$\arg\min_x \|b - cs^x_{L_{1/2}}[b]\|.$$

The L_1 distance of b from $\mathcal{CS}_x$ can be computed as

$$\begin{aligned}
&\|b - cs^x_{L_1}[b]\|_{L_1} = \sum_{A\subseteq\Theta} |b(A) - cs^x_{L_1}[b](A)| \\
&= \sum_{A\not\supseteq\{x\}} |b(A) - 0| + \sum_{A\supseteq\{x\}} \Big|b(A) - \sum_{B\subseteq A, B\supseteq\{x\}} \alpha(B)\Big| \\
&= \sum_{A\not\supseteq\{x\}} b(A) + \sum_{A\supseteq\{x\}} \Big| \sum_{B\subseteq A} m_b(B) + \\
&- \sum_{B\subseteq A, B\supseteq\{x\}} \big(m_b(B) + m_b(B\setminus\{x\})\big)\Big| \\
&= \sum_{A\not\supseteq\{x\}} b(A) + \sum_{A\supseteq\{x\}} \Big| \sum_{B\subseteq A, B\not\supseteq\{x\}} m_b(B) + \\
&- \sum_{B\subseteq A, B\supseteq\{x\}} m_b(B\setminus\{x\})\Big| = \sum_{A\not\supseteq\{x\}} b(A) + \\
&+ \sum_{A\supseteq\{x\}} \Big| \sum_{C\subseteq A\setminus\{x\}} m_b(C) - \sum_{C\subseteq A\setminus\{x\}} m_b(C)\Big| \\
&= \sum_{A\not\supseteq\{x\}} b(A) = \sum_{A\subseteq\{x\}^c} b(A).
\end{aligned} \tag{9}$$

Immediately,

Theorem 2 *The global optimal L_1 consistent approximation of any belief function b is given by*

$$cs_{L_1}[b] \doteq \arg\min_{cs\in\mathcal{CS}} \|b - cs^x_{L_1}[b]\| = cs^{\hat{x}}_{L_1}[b]$$

i.e. the partial approximation associated with the element $\hat{x}$ which minimizes (9):

$$\hat{x} = \arg\min_x \Big\{ \sum_{A\subseteq\{x\}^c} b(A), x\in\Theta \Big\}.$$

6.6 A counterexample

In the binary case (Figure 3) the condition of Theorem 2 reduces to

$$\begin{aligned}
\hat{x} &= \arg\min_x \sum_{A\subseteq\{x\}^c} b(A) = \arg\min_x m_b(\{x\}^c) \\
&= \arg\max_x pl_b(x)
\end{aligned}$$

and the global approximation falls on the component of the consistent complex associated with the element of *maximal plausibility*.

Unfortunately, this is not generally the case for arbitrary frames of discernment Θ. Let us see this in a simple counterexample. Let us first write

$$\begin{aligned}
&\sum_{A\subseteq\{x\}^c} b(A) = \sum_{A\subseteq\{x\}^c} \sum_{B\subseteq A} m_b(B) = \sum_{B\subseteq\{x\}^c} m_b(B)\cdot \\
&\cdot|\{A\subseteq\{x\}^c : A\supseteq B\}| = \sum_{B\subseteq\{x\}^c} m_b(B)\cdot 2^{|\{x\}^c|-|B|}.
\end{aligned} \tag{10}$$

Now, consider a belief function on a frame $\Theta = \{x_1, ..., x_n\}$ of cardinality n, with just two focal elements:

$$\begin{aligned}
m_b(x_1) &= m_x, \\
m_b(\{x_1\}^c) &= m_b(\{x_2, ..., x_n\}) = 1 - m_x.
\end{aligned}$$

If $m_x < 1/2$ all $y \neq x_1$ have maximal plausibility, as $pl_b(x_1) = 1 - b(\{x_1\}^c) = m_x$, while $pl_b(y) = 1 - m_x$ for all $y \neq x$. However, according to (10),

$$\begin{aligned}
\|b - cs^{x_1}_{L_1}[b]\|_{L_1} &= \sum_{A\subseteq\{x_1\}^c} b(A) \\
&= (1 - m_x)2^{n-1-(n-1)} = 1 - m_x,
\end{aligned}$$

where $n = |\Theta|$, while

$$\begin{aligned}
\|b - cs^{y}_{L_1}[b]\|_{L_1} &= \sum_{A\subseteq\{y\}^c} b(A) \\
&= m_x 2^{n-1-1} = m_x 2^{n-2}
\end{aligned}$$

$\forall y \neq x$. But when

$$m_x 2^{n-2} \geq 1 - m_x \equiv n \geq 2 + \log_2\Big(\frac{1-m_x}{m_x}\Big)$$

we have that

$$\|b - cs^{x_1}_{L_1}[b]\|_{L_1} \leq \|b - cs^{y}_{L_1}[b]\|_{L_1} \quad \forall y \neq x_1,$$

and therefore *the global L_1 consistent approximation can fall on a component not associated with the maximal plausibility element.*

6.7 Global optimal solution for L_2

In the L_2 case we get

$$\begin{aligned}
&\|b - cs^x_{L_2}[b]\|^2 = \sum_{A\subseteq\Theta} \Big(b(A) - cs^x_{L_2}[b](A)\Big)^2 = \\
&= \sum_{A\subseteq\Theta} \Big[\sum_{B\subseteq A} m_b(B) - \sum_{B\subseteq A, B\supseteq\{x\}} \alpha(B)\Big]^2 = \\
&= \sum_{A\subseteq\Theta} \Big[\sum_{B\subseteq A} m_b(B) - \sum_{B\subseteq A, B\supseteq\{x\}} m_b(B) + \\
&- \sum_{B\subseteq A, B\supseteq\{x\}} m_b(B\setminus\{x\})\Big]^2 = \\
&= \sum_{A\not\supseteq\{x\}} \big(b(A)\big)^2 + \sum_{A\supseteq\{x\}} \Big[\sum_{B\subseteq A, B\not\supseteq\{x\}} m_b(B) + \\
&- \sum_{B\subseteq A, B\supseteq\{x\}} m_b(B\setminus\{x\})\Big]^2 = \sum_{A\not\supseteq\{x\}} \big(b(A)\big)^2 + \\
&+ \sum_{A\supseteq\{x\}} \Big[\sum_{C\subseteq A\setminus\{x\}} m_b(C) - \sum_{C\subseteq A\setminus\{x\}} m_b(C)\Big]^2
\end{aligned}$$

so that, in analogy with the L_1 case,

$$\|b - cs^x_{L_2}[b]\|^2 = \sum_{A \subseteq \{x\}^c} (b(A))^2.$$

Theorem 3 *The global optimal L_2 consistent approximation of any belief function b is given by*

$$cs_{L_2}[b] \doteq \arg\min_{cs \in \mathcal{CS}} \|b - cs^x_{L_2}[b]\| = cs^{\hat{x}}_{L_2}[b]$$

i.e. the partial approximation associated with the element

$$\hat{x} = \arg\min_x \Big\{ \sum_{A \subseteq \{x\}^c} (b(A))^2, x \in \Theta \Big\}.$$

Other simple counterexamples show that the global L_2 consistent approximation can fall on a component not associated with the maximal plausibility element.

7 Comments and conclusions

Belief functions represent coherent knowledge bases in the theory of evidence. As consistency is not preserved by most operators used to update or elicit evidence, the use of a consistent transformation in conjunction with those combinations rules can be desirable. Consistent transformations are strictly related to the problem of approximating a generic belief function with a consistent one.

In this paper we solved the instance of the consistent approximation problem we obtain when measuring distances between uncertainty measures by means of the classical L_p norms. This makes sense as cs.b.f.s live in a simplicial complex defined in terms of the L_∞ norms, and correspond to possibility distributions. A partial approximation for each component of the complex has to be found. The conclusions of this study are the following:
1. partial L_1/L_2 approximations coincide on each component of the consistent complex;
2. such partial approximation turns out to be the consistent transformation focused on the given element of the frame;
3. the corresponding global solutions have not in general as core the maximal plausibility element, and may lie in general on different components of $\mathcal{CS}$.

The interpretation of the polytope of all L_∞ solutions is worth to be fully investigated in the near future, in the light of the intuition provided by the binary case. In particular its clear analogy with the polytope of consistent probabilities will be interesting matter to study. A natural continuation of this line of research is obviously the solution of the L_p approximation problem for consonant belief functions, as counterparts of possibility measures in the theory of evidence. That will complete our understanding of the relation between geometric norms and evidence consistency.

Proof of Lemma 1

We first note that, by definition of dogmatic belief function b_A (Section 3),

$$\langle b_B, b_C \rangle = \sum_{D \supseteq B,C; D \neq \Theta} 1 = \sum_{E \subsetneq (B \cup C)^c} 1 = 2^{|(B \cup C)^c|} - 1.$$

Hence $\sum_{B \subseteq A} \langle b_B, b_C \rangle (-1)^{|B \setminus A|} =$

$$\begin{aligned} &= \sum_{B \subseteq A} (2^{|(B \cup C)^c|} - 1)(-1)^{|B \setminus A|} \\ &= \sum_{B \subseteq A} 2^{|(B \cup C)^c|} (-1)^{|B \setminus A|} - \sum_{B \subseteq A} (-1)^{|B \setminus A|} \\ &= \sum_{B \subseteq A} 2^{|(B \cup C)^c|} (-1)^{|B \setminus A|}, \end{aligned}$$

as

$$\sum_{B \subseteq A} (-1)^{|B \setminus A|} = \sum_{k=0}^{|B \setminus A|} 1^{|A^c| - k} (-1)^k = 0$$

for Newton's binomial:

$$\sum_{k=0}^{n} p^k q^{n-k} = (p+q)^n. \tag{11}$$

Now, as both $B \supseteq A$ and $C \supseteq A$ the set B can be

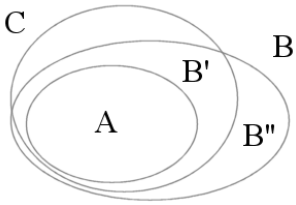


Figure 5: Decomposition of B into $A + B' + B''$ in the proof of Lemma 1.

decomposed into the disjoint sum

$$B = A + B' + B''$$

where

$$\emptyset \subseteq B' \subseteq C \setminus A, \quad \emptyset \subseteq B'' \subseteq (C \cup A)^c$$

(see Figure 5), so that the above quantity can be written as

$$\sum_{\emptyset \subseteq B' \subseteq C \setminus A} \sum_{\emptyset \subseteq B'' \subseteq (C \cup A)^c} 2^{|(A \cup C)|^c - |B''|} (-1)^{|B'| + |B''|} = \sum_{\emptyset \subseteq B' \subseteq C \setminus A} (-1)^{|B'|} \sum_{\emptyset \subseteq B'' \subseteq (C \cup A)^c} (-1)^{|B''|} 2^{|(A \cup C)|^c - |B''|}$$

where

$$\sum_{\emptyset \subseteq B'' \subseteq (C \cup A)^c} (-1)^{|B''|} 2^{|(A \cup C)^c - |B''|} = [2 + (-1)]^{|(A \cup C)^c}$$

$= 1^{|(A \cup C)^c} = 1$, again for Newton's binomial (11). The desired quantity becomes

$$\sum_{\emptyset \subseteq B' \subseteq C \setminus A} (-1)^{|B'|}$$

which is nil for $C \setminus A \neq \emptyset$, equal to 1 when $C \setminus A = \emptyset$, i.e. $C \subseteq A$.

References

[1] P. Baroni, *Extending consonant approximations to capacities*, IPMU, 2004, pp. 1127–1134.

[2] D. Batens, C. Mortensen, and G. Priest, *Frontiers of paraconsistent logic*, Studies in logic and computation (J.P. Van Bendegem, ed.), vol. 8, Research Studies Press, 2000.

[3] L. Caro and A. Babak Nadjar, *Generalization of the Dempster-Shafer theory: a fuzzy-valued measure*, IEEE Transactions on Fuzzy Systems **7** (1999), 255–270.

[4] B.R. Cobb and P.P. Shenoy, *A comparison of Bayesian and belief function reasoning*, Information Systems Frontiers **5** (2003), no. 4, 345–358.

[5] F. Cuzzolin, *An interpretation of consistent belief functions in terms of simplicial complexes*, submitted to Information Sciences (2007).

[6] F. Cuzzolin, *Two new Bayesian approximations of belief functions based on convex geometry*, IEEE Trans. on Systems, Man, and Cybernetics - Part B **37** (2007), no. 4, 993–1008.

[7] F. Cuzzolin, *An interpretation of consistent belief functions in terms of simplicial complexes*, Proc. of ISAIM'08, 2008.

[8] A.P. Dempster, *A generalization of Bayesian inference*, Journal of the Royal Stat. Soc., Series B **30** (1968), 205–247.

[9] D. Dubois and H. Prade, *Possibility theory*, Plenum Press, New York, 1988.

[10] D. Dubois and H. Prade, *Consonant approximations of belief functions*, International Journal of Approximate Reasoning **4** (1990), 419–449.

[11] B.A. Dubrovin, S.P. Novikov, and A.T. Fomenko, *Sovremennaja geometrija. metody i prilozenija*, Nauka, Moscow, 1986.

[12] S. Heilpern, *Representation and application of fuzzy numbers*, Fuzzy Sets and Systems **91** (1997), 259–268.

[13] C. Joslyn, *Towards an empirical semantics of possibility through maximum uncertainty*, Proc. IFSA 1991 (R. Lowen and M. Roubens, eds.), vol. A, 1991, pp. 86–89.

[14] G. J. Klir, W. Zhenyuan, and D. Harmanec, *Constructing fuzzy measures in expert systems*, Fuzzy Sets and Systems **92** (1997), 251–264.

[15] H. Kyburg, *Bayesian and non-Bayesian evidential updating*, Artificial Intelligence **31** (1987), no. 3, 271–294.

[16] J.B. Paris, D. Picado-Muino, and M. Rosefield, *Information from inconsistent knowledge: A probability logic approach*, Interval / Probabilistic Uncertainty and Non-classical Logics, Advances in Soft Computing, vol. 46, Springer-Verlag, Berlin - Heidelberg, 2008.

[17] G. Priest, R. Routley, and J. Norman, *Paraconsistent logic: Essays on the inconsistent*, Philosophia Verlag, 1989.

[18] C. Roemer and A. Kandel, *Applicability analysis of fuzzy inference by means of generalized Dempster-Shafer theory*, IEEE Transactions on Fuzzy Systems **3** (1995), no. 4, 448–453.

[19] G. Shafer, *A mathematical theory of evidence*, Princeton University Press, 1976.

[20] Ph. Smets, *The transferable belief model and possibility theory*, NAFIPS-90 (Kodratoff Y., ed.), 1990, pp. 215–218.

[21] Ph. Smets, *Belief functions : the disjunctive rule of combination and the generalized Bayesian theorem*, International Journal of Approximate Reasoning **9** (1993), 1–35.

[22] Ph. Smets and R. Kennes, *The transferable belief model*, Artificial Intelligence **66** (1994), 191–234.

[23] R. R. Yager, *Class of fuzzy measures generated from a Dempster-Shafer belief structure*, International Journal of Intelligent Systems **14** (1999), 1239–1247.

Epistemic irrelevance in credal networks: the case of imprecise Markov trees

Gert de Cooman and Filip Hermans
SYSTeMS, Ghent University, Belgium
{gert.decooman,filip.hermans}@ugent.be

Alessandro Antonucci and Marco Zaffalon
IDSIA, Switzerland
{alessandro,zaffalon}@idsia.ch

Abstract

We replace strong independence in credal networks with the weaker notion of epistemic irrelevance. Focusing on directed trees, we show how to combine local credal sets into a global model, and we use this to construct and justify an exact message-passing algorithm that computes updated beliefs for a variable in the tree. The algorithm, which is essentially linear in the number of nodes, is formulated entirely in terms of coherent lower previsions. We supply examples of the algorithm's operation, and report an application to on-line character recognition that illustrates the advantages of our model for prediction.

Keywords. Coherence, credal network, epistemic irrelevance, epistemic independence, strong independence, imprecise Markov tree, separation, hidden Markov chain.

1 Introduction

The last twenty years have witnessed a rapid growth of *graphical models* in the fields of artificial intelligence and statistics. These models combine graphs and probability to address complex multivariate problems in a variety of domains, such as medicine, finance, risk analysis, defense, and environment, to name just a few.

Much has been done also on the front of imprecise probability. *Credal networks* [3] have been and still are the subject of intense research. A credal network creates a global model of a domain by combining local uncertainty models using some notion of independence, and then uses this to do inference. The local models represent uncertainty by closed convex sets of probabilities, also called *credal sets*.

The notion of independence used with credal nets in the vast majority of cases is that of *strong independence* (with some exceptions in [6]). Loosely speaking, two variables X,Y are strongly independent if the credal set for (X,Y) can be regarded as originating from a number of precise models in each of which X and Y are stochastically independent. Strong independence is closely related with the *sensitivity analysis* interpretation of credal sets, which regards an imprecise model as arising out of partial ignorance of a precise one. This is a somewhat narrow view, and it does not apply in general.

An alternative and attractive way to express irrelevance that is not committed to the sensitivity analysis interpretation is offered by *epistemic irrelevance* [15]: we say that X is irrelevant to Y if observing X does not affect beliefs about Y. Epistemic irrelevance is defined directly in terms of a subject's beliefs and is therefore very well suited for a behavioural theory of imprecise probability. It is also weaker than strong independence, and it therefore does not lead to overconfident inferences when the sensitivity analysis interpretation is not justified.

At this point the question that we address in this paper should be clear: can we define credal nets based on epistemic irrelevance, and moreover create an exact algorithm to perform efficient inferences with them? We give a fully positive answer to this question in the special case that (i) the graph under consideration is a directed tree, and (ii) the related variables assume only finitely many values. The intuitions that showed us the way towards this result originated in previous work done by some of us on imprecise probability trees [7] and imprecise Markov chains [8].

How do we address this problem? After giving some preliminary notions and introducing the model in Sec. 2, we discuss in Sec. 3 how to combine marginal models into joint ones reflecting certain irrelevance assessments, in a way that is as conservative as possible. We comment on the graphical separation criteria induced by epistemic irrelevance in Sec. 5. We then go on to develop and justify an inference algorithm for treating the model as an expert system in Sec. 6. The algorithm is used to *update* the tree: it computes posterior beliefs about a *target* variable in the tree conditional on the observation of other variables, that are called *instantiated*, meaning that their value is determined. It is based on message passing, as are the traditional algorithms that have been developed for precise graphical models, and it has some remarkable properties: (i) it works in time essentially linear in the size of the tree; (ii) it natively computes posterior lower and upper

previsions (or expectations) rather than probabilities; (iii) it is an algorithm for credal nets developed for the first time exclusively using the formalism of *coherent lower previsions* [15]; and (iv) it is shown to lead to coherent inferences under mild conditions. We give a step-by-step example of the way inferences can be done in our framework in Sec. 7, where we also comment on the intriguing relationship between the failure of certain classical separation properties in our framework, and dilation [10, 14]. The last part of the paper focuses on numerical simulations. In Sec. 8 we empirically measure the amount of imprecision introduced by using epistemic irrelevance rather than strong independence in a credal tree, when propagating inferences backwards (towards the root) from instantiated nodes to the target node; indeed, it can be shown [7] that there is no difference between inferences that go forward from instantiated nodes to target under strong independence and epistemic irrelevance. In Sec. 9 we present an application of our algorithm to on-line character recognition. We learn the probabilities from data and compare the predictions of the our approach with those of its precise probability counterpart. The results are encouraging: they show that the tree can be used for real applications, and that the imprecision it originates is justified.

Due to lack of space, we must assume the reader has a working knowledge of the basics of Walley's [15] theory of coherent lower previsions. We also refrain from giving proofs of technical results for the same reason, and rather stress motivation, simple justifications and examples.

2 Credal trees under epistemic irrelevance

Basic notions and notation. We consider a rooted and directed discrete tree with finite width and depth. We call T the set of its nodes s, and we denote the *root*, or initial, node by $\square$. Consider any node s, then we denote the set of its parents by $P(s)$. Of course, $P(\square)=\emptyset$, and for $s\neq\square$ we have that $P(s)=\{m(s)\}$ where $m(s)$ is the *mother node* of s. Also, for each node s, we denote the set of its *children* by $C(s)$, and the set of its *siblings* by $S(s)$. Clearly, $S(\square)=\emptyset$, and if $s\neq\square$ then $S(s)=C(m(s))\setminus\{s\}$. If $C(s)=\emptyset$, then we call s a *leaf*, or *terminal node*.

For nodes s and t, we write $s\sqsubseteq t$ if s *precedes* t, i.e., if there is a directed segment in the tree from s to t. The relation $\sqsubseteq$ is a special partial order on the set T. $A(s):=\{t\in T: t\sqsubset s\}$ denotes the set of *ancestors* of s, and $D(s):=\{t\in T: s\sqsubset t\}$ its set of *descendants*. Here $s\sqsubset t$ means that $s\sqsubseteq t$ and $s\neq t$. We also use $\uparrow s:=A(s)\cup\{s\}$, $\downarrow s:=D(s)\cup\{s\}$, $\uparrow S:=\bigcup\{\uparrow s: s\in S\}$ and $\downarrow S:=\bigcup\{\downarrow s: s\in S\}$ for any subset $S\subseteq T$.

With each node s of the tree, there is associated a variable X_s assuming values in a finite non-empty set $\mathscr{X}_s$. We denote the set of all real-valued maps (*gambles*) on $\mathscr{X}_s$ by $\mathscr{L}(\mathscr{X}_s)$. We extend this notation to more complicated situations as follows. If S is any subset of T, then we denote by X_S the tuple of variables whose components are the X_s for all $s\in S$. This new joint variable assumes values in the finite set $\mathscr{X}_S:=\times_{s\in S}\mathscr{X}_s$, and the corresponding set of gambles is denoted by $\mathscr{L}(\mathscr{X}_S)$. Generic elements of $\mathscr{X}_s$ are denoted by x_s or z_s. Similarly for x_S and z_S in $\mathscr{X}_S$. Also, if we mention a tuple z_S, then for any $t\in S$, the corresponding element in the tuple will be denoted by z_t. We assume all variables in the tree to be logically independent.

Local uncertainty models. We now add a *local uncertainty model* to each of the nodes s. If s is not the root node, i.e., has a mother $m(s)$, then this local model is a (separately coherent) conditional lower prevision $\underline{Q}_s(\cdot|X_{m(s)})$ on $\mathscr{L}(\mathscr{X}_s)$: for each possible value $z_{m(s)}$ of the variable $X_{m(s)}$ associated with its mother $m(s)$, we have a coherent lower prevision $\underline{Q}_s(\cdot|z_{m(s)})$ for the value of X_s, conditional on $X_{m(s)}=z_{m(s)}$. In the root, we have an unconditional local uncertainty model $\underline{Q}_\square$ for the value of $X_\square$; $\underline{Q}_\square$ is a coherent lower prevision on $\mathscr{L}(\mathscr{X}_\square)$. We use the common generic notation $\underline{Q}_s(\cdot|X_{P(s)})$ for all these local models.

Global uncertainty models. In this and the following two sections, we show how all these local models $\underline{Q}_s(\cdot|X_{m(s)})$ can be combined into *global uncertainty models*. If we generically denote by the symbol $\underline{P}_s$ lower previsions on $\mathscr{L}(\mathscr{X}_{\downarrow s})$, representing information about $X_{\downarrow s}$, then this means we want to end up with an unconditional joint lower prevision $\underline{P}:=\underline{P}_\square$ on $\mathscr{L}(\mathscr{X}_T)$ for all variables in the tree, as well conditional lower previsions $\underline{P}_s(\cdot|X_{m(s)})$ on $\mathscr{L}(\mathscr{X}_{\downarrow s})$ for all non-initial nodes s. Ideally, we want these global (conditional) lower previsions to be coherent with one another, and to reflect the conditional irrelevancies (or Markov-type conditions) that we want the graphical structure of the tree to encode. In addition, we want them to be as conservative (small) as possible.

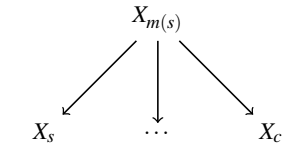
The interpretation of the graphical model. Consider any node s in the tree, and its parent set $P(s)$ [either empty or equal to $\{m(s)\}$]. We also consider the set $\overline{s}:=T\setminus[D(s)\cup P(s)]$ of its non-parent non-descendants. Then *conditional on the parent variables $X_{P(s)}$, the non-parent non-descendant variables $X_{\overline{s}}$ are assumed to be epistemically irrelevant to the variables $X_{\downarrow s}$ associated with s and its descendants.* This interpretation turns the tree into a *credal tree under epistemic irrelevance*, and we shall also use the term *imprecise Markov tree* (IMT) for it.

In terms of the global models, this means that for all $s\in T$, for all $S\subseteq\overline{s}$ and for all $z_{S\cup P(s)}\in\mathscr{X}_{S\cup P(s)}$:

$$\underline{P}_s(\cdot|z_{P(s)})=\underline{P}_s(\cdot|z_{S\cup P(s)}). \tag{1}$$

We discuss the separation properties that accompany this interpretation in some detail in Sec. 5. For now, we focus on one immediate consequence that will help

us go from local to global models in Sec. 4. Consider some non-initial node s. The interpretation of the graphical structure of the tree tells us that for each *sibling* $c \in S(s)$ of s, the variable X_c is epistemically irrelevant to the variable X_s, conditional on $X_{m(s)}$. It even tells us that for any non-empty set $S \subseteq S(s)$ of siblings of s, the variable X_S is epistemically irrelevant to X_s, conditional on $X_{m(s)}$. We conclude that all children of a node are not just epistemically irrelevant to each other: they are even epistemically independent [15, Chapter 9], in some very specific sense.



3 Net-independent natural extension

This leads us to the following small digression. We consider the following problem, the solution of which will help us in our discussion further on. Suppose we have a number of *marginal* lower previsions $\underline{P}_n$ representing beliefs about the values that each of a finite number of (logically independent) variables X_n assume in the respective finite sets $\mathscr{X}_n$, $n \in N$, where N is some finite set.

Net-independent products. We now want to construct a *joint* lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$, where $\mathscr{X}_N = \times_{n\in N}\mathscr{X}_n$, that coincides with the marginals $\underline{P}_n$ on their respective domains $\mathscr{L}(\mathscr{X}_n)$, and such that this joint reflects the following structural assessments: for each $o \in N$ and each non-empty $I \subseteq N \setminus \{o\}$, the variables X_I are epistemically irrelevant to the variable X_o. In other words, learning the value of any number of these variables does not affect beliefs about any single other variable amongst them. We then call the variables X_n, $n \in N$ *net-independent*.

Such irrelevance assessments are useful because they allow us to turn marginal into conditional lower previsions. Indeed, for each $o \in N$ and each $I \subseteq N \setminus \{o\}$ we can use the epistemic irrelevance of X_I to X_o to infer from the marginal lower prevision $\underline{P}_o$ a conditional lower prevision $\underline{P}_o(\cdot|X_I)$ on $\mathscr{L}(\mathscr{X}_o)$ given by:

$$\underline{P}_o(h|x_I) := \underline{P}_o(h) \text{ for all gambles } h \text{ on } \mathscr{X}_o.$$

So we can use the assessment of net-independence of the variables X_n, $n \in N$ to infer from the marginals a family of conditional lower previsions:

$$\mathscr{N}(\underline{P}_n, n \in N) := \{\underline{P}_o(\cdot|X_I)\colon o \in N \text{ and } I \subseteq N \setminus \{o\}\}.$$

Definition 1. *A coherent joint lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$ that coincides with the marginal lower previsions $\underline{P}_n$ on their domains $\mathscr{L}(\mathscr{X}_n)$, $n \in N$ and that is coherent with the family of conditional lower previsions $\mathscr{N}(\underline{P}_n, n \in N)$ is called a* net-independent product *of these marginals. If it exists, then the point-wise smallest such net-independent product is called the* net-independent natural extension *of these marginals, and denoted by $\otimes_{n\in N}\underline{P}_n$.*

Conditioning factorising lower previsions. The following notion of factorisation is intimately linked with that of a net-independent product. It will also play a crucial part in our development of an algorithm for treating an imprecise Markov tree as an expert system.

Definition 2. *We call a coherent lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$* factorising *if for all $o \in N$ and all non-empty $I \subseteq N \setminus \{o\}$, all $g \in \mathscr{L}(\mathscr{X}_o)$ and all non-negative $f_i \in \mathscr{L}(\mathscr{X}_i)$, $i \in I$, $\underline{P}_N(fg) = \underline{P}_N(f\underline{P}_N(g))$, where $f := \prod_{i\in I} f_i$.*

As an important example, the so-called *strong product* [3] $\times_{n\in N}\underline{P}_n$ of the marginal lower previsions $\underline{P}_n$ is factorising. But for any coherent factorising joint lower prevision $\underline{P}_N$, we see that for any non-empty subset I of N:

$$\underline{P}_N(\times_{i\in I}A_i) = \prod_{i\in I}\underline{P}_N(A_i) \text{ and } \overline{P}_N(\times_{i\in I}A_i) = \prod_{i\in I}\overline{P}_N(A_i), \tag{2}$$

where $A_i \subseteq \mathscr{X}_i$ for all $i \in I$. Let us call any real functional Φ on $\mathscr{L}(\mathscr{X})$ *strictly positive* if $\Phi(I_{\{x\}}) > 0$ for all $x \in \mathscr{X}$. Then the following result is immediate from Eq. (2).

Proposition 1. *A factorising coherent lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$ is strictly positive if and only if all its marginals are, and its conjugate upper prevision $\overline{P}_N$ is strictly positive if and only if all its marginals are.*

As a next step, suppose we want to condition a coherent and factorising joint $\underline{P}_N$ on an observation $X_I = x_I$, where I is some proper subset of N. To this end, we calculate the *regular extension* [15, Appendix J]: when $\overline{P}_N(I_{\{x_I\}}) > 0$,

$$\underline{R}(h|x_I) := \max\{\mu \in \mathbb{R}\colon \underline{P}_N(I_{\{x_I\}}[h-\mu]) \geq 0\},$$

where h is any gamble on $\mathscr{X}_O$ and O is any non-empty subset of $N \setminus I$. Otherwise $\underline{R}(\cdot|x_I)$ is vacuous. Then because $\underline{P}_N$ is factorising:

$$\begin{aligned}\underline{P}_N(I_{\{x_I\}}[h-\mu]) &= \underline{P}_N(I_{\{x_I\}}\underline{P}_N(h-\mu)) \\ &= \begin{cases}\underline{P}_N(\{x_I\})(\underline{P}_N(h)-\mu) & \text{if } \underline{P}_N(h) \geq \mu \\ \overline{P}_N(\{x_I\})(\underline{P}_N(h)-\mu) & \text{if } \underline{P}_N(h) \leq \mu,\end{cases}\end{aligned}$$

so we conclude that, quite interestingly,

$$\underline{R}(h|x_I) = \underline{P}_N(h) \text{ as soon as } \overline{P}_N(\{x_I\}) > 0. \tag{3}$$

Because we are working in a finitary context [$\mathscr{X}_N$ is a finite set], the regular extension $\underline{R}(\cdot|X_I)$ is guaranteed to be coherent with the joint lower prevision $\underline{P}_N$ [15, Sec. J3]. This, together with an interesting recent coherence result by Enrique Miranda [11, Theorem 5], leads us to the following conclusion.

Proposition 2. *Any coherent joint lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$ that is factorising and strictly positive,*[1] *is a net-independent product of its marginals.*

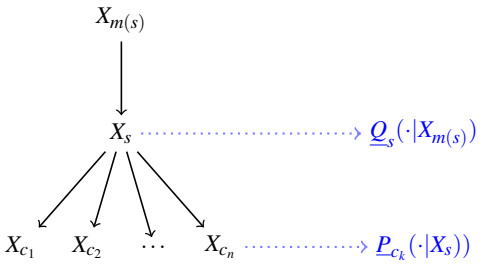
As an immediate consequence, the strong product $\times_{n\in N}\underline{P}_n$ of a collection of strictly positive marginals $\underline{P}_n$, $n \in N$, is also a net-independent product of these marginals, and is therefore coherent with the associated family of conditional lower previsions $\mathscr{N}(\underline{P}_n, n \in N)$. So this family is itself always guaranteed to be coherent, and because all the sets $\mathscr{X}_n$ are finite, we can invoke Walley's Finite Extension Theorem [15, Theorem 8.1.9] to conclude that there always is a point-wise smallest joint lower prevision that is coherent with the family $\mathscr{N}(\underline{P}_n, n \in N)$. This provides the most important step in the proof of the following result. Another crucial step is provided by the fact that, since the strong product is a net-independent product of the marginals $\underline{P}_n$, $n \in N$, it has to dominate the net-independent natural extension: $\times_{n\in N}\underline{P}_n \geq \otimes_{n\in N}\underline{P}_n$.

Proposition 3. *For any collection of strictly positive and coherent marginal lower previsions $\underline{P}_n$ on $\mathscr{L}(\mathscr{X}_n)$, $n \in N$, their net-independent natural extension $\otimes_{n\in N}\underline{P}_n$ exists, and it is a factorising and strictly positive coherent lower prevision on $\mathscr{L}(\mathscr{X}_N)$.*

4 Constructing the most conservative joint

We now show how to construct specific global models for the variables in the tree, and argue that these are the most conservative coherent models that extend the local models and express all conditional irrelevancies (1) encoded in the imprecise Markov tree. In the next section, we will use these global models to construct and justify an algorithm for treating the imprecise Markov tree as an expert system.

The crucial step lies in the recognition that any tree can be constructed recursively from the leaves up to the root, by using basic building blocks of the following type:



The global models are then also constructed in a recursive manner, following the same pattern. Consider a node s and suppose that, in each of its children $c \in C(s)$, we already have a global conditional lower prevision $\underline{P}_c(\cdot|X_s)$ on $\mathscr{L}(\mathscr{X}_{\downarrow c})$. We construct a global conditional lower prevision $\underline{P}_s(\cdot|X_{P(s)})$ on $\mathscr{L}(\mathscr{X}_{\downarrow s})$ by backwards recursion:

$$\underline{P}_s(\cdot|X_s) := \otimes_{c\in C(s)}\underline{P}_c(\cdot|X_s) \tag{4}$$

$$\begin{aligned}\underline{P}_s(\cdot|X_{P(s)}) &:= \underline{Q}_s(\underline{P}_s(\cdot|X_s)|X_{P(s)}) \\ &= \underline{Q}_s(\otimes_{c\in C(s)}\underline{P}_c(\cdot|X_s)|X_{P(s)}),\end{aligned} \tag{5}$$

the conditional lower prevision $\underline{P}_s(\cdot|X_s)$ on $\mathscr{L}(\mathscr{X}_{\downarrow C(s)})$ being the net-independent natural extension of the conditional lower previsions $\underline{P}_c(\cdot|X_s)$ on $\mathscr{L}(\mathscr{X}_{\downarrow c})$, $c \in C(s)$. If we start in leaves t with the 'boundary condition'

$$\underline{P}_t(\cdot|X_{P(t)}) := \underline{Q}_t(\cdot|X_{P(t)}) \text{ for all leaves } t, \tag{6}$$

then the recursion relations (4) and (5) eventually lead to a global model $\underline{P}_s(\cdot|X_{m(s)})$ in all nodes s of the tree, and in particular to a joint model $\underline{P} := \underline{P}_\square$ on $\mathscr{L}(\mathscr{X}_T)$. These are the global (conditional) lower previsions we have been looking for, as the following theorem tells us. Its proof proceeds in a recursive fashion, similar to the construction of the global models. It relies rather heavily on the fact that the net-independent natural extension is factorising, and on the coherence result by Miranda [11, Theorem 5], already mentioned before Proposition 2.

Theorem 4. *If all local models $\underline{Q}_s(\cdot|X_{P(s)})$ on $\mathscr{L}(\mathscr{X}_s)$, $s \in T$ are strictly positive, then the global models $\underline{P}_s(\cdot|X_{P(s)})$ on $\mathscr{L}(\mathscr{X}_{\downarrow s})$, $s \in T$ obtained through Eqs.* (4)–(6)*, constitute the point-wise smallest coherent family of (conditional) lower previsions that (i) extend the local models, and (ii) satisfy the epistemic irrelevance conditions* (1) *encoded in the graphical structure.*

5 Some separation properties

Without going into too much detail, we would like to point out one of the more striking differences between the separation properties in imprecise Markov trees under epistemic irrelevance, and the more usual ones for Bayesian nets [12] and credal nets under strong independence [3].

It is clear from the interpretation of the graphical model described in Sec. 2 that we have the following simple separation results:

where in both cases, X_{i_2} *separates* X_t *from* X_{i_1}: when the value of X_{i_2} is known, additional information about the value of X_{i_1} does not affect beliefs about the value of X_t. In this figure, between i_1 and i_2, and between i_2 and t, there may be other nodes, but the arrows along the path segment through these nodes should all point in the indicated directions. The underlying idea is that t is a (descendant of some) child c of i_2, and conditional on the mother i_2 of c, the non-parent non-descendant i_1 of c is epistemically irrelevant to c and all of its descendants.

[1] We strongly suspect that this proposition, and a number of further results that build on it, such as Proposition 3, can be extended to the case that not $\underline{P}_N$ but $\overline{P}_N$ is strictly positive. We have no proof yet, however.

On the other hand, and in contradistinction with what we are used to in Bayesian nets, we will not generally have separation in the following configuration:

$$X_{i_1} \longleftarrow X_{i_2} \longleftarrow X_t$$

where X_{i_2} *does not necessarily separate* X_t *from* X_{i_1}. We will come across a simple counterexample in Sec. 7. Where does this difference with the case of Bayesian nets originate? It is clear from the reasoning above that X_{i_2} separates X_{i_1} from X_t: conditional on X_{i_2}, X_t is epistemically irrelevant to X_{i_1}. For precise probability models, irrelevance generally implies symmetrical independence, and therefore this will generally imply that conditional on X_{i_2}, X_{i_1} is epistemically irrelevant to X_t as well. But for imprecise probability models no such symmetry is guaranteed [2], and we therefore cannot infer that, generally speaking, X_{i_2} will separate X_{i_1} from X_t. *As a general rule, we can only infer separation if the arrows point from the 'separating' variable* X_{i_2} *towards the 'target' variable* X_t.

6 Algorithm for treating the imprecise Markov tree as an expert system

We now consider the case where the imprecise Markov tree is treated as an expert system: we are interested in making inferences about the value of the variable X_t in some *target node* t, when we know the values x_E of the variables X_E in a set $E \subseteq T \setminus \{t\}$ of *evidence nodes*.

The formulation of the problem. If we assume that the values of the remaining variables are missing at random, then we can do this by conditioning the joint $\underline{P}$ obtained above on the available evidence '$X_E = x_E$'. We will address this problem by updating the lower prevision $\underline{P}$ to the lower prevision $\underline{R}_t(\cdot|x_E)$ on $\mathscr{L}(\mathscr{X}_t)$ using *regular extension* [15, Appendix J]:

$$\underline{R}_t(g|x_E) = \max\{\mu \in \mathbb{R}\colon \underline{P}(I_{\{x_E\}}[g-\mu]) \geq 0\} \quad (7)$$

for all gambles g on $\mathscr{X}_t$, assuming that $\overline{P}(\{x_E\}) > 0$. Consider the map $\rho_g\colon \mathbb{R} \to \mathbb{R}\colon \mu \mapsto \underline{P}(I_{\{x_E\}}[g-\mu])$. By coherence of $\underline{P}$, $|\rho_g(\mu_1) - \rho_g(\mu_2)| \leq |\mu_1 - \mu_2|\overline{P}(\{x_E\})$, which implies that ρ_g is continuous. Coherence of $\underline{P}$ also guarantees that ρ_g is concave and non-increasing. Hence $\{\mu \in \mathbb{R}\colon \rho_g(\mu) \geq 0\} = (-\infty, \underline{R}_t(g|x_E)]$, which shows that the supremum that we should have *a priori* used in (7) is indeed a maximum. $\underline{R}_t(g|x_E)$ is the right-most zero of ρ_g, and it is, again by coherence of $\underline{P}$, guaranteed to lie between $\inf g$ and $\sup g$. If moreover $\underline{P}(\{x_E\}) > 0$, then it is the unique zero. It appears that any algorithm for calculating $\underline{R}_t(g|x_E)$ will benefit from being able to calculate the values of ρ_g, or at least check their signs, efficiently.

Calculating the values of ρ_g recursively. Recall that the joint $\underline{P}$ can be constructed recursively from leaves to root. The idea we now use is that calculating $\rho_g(\mu) = \underline{P}(I_{\{x_E\}}[g-\mu])$ becomes easier if we graft the structure of the tree onto the argument $g^\mu := I_{\{x_E\}}[g-\mu]$ as follows. Define $g^\mu_e := I_{\{x_e\}}$ for all $e \in E$, $g^\mu_t := g - \mu$, and $g^\mu_s := 1$ for $s \in T \setminus (E \cup \{t\})$, whence $g^\mu = \prod_{s \in T} g^\mu_s$. Also define, for any $s \in T$, the gamble ϕ^μ_s on $\mathscr{X}_{\downarrow s}$ by $\phi^\mu_s := \prod_{u \in \downarrow s} g^\mu_u$. Then $\phi^\mu_\square = g^\mu$, $\phi^\mu_s \geq 0$ if $s \not\sqsubseteq t$, and for any $s \in T$:

$$\phi^\mu_s = g^\mu_s \prod_{c \in C(s)} \phi^\mu_c, \quad (8)$$

where we use the convention that $\prod_{u \in \emptyset} \alpha_u = 1$. Eq. (8) is the argument counterpart of Eq. (5). Also, if $s \not\sqsubseteq t$ then g^μ_s and ϕ^μ_s do not depend on μ, nor on g.

First, let us consider any node $s \not\sqsubseteq t$. We define the *messages* $\underline{\pi}_s$ and $\overline{\pi}_s$ recursively by

$$\underline{\pi}_s := \underline{Q}_s\left(g^\mu_s \prod_{c \in C(s)} \underline{\pi}_c | X_{m(s)}\right) \quad \overline{\pi}_s := \overline{Q}_s\left(g^\mu_s \prod_{c \in C(s)} \overline{\pi}_c | X_{m(s)}\right), \quad (9)$$

summarised by the self-explanatory shorthand notation: $\overline{\pi}_s = \overline{\underline{Q}}_s(g^\mu_s \prod_{c \in C(s)} \overline{\underline{\pi}}_c | X_{m(s)})$. There are two possibilities:

$$\overline{\underline{\pi}}_s = \begin{cases} \overline{\underline{Q}}_s\left(\{x_s\}|X_{m(s)}\right) \prod_{c \in C(s)} \overline{\underline{\pi}}_c(x_s) & \text{if } s \in E \\ \overline{\underline{Q}}_s\left(\prod_{c \in C(s)} \overline{\underline{\pi}}_c | X_{m(s)}\right) & \text{if } s \notin E. \end{cases}$$

The messages $\underline{\pi}_s$ and $\overline{\pi}_s$ can be seen as tuples of real numbers, with as many components as there are elements in $\mathscr{X}_{m(s)}$: one for each of the possible values of $X_{m(s)}$. As their notation suggests, they do not depend on the choice of g or μ, but only (at most) on which nodes are *instantiated*, i.e., belong to E, and on which values x_E the variables for these instantiated nodes assume. It then follows from Eqs. (5) and (8) and the factorisation property[2] of the local product lower previsions that:

$$\underline{P}_s(\phi^\mu_s | X_{m(s)}) = \underline{\pi}_s \text{ and } \overline{P}_s(\phi^\mu_s | X_{m(s)}) = \overline{\pi}_s. \quad (10)$$

Next, we turn to nodes $s \sqsubseteq t$. Define the messages π^μ_s by

$$\pi^\mu_s := \underline{Q}_s(\psi^\mu_s | X_{P(s)}), \quad (11)$$

where the gambles ψ^μ_s on $\mathscr{X}_s$ are given by the recursion relations:

$$\psi^\mu_t := \max\{g-\mu, 0\} \prod_{c \in C(t)} \underline{\pi}_c + \min\{g-\mu, 0\} \prod_{c \in C(t)} \overline{\pi}_c, \quad (12)$$

and for each $\square \neq s \sqsubseteq t$, so $m(s)$ exists,

$$\psi^\mu_{m(s)} := g^\mu_{m(s)} \left[\max\{\pi^\mu_s, 0\} \prod_{c \in S(s)} \underline{\pi}_c + \min\{\pi^\mu_s, 0\} \prod_{c \in S(s)} \overline{\pi}_c\right]. \quad (13)$$

[2] This shows that the results of updating the tree (and the algorithm we are deriving) in this way will be exactly the same *for any way* of forming a product of the local models for the children of s, *provided only that this product is factorising*. For instance, using the strong product and the net-independent natural extension will lead to the same inferences.

The messages π_s^μ are again tuples of real numbers, with one component for each of the possible values of $X_{m(s)}$.[3] They depend on the choice of g or μ, as well as on which nodes are instantiated and on which values x_E the variables for these instantiated nodes assume. It then follows from Eqs. (5) and (8) and the factorisation property that

$$\underline{P}_s(\phi_s^\mu | X_{P(s)}) = \pi_s^\mu, \tag{14}$$

and of course $\rho_g(\mu) = \pi_\square^\mu$. We conclude that we can find the value of $\rho_g(\mu)$ by a backwards recursion method consisting in passing messages up to the root of the tree, and in transforming them in each node using the local uncertainty models; see Eqs. (9) and (11)–(13).

There is a further simplification, because we are not necessarily interested in the actual value of $\rho_g(\mu)$, but rather in its sign. It arises whenever there are instantiated nodes above the target node: $E \cap A(t) \neq \emptyset$. Let in that case e_t be the greatest element of the chain $E \cap A(t)$, i.e., the instantiated node closest to t, and let s_t be its successor in the chain $\uparrow t$. If we let $\lambda_g(\mu)$ be the real number

$$\max\{\pi_{s_t}^\mu(x_{e_t}),0\}\prod_{c\in S(s_t)}\underline{\pi}_c(x_{e_t}) + \min\{\pi_{s_t}^\mu(x_{e_t}),0\}\prod_{c\in S(s_t)}\overline{\pi}_c(x_{e_t}),$$

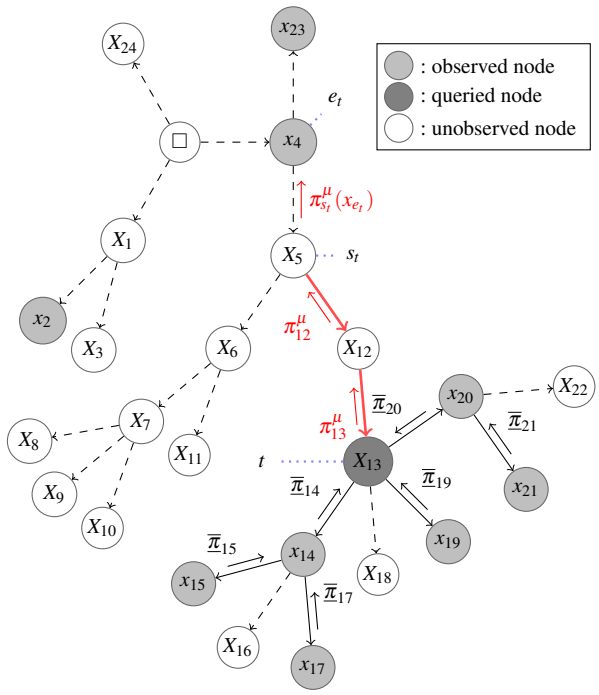
then it follows from Eq. (12) that $\psi_{e_t}^\mu = I_{\{x_{e_t}\}}\lambda_g(\mu)$. If we now continue to use Eqs. (12) and (13) until the root of the tree, we eventually find that

$$\rho_g(\mu) = \begin{cases} \underline{P}(I_{\{x_{e_t}\}})\lambda_g(\mu) & \text{if } \lambda_g(\mu) \geq 0 \\ \overline{P}(I_{\{x_{e_t}\}})\lambda_g(\mu) & \text{if } \lambda_g(\mu) \leq 0. \end{cases}$$

Since we assumed that $\overline{P}(I_{\{x_E\}}) > 0$, it readily follows that $\overline{P}(I_{\{x_{e_t}\}}) > 0$, so we gather from Eq. (7) that $\underline{R}_t(g|x_E) = \max\{\mu \in \mathbb{R}\colon \lambda_g(\mu) \geq 0\}$. In fact, under the assumption that $\overline{P}(I_{\{x_E\}}) > 0$, $\lambda_g(\mu) \geq 0$ can be replaced in this expression by $\pi_{s_t}^\mu(x_{e_t}) \geq 0$. We conclude that in order to do expert system inference of the type described above, we can perform all calculations on the subtree $\downarrow s_t$, where the new root s_t has local model $\underline{Q}_{s_t}(\cdot|x_{e_t})$. This is also borne out by the discussion of the separation properties in Sec. 5.

An algorithm. We now convert these observations into a workable algorithm. Using regular extension and message passing, we are able to compute $\underline{R}_t(g|x_E)$; we (i) choose a $\mu \in [\min g, \max g]$; (ii) calculate the value of $\lambda_g(\mu)$ by sending messages from the terminal nodes towards the root; and (iii) look for the maximal μ that will make this $\lambda_g(\mu)$ zero. But we have seen above that this naive approach can be sped up by exploiting the separation properties of the tree, and the independence of μ for some of the messages. For a start, as we are only interested in the sign of $\rho_g(\mu)$, which is determined by $\pi_{s_t}^\mu(x_{e_t})$, we only have to take nodes into consideration that strictly follow e_t.

[3] Of course, if s is the root node, then $P(s) = \emptyset$ and π_s^μ is just a single real number, which by Eq. (14) is equal to $\rho_g(\mu)$.



The next thing a smarter implementation of the algorithm can do is determine the *trunk* $\tilde{T}$ of the tree: those nodes that precede the queried node t and strictly follow the greatest observed element e_t preceding t. For the tree above for instance, where X_{13} (in grey) is the queried node and the light grey nodes $\{X_2, X_4, X_{14}, X_{15}, X_{17}, X_{19}, X_{20}, X_{21}, X_{23}\}$ are instantiated, the trunk consists of $\tilde{T} = \{X_5, X_{12}, X_{13}\}$.

The start of the algorithm can be implemented with the piece of pseudo-code on the left. Here, the queried node t is known in advance and besides the trunk $\tilde{T}$, also the nodes s_t and e_t are computed. We are especially interested in the nodes that constitute the trunk, because only these nodes will send messages to their parents that depend on μ. As a consequence, we can summarise all the μ-independent messages by propagating all messages until they reach the trunk, which means that they have to be calculated only once.

```
s_t := t
T̃ := {t}
while m(s_t) ∉ E do:
    T̃ := T̃ ∪ m(s_t)
    s_t := m(s_t)
end while
e_t := m(s_t)
```

The following piece of pseudocode does the trick. Both $\underline{\pi}_c$ and $\overline{\pi}_c$ can be calculated in the recursive manner outlined in Eq. (10), where the recursion starts at the leaves and moves up to (but stops right before) the trunk. In the leaves, the local lower and upper previsions of the indicator of the evidence are sent upwards if the leaf is instantiated; if not the constant 1 is sent up,

```
for n ∈ T̃ do:
    for c ∈ C(n) do:
        if c ∉ T̃ then:
            calculate π̄_c
        end if
    end for
    Π̲_n := ∏_{c∈C(s)\T̃} π̄_c
end for
```

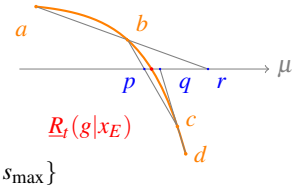
which is equivalent to deleting the node from the tree. We could envisage removing *barren nodes* (all of whose descendants are uninstantiated, such as $X_6, \ldots, X_{11}, X_{16}, X_{18}, X_{22}$ in the example tree above) from the tree beforehand, but we believe the computational overhead created by the search for them will void the gain.

At this point we can calculate $\pi^{\mu}_{s_t}(e_t)$. If we assume that t, s_t, g, $\underline{\Pi}_n$ and $\overline{\Pi}_n$ for $n \in \tilde{T}$ are stored as global variables, the following function will do the job. Now that we have the code to calculate $\pi^{\mu}_{s_t}(e_t)$, we can tackle the final problem: find the maximal μ for which $\pi^{\mu}_{s_t}(e_t) = 0$. In principle, a secant root-finding method could be used, but considering the computational complexity of the getJoint function, and using that $\pi^{\mu}_{s_t}(e_t)$ is concave, we can speed up the calculation of the maximal root drastically as shown in the figure below.

function getJoint(μ)
  $s := t$
  while $s \neq s_t$ **do:**
    calculate ψ^{μ}_{s}
    $\pi^{\mu}_{s} := \underline{Q}_s(\psi^{\mu}_{s}|X_{m(s)})$
    $s := m(s)$
  end while
  calculate $\psi^{\mu}_{s_t}$
  $\pi^{\mu}_{s_t}(e_t) := \underline{Q}_{s_t}(\psi^{\mu}_{s_t}|x_{e_t})$
  return $\pi^{\mu}_{s_t}(e_t)$

If a, b, c, and d are distributed in such a way that $\rho_g(a) \geq \rho_g(b) \geq 0 \geq \rho_g(c) \geq \rho_g(d)$, then the root of ρ_g is in the interval $[s_{\min}, s_{\max}] := [p, \min\{p, r\}]$.

function concaveRoot($a, b, c, d, s_{\min}, s_{\max}$)
  $\mu := \frac{1}{2}(s_{\min} + s_{\max})$
  $f(\mu) :=$ getJoint(μ)
  if $f(\mu) > 0$ **then:**
    $a := b$
    $b := (\mu, f(\mu))$
    $s_{\max} = \min\{b_x - \frac{b_x - a_x}{b_y - a_y} b_y, s_{\max}\}$
  else
    $d := c$
    $c := (\mu, f(\mu))$
    $s_{\max} = \min\{d_x - \frac{d_x - c_x}{d_y - c_y} d_y, s_{\max}\}$
  end if
  $s_{\min} = b_x - \frac{b_x - c_x}{b_y - c_y} b_y$
  if $s_{\max} - s_{\min} <$ tolerance **then:**
    return $s_{\min}$
  else
    return concaveRoot($a, b, c, d, s_{\min}, s_{\max}$)
  end if



Here, $s_{\min}$ is prefered over $s_{\max}$ as return value to stay on the conservative (small) side. If $b_y - a_y = 0$, then we define $\min\{b_x - \frac{b_x - a_x}{b_y - a_y} b_y, s_{\max}\}$ to be equal to $s_{\max}$ and similarly for $d_y - c_y = 0$. Keeping this in mind, we can finalise our algorithm by invoking a call to the following function.

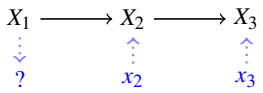
function getLowerPrevisionGivenEvidence(g)
  $a := (\min(g),$ getJoint$(a_x))$
  $d_x := (\max(g),$ getJoint$(a_d))$
  return concaveRoot(a, a, d, d, a_x, d_x)

The complexity of our algorithm is something that should be investigated further. But we can say something taking into account that for a fixed μ each node makes a single local computation and then propagates the result to the mother node: this implies that, with μ fixed, the algorithm is linear in the number of nodes. The iterations on μ create some additional complexity, but the number of iterations is usually small: a quick graphical investigation shows that the computational complexity of our root-finding algorithm must be lower than for the secant and bisection algorithms. We even have some experimental evidence that our root finder can outperform the Newton-Raphson method. Therefore, we can reasonably take the number of iterations to be a small constant for all practical applications, and conclude that the complexity of the algorithm is essentially linear in the number of nodes.

7 A simple example involving dilation

We present a very simple example that allows us to (i) follow the expert system inference method discussed above in a step-by-step fashion; (ii) see that there are separation properties for credal nets under strong independence that fail for credal trees under epistemic irrelevance; and (iii) see that in that case we will typically observe dilation.

Consider the following imprecise Markov chain:



To make things as simple as possible, we suppose that $\mathscr{X}_1 = \{a, b\}$ and that $\underline{Q}_1$ is a linear model Q_1 with mass function q. We also assume that $\underline{Q}_2(\cdot|X_1)$ is a linear model $Q_2(\cdot|X_1)$ with conditional mass function $q(\cdot|X_1)$. We make no such restrictions on the local model $\underline{Q}_3(\cdot|X_2)$. We also use following simplifying notational device: if we have three real numbers $\underline{\kappa}$, $\overline{\kappa}$ and γ, we let

$$\overline{\kappa}\langle\gamma\rangle := \underline{\kappa}\max\{\gamma, 0\} + \overline{\kappa}\min\{\gamma, 0\}.$$

We observe $X_2 = x_2$ and $X_3 = x_3$, and want to make inferences about the target variable X_1: for any $g \in \mathscr{L}(\mathscr{X}_1)$, we want to know $\underline{R}_1(g|x_{\{2,3\}})$. Letting $\underline{r} := \underline{R}_1(\{a\}|x_{\{2,3\}})$ and $\overline{r} := \overline{R}_1(\{a\}|x_{\{2,3\}})$, we infer from coherence that it suffices to calculate $\underline{r}$ and $\overline{r}$, because

$$\underline{R}_1(g|x_{\{2,3\}}) = g(b) + \overline{r}\langle g(a) - g(b)\rangle.$$

We let $g^{\mu} = [I_{\{a\}} - \mu]I_{\{x_2\}}I_{\{x_3\}}$, and apply the approach of the previous section. We see that the trunk $\tilde{T} = \{1\}$, and the instantiated leaf node 3 sends up the messages $\overline{\pi}_3 = \overline{Q}_3(\{x_3\}|X_2)$ to the instantiated node 2, who transforms them into the messages

$$\overline{\pi}_2 = \overline{Q}_2(\{x_2\}|X_1)\overline{\pi}_3(x_2) = q(x_2|X_1)\overline{q}.$$

These are sent up to the (target) root node $t = 1$, which transforms them into the message $\pi_1^\mu = Q_1(\psi_1^\mu)$ with $\psi_1^\mu = q(x_2|X_1)\underline{q}\langle I_{\{a\}} - \mu\rangle$. If we also use that $0 \leq \mu \leq 1$, this leads to

$$\underline{P}_1(g^\mu) = \pi_1^\mu = q(a)q(x_2|a)\underline{q}[1-\mu] + q(b)q(x_2|b)\overline{q}[-\mu],$$

so we find after applying regular extension that

$$\underline{r} = \underline{R}_1(\{a\}|x_{\{2,3\}}) = \frac{q(a)q(x_2|a)\underline{q}}{q(a)q(x_2|a)\underline{q} + q(b)q(x_2|b)\overline{q}}$$
$$\overline{r} = \overline{R}_1(\{a\}|x_{\{2,3\}}) = \frac{q(a)q(x_2|a)\overline{q}}{q(a)q(x_2|a)\overline{q} + q(b)q(x_2|b)\underline{q}}.$$

When $\underline{q} = \overline{q}$, which happens for instance if the local model for X_3 is precise, then we see that, with obvious notations,

$$\overline{r} = \underline{r} = \frac{q(a)q(x_2|a)}{q(a)q(x_2|a) + q(b)q(x_2|b)} =: p(a|x_2) \qquad (15)$$

and therefore X_2 indeed separates X_3 from X_1. But in general, letting $\alpha := q(a)q(x_2|a)$ and $\beta := q(b)q(x_2|b)$, we get

$$\overline{r} - \underline{r} = \frac{\alpha\beta(\overline{q}^2 - \underline{q}^2)}{(\alpha^2+\beta^2)\underline{q}\overline{q} + \alpha\beta(\underline{q}^2 + \overline{q}^2)}$$
$$\overline{r} - p(a|x_2) = \frac{\alpha\beta}{\alpha+\beta}\frac{\overline{q}-\underline{q}}{\alpha\overline{q}+\beta\underline{q}}$$
$$p(a|x_2) - \underline{r} = \frac{\alpha\beta}{\alpha+\beta}\frac{\overline{q}-\underline{q}}{\alpha\underline{q}+\beta\overline{q}}.$$

As soon as $\overline{q} > \underline{q}$, X_2 no longer separates X_3 from X_1, and we witness *dilation* [10, 14] because of the additional observation of X_3!

8 Numerical comparison

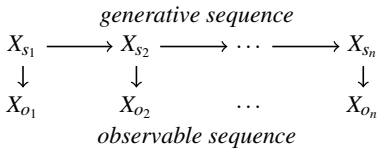
Strong independence implies epistemic irrelevance, but the converse does not generally hold. This implies that inferred probability intervals for epistemic irrelevance will generally include the ones for strong independence [3]. Here, we report on results of a number of numerical tests involving updating the tree. As noted in Sec. 5, the two models have different separation properties: this is particularly important when evidence is back-propagated from leaves to root. For this reason, we compare posterior (lower and upper) probabilities for the root variable of a *chain* when the leaf node variable is instantiated.

We have used the algorithm in Sec. 6 to compute posterior probability intervals in the irrelevance case, while the procedure in [5] is employed in the strong independence case. Inferred intervals for the former turn out to be clearly wider, and a mean square difference of about .2 is observed when considering 100 chains with three or four ternary variables and credal sets with three randomly generated extreme points. For longer chains, the updating with strong independence is too slow and no comparison can be made. Yet, similar results are observed in binary chains, for which the *2U algorithm* [9] can be used for efficient update in the strong independence case. In summary, there is a non-negligible difference between inferences based on the two notions of 'independence'.

9 An application: imprecise HMMs

Hidden Markov models (HMMs, [13]) are popular tools for modelling generative sequences, characterised by an underlying process generating an observable sequence. They have applications in many areas of signal processing, and more specifically in speech and text processing.

Both the generative and the observable sequence are described by sets of variables over the same domain $\mathscr{X}$, denoted respectively by $X_{s_1}, \ldots, X_{s_n}$ and $X_{o_1}, \ldots, X_{o_n}$. The independence assumptions between these variables, which characterise HMMs, are those corresponding to the tree structure below. Informally, this topology states that every element of the generative sequence depends only on its predecessor, while each observation depends only on the corresponding element of the generative sequence.



A local uncertainty model should be defined for each variable. In the more usual case of precise probabilistic assessments, this corresponds to linear versions of the local models $\underline{Q}_{s_1}$, $\underline{Q}_{s_{k+1}}(\cdot|X_{s_k})$ and $\underline{Q}_{o_k}(\cdot|X_{s_k})$, $k = 1, \ldots, n$, where the conditional models are assumed to be *stationary*, i.e., independent of k. These model, respectively, beliefs about the first state in the generative sequence, the transitions between adjacent states, and the observation process.

Bayesian techniques for learning from multinomial data are usually employed for identifying these models. But, especially if only few data are available, other methods leading to imprecise assessments, such as the *imprecise Dirichlet model* (IDM, [16]), might offer a more realistic model of the local uncertainty. For example, for the unconditional local model $\underline{Q}_{s_1}$, applying the IDM leads to the following simple identification:

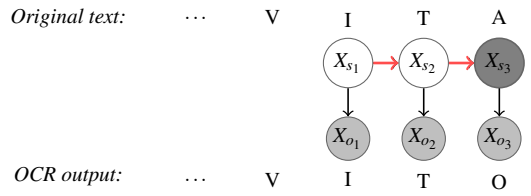
$$\underline{Q}_{s_1}(\{x_1\}) = \frac{n_{x_1}^{s_1}}{s + \sum_{x\in\mathscr{X}} n_x^{s_1}} \quad \overline{Q}_{s_1}(\{x_1\}) = \frac{s + n_{x_1}^{s_1}}{s + \sum_{x\in\mathscr{X}} n_x^{s_1}}, \qquad (16)$$

where $n_{x_1}^{s_1}$ counts the units in the sample for which $X_{s_1} = x_1$, and s is a hyperparameter that expresses the degree of caution in the inferences. For the conditional local models, we can proceed similarly. This leads to the identification of

an *imprecise HMM*, a special credal tree under epistemic irrelevance, like the ones introduced in Sec. 2.

Generally speaking, the algorithm described in Sec. 6 can be used for computing inferences with such imprecise HMMs. Below, we address the more specific problem of *on-line recognition*, which consists in the identification of the most likely value of X_{s_n}, given the evidence for the whole observational sequence $X_{o_1} = x_{o_1}, \ldots, X_{o_n} = x_{o_n}$. For precise local models, this problem requires the computation of the state $\tilde{x}_{s_n} := \operatorname{argmax}_{x_{s_n} \in \mathscr{X}} P(\{x_{s_n}\}|x_{o_1}, \ldots, x_{o_n})$ that is most probable after the observation. For imprecise local models different criteria can be adopted. We consider *maximality*: we order the states by $x_{s_n} > z_{s_n}$ iff $\underline{P}(I_{\{x_{s_n}\}} - I_{\{z_{s_n}\}}|x_{o_1}, \ldots, x_{o_n}) > 0$, and we look for the *undominated* or *maximal* states under this order. This may produce *indeterminate* predictions: the set of the undominated states can have more than one element.

Online character recognition by imprecise HMMs. As a very first application of the imprecise HMM, we have considered a *character recognition* problem. A written text was regarded as a generative sequence, while the observable sequence was obtained by artificially corrupting the text. This is a model for a not perfectly reliable observation process, such as the output of an OCR device. The local models were identified using the IDM, as in (16), by counting the occurrences of single characters and the "transitions" from one character to another in the generative sequence, and by matchings between the elements of the two sequences. By modelling text as a generative sequence, we obviously ignore any correlation there might be between a character and its nth predecessor (with $n \geq 2$). A better, albeit still not completely realistic, model would resort to using n-grams (i.e., clusters of n characters with $n \geq 2$) instead of monograms. Such models might lead to higher accuracy, but they need larger data sets for their quantification, because of the exponentially larger number of possible transitions for which probabilities have to be estimated. The figure below depicts how on-line recognition through HMM might apply to this setup.



The performance of the precise model can be characterised by its *accuracy* (the percentage of correct predictions) alone. The imprecise HMM requires more indicators. We follow [1] in using *determinacy* (percentage of determinate predictions), *set-accuracy* (percentage of indeterminate predictions containing the right state), *single accuracy* (percentage of correct predictions computed considering only determinate predictions), and *indeterminate output size* (average number of states returned when the prediction is indeterminate).

Accuracy	93.96%	(7275/7743)
Accuracy (if imprecise indeterminate)	64.97%	(243/374)
Determinacy	95.17%	(7369/7743)
Set-accuracy	93.58%	(350/374)
Single accuracy	95.43%	(7032/7369)
Indeterminate output size	2.97	over 21

Table 1: Precise vs. imprecise HMMs. Test results obtained by twofold cross-validation on the first two chants of Dante's *Divina Commedia* and $n = 2$. Quantification is achieved by IDM with $s = 2$ and Perks' prior (with the modification suggested in [17]). The single-character output by the precise model is then guaranteed to be included in the set of characters the imprecise HMM identifies.

The recognition using our algorithm is fast: it never takes more than one second for each character. Table 1 reports descriptor values for a large set of simulations, and a comparison with precise model performance. Imprecise HMMs guarantee quite accurate predictions. In contrast with the precise model, there are 'indeterminate' instances for which they do not output a single state. Yet, this happens rarely, and even then we witness a remarkable reduction in the number of undominated states (from the 21 letters of the Italian alphabet to less than three). Interestingly, the instances for which the imprecise probability model returns more than a single state appear to be "difficult" for the precise probability model: the accuracy of the precise models displays a strong decrease if we focus only on these instances, while the imprecise models here display basically the same performance as for other instances, by returning about three characters instead of a single one.

10 Conclusions

We have defined credal trees using Walley's epistemic irrelevance and have developed an efficient exact algorithm for updating beliefs on the tree. Like the algorithms developed for precise graphical models, our algorithm works in a distributed fashion by passing messages along the tree. This leads to computing lower and upper conditional previsions (expectations) with a complexity that is essentially linear in the number of nodes in the tree.

It has been unclear until recently whether an algorithm with the features described above was at all feasible. Epistemic irrelevance is most easily formulated using coherent lower previsions, which have never been used before in the context of credal networks. Moreover, epistemic irrelevance is not as "well-behaved" as strong independence is with respect to the graphoid axioms for propagation of

probability in graphical models [4]. Our results are therefore very encouraging, and they have the potential to open up new avenues of research in credal nets. This is important because strong independence is not always the most suitable notion of independence in an imprecise probability context, and epistemic irrelevance has wider scope, as well as a natural behavioural interpretation.

There is one more issue we would like to clarify at this point. While our algorithm clearly is fully functional as soon as all observations have positive upper probability, we have only proved that it produces coherent inferences when their lower probability is positive; see Theorem 4. At the time of writing this, we have strong indications that our coherence results can be extended to include observations with zero lower but positive upper probability.

Avenues for future research seem to be many. It would be important to extend the algorithm at least to so-called *polytrees*, which are substantially more expressive graphs than trees are. It would be interesting also to study in more detail the separation properties induced by epistemic irrelevance on a graph. For applications, it would be very important to develop statistical methods specialised for credal nets under irrelevance that avoid introducing excessive imprecision in the process of inferring probabilities from data. This could be achieved, for instance, by using a single global IDM over the variables of the tree rather than many local ones, as in our experiments.

Acknowledgements

Research by De Cooman and Hermans has been supported by Flemish BOF-project 01107505. Research by Antonucci and Zaffalon has been partially supported by the Swiss NSF grants n. 200020-116674/1 and n. 200020-121785/1. This paper has benefitted from discussions with Serafín Moral and Fabio Cozman.

References

[1] G. Corani and M. Zaffalon. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *Journal of Machine Learning Research*, 9:581–621, 2008.

[2] I. Couso, S. Moral, and P. Walley. A survey of concepts of independence for imprecise probabilities. *Risk Decision and Policy*, 5:165–181, 2000.

[3] F. G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.

[4] F. G. Cozman and P. Walley. Graphoid properties of epistemic irrelevance and independence. *Annals of Mathematics and Artificial Intelligence*, 45(1-2):173–195, 2005.

[5] C. P. de Campos and F. G. Cozman. Inference in credal networks using multilinear programming. In *Proceedings of the Second Starting AI Researcher Symposium*, pages 50–61, Valencia, 2004. IOS Press.

[6] C. P. de Campos and F. G. Cozman. Computing lower and upper expectations under epistemic independence. *International Journal of Approximate Reasoning*, 44(3):244–260, 2007.

[7] G. de Cooman and F. Hermans. Imprecise probability trees: Bridging two theories of imprecise probability. *Artificial Intelligence*, 172(11):1400–1427, 2008.

[8] G. de Cooman, F. Hermans, and E. Quaeghebeur. Imprecise Markov chains and their limit behaviour. *Probability in the Engineering and Informational Sciences*, 2009. Accepted for publication.

[9] E. Fagiuoli and M. Zaffalon. 2U: an exact interval propagation algorithm for polytrees with binary variables. *Artificial Intelligence*, 106:77–107, 1998.

[10] T. Herron, T. Seidenfeld, and L. Wasserman. Divisive conditioning: further results on dilation. *Philosophy of Science*, 64:411–444, 1997.

[11] E. Miranda. Updating coherent lower previsions on finite spaces. *Fuzzy Sets and Systems*, 2009. In press.

[12] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

[13] L. Rabiner. A tutorial on HMM and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[14] T. Seidenfeld and L. Wasserman. Dilation for sets of probabilities. *The Annals of Statistics*, 21:1139–54, 1993.

[15] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[16] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996. With discussion.

[17] M. Zaffalon. Statistical inference of the naive credal classifier. In G. de Cooman, T. L. Fine, and T. Seidenfeld, editors, *ISIPTA '01 – Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, pages 384–393. Shaker Publishing, Maastricht, 2000.

Exchangeability for sets of desirable gambles

Gert de Cooman & Erik Quaeghebeur
Ghent University, SYSTeMS Research Group
{Gert.deCooman,Erik.Quaeghebeur}@UGent.be

Abstract

Sets of desirable gambles constitute a quite general type of uncertainty model with an interesting geometrical interpretation. We study exchangeability assessments for such models, and prove a counterpart of de Finetti's finite representation theorem. We show that this representation theorem has a very nice geometrical interpretation. We also lay bare the relationships between the representations of updated exchangeable models, and discuss conservative inference (natural extension) under exchangeability.

Keywords. desirability, real desirability, weak desirability, sets of desirable gambles, coherence, exchangeability, representation, natural extension, updating.

1 Introduction

In this paper, we bring together desirability, an interesting approach to modelling uncertainty, with exchangeability, a structural assessment for uncertainty models that is important for inference purposes.

Desirability, or the theory of (coherent) sets of desirable gambles, has been introduced with all main ideas present—as far as our search has unearthed—by Williams [18, 19, 20]. Building on de Finetti's betting framework [6], he considered the 'acceptability' of *one-sided* bets instead of *two-sided* bets. This relaxation leads one to work with cones of bets instead of with linear subspaces of them. The germ of the theory was, however, already present in Smith's work [15, p. 15], who used a (generally) open cone of 'exchange vectors' when talking about currency exchange. Both authors influenced Walley [16, Sec. 3.7 and App. F], who describes three variants (almost, really, and strictly desirable gambles) and emphasises the conceptual ease with which updated and posterior models can be obtained in this framework [17]. Moral [12, 13] then took the next step and applied the theory to study epistemic irrelevance, a structural assessment. De Cooman and Miranda [1] made a general study of transformational symmetry assessments for desirable gambles.

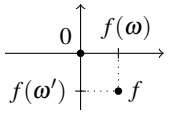
The structural assessment we are interested in here, is exchangeability. Conceptually, it says that the order of the samples in a sequence of them is irrelevant for inference purposes. The first detailed study of this concept was made by de Finetti [4], using the terminology of 'equivalent' events. He proved the now famous Representation Theorem, which is often interpreted as stating that a sequence of random variables is exchangeable if it is conditionally independent and identically distributed. Other important work—all using probabilities or previsions—was done by, amongst many others, Hewitt and Savage [9], Heath and Sudderth [8], and Diaconis and Freedman [7]. Exchangeability in the context of imprecise-probability theory—using lower previsions—was studied by Walley [16, Sec. 9.5] and more in-depth by De Cooman et al. [1–3]. The first embryonic study of exchangeability using desirability was recently performed by Quaeghebeur [14, Sec. 3.1.1].

In this paper, we present the first results of a more matured study of exchangeability using sets of desirable gambles.[1] First, in Sec. 2, we introduce the basics of the theory of desirable gambles. Then, in Sec. 3, we give a desirability-based analysis of finite exchangeable sequences, presenting a Representation Theorem and treating the issues of natural extension and updating under exchangeability.

2 Desirability

Consider a non-empty set Ω describing the possible and mutually exclusive outcomes of some experiment. We also consider a subject, who is uncertain about the outcome of the experiment.

A *gamble* f is a bounded real-valued map on Ω, and it is interpreted as an uncertain reward. When the actual outcome of the experiment is ω, then the corresponding (possibly negative) reward is $f(\omega)$, expressed in units of some pre-determined linear utility. This is illustrated for $\Omega = \{\omega, \omega'\}$. $\mathscr{G}(\Omega)$ denotes the set of all gambles on Ω.



[1] Proofs of this paper's results are included in Appendix A.

We say that a non-zero gamble f is *desirable* to a subject if he accepts to engage in the following transaction, where: (i) the actual outcome ω of the experiment is determined, and (ii) he receives the reward $f(\omega)$, i.e., his capital is changed by $f(\omega)$. The zero gamble is not considered to be desirable.[2]

2.1 Sets of desirable gambles

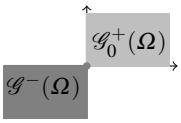
We try and model the subject's beliefs about the outcome of the experiment by considering which gambles are desirable for him. Suppose the subject has a set $\mathscr{R} \subseteq \mathscr{G}(\Omega)$ of desirable gambles.[3]

Definition 1 (Avoiding non-positivity and coherence)**.** *We say that a set of desirable gambles $\mathscr{R}$* avoids non-positivity *if $f \not\leq 0$ for all gambles f in* $\operatorname{coni}(\mathscr{R})$.[4] *Let $\mathscr{K}$ be a linear subspace of $\mathscr{G}(\Omega)$ such that $\mathscr{R} \subseteq \mathscr{K}$. Then we say that $\mathscr{R}$ is* coherent relative to $\mathscr{K}$ *if it satisfies the following* rationality requirements*, for all gambles f_1 and f_2 in $\mathscr{K}$ and all real $\lambda > 0$:*

D1. *if $f = 0$ then $f \notin \mathscr{R}$;*
D2. *if $f > 0$ then $f \in \mathscr{R}$ [accepting partial gain];*
D3. *if $f \in \mathscr{R}$ then $\lambda f \in \mathscr{R}$ [scaling];*
D4. *if $f_1, f_2 \in \mathscr{R}$ then $f_1 + f_2 \in \mathscr{R}$ [combination].*

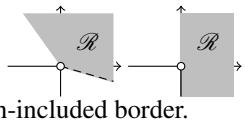
If $\mathscr{R}$ is coherent relative to $\mathscr{G}(\Omega)$, then we simply say that $\mathscr{R}$ is coherent*. We denote the set of coherent sets of desirable gambles by $\mathbb{D}(\Omega)$.*

Requirements D3 and D4 make $\mathscr{R}$ a *cone*: $\operatorname{coni}(\mathscr{R}) = \mathscr{R}$. Due to D2, it includes the positive gambles $\mathscr{G}_0^+(\Omega)$; due to D1, D2 and D4, it excludes the non-positive gambles $\mathscr{G}^-(\Omega)$:



D5. if $f \leq 0$ then $f \notin \mathscr{R}$.

We give two illustrations, the first is a general one and the second models certainty about ω happening. The dashed line indicates a non-included border.



[2] The nomenclature in the literature regarding desirability is somewhat confusing, and we have tried to resolve some of the ambiguity here. Our notion of desirability coincides with Walley's later [17] notion of desirability, initially also used by Moral [12]. Walley in his book [16, App. F] and Moral in a later paper [12] use another notion of desirability. The difference between the two approaches resides in whether the zero gamble is assumed to be desirable or not. We prefer to use the non-zero version here, because it is better behaved in conjunction with our notion of weak desirability in Definition 2.

[3] We use this convention throughout: subscripting a set with zero corresponds to removing zero (or the zero gamble) from the set, if present. For example $\mathbb{R}^+$ ($\mathbb{R}_0^+$) is the set of non-negative (positive) real numbers including (excluding) zero. Further notational conventions: $f \geq g$ iff $f(\omega) \geq g(\omega)$ for all ω in Ω; $f > g$ iff $f \geq g$ and $f \neq g$. The *conical hull operator* coni generates the set of (strictly!) positive linear combinations of elements of its argument set.

[4] A related, but weaker condition, is that $\mathscr{R}$ *avoids partial loss*, meaning that $f \not< 0$ for all gambles f in $\operatorname{coni}(\mathscr{R})$. We need the stronger condition because we have excluded the zero gamble from being desirable.

The intersection $\bigcap_{i\in I} \mathscr{R}_i$ of an arbitrary non-empty family of sets of desirable gambles $\mathscr{R}_i$, $i \in I$, is still coherent. This is the idea behind the following result.

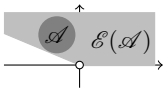
Theorem 1 (Natural extension)**.** *Consider an assessment, a set $\mathscr{A}$ of gambles on Ω, and define its* natural extension

$$\mathscr{E}(\mathscr{A}) := \bigcap \{\mathscr{R} \in \mathbb{D}(\Omega) \colon \mathscr{A} \subseteq \mathscr{R}\} \tag{1}$$
$$= \operatorname{coni}\big(\mathscr{G}_0^+(\Omega) \cup \mathscr{A}\big) \tag{2}$$

Then the following statements are equivalent:

(i) *$\mathscr{A}$ avoids non-positivity;*
(ii) *$\mathscr{A}$ is included in some coherent set of desirable gambles;*
(iii) *$\mathscr{E}(\mathscr{A}) \neq \mathscr{G}(\Omega)$;*
(iv) *$\mathscr{E}(\mathscr{A})$ is a coherent set of desirable gambles;*
(v) *$\mathscr{E}(\mathscr{A})$ is the smallest coherent set of desirable gambles that includes $\mathscr{A}$.*

With a small illustration, we can visualise natural extension as a conical hull operation:



2.2 Weakly desirable gambles, previsions & marginally desirable gambles

We now define *weak desirability*: a useful modification of Walley's [16, Section 3.7] notion of *almost-desirability*. Our conditions for a gamble f to be weakly desirable are more stringent than Walley's for almost-desirability: he only requires that adding any constant strictly positive amount of utility to f should make the resulting gamble desirable. We require that adding anything desirable (be it constant or not) to f should make the resulting gamble desirable. Weak desirability is better behaved under updating: we shall see in Proposition 12 that it makes sure that the exchangeability of a set of desirable gambles, whose definition hinges on the notion of weak desirability, is preserved under updating after observing a sample. This is not necessarily true if weak desirability is replaced by almost-desirability in the definition of exchangeability, as was for instance done in our earlier work [1].

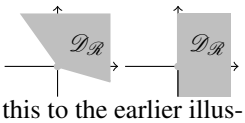
Definition 2 (Weak desirability)**.** *Consider a coherent set $\mathscr{R}$ of desirable gambles. Then a gamble f is called* weakly desirable *if $f + f'$ is desirable for all desirable f', i.e., if $f + f' \in \mathscr{R}$ for all f' in $\mathscr{R}$. We denote the set of weakly desirable gambles by $\mathscr{D}_{\mathscr{R}}$:*

$$\mathscr{D}_{\mathscr{R}} = \{f \in \mathscr{G}(\Omega) \colon f + \mathscr{R} \subseteq \mathscr{R}\}. \tag{3}$$

In particular, every desirable gamble is also weakly desirable, so $\mathscr{R} \subseteq \mathscr{D}_{\mathscr{R}}$.

Proposition 2. *Let $\mathscr{R}$ be a coherent set of desirable gambles, and let $\mathscr{D}_{\mathscr{R}}$ be the associated set of weakly desirable gambles. Then $\mathscr{D}_{\mathscr{R}}$ has the following properties, for all gambles f_1 and f_2 in $\mathscr{G}(\Omega)$ and all real $\lambda \geq 0$:*

WD1. *if $f < 0$ then $f \notin \mathscr{D}_{\mathscr{R}}$ [avoiding partial loss];*[5]
WD2. *if $f \geq 0$ then $f \in \mathscr{D}_{\mathscr{R}}$ [accepting partial gain];*
WD3. *if $f \in \mathscr{D}_{\mathscr{R}}$ then $\lambda f \in \mathscr{D}_{\mathscr{R}}$ [scaling];*
WD4. *if $f_1, f_2 \in \mathscr{D}_{\mathscr{R}}$ then $f_1 + f_2 \in \mathscr{D}_{\mathscr{R}}$ [combination].*



Like $\mathscr{R}$, $\mathscr{D}_{\mathscr{R}}$ is a cone, but it always includes all cone surface gambles (excluding those that incur a partial loss). We have applied this to the earlier illustrations; take note of border changes.

With a set of gambles $\mathscr{A}$, we associate a *lower prevision* $\underline{P}_{\mathscr{A}}$ and an *upper prevision* $\overline{P}_{\mathscr{A}}$ by letting

$$\underline{P}_{\mathscr{A}}(f) = \sup\{\mu \in \mathbb{R} \colon f - \mu \in \mathscr{A}\} \tag{4}$$

$$\overline{P}_{\mathscr{A}}(f) = \inf\{\mu \in \mathbb{R} \colon \mu - f \in \mathscr{A}\} \tag{5}$$

for all gambles f. Observe that $\underline{P}_{\mathscr{A}}$ and $\overline{P}_{\mathscr{A}}$ always satisfy the *conjugacy relation* $\underline{P}_{\mathscr{A}}(-f) = -\overline{P}_{\mathscr{A}}(f)$. We call a real functional $\underline{P}$ on $\mathscr{G}(\Omega)$ a *coherent lower prevision* if and only if there is some coherent set of desirable gambles $\mathscr{R}$ on $\mathscr{G}(\Omega)$ such that $\underline{P} = \underline{P}_{\mathscr{R}}$.

Theorem 3. *Let $\mathscr{R}$ be a coherent set of desirable gambles. Then $\underline{P}_{\mathscr{R}}$ is real-valued, $\underline{P}_{\mathscr{R}} = \underline{P}_{\mathscr{D}_{\mathscr{R}}}$, $\underline{P}_{\mathscr{R}}(f) \geq 0$ for all $f \in \mathscr{D}_{\mathscr{R}}$. Moreover, a real functional $\underline{P}$ is a coherent lower prevision iff it satisfies the following properties, for all gambles f_1 and f_2 in $\mathscr{G}(\Omega)$ and all real $\lambda \geq 0$:*

P1. *$\underline{P}(f) \geq \inf f$ [accepting sure gain];*
P2. *$\underline{P}(f_1 + f_2) \geq \underline{P}(f_1) + \underline{P}(f_2)$ [super-additivity];*
P3. *$\underline{P}(\lambda f) = \lambda \underline{P}(f)$ [non-negative homogeneity].*

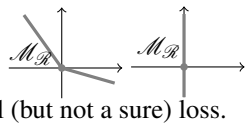
Finally, we turn to marginal desirability. Given a coherent set of desirable gambles $\mathscr{R}$, we define the associated set of *marginally desirable* gambles as

$$\mathscr{M}_{\mathscr{R}} := \{f - \underline{P}_{\mathscr{R}}(f) \colon f \in \mathscr{G}(\Omega)\}. \tag{6}$$

The set of marginally desirable gambles $\mathscr{M}_{\mathscr{R}}$ is completely determined by the lower prevision $\underline{P}_{\mathscr{R}}$. The converse is also true:

Proposition 4. *Let $\mathscr{R}$ be a coherent set of desirable gambles. Then $\underline{P}_{\mathscr{M}_{\mathscr{R}}} = \underline{P}_{\mathscr{R}}$ and*

$$\mathscr{M}_{\mathscr{R}} = \mathscr{M}_{\underline{P}_{\mathscr{R}}} := \{f \in \mathscr{G}(\Omega) \colon \underline{P}_{\mathscr{R}}(f) = 0\}. \tag{7}$$



The set of marginally desirable gambles $\mathscr{M}_{\mathscr{R}}$ is the entire cone surface of $\mathscr{R}$ and $\mathscr{D}_{\mathscr{R}}$, possibly including gambles that incur a partial (but not a sure) loss.

2.3 Updating sets of desirable gambles

Consider a set of desirable gambles $\mathscr{R}$ on Ω. With a non-empty subset B of Ω, we associate an *updated* set of desirable gambles on Ω, as defined by Walley [17]:

$$\mathscr{R}\|B := \{f \in \mathscr{G}(\Omega) \colon I_B f \in \mathscr{R}\}. \tag{8}$$

[5] Compare this to the less stringent requirement for almost-desirability [16, Section 3.7.3]: if $f \in \mathscr{D}_{\mathscr{R}}$ then $\sup f \geq 0$ [avoiding sure loss].

We find it more convenient to work with the following, slightly different but completely equivalent, version:

$$\mathscr{R}|B := \{f \in \mathscr{R} \colon I_B f = f\} = \mathscr{R} \cap \mathscr{G}(\Omega)|B, \tag{9}$$

which completely determines $\mathscr{R}\|B$: for all $f \in \mathscr{G}(\Omega)$,

$$f \in \mathscr{R}\|B \Leftrightarrow I_B f \in \mathscr{R}|B. \tag{10}$$

In our version, updating corresponds to intersecting the cone $\mathscr{R}$ with the linear subspace $\mathscr{G}(\Omega)|B$, which results in a cone $\mathscr{R}|B$ of lower dimension. And since we can uniquely identify a gamble $f = I_B f$ in $\mathscr{G}(\Omega)|B$ with a gamble on B, namely its restriction f_B to B, and *vice versa*, we can also identify $\mathscr{R}|B$ with a set of desirable gambles on B:

$$\mathscr{R}\rfloor B := \{f_B \colon f \in \mathscr{R}|B\} = \{f_B \colon f \in \mathscr{R}\|B\} \subseteq \mathscr{G}(B). \tag{11}$$

Proposition 5. *If $\mathscr{R}$ is a coherent set of desirable gambles on Ω, then $\mathscr{R}|B$ is coherent relative to $\mathscr{G}(\Omega)|B$, or equivalently, $\mathscr{R}\rfloor B$ is a coherent set of desirable gambles on B.*

Our subject takes $\mathscr{R}|B$ (or $\mathscr{R}\rfloor B$) as his set of desirable gambles contingent on observing the event B.

3 Finite exchangeable sequences

Now that we have become better versed in the theory of sets of desirable gambles, we are going to focus on the main topic: reasoning about finite exchangeable sequences. We first show how they are related to count vectors (Sec. 3.1). Then we are ready to give a desirability-based definition of exchangeability (Sec. 3.2) and treat natural extension and updating under exchangeability (Secs. 3.3 and 3.4). After presenting our Finite Representation Theorem (Sec. 3.5), we can show what natural extension and updating under exchangeability look like in terms of the count vector representation (Secs. 3.6 and 3.7).

Consider random variables $X_1, \ldots, X_N$ taking values in a non-empty finite set $\mathscr{X}$,[6] where $N \in \mathbb{N}_0$, i.e., a positive (non-zero) integer. The possibility space is $\Omega = \mathscr{X}^N$.

3.1 Count vectors

We denote by $x = (x_1, \ldots, x_N)$ an arbitrary element of $\mathscr{X}^N$. $\mathscr{P}_N$ is the set of all permutations π of the index set $\{1, \ldots, N\}$. With any such permutation π, we associate a permutation of $\mathscr{X}^N$, also denoted by π, and defined by $(\pi x)_k = x_{\pi(k)}$, or in other words, $\pi(x_1, \ldots, x_N) = (x_{\pi(1)}, \ldots, x_{\pi(N)})$. Similarly, we lift π to a permutation π^t of $\mathscr{G}(\mathscr{X}^N)$ by letting $\pi^t f = f \circ \pi$, so $(\pi^t f)(x) = f(\pi x)$.

[6] A lot of functions and sets introduced below will depend on the set $\mathscr{X}$. We do not indicate this explicitly, not to overburden the notation and because we do not consider different sets of values in this paper.

The permutation invariant atoms $[x] := \{\pi x\colon \pi \in \mathscr{P}_N\}$ are the smallest permutation invariant subsets of $\mathscr{X}^N$. We introduce the *counting map*

$$T^N\colon \mathscr{X}^N \to \mathscr{N}^N\colon x \mapsto T^N(x) \tag{12}$$

where $T^N(x)$ is the $\mathscr{X}$-tuple with components

$$T^N_z(x) := |\{k \in \{1,\dots,N\}\colon x_k = z\}| \quad \text{for all } z \in \mathscr{X}, \tag{13}$$

and the set of possible *count vectors* is given by

$$\mathscr{N}^N := \left\{m \in \mathbb{N}^{\mathscr{X}}\colon \sum_{x\in\mathscr{X}} m_x = N\right\}. \tag{14}$$

If $m = T^N(x)$, then $[x] = \left\{y \in \mathscr{X}^N\colon T^N(y) = m\right\}$, so the atom $[x]$ is completely determined by the count vector m of all its the elements, and is therefore also denoted by $[m]$.

3.2 Defining exchangeability

If a subject assesses that $X_1, \dots, X_N$ are exchangeable, this means that for any gamble f and any permutation π, he finds exchanging $\pi^t f$ for f weakly desirable,[7] because he is indifferent between them [cf. 16, Sec. 4.1.1]. Let

$$\mathscr{D}_{\mathscr{P}_N} := \left\{f - \pi^t f\colon f \in \mathscr{G}(\mathscr{X}^N) \text{ and } \pi \in \mathscr{P}_N\right\}, \tag{15}$$

then we should have that $\mathscr{D}_{\mathscr{P}_N} \subseteq \mathscr{D}_{\mathscr{R}}$. Before we give useful alternative characterisations of exchangeability, we introduce a few notions that will prove crucial further on.

We begin by defining a special linear transformation ex^N of the linear space of gambles $\mathscr{G}(\mathscr{X}^N)$:

$$\mathrm{ex}^N\colon \mathscr{G}(\mathscr{X}^N) \to \mathscr{G}(\mathscr{X}^N)\colon f \mapsto \mathrm{ex}^N(f) := \frac{1}{N!}\sum_{\pi\in\mathscr{P}_N} \pi^t f. \tag{16}$$

Observe that for all gambles f and all permutations π:

$$\mathrm{ex}^N(\pi^t f) = \mathrm{ex}^N(f) \text{ and } \pi^t\left(\mathrm{ex}^N(f)\right) = \mathrm{ex}^N(f). \tag{17}$$

So $\mathrm{ex}^N(f)$ is permutation invariant and therefore constant on the permutation invariant atoms $[m]$, and it assumes the same value for all gambles that can be related to each other through some permutation. What is the value that $\mathrm{ex}^N(f)$ assumes on $[m]$? It is not difficult to see that

$$\mathrm{ex}^N = \sum_{m\in\mathscr{N}^N} \mathrm{MuHy}^N(\cdot|m) I_{[m]}, \tag{18}$$

where we let

$$\mathrm{MuHy}^N(f|m) := \frac{1}{|[m]|}\sum_{y\in[m]} f(y) \tag{19}$$

$$|[m]| = \binom{N}{m} := \frac{N!}{\prod_{z\in\mathscr{X}} m_z!}. \tag{20}$$

$\mathrm{MuHy}^N(\cdot|m)$ is the linear expectation operator associated with the uniform distribution on the invariant atom $[m]$. It characterises a *multivariate hyper-geometric distribution* [10, Sec. 39.2], associated with random sampling without replacement from an urn with N balls of types $\mathscr{X}$, whose composition is characterised by the count vector m. If we also observe that $\mathrm{ex}^N \circ \mathrm{ex}^N = \mathrm{ex}^N$, we see that ex^N is the linear *projection operator* of $\mathscr{G}(\mathscr{X}^N)$ to the linear space

$$\mathscr{G}_{\mathscr{P}_N}(\mathscr{X}^N) := \left\{f \in \mathscr{G}(\mathscr{X}^N)\colon (\forall \pi \in \mathscr{P}_N)\pi^t f = f\right\} \tag{21}$$

of all permutation invariant gambles. We also let

$$\mathscr{D}_{\mathscr{U}_N} := \mathrm{span}(\mathscr{D}_{\mathscr{P}_N}) \tag{22}$$

$$= \left\{f - \mathrm{ex}^N(f)\colon f \in \mathscr{G}(\mathscr{X}^N)\right\} \tag{23}$$

$$= \left\{f \in \mathscr{G}(\mathscr{X}^N)\colon \mathrm{ex}^N(f) = 0\right\}, \tag{24}$$

where 'span' denotes linear span. The linear space $\mathscr{D}_{\mathscr{U}_N}$ is the kernel of the linear projection operator ex^N.

Definition 3 (Exchangeability)**.** *A coherent set $\mathscr{R}$ of desirable gambles on $\mathscr{X}^N$ is called* exchangeable *if any (and hence all) of the following equivalent conditions is (are) satisfied:*

(i) *any gamble in $\mathscr{D}_{\mathscr{P}_N}$ is weakly desirable: $\mathscr{D}_{\mathscr{P}_N} \subseteq \mathscr{D}_{\mathscr{R}}$;*
(ii) *$\mathscr{D}_{\mathscr{P}_N} + \mathscr{R} \subseteq \mathscr{R}$;*
(iii) *any gamble in $\mathscr{D}_{\mathscr{U}_N}$ is weakly desirable: $\mathscr{D}_{\mathscr{U}_N} \subseteq \mathscr{D}_{\mathscr{R}}$;*
(iv) *$\mathscr{D}_{\mathscr{U}_N} + \mathscr{R} \subseteq \mathscr{R}$;*

We call a lower prevision $\underline{P}$ on $\mathscr{G}(\mathscr{X}^N)$ exchangeable *if there is some exchangeable coherent set of desirable gambles $\mathscr{R}$ such that $\underline{P} = \underline{P}_{\mathscr{R}}$.*

The conditions (iii)–(iv) of this definition are quite closely related to the desirability version of a de Finetti-like representation theorem for finite exchangeable sequences in terms of sampling without replacement from an urn. They allow us talk about exchangeability without invoking permutations. This is what we will address in Section 3.5.

A number of useful results follow from this definition:

Proposition 6. *Let $\mathscr{R}$ be a coherent set of desirable gambles. If $\mathscr{R}$ is exchangeable then it is also* permutable*: $\pi^t f \in \mathscr{R}$ for all $f \in \mathscr{R}$ and all $\pi \in \mathscr{P}_N$.*

Proposition 7. *Let $\mathscr{R}$ be a coherent and exchangeable set of desirable gambles. For all gambles f and f' on $\mathscr{X}^N$:*

(i) *$f \in \mathscr{R} \Leftrightarrow \mathrm{ex}^N(f) \in \mathscr{R}$;*
(ii) *If $\mathrm{ex}^N(f) = \mathrm{ex}^N(f')$, then $f \in \mathscr{R} \Leftrightarrow f' \in \mathscr{R}$.*

It follows from this last proposition and Eq. (24) that for any coherent and exchangeable set of desirable gambles $\mathscr{R}$:

$$\mathscr{R} \cap \mathscr{D}_{\mathscr{U}_N} = \emptyset. \tag{25}$$

Theorem 8. *Let $\underline{P}$ be a coherent lower prevision on $\mathscr{G}(\mathscr{X}^N)$. Then the following statements are equivalent:*[8]

[7] Note that the gambles in $\mathscr{D}_{\mathscr{P}_N}$ cannot be assumed to be desirable, because $\mathscr{D}_{\mathscr{P}_N}$ does not avoid non-positivity.

[8] This shows that the exchangeability of a lower prevision can also be expressed using marginally desirable gambles [see 14, Sec. 3.1.1].

(i) $\underline{P}$ *is exchangeable;*
(ii) $\underline{P}(f) = \overline{P}(f) = 0$ *for all* $f \in \mathscr{D}_{\mathscr{P}_N}$*;*
(iii) $\underline{P}(f) = \overline{P}(f) = 0$ *for all* $f \in \mathscr{D}_{\mathscr{U}_N}$.

3.3 Exchangeable natural extension

Let us denote the set of all coherent and exchangeable sets of desirable gambles on $\mathscr{X}^N$ by

$$\mathbb{D}_{\text{ex}}(\mathscr{X}^N) := \left\{\mathscr{R} \in \mathbb{D}(\mathscr{X}^N)\colon \mathscr{D}_{\mathscr{U}_N} + \mathscr{R} \subseteq \mathscr{R}\right\}. \quad (26)$$

This set is closed under arbitrary non-empty intersections. We shall see further on in Corollary 11 that it is also non-empty, and therefore has a smallest element.

Suppose our subject has an assessment, or in other words, a set $\mathscr{A}$ of gambles on $\mathscr{X}^N$ that he finds desirable. Then we can ask if there is some coherent and exchangeable set of desirable gambles $\mathscr{R}$ that includes $\mathscr{A}$. In other words, we want a set of desirable gambles $\mathscr{R}$ to satisfy the requirements: (i) $\mathscr{R}$ is coherent; (ii) $\mathscr{A} \subseteq \mathscr{R}$; and (iii) $\mathscr{D}_{\mathscr{U}_N} + \mathscr{R} \subseteq \mathscr{R}$. Clearly, the intersection $\bigcap_{i\in I}\mathscr{R}_i$ of an arbitrary non-empty family of sets of desirable gambles $\mathscr{R}_i$, $i \in I$ that satisfy these requirements, will satisfy these requirements as well. This is the idea behind the following results.

Proposition 9. *We say that a set* $\mathscr{A}$ *of gambles on* $\mathscr{X}^N$ avoids non-positivity under exchangeability *if the set of gambles* $[\mathscr{G}_0^+(\mathscr{X}^N) \cup \mathscr{A}] + \mathscr{D}_{\mathscr{U}_N}$ *avoids non-positivity. Then: (i)* $\mathscr{R} = \emptyset$ *avoids non-positivity under exchangeability; and (ii) if* $\mathscr{A}$ *is non-empty, then* $\mathscr{A}$ *avoids non-positivity under exchangeability iff* $\mathscr{A} + \mathscr{D}_{\mathscr{U}_N}$ *avoids non-positivity.*

Theorem 10 (Exchangeable natural extension)**.** *Consider a set* $\mathscr{A}$ *of gambles on* $\mathscr{X}^N$*, and define its* exchangeable natural extension $\mathscr{E}^N_{\text{ex}}(\mathscr{A})$ *by*

$$\begin{aligned}\mathscr{E}^N_{\text{ex}}(\mathscr{A}) &:= \bigcap\left\{\mathscr{R} \in \mathbb{D}_{\text{ex}}(\mathscr{X}^N)\colon \mathscr{A} \subseteq \mathscr{R}\right\} && (27)\\ &= \text{coni}\left(\mathscr{D}_{\mathscr{U}_N} + [\mathscr{G}_0^+(\mathscr{X}^N) \cup \mathscr{A}]\right) && (28)\\ &= \mathscr{D}_{\mathscr{U}_N} + \mathscr{E}(\mathscr{A}). && (29)\end{aligned}$$

Then the following statements are equivalent:

(i) $\mathscr{A}$ avoids non-positivity under exchangeability*;*
(ii) $\mathscr{A}$ *is included in some coherent and exchangeable set of desirable gambles;*
(iii) $\mathscr{E}^N_{\text{ex}}(\mathscr{A}) \neq \mathscr{G}(\mathscr{X}^N)$*;*
(iv) $\mathscr{E}^N_{\text{ex}}(\mathscr{A})$ *is a coherent and exchangeable set of desirable gambles;*
(v) $\mathscr{E}^N_{\text{ex}}(\mathscr{A})$ *is the smallest coherent and exchangeable set of desirable gambles that includes* $\mathscr{A}$.

Corollary 11. *The set* $\mathbb{D}_{\text{ex}}(\mathscr{X}^N)$ *is non-empty, and has a smallest element*

$$\mathscr{R}^N_{\text{ex,v}} := \mathscr{E}^N_{\text{ex}}(\emptyset) = \mathscr{D}_{\mathscr{U}_N} + \mathscr{G}_0^+(\mathscr{X}^N). \quad (30)$$

3.4 Updating exchangeable models

Consider an exchangeable and coherent set of desirable gambles $\mathscr{R}$ on $\mathscr{X}^N$, and assume that we have observed the values $\check{x} = (\check{x}_1, \check{x}_2, \ldots, \check{x}_{\check{n}})$ of the first $\check{n}$ variables $X_1, \ldots, X_{\check{n}}$, and that we want to make inferences about the remaining $\hat{n} := N - \check{n}$ variables. To do this, we simply update the set $\mathscr{R}$ with the set $C_{\check{x}} = \{\check{x}\} \times \mathscr{X}^{\hat{n}}$, to obtain the set $\mathscr{R}|C_{\check{x}}$, also denoted as $\mathscr{R}|\check{x} = \left\{f \in \mathscr{R}\colon fI_{C_{\check{x}}} = f\right\}$. As we have seen in Section 2.3, this set can be identified with a coherent set of desirable gambles on $\mathscr{X}^{\hat{n}}$, which we denote by $\mathscr{R}\rfloor\check{x}$. With obvious notations:[9]

$$\mathscr{R}\rfloor\check{x} = \left\{f \in \mathscr{G}(\mathscr{X}^{\hat{n}})\colon fI_{C_{\check{x}}} \in \mathscr{R}\right\}. \quad (31)$$

We already know that updating preserves coherence. We now see that this type of updating on an observed sample also preserves exchangeability.

Proposition 12. *Consider* $\check{x} \in \mathscr{X}^{\check{n}}$ *and a coherent and exchangeable set of desirable gambles* $\mathscr{R}$ *on* $\mathscr{X}^N$*. Then* $\mathscr{R}\rfloor\check{x}$ *is a coherent and exchangeable set of desirable gambles on* $\mathscr{X}^{\hat{n}}$.

We also introduce another type of updating, where we observe a count vector $\check{m} \in \mathscr{N}^{\check{n}}$, and we update the set $\mathscr{R}$ with the set $C_{\check{m}} = [\check{m}] \times \mathscr{X}^{\hat{n}}$, to obtain the set $\mathscr{R}|C_{\check{m}}$, also denoted as $\mathscr{R}|\check{m} = \left\{f \in \mathscr{R}\colon fI_{C_{\check{m}}} = f\right\}$. This set can be identified with a coherent set of desirable gambles on $\mathscr{X}^{\hat{n}}$, which we also denote by $\mathscr{R}\rfloor\check{m}$. With obvious notations:

$$\mathscr{R}\rfloor\check{m} = \left\{f \in \mathscr{G}(\mathscr{X}^{\hat{n}})\colon fI_{C_{\check{m}}} \in \mathscr{R}\right\}. \quad (32)$$

Proposition 13 (Sufficiency of observed count vectors)**.** *Consider* $\check{x}, \check{y} \in \mathscr{X}^{\check{n}}$ *and a coherent and exchangeable set of desirable gambles* $\mathscr{R}$ *on* $\mathscr{X}^N$*. If* $\check{y} \in [\check{x}]$*, or in other words if* $T^{\check{n}}(\check{x}) = T^{\check{n}}(\check{y}) =: \check{m}$*, then* $\mathscr{R}\rfloor\check{x} = \mathscr{R}\rfloor\check{y} = \mathscr{R}\rfloor\check{m}$.

3.5 Finite representation

We now introduce the linear map MuHy^N from the linear space $\mathscr{G}(\mathscr{X}^N)$ to the linear space $\mathscr{G}(\mathscr{N}^N)$, as follows:

$$\begin{aligned}&\text{MuHy}^N\colon \mathscr{G}(\mathscr{X}^N) \to \mathscr{G}(\mathscr{N}^N)\colon\\ &\qquad f \mapsto \text{MuHy}^N(f) := \text{MuHy}^N(f|\cdot), \quad (33)\end{aligned}$$

so $\text{MuHy}^N(f)$ is the gamble on $\mathscr{N}^N$ that assumes the value $\text{MuHy}^N(f|m)$ in the count vector $m \in \mathscr{N}^N$. We also define the linear map T^N from the linear space $\mathscr{G}(\mathscr{N}^N)$ to the linear space $\mathscr{G}_{\mathscr{P}_N}(\mathscr{X}^N)$ as follows:

$$\text{T}^N\colon \mathscr{G}(\mathscr{N}^N) \to \mathscr{G}_{\mathscr{P}_N}(\mathscr{X}^N)\colon g \mapsto \text{T}^N(g) := g \circ T^N, \quad (34)$$

so $\text{T}^N(g)$ is the permutation invariant gamble on $\mathscr{X}^N$ that assumes the constant value $g(m)$ on the invariant atom $[m]$. For all $f \in \mathscr{G}(\mathscr{X}^N)$, $\text{ex}^N(f) = \text{T}^N\big(\text{MuHy}^N(f)\big)$, and similarly, for all $g \in \mathscr{G}(\mathscr{N}^N)$, $\text{MuHy}^N\big(\text{T}^N(g)\big) = g$. Hence:

$$\text{ex}^N = \text{T}^N \circ \text{MuHy}^N \text{ and } \text{MuHy}^N \circ \text{T}^N = \text{id}_{\mathscr{G}(\mathscr{N}^N)}. \quad (35)$$

[9] Here and further on we silently use cylindrical extension on gambles, i.e., let them 'depend' on extra variables whose value does not influence the value they take.

If we invoke Eq. (17) we find that

$$\mathrm{MuHy}^N(\pi^t f) = \mathrm{MuHy}^N(f). \tag{36}$$

Also taking into account the linearity of MuHy^N and Eq. (16), this leads to

$$\mathrm{MuHy}^N(\mathrm{ex}^N(f)) = \mathrm{MuHy}^N(f). \tag{37}$$

The relationships between the three important linear maps we have introduced above are clarified by the commutative diagram in Fig. 1.

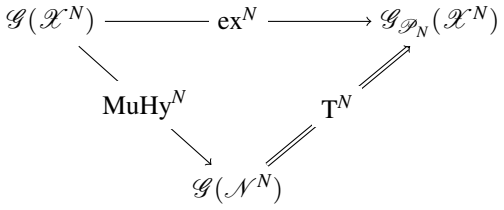


Figure 1: Single sequence length commutative diagram. Double arrows indicate a linear isomorphism.

For every gamble f on $\mathscr{X}^N$, $f = \mathrm{ex}^N(f) + [f - \mathrm{ex}^N(f)]$, so it can be decomposed as a sum of a permutation invariant gamble $\mathrm{ex}^N(f)$ and an element $f - \mathrm{ex}^N(f)$ of the kernel $\mathscr{D}_{\mathscr{U}_N}$ of the linear projection operator ex^N. Since we know that MuHy^N is a linear isomorphism between the spaces $\mathscr{G}_{\mathscr{P}_N}(\mathscr{X}^N)$ and $\mathscr{G}(\mathscr{N}^N)$, we now investigate whether we can represent coherent and exchangeable $\mathscr{R}$ by some set of desirable *count gambles* on $\mathscr{N}^N$.

Theorem 14 (Finite Representation)**.** *A set of desirable gambles $\mathscr{R}$ on $\mathscr{X}^N$ is coherent and exchangeable iff there is some coherent set $\mathscr{S}$ of desirable gambles on $\mathscr{N}^N$ such that*

$$\mathscr{R} = (\mathrm{MuHy}^N)^{-1}(\mathscr{S}), \tag{38}$$

and in that case this $\mathscr{S}$ is uniquely determined by

$$\mathscr{S} = \left\{g \in \mathscr{G}(\mathscr{N}^N)\colon\ \mathrm{T}^N(g) \in \mathscr{R}\right\} = \mathrm{MuHy}^N(\mathscr{R}). \tag{39}$$

Corollary 15. *A lower prevision $\underline{P}$ on $\mathscr{G}(\mathscr{X}^N)$ is coherent and exchangeable iff there is some coherent lower prevision $\underline{Q}$ on $\mathscr{G}(\mathscr{N}^N)$ such that $\underline{P} = \underline{Q} \circ \mathrm{MuHy}^N$. In that case $\underline{Q}$ is uniquely determined by $\underline{Q} = \underline{P} \circ \mathrm{T}^N$.*

We call the set $\mathscr{S}$ and the lower prevision $\underline{Q}$ the *count representations* of the exchangeable set $\mathscr{R}$ and the exchangeable lower prevision $\underline{P}$, respectively. Our Finite Representation Theorem allows us to give an appealing geometrical interpretation to the notions of exchangeability and representation. The exchangeability of $\mathscr{R}$ means that it is completely determined by its count representation $\mathrm{MuHy}^N(\mathscr{R})$, or what amounts to the same thing since T^N is a linear isomorphism: by its projection $\mathrm{ex}^N(\mathscr{R})$ on the linear space of all permutation invariant gambles. This turns count vectors into useful sufficient statistics (compare with Proposition 13), because the dimension of $\mathscr{G}(\mathscr{N}^N)$ is typically much smaller than that of $\mathscr{G}(\mathscr{X}^N)$.

3.6 Exchangeable natural extension and representation

The exchangeable natural extension is easy to calculate using natural extension in terms of count representations, and the following simple result therefore has important consequences for practical implementations of reasoning and inference under exchangeability.

Theorem 16. *Let $\mathscr{A}$ be a set of gambles on $\mathscr{X}^N$, then*

(i) *$\mathscr{A}$ avoids non-positivity under exchangeability iff $\mathrm{MuHy}^N(\mathscr{A})$ avoids non-positivity.*
(ii) $\mathrm{MuHy}^N(\mathscr{E}^N_{\mathrm{ex}}(\mathscr{A})) = \mathscr{E}(\mathrm{MuHy}^N(\mathscr{A}))$.

3.7 Updating and representation

Suppose, as in Section 3.4, that we update a coherent and exchangeable set of desirable gambles $\mathscr{R}$ after observing a sample $\check{x}$ with count vector $\check{m}$. This leads to an updated coherent and exchangeable set of desirable gambles $\mathscr{R}\rfloor\check{x} = \mathscr{R}\rfloor\check{m}$ on $\mathscr{X}^{\hat{n}}$. Here, we take a closer look at the corresponding set of desirable gambles on $\mathscr{N}^{\hat{n}}$, which we denote (symbolically) by $\mathscr{S}\rfloor\check{m}$ (but we do not want to suggest with this notation that this is in some way an updated set of gambles!). The Finite Representation Theorem 14 tells us that $\mathscr{S}\rfloor\check{m} = \mathrm{MuHy}^{\hat{n}}(\mathscr{R}\rfloor\check{m})$, but is there a direct way to infer the count representation $\mathscr{S}\rfloor\check{m}$ of $\mathscr{R}\rfloor\check{m}$ from the count representation $\mathscr{S} = \mathrm{MuHy}^N(\mathscr{R})$ of $\mathscr{R}$?

To show that there is, we need to introduce two new notions: the *likelihood function*

$$L_{\check{m}}\colon \mathscr{N}^{\hat{n}} \to \mathbb{R}\colon \hat{m} \mapsto L_{\check{m}}(\hat{m}) := \frac{|[\check{m}]|\,|[\hat{m}]|}{|[\check{m}+\hat{m}]|}, \tag{40}$$

associated with sampling without replacement, and the linear map $+_{\check{m}}$ from the linear space $\mathscr{G}(\mathscr{N}^{\hat{n}})$ to the linear space $\mathscr{G}(\mathscr{N}^N)$ given by

$$+_{\check{m}}\colon \mathscr{G}(\mathscr{N}^{\hat{n}}) \to \mathscr{G}(\mathscr{N}^N)\colon\ g \mapsto +_{\check{m}} g \tag{41}$$

where

$$+_{\check{m}}\, g(M) = \begin{cases} g(M-\check{m}) & \text{if } M \geq \check{m} \\ 0 & \text{otherwise.} \end{cases} \tag{42}$$

Proposition 17. *Consider a coherent and exchangeable set of desirable gambles $\mathscr{R}$ on $\mathscr{X}^N$, with count representation $\mathscr{S}$. Let $\mathscr{S}\rfloor\check{m}$ be the count representation of the coherent and exchangeable set of desirable gambles $\mathscr{R}\rfloor\check{m}$, obtained after updating $\mathscr{R}$ with a sample $\check{x}$ with count vector $\check{m}$. Then*

$$\mathscr{S}\rfloor\check{m} = \left\{g \in \mathscr{G}(\mathscr{N}^{\hat{n}})\colon\ +_{\check{m}}(L_{\check{m}} g) \in \mathscr{S}\right\}. \tag{43}$$

4 Conclusions

We have shown that modelling an exchangeability assessment using sets of desirable gambles is not only possible, but also elegant.

Our results indicate that, using sets of desirable gambles, it is conceptually easy to reason about exchangeable sequences. Calculating the natural extension and updating are but simple geometrical operations: taking unions, sums and conical hulls and taking intersections, respectively. This approach has the added advantage that the exchangeability assessment is preserved under updating, also when the conditioning event has lower probability zero, which does not hold when using (lower) previsions (although this might be remedied by using full conditional measures).

Moreover, using our Finite Representation Theorem, reasoning about exchangeable sequences can be reduced to reasoning about count vectors. Working with this representation automatically guarantees that exchangeability is satisfied. The representation for the natural extension and for updated models can be derived directly from the representation of the original model, without having to go back to the (more complex) world of sequences. We have also looked at the problem of representation for infinite sequences, but will report this elsewhere.

The conceptual techniques employed in this paper are not restricted in use to a treatment of exchangeability. They could be applied to other structural assessments, e.g., invariance assessments, as long as this assessment allows us to identify a characterising set of weakly desirable gambles that is sufficiently well-behaved (cf. the first paragraph of Sec. 3.2). This idea was briefly taken up by one of us in another paper [1], but clearly merits further attention.

Thinking in even broader terms, we feel that using sets of desirable gambles can provide a refreshing and fruitful approach to many problems in uncertainty modelling, not only those related to structural assessments.

References

[1] G. de Cooman and E. Miranda. Symmetry of models versus models of symmetry. In W. L. Harper and G. R. Wheeler, editors, *Probability and Inference: Essays in Honor of Henry E. Kyburg, Jr.*, pages 67–149. King's College Publications, 2007.

[2] G. de Cooman, E. Quaeghebeur, and E. Miranda. Representing and assessing exchangeable lower previsions. In *Bulletin of the International Statistical Institute 56th Session – Proceedings*, number 1556, Lisboa, 2007. URL `http://hdl.handle.net/1854/8320`.

[3] G. de Cooman, E. Quaeghebeur, and E. Miranda. Exchangeable lower previsions. *Bernoulli*, 2009. Accepted for publication.

[4] B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68, 1937. English translation in [11].

[5] B. de Finetti. *Teoria delle Probabilità*. Einaudi, Turin, 1970.

[6] B. de Finetti. *Theory of Probability: A Critical Introductory Treatment*. John Wiley & Sons, Chichester, 1974–1975. English translation of [5], two volumes.

[7] P. Diaconis and D. Freedman. Finite exchangeable sequences. *The Annals of Probability*, 8:745–764, 1980.

[8] D. C. Heath and W. D. Sudderth. De Finetti's theorem on exchangeable variables. *The American Statistician*, 30:188–189, 1976.

[9] E. Hewitt and L. J. Savage. Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, 80:470–501, 1955.

[10] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Discrete Multivariate Distributions*. Wiley Series in Probability and Statistics. John Wiley and Sons, New York, 1997.

[11] H. E. Kyburg, Jr. and H. E. Smokler, editors. *Studies in Subjective Probability*. Wiley, New York, 1964. Second edition (with new material) 1980.

[12] S. Moral. Epistemic irrelevance on sets of desirable gambles. In Gert de Cooman, Terrence L. Fine, and Teddy Seidenfeld, editors, *ISIPTA '01 – Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, pages 247–254. Shaker Publishing, Maastricht, 2000.

[13] S. Moral. Epistemic irrelevance on sets of desirable gambles. *Annals of Mathematics and Artificial Intelligence*, 45:197–214, 2005.

[14] E. Quaeghebeur. *Learning from samples using coherent lower previsions*. PhD thesis, Ghent University, 2009.

[15] C. A. B. Smith. Consistency in statistical inference and decision. *Journal of the Royal Statistical Society, Series A*, 23:1–37, 1961.

[16] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[17] P. Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24:125–148, 2000.

[18] P. M. Williams. Notes on conditional previsions. Technical report, School of Mathematical and Physical Science, University of Sussex, UK, 1975. Revised journal version: [21].

[19] P. M. Williams. Coherence, strict coherence and zero probabilities. In *Proceedings of the Fifth International Congress on Logic, Methodology and Philosophy of Science*, volume VI, pages 29–33. Dordrecht, 1975. Proceedings of a 1974 conference held in Warsaw.

[20] P. M. Williams. Indeterminate probabilities. In M. Przelecki, K. Szaniawski, and R. Wojcicki, editors, *Formal Methods in the Methodology of Empirical Sciences*, pages 229–246. Reidel, Dordrecht, 1976. Proceedings of a 1974 conference held in Warsaw.

[21] P. M. Williams. Notes on conditional previsions. *International Journal of Approximate Reasoning*, 44: 366–383, 2007. Revised journal version of [18].

A Proofs

We provide proofs for the more involved results.

Proof of Proposition 4. Since it follows from Theorem 3 that $\underline{P}_{\mathscr{R}}(f - \underline{P}_{\mathscr{R}}(f)) = \underline{P}_{\mathscr{R}}(f) - \underline{P}_{\mathscr{R}}(f) = 0$ for all gambles f, it follows that $\mathscr{M}_{\mathscr{R}} \subseteq \{f \in \mathscr{G}(\Omega) : \underline{P}_{\mathscr{R}}(f) = 0\}$. For the converse inequality, assume that $\underline{P}_{\mathscr{R}}(f) = 0$ holds; then $f = f - \underline{P}_{\mathscr{R}}(f) \in \mathscr{M}_{\mathscr{R}}$.

This also means that $\underline{P}_{\mathscr{R}}(g) = 0$ iff $g \in \mathscr{M}_{\mathscr{R}}$, so for every gamble f we can write:

$$\underline{P}_{\mathscr{M}_{\mathscr{R}}}(f) = \sup\{\mu \in \mathbb{R} : f - \mu \in \mathscr{M}_{\mathscr{R}}\} \quad (44)$$
$$= \sup\{\mu \in \mathbb{R} : \underline{P}_{\mathscr{R}}(f - \mu) = 0\} \quad (45)$$
$$= \sup\{\mu \in \mathbb{R} : \mu = \underline{P}_{\mathscr{R}}(f)\} = \underline{P}_{\mathscr{R}}(f), \quad (46)$$

which proves the equality of $\underline{P}_{\mathscr{M}_{\mathscr{R}}}$ and $\underline{P}_{\mathscr{R}}$. □

Proof of Proposition 5. We need to prove that D1–D4 hold for $\mathscr{R}|B$. For D1, consider $f \in \mathscr{G}(\Omega)|B$ and assume that $f = 0$. Then by coherence $f \notin \mathscr{R}$ and hence $f \notin \mathscr{R}|B$. For D2, consider $f \in \mathscr{G}(\Omega)|B$ and assume that $f > 0$. Then by coherence $f \in \mathscr{R}$ and hence $f \in \mathscr{R}|B$. The proof for D3 is similar to the one for D4. For D4, consider $f_1, f_2 \in \mathscr{R}|B$, then on the one hand $f_1, f_2 \in \mathscr{R}$ and therefore $f_1 + f_2 \in \mathscr{R}$ by coherence; and on the other hand $f_1, f_2 \in \mathscr{G}(\Omega)|B$ and therefore $f_1 + f_2 = I_B f_1 + I_B f_2 = I_B(f_1 + f_2)$, so $f_1 + f_2 \in \mathscr{G}(\Omega)|B$ and hence $f_1 + f_2 \in \mathscr{R}|B$. □

Proof of the equivalences in Definition 3. That (i)⇔(ii) and (iii)⇔(iv) is an immediate consequence of the definition of weak desirability. We continue to show that (i)⇔(iii). For the '⇒' part, observe that $f - \mathrm{ex}^N(f) = \frac{1}{N!}\sum_{\pi \in \mathscr{P}_N}[f - \pi^t f] \in \mathscr{D}_{\mathscr{R}}$, since $\mathscr{D}_{\mathscr{R}}$ is a convex cone by Proposition 2. For the '⇐' part, consider any $f \in \mathscr{G}(\mathscr{X}^N)$ and $\pi \in \mathscr{P}_N$. Consider any $f' \in \mathscr{R}$. Then by assumption both $f - \mathrm{ex}^N(f) + f'/2$ and $\pi^t(-f) - \mathrm{ex}^N(\pi^t(-f)) + f'/2$ belong to $\mathscr{R}$. Hence, because $\mathscr{R}$ is closed under addition, their sum $f - \pi^t f + f'$, obtained using Eq. (17), also belongs to $\mathscr{R}$. Hence $f - \pi^t f$ is weakly desirable. □

Proof of Proposition 6. Consider $f \in \mathscr{R}$. Since $\pi^t f - f = (-f) - \pi^t(-f) \in \mathscr{D}_{\mathscr{P}_N}$, we see that $\pi^t f = f + \pi^t f - f \in \mathscr{R} + \mathscr{D}_{\mathscr{P}_N} \subseteq \mathscr{R}$, using the exchangeability condition of Def. 3(ii). □

Proof of Proposition 7. The first statement is a consequence of the second, with $f' = \mathrm{ex}^N(f)$, because then $\mathrm{ex}^N(f') = \mathrm{ex}^N(\mathrm{ex}^N(f)) = \mathrm{ex}^N(f)$. For the second statement, consider arbitrary gambles f and f' on $\mathscr{X}^N$ such that $\mathrm{ex}^N(f) = \mathrm{ex}^N(f')$, and assume that $f \in \mathscr{R}$. We prove that then also $f' \in \mathscr{R}$. Since $\mathrm{ex}^N(f) - f = (-f) - \mathrm{ex}^N(-f) \in \mathscr{D}_{\mathscr{R}}$ and $f' - \mathrm{ex}^N(f') \in \mathscr{D}_{\mathscr{R}}$, we see that $f' - f \in \mathscr{D}_{\mathscr{R}}$ by WD4, and therefore $f' = f + f' - f \in \mathscr{R} + \mathscr{D}_{\mathscr{R}} \subseteq \mathscr{R}$. □

Proof of Theorem 8. We give a circular proof. We first show that (ii) holds if $\underline{P}$ is exchangeable, i.e., if there is some coherent and exchangeable $\mathscr{R}$ such that $\underline{P} = \underline{P}_{\mathscr{R}}$. We already know from Theorem 3 that $\underline{P} = \underline{P}_{\mathscr{R}}$ satisfies P1–P3, because $\mathscr{R}$ is coherent. Consider any $f \in \mathscr{D}_{\mathscr{P}_N}$. Since $\mathscr{D}_{\mathscr{P}_N} \subseteq \mathscr{D}_{\mathscr{R}}$, it also follows from Theorem 3 that $\underline{P}_{\mathscr{R}}(f) \geq 0$ and similarly $-\overline{P}_{\mathscr{R}}(f) = \underline{P}_{\mathscr{R}}(-f) \geq 0$ because also $-f \in \mathscr{D}_{\mathscr{P}_N}$. Hence indeed $0 \leq \underline{P}_{\mathscr{R}}(f) \leq \overline{P}_{\mathscr{R}}(f) \leq 0$, where the second inequality is a consequence of P1 and P2.

That (ii) implies (iii) follows the super-additivity of $\underline{P}$ and the sub-additivity of $\overline{P}$.

Finally, we show that (iii) implies that $\underline{P}$ is exchangeable. The standard argument in [17, Section 6] tells us that $\mathscr{R}' := \{f \in \mathscr{G}(\mathscr{X}^N) : f > 0 \text{ or } \underline{P}(f) > 0\}$ is a coherent set of desirable gambles such that $\underline{P}_{\mathscr{R}'} = \underline{P}$. Now consider the set $\mathscr{R} := \mathscr{R}' + \mathscr{D}_{\mathscr{U}_N}$. We show that this $\mathscr{R}$ is a coherent and exchangeable set of desirable gambles, and that $\underline{P}_{\mathscr{R}} = \underline{P}$. It is clear from its definition that $\mathscr{R}$ satisfies D2, D3 and D4, so let us assume *ex absurdo* that $0 \in \mathscr{R}$, meaning that there is some $f \in \mathscr{R}'$ such that $f' := -f \in \mathscr{D}_{\mathscr{U}_N}$. There are two possibilities. Either $f > 0$, so $f' < 0$, which contradicts Lemma 18. Or $\underline{P}(f) > 0$. But it follows from the coherence of the lower prevision $\underline{P}$ and the assumption that $0 = \underline{P}(f + f') = \underline{P}(f) > 0$, a contradiction too. So $\mathscr{R}$ satisfies D1 as well, and is therefore coherent. It is obvious that $\mathscr{R}$ is exchangeable: $\mathscr{R} + \mathscr{D}_{\mathscr{U}_N} = \mathscr{R}' + \mathscr{D}_{\mathscr{U}_N} + \mathscr{D}_{\mathscr{U}_N} = \mathscr{R}' + \mathscr{D}_{\mathscr{U}_N} = \mathscr{R}$. The proof is complete if we can show that $\underline{P} = \underline{P}_{\mathscr{R}}$. Fix any gamble f. Observe that $f - \alpha \in \mathscr{R}$ iff there are $f' \in \mathscr{R}$ and $f'' \in \mathscr{D}_{\mathscr{U}_N}$ such that $f - \alpha = f' + f''$. But then it follows from the coherence of $\underline{P}$ and the assumption that $\underline{P}(f) = \alpha + \underline{P}(f' + f'') = \alpha + \underline{P}(f') \geq \alpha$, and therefore $\underline{P}_{\mathscr{R}}(f) \leq \underline{P}(f) = \underline{P}_{\mathscr{R}'}(f)$. For the converse

inequality, we infer from $0 \in \mathscr{D}_{\mathscr{U}_N}$ that $\mathscr{R}' \subseteq \mathscr{R}$, and therefore $\underline{P}_{\mathscr{R}'} \leq \underline{P}_{\mathscr{R}}$. □

Lemma 18. *For all f in $\mathscr{D}_{\mathscr{U}_N}$, $f \not< 0$.*

Proof. First of all, observe that for any gamble f' on $\mathscr{X}^N$, if $f' > 0$ then also $\mathrm{ex}^N(f') > 0$. Now consider $f \in \mathscr{D}_{\mathscr{U}_N}$ and assume *ex absurdo* that $f < 0$. Then $-f > 0$ and therefore $-\mathrm{ex}^N(f) = \mathrm{ex}^N(-f) > 0$, whence $\mathrm{ex}^N(f) < 0$. But since $f \in \mathscr{D}_{\mathscr{U}_N}$ we also have that $\mathrm{ex}^N(f) = 0$, a contradiction. □

Proof of Proposition 9. For the first statement, we have to prove that $\mathscr{G}_0^+(\mathscr{X}^N) + \mathscr{D}_{\mathscr{U}_N}$ avoids non-positivity. Consider any $f' \in \mathscr{D}_{\mathscr{U}_N}$ and any $f'' \in \mathscr{G}_0^+(\mathscr{X}^N)$, then we have to prove that $f := f' + f'' \not\leq 0$. There are two possibilities. Either $f' = 0$ and then $f = f'' > 0$. Or $f' \neq 0$, and then Lemma 18 tells us that $f' \not< 0$, whence $f' \not\leq 0$ and therefore *a fortiori* $f \not\leq 0$.

For the second statement, it clearly suffices to prove the 'if' part. Assume therefore that $\mathscr{A} + \mathscr{D}_{\mathscr{U}_N}$ avoids non-positivity. Consider any f in $\mathrm{coni}([\mathscr{G}_0^+(\mathscr{X}^N) \cup \mathscr{A}] + \mathscr{D}_{\mathscr{U}_N})$, so there are $n \geq 1$, $\lambda_k \in \mathbb{R}^+$, $f' \in \mathscr{D}_{\mathscr{U}_N}$, $f_k \in \mathscr{G}_0^+(\mathscr{X}^N) \cup \mathscr{A}$ such that $f = f' + \sum_{k=1}^n \lambda_k f_k$. Let $I := \{k \in \{1,\ldots,n\} : f_k > 0\}$ then $f_\ell \in \mathscr{A}$ for all $\ell \notin I$. By assumption $f' + \sum_{\ell \notin I} \lambda_\ell f_\ell \not\leq 0$, and therefore *a fortiori* $f \not\leq 0$. □

Proof of Theorem 10. It is immediately clear from the fact that $\mathbb{D}_{\mathrm{ex}}(\mathscr{X}^N)$ is closed under arbitrary non-empty intersections, the definition of $\mathscr{E}_{\mathrm{ex}}^N(\mathscr{A})$, and the fact that $\mathscr{G}(\mathscr{X}^N)$ is not a coherent set of desirable gambles, that the last four statements are equivalent. We now prove (i)⇔(ii).

First, assume that $\mathscr{A}$, and therefore also $\mathscr{G}_0^+(\mathscr{X}^N) \cup \mathscr{A}$, is included in some coherent and exchangeable set of desirable gambles $\mathscr{R}$. By exchangeability, $[\mathscr{G}_0^+(\mathscr{X}^N) \cup \mathscr{A}] + \mathscr{D}_{\mathscr{U}_N} \subseteq \mathscr{R} + \mathscr{D}_{\mathscr{U}_N} \subseteq \mathscr{R}$. Since $\mathrm{coni}(\mathscr{R}) = \mathscr{R}$ avoids non-positivity, so does any of its subsets, and therefore in particular $[\mathscr{G}_0^+(\mathscr{X}^N) \cup \mathscr{A}] + \mathscr{D}_{\mathscr{U}_N}$. This means that $\mathscr{A}$ indeed avoids non-positivity under exchangeability.

Conversely, assume that $\mathscr{A}$ avoids non-positivity under exchangeability. For the sake of convenience, denote the set on the right-hand side of Eq. (28) by $\mathscr{R}^*$. It is clear that $\mathscr{R}^*$ satisfies D2, D3 and D4. Consider any $f \in \mathscr{R}^*$, then $f \not\leq 0$, precisely because $\mathscr{A}$ avoids non-positivity under exchangeability. Hence $\mathscr{R}^*$ also satisfies D1, and is therefore coherent. To show that $\mathscr{R}^*$ is exchangeable, again consider any $f \in \mathscr{R}^*$, so there are $n \geq 1$, $\lambda_k \in \mathbb{R}^+$, $f' \in \mathscr{D}_{\mathscr{U}_N}$, $f_k \in \mathscr{G}_0^+(\mathscr{X}^N) \cup \mathscr{A}$ such that $f = f' + \sum_{k=1}^n \lambda_k f_k$. Then for any $f'' \in \mathscr{D}_{\mathscr{U}_N}$ we see that $f' + f'' \in \mathscr{D}_{\mathscr{U}_N}$ and therefore indeed $f + f'' = (f' + f'') + \sum_{k=1}^n \lambda_k f_k \in \mathscr{R}^*$.

Since $\mathscr{A} \subseteq \mathscr{R}^*$, the proof of the equivalences is complete. We now turn to the proof of Eq. (28), i.e., we prove that $\mathscr{E}_{\mathrm{ex}}^N(\mathscr{A}) = \mathscr{R}^*$. It is clear that any coherent and exchangeable set of desirable gambles that includes $\mathscr{A}$, must also include $\mathscr{R}^*$, by the axioms D2, D3, and D4. Since we have just proved that $\mathscr{R}^*$ is coherent and exchangeable, it is the smallest coherent and exchangeable set of desirable gambles that includes $\mathscr{A}$. The desired equality now follows because we have assumed that (i) holds, and we have just proved that (i) implies (v).

Eq. (29) follows from Eq. (28) and Theorem 1, since $\mathscr{D}_{\mathscr{U}_N}$ is a cone. □

Proof of Corollary 11. This is an immediate consequence of Proposition 9(i) and Theorem 10. □

Proof of Proposition 12. The coherence of $\mathscr{R}\rfloor\check{x}$ is guaranteed by Proposition 5. We show that $\mathscr{R}\rfloor\check{x}$ is exchangeable. Consider arbitrary $f \in \mathscr{G}(\mathscr{X}^{\hat{n}})$, $\hat{\pi} \in \mathscr{P}_{\hat{n}}$ and $f_1 \in \mathscr{R}\rfloor\check{x}$. Then we must show that $f_1 + f - \hat{\pi}^t f \in \mathscr{R}\rfloor\check{x}$, or in other words that $I_{C_{\check{x}}}[f_1 + f - \hat{\pi}^t f] \in \mathscr{R}$. But since $f_1 \in \mathscr{R}\rfloor\check{x}$, we know that $I_{C_{\check{x}}} f_1 \in \mathscr{R}$. And if we consider the permutation $\pi \in \mathscr{P}_N$ defined by

$$\pi(k) := \begin{cases} k & 1 \leq k \leq \check{n} \\ \check{n} + \hat{\pi}(k - \check{n}) & \check{n}+1 \leq k \leq N, \end{cases} \tag{47}$$

then clearly $I_{C_{\check{x}}} \hat{\pi}^t f = \pi^t(I_{C_{\check{x}}} f)$ and therefore $I_{C_{\check{x}}}[f_1 + f - \hat{\pi}^t f] = I_{C_{\check{x}}} f_1 + I_{C_{\check{x}}} f - \pi^t(I_{C_{\check{x}}} f)$ and this gamble belongs to $\mathscr{R}$ because $\mathscr{R}$ is exchangeable. □

Proof of Proposition 13. Consider $\check{\pi} \in \mathscr{P}_{\check{n}}$ and any gamble f on $\mathscr{X}^{\hat{n}}$. Assume that $I_{C_{\check{x}}} f \in \mathscr{R}$.

We first prove that $I_{C_{\check{\pi}\check{x}}} f \in \mathscr{R}$. Consider the permutation $\pi \in \mathscr{P}_N$ defined by

$$\pi(k) := \begin{cases} \check{\pi}^{-1}(k) & 1 \leq k \leq \check{n} \\ k & \check{n}+1 \leq k \leq N, \end{cases} \tag{48}$$

then clearly $\pi^t(I_{C_{\check{x}}} f) = (I_{C_{\check{x}}} f) \circ \pi = (I_{C_{\check{x}}} \circ \check{\pi}^{-1}) f = I_{C_{\check{\pi}\check{x}}} f$, so it follows from Proposition 6 that indeed $I_{C_{\check{\pi}\check{x}}} f \in \mathscr{R}$. This already implies that $\mathscr{R}\rfloor\check{x} = \mathscr{R}\rfloor\check{\pi}\check{x}$, and therefore also that $\mathscr{R}\rfloor\check{x} = \mathscr{R}\rfloor\check{y}$.

Since $\mathscr{R}$ is coherent, it also follows from $I_{C_{\check{x}}} f \in \mathscr{R}$ and the reasoning above that $I_{C_{\check{m}}} f = \sum_{\check{y} \in [\check{m}]} I_{C_{\check{y}}} f \in \mathscr{R}$, whence $\mathscr{R}\rfloor\check{x} \subseteq \mathscr{R}\rfloor\check{m}$. To prove the converse inequality, assume that $I_{C_{\check{m}}} f \in \mathscr{R}$. We know that $[\check{m}] = \{\check{\pi}\check{x} : \check{\pi} \in \mathscr{P}_{\check{n}}\}$, and therefore for any $\check{y} \in [\check{m}]$ we can pick a $\check{\pi}_{\check{y}} \in \mathscr{P}_{\check{n}}$ such that $\check{\pi}_{\check{y}}\check{x} = \check{y}$. With this $\check{\pi}_{\check{y}}$ we construct a permutation $\pi_{\check{y}} \in \mathscr{P}_N$ in the manner described above, which satisfies $\pi_{\check{y}}^t(I_{C_{\check{x}}} f) = I_{C_{\check{y}}} f$. But then the exchangeability and coherence of $\mathscr{R}$ tell us that

$$\begin{aligned} I_{C_{\check{m}}} f + \sum_{\check{y} \in [\check{m}]} [(I_{C_{\check{x}}} f) - \pi_{\check{y}}^t(I_{C_{\check{x}}} f)] &= I_{C_{\check{m}}} f + f \sum_{\check{y} \in [\check{m}]} [I_{C_{\check{x}}} - I_{C_{\check{y}}}] \\ &= |[\check{m}]| f I_{C_{\check{x}}} \end{aligned} \tag{49}$$

belongs to $\mathscr{R}$, whence also $I_{C_{\check{x}}} f \in \mathscr{R}$, by coherence. □

Proof of Theorem 14. We begin with the sufficiency part. Assume that there is some coherent set $\mathscr{S}$ of desirable gambles on $\mathscr{N}^N$ such that $\mathscr{R} = (\mathrm{MuHy}^N)^{-1}(\mathscr{S})$. We show that $\mathscr{R}$ is coherent and exchangeable, and that $\mathscr{S} = \mathrm{MuHy}^N(\mathscr{R})$.

We first show that $\mathscr{R}$ is coherent. For D1, consider $f \in \mathscr{G}(\mathscr{X}^N)$ with $f = 0$. Then obviously also $\mathrm{MuHy}^N(f) = 0$ and therefore $\mathrm{MuHy}^N(f) \notin \mathscr{S}$. Hence $f \notin \mathscr{R}$. For D2, let $f > 0$. Then obviously also $\mathrm{MuHy}^N(f) > 0$, and therefore $\mathrm{MuHy}^N(f) \in \mathscr{S}$. Hence $f \in \mathscr{R}$. The proof for D3 is similar to the one for D4. For D4, let $f_1, f_2 \in \mathscr{R}$. Then $g_1 := \mathrm{MuHy}^N(f_1) \in \mathscr{S}$ and $g_2 := \mathrm{MuHy}^N(f_2) \in \mathscr{S}$. This implies that $\mathrm{MuHy}^N(f_1 + f_2) = g_1 + g_2 \in \mathscr{S}$, so again $f_1 + f_2 \in \mathscr{R}$.

To show that $\mathscr{R}$ is exchangeable, consider any $f \in \mathscr{R}$ and $f' \in \mathscr{D}_{\mathscr{U}_N}$. We have to show that $f + f' \in \mathscr{R}$. It is clear that $\mathrm{MuHy}^N(f+f') = \mathrm{MuHy}^N(f) + 0 = \mathrm{MuHy}^N(f) \in \mathscr{S}$. Hence $f + f' \in (\mathrm{MuHy}^N)^{-1}(\mathscr{S})$, so indeed $f + f' \in \mathscr{R}$.

We show that $\mathscr{S} = \mathrm{MuHy}^N(\mathscr{R})$. Consider any $g \in \mathscr{G}(\mathscr{N}^N)$, then using Eq. (35), $\mathrm{MuHy}^N(\mathrm{T}^N(g)) = g$. Since by assumption $\mathscr{R} = (\mathrm{MuHy}^N)^{-1}(\mathscr{S})$, we see that

$$g \in \mathscr{S} \Leftrightarrow \mathrm{MuHy}^N(\mathrm{T}^N(g)) \in \mathscr{S} \Leftrightarrow \mathrm{T}^N(g) \in \mathscr{R}. \quad (50)$$

This shows that $\mathscr{S} = \{g \in \mathscr{G}(\mathscr{N}^N)\colon \mathrm{T}^N(g) \in \mathscr{R}\}$. We show that also $\mathscr{S} = \mathrm{MuHy}^N(\mathscr{R})$. Let $g \in \mathscr{S}$, then we have just proved that $\mathrm{T}^N(g) \in \mathscr{R}$, and therefore, using Eq. (35), $g = \mathrm{MuHy}^N(\mathrm{T}^N(g)) \in \mathrm{MuHy}^N(\mathscr{R})$. Conversely, let $g \in \mathrm{MuHy}^N(\mathscr{R})$. Then there is some $f \in \mathscr{R}$ such that $g = \mathrm{MuHy}^N(f)$ and therefore $\mathrm{T}^N(g) = \mathrm{T}^N(\mathrm{MuHy}^N(f)) = \mathrm{ex}^N(f)$, where the last equality follows from Eq. (35). Now Proposition 7 tells us that $\mathrm{ex}^N(f) \in \mathscr{R}$, because $f \in \mathscr{R}$ and $\mathscr{R}$ is exchangeable. Hence $\mathrm{T}^N(g) \in \mathscr{R}$ and therefore $g \in \mathscr{S}$.

Next, we turn to the necessity part. Suppose that $\mathscr{R}$ is coherent and exchangeable. It suffices to prove that $\mathscr{S} := \mathrm{MuHy}^N(\mathscr{R})$ is a coherent set of desirable gambles on $\mathscr{N}^N$, and that Eq. (38) is satisfied for this choice of $\mathscr{S}$.

We begin with the coherence of $\mathrm{MuHy}^N(\mathscr{R})$. For D1, consider $g \in \mathscr{G}(\mathscr{N}^N)$ with $g = 0$. Assume *ex absurdo* that $g \in \mathrm{MuHy}^N(\mathscr{R})$, meaning that there is some $f \in \mathscr{R}$ such that $0 = g = \mathrm{MuHy}^N(f)$, or in other words $f \in \mathscr{D}_{\mathscr{U}_N}$. This is impossible, due to Eq. (25). For D2, let $g \geq 0$. Then obviously also $f := \mathrm{T}^N(g) \geq 0$. Therefore $f \in \mathscr{R}$ and, because of Eq. (35), $g = \mathrm{MuHy}^N(\mathrm{T}^N(g)) = \mathrm{MuHy}^N(f) \in \mathrm{MuHy}^N(\mathscr{R})$. The proof for D3 is similar to the one for D4. For D4, let $g_1, g_2 \in \mathrm{MuHy}^N(\mathscr{R})$, so there are $f_1, f_2 \in \mathscr{R}$ such that $g_1 = \mathrm{MuHy}^N(f_1)$ and $g_2 = \mathrm{MuHy}^N(f_2)$. Then by coherence of $\mathscr{R}$, $f_1 + f_2 \in \mathscr{R}$, and therefore, by linearity of MuHy^N,

$$\begin{aligned} g_1 + g_2 &= \mathrm{MuHy}^N(f_1) + \mathrm{MuHy}^N(f_2) \\ &= \mathrm{MuHy}^N(f_1 + f_2) \in \mathrm{MuHy}^N(\mathscr{R}). \end{aligned} \quad (51)$$

Finally, we show that $\mathscr{R} = (\mathrm{MuHy}^N)^{-1}(\mathrm{MuHy}^N(\mathscr{R}))$. Consider $f \in \mathscr{R}$, then $\mathrm{MuHy}^N(f) \in \mathrm{MuHy}^N(\mathscr{R})$ and therefore $f \in (\mathrm{MuHy}^N)^{-1}(\mathrm{MuHy}^N(\mathscr{R}))$. Conversely, consider f in $(\mathrm{MuHy}^N)^{-1}(\mathrm{MuHy}^N(\mathscr{R}))$. Then $g := \mathrm{MuHy}^N(f) \in \mathrm{MuHy}^N(\mathscr{R})$, so we infer that there is some $f' \in \mathscr{R}$ such that $g = \mathrm{MuHy}^N(f) = \mathrm{MuHy}^N(f')$. Hence $\mathrm{MuHy}^N(f - f') = 0$, so $f - f' \in \mathscr{D}_{\mathscr{U}_N}$ and therefore $f = f' + f - f' \in \mathscr{R} + \mathscr{D}_{\mathscr{U}_N}$. This implies that $f \in \mathscr{R}$, since $\mathscr{R}$ is exchangeable. □

Proof of Corollary 15. This result can be easily proved as an immediate consequence of Theorem 14 and Eq. (4). As an illustration, we give a more direct proof of the necessity part, based on Theorem 8. This theorem, together with Eq. (35), tells us that for any gamble f on $\mathscr{X}^N$, $\underline{P}(f) = \underline{P}\big(\mathrm{ex}^N(f)\big) = \underline{P}\big(\mathrm{T}^N(\mathrm{MuHy}^N(f))\big) = \underline{Q}\big(\mathrm{MuHy}^N(f)\big)$. □

Proof of Theorem 16. We begin with the second statement. Recall that $\mathscr{E}^N_{\mathrm{ex}}(\mathscr{A}) = \mathscr{D}_{\mathscr{U}_N} + \mathscr{E}^N_{\mathrm{ex}}(\mathscr{A})$ from Theorem 10. Since MuHy^N is a linear operator, it commutes with the coni operator, and therefore:

$$\begin{aligned} &\mathrm{MuHy}^N(\mathscr{E}^N_{\mathrm{ex}}(\mathscr{A})) \\ &\quad = \mathrm{MuHy}^N(\mathscr{D}_{\mathscr{U}_N}) + \mathrm{MuHy}^N(\mathscr{E}^N_{\mathrm{ex}}(\mathscr{A})) \\ &\quad = \mathrm{MuHy}^N(\mathscr{E}^N_{\mathrm{ex}}(\mathscr{A})) \\ &\quad = \mathrm{coni}(\mathrm{MuHy}^N(\mathscr{G}^+_0(\mathscr{X}^N) \cup \mathscr{A})) \\ &\quad = \mathrm{coni}(\mathrm{MuHy}^N(\mathscr{G}^+_0(\mathscr{X}^N)) \cup \mathrm{MuHy}^N(\mathscr{A})) \\ &\quad = \mathrm{coni}(\mathscr{G}^+_0(\mathscr{N}^N) \cup \mathrm{MuHy}^N(\mathscr{A})) \\ &\quad = \mathscr{E}(\mathrm{MuHy}^N(\mathscr{A})), \end{aligned}$$

where the second equality follows from $\mathrm{MuHy}^N(\mathscr{D}_{\mathscr{U}_N}) = \{0\}$, the third from Theorem 10, and the last from Theorem 1. The first statement is an immediate consequence of the second and Theorems 1, 10 and 14. □

Proof of Proposition 17. Recall that $g \in \mathscr{S}\rfloor\check{m}$ iff there is some $f \in \mathscr{G}(\mathscr{X}^{\hat{n}})$ such that at the same time $g = \mathrm{MuHy}^{\hat{n}}(f)$ and $I_{C_{[\check{m}]}} f \in \mathscr{R}$, or in other words $\mathrm{MuHy}^N(I_{C_{[\check{m}]}} f) \in \mathscr{S}$. We therefore consider $M \in \mathscr{N}^N$ and observe that

$$\mathrm{MuHy}^N(I_{C_{[\check{m}]}} f | M) = \frac{1}{|[M]|} \sum_{x \in [M]} (I_{C_{[\check{m}]}} f)(x) \quad (52)$$

$$= \frac{1}{|[M]|} \sum_{\substack{\check{x} \in [\check{m}], \hat{x} \in \mathscr{X}^{\hat{n}} \\ (\check{x}, \hat{x}) \in [M]}} f(\hat{x}), \quad (53)$$

so this value is zero unless $M \geq \check{m}$. In that case we can write $M = \check{m} + \hat{m}$, where $\hat{m} := M - \check{m}$ is a count vector in $\mathscr{N}^{\hat{n}}$; so we find that

$$\mathrm{MuHy}^N(I_{C_{[\check{m}]}} f | \check{m} + \hat{m}) = \frac{1}{|[\check{m} + \hat{m}]|} \sum_{\check{x} \in [\check{m}], \hat{x} \in [\hat{m}]} f(\hat{x}) \quad (54)$$

$$= \frac{|[\check{m}]|\,|[\hat{m}]|}{|[\check{m} + \hat{m}]|} \mathrm{MuHy}^{\hat{n}}(f | \hat{m}). \quad (55)$$

Hence indeed $g \in \mathscr{S}\rfloor\check{m}$ iff $+_{\check{m}}(L_{\check{m}} g) \in \mathscr{S}$. □

Representing and Solving Factored Markov Decision Processes with Imprecise Probabilities

Karina Valdivia Delgado
Instituto de Matemática e Estatística
Universidade de São Paulo
São Paulo – Brazil
kvd@ime.usp.br

Leliane Nunes de Barros
Instituto de Matemática e Estatística
Universidade de São Paulo
São Paulo – Brazil
leliane@ime.usp.br

Fabio Gagliardi Cozman
Escola Politécnica
Universidade de São Paulo
São Paulo – Brazil
fgcozman@usp.br

Ricardo Shirota
Escola Politécnica
Universidade de São Paulo
São Paulo – Brazil
ricardo.shirota@poli.usp.br

Abstract

This paper investigates Factored Markov Decision Processes with Imprecise Probabilities; that is, Markov Decision Processes where transition probabilities are imprecisely specified, and where their specification does not deal directly with states, but rather with factored representations of states. We first define a Factored MDPIP, based on a multilinear formulation for MDPIPs; then we propose a novel algorithm for generation of Γ-maximin policies for Factored MDPIPs. We also developed a representation language for Factored MDPIPs (based on the standard PPDDL language); finally, we describe experiments with a problem of practical significance, the well-known System Administrator Planning problem.

Keywords. Imprecise Markov Decision Processes (MDPIPs), Probabilistic Planning and PPDDL, Knowledge Representation Languages, Multilinear programming.

1 Introduction

Sequential decision making is an essential activity in many domains, ranging from operations research [22] to robotics [29]. The last forty years have seen steady interest in Markov Decision Processes with Imprecise Probabilities (MDPIPs), since the seminal work by Satia and Lave Jr. [24]. Several algorithms have been developed for "flat" representations of MDPIPs, that is, representations that explicitly deal with individual states and transitions between states [13, 28, 34].

In this paper we focus on *factored* representations for MDPIPs. A factored representation deals with state variables that compactly encode a possibly large set of states. Factored versions of Markov Decision Processes (MDPs), where all probabilities are precisely specified, have received considerable attention [3], particularly in connection with large planning problems that arise in artificial intelligence. In fact, the leading representation language for probabilistic planning, PPDDL, is essentially a fragment of first-order logic that can specify Factored MDPs by using predicates to encode states [35]. In our previous work [28], we have briefly discuss Factored MDPIPs as we examined algorithms for "flat" MDPIPs. In the present paper we aim to: (1) give a definition of a factored MDPIP; (2) present an algorithm for policy generation; (3) propose a language for compact specification of factored MDPIP and (4) show some experiments with a well-known practical problem.

In Section 2 we review basic concepts on Factored MDPs. In Section 3, we describe the theory of "flat" MDPIPs and the relevant literature. In Section 4 we define factored representations and the PDL_1 language, a variant on PPDDL. In Section 5 we present an algorithm, which we call FACTOREDMPA (Factored Multilinear programming-based approximation), that produces Γ-minimax policies by resorting to Approximate Nonlinear Programming, and we show the performance of this algorithm in a well-known sequential decision problem described in PDL_1, the System Administrator Planning problem.

2 Markov Decision Processes and their Factored Representations

In this section we review basic facts about MDPs and the high-level representation language PPDDL.

Markov Decision Processes (MDPs) encode possibly infinite sequences of decisions under uncertainty [1, 22]. We are interested in MDPs that consist of (i) a countable set $\mathcal{T}$ of *stages*, such that a decision is made at each stage; (ii) a finite set $\mathcal{S}$ of *states*; (iii) a finite set of *actions* $\mathcal{A}(s)$ for each state s; (iv) a conditional probability distribution P_t that specifies the probability of transition from state s to state s' given action a at stage t, such that probabilities are stationary (do no depend on t) and written $P(s'|s,a)$; (v) a *reward* function R_t that indicates how much is gained (or lost, by using a negative value) when action a is selected in state s at stage t, such that the reward function is stationary and written $R(s,a)$.

The state obtained at stage t is denoted s_t; the action selected at stage t is denoted a_t. The history h_t of an MDP at stage t is the sequence of states and actions visited by the process, $[s_1, a_1, \ldots, a_{t-1}, s_t]$. The *Markov assumption* for MDPs adopts $P(s_t|h_{t-1}, a_t) = P(s_t|s_{t-1}, a_t)$. The main consequence of the Markov condition is that $P(h_t|s_1)$ factorizes as $P(s_t|s_{t-1}, a_{t-1})\, P(s_{t-1}|s_{t-2}, a_{t-2}) \ldots \times P(s_3|s_2, a_2)\, P(s_2|s_1, a_1)$. A *decision rule* $d_t(s,t)$ indicates the action that is to be taken in state s at stage t. A *policy* π is a sequence of decision rules, one for each stage. A policy may be *deterministic* or *randomized*; that is, it may prescribe actions with certainty, or rather it may just prescribe a probability distribution over the actions. A policy may be *history-dependent* or not; that is, it may depend on all states and actions visited in previous stages, or just on the current state. A policy that is not history-dependent is called *Markovian*. A Markovian policy induces a unique probability distribution over histories. Moreover, a Markovian policy needs only specify the prescribed action for each state: $\pi : S \to \mathcal{A}(s)$, where $\pi(s)$ is the action recommended by the policy π for the state s.

To compare different policies we adopt the discounted expected reward with infinite horizon [22]; in this case the solution is given by the *Bellman equation* as follows. First, introduce the concept of *value function* $V_\pi : S \to \mathbb{R}$, that defines the value of state s based on the values of the possible successor states $s' \in S$:

$$V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s' \in S} P(s'|s, \pi(s)) V_\pi(s').$$

The factor γ in this expression is called the *discount* factor of the MDP [22, p. 125].

For MDPs the *optimal value function*, represented by V^*, is the value function associated with any optimal policy. Then, the Bellman equation is [14]:

$$V^*(s) = \max_{a \in \mathcal{A}(s)} \{R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s')\}.$$

The Bellman equation can be also formulated as a linear program [18]:

$$\begin{aligned} \min_{V^*} \quad & : \quad \sum_s V^*(s) \qquad (1) \\ s.t. \quad & : \quad V^*(s) \geq R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s'), \\ & \qquad \forall s \in S, a \in \mathcal{A}(s). \end{aligned}$$

Basically, we force $V^*(s)$ to be greater than or equal to $\max_a \{R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s')\}$; by minimizing $\sum_s V^*(s)$, we obtain the maximum value of the righthand side.

We now consider factored representations of MDPs; that is, MDPs where states are compactly specified using variables/predicates. In a *Factored MDP*, states $\vec{x}$ are represented by a set $\Lambda = \{X_1, X_2, ..., X_n\}$ of variables. Thus, a state $\vec{x} \in S$ is represented as a tuple $\{x_1, x_2, ..., x_n\}$ where x_i is the value of the state variable X_i. Note that the size of S is exponential in the number n of variables.[1]. Recent results have shown that it is possible to solve a Factored MDP with billions of states [12, 3].

In a Factored MDP, the reward function $R(\vec{x}, a)$ can be defined by the sum of local-rewards $R_i(\vec{x}, a)$.

$$R(\vec{x}, a) = \sum_{j=1}^{k_R} R_j(\vec{x}, a). \qquad (2)$$

The scope of each local-reward function R_j is typically restricted to some subset of variables $D_j \subset \Lambda = \{X_1, ..., X_n\}$, defined for each pair $\vec{x} \in S$ and $a \in \mathcal{A}(\vec{x})$.

The next step is to encode the transition probabilities. For each action a we define probabilities using a Dynamic Bayesian Network (DBN); that is, a directed acyclic graph with two layers: one representing the actual state and other representing the next state (Figure 1a). The nodes are denoted by X_i and X_i' for variables in the actual state and next state, respectively. Edges are allowed *from* nodes in the first layer *into* the second layer, and also between nodes in the second layer. We denote by $\mathrm{pa}(X_i')$ the parents of X_i' in the graph. The graph is assumed endowed with the following Markov condition: a variable X_i' is conditionally independent of its nondescendants given its parents. This implies the following factorization of transition probabilities:

$$P(\vec{x}'|\vec{x}, a) = \prod_{i=1}^{n} P(x_i'|\mathrm{pa}(X_i'), a); \qquad (3)$$

[1] Although the complexity of an MDP is P-Complete, i.e, an MDP problem is solved in a polynomial time in the size of the state space, it is exponential in the number of variables [20, 21].

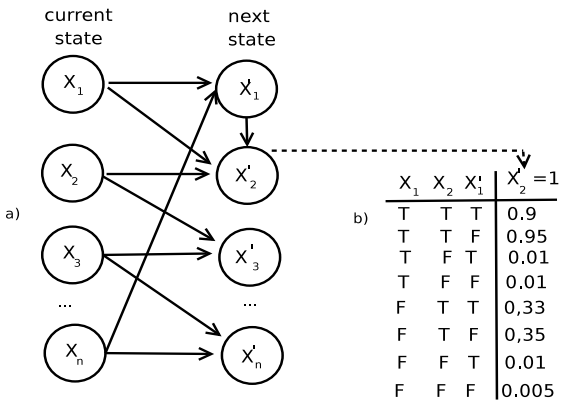


Figure 1: a) A DBN for an action a; b) a conditional probability table for the state variable X_2'.

that is, the probability to go to $\vec{x}' \in S$, given the agent is in state $\vec{x} \in S$ and executes the action $a \in \mathcal{A}(\vec{x})$, is the product of the conditional probability of the agent being in a state where $X_i' = x_i'$ given the parents of X_i' and the action $a \in \mathcal{A}(\vec{x})$ (Figure 1 b). We call P_a the set of conditional probability tables of a DBN for action a.

There are many methods to generate exact and approximate optimal policies for MDPs and Factored MDPs, including value and policy iteration. The technique of Approximate Linear Programming (ALP) [25] has recently been revisited as one of the most promising methods for solving complex Factored MDPs. Refinements for the ALP approach, geared towards Factored MDPs, have been developed over the past few years. The basic idea is to solve an MDP, formulated as Problem (1), by defining a set of basis functions and by using them to construct an approximation of the optimal value function, denoted by $\widehat{V}(\vec{x})$. Basis functions are provided by domain experts or automatically generated [21, 17]. Given $\vec{x} \in S$ and a set of of basis functions $H = \{h_1, ..., h_k\}$, $V^*(\vec{x})$ can be approximated using a linear combination of H:

$$\widehat{V}(\vec{x}) = \sum_{j=1}^{k} w_j h_j(\vec{x}). \tag{4}$$

The quality of the approximation depends on the algorithm used to find $\mathbf{w} = (w_1, ..., w_k)$, such that Equation (4) is a good approximation for $V^*(\vec{x})$. Thus, the ALP formulation of an MDP, given (1), (2) and (4), is the linear program:

$$\min_{w} \quad : \quad \sum_{\vec{x}} \sum_{i=1}^{k} w_i h_i(\vec{x}) \tag{5}$$

$$\text{s.t.} \quad : \quad \sum_{i=1}^{k} w_i h_i(\vec{x}) \geq \sum_{j=1}^{k_R} R_j(\vec{x}, a) + \gamma \sum_{\vec{x}' \in S} P(\vec{x}'|\vec{x}, a) \sum_{i=1}^{k} w_i h_i(\vec{x}'),$$
$$\forall \vec{x} \in S, a \in \mathcal{A}(\vec{x}).$$

The number of variables in the linear program (5) can be smaller than $|S|$, depending on the number of basis functions we have. However, the number of constraints does not change. ALP does not provide computational gains if we do not exploit the factored structure. In Section 5 we will discuss this fact in more detail.

In the last few years, many knowledge representation languages have been proposed for specifying factored MDPs. The most popular such language is Probabilistic Planning Domain Description Language (PPDDL)[35], a language based on first-order logic that has been applied to practical planning problems. In Section 4 we extend PPDDL to factored MDPIPs, and there we present the language in more detail.

3 Markov Decision Processes with Imprecise Probabilities

An MDPIP is simply an MDP where transition probabilities may be imprecisely specified. Note that the term MDPIP was proposed by White III and Eldeib [34], while Satia and Lave Jr. [24] adopt instead the term *MDP with Uncertain Transition Probabilities*.

To specify an MDPIP, one must specify all elements of an MDP except the transition probabilities; now one must specify a *set* of probabilities for each transition between states. We refer to these sets as *transition credal sets*. We assume stationarity for the transition credal sets $K(s'|s, a)$. We also assume that each history h_t is associated with stationary probability distributions $P(s_t|s_{t-1}, a_{t-1})$ that themselves satisfy the Markov condition (and of course $P(s_t|s_{t-1}, a_{t-1}) \in K(s_t|s_{t-1}, a_{t-1})$). That is, our MDPIPs are *elementwise-stationary* [28].

A few definitions are needed. We adopt elementwise conditioning: $K(X|A)$ is obtained from $K(X)$ by conditioning every distribution in the credal set $K(X)$ on the event A. The notation $K(X|Y)$ represents a *set* of credal sets: there is a credal set $K(X|Y = y)$ for each nonempty event $\{Y = y\}$. Given a credal set $K(X)$, we can compute *lower* and *upper* probabilities respectively as $\underline{P}(A) = \inf_{P \in K} P(A)$ and $\overline{P}(A) = \sup_{P \in K} P(A)$. We can also compute *lower* and *upper* expectations for any bounded function $f(X)$ as $\underline{E}[f] = \inf_{P \in K} E[f]$ and $\overline{E}[f] = \sup_{P \in K} E[f]$,

and likewise for conditional lower/upper probabilities/expectations. We assume all credal sets to be closed, so infima and suprema can be replaced by minima and maxima.

There are several criteria of choice for selecting policies in a given MDPIP, even if we fix a single utility and focus on discounted infinite horizon. The Γ-maximin criterion selects a policy that yields the supremum of lower expected reward. In the context of discounted infinite horizon, there is always a deterministic stationary policy that is Γ-maximin [24]; moreover, this policy induces a value function that is the unique solution of

$$V^*(s) = \sup_a \inf_P \left(R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V^*(s') \right), \tag{6}$$

subject to the fact that probabilities must belong to given transition credal sets. Given our assumption that sets of actions are finite and credal sets are closed, we can replace sup and inf respectively by max and min in this expression.

A few other criteria of choice are worth mentioning. The Γ-maximax criterion [24] selects a policy that yields the supremum of upper expected reward [33], while the Γ-maximix criterion selects a policy that yields the maximum of $\alpha(\max_P V_\pi) + (1-\alpha)(\min_P V_\pi)$, for some $\alpha \in (0,1)$. Other criteria seek *sets* of admissible policies, such as the Interval Dominance, Maximality and E-admissible criteria [15]. There are strong foundational reasons to side with the most restrictive of the last three criteria; that is, to adopt E-admissibility [26]. However, in this paper we adopt the Γ-maximin criterion due to its popularity in the existing literature on MDPIPs. We certainly hope to examine other criteria in our future work on Factored MDPIPs.

There are algorithms for solving flat MDPIPs based on dynamic programming [24, 34]. Additional algorithms have been proposed to solve special cases of MDPIPs [10, 31].

4 Defining and representing Factored MDPIPs

We define a *Factored MDPIP*, intuitively enough, as an MDPIP where states are compactly specified using variables/predicates. Thus we have a Factored MDP where transition probabilities are not unique, but rather given by transition credal sets. The challenge then is to specify such transition credal sets in a manner that is itself compact. We suggest that *Dynamic Credal Networks* (DCNs) offer the adequate language to express transition credal sets.

A DCN has the same structure as a DBN (Figure 1), but now each variable X is associated with a set of conditional credal sets; that is a credal set $K(X|\text{pa}(X) = k)$ for each value k of $\text{pa}(X)$. In this paper we assume that every DCN represents a joint credal set over all of its variables, and this joint credal set is exactly the *strong extension* of the credal network [5, 6]. That is, the DCN represents a joint credal set where each distribution satisfies the following expression:

$$P(\vec{x}'|\vec{x},\vec{p},a) = \prod_{i=1}^{n} P(x_i'|\text{pa}(X_i'),\vec{p},a); \tag{7}$$

where each $P(x_i'|\text{pa}(X_i'),\vec{p},a)$ comes from an appropriate credal set associated with the DCN.

Consider the generation of Γ-maximin policies; that is, solution of Equation (6). It does not seem possible to produce a linear programming solution like the linear programming for MDP (Problem 1). However in our previous work [28] we have shown that it is possible to generate solutions using well know programming problems. First, the Equation (6) can be reduced to a *bilevel programming problem*:

$$\begin{aligned}
\min_{V^*} \quad &: \quad \sum_s V^*(s) \qquad (8)\\
\text{s.t.} \quad &: \quad V^*(s) \geq R(s,a) + \\
&\qquad \gamma \sum_{s' \in S} P(s'|s,a) V^*(s'), \forall s \in S, a \in A;\\
&\qquad P \in \arg\min \sum_{s' \in S} P(s'|s,a) V^*(s'),\\
&\qquad \text{s.t.} : P(s'|s,a) \in K_a(s'|s)
\end{aligned}$$

Then, the bilevel problem (8) can be transformed in an equivalent *multilinear programming problem*:

$$\begin{aligned}
\min_{V^*,P} \quad &: \quad \sum_s V^*(s) \qquad (9)\\
\text{s.t.} \quad &: \quad V^*(s) \geq R(s,a) + \\
&\qquad \gamma \sum_{s' \in S} P(s'|s,a) V^*(s'),\\
&\qquad \forall s \in S, a \in \mathcal{A}(s), P(s'|s,a) \in K_a(s'|s,a).
\end{aligned}$$

Note that the solution of multilinear programs is far from trivial, thus our previous solution can only deal with relatively small flat MDPIPs.

We now specialize Problem (9) for Factored MDPIPs. The *factored value function* of a Factored MDPIP is given by Equation (4) restricting the scope of each basis function to some small subset of state variables $C_i \subset \Lambda = \{X_1, ..., X_n\}$. We can use the factored value function (4), the reward function (2), the transition probabilities (7) and replace them in Problem (9) in

order to obtain the factored multilinear programming problem:

$$\begin{aligned} \min_{w,\vec{p}} \quad & : \quad \sum_s \sum_i w_i h_i(\vec{\mathbf{x}}) \qquad (10) \\ s.t. \quad & : \quad \sum_i w_i h_i(\vec{x}) \geq \sum_{j=1}^{kR} R_j(\vec{x}, a) + \\ & \qquad \gamma \sum_{\vec{x}' \in S} P(\vec{x}'|\vec{x}, \vec{p}, a) \sum_i w_i h_i(\vec{x}'), \\ & \quad \forall \vec{x} \in S, a \in \mathcal{A}(\vec{x}). \\ & \quad P(\vec{x}'|\vec{x}, \vec{p}, a) \in K_a(\vec{x}'|\vec{x}) \end{aligned}$$

where:

$$P(\vec{x}'|\vec{x}, \vec{p}, a) = \prod_{i=1}^{n} P(x_i'|\mathrm{pa}(X_i'), \vec{p}, a)$$

This particular nonlinear program will be studied in the next section; the main contribution of this paper is an algorithm for the generation of Γ-maximin policies in Factored MDPIPs that solves Problem (10). Before we plunge into that, we spend the remainder of this section discussing the representation of Factored MDPIPs.

As we have mentioned before, the *Probabilistic Planning Domain Description Language* (PPDDL) [2] is a high-level language for the specification of Factored MDPs, with a relatively simple syntax. Every planning problem is expressed in two parts: the `domain` contains directives, constants, and descriptions of actions; the `problem` basically contains a description of the initial state and the desired goal. We wish to focus on the syntax and semantics of `domain`s, so we present the relevant pieces of the syntax here. The basic BNF for `domain`s is:

```
<domain> ::= (define (domain <NAME>)
    (:requirements :adl)
      [<types>][<constants>][<predicates>]
       <action>*)
<action> ::= (:action <NAME>
       [<param>] [<prec>] [<effect>])
<prec> ::= (:precondition <p-formula>)
<effect> ::= (:effect {<nd-eff>|<det-eff>})
<nd-eff> ::= <prob>|<one-of>
<prob> ::= (probabilistic <p-eff>+)
<p-eff> ::= <RATIONAL> <det-eff>
<one-of> ::= (oneof <det-eff>+),
```

where: `<types>`, `<constants>`, `<predicates>` and `<param>` refer to lists of names or logical variables (possibly typed); `<RATIONAL>` denotes a rational number; `<p-formula>` is a formula containing either atoms, or conjunction of `p-formula`s, or universal quantification over `p-formula`s, or inequality of two given names as `(not (= <NAME> <NAME>))`; and a `<det-eff>` is a formula containing either atoms, or negation of atoms, or conjunction of `det-eff`s, or universal quantification over `det-eff`s, or the *conditional* operator `when`. This conditional operator has syntax `(when <p-formula> <simple-eff>)`, where `simple-eff` is a formula containing either atoms, or negations of atoms, or conjunction of `det-eff`s, or universal quantification over `simple-eff`s.

In PPDDL, a probabilistic action is understood as a probabilistic transition given by a Dynamic Bayesian Network [35]. PPDDL also allows an action to contain `oneof` elements, where a nondeterministic choice is made and one of the effects listed in the scope of the `oneof` element is selected and pursued. There are no probabilities attached to such nondeterministic choices. We call these conventions the *standard semantics* of PPDDL. Note that the standard semantics of PPDDL takes us beyond Markov Decision Processes (MDPs) given the presence of nondeterminism; however the expressivity of PPDDL is still far from general MDPIPs, because in PPDDL each action may contain *either* a probabilistic effect *or* a nondeterministic effect. For instance, a domain may contain two actions, one with probabilistic effects, and the other with nondeterministic effects. What is not allowed in PPDDL is the mixture of probabilistic and nondeterministic effects *in the same action*.

In a previous publication we have explored the facilities of PPDDL to express planning problems where probabilistic and nondeterministic choices (in the PPDDL sense) are mixed, but only allowing that all probabilistic choices precede all nondeterministic choices in an action [30]. The reason for this restriction is that the ensuing planning problems are instances of MDPIPs were all transition credal sets are given by infinitely monotone Choquet capacities. We refer to PPDDL with this added flexibility as PDL_2, and we refer to PPDDL with no restrictions on the combination of probabilistic and nondeterministic choices as PDL_1. We note that PDL_1 can specify Factored MDPIPs with clear syntax; to illustrate this fact, we consider the well-known System Administrator Problem [12].[2]

Example 1 *Consider the problem of optimizing the behavior of a system administrator that works with a network of computers. There are many possible configurations; for example, the cycle network where com-*

[2] This example is actually a variant on the original Factored MDP for the System Administrator problem, because some additive aspects cannot be encoded in PPDDL [23]. The way we solve this limitation was to start with a high reward value and decrease every time the action `reboot` was executed (effect `(decrease (reward) 1)` from Figure 2).

puter i is connected to computer $i+1$. One of these computers is designated as server, while the rest are clients. Each computer is associated to a binary variable X_i, the value of a variable indicates whether the respective machine is up (1) or down (0). At each time step the administrator receives a payment (reward) for each machine that is working. Since the server is the most important computer in the network it is given a greater reward if it is working. The job of the system administrator is deciding which of the machines should reboot. So, there are n+1 possible actions at each step: reboot one of the n machines or not reboot any machine. After executing the action reboot the machine i, the probability of machine i to be working on the next step is high. At each step each computer has a low probability to stop working, which grows dramatically if their neighbors are not working. The machines can begin working spontaneously with a small probability.

In the original PPDDL for the System Administrator Domain [23], the probability of a computer i start working in the next state, given that the action `reboot(i)` was executed, is 0.9; and with the probability 0.1 the state remains unchanged. Also, with probability 0.6 the state variable x_i (computer i) becomes false in the next state, when it is connected with other computer that it is not working (and the computer i has not been rebooted); and with probability 0.4 the state remains unchanged.

Figure 2 presents a PDL_1 specification of a Factored MDPIP that represents the System Administrator Problem, where experts disagree on the probability distributions. With probability between 0.6 and 0.8 the state variable x_i (computer i) becomes false in the next state, if it is connected with other computer that it is not working (and the computer i has not been rebooted). Considering an instance of the domain described in Figure 2 with 3 state variables, which implies 8 states and 3 actions: a_1 for *reboot computer 1*, a_2 for *reboot computer 2* and a_3 for *reboot computer 3*, the corresponding factored MDPIP have the following set of constraints:

$$P(X_i' = 1 | X_1 = 0, a_i) = 0.9$$
$$P(X_i' = 1 | X_1 = 1, a_i) = 1$$

And for $i \neq j$ we have:

$$P(X_i' = 0 | X_{i-1} = 0, X_i = 0, a_j) = 1$$
$$0.6 <= P(X_i' = 0 | X_{i-1} = 0, X_i = 1, a_j) <= 0.8$$
$$P(X_i' = 0 | X_{i-1} = 1, X_i = 0, a_j) = 1$$
$$P(X_i' = 0 | X_{i-1} = 1, X_i = 1, a_j) = 0$$

Instances such these are solved by the algorithm presented in the next section.

```
(define (domain sysadmin)
(:requirements :adl)
(:types comp)
(:predicates (up ?c)(conn ?c ?d))
(:action reboot
  :parameters (?x - comp)
  :effect
     (and (decrease (reward) 1)
          (probabilistic 0.9 (up ?x))
          (oneof
          (forall (?d - comp)
            (probabilistic
              0.6 (when (exists (?c - comp)
                      (and (conn ?c ?d)
                         (not (up ?c))
                         (not (= ?x ?d))))
                  (not (up ?d))
          )))
          (forall (?d - comp)
            (probabilistic
              0.8 (when (exists (?c - comp)
                      (and (conn ?c ?d)
                          (not (up ?c))
                          (not (= ?x ?d))))
                  (not (up ?d))
     )))))

)
(define
 (problem sysadmin-3)
   (:domain sysadmin)
   (:objects x1 - comp x2
                - comp x3
                - comp)
(:init (conn x1 x2)
       (conn x2 x3)
       (conn x3 x1))
(:goal (forall (?c - comp)
               (up ?c)))
(:goal-reward 500)
)
```

Figure 2: The System Administrator domain in PDL_1, with action *reboot* (probabilistic and nondeterministic). This domain defines a Factored MDPIP (adapted from [23]). One limitation of this language is do not allow to express local-reward and basis functions for approximated solutions of MDPs and MDPIPs.

5 FACTOREDMPA: Solving a Factored MDPIP

Koller and Parr [16] show that if we are working with a Factored MDP (Problem 5), a necessary condition to efficiently apply the ALP technique is to restrict the scope of each basis function to some small subset of state variables $C_i \subset \Lambda = \{X_1, ..., X_n\}$ and also to assume small dependency in the DBN[3]. Guestrin et al. [12] then exploited these conditions and developed an efficient algorithm for Factored MDPs. The success of their FACTOREDLPA algorithm is due to: (i) the use of a method to simplify the computation of each constraint of the ALP problem, named *Backprojection* algorithm [16]; and (ii) the *FactoredLP* algorithm that creates a new and smaller set of equivalent constraints for the linear programming problem (5). There are other efficient algorithms that use general techniques to solve linear problems with large number of constraints [21, 8, 9] (e.g., constraints generation), and that somehow, have improved the approach proposed by Guestrin [12].

Based on those ideas, we want to solve a Factored MDPIP formulated as an Approximated Multilinear Programming (Problem 10). First, the same efficient and general techniques that solve linear problems with large number of constraints [21, 8, 9] cannot be applied directly on the multilinear problem. However, the FACTOREDLPA algorithm can be adapted to solve a factored MDPIP as we show in this section. The new algorithm we will name as FACTOREDMPA.

Shortly, FACTOREDMPA first simplifies the computation of each constraint applying the same *Backprojection* algorithm used by Guestrin for factored MDP, then it calls the *FactoredMP* algorithm to create a new and smaller equivalent set of constraints for the Multilinear Programming (Problem 10). Finally, in order to obtain w_i and $\vec{p}$, it calls a nonlinear solver with the new equivalent problem.

5.1 Simplifying the computation of each constraint

We can also take advantage of the fact that the transition model for MDPIPs is factored and the basis functions have scope restricted to a small set of variables in order to efficiently compute the constraints.

From problem (10), given $\vec{x} \in S$ and $a \in \mathcal{A}(\vec{x})$, we have the following constraint:

$$\sum_i w_i h_i(\vec{x}) \geq \sum_{j=1}^{kR} R_j(\vec{x}, a) + \gamma \sum_{\vec{x}' \in S} P(\vec{x}'|\vec{x}, \vec{p}, a) \sum_i w_i h_i(\vec{x}')$$

[3] Although this assumption seems too restrictive, there is a large set of applications that it can be done [11].

Now, we can reorder the sum and obtain:

$$\sum_i w_i h_i(\vec{x}) \geq \sum_{j=1}^{kR} R_j(\vec{x}, a) + \gamma \sum_i w_i \underbrace{\sum_{\vec{x}' \in S} P(\vec{x}'|\vec{x}, \vec{p}, a) h_i(\vec{x}')}$$

Let the underbrace term be renamed as $g_i^a(\vec{x}, \vec{p})$. Note that, for MDPIPs, $g_i^a(\vec{x}, \vec{p})$ is a polynomial expression, i.e. it is described in terms of probability variables and has the following canonical form (with $d_0 = 0$ and d_i a constant):

$$d_0 + \sum_i d_i \prod p_{ij} \tag{11}$$

This term can be precomputed in a efficient way using the *Backprojection* algorithm [16]. For a further computation improvement, the set of constraints can be rewritten as:

$$0 \geq \sum_{j=1}^{kR} R_j(\vec{x}, a) + \sum_i w_i \left(\underbrace{\gamma g_i^a(\vec{x}, \vec{p}) - h_i(\vec{x})} \right)$$

Again, let the underbrace term be renamed as $c_i^a(\vec{x}, \vec{p})$. This term can be precomputed resulting also in the polynomial form (11). Finally $\forall \vec{x} \in S, a \in \mathcal{A}(\vec{x})$ we obtain:

$$0 \geq \sum_{j=1}^{kR} R_j(\vec{x}, a) + \sum_i w_i c_i^a(\vec{x}, \vec{p}) \tag{12}$$

Even with this simplified form to rewrite constraints for the Approximate Multilinear Programming, we are still working with the complete set of constraints ($|S| * |A| + m_2$), where m_2 is the number of constraints related to the probabilities p_{ij}. Since the direct use of general non-linear solvers [19], geometric solvers [4] or multilinear solvers [27] for Problem (10), can only solve problems with small state space, we have to find a way to reduce the number of constraints.

5.2 The *FactoredMP* algorithm

We extend the *FactoredLP* technique proposed by Guestrin [12] in order to obtain a new and smaller equivalent multilinear program for Problem (10). We call this new algorithm *FactoredMP*.

The basic idea is to replace the set of constraints in (12) by an equivalent set of non-linear constraints $\forall a \in \mathcal{A}(\vec{x})$, given by:

$$0 \geq \max_{\vec{x}} \left\{ \sum_{j=1}^{kR} R_j(\vec{x}, a) + \sum_i w_i c_i^a(\vec{x}, \vec{p}) \right\}$$

So, for an action a, we have to compute the following maximization:

$$0 \geq \max_{\vec{x}} \left\{ \sum_{j=1}^{kR} R_j(\vec{x}) + \sum_i w_i c_i(\vec{x}, \vec{p}) \right\} \tag{13}$$

Note that $R_j(\vec{x})$ and $c_i(\vec{x},\vec{p})$ are functions of $\vec{x}$ and we want to do max over $\vec{x}$. Now we can, instead of adding all terms and do the maximization, do the maximization over state variables one by one. To do so we use a modification of the general variable elimination algorithm proposed by Guestrin [12].

For example, if we want to eliminate variable X_1 we do as following. If R_1 is the only local-reward function that depends on X_1 and c_1 is a function that depends on $(X_1,\ X_4)$ and there is no other function c_i that depends on X_1, we can push the maximization over X_1 inwards to obtain:

$$0 \geq \max_{X_2 \ldots X_n} \left\{ \sum_{j=2}^{kR} R_j(\vec{x}) + \sum_{i=2} w_i c_i(\vec{x},\vec{p}) + \max_{X_1} \{R_1(X_1) + w_1 c_1(X_1, X_4, \vec{p})\} \right\}$$

For each variable X_l we want to eliminate, *FactoredMP* selects L relevant functions, renamed as u^{e_1} ... u^{e_L}. A relevant function is the one whose scope contains X_l. We can now replace the maximization over the relevant functions for X_l by the following new function:

$$u_Z^{e_{new}} = \max_{X_l} \sum_{j=1}^{L} u^{e_j} \tag{14}$$

Where Z is the union of all variables in functions u^{e_1} ... u^{e_L} minus X_l. In the above example, the relevant functions are $u_{X_1}^{e_1} = R_1(X_1)$ and $u_{X_1,X_4}^{e_2} = w_1 c_1(X_1, X_4, \vec{p})$. The term $u_Z^{e_{new}}$ is $u_{X_4}^{e_{new}} = \max_{X_1} \left\{ u_{X_1}^{e_1} + u_{X_1,X_4}^{e_2} \right\}$, resulting in the following constraint:

$$0 \geq \max_{X_2 \ldots X_n} \left\{ \sum_{j=2}^{kR} R_j(\vec{x}) + \sum_{i=2} w_i c_i(\vec{x},\vec{p}) + u_{X_4}^{e_{new}} \right\}$$

In order to enforce the definition of $u_Z^{e_{new}}$ as the maximum over X_l (Eq. 14), *FactoredMP* introduces the following set of constraints for any assignment z to Z:

$$u_Z^{e_{new}} \geq \sum_{j=1}^{L} u^{e_j} \forall x_l$$

In the example we need to introduce four constraints: one constraint for each configuration of X_4 and for each configuration of X_1.

This procedure is repeated until all variables have been eliminated. At the end, all the remaining functions u^{e_i} will have empty scope and the following constraint must be added:

$$0 \geq \sum_{j=i} u^{e_i}$$

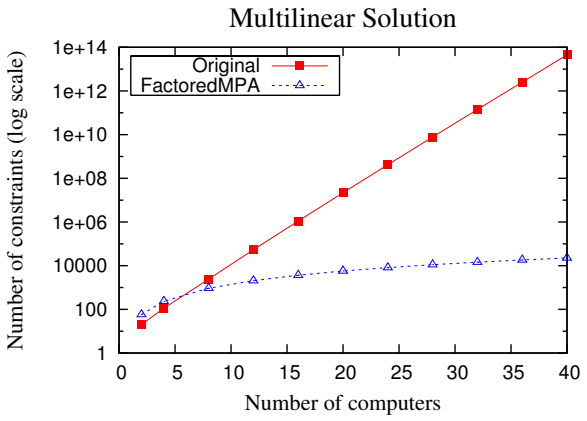


Figure 3: The number of constraints for the System Administrator domain, with imprecise probabilities, for problems with n computers; there are 2^n number of states and $n+1$ actions in this problem.

Notice that, the same process must be applied for all actions $a \in A$. The *FactoredMP* algorithm reduces a structured multilinear programming problem (10) with exponentially many constraints to a new smaller equivalent set of constraints. This property is inherited from the *FactoredLP* procedure.

5.3 Experimental Results

In order to analyze the scalability of the proposed algorithm, we have calculated the original number of constraints and the number of constraints after applying the algorithm FACTOREDMPA for the problem (10). In order to do this, we consider the System Administrator domain (described in the previous section). Figure 3 shows the result for problems varying the number of computers from 2 to 40. The graph shows the original number of constraints grows exponentially while the constraints generated by the FACTOREDMPA algorithm grows quadratically with the number of computers.

We have implemented the FACTOREDMPA algorithm using Matlab as frontend, and MINOS as the nonlinear solver (to handle the reduced multilinear programs). In Figure 4 we show the running times for the System Administrator domain described as in Figure 2 using a simple set of basis functions: the constant function $h_0 = 1$ and $h_i(X_i = 1) = 1$ and $h_i(X_i = 0) = 0$. These results show that with the FACTOREDMPA algorithm it is possible to solve large problems, e.g. we solve in 300 seconds a problem with 40 computers which in the original AMP formulation would have more than $2^{40} * 41$ constraints.

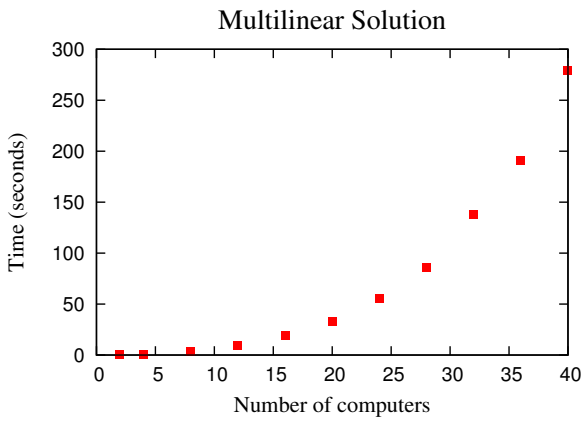


Figure 4: Running time of FACTOREDMPA for the System Administrator Domain Example 1

6 Conclusion

In this paper we have investigated Markov Decision Processes with Imprecise Probabilities, a class of models that adds considerable flexibility and realism to the popular Markov Decision Processes. We have defined a **Factored MDPIP** problem based on a multilinear formulation for MDPIPs [28] and Factored MDPs [12]. We also developed a representation language to specify Factored MDPIPs, named PDL_1, which extends PPDDL by allowing free mixtures of probabilistic and nondeterministic operators. Although PDL_1 does not allow to specify basis and local-reward functions, it is an original and practical high-level language to express factored MDPIPs. Further, we can take advantage of the fact the PPDDL language has largely been used as a benchmark language to solve probabilistic planning problems and, with a simple modification on those problems to obtain a PDL_1 specification, we can have a variety of MDPIP problems.

Our main contribution is a new algorithm, named FACTOREDMPA, to find Γ-maximin policies for Factored MDPIPs. The algorithm is an adaptation of the FACTOREDLPA (Factored Linear Programming-based Approximation) algorithm used to solve Factored MDPs [12, 21]. To evaluate the FACTOREDMPA algorithm, we have modified the System Administrator problem by introducing imprecision in probability values. We thus obtain Factored MDPIPs with varying sizes. Our experiments show that by exploiting the factored representation of a sequential decision problem, and by making the assumption of a restrict scope for variable dependences, relatively large problems can be solved (note that the number of constraints and cpu-time grows quadratically with the number of variables).

Acknowledgements

This work has been supported by FAPESP grant 2008/03995-5; the first author is supported by CAPES; the third author is partially supported by CNPq; and the fourth author is supported by FAPESP. Tests were run in MATLAB and AMPL that calls a multilinear programming solver MINOS.

References

[1] D. P. Bertsekas and J. N. Tsitsiklis. An analysis of stochastic shortest path problems. *Math. Oper. Res.*, 16(3):580–595, 1991.

[2] B. Bonet and R. Givan. International Planning Competition: Non-deterministic track — Call for Participation, December 2005.

[3] C. Boutilier, T. Dean, and S. Hanks. Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11:1–94, 1999.

[4] S. Boyd, S.-J. Kim, L. Vandenberghe, and A. Hassibi. A tutorial on geometric programming, 2004.

[5] F. G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.

[6] F. G. Cozman. Graphical models for imprecise probabilities. *International Journal of Approximate Reasoning*, 39(2-3):167–184, 2005.

[7] G. de Cooman and M. C. M. Troffaes. Dynamic programming for deterministic discrete-time systems with uncertain gain. *International Journal Approximate Reasoning*, 3(2-3):257–278, 2005.

[8] D. P. de Farias and B. Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. *Math. Oper. Res.*, 29(3):462–478, 2004.

[9] D. A. Dolgov and E. H. Durfee. Symmetric approximate linear programming for factored MDPs with application to constrained problems. *Ann. Math. Artif. Intell.*, 47(3-4):273–293, 2006.

[10] R. Givan, S. Leach, and T. Dean. Bounded-parameter Markov decision processes. *Artificial Intelligence*, 122:71–109(39), 2000.

[11] C. Guestrin, D. Koller, R. Parr, and S. Venkataraman. Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research (JAIR)*, 19:399–468, 2003.

[12] C. E. Guestrin. *Planning under Uncertainty in Complex Structured Environments*. PhD thesis, Stanford University, 2003. Adviser-Daphne Koller.

[13] D. Harmanec. Generalizing Markov decision processes to imprecise probabilities. *Journal of Statistical Planning and Inference*, 105:199–213, 2002.

[14] R. A. Howard. *Dynamic Programming and Markov Process*. The MIT Press, 1960.

[15] D. Kikuti, F. G. Cozman, and C. P. de Campos. Partially ordered preferences in decision trees: computing strategies with imprecision in probabilities. In *IJCAI Workshop on Advances in Preference Handling*, pages 118–123, Edinburgh, United Kingdom, 2005.

[16] D. Koller and R. Parr. Computing factored value functions for policies in structured MDPs. In *IJCAI*, pages 1332–1339, 1999.

[17] S. Mahadevan. Samuel meets Amarel: Automating value function approximation using global state space analysis. In *AAAI*, pages 1000–1005, 2005.

[18] A. S. Manne. Linear programming and sequential decision models. In *Management Science*, volume 6(3), pages 259–267, 1960.

[19] B. A. Murtagh, M. A. Saunders, W. Murray, P. E.Gill, R. Raman, and E. Kalvelagen. Minos: A solver for large-scale.

[20] C. Papadimitriou and J. N. Tsitsiklis. The complexity of Markov decision processes. *Math. Oper. Res.*, 12(3):441–450, 1987.

[21] R.-E. Patruscu. *Linear Approximations for Factored Markov Decision Processes*. PhD thesis, University of Waterloo, 2004.

[22] M. L. Puterman. *Markov Decision Processes*. Wiley series in probability and mathematical statistics. John Wiley and Sons, New York, 1994.

[23] S. Sanner. How to spice up your planning under uncertainty research life, 2008. In Workshop on A Reality Check for Planning and Scheduling Under Uncertainty at ICAPS, 2008.

[24] J. K. Satia and R. E. Lave Jr. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21:728–740, 1970.

[25] P.J. Schweitzer and A. Seidmann. Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications*, 110:568–582, 1985.

[26] T. Seidenfeld. A contrast between two decision rules for use with (convex) sets of probabilities: γ-maximin versus E-admissibility. *Synthese*, 140(1-2), 2004.

[27] H. D. Sherali and C. H. Tuncbilek. A global optimization algorithm for polynomial programming problems using a reformulation-linearization technique. *Global Optimization*, 2:101–112, 1992.

[28] R. Shirota, F. G. Cozman, F. W. Trevizan, C. P. de Campos, and L. N. de Barros. Multilinear and integer programming for markov decision processes with imprecise probabilities. In *5th ISIPTA*, pages 395–404, Prague,Czech Republic, 2007.

[29] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2005.

[30] F. W. Trevizan, F. G. Cozman, and L. N. de Barros. Mixed probabilistic and nondeterministic factored planning through Markov decision processes with set-valued transitions. In Workshop on A Reality Check for Planning and Scheduling Under Uncertainty at ICAPS, 2008.

[31] F. W. Trevizan, F. G. Cozman, and L. N. de Barros. Planning under Risk and Knightian Uncertainty. In *Proc. of IJCAI*, pages 2023 – 2028, Hyderabad, India, 01 2007. AAAI.

[32] M. C. M. Troffaes. Learning and optimal control of imprecise Markov decision processes by dynamic programming using the imprecise Dirichlet model. pages 141–148, Berlin, 2004. Springer.

[33] L. V. Utkin and T. Augustin. Powerful algorithms for decision making under partial prior information and general ambiguity attitudes. *ISIPTA*, pp. 349–358, 2005.

[34] C. C. White III and H. K. El-Deib. Markov decision processes with imprecise transition probabilities. *Operations Research*, 42(4):739–749, July-August 1994.

[35] H. L. S. Younes, M. L. Littman, D. Weissman, and J. Asmuth. The first probabilistic track of the international planning competition. *Journal of Artificial Intelligence Research*, 24:851–887, 2005.

The role of generalised p-boxes in imprecise probability models

S. Destercke
Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD)
Ingénierie des Agropolymères et Technologies Emergentes (IATE)
sebastien.destercke@cirad.fr

D. Dubois
Institut de Recherche en Informatique de Toulouse (IRIT)
CNRS & Université de Toulouse
dubois@irit.fr

Abstract

Recently, we have introduced an uncertainty representation generalising imprecise cumulative distributions to any (pre-)ordered space as well as possibility distributions: generalised p-boxes. This representation has many attractive features, as it remains quite simple while having an interesting interpretation in terms of lower and upper confidence bounds over nested sets. However, the merits of this representation in various uncertainty processing tasks still have to be evaluated. This is the topic of this paper, in which the handling of information modelled by generalised p-boxes is studied, from the point of view of elicitation, propagation, conditioning and fusion.

Keywords. Generalized p-boxes, comonotonic clouds, fusion, conditioning, propagation.

1 Introduction

When modelling and processing uncertainty in the presence of imprecision and incompleteness, it is often desirable to use approaches whose complexity remains low rather than full-fledged generic models. The benefits of using the former is that their manipulation is often easier, implying a lower computational cost. They can also be easier to explain to non-experts, thus being useful at the elicitation and post-processing stages. The disadvantage of such simple models is that in some situations they may not be sufficient to represent the available knowledge nor to faithfully address a given problem.

Recently, we have introduced an uncertainty representation generalising imprecise cumulative distributions to any (pre-)ordered space as well as possibility distributions [5]: generalised p-boxes. We showed that this representation is quite simple, can be modeled by random sets and has strong connections with many other recent uncertainty representations such as clouds [16]. The interpretation of generalised p-boxes as collection of nested sets with associated lower and upper confidence bounds makes them promising for uncertainty elicitation. Note that general clouds, of which generalized p-boxes constitute a subfamily, are more complex, hence less attractive, in this respect [5].

However, for a given representation to be useful in uncertainty analysis, one has to study its stability across computations, and their computational complexity. Such a study has already partially been done for generalised p-boxes, whose propagation through a model and use in optimisation procedures under uncertainty have been considered previously [4, 11]. In this paper, we recall some of these previous results and complete this study by investigating other aspects of generalised p-boxes manipulation, such as conditioning or merging. When possible, we link our results with other ones originating from the frameworks of probability sets [18], belief functions [17] and possibility theory [7]. Since generalised p-boxes constitute a subfamily of Neumaier's clouds [5,16], this study also provides some answers to questions regarding the manipulation of these clouds (in particular with respect to their merging).

Section 2 recalls basics about generalised p-boxes and their links with other uncertainty representations and frameworks. In the following sections, we explore the problems of computing probability bounds, information elicitation, propagation, conditioning and merging with generalised p-boxes. We conclude that their main practical interest lies in their simplicity for information representation and elicitation.

2 Preliminaries

Let X be a variable taking its value on a finite space $\mathcal{X}$ having N elements. First recall that two mappings f and f' from a finite indexed set $\mathcal{X} = \{x_1, \dots, x_N\}$ to the real line $\mathbb{R}$ are said to be *comonotonic* if there is a common permutation σ of $\{1, 2, \dots, N\}$ such that $f(x_{\sigma(1)}) \geq f(x_{\sigma(2)}) \geq \dots \geq f(x_{\sigma(N)})$ and $f'(x_{\sigma(1)}) \geq$

$f'(x_{\sigma(2)}) \geq \cdots \geq f'(x_{\sigma(N)})$. In other words, f and f' are comonotonic if and only if for any pair of elements $x, y \in \mathcal{X}$, $f(x) < f(y) \Rightarrow f'(x) \leq f'(y)$ (and $f'(x) < f'(y) \Rightarrow f(x) \leq f(y)$). Note that comonotonicity is not a transitive relation[1]. We consider here that uncertainty about X is modelled by a generalised p-box $[\underline{F}, \overline{F}]$, defined as follows:

Definition 1. *A generalised p-box $[\underline{F}, \overline{F}]$ over a finite space $\mathcal{X}$ is a pair of comonotonic mappings $\underline{F}, \overline{F}$, $\underline{F} : \mathcal{X} \to [0,1]$ and $\overline{F} : \mathcal{X} \to [0,1]$ from $\mathcal{X}$ to $[0,1]$ such that $\underline{F}$ is pointwise less than $\overline{F}$ (i.e. $\underline{F} \leq \overline{F}$) and there is at least one element x in $\mathcal{X}$ for which $\overline{F}(x) = \underline{F}(x) = 1$,*

These limit conditions ensure that $[\underline{F}, \overline{F}]$ characterizes a so-called coherent lower probability. To make notations easier, we introduce an additional element x_0 to $\mathcal{X}$, such that $\overline{F}(x_0) = \underline{F}(x_0) = 0$. As many applications involve variables taking values on the real line $\mathbb{R}$, we also consider generalised p-boxes defined on this space or on one of its product spaces. We limit ourselves to Borel sets and to discrete generalised p-boxes (i.e., when $\overline{F}, \underline{F}$ only takes a finite number of values), other situations being seldom encountered in practice. This allows us, by a proper partition, to come back to the finite space case.

A generalised p-box $[\underline{F}, \overline{F}]$ induces a particular weak order $\leq_{[\underline{F},\overline{F}]}$ between elements of $\mathcal{X}$, such that $x \leq_{[\underline{F},\overline{F}]} y$ iff $\overline{F}(x) \leq \overline{F}(y)$ and $\underline{F}(x) \leq \underline{F}(y)$. In the sequel, for sake of clarity, we assume that each distribution $\overline{F}, \underline{F}$ takes distinct values for each element $x \in \mathcal{X}$, and we consider that these elements are indexed in agreement with the ordering induced by the generalised p-box representing the uncertainty about the value of X. That is, elements $x_1, \ldots, x_N$ are indexed such that $i < j \to \overline{F}(x_i) \leq \overline{F}(x_j)$ and $\underline{F}(x_i) \leq \underline{F}(x_j)$. Given a generalised p-box $[\underline{F}, \overline{F}]$ over $\mathcal{X}$, we define $[\underline{F}, \overline{F}]$-*connected* subsets and $\sqsubseteq_{[\underline{F},\overline{F}]}$-*ordering* as follows:

Definition 2. *Given a generalised p-box $[\underline{F}, \overline{F}]$ over $\mathcal{X}$, a subset $C \subseteq \mathcal{X}$ is called* $[\underline{F}, \overline{F}]$-connected *if it can be expressed as a union of consecutive elements x_k, that is*

$$C = \{x_k \in \mathcal{X} | 0 < i \leq k \leq j \leq N\}.$$

Definition 3. *Let $A = \{x_i, \ldots, x_j\}, B = \{x_{i'}, \ldots, x_{j'}\} \subseteq \mathcal{X}$ be two $[\underline{F}, \overline{F}]$-connected sets. The* $\sqsubseteq_{[\underline{F},\overline{F}]}$-ordering *between these sets if defined as follows*

$$A \sqsubseteq_{[\underline{F},\overline{F}]} B \text{ if and only if } \begin{cases} i \leq i' \\ j \leq j'. \end{cases}$$

When $A \sqsubseteq_{[\underline{F},\overline{F}]} B$ and $B \not\sqsubseteq_{[\underline{F},\overline{F}]} A$, we denote it by $A \sqsubset_{[\underline{F},\overline{F}]} B$.

[1] Otherwise all mappings would be comonotonic, since all mappings are comonotonic with the constant mapping.

2.1 Link with convex sets of probability

Convex sets of probabilities constitute one of the most general uncertainty model available nowadays. Their use has been popularised by Walley [18] and studied by numerous authors (see Miranda [13] for a recent review). In this paper, we will restrict ourselves to sets of probabilities $\mathcal{P}_{\underline{P}}$ induced by lower probabilities. Given a probability set $\mathcal{P}$, its lower probability $\underline{P}$ on an event $A \subseteq \mathcal{X}$ is defined as $\underline{P}(A) = \inf_{P \in \mathcal{P}} P(A)$. Upper probability can be defined similarly (i.e., $\overline{P}(A) = \sup_{P \in \mathcal{P}} P(A)$) and the two measures are dual, in the sense that, for any event $A \subseteq \mathcal{X}$, $\underline{P}(A) = 1 - \overline{P}(A^c)$, where A^c is the complement of A. Then $\mathcal{P}_{\underline{P}} = \{P \geq \underline{P}\}$. The lower probability $\underline{P}$ is called coherent if $\underline{P}(A) = \inf\{P(A), P \in \mathcal{P}_{\underline{P}}\}, \forall A$. The probability set $\mathcal{P}_{\underline{P}}$ is then called a *credal set.*

A generalised p-box $[\underline{F}, \overline{F}]$ induces a particular credal set $\mathcal{P}_{[\underline{F},\overline{F}]}$ such that

$$\mathcal{P}_{[\underline{F},\overline{F}]} = \{P \in \mathbb{P}_{\mathcal{X}} | \underline{F}(x_i) \leq P(\{x_1, \ldots, x_i\}) \leq \overline{F}(x_i)\}$$

with $\mathbb{P}_{\mathcal{X}}$ the set of all probability measures over $\mathcal{X}$. When $\mathcal{X}$ is the real line ($\mathcal{X} = \mathbb{R}$) and when sets A_i are of the type $(-\infty, x_i]$ with $x_i < x_j$ when $i \leq j$, we retrieve classical p-boxes [10].

2.2 Link with random sets

Formally, a random set [2] is a mapping from a probability space to the power set of another space. In the discrete case [17], a random set can also be constructed as a mass assignment $m : \wp(\mathcal{X}) \to [0,1]$ s.t. $\sum_{E \in \wp(\mathcal{X})} m(E) = 1$. In this case, subsets E having a strictly positive mass are called focal sets. We denote the set of focal sets by $\mathcal{F}$, and a random set by $(m, \mathcal{F})$. From a random set, we can define two uncertainty measures, respectively the belief and plausibility functions, which reads, for all $A \subseteq X$:

$$Bel(A) = \sum_{E, E \subseteq A} m(E); \; Pl(A) = \sum_{E, E \cap A \neq \emptyset} m(E).$$

The belief function quantifies our credibility in event A, by summing all the masses that **surely** support A, while the plausibility function measures the maximal confidence that can be given to event A, by summing all masses that **could** support A. They are dual measures, in the sense that for all events A, we have $Bel(A) = 1 - Pl(A^c)$. The belief function can be interpreted as a lower probability, and in this case induces a credal set $\mathcal{P}_{(m,\mathcal{F})} = \{P \in \mathbb{P}_{\mathcal{X}} | P \geq Bel\}$, and $Bel(A) = \underline{P}(A)$, $Pl(A) = \overline{P}(A)$ for any event $A \subseteq \mathcal{X}$.

A generalised p-box $[\underline{F}, \overline{F}]$ also induces a particular random set [5]. This random set can be built by the following procedure: consider a threshold $\theta \in [0,1]$.

When $\underline{F}(x_{i+1}) > \theta \geq \underline{F}(x_i)$ and $\overline{F}(x_{j+1}) > \theta \geq \overline{F}(x_j)$, then, the corresponding focal set is $A_{i+1} \setminus A_j$, with weight

$$m(A_{i+1} \setminus A_j) = \min(\underline{F}(x_{i+1}), \overline{F}(x_{j+1})) - \max(\underline{F}(x_i), \overline{F}(x_j)). \quad (1)$$

This allows to apply results concerning random sets to generalised p-boxes. Let us note $(m, \mathcal{F})_{[\underline{F},\overline{F}]}$ the random set induced by a generalised p-box $[\underline{F}, \overline{F}]$. The focal sets of $(m, \mathcal{F})_{[\underline{F},\overline{F}]}$ have very specific features, which can be summarised as follows:

$[\underline{F}, \overline{F}]$-connectedness: If $A \in \mathcal{F}_{[\underline{F},\overline{F}]}$, then A is $[\underline{F}, \overline{F}]$-connected.

$[\underline{F}, \overline{F}]$-ordered: focal sets are completely ordered with respect to ordering $\sqsubseteq_{[\underline{F},\overline{F}]}$, i.e., for any two distinct sets $A, B \in \mathcal{F}_{[\underline{F},\overline{F}]}$, either $A \sqsubset_{[\underline{F},\overline{F}]} B$ or $B \sqsubset_{[\underline{F},\overline{F}]} A$.

2.3 Link with possibility distributions and clouds

A possibility distribution [7] is a mapping $\pi : \mathcal{X} \to [0, 1]$ from a space $\mathcal{X}$ to the unit interval such that $\pi(x) = 1$ for at least one element x in $\mathcal{X}$. Formally, a possibility distribution is equivalent to the membership function of a fuzzy set. From this possibility distribution are defined two uncertainty measures, respectively the possibility and necessity functions, which reads, for all $A \subset X$:

$$\Pi(A) = \sup_{x \in A} \pi(x); \; N(A) = 1 - \Pi(A^c).$$

Given a possibility distribution π and a degree $\alpha \in [0, 1]$, the strong and regular α-cuts are subsets respectively defined as the sets $E_{\overline{\alpha}} = \{x \in \mathcal{X} | \pi(x) > \alpha\}$ and $E_\alpha = \{x \in \mathcal{X} | \pi(x) \geq \alpha\}$. These α-cuts are nested, since if $\alpha > \beta$, then $E_\alpha \subseteq E_\beta$. In the finite case, a possibility distribution takes at most N values. Let us denote these N values by $\alpha_0 = 0 < \alpha_1 < \ldots < \alpha_N = 1$. We denote the set of probabilities $\mathcal{P}_\pi = \{P \in \mathbb{P}_\mathcal{X} | P \geq N\}$ associated to a possibility distribution π by $\mathcal{P}_\pi$.

Possibility distributions can also be interpreted as particular random sets: they are equivalent to random sets whose focal elements are nested. A belief function (resp. a plausibility function) is a necessity measure (resp a possibility measure) if and only if it derives from a mass function with nested focal sets. Such a random set is called consonant by Shafer [17]. Given a possibility distribution π, the corresponding random set will have the following focal sets E_i with masses $m(E_i)$, $i = 1, \ldots, N$:

$$\begin{cases} E_i = \{x \in X | \pi(x) \geq \alpha_i\} = E_{\alpha_i} \\ m(E_i) = \alpha_i - \alpha_{i-1}. \end{cases} \quad (2)$$

Uncertainty modelled by a generalised p-box $[\underline{F}, \overline{F}]$ can also be modelled by a pair of possibility distributions $\pi_{\overline{F}}, \pi_{\underline{F}}$ such that, for $i = 1, \ldots, N$,

$$\pi_{\overline{F}}(x_i) = \overline{F}(x_i), \quad (3)$$
$$\pi_{\underline{F}}(x_i) = 1 - \underline{F}(x_{i-1}), \quad (4)$$

in the sense that $\mathcal{P}_{[\underline{F},\overline{F}]} = \mathcal{P}_{\pi_{\underline{F}}} \cap \mathcal{P}_{\pi_{\overline{F}}}$. The random sets with mass assignments $m_{\pi_{\overline{F}}}$ and $m_{\pi_{\underline{F}}}$ modeling the uncertainty of distributions $\pi_{\overline{F}}, \pi_{\underline{F}}$ are such that, for $i = 0, \ldots, N-1$,

$$m_{\pi_{\overline{F}}}(A_i^c) = \overline{F}(x_{i+1}) - \overline{F}(x_i)$$
$$m_{\pi_{\underline{F}}}(A_{i+1}) = \underline{F}(x_{i+1}) - \underline{F}(x_i).$$

If we denote the M distinct values taken by $\overline{F}, \underline{F}$ by $0 = \gamma_0 < \gamma_1 < \ldots < \gamma_M = 1$, then the following random set, defined for $j = 1, \ldots, M$ as

$$\begin{cases} E_j = \{x_i \in X | (\pi_{\overline{F}}(x_i) \geq \gamma_j) \wedge (1 - \pi_{\underline{F}}(x_i) < \gamma_j)\}, \\ m(E_j) = \gamma_j - \gamma_{j-1}. \end{cases} \quad (5)$$

is the same as the random set given by Eq. (1).

Due to their links with possibility distributions, generalised p-boxes also have strong connections with clouds, an uncertainty representation introduced by Neumaier [16]. A cloud $[\pi, \delta]$ is a pair of distributions δ, π from $\mathcal{X}$ to $[0, 1]$ such that $\delta \leq \pi$, $\pi(x) = 1$ for at least one $x \in \mathcal{X}$ and $\delta(y) = 0$ for at least one element $y \in \mathcal{X}$. A cloud $[\pi, \delta]$ induces a set of probabilities $\mathcal{P}_{[\pi,\delta]} = \{P \in \mathbb{P}_X | P(\delta_\alpha) \leq 1 - \alpha \leq P(\pi_{\overline{\alpha}})\}$, with $\delta_\alpha = \{x | \delta(x) \geq \alpha\}$ and $\pi_{\overline{\alpha}} = \{x | \pi(x) > \alpha\}$. It can be shown that clouds whose distributions δ, π are comonotonic are equivalent to generalised p-boxes [5], in the sense that they model exactly the same family of probability sets. A so-called comonotonic cloud $[\pi, \delta]$ models the same uncertainty as the generalised p-box $[\underline{F}, \overline{F}]$ for which $\pi_{\overline{F}} = \pi$ and $\pi_{\underline{F}} = 1 - \delta$, and conversely. That is, for any cloud $[\pi, \delta]$, we have $\mathcal{P}_{[\pi,\delta]} = \mathcal{P}_\pi \cap \mathcal{P}_{1-\delta}$, with $\pi, 1 - \delta$ possibility distributions. Using the fact that clouds $[\pi, \delta]$ and $[1 - \delta, 1 - \pi]$ represent the same uncertainty, in the sense that $\mathcal{P}_{[\pi,\delta]} = \mathcal{P}_{[1-\delta,1-\pi]}$, it is immediate that a generalised p-box $[\underline{F}, \overline{F}]$ represents the same uncertainty as the generalised p-box $[\underline{F}_*, \overline{F}_*]$, where, for $i = 1, \ldots, N$

$$\underline{F}_*(x_i) = 1 - \overline{F}(x_{i-1}) \text{ and } \overline{F}_*(x_i) = 1 - \underline{F}(x_{i-1})$$

with the ordering $\leq_{[\underline{F}_*,\overline{F}_*]}$ being the reverse of $\leq_{[\underline{F},\overline{F}]}$.

3 Computing probability bounds

Given a generalised p-box $[\underline{F}, \overline{F}]$, computing lower and upper probabilities over any event $A \subseteq \mathcal{X}$ is

rather easy. First, we consider events forming $[\underline{F}, \overline{F}]$-*connected* sets $C = \{x_k \in \mathcal{X} | 0 < i \leq k \leq j \leq N\}$ where x_i, x_j are respectively the two elements of C with least and greatest index with respect to ordering $\leq_{[\underline{F},\overline{F}]}$. The lower probability of such a set is clearly obtained as [5]

$$\underline{P}(C) = \max\{0, \underline{F}(x_j) - \overline{F}(x_{i-1})\}.$$

Now the focal sets induce, via their intersections, a partition of $\mathcal{X}$. Any subset $E \in \mathcal{X}$ in the Boolean sub-algebra $\mathcal{H}$ induced by this partition is made of a disjoint union of $[\underline{F}, \overline{F}]$-*connected* sets $C_k : E = C_1 \cup \ldots \cup C_M$. Then [5]:

$$\underline{P}_{[\underline{F},\overline{F}]}(E) = \sum_{k=1}^{M} \underline{P}_{[\underline{F},\overline{F}]}(C_k).$$

Now we can compute the lower and upper probabilities of any event $A \subseteq \mathcal{X}$. Namely, let A_* be the lower approximation of A in $\mathcal{H}$ (i.e. the maximal subset $A_* \subseteq A$ in $\mathcal{H}$). It can be proved [5, 14] that $\underline{P}(A) = \underline{P}(A_*)$, hence, if $A_* = C_1 \cup \ldots \cup C_M$ and $C_i = \{x_{\underline{i}}, x_{\underline{i}+1}, \ldots, x_{\overline{i}}\}$, that

$$\underline{P}(A) = \sum_{i=1}^{M} \max\{0, \underline{F}(x_{\overline{i}}) - \overline{F}(x_{\underline{i}-1})\}.$$

Upper probabilities are easily retrieved by duality. In particular, if $C = \{x_k \in \mathcal{X} | 0 < i \leq k \leq j \leq N\}$ is a $[\underline{F}, \overline{F}]$-*connected* subset, then

$$\overline{P}(C) = \overline{F}(x_j) - \underline{F}(x_{i-1}). \tag{6}$$

Note that these bounds always coincide with the lower envelope of $\mathcal{P}_{[\underline{F},\overline{F}]}$, contrary to other conservative bounds using the relations between possibility distributions and ordinary p-boxes [1] or clouds and possibility distributions [16] in their respective computations.

4 Elicitation of generalised p-boxes

To shorten notations, we adopt, from now on, the following notation: for $i = 1, \ldots, N$, let $\alpha_i := \overline{F}(x_i)$ and $\beta_i := \underline{F}(x_i)$ be the lower and upper probability bounds of sets $\{x_1, \ldots, x_i\}$, themselves denoted by A_i. A generalised p-box can then be described as a set of N probabilistic constraints on nested sets

$$i = 1, \ldots, N, \quad \alpha_i \leq P(A_i) \leq \beta_i.$$

Hence, generalised p-boxes can be elicited by asking an expert to provide upper and lower uncertainty bounds over a finite set of nested sets or intervals. There are many situations where asking information under this form is more natural than asking for a set of (imprecise) quantiles, as would be done for ordinary p-boxes. A typical situation is when a parameter or physical quantity θ can be assumed to have an unknown but constant value: in such cases, it sounds natural to ask for confidence bounds around a best estimate $\widehat{\theta}$. Other situations where generalised p-boxes may prove interesting is: (1) when working with categorical variables for which a natural ordering does exist and; (2) when $\theta \in \mathbb{R}^n$ and when sets A_i are convex nested regions of $\mathbb{R}^n$, in which case generalised p-boxes may fit in, while ordinary p-boxes does not.

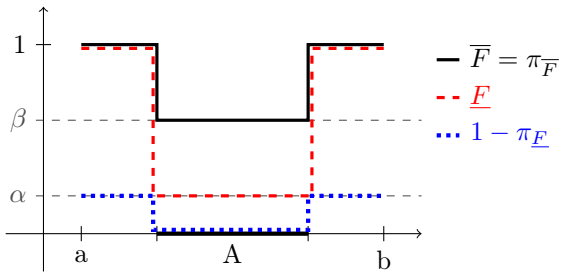


Figure 1: Illustration of $[\underline{F}, \overline{F}]$ and associated cloud $[\pi_{\overline{F}}, 1 - \pi_{\underline{F}}]$ of Example 1

Example 1. *Given a parameter $\theta \in [a, b]$, with $[a, b] \subseteq \mathbb{R}$, an expert provides an interval A as a best guess about the value of θ, together with upper and lower confidence estimates whether θ is in A. This answer (which can be given, for example, as a level on a symbolic scale) is translated into confidence bounds α, β such that $\alpha \leq P(A) \leq \beta$. Define $\mathcal{X}$ as $\{A, [a, b] \setminus A\}$. This information can be translated into a generalised p-box taking values $\overline{F}(x) = \underline{F}(x) = 1$ if $x \in [a, b] \setminus A (= x_2)$ and $\overline{F}(x) = \beta, \underline{F}(x) = \alpha$ if $x \in A (= x_1)$. Note that this is a generalisation of the so-called simple support function [17], where an upper confidence bound (β) is given in addition to a lower one. Figures 1 and 2 provides a graphical illustration of this simple generalised p-box, while its induced random set is such that*

$$m(A) = \alpha;\ m([a, b]) = \beta - \alpha;\ m([a, b] \setminus A) = 1 - \beta.$$

Note that, from a purely practical viewpoint, the cloud $\pi_{\underline{F}}, 1 - \pi_{\overline{F}}$ on figure 2 looks more self-explanatory, at least graphically.

The next example is more complex, illustrating how p-boxes can help in uncertainty elicitation.

Example 2. *Consider an expert having to assess a pH value in a certain situation. His best guess is $pH \in [4.5, 5.5]$. He is not very certain about these bounds and only judges them fairly plausible. He provides another wider interval $[4, 6]$ about which he feels*

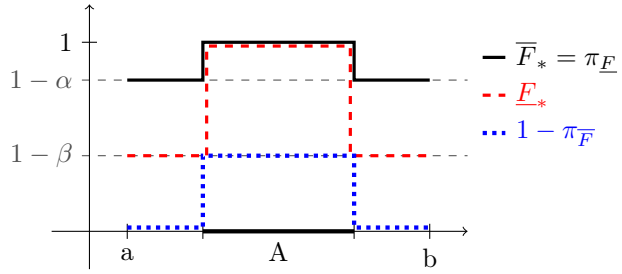


Figure 2: Illustration of , $\underline{F}_*, \overline{F}_*$ and associated cloud $[\pi_{\underline{F}}, 1 - \pi_{\overline{F}}]$ of Example 1

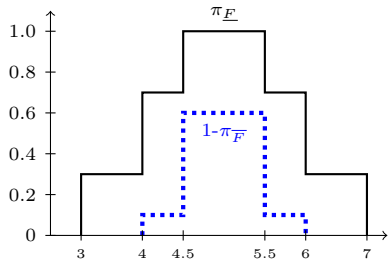


Figure 3: $\pi_{\underline{F}}, 1-\pi_{\overline{F}}$ of generalised p-box of Example 2

more confident. He is however absolutely sure that pH values outside $[3, 7]$ *are impossible. His opinion can be modelled as follows:*

- $0.3 \leq P(pH \in [4.5, 5.5]) \leq 0.6$,
- $0.7 \leq P(pH \in [4, 6]) \leq 0.9$,
- $1 \leq P(pH \in [3, 7]) \leq 1$.

The resulting distributions $\pi_{\underline{F}}, 1 - \pi_{\overline{F}}$ *of this generalised p-box are pictured in Figure 3.*

5 Propagating generalised p-boxes

Let f be a function of variable X such that $f(X) = Y$, with Y a variable taking values on a space $\mathcal{Y}$. Recall that X can be any pre-ordered space (e.g., the discretization of a multi-dimensional continuous space). First recall that the propagation of a random set $(m, \mathcal{F})$, and of its induced set of probabilities $\mathcal{P}_{(m,\mathcal{F})}$, comes down to computing, for every focal set $A \in \mathcal{F}$, the image $f(A)$ and to assigning the same mass to this set as to A in the original random set $(m, \mathcal{F})$. In a previous paper [4], we studied how to propagate a generalised p-box $[\underline{F}, \overline{F}]$ on $\mathcal{X}$, defined by the constraints $\alpha_i \leq P(A_i) \leq \beta_i$ for $i = 1, \ldots, N$, through the model f. We compared three different methods:

- by computing the image of each focal set of $(m, \mathcal{F})_{[\underline{F},\overline{F}]}$, ending up with the random set denoted by $(m, \mathcal{F})_{f((m,\mathcal{F}))}$ and such that to any threshold $\theta \in [0, 1]$ corresponds the focal set

$$\left.\begin{array}{l}\alpha_{i+1} > \theta \geq \alpha_i \\ \beta_{j+1} > \theta \geq \beta_j\end{array}\right\} \begin{array}{l} m(f(A_{i+1} \setminus A_j)) = \\ \min(\alpha_{i+1}, \beta_{j+1}) - \max(\alpha_i, \beta_j); \end{array}$$

- by considering constraints $\alpha_i \leq P(f(A_i)) \leq \beta_i$ on the probabilities of images of sets A_i. Sets $f(A_i)$ being still nested, these constraints again correspond to a generalized p-box, inducing the random set denoted by $(m, \mathcal{F})_{f([\underline{F},\overline{F}])}$ and such that to any threshold $\theta \in [0, 1]$ corresponds the focal set

$$\left.\begin{array}{l}\alpha_{i+1} > \theta \geq \alpha_i \\ \beta_{j+1} > \theta \geq \beta_j\end{array}\right\} \begin{array}{l} m(f(A_{i+1}) \setminus f(A_j)) = \\ \min(\alpha_{i+1}, \beta_{j+1}) - \max(\alpha_i, \beta_j). \end{array}$$

  Note that $f(A_{i+1}) \setminus f(A_j) \subseteq f(A_{i+1} \setminus A_j)$ the former possibly being empty ;

- by separately propagating the focal sets of each possibility distributions $\pi_{\overline{F}}, \pi_{\underline{F}}$, ending up with two propagated random sets $(m, \mathcal{F})_{f(\pi_{\underline{F}})}$ and $(m, \mathcal{F})_{f(\pi_{\overline{F}})}$ which respectively have, for $i = 0, \ldots, N-1$, mass assignments $m(f(A_i^c)) = \beta_{i+1} - \beta_i$ and $m(f(A_{i+1})) = \alpha_{i+1} - \alpha_i$. Rearranging them as in the original generalised p-box, we end up with the random set denoted by $(m, \mathcal{F})_{f(\pi_{\underline{F}},\pi_{\overline{F}})}$ and such that to any threshold $\theta \in [0, 1]$ corresponds the focal set

$$\left.\begin{array}{l}\alpha_{i+1} > \theta \geq \alpha_i \\ \beta_{j+1} > \theta \geq \beta_j\end{array}\right\} \begin{array}{l} m(f(A_{i+1}) \setminus f(A_j^c)^c) = \\ \min(\alpha_{i+1}, \beta_{j+1}) - \max(\alpha_i, \beta_j). \end{array}$$

  Here, $f(A_{i+1} \setminus A_j) \subseteq f(A_{i+1}) \setminus f(A_j^c)^c)$.

If we respectively denote the probability sets, induced by the three propagated random sets, by $\mathcal{P}_{f((m,\mathcal{F}))}, \mathcal{P}_{f(\pi_{\underline{F}},\pi_{\overline{F}})}$, and $\mathcal{P}_{f([\underline{F},\overline{F}])}$, we have the following inclusion relations:

$$\mathcal{P}_{f([\underline{F},\overline{F}])} \subseteq \mathcal{P}_{f((m,\mathcal{F}))} \subseteq \mathcal{P}_{f(\pi_{\underline{F}},\pi_{\overline{F}})},$$

with the inclusions being usually strict. The above relations turn into equalities when f is an injective function, however restricting oneself to such functions is very limiting. When f is not injective, only the second set $\mathcal{P}_{f((m,\mathcal{F}))}$ provides the correct result.

6 Conditioning with generalised p-boxes

Since the lower probability $\underline{P}_{[\underline{F},\overline{F}]}$ induced by a generalised p-box is also a belief function, there are two main ways of conditioning $\underline{P}_{[\underline{F},\overline{F}]}$ when uncertainty on X is modelled by a generalised p-box $[\underline{F}, \overline{F}]$: the first is Dempster's rule of conditioning, while the second

is Walley's rule of conditioning. Both extend classical Bayes conditioning, but correspond to different interpretations of conditioning [8]. In this section, we study whether the conditional uncertainty measures obtained by both conditionings can still be modelled by generalised p-boxes.

6.1 Dempster conditioning

Given a random set $(m, \mathcal{F})$ and a conditioning event $B = \{x_{b_1}, \dots, x_{b_M}\}$, we denote the conditional (plausibility and belief) measures obtained by Dempster conditioning [2]by $\overline{P}_{[B]}, \underline{P}_{[B]}$. These conditional measures, which are still belief and plausibility measures, can be obtained by computing, for each event $A \subseteq \mathcal{X}$

$$\overline{P}_{[B]}(A) = \frac{\overline{P}(A \cap B)}{\overline{P}(B)},$$

where $\overline{P}$ is the plausibility measure of $(m, \mathcal{F})$. Since $\overline{P}_{[B]}$ is a plausibility function, it has positive mass assignment $m_{[B]}$, which can also be built from the initial distribution m, by transferring it to subsets of B, computing for every subset $A \in \mathcal{X}$,

$$m_{[B]}(A) = \begin{cases} \frac{\sum_{C \subseteq \{\mathcal{X} \setminus B\}} m(A \cup C)}{1 - \sum_{A \subseteq B^c} m(A)}, & \text{if } A \subseteq B \\ 0, & \text{otherwise.} \end{cases}$$

This means that the mass $m(A)$ is transferred to $A \cap B$ if $A \cap B \neq \emptyset$, and that the masses given to non-empty sets are then normalised (so that $\sum_{A \subseteq \mathcal{X}} m_{[B]}(A) = 1$). Now, the question is to know whether the upper and lower measures $\overline{P}_{[B]}, \underline{P}_{[B]}$ are still induced by a generalised p-box? The answer is yes, as the following proposition indicates.

Proposition 1. *Let $\underline{P}_{[\underline{F},\overline{F}]}$ be the lower probability induced by a generalised p-box, and B a conditioning event, then the lower measure $\underline{P}_{[B]}$ obtained by Dempster's conditioning stems from a generalised p-box $[\underline{F}, \overline{F}]_{[B]}$ defined on $\mathcal{X} \cap B$ and yielding the restriction of weak order $\leq_{[\underline{F},\overline{F}]}$ of the original p-box to elements $x \in B$.*

Proof. Since $\underline{P}_{[B]}$ is still induced by a random set, it suffices to shows that $m_{[B]}$ remains a mass assignment induced by a generalised p-box, that is that focal sets of $m_{[B]}$ are $[\underline{F}, \overline{F}]$-connected and $[\underline{F}, \overline{F}]$-ordered on B with pre-ordering $\leq_{[\underline{F},\overline{F}]}$.

First, as we consider the weak ordering $\leq_{[\underline{F},\overline{F}]}$ restricted to elements of B, and as any focal set $A = \{x_i, \dots, x_j\}$ is transformed after conditioning to the focal set $A \cap B$, thus retaining all elements in A and B, $A \cap B$ is still $[\underline{F}, \overline{F}]$-connected if the (pre)-ordering is restricted to elements of B.

We have then to show that two distinct focal sets A, A' remain $[\underline{F}, \overline{F}]$-ordered after conditioning on B. Assume $A = \{x_i, \dots, x_j\} \sqsubset_{[\underline{F},\overline{F}]} A' = \{x_k, \dots, x_l\}$, meaning that $i \leq k$ and $j \leq l$, with at least one of the two inequalities strict. Let us consider an element $x_{b_i} \in B$ and the sets $A \setminus x_{b_i}$,$A' \setminus x_{b_i}$. If $x_{b_i} \in A \cap A'$, then $k \leq b_i \leq j$, and $A \setminus x_{b_i} \sqsubset_{[\underline{F},\overline{F}]} A' \setminus x_{b_i}$. If $x_{b_i} \in A \setminus A'$, then $i \leq b_i < k$, and either $A \setminus x_{b_i} = A' \setminus x_{b_i}$ or $A \setminus x_{b_i} \sqsubset_{[\underline{F},\overline{F}]} A' \setminus x_{b_i}$, as A, A' are $[\underline{F}, \overline{F}]$-connected, thus we have $A \setminus x_{b_i} \sqsubseteq_{[\underline{F},\overline{F}]} A' \setminus x_{b_i}$. As we can do it repeatedly for each element $x \in B$, this finishes the proof. □

The above proposition indicates that all the information contained in conditional measures $\overline{P}_{[B]}, \underline{P}_{[B]}$ is captured by a generalised p-box. If $B = \{x_{b_1}, \dots, x_{b_M}\}$, with elements indexed accordingly to $\leq_{[\underline{F},\overline{F}]}$, and if we let $B_i = \{x_{b_1}, \dots, x_{b_i}\}$, then it is sufficient to compute $\overline{P}_{[B]}(B_i), \underline{P}_{[B]}(B_i)$ for $i = 1, \dots, M$ and to consider the induced generalised p-box $[\underline{F}, \overline{F}]_{[B]}$ to model all the conditional information. Let us consider the case (which is the most likely to happen in practice) of conditioning on a $[\underline{F}, \overline{F}]$-connected set $B = \{x_{b_i} | b_1 \leq b_i \leq b_M\}$, then the conditioned generalised p-box is easy to compute, as we have, for $i = 1, \dots, M$ (Using Eq. (6))

$$\begin{aligned} \overline{P}_{[B]}(B_i) &= \frac{\overline{P}(B_i \cap B)}{\overline{P}(B)} = \frac{\overline{P}(\{x_{b_1}, \dots, x_{b_i}\})}{\overline{P}(\{x_{b_1}, \dots, x_{b_M}\})} \\ &= \frac{\overline{F}(x_{b_i}) - \underline{F}(x_{b_1 - 1})}{\overline{F}(x_{b_M}) - \underline{F}(x_{b_1 - 1})} = \overline{F}_{[B]}(x_{b_i}), \\ \underline{P}_{[B]}(B_i) &= \underline{P}_{[B]}(B_i) = 1 - \overline{P}_{[B]}(B_i^c) = 1 - \frac{\overline{P}(B_i^c \cap B)}{\overline{P}(B)} \\ &= 1 - \frac{\overline{P}(\{x_{b_i+1}, \dots, x_{b_M}\})}{\overline{P}(\{x_{b_1}, \dots, x_{b_M}\})} \\ &= \frac{\underline{F}(x_{b_i}) - \underline{F}(x_{b_1 - 1})}{\overline{F}(x_{b_M}) - \underline{F}(x_{b_1 - 1})} = \underline{F}_{[B]}(x_{b_i}). \end{aligned}$$

6.2 Walley's conditioning

Let us now study Walley's conditioning [18]. Given a set of probabilities $\mathcal{P}$, its associated lower and upper probabilities $\underline{P}, \overline{P}$ and a conditioning event B for which $\underline{P}(B) > 0$,[2] we denote the (dual) measures obtained after applying Walley's conditioning by $\underline{P}_{|B}$ and $\overline{P}_{|B}$. For any event $A \subseteq \mathcal{X}$, $\underline{P}_{|B}(A)$ is

$$\underline{P}_{|B}(A) = \inf_{P \in \mathcal{P}} \frac{P(A \cap B)}{P(B)}.$$

[2] We avoid dealing with the case where there are $P \in \mathcal{P}$ such that $P(B) = 0$, which requires more caution (See Miranda [13] for an introduction)

	x_1	x_2	x_3	x_4
$\overline{F}$	0.3	0.5	0.9	1
$\underline{F}$	0.1	0.4	0.7	1

Table 1: Generalised p-box $[\underline{F}, \overline{F}]$ of Example 3

When lower probabilities are belief functions, $\underline{P}_{|B}(A)$ can be computed by the following formula:

$$\underline{P}_{|B}(A) = \frac{\underline{P}(A \cap B)}{\underline{P}(A \cap B) + \overline{P}(A^c \cap B)}.$$

We can then ask ourselves the same question as for Dempster's conditioning: can the information of $\underline{P}_{|B}$, which is known to still be a belief function [12], be totally captured by a generalised p-box? The next example shows that this is not the case.

Example 3. *Consider the space $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$ and the p-box $[\underline{F}, \overline{F}]$ summarized in Table 1. Consider now the conditioning event $B = \{x_1, x_2, x_4\}$. Computing the conditional measure $\underline{P}_{|B}$ for $\{x_1\}$, $\{x_4\}$ and $\{x_1, x_4\}$, we get*

$$\underline{P}_{|B}(\{x_1\}) = 1/8; \underline{P}_{|B}(\{x_4\}) = 1/6; \underline{P}_{|B}(\{x_1, x_4\}) = 2/6.$$

Were $\underline{P}_{|B}$ induced by a generalised p-box, it would satisfy $\underline{P}_{|B}(\{x_1, x_4\}) = \underline{P}_{|B}(\{x_1\}) + \underline{P}_{|B}(\{x_4\})$ (after Section 3), since $\{x_1\}$ and $\{x_4\}$ are disjoint $[\underline{F}, \overline{F}]$-connected sets. It is not the case here, hence $\underline{P}_{|B}$ cannot be modelled by a generalised p-box

This example shows that generalised p-box models are not preserved under Walley's conditioning. However, lower conditional probabilities remain easy to compute. Also, conditional probabilities $\underline{P}_{|B}, \overline{P}_{|B}$ should not be further used in iterated procedures (contrary to $\underline{P}_{[B]}, \overline{P}_{[B]}$). Indeed this type of conditioning is tailored to question-answering of statistical knowledge modelled by credal sets, on the basis of singular information B [8]. If additional singular information C comes up, one has to compute $\underline{P}_{|B\cap C}, \overline{P}_{|B\cap C}$ directly from $\underline{P}, \overline{P}$, therefore the non-preservation of generalised p-boxes under this kind of conditioning is not really an issue.

7 Merging generalised p-boxes

In this section, we assume that S different generalised p-boxes $[\underline{F}, \overline{F}]_1, \ldots, [\underline{F}, \overline{F}]_S$ are available to model our uncertainty about X. They can be provided by different experts, sensors, or any other source of information. In such cases, it is desirable to provide rules to merge uncertain information, possibly taking into account source dependencies. In the following, we say that generalised p-boxes $[\underline{F}, \overline{F}]_1, \ldots, [\underline{F}, \overline{F}]_S$ form a comonotonic set if $\underline{F}_i, \overline{F}_i$, $i = 1, \ldots, S$, are all comonotonic (i.e. all orderings $\leq_{[\underline{F},\overline{F}]_i}$ are the same).

7.1 Idempotent merging rules

When dependencies between sources are not well known, it is usual to use merging rules satisfying the property of idempotence, as this ensures that the merging of two identical information items $[\underline{F}, \overline{F}]_1, [\underline{F}, \overline{F}]_2$ will result in the same representation (thus not adding unwarranted information). Given the strong connections between generalised p-boxes, p-boxes and possibility distributions, it appears natural to define idempotent merging rules as follows:

Conjunction: we define the conjunctively merging $[\underline{F}, \overline{F}]_\cap$ of generalised p-boxes, for any $x \in \mathcal{X}$ as the following pair of mappings

$$\underline{F}_\cap(x) = \max_{i=1,S} \underline{F}_i(x) \text{ and } \overline{F}_\cap(x) = \min_{i=1,S} \overline{F}_i(x). \quad (7)$$

We say that the conjunction is empty when $\underline{F}_\cap(x) > \overline{F}_\cap(x)$ for at least one $x \in \mathcal{X}$

Disjunction: we define the conjunctively merging $[\underline{F}, \overline{F}]_\cup$ of generalised p-boxes as the pair of mappings $\underline{F}_\cap, \overline{F}_\cap$ such that, for any $x \in \mathcal{X}$

$$\underline{F}_\cup(x) = \min_{i=1,S} \underline{F}_i(x) \text{ and } \overline{F}_\cup(x) = \max_{i=1,S} \overline{F}_i(x). \quad (8)$$

Convex combination: Let $\lambda_1, \ldots, \lambda_S$ be non negative weights summing up to one ($\sum_{i=1}^S \lambda_i = 1$) and associated to sources. We then define the arithmetic weighted mean $[\underline{F}, \overline{F}]_\Sigma$ as the pair of mappings $\underline{F}_\Sigma, \overline{F}_\Sigma$ such that, for any $x \in \mathcal{X}$

$$\underline{F}_\Sigma(x) = \sum_{i=1}^S \lambda_i \underline{F}_i(x) \text{ and } \overline{F}_\Sigma(x) = \sum_{i=1}^S \lambda_i \overline{F}_i(x). \quad (9)$$

One can check that, when generalised p-boxes are restricted to ordinary p-boxes, idempotent fusion rules proposed by Ferson's et al. [10] are retrieved. The merging results $[\underline{F}, \overline{F}]_\cup$, $[\underline{F}, \overline{F}]_\Sigma$ and $[\underline{F}, \overline{F}]_\cap$ are not guaranteed to be generalised p-boxes (as comonotonicity can be lost), but they can still be interpreted as clouds (thus offering a possible answer as how to merge clouds [16]). However, when generalised p-boxes form a comonotonic set, the fact that the maximum, minimum and mean operators are non-decreasing in their arguments ensures that the result will still be a generalised p-box with the same induced ordering.

It is also useful to notice that the possibility distribution pairs induced by $[\underline{F}, \overline{F}]_\cup$, $[\underline{F}, \overline{F}]_\cap$ and $[\underline{F}, \overline{F}]_\Sigma$ are such that

- $\pi_{\underline{F}_\cup} = \max_{i=1,S} \pi_{\underline{F}_i}$ and $\pi_{\overline{F}_\cup} = \max_{i=1,S} \pi_{\overline{F}_i}$,
- $\pi_{\underline{F}_\cap} = \min_{i=1,S} \pi_{\underline{F}_i}$ and $\pi_{\overline{F}_\cap} = \min_{i=1,S} \pi_{\overline{F}_i}$,
- $\pi_{\underline{F}_\Sigma} = \sum_{i=1,S} \lambda_i \pi_{\underline{F}_i}$ and $\pi_{\overline{F}_\Sigma} = \sum_{i=1,S} \lambda_i \pi_{\overline{F}_i}$.

The proposed idempotent merging rules are therefore equivalent to applying the classical idempotent rules of possibility theory twice (those rules are retrieved when p-boxes reduce to single possibility distributions).

With regard to sets of probabilities, these merging rules can be used as approximations of exact computations. Let $[\pi_{\overline{F}_\cup}, 1 - \pi_{\underline{F}_\cup}]$, $[\pi_{\overline{F}_\cap}, 1 - \pi_{\underline{F}_\cap}]$ denote the clouds resulting from disjunctions, conjunctions of generalised p-boxes $[\underline{F}, \overline{F}]_1, \ldots, [\underline{F}, \overline{F}]_S$, and $\mathcal{P}_{[\underline{F},\overline{F}]_\cup}, \mathcal{P}_{[\underline{F},\overline{F}]_\cap}$ their induced sets of probabilities (possibly empty). The following proposition holds:

Proposition 2. *Let $\mathcal{P}_{[\underline{F},\overline{F}]_1}, \ldots, \mathcal{P}_{[\underline{F},\overline{F}]_S}$ be the sets of probabilities induced by $[\underline{F}, \overline{F}]_1, \ldots, [\underline{F}, \overline{F}]_S$. Then, the following inclusions hold*

$$\mathcal{P}_{[\underline{F},\overline{F}]_\cap} \subseteq \bigcap_{i=1}^{S} \mathcal{P}_{[\underline{F},\overline{F}]_i},$$
$$\mathcal{P}_{[\underline{F},\overline{F}]_\cup} \supseteq \bigcup_{i=1}^{S} \mathcal{P}_{[\underline{F},\overline{F}]_i},$$

the first inclusion turning into an equality when generalised p-boxes form a comonotonic set.

Proof. First recall that, if π_1, π_2 are two possibility distributions, $\min\{\pi_1, \pi_2\}$, $(\max\{\pi_1, \pi_2\})$ their minimum (maximum) and $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_{\min_{12}}$ $(\mathcal{P}_{\max_{12}})$ their induced sets of probabilities, then $\mathcal{P}_{\min_{12}} \subseteq \mathcal{P}_1 \cap \mathcal{P}_2$ $(\mathcal{P}_1 \cup \mathcal{P}_2 \subseteq \mathcal{P}_{\max_{12}})$.

Using the relation between clouds and sets of probabilities, we have, for conjunction:

$$\mathcal{P}_{[\underline{F},\overline{F}]_\cap} = \mathcal{P}_{\pi_{\overline{F}_\cap}} \cap \mathcal{P}_{\pi_{\underline{F}_\cap}},$$

and since $\pi_{\underline{F}_\cap} = \min_{i=1,S} \pi_{\underline{F}_i}$ and $\pi_{\overline{F}_\cap} = \min_{i=1,S} \pi_{\overline{F}_i}$, we have

$$\bigcap_{i=1}^{S} \left(\mathcal{P}_{\pi_{\overline{F}_i}} \cap \mathcal{P}_{\pi_{\underline{F}_i}} \right) \supseteq \left(\mathcal{P}_{\min_{i=1,S} \pi_{\overline{F}_i}} \cap \mathcal{P}_{\min_{i=1,S} \pi_{\underline{F}_i}} \right)$$

and since $\mathcal{P}_{\pi_{\overline{F}_i}} \cap \mathcal{P}_{\pi_{\underline{F}_i}} = \mathcal{P}_{[\underline{F},\overline{F}]_i}$, this shows the inclusion relation for the conjunction. If we consider now the case where generalised p-boxes form a comonotonic set, then it means that all constraints bear on the same events A_i, $i = 1, \ldots, N$, and are of the kind $\alpha_{i,j} \leq P(A_i) \leq \beta_{i,j}$, where $\alpha_{i,j}, \beta_{i,j}$ are the lower and upper bounds of p-box $[\underline{F}, \overline{F}]_j$ for the set A_i. Thus, the intersection $\cap_{i=1}^{S} \mathcal{P}_{[\underline{F},\overline{F}]_i}$ is induced by the set of following constraints:

$$\max_{i=1,S} \alpha_i \leq P(A_i) \leq \min_{i=1,S} \beta_i,$$

and these constraints exactly describe the generalised p-box $[\underline{F}, \overline{F}]_\cap$. So, $\mathcal{P}_{[\underline{F},\overline{F}]_\cap} = \bigcap_{i=1}^{S} \mathcal{P}_{[\underline{F},\overline{F}]_i}$.

To see the inclusion relation for disjunction, it is sufficient to note that $\cup_{i=1}^{S} (\mathcal{P}_{\pi_{\underline{F}_i}} \cap \mathcal{P}_{\pi_{\overline{F}_i}}) \subseteq (\cup_{i=1}^{S} \mathcal{P}_{\pi_{\underline{F}_i}}) \cap (\cup_{i=1}^{S} \mathcal{P}_{\pi_{\overline{F}_i}})$, for any $i = 1, \ldots, S$. The first probability set is sometimes not convex even in the comonotonic case. □

In particular, Proposition 2 indicates that the conjunction of sets of probabilities induced by ordinary p-boxes or of sets of comonotonic possibility distributions is induced by the result of the proposed merging rule. The conjunctive and disjunctive merging rules can also be interpreted in terms of random sets, as the next proposition indicates. It shows that merging rules can be associated to a random set merging applying a commensuration process [9], with a hypothesis of level-wise merging (i.e. correlation between the sources).

Proposition 3. *Consider the set $\{\gamma_1, \ldots, \gamma_M\} = \bigcup_{i=1}^{S} \{\overline{F}_i(x), \underline{F}_i(x) | x \in \mathcal{X}\}$ of distinct values taken by the generalised p-box $[\underline{F}, \overline{F}]_i$, $i = 1, \ldots, S$, and indexed such that $0 = \gamma_0 < \gamma_1 < \ldots < \gamma_M = 1$. Assume $[\underline{F}, \overline{F}]_\cap$ and $[\underline{F}, \overline{F}]_\cup$ are generalised p-boxes, then they respectively induce the random sets $(m, \mathcal{F})_\cap$,$(m, \mathcal{F})_\cup$ having, for $j = 1, \ldots, M$, the following focal sets:*

$$m_\cap(\cap_{i=1}^{S} E_{i,j}) = \gamma_j - \gamma_{j-1} \tag{10}$$

and

$$m_\cup(\cup_{i=1}^{S} E_{i,j}) = \gamma_j - \gamma_{j-1}, \tag{11}$$

with $E_{i,j} = \{x \in \mathcal{X} | (\pi_{\overline{F}_i}(x) \geq \gamma_j) \wedge (1 - \pi_{\underline{F}_i}(x) < \gamma_j)\}$ the set obtained by Eq. (5) for $[\underline{F}, \overline{F}]_i$.

Proof. Again, we provide only the proof for $[\underline{F}, \overline{F}]_\cup$. If we consider $[\underline{F}, \overline{F}]_\cup$ and the induced pair of possibility distributions $\pi_{\underline{F}_\cup}, \pi_{\overline{F}_\cup}$, the induced random $(m, \mathcal{F})$ have, for $j = 1, \ldots, M$, masses $m(E_j) = \gamma_j - \gamma_{j-1}$ assigned to focal sets such that

$$\begin{aligned}
E_j &= \{x | \pi_{\overline{F}_\cup}(x) \geq \gamma_j \wedge (1 - \pi_{\underline{F}_\cup}(x) < \gamma_j)\} \\
&= \left\{ x | \max_{i=1,S} \pi_{\overline{F}_i}(x) \geq \gamma_j \ \wedge (1 - \max_{i=1,S} \pi_{\underline{F}_i}(x) < \gamma_j) \right\} \\
&= \left\{ x | \max_{i=1,S} \pi_{\overline{F}_i}(x) \geq \gamma_j \right\} \cap \left\{ x | \max_{i=1,S} \pi_{\underline{F}_i} \geq 1 - \gamma_j \right\} \\
&= \cup_{i=1,S} \left\{ x | \pi_{\overline{F}_i}(x) \geq \gamma_j \right\} \cap \cup_{i=1,S} \left\{ x | \pi_{\underline{F}_i} \geq 1 - \gamma_j \right\} \\
&= \bigcup_{i=1,S} \left\{ x | \pi_{\overline{F}_i}(x) \geq \gamma_j \wedge \pi_{\underline{F}_i} \geq 1 - \gamma_j \right\} = \bigcup_{i=1,S} (E_{i,j}).
\end{aligned}$$

This ends the proof. The fourth equality following from known relation between possibilistic disjunction with maximum rule and random sets combination (namely, that the maximum of a set of possibility distributions boils down to computing level-wise unions of their α-cuts [9]). □

Note that when $[\underline{F},\overline{F}]_\cap$ and $[\underline{F},\overline{F}]_\cup$ are only clouds, the random sets in the Proposition are only inner approximations [5]. Finding out a relation between $[\underline{F},\overline{F}]_\Sigma$ and the convex mixture of sets of probabilities (i.e. $\mathcal{P}_\Sigma = \{\sum_{i=1}^S \lambda_i P_i | P_i \in \mathcal{P}_{[\underline{F},\overline{F}]_i}\}$) or of random sets (the two procedure inducing the same set of probabilities) looks harder, except when generalised p-boxes form a comonotonic set, in which case $[\underline{F},\overline{F}]_\Sigma$ can be seen as an approximation of the result that is exact on sets A_i (due to the fact that $\underline{P}_\sigma(A_i) = \sum_{i=1}^S \lambda_i \underline{P}_{[\underline{F},\overline{F}]_i}(A_i) = \sum_{i=1}^S \lambda_i \overline{F}(x_i)$).

8 Other merging rules

In cases where the independence of sources or some dependence structures between them can be assumed, the property of idempotence can be dropped, and it is desirable to use merging rules reflecting the known (in)dependence structure. We are not aware of merging rules exploiting such information in settings emphasising the use of probability sets, but such rules do exist in the settings of possibility theory and of random sets. Exploiting the links between generalised p-boxes and possibility distributions, we can therefore propose an extension of the idempotent merging rules proposed in Section 7.1, such that conjunctive and disjunctive rules respectively become

$$\underline{F}_\top(x) = \bot_{i=1,S}\underline{F}_i(x);\ \overline{F}_\top(x) = \top_{i=1,S}\overline{F}_i(x). \quad (12)$$

$$\underline{F}_\bot(x) = \top_{i=1,S}\underline{F}_i(x);\ \overline{F}_\bot(x) = \bot_{i=1,S}\overline{F}_i(x). \quad (13)$$

with $\top$ a triangular norm and $\bot$ its dual triangular conorm, possibly restricted to associative copulas [15][3] if a probabilistic interpretation is to be preserved. A t-norm is a function $\top : [0,1]^2 \to [0,1]$ that is associative, commutative, non-decreasing in each variable and $\top(x,1) = x$, $\top(x,0) = 0$. The dual t-conorm of a t-norm is such that $\bot(x,y) = 1 - \top(1-x, 1-y)$ for any $(x,y) \in [0,1]^2$. For instance, if all sources can be judged independent, it makes sense to use the product t-norm and its associated t-conorm $\bot(x,y) = x + y - x \cdot y$. Note that this rule is still equivalent to a pair-wise application of the t-norm to possibility distributions $\pi_{\overline{F}_i}, \pi_{\underline{F}_i}$, and that inclusions in Proposition 2 remain valid and are, in this case, always strict.

[3] t-norms $\top$ satisfying $\top(c,d) - \top(c,b) - \top(a,d) + \top(a,b) \geq 0$ for any $(a,b,c,d) \in [0,1]^4$ such that $a \leq c$, $b \leq d$

	A	A^c	$\mathcal{X}$
B	$A \cap B$	$A^c \cap B$	B
B^c	$A \cap B^c$	$A^c \cap B^c$	B^c
$\mathcal{X}$	A	B	$\mathcal{X}$

Table 2: Dempster's rule allocation for Example 4.

As generalised p-boxes constitute particular instances of random sets, it is also possible to merge their induced random sets by families of rules used in this setting [3]. For example, one can apply unnormalised Dempster's rule if sources can be judged independent. Given two random sets with mass assignments m_1, m_2 on $\mathcal{X}$, the random set with mass assignment m_{12} resulting from unnormalised Dempster's rule is such that, for any $A \subseteq \mathcal{X}$,

$$m_{12}(A) = \sum_{\substack{B \cap C = A \\ B,C \subseteq \mathcal{X}}} m_1(B) \cdot m_2(C).$$

The disjunctive rule is obtained by replacing $\cap$ with $\cup$ in the formula. As for possibility distributions [9], applying this rule to random sets induced by a set of generalised p-boxes does not, in general, result in a random set induced by a generalised p-box as the next example indicates.

Example 4. *Let us consider two generalised p-boxes as in Example 1, such that the first source provide bounds α_1, β_1 on set A and the second source provides bounds α_2, β_2 for a distinct set B, such that $B \cap A \neq \{A, B, \emptyset\}$. Table 2 summarises which sets receive a positive mass for the conjunctive allocation.*

Since $A \cap B$, $A \cap B^c$, $A^c \cap B^c$, $A \cap B^c$ are disjoint focal sets strictly included in $\mathcal{X}$, the result is not a generalised p-box, since there are no weak order on elements of $\mathcal{X}$ such that all focal sets are connected *and* ordered. *The same argument holds for the disjunctive counterpart of Dempster's rule.*

9 Summary and Conclusions

This paper suggests that generalised p-boxes are not very stable uncertainty representations, in the sense that most information processing tasks (e.g. propagation, conditioning), once applied to generalised p-boxes, result in representations that are no longer generalised p-boxes. However, even in such situations, using these representations can alleviate the computational burden (e.g., by using quick approximation). There are also specific processing tasks (i.e. propagation through injective functions, dempsterian conditioning, merging of comonotonic sets of generalised p-boxes) where the final result is still a generalised p-box.

Consequently, processing information solely by the means of generalised p-boxes appears of poor interest when one want to make exact computations, as their expressive power remains limited (they can be, however, useful to provide quick approximations). It thus appears that the main interest of generalised p-boxes lies in the elicitation and post-processing stages. Indeed, assigning lower and upper confidence bounds to a set of nested sets is a quite natural way to characterise and to represent information tainted with uncertainty. Recent works on comonotonic clouds [11] also show that generalised p-boxes (which have an expressive power equivalent to comonotonic clouds, as they can model the same sets of probabilities) are convenient for modelling uncertainty in high-dimensional spaces and facilitate optimisation tasks (exploiting the convexity of confidence regions).

Concerning future works, there are still a number of practical results concerning ordinary p-boxes and possibility distributions whose extensions to generalised p-boxes need to be explored. Among these results are fuzzy [6] and probabilistic [19] arithmetic, respectively allowing easy propagation of fuzzy sets and ordinary p-boxes under different (in)dependence assumptions.

Acknowledgements

The second author's work has been partially supported by French Research National Agency (ANR) through CO2 program (project CRISCO2, ANR-06-CO2-003)

References

[1] C. Baudrit and D. Dubois. Practical representations of incomplete probabilistic knowledge. *Comp. Stat. & Data Anal.*, 51(1):86–108, 2006.

[2] A. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[3] T. Denoeux. Conjunctive and disjunctive combination of belief functions induced by non-distinct bodies of evidence. *Artificial Intelligence*, 172:234–264, 2008.

[4] S. Destercke, D. Dubois, and E. Chojnacki. Computing with generalized p-boxes: preliminary results. In *Proc. Information Processing and Management of Uncertainty in Knowledge-based systems (IPMU)*, 2008.

[5] S. Destercke, D. Dubois, and E. Chojnacki. Unifying practical uncertainty representations: I generalized p-boxes, II clouds. *Int. J. of Approx. Reas.*, 49:649–677, 2008.

[6] D. Dubois, E. Kerre, R. Mesiar, and H. Prade. Fuzzy interval analysis. In *Fundamentals of fuzzy sets*, pages 483–581. Kluwer, Boston, 2000.

[7] D. Dubois and H. Prade. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York, 1988.

[8] D. Dubois and H. Prade. Focusing vs. belief revision. In *Qualitative and Quantitative Practical Reasoning*, pages 96–107. Springer, 1997.

[9] D. Dubois and R. Yager. Fuzzy set connectives as combination of belief structures. *Information Sciences*, 66:245–275, 1992.

[10] S. Ferson, L. Ginzburg, V. Kreinovich, D. Myers, and K. Sentz. Constructing probability boxes and Dempster-Shafer structures. Technical report, Sandia National Laboratories, 2003.

[11] M. Fuchs and A. Neumaier. Potential based clouds in robust design optimization. *Journal of Statistical Theory and Practice*, 3:225–238, 2009.

[12] J. Jaffray. Bayesian updating and belief functions. *IEEE Trans. on Systems, Man and Cybernetics*, 22:1144–1152, 1992.

[13] E. Miranda. A survey of the theory of coherent lower previsions. *Int. J. of Approximate Reasoning*, 48:628–658, 2008.

[14] E. Miranda, M. Troffaes, and S. Destercke. Generalised p-boxes on totally ordered spaces. In *SMPS 2008 Conference, Toulouse*, pages 234–242, 2008.

[15] R. Nelsen. *An Introduction to Copulas*, volume 139 of *Lecture notes in statistics*. Springer-Verlag, New York, 1999.

[16] A. Neumaier. Clouds, fuzzy sets and probability intervals. *Reliable Computing*, 10:249–272, 2004.

[17] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, New Jersey, 1976.

[18] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.

[19] R. Williamson and T. Downs. Probabilistic arithmetic i : Numerical methods for calculating convolutions and dependency bounds. *I. J. of Approximate Reasoning*, 4:8–158, 1990.

Boundary linear utility and sensitivity of decisions with imprecise utility trade-off parameters

Malcolm Farrow
Newcastle University, UK
Malcolm.Farrow@newcastle.ac.uk

Michael Goldstein
Durham University, U.K.
Michael.Goldstein@durham.ac.uk

Abstract

In earlier work we have developed methods for analysing decision problems based on multi-attribute utility hierarchies, structured by mutual utility independence, which are not precisely specified due to unwillingness or inability of an individual or group to agree on precise values for the trade-offs between the various attributes. Our analysis is based on whatever limited collection of preferences we may assert between attribute collections. In this paper we show how to assess the robustness of our selected decision using the properties of boundary linear utility.

Keywords. Robust decisions, imprecise utilities, utility hierarchies, mutual utility independence, boundary linear utility, sensitivity analysis.

1 Introduction

In two earlier papers we have developed a methodology for decision analysis with multi-attribute utilities which does not require the specification of precise trade-offs between different risks. Multi-attribute utilities may be imprecisely specified, due to an unwillingness or inability on the part of a client to specify fixed risk trade-offs or because of disagreement within a group with responsibility for the decision.

In [3] we introduced our approach to constructing imprecise multi-attribute utility hierarchies and finding the Pareto optimal rules. We described the structure which we use, which is based on a utility hierarchy with utility independence at each node, explained the notion of imprecise utility trade-offs for such a hierarchy, based on limited collections of stated preferences between outcomes, and used Pareto optimality, over the set of possible trade-off specifications, to reduce the set of alternatives. These methods and some associated theory are summarised in Section 2 of this paper.

We are particularly concerned with problems where the number of alternatives among which we must choose is large. Many real decision problems, for example in experimental design, have very large spaces of possible choices. Relaxing the requirement for precise trade-off specification reduces our ability to eliminate rules, i.e. choices, by dominance and can leave us with a large class of rules, none of which is dominated by any other over the whole range of possible trade-offs allowed by the imprecise specification. We are therefore faced with the need for practical ways to reduce the decision space which are tractable even when the decision space is very large and there is a complicated multi-attribute utility structure to consider. In [4] we described ways to reduce further the class of alternatives that we must consider, by eliminating rules which are "ε-dominated" and combining rules which are "ε-equivalent." We explored the effects of different values of ε and of different parts of the hierarchy to see when and why rules are eliminated.

To choose a single rule d^* from our reduced list, we can use the boundary linear utility approach described in [3], or choose the rule which is the last to be eliminated as we increase the value of our ε criterion as described in [4]. We can then find the set D^* of rules which are "almost equivalent" to d^* and perhaps use secondary considerations to choose among them. We review boundary linear utility in Section 3 of this paper.

In Section 4 we describe methods, based on the boundary linear utility, for exploring the sensitivity of possible choices to variation in the utility trade-offs. This helps us to find a decision which, as far as possible, is a good choice over the whole range of possible trade-offs.

The practical implementation of our approach is illustrated throughout by an example concerning the introduction of a new course module at a university, which we first described in [4].

2 Mutually utility independent hierarchies and imprecise utility tradeoffs

2.1 Mutually utility independent hierarchies

In [3] we proposed a general class of multi-attribute utility functions. This uses the concept of mutual utility independence among sets of attributes in order to impose a structure on the utility function. Attributes $\underline{Y} = (Y_1, ..., Y_k)$ are *utility independent* of the attributes $\underline{Z} = (Z_1, ..., Z_r)$ if conditional preferences over lotteries with differing values of $\underline{Y}$ but fixed values, $\underline{z}$, of $\underline{Z}$, do not depend on the particular choice of $\underline{z}$. Attributes $\underline{X} = (X_1, ..., X_s)$ are *mutually utility independent* if every subset of $\underline{X}$ is utility independent of its complement. If attributes $\underline{X}$ are mutually utility independent, then the utility function for $\underline{X}$ must be given by the *multiplicative form*

$$U(\underline{X}) = B^{-1}\left\{\prod_{i=1}^{s}[1 + ka_iU_i(X_i)] - 1\right\}, \quad (1)$$

where B does not depend on $U_1(X_1), \ldots, U_s(X_s)$, or the *additive form*

$$U(\underline{X}) = \sum_{i=1}^{s} a_iU_i(X_i), \quad (2)$$

(see [6]) where $\mathrm{U}_i(\mathrm{X}_i)$ is a conditional utility function for attribute X_i, namely an evaluation of the utility of X_i for fixed values of the other attributes. The coefficients in (1) and (2) are the *trade-off parameters*; the a_i reflect the relative importance of the attributes and k reflects the degree to which rewards may be regarded as complementary, if $k > 0$, or as substitutes, if $k < 0$.

The assumption of mutual utility independence, which many people would often be prepared to make, is enough in itself to reduce the problem to one of considering a finite number of parameters.

Keeney and Raiffa [6] also describe the idea of a hierarchy of utilities, as follows. We form an overall multi-attribute utility from marginal utilities for the various attributes by a hierarchical structure in which, at each node, several utilities are merged into a combined utility. This combined utility is merged with others at a node in the next level until, finally, one overall utility function is formed. If, at each node, we have mutual utility independence for the utilities combined at that node, then we term such a utility function a *Mutually Utility Independent Hierarchic (MUIH)* utility. Thus, in a MUIH utility, at each node we combine utilities using either (1) or (2).

Our hierarchical structure allows us to relax the requirement for overall mutual utility independence by allowing the user to specify utility independence just at the nodes of the hierarchy and, of course, the user can choose this structure.

In our utility hierarchy we consider the overall utility node to be at the "top" level and the predecessors of a node to be at "lower" levels. We refer to the nodes corresponding to the individual attributes, that is nodes which have no predecessors, as *marginal nodes*. We refer to a direct predecessor of a node as a *parent* and a direct successor as a *child*. For each node n, we denote by $H(n)$, the *sub-hierarchy* under n, where $H(n)$ is the set of nodes containing n and all of its predecessors. We divide the child nodes in the hierarchy into the following three types:

1. an *additive node*, where utilities are combined as in (2) with $\sum_{i=1}^{s} a_i \equiv 1$ and $a_i > 0$ for $i = 1, \ldots, s$;
2. a *binary node*, where precisely two utilities are combined, where we rescale the combined utility as
$$U = a_1U_1 + a_2U_2 + hU_1U_2 \quad (3)$$
where $0 < a_i < 1$ and $-a_i \leq h \leq 1 - a_i$, for $i = 1, 2$, and $a_1 + a_2 + h \equiv 1$. Note that (3) is derived by setting $s = 2$ and $h = ka_1a_2$ in (1).
3. a *multiplicative node*, where more than two utilities are combined and the parameter k in (1) may be nonzero. We scale the utility using
$$B = \prod_{i=1}^{s}(1 + ka_i) - 1 \quad (4)$$
with $a_1 \equiv 1, k > -1$ and, for $i = 1, \ldots, s$, we have $a_i > 0$ and $ka_i > -1$.

For each child node n, we denote by $\underline{\phi}_n = (\phi_{n,1}, \ldots, \phi_{n,m(n)})$ the collection of trade-off parameters which determine how the parent utilities at node n are combined to give the value at the child node. Thus, each $\phi_{n,j}$ corresponds to an a_i in (2) an a_i or h term in (3), or an a_i or k in (1). If there are N child nodes, then we denote by $\underline{\theta} = (\underline{\phi}_1, \ldots, \underline{\phi}_N)$ the collection of all the trade-off parameters in the hierarchy. If we allow imprecision in some of the elements of $\underline{\theta}$, then we refer to the resulting utility specification as an *imprecise independence hierarchy (IIH)*. If the hierarchy contains only additive and binary nodes, then we refer to the specification as a *simple imprecise independence hierarchy (SIIH)*

The utility at each child node is determined both by the values of the utilities at the marginal nodes and

also by the choice of trade-off parameters. As we shall vary the trade-off parameters, and thus the utilities at the child nodes, we require a standard scale for all utilities in the IIH, whose interpretation does not depend on the choice of trade-off parameters. This is constructed as follows.

As the marginal utility at each marginal node is expressed in a utility scale, we norm all the marginal utilities to lie between 0, the worst outcome that we shall consider for the problem, and 1, the best outcome. The effect of the scalings that we have chosen for additive, binary and multiplicative nodes is that, at each node n in the hierarchy, the utilities of C_n and c_n are 1 and 0 respectively, where C_n is an outcome such that all marginal predecessor nodes have utility 1, and c_n is an outcome such that all marginal predecessor nodes have utility 0. Therefore, a utility value of u at node n may always be interpreted as the utility of a gamble giving C_n with probability u and c_n with probability $1-u$, irrespective of the chain of trade-off parameters in the hierarchy. This utility scale is termed the *standard scale* for the hierarchy. Throughout this paper, all utilities are assumed to be on the standard scale.

2.2 Example: Designing a new course module at a university

In [4] we introduced an example concerning the design of a new course module at a university. We use the same example here to illustrate our approach. The module is to contain six units, or topics, each of which may, for the purpose of this example, be considered to be of the same size in the sense that, given the same teaching method, they would require the same length of time. Each topic could be taught by any one of three teaching methods, denoted as follows:

Lect : a traditional course of lectures and tutorials.

Lab : a laboratory-based course using a computer algebra package.

OL : an "open learning" course without lectures or formal laboratory sessions.

Thus we have $3^6 = 729$ possible choices of combinations of teaching methods. We can denote a choice $(\mu_1, \ldots, \mu_6)$ where $\mu_i = 1$,2 or 3 according to which method is used for unit i. (In practice there are additional choices to be made, but we do not wish to introduce unnecessary complexity into this example). The attributes which we consider in our analysis are as follows. Further details are given in [4].

- For students:

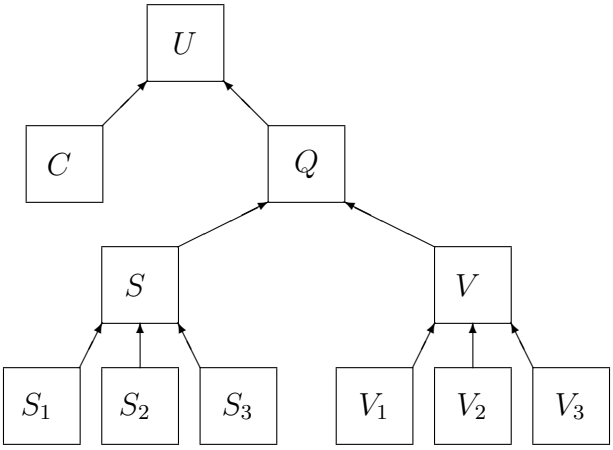


Figure 1: Utility hierarchy for the course design example.

  S_1 short term learning,
  S_2 longer-term learning,
  S_3 satisfaction,

- For the university and staff:

  V_1 staff satisfaction,
  V_2 institutional benefits,
  V_3 staff development,
  C financial cost.

As for many decision problems, the attributes of interest are in very different units and it may be difficult to establish precise trade-offs between the attributes in order to rank the various teaching choices.

2.3 Example: Utility hierarchy

The utility hierarchy is shown in Figure 1.

The overall utility node U is a binary node, combining the utility U_C for cost and the utility U_Q for quality. So the overall utility is

$$U = a_{UQ}U_Q + a_{UC}U_C + h_U U_Q U_C.$$

The "module quality" utility U_Q is formed at a binary node and is given by

$$U_Q = a_{QS}U_S + a_{QV}U_V + h_Q U_S U_V,$$

where U_S and U_V are the utilities for "Students" and "University". Each of these is an additive node which depends on three marginal utilities:

$$\begin{aligned} U_S &= a_{S1}U_{S1} + a_{S2}U_{S2} + a_{S3}U_{S3}, \\ U_V &= a_{V1}U_{V1} + a_{V2}U_{V2} + a_{V3}U_{V3}. \end{aligned}$$

	Vertex			
	$\underline{\phi}_{S1}$	$\underline{\phi}_{S2}$	$\underline{\phi}_{S3}$	$\underline{\phi}_{S0}$
a_{S1}	0.2	0.2	0.5	0.3
a_{S2}	0.4	0.7	0.4	0.5
a_{S3}	0.4	0.1	0.1	0.2

Node S (Students).

	Vertex			
	$\underline{\phi}_{V1}$	$\underline{\phi}_{V2}$	$\underline{\phi}_{V3}$	$\underline{\phi}_{V0}$
a_{V1}	0.05	0.05	0.20	0.10
a_{V2}	0.50	0.75	0.55	0.60
a_{V3}	0.45	0.20	0.25	0.30

Node V (University).

	Vertex				
	$\underline{\phi}_{Q1}$	$\underline{\phi}_{Q2}$	$\underline{\phi}_{Q3}$	$\underline{\phi}_{Q4}$	$\underline{\phi}_{Q0}$
a_S	0.890	0.500	0.890	0.500	0.695
a_V	0.110	0.500	0.305	0.305	0.305
h_Q	0.000	0.000	-0.196	0.195	0.000

Node Q (Module Quality).

	Vertex				
	$\underline{\phi}_{U1}$	$\underline{\phi}_{U2}$	$\underline{\phi}_{U3}$	$\underline{\phi}_{U4}$	$\underline{\phi}_{U0}$
a_C	0.7	0.5	0.7	0.5	0.6
a_Q	0.3	0.5	0.4	0.4	0.4
h_U	0.0	0.0	-0.1	0.1	0.0

Node U (Overall Utility).

Table 1: Trade-off parameter values.

The marginal utilities U_{S1}, U_{S2}, U_{S3}, U_{V1}, U_{V2}, U_{V3} are associated with the attributes S_1, S_2, S_3, V_1, V_2, V_3. Details of the evaluation of expected marginal utilities are given in [4].

The utility function is fully specified when we assign values to all of the trade-off parameters in the above relations. In this paper, we shall consider how to analyse the problem as an SIIH, when we are unwilling to give precise values to these trade-offs.

2.4 Using Imprecise Trade-off Parameters

One of the most difficult tasks in specifying a mutually utility independent structure is the quantification of the various trade-off parameters in the forms (2), (3) and (1), as this typically requires the comparison of intrinsically different types of costs and benefits. Therefore, it is of fundamental interest to consider problems where we are unwilling to fix on particular trade-off values or where a group of individuals must make a joint decision, and there is broad agreement on the marginal utilities, but different members of the group have different priorities when trading risks.

Although we are unwilling to place strict values on the trade-offs, there will be certain combinations of outcomes over which we are prepared to state preferences and these comparisons establish the region of the space of trade-off parameters which we must consider. We choose to elicit our imprecision in the values of the trade-off parameters $\underline{\theta}$ based on our stated preferences over utility combinations for outcomes, as this is usually more meaningful than considering directly the imprecision in the elements of $\underline{\theta}$. So, for each child node, we make a collection of pairwise comparisons between vectors of values of parent utilities (or, equivalently, the corresponding vectors of attribute values). Details are given in [3].

Some authors also consider imprecision in the marginal utility functions. Recent examples include [7] who describe a decision support system in which the imprecise multi-attribute utility function is additive and [5], who allow a multiplicative function in which a range for the value of k in (1) is determined by considering the values implied by ranges given for $a_1, \ldots, a_s$. In both cases ranges for the trade-off parameters are combined to form a rectangular space. In this paper we only consider imprecision in trade-offs and assume that the necessary expectations of marginal utilities, and in some cases their products, can be agreed. However we do not impose an arbitrary probability distribution over ranges of imprecision, or over attributes, nor do we assume a rectangular shape for the space of trade-off parameters allowed by the imprecise specification resulting from a careful elicitation process.

For each additive or binary child node, we state whichever preferences we wish between pairs of utility vectors for the parent nodes. Each stated preference places a linear constraint on the allowable choices for the trade-off parameters $\underline{\phi_i}$. We term the collection, R, of all sets of trade-off parameters consistent with each of the stated preferences the *feasible* region of choices for the trade-off parameters. In [3] we showed that the shape of the region of trade-off parameters resulting from the above elicitation scheme for an SIIH is as follows. At each additive or binary node n, we obtain a convex polyhedron R_n for the allowable values of $\underline{\phi}_n$. The regions $R_1, \ldots, R_N$ together define a

d_1	1,	3,	1,	1,	3,	2
d_2	2,	3,	1,	3,	3,	2
d_3	1,	3,	1,	3,	3,	2
d_4	1,	3,	1,	1,	1,	3
d_5	1,	3,	1,	1,	3,	3
d_6	2,	3,	1,	1,	3,	2

Table 2: Alternatives for comparison.

region R in the combined space of parameters $\underline{\theta}$, where $\underline{\theta} \in R$ if and only if $\underline{\phi}_n \in R_n$ for $n = 1, \ldots, N$. The vertices of R_n are denoted $\underline{\phi}_n^{(1)}, \ldots, \underline{\phi}_n^{(r_n)}$ and those of R are denoted $\underline{\theta}^{(1)}, \ldots, \underline{\theta}^{(r)}$. Let P be the set of vertices of R and P_n be the set of vertices of R_n.

We explained in [3] that, in the case of an IIH containing multiplicative nodes where the utilities are combined using (1) and (4), we must modify the elicitation procedure. We also described the shape of the resulting feasible set. If we are willing to choose a fixed value for k then, at each multiplicative node n, we obtain a bounded rectangular region $R_n(k)$, with vertices $\underline{\phi}_n^{(1)}, \ldots, \underline{\phi}_n^{(r_n)}$, for the remaining elements of $\underline{\phi}_n$. The shape is somewhat more complicated if the value of k is also treated as imprecise.

2.5 Example: Imprecise trade-offs

The specification of imprecise utility trade-offs in this example was described in more detail in [4]. Table 1 gives the vertex set P_i for the feasible region R_i, at each node i. For each node, a central value $\underline{\phi}_{i0}$ is also listed, which is the average of the values at each vertex.

In [4] we found that there were 50 Pareto optimal choices in this example. Of these, 37 could be eliminated because they were equivalent to other choices which were retained. We chose the value $\varepsilon = 0.012$ and, by applying our ideas of almost-preference with this value of ε, reduced the list to the six alternatives listed in Table 2. These are ordered according to our ε-preference procedure, d_6 being eliminated before d_5 and so on. The last remaining choice is d_1.

3 Boundary linear utility

3.1 Definitions and motivation

The feasible region for the trade-off parameters in a SIIH is the convex hull of a finite collection of trade-off parameters $\underline{\theta}^{(i)} \in P, i = 1, \ldots, r$. We now need a way to compare non-dominated choices over this region. Let U_i be the utility function determined by the choice of trade-offs $\underline{\theta}^{(i)} \in P, i = 1, \ldots, r$. Any function of the form

$$\bar{U}_\lambda = \sum_{i=1}^{r} \lambda_i U_i \tag{5}$$

where $\lambda = (\lambda_1, \ldots, \lambda_r)$ are non-negative constants such that $\sum_{i=1}^r \lambda_i = 1$ is termed a *boundary linear utility*. For any such $\bar{U}_\lambda$, we may identify the rule which maximises $\bar{U}_{d,\lambda} = \sum_{i=1}^r \lambda_i U_{d,i}$, where $U_{d,i}$ is the utility of alternative d with trade-off $\underline{\theta}^{(i)}$.

In [3] the boundary linear form is motivated by various axiomatic and natural requirements for the combination of group preferences. In addition to such theoretical support, the boundary linear form is easy to interpret, gives a clear comparison between different choices and leads to tractable procedures even for large numbers of alternative decisions. The choice of the λ weights can be used to emphasise or de-emphasise the importance of a particular attribute by putting more or less weight on vertices corresponding to different values for a particular trade-off.

While the set of λ weights is formally equivalent to a probability distribution over the points in P, our interpretation of the λ weights is not probabilistic but is in terms of the properties of the boundary linear utility described below and as a means for exploring the robustness of alternatives. A probability distribution over possible sets of attribute weights is used in [1] as a means of exploring sensitivity. In [2] a weight specification, known as a second order belief specification, over the ranges of imprecisely specified probabilities and expected utilities in a decision tree is used to help make a unique choice of alternative.

3.2 Properties of the boundary linear utility

Let us consider first the case of a SIIH.

There is a natural relation between Pareto optimality and Bayes rules for boundary linear utilities. In [3] we showed that, for a SIIH, a decision which is either (i) a unique Bayes decision for some $\bar{U}_\lambda$, or (ii) a Bayes decision for some $\bar{U}_\lambda$ with $\lambda_i > 0$ for $i = 1, \ldots, r$, is Pareto optimal over R.

Each weight λ_i corresponds to a complete parameter specification $\underline{\theta}^{(i)}$. It is useful to be able to relate this to weights applied to parameter specifications at individual nodes. Denote by $\lambda(i_1, \ldots, i_N)$ the weight applied to the combination of vertices $\underline{\phi}_1^{(i_1)}, \ldots, \underline{\phi}_N^{(i_N)}$ at nodes $1, \ldots, N$ respectively. Denote by $\lambda_{n,i}$ the weight applied to vertex $\underline{\phi}_n^{(i)}$ at node n. If we require that the weights applied to vertices at node n should not change if we combine this vertex with a different vertex at another node then we require

$$\frac{\lambda(i_1,\ldots,i_n,\ldots,i_N)}{\lambda(i_1,\ldots,i'_n,\ldots,i_N)} = \frac{\lambda_{n,i_n}}{\lambda_{n,i'_n}}$$

for two different vertices i_n and i'_n at node n, with $\lambda_{n,i'_n} \neq 0$. It follows that $\lambda(i_1,\ldots,i_N) = \prod_{n=1}^{N} \lambda_{n,i_n}$. Such a weight specification is called a *multiplicative weighting*. For such a specification, we may vary the weights at each node separately.

It is often helpful to equate the boundary linear form with the utility at interior trade-off values. It follows directly from the fact that R_i is a convex polyhedron that, for any $\underline{\theta}$ in R, there exists a multiplicative weighting λ such that $\underline{\theta} = \underline{\bar{\theta}}_\lambda$ and, for any multiplicative weighting λ, there exists a $\underline{\theta}$ in R such that $\underline{\theta} = \underline{\bar{\theta}}_\lambda$, where $\underline{\bar{\theta}}_\lambda = \sum_j \lambda_j \underline{\theta}^{(j)}$ and the sum is taken over all of the vertices of R. In [3] we showed that, in a SIIH, if λ is a multiplicative weighting then $\bar{U}_\lambda = U(\underline{\bar{\theta}}_\lambda)$. This result establishes a correspondence between the elements of R and the multiplicative boundary linear utilities.

From this we know that, for any $\underline{\theta}$ in R, we can find $\lambda_1,\ldots,\lambda r$ such that $U(\underline{\theta}) = \sum_i \lambda_i U_i$. Values of $\underline{\theta}$ not on the boundary of R will give $\underline{\lambda}$ values satisfying $\lambda_i > 0$ for $i = 1,\ldots,r$. Rules which are Bayes for such internal $\underline{\theta}$ values will therefore be Pareto optimal over R.

For illustration of the multiplicative weighting, consider a simple example with three marginal utilities and two additive nodes where

$$U_0 = \phi_{01}U_1 + \phi_{02}U_2,\ U_1 = \phi_{13}U_3 + \phi_{14}U_4$$

and at each of the two nodes we have two alternative parameter specifications, corresponding to the vertex values. The two values for ϕ_{01} are ϕ_{011} and ϕ_{012} etc. Thus R has four vertices. Assign weight λ_{jk} to the vertex where node 0 takes parameter specification j and node 1 takes parameter specification k. The coefficient of U_3 in U_0 is now

$$\begin{aligned}\Phi_3 &= \{(\lambda_{11}+\lambda_{12})\phi_{011} + (\lambda_{21}+\lambda_{22})\phi_{012}\} \\ &\quad \times \{(\lambda_{11}+\lambda_{21})\phi_{131} + (\lambda_{12}+\lambda_{22})\phi_{132}\}.\end{aligned}$$

Now introduce weights on the parameter values at the individual nodes and calculate the overall weights from these so that $\lambda_{11} = \lambda^\star_{01}\lambda^\star_{11}, \lambda_{12} = \lambda^\star_{01}\lambda^\star_{12}, \lambda_{21} = \lambda^\star_{02}\lambda^\star_{11}, \lambda_{22} = \lambda^\star_{02}\lambda^\star_{12}$, where $\lambda^\star_{01}$ is the weight on the first parameter set at node 0 etc. and the weights at each node sum to 1. The coefficient of U_3 now simplifies to

$$\begin{aligned}\Phi_3 &= \{\lambda^\star_{01}\phi_{011} + \lambda^\star_{02}\phi_{012}\}\{\lambda^\star_{11}\phi_{131} + \lambda^\star_{12}\phi_{132}\} \\ &= \lambda^\star_{01}\lambda^\star_{11}\phi_{011}\phi_{131} + \lambda^\star_{01}\lambda^\star_{12}\phi_{011}\phi_{132} \\ &\quad + \lambda^\star_{02}\lambda^\star_{11}\phi_{012}\phi_{131} + \lambda^\star_{02}\lambda^\star_{12}\phi_{012}\phi_{132} \\ &= \lambda_{11}\phi_{011}\phi_{131} + \lambda_{12}\phi_{011}\phi_{132} + \lambda_{21}\phi_{012}\phi_{131} \\ &\quad + \lambda_{22}\phi_{012}\phi_{132}\end{aligned}$$

a weighted average of the coefficients at the four vertices, as required.

3.3 Boundary linear utility in a general IIH

The boundary linear utility is easily extended to the case where a hierarchy contains multiplicative nodes where the utilities are combined as in (1) and (4) provided that a precise value of the parameter k is used. The extension to the case where k is imprecisely specified is discussed in [3] where we showed that, in any IIH, for any $\underline{\theta}$ in R there exists a multiplicative weighting λ such that $U(\underline{\theta}) = \bar{U}_\lambda$, thus generalising the correspondence between the elements of R and the multiplicative boundary linear utilities.

3.4 Example: Boundary linear utility

With equal λ weights on all vertices, the alternative which maximises $\mathrm{E}(\bar{U}_\lambda)$ is rule d_1 which gives $\mathrm{E}(\bar{U}_\lambda) = 0.5120$. The central point $\underline{\theta}_0$, at which $U(\underline{\theta}) = \bar{U}_\lambda$, is given by the centres of each range as given in Table 1.

The λ weights could be varied to change the emphasis on different attributes. For example, at node U the coefficient of financial cost varies between 0.5 and 0.7. Putting more weight on all vertices where the coefficient was 0.7 would emphasise this attribute, whereas more on all vertices where it was 0.5 would de-emphasise it. For illustration we changed the weights to 2:1 in favour of 0.7 and 2:1 in favour of 0.5. In each case rule d_1 maximised $\mathrm{E}(\bar{U}_\lambda)$ giving values of 0.5096 and 0.5144 respectively. This increases our confidence in the choice of d_1.

Sometimes, we may uniquely choose a collection of λ weights under the guidance of one of the formal arguments in [3]. However, usually we will want to consider the robustness of our choice to variation in λ, which we now address more formally.

4 Exploring sensitivity

4.1 General comments

The boundary linear utility gives us an approach to choosing between alternative rules. However, while

any given boundary linear utility function identifies a "best" alternative, we would usually prefer an alternative which is robust in the sense that it behaves well compared to most alternatives over most of the range of trade-off parameters. We now consider how such robustness may be assessed.

When we have chosen a multiplicative boundary linear utility $\bar{U}_\lambda = \sum_{i=1}^{r} \lambda_i U_i$, we find the decision $d^\star$ which maximises expected utility, under $\bar{U}_\lambda$. We also define a 'central' parameter specification $\underline{\theta}_0 = \sum_{i=1}^{r} \lambda_j \underline{\theta}^{(j)}$ where this sum is taken over the elements of P. From Section 3.2 we know that, in a SIIH, when λ is a multiplicative weighting, $U(\underline{\theta}_0) = \bar{U}_\lambda$. Thus, we can explore sensitivity in two ways. First, we can see how much we must change the λ weightings in order to alter our choice of best decision and secondly, at least in a SIIH, we can see how far we must move away from the central value θ_0, to alter our choice. Effectively, this establishes two separate but related sensitivity metrics. The former is concerned solely with the relative importance of the various vertices of the trade-off space, irrespective of their Euclidean values, while the latter reflects the actual Euclidean distances between alternative trade-off parameters.

The investigations described below are designed to assess the robustness of our decision to the choice of trade-off. At each step, if the analysis suggests that there are other alternatives which perform substantially better than our selected rule over much of the trade-off space, then we may repeat the steps, substituting the suggested alternatives, to see whether a more robust choice of rule may be found.

Suppose, in what follows, that we have a set D of alternatives for comparison with $d^\star$. This set may be a subset of the Pareto optimal choices formed using the methods in [4]. Suppose also that we have chosen a small increment $\varepsilon > 0$ which we tolerate in comparing utilities, as discussed in [4].

4.2 Volume sensitivity

A first general robustness measure is as follows. For each alternative in D, we compute the volume of λ-space, as a proportion of the total volume within which $\sum \lambda_j = 1$, over which the difference in utility between that alternative and $d^\star$ is at least ε. If this proportion is very small, then this suggests that $d^\star$ is robust against that alternative.

Having assessed global sensitivity over the whole hierarchy, we may repeat the analysis in any subhierarchy. For any child node i, with utility U_i, we may find the proportion of the permissible λ-space for the vertices of the feasible region of parameters in the sub-hierarchy under i in which the difference in expectations of U_i between an alternative and $d^\star$ is at least ε.

To do these analyses we need to be able to compute the volume of λ-space which satisfies a condition

$$g(d_1, d_2) = \bar{U}_\lambda(d_1) - \bar{U}_\lambda(d_2) > x \tag{6}$$

for some specified x, where d_1 and d_2 are two choices. Let $\underline{d} = (d_1, \ldots, d_e)$ and $\underline{\bar{U}}_\lambda^{(n)}(\underline{d}) = (\bar{U}_\lambda^{(n)}(d_1), \ldots, \bar{U}_\lambda^{(n)}(d_e))$, where $\bar{U}_\lambda^{(n)}(d_j)$ is the boundary linear utility evaluated at node n with weights $\underline{\lambda}$ over the subhierarchy $H(n)$ under n. To evaluate the volume satisfying (6), we can make use of the following analogy.

If we gave $\underline{\lambda}$ a uniform distribution over its feasible region then the required volume would be the probability that (6) is satisfied. The utility hierarchy can then be interpreted as a graph in which the probability distribution of the utility difference between any two decisions at a child node, given the values of the parent utilities, would depend only on the distribution of the tradeoff parameters at the child node. Thus we can evaluate the distribution of $\underline{\bar{U}}_\lambda^{(n)}(\underline{d})$ higher in the hierarchy through a chain of conditional distributions. See, e.g., [9].

Specifically, the density of $\underline{\bar{U}}_\lambda^{(n)}(\underline{d})$, the values at a child node n with parents $n_1, \ldots, n_s$, is

$$f_n(\underline{\bar{U}}_\lambda^{(n)}(\underline{d})) = \int \cdots \int \left\{ f_{n|H(n)}[\underline{\bar{U}}_\lambda^{(n)}(\underline{d}) \mid \underline{\bar{U}}_\lambda^{(n^*)}(\underline{d})] \prod_{i=1}^{s} f_{n_i}(\underline{\bar{U}}_\lambda^{(n_i)}(\underline{d})) \right\} d\underline{\bar{U}}_\lambda^{(n^*)}(\underline{d}) \tag{7}$$

and

$$\Pr(g_n > x) = \int \cdots \int \left\{ \Pr[g_n > x \mid \underline{\bar{U}}_\lambda^{(n^*)}(\underline{d})] \prod_{i=1}^{s} f_{n_i}[\underline{\bar{U}}_\lambda^{(n_i)}(\underline{d})] \right\} d\underline{\bar{U}}_\lambda^{(n^*)}(\underline{d}) \tag{8}$$

where $\underline{\bar{U}}_\lambda^{(n^*)}(\underline{d}) = \underline{\bar{U}}_\lambda^{(n_1)}(\underline{d}), \ldots, \underline{\bar{U}}_\lambda^{(n_s)}(\underline{d})$ and $f_{n|H(n)}[\underline{\bar{U}}_\lambda^{(n)}(\underline{d}) \mid \underline{\bar{U}}_\lambda^{(n^*)}(\underline{d})]$ is the conditional density given the values of the boundary linear utilities evaluated at the parent nodes for the elements of $\underline{d}$.

Starting with the children of the marginal nodes, the distribution of $\underline{\bar{U}}_\lambda^{(n)}(\underline{d})$ is evaluated node-by-node up

the hierarchy using (7). At a child node n with r_n vertices we have, from (5) and (6),

$$g_n(d_1, d_2) = \sum_{i=1}^{r_n} \lambda_{n,i}[U_i^{(n)}(d_1) - U_i^{(n)}(d_2)]$$

where $U_i^{(n)}(d)$ is evaluated at vertex i of node n. Thus $g_n(d_1, d_2) = x$ defines a plane in λ-space which may cut the feasible region. The conditional probability $\Pr[g_n > x \mid \underline{\bar{U}}_\lambda^{(n^*)}(\underline{d})]$ in (8) is then a proportion of the volume of the feasible polyhedron which can be determined by finding where $g_n = x$ cuts the edges.

Similarly, the conditional probability $\Pr[\bar{U}_\lambda^{(n)}(d_1) < x \mid \underline{\bar{U}}_\lambda^{(n^*)}(\underline{d})]$ is the proportion of the volume of the feasible polyhedron at node n cut off by $\bar{U}_\lambda^{(n)}(d_1) = x$, with the parent utilities fixed. Differentiating this probability with respect to x gives the conditional density of $\bar{U}_\lambda^{(n)}(d_1)$. Then fixing $\bar{U}_\lambda^{(n)}(d_1) = x_1$ imposes a linear constraint on $\lambda_{n,1}, \ldots, \lambda_{n,r_n}$ and reduces the dimension of the feasible region by 1. By considering the intersection of $\bar{U}_\lambda^{(n)}(d_2) = x_2$ with this reduced region we can find the conditional distribution of $\bar{U}_\lambda^{(n)}(d_2)$ given $\bar{U}_\lambda^{(n)}(d_1) = x_1$. If required, we can continue this process for $d_3, \ldots, d_{r_n-1}$. For $j > r_n - 1$, $\bar{U}_\lambda^{(n)}(d_j)$ is then a deterministic function of $\underline{\bar{U}}_\lambda^{(n)}(d_1, \ldots, d_{r_n-1})$. In this way we can find the conditional density $f_{n|H(n)}[\underline{\bar{U}}_\lambda^{(n)}(\underline{d}) \mid \underline{\bar{U}}_\lambda^{(n^*)}(\underline{d})]$ in (7).

4.3 Example: Volume sensitivity

We have identified the choice d_1 under the utility with equal weightings at each vertex in P. We now consider the sensitivity of that choice, following the steps in Section 4.

We computed the volume of λ-space, as a proportion of the total volume within which $\sum \lambda_j = 1$, over which the difference in utility between alternative d_1 and each of the other retained alternatives is at least $-\varepsilon$, at our chosen value of 0.012. We concluded that the volume over which the difference in favour of any alternative over d_1 is greater than ε is less than 0.01% of the total volume and therefore that d_1 is a robust choice. (The proportion is nonzero, since we know that the difference is greater than ε at some of the vertices. However the region of λ space which we are exploring is a simplex of very high dimension and the neighbourhoods of the vertices of this simplex contribute only a tiny fraction of the total volume.)

Next we computed the proportions of λ-volume over which each alternative's boundary linear utility exceeded that of d_1 by at least ε for each of the non-marginal nodes in the hierarchy. Table 3 gives the results. The results show that the challenge to d_1 seems to be based in node V. The apparent main challengers, rules d_2 and d_6, differ only in Unit 4 which is given by lectures in d_6 and open learning in d_2. According to the elicited expectations, d_6 thus favours the students more.

		Node			
		U	Q	S	V
Rule	d_2	0.000	1.000	0.103	1.000
	d_3	0.000	0.001	0.000	1.000
	d_4	0.000	0.000	0.000	0.000
	d_5	0.000	0.000	0.000	0.000
	d_6	0.000	1.000	1.000	1.000

Table 3: Proportions of λ volume where the utility difference is at least ε at non-marginal nodes.

	Choice					
	d_1	d_2	d_3	d_4	d_5	d_6
C	0.484	0.416	0.476	0.544	0.536	0.424
S_1	0.497	0.447	0.463	0.433	0.400	0.480
S_2	0.578	0.577	0.528	0.420	0.370	0.627
S_3	0.800	0.900	0.800	0.800	0.800	0.900
V_1	0.533	0.467	0.433	0.500	0.400	0.567
V_2	0.433	0.667	0.533	0.300	0.400	0.567
V_3	0.467	0.717	0.583	0.333	0.450	0.600

Table 4: Values of expected utilities at the marginal nodes.

4.4 Distances in λ-space

Next, for each alternative in D, we identify those vertices where the difference in utility between that choice and $d^\star$ is at least ε. For each of these vertices, we find the distance, in λ-space, in the direction of the vertex, between $\underline{\lambda}_0$, our original λ specification, and the point where the difference in boundary linear utility between that choice and $d^\star$ first reaches ε. Let $\| \underline{\lambda} \| = \sqrt{\underline{\lambda}'\underline{\lambda}}$, where $\underline{\lambda}'$ is the transpose of $\underline{\lambda}$. We find $t \| \underline{\lambda}_v - \underline{\lambda}_0 \|$, where $\underline{\lambda}_v$ is the λ vector for a vertex, $t = \{\delta(\underline{\lambda}_0) + \varepsilon\}/\{\delta(\underline{\lambda}_0) - \delta(\underline{\lambda}_v)\}$ and $\delta(\underline{\lambda})$ is the difference in boundary linear utility at $\underline{\lambda}$. Large values of these distances suggest robustness of $d^\star$. In this metric, the distance between any two vertices is $\sqrt{2}$.

4.5 Example: Distances in λ-space

Table 4 shows the values of the expected marginal-node utilities for the members of D and Table 6 shows at which marginal node each alternative is superior to $d^\star$. Table 5 lists the vertices where the difference in utility between one of the other alternatives and

$d^\star$ is at least ε. The vertices are numbered for easy reference. The vertices can be identified using the numbering of the vertices at each node, which is the same as in Table 1. Table 5 then gives the distances, from the original $\underline{\lambda}$ specification towards these vertices, to reach points where the difference in utility between one of the other alternatives and $d^\star$ is at least ε. Most of the distances are large. There are a few exceptions, notably for rule d_2 at vertices 46 and 47. Rule d_2 is the retained option with the least dependence on traditional lectures and at these vertices relatively little weight is placed on financial cost but relatively great weight is placed on institutional benefit. To put the distances in context, observe that each original λ value is approximately 0.007. The move required for vertex 46 changes λ_{46} to approximately 0.6 and therefore the average of the other λ values is less than 0.003 or 0.5% of λ_{46}. There seems to be little reason here to change our conclusion that d_1 is a robust choice. Notice also how the pattern of marginal nodes in common between rules in Table 6 tends to be repeated with vertices in common in Table 5.

4.6 Sensitivity in the θ-metric

We can quantify sensitivity in the θ-metric by looking at the effect of general movement away from $\underline{\theta}_0$ as follows. Let the elements of P be $\underline{\theta}^{(1)}, \ldots, \underline{\theta}^{(r)}$. Define the *scaled range* R_t to be the convex hull of P_t, the elements of which are given by $\underline{\theta}_t^{(i)} = \underline{\theta}_0 + t(\underline{\theta}^{(i)} - \underline{\theta}_0)$ for $t \geq 0$. We may think of this as expanding a volume (in the θ-metric) centred on $\underline{\theta}_0$ until a boundary of the region of optimality of $d^\star$ is reached. An obvious extension of Lemma 2 in [3] shows that this boundary will be reached first at an element of P_t so we only need to make comparisons at the vertices. For each element of D we evaluate, at each of a range of values of t up to 1, the maximum over P_t of the difference in expected utility compared with $d^\star$ and plot these values against t. This plot will serve as an indication of over how large a range around $\underline{\theta}_0$ we can judge $d^\star$ to be robust. This approach may be compared with that of [8] in which the sensitivity of a preferred alternative is measured using the minimum distance (in some metric) to a point in the parameter space at which another alternative becomes preferable.

4.7 Example: Sensitivity in the θ-metric

Figure 2, shows one of the range expansion plots. The horizontal axis is the expansion factor t. The vertical axis is the difference in expected utility between an alternative, in this case d_2, and $d^\star$, in this case d_1. At each value of the expansion factor the values at the 144 vertices of the range were calculated and the maximum, minimum, median and upper and lower quartiles of these 144 values are plotted.

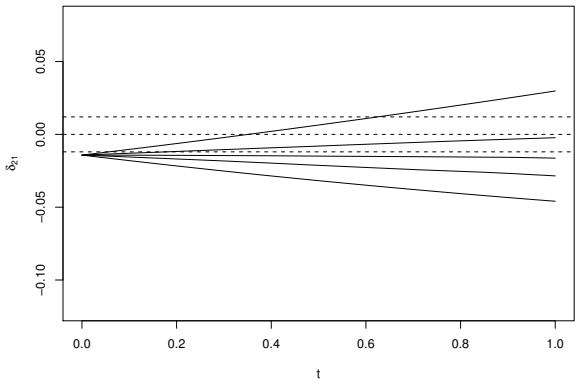


Figure 2: Expansion with respect to all parameters. Maximum, quartiles and minimum of the difference in expected utility between d_2 and d_1 at 144 vertices, against expansion factor t. Reference lines are given at zero and $\pm\varepsilon$.

Choice	Marginal Node					
d_2			S_3		V_2	V_3
d_3					V_2	V_3
d_4	C					
d_5	C					
d_6		S_2	S_3	V_1	V_2	V_3

Table 6: Marginal nodes at which alternatives are superior to d_1.

From Figure 2, we see that d_2 does substantially worse than d_1 over most of the range but possibly better for large t. Similar plots for the other alternatives show that none of the other rules does much better than d_1 over any of the range and some do much worse in some of the range. Generally the maximum difference only exceeds ε towards the end of the range. We conclude that d_2 is the only alternative to d_1 worth further consideration.

5 Conclusion

In [3], [4] and this paper we have described an approach to multi-attribute decision analysis where the trade-offs between attributes are not precisely specified. Imposing the condition of utility independence makes the dimensionality of the trade-off specification finite and allows us to work in terms of ranges for trade-off parameters. However, by imposing this condition only at the nodes of a utility hierarchy we can relax the requirement for mutual utility indepen-

	Vertex for Node				Distance for Alternative			Vertex for Node				Distance for Alternative		
Vertex	U	Q	S	V	d_4	d_5	Vertex	U	Q	S	V	d_2	d_3	d_6
1	1	1	1	1	0.591		46	2	2	1	1	0.354	0.815	0.710
2	1	1	1	2	0.591		47	2	2	1	2	0.371	0.919	0.710
3	1	1	1	3	0.567		48	2	2	1	3	0.672		0.967
7	1	1	3	1	0.954		49	2	2	2	1	0.518		0.978
8	1	1	3	2	0.954		50	2	2	2	2	0.548		0.978
9	1	1	3	3	0.906		52	2	2	3	1	0.638		
10	1	2	1	1		0.698	53	2	2	3	2	0.681		
11	1	2	1	2		0.732	64	2	4	1	1	0.468		0.779
12	1	2	1	3	0.843	0.867	65	2	4	1	2	0.492		0.782
16	1	2	3	1		0.994	66	2	4	1	3	0.892		
19	1	3	1	1	0.654		67	2	4	2	1	0.929		
20	1	3	1	2	0.656		68	2	4	2	2	0.992		
21	1	3	1	3	0.605		118	4	2	1	1	0.637		
28	1	4	1	1	0.942	0.809	119	4	2	1	2	0.673		
29	1	4	1	2	0.938	0.840	121	4	2	2	1	0.991		
30	1	4	1	3	0.782	0.983	136	4	4	1	1	0.878		
							137	4	4	1	2	0.930		

Table 5: Distances to points where the utility difference is at least ε.

dence between all attributes. In our earlier papers we discussed how to reduce the number of alternatives for consideration and how to make a robust choice. In this paper we have considered the examination of sensitivity of our choice, in particular using the boundary linear utility.

The example illustrated the use of our methods. We gained a better understanding of the issues which are important in making our choice and greater confidence in our selection of d_1. We saw that d_2 posed the most important challenge to the choice of d_1 and identified node V as the main basis for this challenge.

We believe that, in many difficult decision problems where a range of trade-off specifications must be considered, our methods could lead to the selection of a choice which is, in practical terms, close to optimal everywhere in the range.

References

[1] J. Butler, J. Jia and J. Dyer. Simulation techniques for the sensitivity analysis of multi-criteria decision models. *European Journal of Operational Research*, 103:531–546, 1997.

[2] M. Danielson, L. Ekenberg and A. Larsson. Distribution of expected utility in decision trees. *International Journal of Approximate Reasoning*, 46:387–407, 2007.

[3] M. Farrow and M. Goldstein. Trade-off sensitive experimental design: a multicriterion, decision theoretic, Bayes linear approach. *Journal of Statistical Planning and Inference*, 136:498–526, 2006.

[4] M. Farrow and M. Goldstein. Almost-Pareto decision sets in imprecise utility hierarchies. *Journal of Statistical Theory and Practice*, 3:137–155, 2009.

[5] A. Jiménez, S. Ríos-Insua and A. Mateos. A decision support system for multiattribute utility evaluation based on imprecise assignments. *Decision Support Systems*, 36:65–79, 2003.

[6] R. L. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Trade-offs*, New York: John Wiley & Sons, 1976.

[7] A. Mateos, A. Jiménez and S. Ríos-Insua. Modelling individual and global comparisons for multiattribute preferences. *Journal of Multi-Criteria Decision Analysis*, 12:177–190, 2003.

[8] D. Ríos-Insua and S. French. A framework for sensitivity analysis in discrete multi-objective decision-making. *European Journal of Operational Research*, 54:176–190, 1991.

[9] J. Q. Smith. Statistical principles on graphs. In *Influence Diagrams, Belief Nets and Decision Analysis*, R. M. Oliver and J. Q. Smith (eds), New York: John Wiley & Sons, 89–120, 1990.

Multivariate Models and Confidence Intervals: A Local Random Set Approach

Thomas Fetz
Unit for Engineering Mathematics
University of Innsbruck, Austria
Thomas.Fetz@uibk.ac.at

Abstract

This article is devoted to the propagation of families of confidence intervals obtained by non-parametric methods through multivariate functions comprising the semantics of confidence limits. At fixed confidence level, local random sets are defined whose aggregation admits the calculation of upper probabilities of events. In the multivariate case, a number of ways of combinations is highlighted to encompass independence and unknown interaction using random set independence and Fréchet bounds. For all cases we derive formulas for the corresponding upper probabilities and elaborate how they relate. The methods are exemplified by means of an example from structural mechanics.

Keywords. Confidence intervals, non-parametric models of uncertainty, random sets, fuzzy sets, upper probability, independence, unknown interaction, Fréchet bounds.

1 Introduction

In order to render models of *imprecise probability theory* operative, their *semantics* have to be developed. It has been observed [5, 10, 11] that the idea of *confidence limits* can provide a workable basis for constructing imprecise probability models. In particular, it has been argued in [12, 13] that random sets constructed by Tchebycheff's inequality can serve as a non-parametric model of the variability of a parameter, given its mean value and variance as sole information.

This article develops the concept of using confidence limits for estimating upper and lower probabilities of events. While the papers [5, 12, 13] addressed the univariate case only, it is demonstrated in [10] how to generate joint fuzzy sets from families of marginal confidence intervals using the product t-norm for independence and t-norms based on Fréchet bounds for unknown dependency. In this paper we demonstrate how multivariate input can be treated using a *local random set* approach.

Suppose we are given confidence intervals I_α of some parameter at level α, $0 < \alpha \leq 1$. Then the probability of I_α is bigger than $1 - \alpha$, while the probability of its complement is less than α. The key idea is to define local random sets at level α, formed by I_α and I_α^{c} with weights consistent with the confidence limits. In this way, the upper probability of an event A can be computed as the smallest α for which A lies outside the confidence interval I_α. This procedure gives a conclusive interpretation of upper probabilities in terms of confidence limits.

The plan of this paper is as follows:

In Section 2 families of non-parametric confidence intervals are generated by means of Tchebycheff's inequality.

In Section 3 we introduce the concept of local random sets and its semantics.

In Section 4 it is described how to propagate this kind of uncertainty through univariate functions and it is shown that the local random set approach is consistent with the fuzzy and random set approaches.

In Section 5 we address the multivariate case and generate local joint random sets in various ways consistent with the confidence interpretation. This leads to different estimates for the upper probabilities of events. We derive computational formulas for all cases and show how the results relate to each other and to random set and fuzzy set independence and to the case where nothing is known about how the variables interact.

In Section 6, the method is applied to compute upper distribution functions for the limit state of a beam bedded on two springs where the uncertainty of the spring constants is modelled by families of confidence intervals.

2 Non-parametric models of the variability of a parameter X

In this article we model the variability of a parameter X by a family $\mathbf{I}$ of non-parametric confidence intervals I_α using Tchebycheff's inequality, cf. [5, 13].

Let a random variable X be given with expectation $\mu = \mathsf{E}(X)$ and variance $\sigma^2 = \mathsf{V}(X)$. *Tchebycheff's inequality*

$$P(|X-\mu| > \frac{\sigma}{\sqrt{\alpha}}) \le \alpha, \quad \alpha \in (0,1]$$

leads to *non-parametric confidence intervals*

$$I_\alpha = \left[\mu - \frac{\sigma}{\sqrt{\alpha}}, \mu + \frac{\sigma}{\sqrt{\alpha}}\right], \quad \alpha \in (0,1]$$

for the variability of X at confidence level $1-\alpha$, given its expectation and variance as sole information. This follows from the fact that the complement I_α^{c} of I_α is the set used as the argument of P in Tchebycheff's inequality and by

$$P(I_\alpha^{\mathsf{c}}) \le \alpha, \quad P(I_\alpha) = 1 - P(I_\alpha^{\mathsf{c}}) \ge 1-\alpha. \qquad (1)$$

Then the confidence we have in I_α is $1-\alpha$ or greater. All these confidence intervals together are a family denoted by $\mathbf{I} = \{I_\alpha\}_{\alpha\in(0,1]}$ and they are nested, since $I_\alpha \supseteq I_\beta$ if $\alpha \le \beta$. This property will be also important in the multivariate case later on. A family $\mathbf{I}$ is visualized by plotting in Fig. 1 the interval bounds of I_α, $\alpha \in (0,1]$, at levels α.

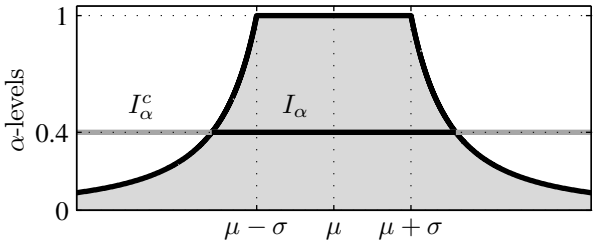


Figure 1: Example of a family $\mathbf{I}$.

3 The univariate case

Let a family $\mathbf{I}$ of confidence intervals I_α, $\alpha \in (0,1]$, generated as in the previous Section be given.

3.1 Local random sets at level α

We assume that $\alpha \in (0,1]$ is fixed. Equipping the two intervals I_α and I_α^{c} with weights $m(I_\alpha)$ and $m(I_\alpha^{\mathsf{c}})$ we get a finite random set. The possible values of these weights are determined by

$$m(I_\alpha) = P(I_\alpha) \;\text{ and }\; m(I_\alpha^{\mathsf{c}}) = P(I_\alpha^{\mathsf{c}})$$

and the inequalities (1) where the weight $m(I_\alpha)$ of I_α corresponds to the confidence we have in the set I_α. We call such a random set corresponding to a certain level α *local random set*.

For an arbitrary event A there are three possibilities for the relations to the two focal sets. These relations and the corresponding upper probabilities $\overline{P}_\alpha$ are shown in the following table:

	Cases	$\overline{P}_\alpha(A) \in$
(i)	$A \cap I_\alpha = \varnothing$	$[0,\alpha]$
(ii)	$A \cap I_\alpha^{\mathsf{c}} = \varnothing$	$[1-\alpha,1]$
(iii)	$A \cap I_\alpha \neq \varnothing$ and $A \cap I_\alpha^{\mathsf{c}} \neq \varnothing$	1

The *local upper probability* $\overline{P}_\alpha(A)$ at level α for an event A is obtained by

$$\overline{P}_\alpha(A) = m(I_\alpha)\,\chi(A \cap I_\alpha \neq \varnothing) + \\ + m(I_\alpha^{\mathsf{c}})\,\chi(A \cap I_\alpha^{\mathsf{c}} \neq \varnothing)$$

where $\chi : \mathbb{R} \to \{0,1\}$ is the indicator function. Here the upper probabilities are intervals because of the inequalities (1) for the weights.

If A has the role of the "bad" and undesired event, case (i) is the most interesting one, because its meaning is:

If A is outside the confidence interval I_α at confidence level $1-\alpha$, then we can say for sure that A occurs only with probability α, at most.

To avoid interval-valued weights and upper probabilities we take always the upper bounds of $\overline{P}_\alpha$ in the above table, that means

$$\overline{P}_\alpha(A) := \begin{cases} \alpha & \text{if } A \cap I_\alpha = \varnothing, \\ 1 & \text{otherwise.} \end{cases}$$

Then we are on the safe side in all three cases.

In general we are not in the interesting case (i) for a given A, but we can try to achieve the situation of case (i) by increasing α. On the other hand, if we are already in case (i), we should try to decrease α to get a smaller upper probability $\overline{P}_\alpha$. This leads to the following rule for the upper probability $\overline{P}(A)$, cf. Fig. 2:

Find the confidence interval I_{α^*} with the smallest $\alpha^* \in (0,1]$ among those confidence intervals I_α with $I_\alpha \cap A = \varnothing$. Then $\overline{P}(A) = \alpha^*$. If we do not find such an interval I_{a^*}, then $\overline{P}(A) = 1$.

With $\inf\{\varnothing\} = 1$ to encompass the case where no I_{a^*} can be found, we get the following formula for the

upper probability:

$$\overline{P}(A) = \inf\{\alpha \in (0,1] :\ I_\alpha \cap A = \varnothing\} = \alpha^*. \qquad (2)$$

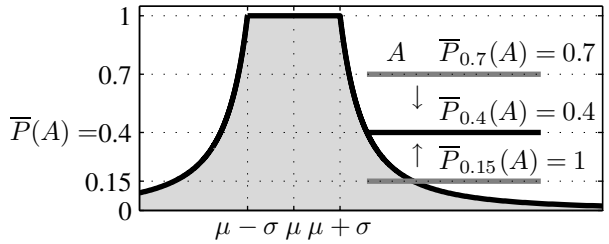


Figure 2: Computation of $\overline{P}(A)$.

3.2 Interpretation of I as a random set and fuzzy set

Together with the uniform distribution on the interval $(0,1]$, the family $\mathbf{I} = \{I_\alpha\}_{\alpha\in(0,1]}$ of confidence intervals is an infinite random set [3, 4, 11]. Note that now all $I_\alpha \in \mathbf{I}$ together play the role of focal sets and not only two sets $I_\alpha, I_\alpha^{\mathsf{c}}$ for fixed α as before. Then for the upper probability $\overline{P}(A)$ (or Plausibility) we get

$$\overline{P}(A) = \mathrm{Pl}(A) = \int_{\beta:\ I_\beta\cap A\neq\varnothing} \mathrm{d}\beta = 1 - \int_{\beta:\ I_\beta\cap A=\varnothing} \mathrm{d}\beta =$$
$$= 1 - \int_{\inf\{\beta:\ I_\beta\cap A=\varnothing\}=\alpha^*}^{1} \mathrm{d}\beta = \alpha^*,$$

because the confidence intervals I_β are nested.

Now we interpret the family $\mathbf{I}$ of nested confidence intervals I_α as fuzzy numbers [14] defined by the α-level sets I_α. The membership function μ is given by the endpoints of the intervals I_α as in Fig. 1. Then the upper probability $\overline{P}(A)$ (or Possibility) is given by

$$\begin{aligned}\overline{P}(A) = \mathrm{Pos}(A) &= \sup\{\mu(x) :\ x \in A\} = \\ &= \sup\{\alpha \in (0,1] :\ I_\alpha \cap A \neq \varnothing\} = \\ &= \inf\{\alpha \in (0,1] :\ I_\alpha \cap A = \varnothing\} = \alpha^*\end{aligned}$$

where $\sup\{\varnothing\} = 0$.

So all three interpretations lead to the same result for the upper probability $\overline{P}(A)$.

4 Propagation of uncertainty trough a univariate function g

4.1 Preliminaries

Let a continuous function

$$g : D \subseteq \mathbb{R} \longrightarrow \mathbb{R} : x \mapsto g(x)$$

and a family $\mathbf{I}$ of confidence intervals I_α be given where we assume that $I_\alpha \subseteq D$ which we achieve simply by truncating I_α if necessary.

Further we are using in the following that

$$\begin{aligned}&P(g(X) \in A) = P(X \in g^{-1}(A)),\\ &I_\alpha \cap g^{-1}(A) = \varnothing \Longleftrightarrow g(I_\alpha) \cap A = \varnothing \text{ and}\\ &I_\alpha \cap g^{-1}(A) \neq \varnothing \Longleftrightarrow g(I_\alpha) \cap A \neq \varnothing\end{aligned}$$

where $g(I_\alpha) = \{g(x) :\ x \in I_\alpha\}$ is the image of I_α under g and $g^{-1}(A) = \{x :\ g(x) \in A\}$ the inverse image of A.

Now we compute $\overline{P}(g(X) \in A)$ for the local random set approach and show that we get the same result as for the random set and for the fuzzy set interpretation.

4.2 Local random set approach

For the local random set approach we have

$$\begin{aligned}\overline{P}(g(X) \in A) &= \overline{P}(X \in g^{-1}(A)) = \\ &= \inf\{\alpha \in (0,1] :\ I_\alpha \cap g^{-1}(A) = \varnothing\} = \\ &= \inf\{\alpha \in (0,1] :\ g(I_\alpha) \cap A = \varnothing\} = \alpha^*.\end{aligned}$$

The only difference to Eq. (2) is that now $g(I_\alpha)$ is used instead of I_α. This motivates the definition

$$g(\mathbf{I}) = \{g(I_\alpha)\}_{\alpha\in(0,1]}$$

which is the family of the images of all confidence intervals. Propagating $\mathbf{I}$ through a function in the univariate case means simply replacing $\mathbf{I}$ by $g(\mathbf{I})$ and applying formula (2), cf. Fig. 3.

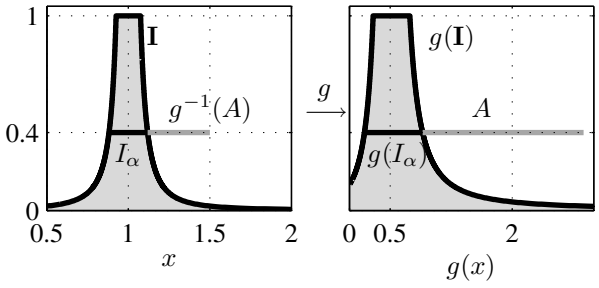


Figure 3: Computation of $\overline{P}(g(X) \in A)$.

4.3 Random set and fuzzy set approach

By the arguments presented in the preliminaries and in Section 3 we get for the random set interpretation

$$\overline{P}(g(X) \in A) = \mathrm{Pl}(g(X) \in A) = \int_{\beta:\ g(I_\beta)\cap A\neq\varnothing} \mathrm{d}\beta =$$
$$= 1 - \int_{\beta:\ g(I_\beta)\cap A=\varnothing} \mathrm{d}\beta = 1 - \int_{\alpha^*}^{1} \mathrm{d}\beta = \alpha^*$$

since again the $g(I_\beta)$ are nested. Further we get for the fuzzy set interpretation of $\mathbf{I}$ with α-level sets I_α:

$$\begin{aligned}\overline{P}(g(X) \in A) &= \mathrm{Pos}(g(X) \in A) = \\ &= \sup\{\alpha \in (0,1] :\ g(I_\alpha) \cap A \neq \varnothing\} \\ &= \inf\{\alpha \in (0,1] :\ g(I_\alpha) \cap A = \varnothing\} = \alpha^*.\end{aligned}$$

4.4 Summary

As we have seen the local random set approach preserves the method of searching for the "best" confidence interval when applying a univariate function g.

More important is, that the result is consistent with the random set and fuzzy set interpretation of the family of confidence intervals. But this is only true in the univariate case. It is a wellknown fact that the random set and the fuzzy set approach lead to different results in the multivariate case which will also have consequences for the local random set version.

5 The multivariate case

Here we assume that for n random variables $X_1, \dots, X_n$ families $\mathbf{I}_1, \dots, \mathbf{I}_n$ of confidence intervals are given. Then we have to determine the joint uncertainty of all these variables which will be done by means of local joint random sets obtained by combining confidence intervals $I_{1,\alpha_1} \in \mathbf{I}_1, \dots, I_{n,\alpha_n} \in \mathbf{I}_n$.

The goal of this and the next Section is to get a formula similar to the univariate version

$$\overline{P}(g(X) \in A) = \inf\{\alpha \in (0,1] :\ g(I_\alpha) \cap A = \varnothing\}.$$

But such a formula will not be uniquely defined because we have several possibilities of choice

- for the set of confidence intervals considered to be combined and
- for the weights used for the local joint random set.

5.1 Combination of marginal confidence intervals

Let the *joint confidence set* $J_{\boldsymbol{\alpha}}$ be given by

$$J_{\boldsymbol{\alpha}} = I_{1,\alpha_1} \times \dots \times I_{n,\alpha_n}$$

with $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)$. Then $\mathbf{J} = \{J_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in S}$ is the family of all joint confidence sets depending on which set S of indices $\boldsymbol{\alpha}$ is considered.

If $S = S_{\mathrm{R}} = (0,1]^n$ then all possible combinations of confidence intervals are used, exactly as the joint focal sets are generated for random set independence.

A second possibility is to combine only confidence intervals of the same level α similar to the combination of the α-level sets for fuzzy set independence. In this case we have the set

$$S = S_{\mathrm{F}} = \{\boldsymbol{\alpha} \in (0,1]^n :\ \alpha_1 = \alpha_2 = \dots = \alpha_n\} \subseteq S_{\mathrm{R}}$$

which has the advantage that the number of joint confidence sets does not grow with the number of variables. For simplification we will then also use the notation

$$\mathbf{J} = \{J_\alpha\}_{\alpha \in [0,1]} = \{I_{1,\alpha} \times \dots \times I_{n,\alpha}\}_{\alpha \in [0,1]}$$

for the family of joint confidence sets.

5.2 Local joint random sets

For two variables X_1 and X_2 the combination of a confidence interval $I_{1,\alpha_1} \in \mathbf{I}_1$ at confidence level $1-\alpha_1$ with a confidence interval $I_{2,\alpha_2} \in \mathbf{I}_2$ at confidence level $1-\alpha_2$ means to generate a local joint random set with focal sets

$$I_{1,\alpha_1} \times I_{2,\alpha_2},\ I^{\mathsf{c}}_{1,\alpha_1} \times I_{2,\alpha_2},\ I_{1,\alpha_1} \times I^{\mathsf{c}}_{2,\alpha_2},\ I^{\mathsf{c}}_{1,\alpha_1} \times I^{\mathsf{c}}_{2,\alpha_2}$$

from the marginal local random set at level α_1 with focal sets I_{1,α_1}, $I^{\mathsf{c}}_{1,\alpha_1}$ for the first variable and from the marginal local random set at level α_2 with focal sets I_{2,α_2}, $I^{\mathsf{c}}_{2,\alpha_2}$ for the second one, cf. Fig. 4. The focal set $J_{\boldsymbol{\alpha}} = I_{1,\alpha_1} \times I_{2,\alpha_2}$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$, is then the joint confidence set.

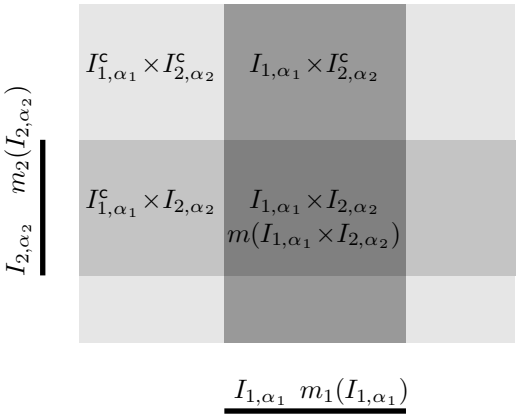


Figure 4: Joint focal sets $I_{1,\alpha_1} \times I_{2,\alpha_2}$, $I^{\mathsf{c}}_{1,\alpha_1} \times I_{2,\alpha_2}$, $I_{1,\alpha_1} \times I^{\mathsf{c}}_{2,\alpha_2}$ and $I^{\mathsf{c}}_{1,\alpha_1} \times I^{\mathsf{c}}_{2,\alpha_2}$.

Computation of the local upper probability $\overline{P}_{\boldsymbol{\alpha}}(A)$:

In the following we do not care about how an event A with empty intersection with the joint confidence set $I_{1,\alpha_1} \times I_{2,\alpha_2}$ hits the remaining three focals sets. We

assume the worst case (hitting all three focals sets), that means

$$\begin{aligned}\overline{P}_{\boldsymbol{\alpha}}(A) &= m(I_{1,\alpha_1} \times I_{2,\alpha_2}^{\mathsf{c}}) + m(I_{1,\alpha_1} \times I_{2,\alpha_2}^{\mathsf{c}}) + \\ &\quad + m(I_{1,\alpha_1}^{\mathsf{c}} \times I_{2,\alpha_2}) \\ &= 1 - m(I_{1,\alpha_1} \times I_{2,\alpha_2}) = \overline{P}((I_{1,\alpha_1} \times I_{2,\alpha_2})^{\mathsf{c}})\end{aligned}$$

which relieves us from computing images of sets where the complements are involved.

For n variables we have then

$$\overline{P}_{\boldsymbol{\alpha}}(A) = 1 - m(I_{1,\alpha_1} \times \cdots \times I_{n,\alpha_n})$$

for $(I_{1,\alpha_1} \times \cdots \times I_{n,\alpha_n}) \cap A = \varnothing$.

5.3 The local joint weight

The main task is to determine the weight $m(I_{1,\alpha_1} \times I_{2,\alpha_2})$ of the joint confidence set $I_{1,\alpha_1} \times I_{2,\alpha_2}$ which represents the confidence we have in this set.

This weight is not uniquely determined, because joint probability distributions are not unique in general. The weights of all four joint focal sets (see Fig. 4) has to be chosen in a way that the horizontal and vertical sums in the following table lead to the marginal weights m_i which are either α_1 or $1-\alpha_1$ for the first variable and either α_2 or $1-\alpha_2$ for the second one:

$m_2(I_{2,\alpha_2}^{\mathsf{c}}) = \alpha_2$	$m(I_{1,\alpha_1} \times I_{2,\alpha_2}^{\mathsf{c}})$	$m(I_{1,\alpha_1}^{\mathsf{c}} \times I_{2,\alpha_2}^{\mathsf{c}})$
$m_2(I_{2,\alpha_2}) = 1-\alpha_2$	$m(I_{1,\alpha_1} \times I_{2,\alpha_2})$	$m(I_{1,\alpha_1}^{\mathsf{c}} \times I_{2,\alpha_2})$
	$m_1(I_{1,\alpha_1}) = 1-\alpha_1$	$m_1(I_{1,\alpha_1}^{\mathsf{c}}) = \alpha_1$

5.3.1 Random set independence

In the case of random set independence the weight of the joint confidence set is given by the product of the marginal weights. For n variables we have then

$$m(I_{1,\alpha_1} \times \cdots \times I_{n,\alpha_n}) = \prod_{i=1}^{n} m_i(I_{i,\alpha_i}) = \prod_{i=1}^{n}(1-\alpha_i)$$

which leads to the local upper probability

$$\overline{P}_{\boldsymbol{\alpha}}(A) = 1 - m(I_{1,\alpha_1} \times \cdots \times I_{n,\alpha_n}) = 1 - \prod_{i=1}^{n}(1-\alpha_i)$$

if $(I_{1,\alpha_1} \times \cdots \times I_{n,\alpha_n}) \cap A = \varnothing$.

If it is known that the uncertain variables are independent, random set independence is one possibility to take the independence of the variables into account. We note that there are other notions of independence such as strong independence and epistemic independence [2, 6, 7, 8].

5.3.2 Lower and upper bounds for the focal weights $m(I_{1,\alpha_1} \times \cdots \times I_{n,\alpha_n})$

Using the bounds of Fréchet [9] for joint probability distributions we get in the 2-dimensional case for the joint weight $m(I_{1,\alpha_1} \times I_{2,\alpha_2})$

$$\begin{aligned}\max(m(I_{1,\alpha_1}) + m(I_{2,\alpha_2}) - 1, 0) &\le m(I_{1,\alpha_1} \times I_{2,\alpha_2}) \le \\ &\le \min(m(I_{1,\alpha_1}), m(I_{2,\alpha_2}))\end{aligned}$$

and with $m(I_{i,\alpha_i}) = 1-\alpha_i$

$$\begin{aligned}\max(1-\alpha_1-\alpha_2, 0) &\le m(I_{1,\alpha_1} \times I_{2,\alpha_2}) \le \\ &\le \min(1-\alpha_1, 1-\alpha_2).\end{aligned}$$

Further using that the local upper probability $\overline{P}_{\boldsymbol{\alpha}}(A) = 1 - m(I_{1,\alpha_1} \times I_{2,\alpha_2})$ for $(I_{1,\alpha_1} \times I_{2,\alpha_2}) \cap A = \varnothing$ leads to lower and upper bounds

$$\max(\alpha_1, \alpha_2) \le \overline{P}_{\boldsymbol{\alpha}}(A) \le \min(\alpha_1 + \alpha_2, 1)$$

for $\overline{P}_{\boldsymbol{\alpha}}(A)$.

With Frechet's version of the inequality for n variables we get then the bounds

$$\max_{i=1,\ldots,n}(\alpha_i) \le \overline{P}_{\boldsymbol{\alpha}}(A) \le \min(\alpha_1 + \cdots + \alpha_n, 1).$$

We use these bounds if nothing is known about how the uncertain variables interact.

5.4 Levels of the joint confidence set

These different approaches have only an influence on the level of the joint confidence sets, but not on the sets itself.

For the three different approaches (random set independence, lower bound and upper bound) we have different levels described by the level function

$$\ell(\boldsymbol{\alpha}) = \begin{cases} \max\limits_{i=1,\ldots,n}(\alpha_i) & \text{lower bound,} \\ 1 - \prod_{i=1}^{n}(1-\alpha_i) & \text{random set independence,} \\ \min(\alpha_1 + \cdots + \alpha_n, 1) & \text{upper bound} \end{cases}$$

which leads to the upper probability

$$\overline{P}_{\ell}^{S}(A) = \inf_{\boldsymbol{\alpha} \in S}\{\ell(\boldsymbol{\alpha}) : J_{\alpha} \cap A = \varnothing\}$$

where the subscript ℓ indicates the level function and the superscript S the set of the $(\alpha_1, \ldots, \alpha_n)$ considered.

5.5 Propagating uncertainty through a multivariate function g

Let a continuous multivariate function

$$g : D \subseteq \mathbb{R}^n \to \mathbb{R} : x \mapsto g(x)$$

be given.

Using the same ideas as in the univariate case we get now the desired formula for the upper probability

$$\begin{aligned}\overline{P}_\ell^S(g(X)\in A) &= \\ &= \inf_{\boldsymbol{\alpha}\in S}\{\ell(\boldsymbol{\alpha}):\ J_{\boldsymbol{\alpha}}\cap g^{-1}(A)=\varnothing\} = \\ &= \inf_{\boldsymbol{\alpha}\in S}\{\ell(\boldsymbol{\alpha}):\ g(J_{\boldsymbol{\alpha}})\cap A=\varnothing\}.\end{aligned}$$

We note that this is the same formula as in the univariate case with the only difference that the level $\ell(\alpha)$ of the resulting interval $g(J_{\boldsymbol{\alpha}}) = g(I_{1,\alpha_1}\times\cdots\times I_{n,\alpha_n})$ may change according to the chosen level function ℓ and that the upper probability depends on the set of confidence intervals considered for combination which is indicated again by ℓ and S.

5.6 Notations

We introduce the following notations for the upper probability $\overline{P}_\ell^S(A) = \inf_{\boldsymbol{\alpha}\in S}\{\ell(\boldsymbol{\alpha}) : J_\alpha\cap A = \varnothing\}$ depending on ℓ and S.

If all possible combinations of confidence intervals are allowed, $S = S_{\mathrm{R}}$, we indicate this by the superscript R:

Notation	level $\ell(\boldsymbol{\alpha})$	
$\overline{P}^{\mathrm{R}}_{\mathrm{lower}}$	$\max\limits_{i=1,\ldots,n}(\alpha_i)$	lower Fréchet bound
$\overline{P}^{\mathrm{R}}_{\mathrm{indep}}$	$1-\prod_{i=1}^n(1-\alpha_i)$	random set independence
$\overline{P}^{\mathrm{R}}_{\mathrm{upper}}$	$\min\left(\sum_{i=1}^n\alpha_i, 1\right)$	upper Fréchet bound

If we consider only combinations of confidence intervals of the same level α, $S = S_{\mathrm{F}}$, we indicate this by the superscript F:

Notation	level $\ell(\alpha)$	
$\overline{P}^{\mathrm{F}}_{\mathrm{lower}}$	α	lower Fréchet bound
$\overline{P}^{\mathrm{F}}_{\mathrm{indep}}$	$1-(1-\alpha)^n$	random set independence
$\overline{P}^{\mathrm{F}}_{\mathrm{upper}}$	$\min(n\alpha, 1)$	upper Fréchet bound

Now we recall the definitions of the upper probabilities for random set independence, fuzzy set independence and unknown interaction in the multivariate case where the notations are given in the following table:

Notation	
$\overline{P}_{\mathrm{R}}$	random set independence
$\overline{P}_{\mathrm{F}}$	fuzzy set independence
$\overline{P}_{\mathrm{U}}$	unknown interaction

The upper probability for random set independence (joint plausibility measure) is defined by

$$\overline{P}_{\mathrm{R}}(A) = \int\limits_{(0,1]^n}\chi(J_{\boldsymbol{\beta}}\cap A\neq\varnothing)\,\mathrm{d}\boldsymbol{\beta}$$

where the $J_{\boldsymbol{\beta}} = I_{1,\beta_1}\times\cdots\times I_{n,\beta_n}$ has the meaning of joint focal sets.

The upper probability for fuzzy set independence (joint possibility measure) is given by

$$\overline{P}_{\mathrm{F}}(A) = \sup\{\alpha\in(0,1]:\ J_\alpha\cap A\neq\varnothing\}$$

where $J_\alpha = I_{1,\alpha}\times\cdots\times I_{n,\alpha}$ are now the joint α-level sets.

In the case where we do not know how the variables are correlated or interact the upper probability for unknown interaction is defined by

$$\overline{P}_{\mathrm{U}}(A) = \sup\{P(A):\ P\in\mathcal{M}_{\mathrm{U}}\}$$

where $\mathcal{M}_{\mathrm{U}}$ is the biggest set of all joint probability measures generated by marginal probability measures compatible with the families of confidence intervals.

For upper distribution functions defined by

$$\overline{F}_\ell^S(x) = \overline{P}_\ell^S((\infty, x])$$

we use the analogous notation as presented in the above tables, e.g. $\overline{F}^R_{\mathrm{indep}}$ is the upper distribution function for $\ell(\boldsymbol{\alpha}) = 1-\prod_{i=1}^n(1-\alpha_i)$ and $S = S_{\mathrm{R}}$.

5.7 The ordering of the upper probabilities

With $\boldsymbol{\beta}\leq\boldsymbol{\alpha}$ defined by $\beta_i\leq\alpha_i$, $i = 1,\ldots,n$, we have the order relation

$$J_{\boldsymbol{\alpha}}\subseteq J_{\boldsymbol{\beta}}\Longleftrightarrow\boldsymbol{\beta}\leq\boldsymbol{\alpha}$$

since all $\mathbf{I}_i$ are families of nested confidence intervals.

Let an event A be given. Then we have always an $\boldsymbol{\alpha}\in S_{\mathrm{R}}$ such that

$$\overline{P}_\ell^{S_{\mathrm{R}}}(A) = \inf_{\boldsymbol{\beta}\in S_{\mathrm{R}}}\{\ell(\boldsymbol{\beta}):\ J_{\boldsymbol{\beta}}\cap A=\varnothing\} = \ell(\boldsymbol{\alpha}),$$

because all level functions ℓ are continuous.

Inspired by a figure in [1] used in a different context we define for above A and $\boldsymbol{\alpha}$ the sets :

$$\begin{aligned}S_{\mathrm{hit}}(A) &= \{\boldsymbol{\alpha}\in S_{\mathrm{R}}:\ J_{\boldsymbol{\alpha}}\cap A\neq\varnothing\},\\ \underline{S}(\boldsymbol{\alpha}) &= \{\boldsymbol{\beta}\in S_{\mathrm{R}}:\ \ell(\boldsymbol{\beta})\leq\ell(\boldsymbol{\alpha}) = \overline{P}_\ell^{S_{\mathrm{R}}}(A)\}\end{aligned}$$

and

$$\overline{S}(\boldsymbol{\alpha}) = (0,1]^n\setminus((\alpha_1,1]\times\cdots\times(\alpha_n,1]),$$

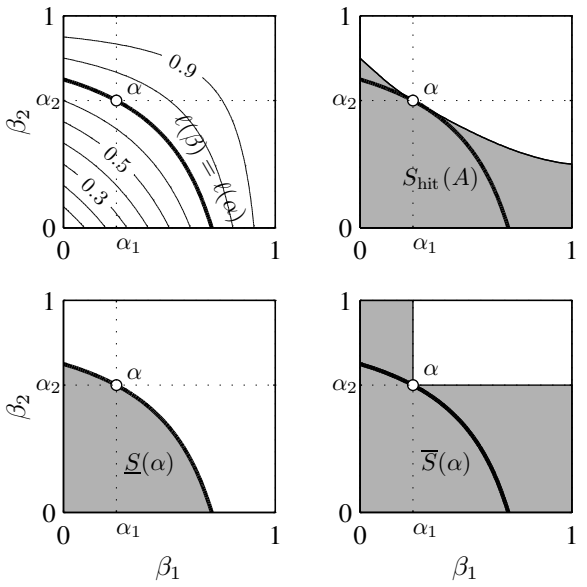


Figure 5: Contourlines of $1-(1-\alpha_1)(1-\alpha_2)$ and the sets $S_{\text{hit}}(A)$, $\underline{S}(\boldsymbol{\alpha})$ and $\overline{S}(\boldsymbol{\alpha})$ for a given $\boldsymbol{\alpha}$.

cf. Fig. 5.

For $\boldsymbol{\alpha} \in S_{\text{hit}}(A)$ the set $S_{\text{hit}}(A)$ has the property

$$\boldsymbol{\beta} \le \boldsymbol{\alpha} \implies \boldsymbol{\beta} \in S_{\text{hit}}(A).$$

Since all level functions are increasing in all directions $\underline{S}(\boldsymbol{\alpha})$ and obviously $\overline{S}(\boldsymbol{\alpha})$ also have this property.

Then we have $\underline{S}(\boldsymbol{\alpha}) \subseteq S_{\text{hit}}(A) \subseteq \overline{S}(\boldsymbol{\alpha})$. See Fig. 5.

Since $S_{\text{F}} \subseteq S_{\text{R}}$ we have always $\overline{P}_\ell^{S_{\text{R}}}(A) \le \overline{P}_\ell^{S_{\text{F}}}$.

5.7.1 $\overline{P}_{\text{R}}(A) \le \overline{P}^{\text{R}}_{\text{indep}}(A) \le \overline{P}^{\text{F}}_{\text{indep}}(A)$

Let $\boldsymbol{\alpha} \in S_{\text{R}}$, such that

$$\overline{P}^{\text{R}}_{\text{indep}}(A) = \ell(\boldsymbol{\alpha}) = 1 - \prod_{i=1}^{n}(1-\alpha_i).$$

Then

$$\overline{P}_{\text{R}}(A) = \int_{S_{\text{hit}}(A)} \mathrm{d}\boldsymbol{\beta} \le \int_{\overline{S}(\boldsymbol{\alpha})} \mathrm{d}\boldsymbol{\beta} = \\ = 1 - \prod_{i=1}^{n}(1-\alpha_i) = \overline{P}^{\text{R}}_{\text{indep}}(A).$$

5.7.2 $\overline{P}_{\text{U}}(A) \le \overline{P}^{\text{R}}_{\text{upper}}(A) \le \overline{P}^{\text{F}}_{\text{upper}}(A)$

Let $p_{[0,1]}$ the probability measure representing the uniform distribution on $[0,1]$, $\mathcal{M}'_{\text{U}}$ the set of all probability measures p on $(0,1]^n$ whose marginals are $p_{[0,1]}$ and $\overline{p}(S) = \sup\{p(S) : p \in \mathcal{M}'_{\text{U}}\}$. Then we have

$$\overline{p}(S_{\text{hit}}(A)) = \overline{P}_{\text{U}}(A) \le \overline{p}(\overline{S}(\boldsymbol{\alpha})).$$

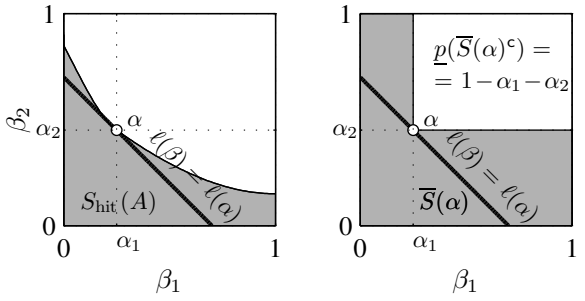


Figure 6: 2-dimensional visualization of the proof in Sec. 5.7.2

The least probability we can concentrate in $\overline{S}(\boldsymbol{\alpha})^{\mathsf{c}}$ is given by

$$\underline{p}(\overline{S}(\boldsymbol{\alpha})^{\mathsf{c}}) = \max\left(\sum_{i=1}^{n} p_{[0,1]}((\alpha_i, 1]) - (n-1), 0\right) = \\ = \max\left(\sum_{i=1}^{n}(1-\alpha_i) - (n-1), 0\right)$$

using the lower Fréchet bound which leads to

$$\overline{p}(\overline{S}(\boldsymbol{\alpha})) = 1 - \underline{p}(\overline{S}(\boldsymbol{\alpha})^{\mathsf{c}}) = \\ = \min\left(\sum_{i=1}^{n}\alpha_i, 1\right) = \overline{P}^{\text{R}}_{\text{upper}}(A),$$

cf. Fig. 6 for the 2-dimensional case.

5.7.3 $\overline{P}^{\text{R}}_{\text{lower}}(A) = \overline{P}^{\text{F}}_{\text{lower}}(A) = \overline{P}_{\text{F}}(A)$

First we show that $\overline{P}^{\text{F}}_{\text{lower}}(A) = \overline{P}_{\text{F}}(A)$:

$$\overline{P}_{\text{F}}(A) = \sup_{\alpha\in(0,1]}\{J_\alpha \cap A \neq \varnothing\} = \\ = \inf_{\alpha\in(0,1]=S_{\text{F}}}\{J_\alpha \cap A = \varnothing\} = \overline{P}^{\text{F}}_{\text{lower}}(A)$$

with $J_\alpha = I_{1,\alpha} \times \cdots \times I_{n,\alpha}$ which is both the joint confidence at level α and the corresponding joint α-level set.

Again let $\boldsymbol{\alpha} \in S_{\text{R}}$, such that

$$\overline{P}^{\text{R}}_{\text{lower}}(A) = \ell(\boldsymbol{\alpha}) = \max_{i=1,\ldots,n}(\alpha_i) =: \alpha.$$

But then also $(\alpha, \ldots, \alpha) \in \underline{S}(\boldsymbol{\alpha})$ and $\overline{P}^{\text{F}}_{\text{lower}}(A) = \alpha$ which proves the first equality.

5.8 The special case $S = S_{\text{F}}$

In the case of $S = S_{\text{F}}$ the joint confidence sets are nested. Let

$$G_\alpha = g(J_\alpha),\ \alpha \in (0,1]$$

be the image of the joint confidence set J_α under g. The index α does correspond only to the case where $\ell(\alpha) = \alpha$. But if we lift the images of the joint confidence sets to the right level by the transformation $H_\alpha = G_{\ell^{-1}(\alpha)}$, $\alpha = 1, \ldots, n$, we get the family

$$\mathbf{H}_\ell = \{H_\alpha\}_{\alpha \in (0,1]}$$

where ℓ indicates the level function used for the transformation. Then the upper probability corresponding to ℓ is simply obtained by

$$\overline{P}_\ell(A) = \inf_{\alpha \in (0,1]} \{H_\alpha \cap A = \varnothing\}$$

as in Section 3 and 4 where no ℓ appears in the formula. In Fig. 7 families $\mathbf{H}_\ell$ are plotted for the three different level functions ℓ presented in this paper.

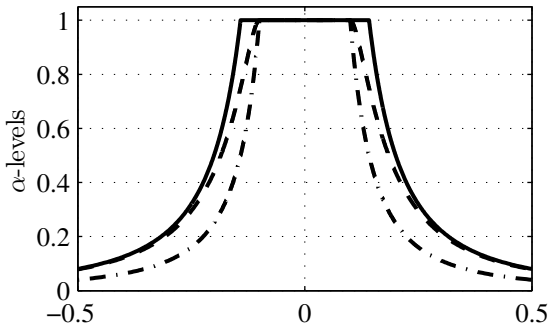


Figure 7: An example of families $\mathbf{H}_\ell$ for ℓ corresponding to the upper bound (solid) and lower bound (dash-dotted) and for random set independence (dashed).

6 A Numerical Example

As a numerical example we consider a beam of length 3 m supported on both ends and additionally bedded on two springs, cf. Fig. 8. The values of the beam rigidity $EI = 1$ kNm^2 and of the equally distributed load $f(x) = 100$ kN/m are assumed to be deterministic, but the values of the two spring constants λ_1 and λ_2 are uncertain.

In this example we assume that the expectations and variances of the two variables λ_1 and λ_2 are given as in the following table.

variable	expectation	variance
λ_1	30	2
λ_2	35	1.5

The corresponding families of confidence intervals generated by means of Tchebycheff's inequality are truncated by the interval $[0, 50]$ and depicted in Fig. 9.

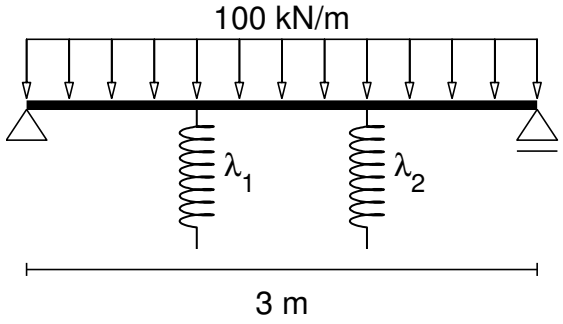


Figure 8: A beam bedded on two springs.

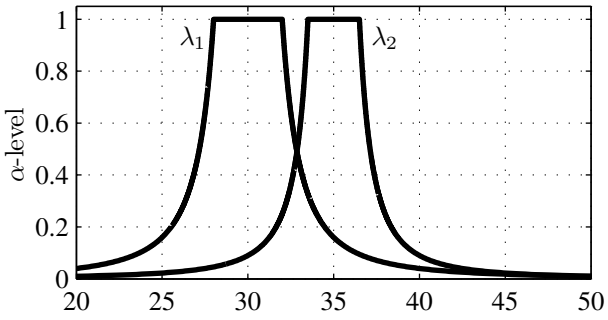


Figure 9: Families of confidence intervals for the two spring constants.

Now we want to compute the upper probability of failure of the beam. The criterion of failure is

$$\max_{x \in [0,3]} |M(x, \lambda_1, \lambda_2)| \geq M_{\mathsf{yield}}$$

where $M(x)$ is the bending moment at point $x \in [0, 3]$ depending on the two spring constants λ_1, λ_2 and $M_{\mathsf{yield}} = 12$ kNm the elastic limit moment. We reformulate the failure criterion as failure function

$$g(\lambda_1, \lambda_2) = M_{\mathsf{yield}} - \max_{x \in [0,3]} |M(x, \lambda_1, \lambda_2)|$$

where now $g(\lambda_1, \lambda_2) \leq 0$ means failure. In Fig. 10 the failure function g is depicted as a contour plot for values $(\lambda_1, \lambda_2) \in [10, 45] \times [10, 45]$ where we can see that g is a concave function in both directions.

Since we want to know if $g(\lambda_1, \lambda_2)$ becomes zero it is sufficient to have only the lower bounds of the images

$$G_{\boldsymbol{\alpha}} = [\underline{G}_{\boldsymbol{\alpha}}, \overline{G}_{\boldsymbol{\alpha}}] = g(J_{\boldsymbol{\alpha}})$$

of the joint confidence sets $J_{\boldsymbol{\alpha}} = \lambda_{1,\alpha_1} \times \lambda_{2,\alpha_2}$. These lower bounds can be easily obtained by minimizing the function values at the vertices of the joint confidence set which is not true for the upper bounds.

The function values $g(\lambda_1, \lambda_2)$ are computed by the finite element method. To omit a large number of function evaluations for a large number of joint confidence

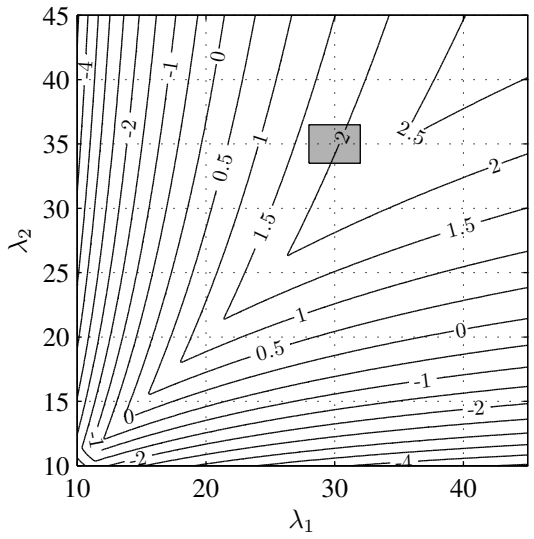


Figure 10: Contour plot of the failure function g. The gray rectangle is the joint confidence set $\lambda_{1,\alpha_1} \times \lambda_{2,\alpha_2}$ for $(\alpha_1, \alpha_2) = (1, 1)$.

sets to be considered we evaluate g on grid points on $[0, 50] \times [0, 50]$ and get $g(\lambda_1, \lambda_2)$ using interpolation.

We get the upper probability distribution functions $\overline{F}_\ell^S$ by

$$\overline{F}_\ell^S(x) = \overline{P}_\ell^S((\infty, x]) = \inf_{\boldsymbol{\alpha} \in S} \{\ell(\boldsymbol{\alpha}) : \underline{G}_{\boldsymbol{\alpha}} > x\}.$$

The results are plotted for $x \in [-0.5, 1.75]$ in Fig. 11. The upper probabilities $\overline{P}_\ell^S((\infty, 0])$ of failure are given by the upper distribution functions at zero.

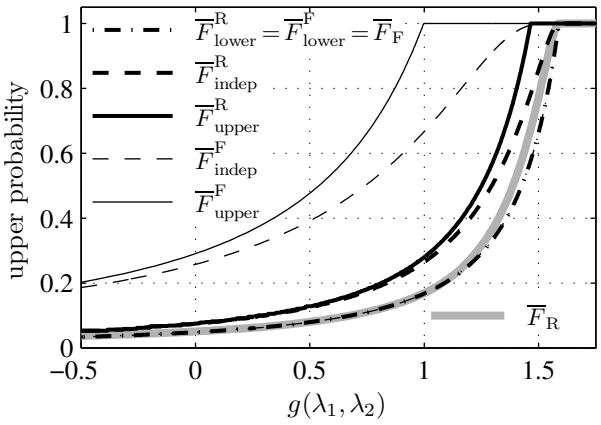


Figure 11: Upper probability distribution functions.

Conclusion

The notion of local random sets was introduced in this article in order to provide a conclusive semantic connection between confidence intervals and random sets. We showed how upper probabilities of events can be calculated from families confidence intervals. The upper probabilities are unique in the univariate case, while in the multivariate case different methods of combinations leading to different upper probabilities are admissible. Further we gave computational formulas for all cases and showed how the resulting upper probabilities are ordered. We demonstrated how the method can be applied in an example from structural mechanics.

References

[1] C. Baudrit and D. Dubois. Comparing methods for joint objective and subjective uncertainty propagation with an example in a risk assessment. In *F. Gagliardi Cozman, R. Nau, T. Seidenfeld (Eds.): ISIPTA '05, Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*, Carnegie Mellon University, Pittsburgh, 2005.

[2] I. Couso, S. Moral, and P. Walley. Examples of independence for imprecise probabilities. In G. de Cooman, G. Cozman, S. Moral, and P. Walley, editors, *Proceedings of the first international symposium on imprecise probabilities and their applications*, pages 121–130, Ghent, 1999. Universiteit Gent.

[3] A.P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.*, 38:325–339, 1967.

[4] A.P. Dempster. Upper and lower probabilities generated by a random closed interval. *Ann. Math. Statistics*, 39:957–966, 1968.

[5] D. Dubois, L. Foulloy, G. Mauris, and H. Prade. Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliable Computing*, 10:273–297, 2004.

[6] Th. Fetz. Sets of joint probability measures generated by weighted marginal focal sets. In *G. de Cooman, T. Fine, T. Seidenfeld (Eds.), ISIPTA'01, Proceedings of the Second Symposium on Imprecise Probabilities and Their Applications*, pages 171–178, Maastricht, 2001. Shaker Publ. BV.

[7] Th. Fetz. Multi-parameter models: rules and computational methods for combining uncertainties. In W. Fellin, H. Lessman, R. Vieider, and M. Oberguggenberger, editors, *Analyzing Uncertainty in Civil Engineering*. Springer, Berlin, 2004.

[8] Th. Fetz and M. Oberguggenberger. Propagation of uncertainty through multivariate functions in the framework of sets of probability measures. *Reliability Engineering and System Safety*, 85(1-3):73–87, 2004.

[9] M. Fréchet. Généralisations du théorème des probabilités totales. *Fundamenta Mathematica*, 25:379–387, 1935.

[10] V. Kreinovich, H. T. Nguyen, S. Ferson, and L. Ginzburg. From computation with guaranteed intervals to computation with confidence intervals: A new application of fuzzy techniques. In *Annual Meeting of the North American Fuzzy Information Processing Society, Proceedings NAFIPS 2002*, pages 418–423, 2002.

[11] H. T. Nguyen. *An Introduction to Random Sets*. Chapman & Hall/CRC, Boca Raton, 2006.

[12] M. Oberguggenberger and W. Fellin. Reliability bounds through random sets: nonparametric methods and geotechnical applications. *Computers & Structures*, 86:1093–1101, 2008.

[13] M. Oberguggenberger, J. King, and B. Schmelzer. Imprecise probability methods for sensitivity analysis in engineering. In G. de Cooman, J. Vejnarova, and M. Zaffalon, editors, *ISIPTA'07, Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications*, pages 317–326, Prague, 2007. Action M Agency, SIPTA.

[14] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

A Minimum Distance Estimator in an Imprecise Probability Model – Computational Aspects and Applications

Robert Hable
University of Bayreuth, Germany
Robert.Hable@uni-bayreuth.de

Abstract

The present article considers estimating a parameter θ in an imprecise probability model $(\overline{P}_\theta)_{\theta\in\Theta}$ which consists of coherent upper previsions $\overline{P}_\theta$. After the definition of a minimum distance estimator in this setup and a summarization of its main properties, the focus lies on applications. It is shown that approximate minimum distances on the discretized sample space can be calculated by linear programming. After a discussion of some computational aspects, the estimator is applied in a simulation study consisting of two different models. Finally, the estimator is applied on a real data set in a linear regression model.

Keywords. Imprecise probabilities, coherent lower previsions, minimum distance estimator, empirical measure, R Project for Statistical Computing.

1 Introduction

1.1 Motivation

In classical statistics, it is common to assume complete knowledge about a statistical model which consists of a (smooth parametric) family $(P_\theta)_{\theta\in\Theta}$ of (precise) probability measures. The task is to make assertions about the true parameter $\theta_0 \in \Theta$. Most often, it is assumed that such assertions can be based on data $x_1, \ldots, x_n$ from random variables which are independent identically distributed according to the true distribution P_{θ_0}. That is, the data analyst already knows that the real distribution P_0 can only be a member of a very special family of probability measures $(P_\theta)_{\theta\in\Theta}$ and the only thing which is not one hundred percent sure is the correct parameter $\theta_0 \in \Theta$. Since this assumption is much to strong for many real applications, generalizations of this probabilistic setup are needed. Suitable generalizations of the concept of probability have been developed, among others, by [12] (coherent lower/upper prevision) and [15] (F-probability). Here, the probability of an event is no longer a number $p \in [0,1]$ but an interval $[\underline{p}, \overline{p}] \subset [0,1]$. In order to generalize the setup of classical statistics to a (more realistic) imprecise probability setup, it is natural to replace the precise model $(P_\theta)_{\theta\in\Theta}$ by an imprecise model $(\overline{P}_\theta)_{\theta\in\Theta}$ which consists of such coherent upper previsions $\overline{P}_\theta$.

The classical frequentist theory of statistics is, in large part, concerned with hypothesis testing (in the sense of Neyman-Pearson) and estimating a parameter. While Neyman-Pearson testing under imprecise probabilities has been extensively studied (cf. e.g. [1] and [2]), estimating a parameter has hardly been considered explicitly within the theory of coherent lower previsions so far. There are a few articles which are concerned with it in Bayesian models (primarily associated with Walley's Imprecise Dirichlet Model), e.g. [13], [9], [7] and [14]. In addition, there are a few articles which address very special applications, e.g. [8] (climate projections) and [3] (prediction of the next influenza pandemic). However, general investigations about frequentist estimation of a parameter using coherent lower/upper previsions are still missing. A first attempt is made in [6] where a minimum distance estimator is developed, and its asymptotic properties are investigated.

The present article focuses on applications of this estimator; for the theoretical validation of the estimator, it is referred to [6]. After a recollection of the definition and the basic properties of the minimum distance estimator in Section 2, Section 3 investigates the concrete calculation of the estimator. At first, the sample space has to be suitable discretized, then the distances between the empirical measure and the coherent upper previsions can be approximately calculated by linear programming. An explicit linear program is developed in Subsection 3.2. The minimum distance estimator is already implemented in the (open source) statistical programming language R; it is publicly available as R-package "imprProbEst" [5]. Subsection 3.3 explains some details about this implementation in R. Next, Section 4 presents a simulation

study where the estimator is applied in two different models and compared to classical estimators. This simulation study exemplifies that the proposed estimator can also be calculated for large sample sizes. This meets objections that, due to high computational costs, imprecise probabilities could not be used for practical purposes. Finally, the minimum distance estimator is applied on a real data set in Section 5. Section 6 contains some concluding remarks.

1.2 Setup and Notation

Let $\mathcal{X}$ be a set with σ-algebra $\mathcal{B}$. Then, $\mathcal{L}_\infty(\mathcal{X},\mathcal{B})$ denotes the set of all bounded, $\mathcal{B}$-measurable real functions $f : \mathcal{X} \to \mathbb{R}$. The supremum norm on $\mathcal{L}_\infty(\mathcal{X},\mathcal{B})$ is denoted by $\|f\| = \sup_{x\in\mathcal{X}} |f(x)|$. The set of all bounded, finitely additive, signed measures is denoted by $\mathrm{ba}(\mathcal{X},\mathcal{B})$ and can be identified with the dual space of $\mathcal{L}_\infty(\mathcal{X},\mathcal{B})$; cf. [4, Theorem IV.5.1]. Finally, $\mathrm{ba}_1^+(\mathcal{X},\mathcal{B})$ denotes the set of all finitely additive probability measures. Integrals with respect to $\mu \in \mathrm{ba}(\mathcal{X},\mathcal{B})$ are denoted by $\mu[f]$.
In accordance with [12, § 2.5.1], a coherent upper prevision on $(\mathcal{X},\mathcal{B})$ is a map

$$\overline{P} \;:\; \mathcal{L}_\infty(\mathcal{X},\mathcal{B}) \;\to\; \mathbb{R}, \qquad f \;\mapsto\; \overline{P}[f]$$

such that there is a (non-empty) set $\mathcal{V} \subset \mathrm{ba}_1^+(\mathcal{X},\mathcal{B})$ and $\overline{P}[f] = \sup_{P\in\mathcal{V}} P[f]$ for every $f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{B})$; cf. also [12, § 3.3.3] and [6, § 2.3]. The non-empty set $\mathcal{M} := \left\{P \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{B}) \;\middle|\; P[f] \le \overline{P}[f] \;\; \forall f\right\}$ is called *credal set of* $\overline{P}$ then.
A coherent upper prevision $\overline{P}$ is called *finitely generated* if there is a finite set $\{f_1,\ldots,f_s\} \subset \mathcal{L}_\infty(\mathcal{X},\mathcal{B})$ such that $\overline{P}$ is the natural extension of a coherent upper prevision on $\{f_1,\ldots,f_s\} \subset \mathcal{L}_\infty(\mathcal{X},\mathcal{B})$. That is, $P \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{B})$ is in the credal set of $\overline{P}$ if and only if $P[f_j] \le \overline{P}[f_j]$ for every $j \in \{1,\ldots,s\}$. Such coherent upper previsions naturally arise in applications whenever a practitioner is only able to specify upper (or lower) bounds on the probability or expectation of a finite number of events or functions respectively. A finitely generated, coherent upper prevision $\overline{P}$ is called *regular* if, in addition, $\overline{P}[f_j] > \underline{P}[f_j] \;\; \forall j \in \{1,\ldots,s\}$ where $\underline{P}$ denotes the coherent lower prevision corresponding to $\overline{P}$; i.e. $\underline{P}[f] = -\overline{P}[-f] = \inf_{P\in\mathcal{M}} P[f]$ for every $f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{B})$.

2 A minimum distance estimator for imprecise models

2.1 Assumptions

In order to state the definition of the minimum distance estimator, the following fixings and assumptions are made. These are valid throughout the rest of the article:

$(\mathcal{X},\mathcal{B})$ is a measurable space and Θ is a finite[1] index set. The data $x_1,\ldots,x_n$ stem from random variables which are independent identically distributed according to a probability measure P_0. For every $\theta \in \Theta$, let $\overline{P}_\theta$ be a coherent upper previsions on $(\mathcal{X},\mathcal{B})$ with credal set $\mathcal{M}_\theta$; $(\overline{P}_\theta)_{\theta\in\Theta}$ is called *imprecise model*. It is only assumed that the true probability measure P_0 is contained in $\mathcal{M}_{\theta_0}$ where $\theta_0 \in \Theta$ is the unknown true parameter. The task is to estimate θ_0.[2]

The following fundamental assumptions on the coherent upper previsions are made:

There is a finite subset $\mathcal{K} = \{f_1,\ldots,f_s\} \subset \mathcal{L}_\infty(\mathcal{X},\mathcal{B})$ such that

$$\mathcal{M}_\theta \;=\; \left\{P_\theta \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{B}) \;\middle|\; P_\theta[f_j] \le \overline{P}_\theta[f_j] \;\; \forall f_j \in \mathcal{K}\right\}$$

for every $\theta \in \Theta$. Furthermore, it is assumed that

$$\overline{P}_\theta[f_j] - \underline{P}_\theta[f_j] \;>\; 0 \qquad \forall f_j \in \mathcal{K} \tag{1}$$

where $\underline{P}_\theta$ is the corresponding lower coherent prevision. In particular, each $\overline{P}_\theta$ is a regular, finitely generated coherent upper previsions.[3]
In the following, it is always assumed that each $f_j \in \mathcal{K}$ is standardized; i.e. $\inf f_j = 0$ and $\sup f_j = 1$. Of course, this is no restriction since every bounded, non-constant function f' can be standardized by

$$f \;:=\; \frac{f' - \inf f'}{\sup f' - \inf f'}$$

and, for every $P_\theta \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{B})$, we have

$$P_\theta[f] \;\le\; \overline{P}_\theta[f] \qquad \Leftrightarrow \qquad P_\theta[f'] \;\le\; \overline{P}_\theta[f']$$

2.2 Definition and asymptotic properties of the minimum distance estimator

The idea of the minimum distance estimator developed in [6, § 6] is very simple: The data $x_1, \ldots, x_n$ are used to build the empirical measure $\mathbb{P}^{(n)}$. Then, the minimum distance estimator is that $\hat{\theta} \in \Theta$ such

[1] Finiteness of the index set is not crucial for the definition and basic properties of the estimator (see [6, § 6]) but the algorithm which calculates the estimator is based on this assumption (see § 3).

[2] This approach corresponds to the use of the type-2 product of coherent upper previsions [12, § 9.3.5]. The type-2 product of coherent upper previsions is consistent with a strict sensitivity analyst's point of view on imprecise probabilities.

[3] Though credal sets may also contain elements P which are not σ-additive, the above assumptions include that P_0 is σ-additive. In case of regular, finitely generated coherent upper previsions, this assumption is justified by [6, Prop. 6.4] which states that these previsions can be represented by sets of (σ-additive) probability measures.

that $\mathbb{P}^{(n)}$ lies next to $\mathcal{M}_{\hat{\theta}}$. That is, we calculate the distance between $\mathbb{P}^{(n)}$ and $\mathcal{M}_\theta$ for every $\theta \in \Theta$ and pick that $\hat{\theta}$ where the distance is minimal.

The empirical measure $\mathbb{P}^{(n)}$ is defined to be the map

$$\mathbb{P}^{(n)} : \ \mathcal{X}^n \ \to \ \mathrm{ba}_1^+(\mathcal{X},\mathcal{B}), \quad x \ \mapsto \ \mathbb{P}_x^{(n)} \ = \ \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$

where $x = (x_1, \ldots, x_n)$ and δ_{x_i} denotes the Dirac measure in $x_i \in \mathcal{X}$. Appropriately to the sensitivity analyst's point of view, the distance between a measure P' and a coherent upper prevision $\overline{P}$ is defined to be

$$\big\|P' - \overline{P}\big\| \ := \ \inf_{P \in \mathcal{M}} \|P' - P\| \tag{2}$$

where $\mathcal{M}$ denotes the credal set of $\overline{P}$ and $\|P' - P\|$ the operator norm

$$\|P' - P\| \ = \ \sup_{f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{B})} \frac{\big|\, P'[f] - P[f] \,\big|}{\|f\|}$$

The *minimum distance estimator* $\hat{\theta}_n$ is defined to be

$$\hat{\theta}_n \ : \ x \ \mapsto \ \arg\min_{\theta \in \Theta} \big\|\mathbb{P}_x^{(n)} - \overline{P}_\theta\big\|$$

Note that the minimizing θ is not always unique; in this case, the minimum distance estimator may pick any minimizing θ.

Now, let us turn over to the asymptotic properties of the minimum distance estimator according to [6, § 6.4]. Firstly, note that the use of the operator norm together with the empirical measure is not unproblematic in classical statistics: Though several distances d provide the desirable property that

$$d\big(\mathbb{P}_x^{(n)}, P_0\big) \quad \xrightarrow[n\to\infty]{} \quad 0 \tag{3}$$

almost surely, this is not necessarily true for the operator norm (e.g. in case of the standard normal distribution). However, this annoying difficulty totally disappears in our imprecise probability setup (Subsections 1.2 and 2.1). If we replace P_0 by a regular, finitely generated coherent upper prevision $\overline{P}$, we get that

$$\big\|\mathbb{P}_x^{(n)} - \overline{P}\big\| \quad \xrightarrow[n\to\infty]{} \quad 0 \tag{4}$$

almost surely if P_0 lies in the credal set $\mathcal{M}$ of $\overline{P}$; cf. [6, Theorem 6.6].

A true parameter θ_0 is any $\theta_0 \in \Theta$ such that

$$P_0 \ \in \ \mathcal{M}_{\theta_0}$$

Since it is not assumed that the credal sets are disjoint, there may be several true parameters.

According to [6, Theorem 6.10], the probability of the event

$$\Big\{ x \in \mathcal{X}^n \ \Big|\ P_0 \notin \mathcal{M}_{\hat{\theta}_n(x)} \Big\} \tag{5}$$

tends to zero for increasing sample size n if the index set Θ is finite.

The mathematically rigorous statements about these asymptotic properties are more involved and have to be formulated in terms of random variables and image measures. This is because the expressions in (4) and (5) will not be measurable in general. For the treatment of unmeasurable maps in asymptotic statistics, confer e.g. [11, §18].

3 Calculation of the minimum distance estimator

3.1 Discretization of the sample space

As seen in the previous section, it is not necessary to discretize the sample space in order to define the minimum distance estimator based on the total variation norm in a sensible way. Since this is not possible for precise probabilities, going over to imprecise probabilities, in a sense, turns out to be a simplification.
Of course, if we want to calculate the estimator by use of computers, the sample space has to be discretized – at least implicitly. However, it is one of the most striking properties of the above presented minimum distance estimator, that this is only a practical need which is irrelevant for theoretical investigations. That is, we can also deal with infinite sample spaces $(\mathcal{X},\mathcal{B})$. In case of precise probabilities, discretization would even be part of the definition of the minimum distance estimator.

Recall our assumptions given in Subsection 2.1. In order to calculate the minimum distance estimator, we have to calculate

$$\big\|\mathbb{P}_x^{(n)} - \overline{P}_\theta\big\| \ = \inf_{P_\theta \in \mathcal{M}_\theta} \sup_{f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{B})} \frac{\big|\, \mathbb{P}_x^{(n)}[f] - P_\theta[f] \,\big|}{\|f\|}$$

for $\theta \in \Theta$. Though $\mathcal{M}_\theta$ is a large subset of $\mathrm{ba}_1^+(\mathcal{X},\mathcal{B})$, these values can nevertheless be calculated with arbitrary accuracy as explained in the following:

At first, fix any accuracy $\varepsilon > 0$. Then, the sample space $(\mathcal{X},\mathcal{B})$ may be discretized as follows:

For $\theta \in \Theta$, let $\mathcal{K}_\theta$ be the smallest subset of $\mathcal{K}$ such that

$$\mathcal{M}_\theta \ = \big\{ P_\theta \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{B}) \,\big|\, P_\theta[f_j] \le \overline{P}_\theta[f_j] \ \forall\, f_j \in \mathcal{K}_\theta \big\}$$

and put $\mathcal{I}_\theta = \{j \in \{1,\ldots,s\} | f_j \in \mathcal{K}_\theta\}$. That is, $\mathcal{K}_\theta = \{f_j \in \mathcal{K} | j \in \mathcal{I}_\theta\}$. Furthermore, put

$$\varepsilon_\theta^{(j)} := \frac{\overline{P}_\theta[f_j] - \underline{P}_\theta[f_j]}{2s} \cdot \varepsilon \qquad \forall\, j \in \mathcal{I}_\theta$$

and choose simple functions $h_\theta^{(j)}$ such that

$$f_j \le h_\theta^{(j)} \le f_j + \varepsilon_\theta^{(j)} \qquad \forall\, j \in \mathcal{I}_\theta \tag{6}$$

Then, let $\mathcal{C}_\theta$ be the smallest σ-algebra on $\mathcal{X}$ such that the simple functions $h_\theta^{(j)}$, $j \in \mathcal{I}_\theta$, are $\mathcal{C}_\theta/\mathbb{B}$-measurable. Note that $\mathcal{C}_\theta$ is a finite subset of $\mathcal{B}$. So, there is a finite partition $\{C_\theta^{(1)},\ldots,C_\theta^{(r)}\}$ of $\mathcal{X}$ such that every event $C \in \mathcal{C}_\theta$ is a (finite) union of elements of the partition $\{C_\theta^{(1)},\ldots,C_\theta^{(r)}\}$.

Now, let $\overline{Q}_\theta$ be the coherent upper prevision on $\mathcal{L}_\infty(\mathcal{X},\mathcal{C}_\theta)$ which corresponds to the credal set

$$\mathcal{N}_\theta = \left\{ Q_\theta \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{C}_\theta) \;\middle|\; \begin{array}{c} Q_\theta[h_\theta^{(j)}] \le \overline{P}_\theta[f_j] + \varepsilon_\theta^{(j)} \\ \forall\, j \in \mathcal{I}_\theta \end{array} \right\}$$

According to [6, Theorem 6.11], we have the following inequalities for every $x \in \mathcal{X}^n$:

$$\left\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\right\| \le \left\|\mathbb{P}_x^{(n)} - \overline{P}_\theta\right\| \le \left\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\right\| + \varepsilon \tag{7}$$

3.2 Approximate calculation of the distance by linear programming

According to (7), it is possible to calculate

$$\left\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\right\| = \inf_{Q_\theta \in \mathcal{N}_\theta} \sup_{f \in \mathcal{L}(\mathcal{X},\mathcal{C}_\theta)} \frac{\left|\mathbb{P}_x^{(n)}[f] - Q_\theta[f]\right|}{\|f\|} \tag{8}$$

in order to approximately calculate $\|\mathbb{P}_x^{(n)} - \overline{P}_\theta\|$, where $\overline{Q}_\theta$ is a coherent upper prevision on the finite space $(\mathcal{X},\mathcal{C}_\theta)$. So, we have to minimize the convex function

$$\mathcal{N}_\theta \to \mathbb{R}, \qquad Q_\theta \mapsto \sup_{f \in \mathcal{L}(\mathcal{X},\mathcal{C}_\theta)} \frac{\left|\mathbb{P}_x^{(n)}[f] - Q_\theta[f]\right|}{\|f\|}$$

Though this is a convex optimization problem, the optimal solution can be found by solving one single linear program.

In order to formulate this linear program, choose any $c_j \in C_\theta^{(j)}$ for every element $C_\theta^{(j)}$ of the partition $\{C_\theta^{(1)},\ldots,C_\theta^{(r)}\}$ of $\mathcal{X}$ which generates $\mathcal{C}_\theta$. Furthermore, put

$$N_j = \left\{ i \in \{1,\ldots,n\} \;\middle|\; x_i \in C_\theta^{(j)} \right\}$$

and let n_j be the number of elements in N_j for every $j \in \{1,\ldots,r\}$. In addition, put

$$\mathcal{J}_0 = \left\{ j \in \{1,\ldots,r\} \;\middle|\; n_j = 0 \right\}$$

and

$$\mathcal{J}_1 = \left\{ j \in \{1,\ldots,r\} \;\middle|\; n_j > 0 \right\}$$

Now, consider the following linear program:

$$\sum_{j \in \mathcal{J}_1} q_j - \gamma_j \longrightarrow \max! \tag{9}$$

where

$$\sum_{j=1}^{r} q_j = 1, \tag{10}$$

$$\sum_{j=1}^{r} q_j h_\theta^{(k)}(c_j) \le \overline{P}_\theta[f_k] + \varepsilon_\theta^{(k)} \quad \forall\, k \in \mathcal{I}_\theta \tag{11}$$

and

$$q_j - \gamma_j \le \frac{n_j}{n} \qquad \forall\, j \in \mathcal{J}_1 \tag{12}$$

for the variables

$$(q_1,\ldots,q_r) \in \mathbb{R}^r, \quad q_j \ge 0 \quad \forall\, j \in \{1,\ldots,r\} \tag{13}$$

and

$$(\gamma_j)_{j \in \mathcal{J}_1} \subset \mathbb{R}, \qquad \gamma_j \ge 0 \quad \forall\, j \in \mathcal{J}_1 \tag{14}$$

Let β_θ be the optimal value of the above linear program. Then, Proposition 3.1 below shows that

$$\left\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\right\| = 2 \cdot \left(1 - \beta_\theta\right) \tag{15}$$

Hence, it is, in fact, enough to solve one single linear program in order to obtain the distance $\left\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\right\|$. Of course, this was useless in applications if this linear program would tend to be unsolvable because of exceedingly high computational costs. So let us take a closer look on the size of the above linear program:

Since the number of elements in $\mathcal{J}_1$ is not larger than $\min\{r,n\}$, we have the following upper bounds:

Number of variables: $r + \min\{r,n\}$

Number of inequalities: $2 + \sharp(\mathcal{K}_\theta) + \min\{r,n\}$

Similar to the discretization method presented in [6, § 5.4] in data-based decision theory, r can – in general – exceed beyond all reasonable bounds but will stay within a reasonable order of magnitude in most applications. In particular, the latter statement is true if the functions $f_j \in \mathcal{K}_\theta$ are convex, concave or indicator functions of (finite unions of) intervals; confer [6, Prop. 5.16]. Though the number n of observations may be very large, it will hardly reach astronomical orders of magnitude in real applications. The size of

the number of elements in $\mathcal{K}_\theta$ (i.e. the number of elements in $\mathcal{I}_\theta$) will usually be negligible.

Note that a very large r will usually result from small values $\varepsilon_\theta^{(j)}$. However, in most real applications, $\overline{P}_\theta$ cannot be specified so accurately that too small values $\varepsilon_\theta^{(j)}$ are meaningful. Furthermore, such small values $\varepsilon_\theta^{(j)}$ indicates that the imprecise model $(\overline{P}_\theta)_{\theta\in\Theta}$ is in danger of being instable – confer [6, § 5.2]. In this case it might be justified to replace $\varepsilon_\theta^{(j)}$ by a larger value. In doing so, we end up with a linear program of a smaller size but, then, it is not guaranteed that $\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\|$ still is an approximation of $\|\mathbb{P}_x^{(n)} - \overline{P}_\theta\|$. However, replacing $\varepsilon_\theta^{(j)}$ by a larger value corresponds to a more conservative proceeding. If this has a large effect on $\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\|$, this means that small changes of $\overline{P}_\theta[f_j]$, $j \in \mathcal{I}_\theta$, have large effects on $\overline{P}_\theta[f]$ for some $f \notin \mathcal{K}_\theta$. In this unstable case, it seems to be a good idea to be more conservative because this may save from arbitrary results.[4]

The following proposition says that $\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\|$ can indeed be calculated by solving the linear program given by (9) – (14):

Proposition 3.1 *Let β_θ be the optimal value of the linear program given by (9) – (14). Then, $\left\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\right\|$ is given by (15).*

Proof:
STEP 1: Firstly, it is shown that, for every $Q \in \mathcal{N}_\theta$,

$$\|\mathbb{P}_x^{(n)} - Q\| = 2\sum_{j\in\mathcal{J}_1}\Big(\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)})\Big)^+ \qquad (16)$$

To this end, fix any $Q \in \mathcal{N}_\theta$ and note that – due to finiteness of $\mathcal{C}_\theta$ – the total variation distance is equal to

$$\|\mathbb{P}_x^{(n)} - Q\| \;=\; \sum_{j=1}^{r}\big|\,\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)})\big| \qquad (17)$$

Since $\{C_\theta^{(1)},\ldots,C_\theta^{(r)}\}$ is a partition of $\mathcal{X}$, we have

$$\begin{aligned} 0 \;&=\; \mathbb{P}_x^{(n)}(\mathcal{X}) - Q(\mathcal{X}) \;=\; \sum_{j=1}^{r}\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)}) \\ &=\; \sum_{j=1}^{r}\Big(\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)})\Big)^+ - \\ &\qquad -\sum_{j=1}^{r}\Big(\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)})\Big)^- \end{aligned}$$

Hence,

$$\begin{aligned} \|\mathbb{P}_x^{(n)} - Q\| \;&\overset{(17)}{=}\; \sum_{j=1}^{r}\big|\,\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)})\big| \\ &=\; 2\cdot\sum_{j=1}^{r}\Big(\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)})\Big)^+ \end{aligned}$$

Note that $\mathbb{P}_x^{(n)}(C_\theta^{(j)}) = 0$ if $j \notin \mathcal{J}_1$ and, therefore,

$$\Big(\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)})\Big)^+ \;=\; 0 \qquad \forall\, j \notin \mathcal{J}_1$$

This proves (16).

STEP 2: Secondly, it is shown that, for every $Q \in \mathcal{N}_\theta$ and every $j \in \mathcal{J}_1$,

$$\begin{aligned} &\Big(\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)})\Big)^+ \;= \\ &\quad=\; \inf_{\gamma_j\in\Gamma_j(Q)}\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)}) + \gamma_j \qquad (18) \end{aligned}$$

where

$$\Gamma_j(Q) := \left\{\gamma_j \in \mathbb{R}\;\middle|\; \begin{array}{c}\gamma_j \geq 0\,,\\ Q(C_\theta^{(j)}) - \gamma_j \;\leq\; \mathbb{P}_x^{(n)}(C_\theta^{(j)})\end{array}\right\}$$

In case of $\mathbb{P}_x^{(n)}(C_\theta^{(j)}) \geq Q(C_\theta^{(j)})$, it is easy to see that the infimum is attained in $\tilde{\gamma}_j = 0 \in \Gamma_j(Q)$ and, therefore, (18) is fulfilled.
In case of $\mathbb{P}_x^{(n)}(C_\theta^{(j)}) < Q(C_\theta^{(j)})$, it is easy to see that the infimum is attained in $\tilde{\gamma}_j = Q(C_\theta^{(j)}) - \mathbb{P}_x^{(n)}(C_\theta^{(j)}) \in \Gamma_j(Q)$ and, therefore, (18) is again fulfilled.

STEP 3: Finally, put

$$\mathbb{M} \;=\; \left\{(Q,\gamma) \in \mathcal{N}_\theta \times \mathbb{R}^{\sharp(\mathcal{J}_1)}\;\middle|\; \begin{array}{c}\gamma = (\gamma_j)_{j\in\mathcal{J}_1}\,,\\ \gamma_j \in \Gamma_j(Q)\;\forall\, j \in \mathcal{J}_1\end{array}\right\}$$

Then, it follows from *STEP 1* and *STEP 2* that

$$\begin{aligned} &\inf_{Q\in\mathcal{N}_\theta}\|\mathbb{P}_x^{(n)} - Q\| \;= \qquad (19)\\ &\quad=\; 2\cdot\inf_{(Q,\gamma)\in\mathbb{M}}\sum_{j\in\mathcal{J}_1}\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)}) + \gamma_j \end{aligned}$$

The definition of $\mathcal{J}_1$ implies $\sum_{j\in\mathcal{J}_1}\mathbb{P}_x^{(n)}(C_\theta^{(j)}) = 1$. Hence,

$$\begin{aligned} &\inf_{Q\in\mathcal{N}_\theta}\|\mathbb{P}_x^{(n)} - Q\| \;= \\ &\qquad\overset{(19)}{=}\; 2\cdot\Big(1 - \sup_{(Q,\gamma)\in\mathbb{M}}\sum_{j\in\mathcal{J}_1}\big(Q(C_\theta^{(j)}) - \gamma_j\big)\Big) \end{aligned}$$

For every $j \in \{1,\ldots,r\}$, identify $Q(C_\theta^{(j)})$ with the variable q_j in the linear program. Then, it follows from the definitions of $\mathcal{N}_\theta$ and $\mathbb{M}$ that

$$\sup_{(Q,\gamma)\in\mathbb{M}}\sum_{j\in\mathcal{J}_1}\big(Q(C_\theta^{(j)}) - \gamma_j\big) \;=\; \beta_\theta$$

and, therefore, $\inf_{Q\in\mathcal{N}_\theta}\|\mathbb{P}_x^{(n)} - Q\| = 2\cdot(1-\beta_\theta)$. □

[4] Confer [6, § 5.2] for more details on the stability of coherent upper previsions and the potential instability of the natural extension.

3.3 Implementation in the statistical programming language R

The minimum distance estimator is implemented in the (open source) statistical programming language R [10] and is publicly available as R-package "imprProbEst" [5]. In order to calculate the estimator, the program has to do the following steps:

1. for "some" $\theta \in \Theta$, (approximately) calculate the distance $\|\mathbb{P}^{(n)} - \overline{Q}_\theta\|$, i.e.
   - discretize the sample space
   - solve the linear program given by (9)- (14)
2. choose that $\hat{\theta}$ which minimizes $\|\mathbb{P}^{(n)} - \overline{Q}_\theta\|$

The inputs are the observations $x = (x_1, \ldots, x_n)$ and the imprecise model given by the (standardized) functions $f_j \in \mathcal{K}_\theta$ and the previsions $\overline{P}_\theta[f_j]$, $f_j \in \mathcal{K}_\theta$, for every $\theta \in \Theta$.

Note that we do not assume any condition of regularity for the map $\theta \mapsto \overline{P}_\theta$. Therefore, one might suppose that we have to calculate $\|\mathbb{P}^{(n)} - \overline{Q}_\theta\|$ for every $\theta \in \Theta$ in order to find the minimizing $\hat{\theta}$. Though this is possible since Θ is assumed to be finite here, such a proceeding is very cumbersome because the calculation of $\|\mathbb{P}^{(n)} - \overline{Q}_\theta\|$ is computationally costly. Fortunately, it usually suffices to calculate $\|\mathbb{P}^{(n)} - \overline{Q}_\theta\|$ only for very few elements of Θ: Put

$$t(\theta) \;=\; 2 \cdot \big(\max_{j \in \mathcal{I}_\theta} \mathbb{P}_x^{(n)}[h_\theta^{(j)}] - \overline{P}_\theta[f_j] - \varepsilon_\theta^{(j)} \big)$$

and $\Theta = \{\theta_1, \ldots, \theta_m\}$. Then, for every $\theta_l \in \Theta$,

$$\begin{aligned} &\|\mathbb{P}^{(n)} - \overline{Q}_{\theta_l}\| \;\geq \\ &\overset{(*)}{\geq} \max_{j \in \mathcal{I}_\theta} \mathbb{P}^{(n)}\big[f_j - (1 - f_j)\big] - \overline{Q}_{\theta_l}\big[f_j - (1 - f_j)\big] \\ &= \; 2 \cdot \big(\max_{j \in \mathcal{I}_\theta} \mathbb{P}^{(n)}[f_j] - \overline{Q}_{\theta_l}[f_j] \big) \overset{(**)}{\geq} t(\theta_l) \end{aligned}$$

where $(*)$ is valid since the standardization of f_j implies $\|f_j - (1 - f_j)\| = 1$, and $(**)$ follows from the definition of $t(\theta_l)$ and (6). Hence, the algorithm only has to calculate the subsequent value $\|\mathbb{P}^{(n)} - \overline{Q}_{\theta_l}\|$ if

$$t(\theta_l) \;\leq\; \min_{k \in \{1, \ldots, l-1\}} \|\mathbb{P}^{(n)} - \overline{Q}_{\theta_k}\| \qquad (20)$$

is fulfilled. If (20) is not fulfilled, we do not have to calculate $\|\mathbb{P}^{(n)} - \overline{Q}_{\theta_l}\|$ because, in this case, it follows from the above calculation that θ_l is already known to be not a minimizer. The simulation studies described in Section 4 showed that, in this way, usually only a very small number of distances $\|\mathbb{P}^{(n)} - \overline{Q}_\theta\|$ has to be calculated.

4 A simulation study

4.1 Model 1: A first example

Model 1 is intended to demonstrate two aspects of the proposed estimator: Firstly, the estimator can really be calculated even for large numbers of observations. In the simulation study, the estimator is applied for sample sizes $n = 30$, $n = 100$, $n = 1000$, $n = 10000$. For each number of observations, the estimator is evaluated 500 times. Secondly, the estimator can provide good results even though it is developed for the rather large imprecise models given by finitely generated coherent upper previsions. In order to demonstrate this, the imprecise Model 1 contains a nice precise parametric model so that the estimator can be compared with a maximum likelihood estimator. While the maximum likelihood estimator is applied by using complete knowledge of the precise parametric model, our minimum distance estimator is only based on the knowledge of a large imprecise model. Since the simulated data exactly stem from the ideal parametric model, this is a rather unequal situation which favors the maximum likelihood estimator and, therefore, the maximum likelihood estimator should clearly beat our estimator. Nevertheless, the performance of our estimator appears to be almost as good as the one of the maximum likelihood estimator in the simulation study. In this way, it can be seen that going over to a large imprecise model does not necessarily mean to loose a lot of efficiency even if the ideal parametric model was precisely true.

Here is a detailed description of Model 1: The sample space is $(\mathcal{X}, \mathcal{B})$ where $\mathcal{X}$ is equal to $[0, 1]$ and $\mathcal{B}$ is the Borel-σ-algebra. The precise parametric model $(P_\theta)_{\theta \in \Theta}$ is given by $dP_\theta = p_\theta \, d\lambda$, $\theta \in \Theta := [-2, 2]$ where the Lebesgue-densities p_θ are

$$p_\theta(x) = 1 + \theta\big(x - 0.5\big) I_{[0,0.5]}(x) + \theta\big(0.75 - x\big) I_{(0.5,1]}(x)$$

for every $x \in [0, 1]$. Despite of this confusing formula, the densities p_θ are very simple and natural as can be seen from Figure 1. In order to define the imprecise model, the parameter set Θ is discretized as follows:

$$\Theta_0 := \big\{ \theta \in \Theta \,\big|\, \theta = -2 + 0.1k - 0.05\,,\; k \in \{1, \ldots, 40\} \big\}$$

That is, $\theta_0 \in \Theta_0$ corresponds to the interval $(\theta_0 - 0.05\,,\, \theta_0 + 0.05]$ with center θ_0. The imprecise model $(\overline{P}_\theta)_{\theta \in \Theta_0}$ is given by credal sets

$$\mathcal{M}_\theta = \big\{ Q_\theta \,\big|\, Q_\theta[f_j] \leq \overline{P}_\theta[f_j] \;\; \forall\, f_j \in \mathcal{K} \big\} \quad \forall\, \theta \in \Theta_0$$

Here, $\mathcal{K}$ is the finite set $\mathcal{K} = \big\{f_1, \ldots, f_{10}\big\}$ which consists of the (rather arbitrarily chosen) functions $f_j : [0, 1] \to \mathbb{R}$, $x \mapsto f_j(x)$ given by

$$f_1(x) \;=\; x\,, \qquad f_2(x) \;=\; 1 - x\,, \qquad f_3(x) \;=\; x^2\,,$$

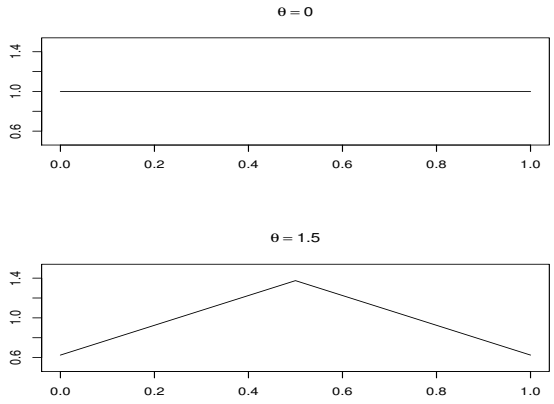


Figure 1: Graphs of p_θ for $\theta = 0$ (the uniform distribution) and $\theta = 1.5$ in Model 1

$$f_4(x) = x^3, \quad f_5(x) = I_{\left[\frac{1}{4},\frac{3}{4}\right]}(x), \quad f_6(x) = I_{\left[0,\frac{1}{4}\right]}(x),$$
$$f_7(x) = I_{\left[\frac{3}{4},1\right]}(x), \qquad f_8(x) = \sqrt{x},$$
$$f_9(x) = x + \tfrac{1}{2} I_{\left[\frac{1}{4},\frac{1}{2}\right]}(x), \qquad f_{10}(x) = 4(x - x^2)$$

and the upper previsions on these functions are defined by

$$\overline{P}_{\theta_0}[f_j] = \sup_{\theta \in [\theta_0 - 0.05\,,\, \theta_0 + 0.05]} \int_0^1 f_j(x) p_\theta(x)\, \lambda(dx)$$

for every $j \in \{1, \ldots, 10\}$ and $\theta_0 \in \Theta_0$.

In the simulation study, the data $x_1, \ldots, x_n$ stem from the uniform distribution $P_0 = \mathrm{Unif}\big([0,1]\big)$. That is, $\theta = 0$ is the true parameter which has to be estimated.

For the estimation, the proposed minimum distance estimator and the maximum likelihood estimator

$$\hat{\theta}_{n,\mathrm{MaxLikelihood}}(x_1, \ldots, x_n) = \arg \max_{\theta \in [-2,2]} \prod_{i=1}^{n} p_\theta(x_i)$$

are applied. Note that – due to the discretization of Θ – our minimum distance estimator does not specify a precise value θ as an estimation but an interval $[\theta_0 - 0.05, \theta_0 + 0.05]$. In order to compare the results between both estimators, these intervals $[\theta_0 - 0.05, \theta_0 + 0.05]$ are recorded by their center θ_0.

Table 1 shows the empirical mean squared error (MSE)

$$\frac{1}{500} \sum_{j=1}^{500} \big(\hat{\theta}_n^{(j)} - 0\big)^2$$

of the estimations $\hat{\theta}_n^{(j)}$ calculated over all runs $j = 1, \ldots, 500$ for the proposed minimum distance estimator (MinDistance) and the classical maximum likelihood estimator (MaxLikelihood); these values are

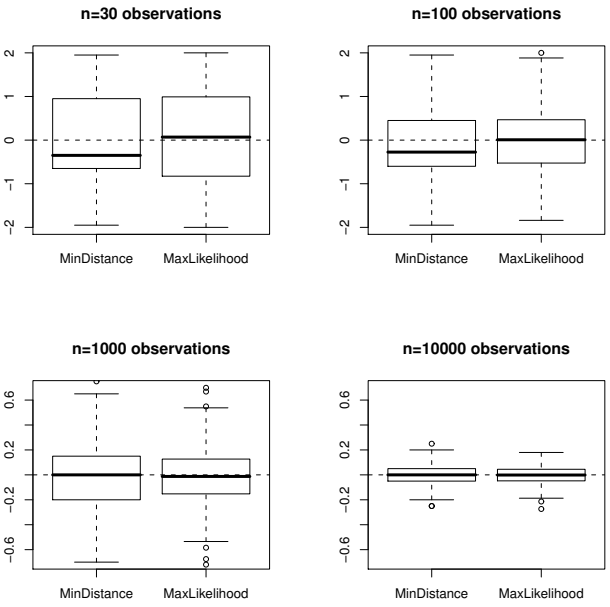


Figure 2: Boxplots of the estimations obtained in 500 runs for each number of observations in Model 1

n	MinDistance	MaxLikelihood
30	1.29943	1.35598
100	0.59675	0.49674
1000	0.06753	0.04692
10000	0.00711	0.00482

Table 1: Empirical mean squared error calculated over the estimations obtained in 500 runs for each number of observations in Model 1

similar for both estimators. Figure 2 shows the boxplots of the estimations. These results demonstrate that, in Model 1, the maximum likelihood estimator is not much better than the minimum distance estimator even though the unequal situation of Model 1 highly privilege the maximum likelihood estimator as explained above.

4.2 Model 2: Approximate Poisson distributions

In Model 2, the sample space is $(\mathbb{N}_0, 2^{\mathbb{N}_0})$ and it is assumed that the data "approximately" stem from a Poisson distribution $\mathrm{Poi}(\theta)$ where the parameter set is $\Theta = (0, 50]$. The parameter set is again discretized:

$$\Theta_0 := \big\{\theta \in \Theta \,\big|\, \theta = 0.1 + 0.05k\,,\ k \in \{0, \ldots, 998\}\big\}$$

The imprecise model $(\overline{P}_\theta)_{\theta \in \Theta_0}$ is given by credal sets

$$\mathcal{M}_\theta = \big\{Q_\theta \mid Q_\theta[f_\theta^{(j)}] \leq \overline{P}_\theta[f_\theta^{(j)}] \ \ \forall\, f_\theta^{(j)} \in \mathcal{K}_\theta\big\}$$

and $\mathcal{K}_\theta$ is the finite set $\mathcal{K}_\theta = \{f_\theta^{(1)}, \ldots, f_\theta^{(56)}\}$ which consists of the following functions:

$$f_\theta^{(j)} \;=\; I_{\{4(j-1),\ldots,4j-1\}} \qquad \forall\, j \in \{1,\ldots,25\}$$

$$f_\theta^{(25+j)} \;=\; 1 - f_\theta^{(j)} \qquad \forall\, j \in \{1,\ldots,25\}$$

$$f_\theta^{(51)}(x) = \frac{x}{100} I_{\{0,\ldots,100\}}(x)\,, \quad f_\theta^{(52)} = 1 - f_\theta^{(51)}$$

$$f_\theta^{(53)}(x) = \Big(\frac{x}{100}\Big)^2 I_{\{0,\ldots,100\}}(x)\,, \quad f_\theta^{(54)} = 1 - f_\theta^{(53)}$$

$$f_\theta^{(55)} = I_{(\theta-1,\theta]}\,, \quad f_\theta^{(56)} = 1 - f_\theta^{(55)}$$

The upper previsions on these functions are defined by

$$\overline{P}_{\theta_0}[f_{\theta_0}^{(j)}] = (1-r) \sup_{\theta \in [\theta_0 - 0.025\,,\, \theta_0 + 0.025]} \mathrm{Pois}(\theta)[f_{\theta_0}^{(j)}] + r$$

for every $j \in \{1,\ldots,56\}$ and $\theta_0 \in \Theta_0$. In the simulation study, we put $r = 0.01$. [5]

For the estimation, our minimum distance estimator and the maximum likelihood estimator

$$(x_1,\ldots,x_n) \;\mapsto\; \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} \mathrm{Pois}(\theta)\big(\{x_i\}\big)\,,$$

are applied. The simulation study consists of 500 runs with different sample sizes $n = 20$, $n = 100$ and $n = 250$. The real distribution which generates the data is equal to

$$P_0 \;=\; (1-c)\mathrm{Pois}(12.5) + c\,\mathrm{Unif}\big(\{0,\ldots,100\}\big)$$

for $c = 0$, $c = 0.01$ and $c = 0.1$ where $c = 0$ is the "ideal situation" and $c \in \{0.01; 0.1\}$ stands for (very) small deviations of the "ideal situation". Figure 3 shows the boxplots for $c = 0$ and $c = 0.01$ (only sample sizes $n = 20$ and $n = 250$); Figure 4 shows the boxplots for $c = 0.1$. Table 2 gives the empirical mean squared errors. In the ideal situation, the maximum likelihood estimator is only slightly better than the (imprecise probability) minimum distance estimator. However, very small deviations from the ideal situation are enough so that the minimum distance estimator beats the maximum likelihood estimator. In particular, this is true even for $c = 0.01$ and $n = 20$ though, in this case, most samples $x_1,\ldots,x_{20}$ will not contain any "wrong" observation – i.e. will be "ideal".

[5] Though this looks very similar to contamination neighborhoods (which are quite common in robust statistics), these upper previsions lead to much bigger credal sets than contamination neighborhoods. This is because, here, the definition of the upper previsions only involves a finite number of functions $f_{\theta_0}^{(j)}$, while the definition of contamination neighborhoods involves all functions $f \in \mathcal{L}_\infty(\mathbb{N}_0, 2^{\mathbb{N}_0})$.

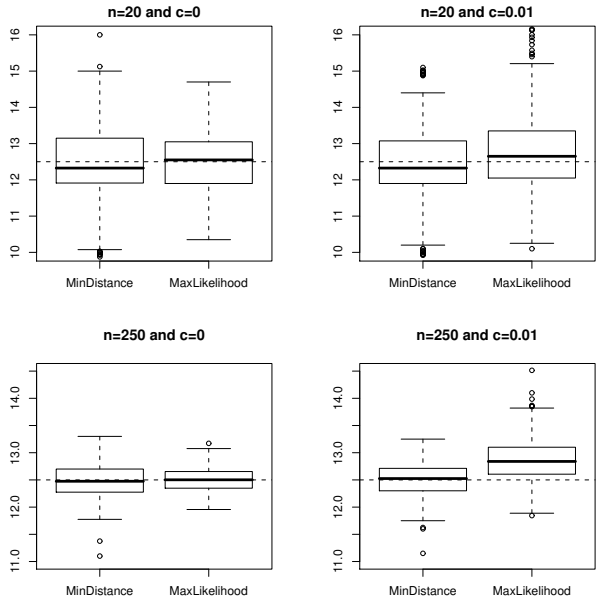


Figure 3: boxplots of the estimations obtained in 500 runs for sample size $n = 20$ and $n = 250$ in Model 2

$n = 20$	c=0	c=0.01	c=0.10
MinDistance	1.22	1.15	1.20
MaxLikelihood	0.65	1.88	24.99
$n = 100$	c=0	c=0.01	c=0.10
MinDistance	0.24	0.29	0.22
MaxLikelihood	0.12	0.52	16.24
$n = 250$	c=0	c=0.01	c=0.10
MinDistance	0.10	0.10	0.12
MaxLikelihood	0.05	0.29	15.27

Table 2: Empirical mean squared error calculated over the estimations obtained in 500 runs in Model 2

5 Application on a real data set

Finally, the estimator is applied on a real data set for linear regression. The data set consists of 200 data

$$x_i = (y_i, z_i) \;\in\; [0,\infty) \times [160,\infty)\,, \quad i \in \{1,\ldots,200\}$$

from the *National Health and Nutrition Examination Survey* (NHANES) from the years 2005–2006 which records the health and nutritional status of adults and children in the United States of America. [6] Every observation x_i corresponds to a person where y_i specifies the person's weight (in kilograms) and z_i specifies the person's height (in centimeters). [7] The following rela-

[6] The data are publicly available in the Internet on the website of the *Centers for Disease Control and Prevention*: `http://www.cdc.gov/nchs/nhanes.htm`

[7] The original data set contains many additional variables which have been omitted here. The 200 persons whose data

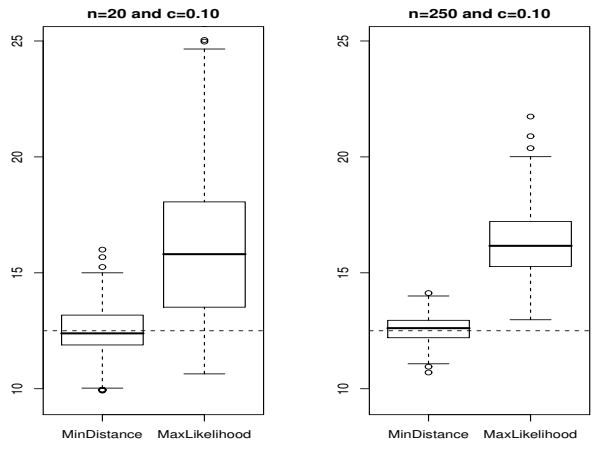


Figure 4: Boxplots of the estimations obtained in 500 runs for sample size $n = 20$ and $n = 250$ in Model 2

tion is assumed

$$y_i \;=\; \theta_1 + \theta_2(z_i - 160) + \varepsilon_i\,, \qquad i \in \{1, \ldots, 200\}$$

for persons with a height of at least 160 cm. Accordingly, only persons have been considered who fulfill this condition. The set of possible parameters is bounded and may be given by $\Theta = [25, 100] \times [0.5, 1.5]$ In order to apply the minimum distance estimator, Θ is again discretized:

$$\Theta_0 \;=\; \left\{ (\theta_1, \theta_2) \;\middle|\; \begin{array}{l} \theta_1 \in \{25, 26, \ldots, 100\}\,, \\ \theta_2 \in \{0.5, 0.55, 0.6, \ldots, 1.45, 1.5\} \end{array} \right\}$$

As an imprecise distribution of the i.i.d errors ε_i, we take the coherent upper prevision $\overline{E}_\sigma$, which is based on the normal distribution $\mathcal{N}(0, \sigma^2)$ in the following way: Take $h_0 = I_{(-\infty, -20]}$,

$$h_1 = I_{(-20, -15]},\;\; h_2 = I_{(-15, -10]},\;\; \ldots\;,\;\; h_{12} = I_{(35\,,\,40]}$$

$$h_{12+j} \;=\; 1 - h_j \qquad \forall\, j \in \{1, \ldots, 12\}$$

and $h_{25} \;=\; I_{(40, \infty)}$. Put $S_0 = \{1, 2, \ldots, 30\}$. The error distribution $\overline{E}_{\sigma_0}$ is assumed to be the coherent upper prevision whose credal set consists of all probability charges E on $\mathbb{R}$ such that for every $j \in \{0, \ldots, 25\}$

$$E[h_j] \;\leq\; (1 - r) \sup_\sigma \mathcal{N}(0, \sigma^2)[h_j] + r \sup h_j I_{(0, \infty)}$$

where the supremum is over $\sigma \in [\sigma_0 - 0.5, \sigma_0 + 0.5]$, $r = 0.05$ and $\sigma_0 \in S$. (Roughly speaking, this means that E is "approximately" a normal distribution but overweight is more likely than underweight. Then, the imprecise model is given by

$$\overline{P}_{\theta_0, \sigma_0} \;=\; \overline{S}_{\sigma_0}[f^{(j)}_{\theta_0}] \qquad \forall\, j \in \{0, \ldots, 25\}$$

are analyzed here have been randomly picked out of the data from the National Health and Nutrition Examination Survey.

	MinDistance	LeastSquares
θ_1	59	67.8
θ_2	0.95	1.03
σ_0	17	—

Table 3: Results of the estimators for the real data set NHANES; the nuisance parameter σ_0 is only estimated by the minimum distance estimator

where $f^{(j)}_{\theta_0} : (y, z) \mapsto h_j\big(y - \theta_1 - \theta_2(z_i - 160)\big)$. The parameter of interest is $\theta_0 = (\theta_1, \theta_2)$; σ_0 is a nuisance parameter.

Our minimum distance estimator is compared to the classical least-squares estimator. The results are given in Table 3, and Figure 5 illustrates the corresponding regression lines. By definition, the least-squares estimator fits the data best with respect to the squared residuals. However, this also leads to the fact that this estimator is sensitive to outliers. This effect is also visible in Figure 5: The least-squares estimator seems to be more influenced by a relatively small number of considerably overweight persons than the minimal distance estimator.

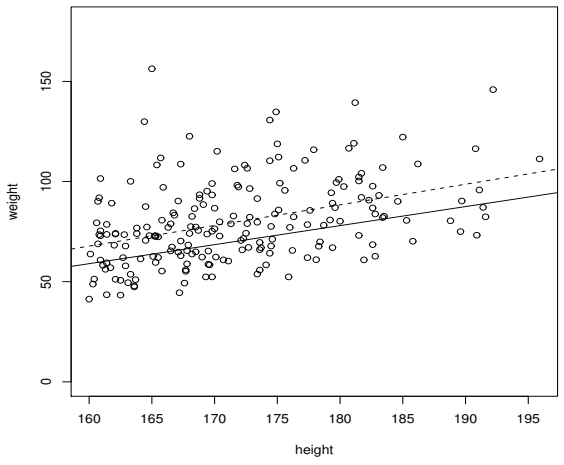


Figure 5: Regression lines for the real data set NHANES obtained by the minimum distance estimator (solid line) and by the least-squares estimator (dashed line)

6 Concluding remarks

The present article considers estimating a parameter in an imprecise probability model – a topic which has hardly been considered explicitly within the theory of coherent upper previsions so far. In this setup, a minimum distance estimator is presented and an algorithm for calculating the estimator is given which is based on

linear programming. The applicability of the estimator is verified by a simulation study and on a real data set. In particular, the simulation study shows that the proposed estimator can even be used for large sample sizes and may, in fact, lead to good results in realistic situations. This meets objections that imprecise probabilities could not be used for practical purposes. The estimator has been programmed in R and has already been made publicly available as (open source) R package "imprProbEst"; cf. [5]. However, future research should also develop alternative estimators so that the proposed minimum distance estimator can be compared to other estimators under imprecise probabilities.

Acknowledgments

I would like to thank Thomas Augustin for valuable suggestions and Matthias Kohl for his help with programming in R. In addition, I thank reviewers for helpful comments.

References

[1] T. Augustin. *Optimale Tests bei Intervallwahrscheinlichkeit*. Vandenhoeck & Ruprecht, Göttingen, 1998.

[2] T. Augustin. Neyman-Pearson testing under interval probability by globally least favorable pairs – reviewing Huber-Strassen theory and extending it to general interval probability. *Journal of Statistical Planning and Inference*, 105:149–173, 2002.

[3] M. Bickis and U. Bickis. Predicting the next pandemic: An exercise in imprecise hazards. In G. de Cooman, J. Vejnarová, and M. Zaffalon, editors, *Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications*, pages 41–46. SIPTA, Prague, 2007.

[4] N. Dunford and J.T. Schwartz. *Linear operators. I. General theory*. Wiley-Interscience Publishers, New York, 1958.

[5] R. Hable. *imprProbEst: Minimum distance estimation in an imprecise probability model*, 2008. Contributed R-Package on CRAN, Version 1.0, 2008-10-23; maintainer Hable, R.

[6] R. Hable. *Data-based decisions under complex uncertainty*. PhD thesis, Ludwig-Maximilians-Universität (LMU) Munich, 2009. http://edoc.ub.uni-muenchen.de/9874/.

[7] M. Hutter. Practical robust estimators for the imprecise Dirichlet model. *International Journal of Approximate Reasoning*, 50:231–242, 2009.

[8] E. Kriegler and H. Held. Climate projections for the 21st century using random sets. In J.M. Bernard, T. Seidenfeld, and M. Zaffalon, editors, *ISIPTA'03, Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications, Lugano*, pages 345–360. Carleton Scientific, Waterloo, 2003.

[9] E. Quaeghebeur and G. de Cooman. Imprecise probability models for inference in exponential families. In F.G. Cozman, R. Nau, and T. Seidenfeld, editors, *ISIPTA'05, Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications, Pittsburg*, pages 287–296. SIPTA, Manno, 2005.

[10] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[11] A.W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, Cambridge, 1998.

[12] P. Walley. *Statistical reasoning with imprecise probabilities*. Chapman & Hall, London, 1991.

[13] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society. Series B. Methodological*, 58(1):3–57, 1996. With discussion and a reply by the author.

[14] G. Walter and T. Augustin. Imprecision and prior-data conflict in generalized Bayesian inference. *Journal of Statistical Theory and Practice: Special Issue on Imprecision*, 3:255–271, 2009.

[15] K. Weichselberger. The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24:149–170, 2000.

How can we get new knowledge?

Frank Hampel
ETH Zurich
Zurich, Switzerland
hampel@stat.math.ethz.ch

Abstract

The paper discusses the (common, important, and yet neglected) situation of a (strong or full) conflict of evidence in scientific and everyday inference (which may lead to valuable new knowledge and even an unexpected scientific breakthrough). It analyses the structure and role of the background knowledge we are using and may have to change, and the many aspects of new information and its interpretation. A number of real life examples follows, which also bring up some more subtle points of inductive thinking.

Keywords. Background knowledge, new information, conflict of evidence, change of model/paradigm, common sense thinking, scientific breakthrough, philosophical foundations of inductive inference, real life examples.

1 Introduction

In observing the various theories using something like upper and lower probabilities, such as in Shafer (1976), Dubois and Prade (1988) and Zadeh (1965), I am still wondering what the precise numerical interpretation of the numbers occurring there is supposed to be, apart from situations with symmetry (cf., e.g., Coolen, 1998) or the start with "total ignorance" which all these theories can deal with (contrary to the neo-Bayesian theory). However, it may well be that only the vague, approximate interpretation of the numerical fixation is relevant, that (like in the Neyman-Pearson theory) in a single situation only values "near" 0 or 1 have a direct practical interpretation, and that perhaps in similar situations the different theories, as far as they are "objective", may lead to somewhat similar values.

A (rare?) example where such a comparison is possible, are enforced fair bets on k independent tosses of a biased coin, starting with total ignorance about the probability of success, evaluated by Smets's pignistic transformation of the Dempster-Shafer belief function theory (Smets, 1990, Smets, 1991, Smets, 1993) and by my own frequentist theory (Hampel, 1993a, Hampel, 1993b, Hampel, 1998, Hampel, 2001; cf. also Hampel, 2002, Hampel, 2005). The enforced probabilities of $(0, \ldots, k)$ successes are for $k = 1$ (1/2, 1/2) (for both and many other theories), for $k = 2$ (5/12, 2/12, 5/12) (Smets) and (1/2, 0, 1/2) (Hampel), for $k = 3$ (157/432 = 0.363, 59/432=0.137, 59/432, 157/432) (Smets) and, for the symmetric solution, (5/12 = 0.417, 1/12 = 0.083, 1/12, 5/12) (Hampel). The numbers are clearly different, but still show some superficial similarity.

Exact numbers may be needed in intermediate calculations (to avoid rounding errors), and they are important in well-developed quantitative theories, where the aims are "only" numerical refinements within a given frame. (The "only" should not be misleading; most research is of this type, and also within a fixed frame there is qualitative progress possible, as by tests.) But when I look at everyday learning and also at scientific breakthroughs, I find that progress often comes by abandoning an old framework or paradigm and replacing it by a new one (cf. also Kuhn, 1962). Such a replacement should obviously be considered when there is a contradiction between the old framework and a new observation.

But in the literature (as far as I know it) I find discussion of this model change conspicuously absent. Only top applied statisticians like John Tukey or Cuthbert Daniel dare to "change the horses in the middle of the stream" (C.D.). The Neyman-Pearson theory is very anxious not to change the assumed model, because then some probabilities would be changed; but these probabilities may have become completely irrelevant. Neo-Bayesians renormalize their posteriors, no matter how small they are without renormalization. For example, depending on circumstances, a single outlier can play havoc with their results. (The (in)famous

dictum by de Finetti: "There is no Bayesian problem of outliers, because there are no outliers" assumes an omniscient, God-like attitude: in real life, we just don't know everything.) Shafer (1976) discusses at length a quantitative measure of the degree of contradiction between two claims; but he does not say what to do when the contradiction is large (except renormalizing) or even complete.

Now one reason for the silence about such a crucial point is probably that it may be (or may seem) impossible to build an exact, numerical mathematical theory about it. However, my impression is that most arguments in real life and many arguments in top science are even on a 0-1-scale (not involving degrees in between), but involving changes of background beliefs; therefore I consider it worthwhile to study the (existing) qualitative (and perhaps even semi-quantitative) structures which we can find when we analyse our corresponding thinking in more detail. Such an attempt is what the paper is about.

This paper is closely related to the short outline in Hampel (2007); cf. also Hampel (2009). After an introductory example, the interplay between background knowledge and new information, the structure of background knowledge in real life, and the interpretation and surrounding structure of new information are discussed; then a number of real life examples are given, including the discussion of some finer points in inductive logic.

While in Shafer (1976) the two (or more) sources of new information are treated symmetrically, keeping the background model fixed, here one of the sources of information is the background model itself which may have to be changed by the new information.

Although not discussed here, the new concepts can easily be applied to the problem of model change in applied statistics or data analysis, using the great experience of top applied statisticians.

We also note that probably quite a few scientific breakthroughs, like the discovery of penicillin by Fleming, or the discovery of the effect of rubella on pregnancy, have their root in an unexpected (and first unexplainable) observation which was then thoroughly analyzed.

One referee kindly alerted me to the danger that my approach might be confused with the (relatively popular) work on "belief revision" as exemplified in the classic paper by Alchourron et al. (1985). But that paper deals merely with the logical consequences if in a complex logical system one statement is being contradicted (or another statement is being added). This is certainly a legitimate topic of research, but it is entirely restricted to deductive logic, working out the (intricate) logical consequences of partial knowledge. By contrast, I am considering the situation that a former belief is *entirely* wrong, and a new belief has to be created on the basis of inductive guesswork (based on "life experience"). It is one of my main points that such arguments cannot be derived by pure deductive logic (except, of course, if one believes to be omniscient, like God or some Bayesians). Nevertheless, I do describe a rich new structure of inductive thinking. And while my paper abounds in real life examples, I cannot find a single real life example in the 20 or so pages of Alchourron et al. (1985). (In addition, I believe that in practice often the contradictory new observation in their paper needs one or more detailed interpretations in order to allow meaningful logical deductions.) Overall, I think that my approach is often much closer to real life problems (and to Kuhn, 1962) than the approach in Alchourron et al. (1985).

2 Oregon and Dolomites

The following story (which may well be more widely known) was told to me in 1984 by the late Philippe Smets.

A couple (perhaps from the US East Coast) was planning their holiday travel. The wife had found a very enticing article about the Dolomites in a travel journal and wanted to go there. But the husband found in the same journal a highly commendatory article about a dry and sunny place in Oregon, and everybody knows that in Oregon it always rains. Later, the husband found out that there are indeed dry and sunny places in Oregon, and both went to the Dolomites.

Let us now analyze this little story in more detail.

The wife got and accepted a "new information" from the travel journal, namely that the Dolomites would be a very nice place to visit. The husband was more sceptical, but since he could not directly judge the article on the Dolomites, he tried to get some more general "information" on the overall reliability and quality of the travel journal, from which to "extrapolate" to the Dolomites article. And he found an article (on some place in Oregon) which was in contradiction with his "background knowledge." Believing his "background knowledge" more than the unknown quality of the journal, he at least cast some doubt on the praise of the Dolomites. (Or perhaps he found the Oregon contradiction just by accident; the end result would be pretty much the same.)

But then he happened to learn (reliably) about the cold desert in the thinly populated (and hence often forgotten) East of Oregon, refining and in this par-

ticular case correcting his "background knowledge"; thus there was no contradiction and no reason to mistrust the journal anymore, and the couple decided to trust the recommendation for the Dolomites. (It could even be argued that a journal talking about dry spots in Oregon is rather sophisticated and not just citing mainstream beliefs and hence trustworthy, once the existence of these dry spots is acknowledged.)

3 Background knowledge and new information

I think this story is an (already somewhat intricate) example of the following general scheme. We all have accumulated, throughout our lifetime up to the present, a large body of "background knowledge" which we use, often subconsciously, to judge our present surroundings. (The structure of our background knowledge is itself very interesting and important, see Section 4 below.) When we now get some "new information" (be it by words, by experiment, or by observation and experience), we compare it with the "extrapolation" from a pertaining part of our background knowledge; if there is a contradiction, then (apart from the chance of later getting new, clarifying information) we have to dismiss either the new information or some part of our background knowledge; or at least we have to "reinterprete" the one or the other or both, in order to make them compatible again.

This is the classical, rational, "scientific" procedure. However, we might also try to live with a contradiction in our "new background knowledge"; and this not only due to irrational or confused thinking, or in fields like religion ("credo quia absurdum"), but also in pure science such as quantum mechanics (like in the saying about the physicist who believes light is a particle on Mondays, Wednesdays, Fridays, and a wave on Tuesdays, Thursdays, Saturdays, and on Sundays he prays).

A rational way of living with a contradiction is to transcend two contradictory claims A and B by not believing either A or B, but by merely noting that both claims exist, without committing oneself. This is possible in pure inference, as opposed to decisions, in view of the necessary action there; but as long as no action is necessary, and even after a necessary action, it makes sense to consider both A and B possible (as can be done in all theories with something like upper and lower probabilities, beliefs, and so on). If later we are reliably told that the chances, or something similar, of A over B are 999:1, then we most strongly keep A, even if before we had made the (apparently bad, on hindsight) decision B. A theory (like the Bayesian one) which would make any decision, however shaky, automatically part of our new background knowledge, would (with a suitable formalism) still keep B; it would weigh internal consistency over time (!) higher than eventual truthfulness.

4 Background knowledge and real life

But where does our background knowledge come from? It has a wide variety of sources: wishful thinking; prejudice; emotions; belief from hearsay, especially from "authorities"; belief from the media, including the internet; a more or less detailed "official" scientific knowledge (which, as experience shows, is mostly fairly stable, but still continuously and sometimes even fundamentally revised, and which contains many bold extrapolations which are hard to judge for the outsider and which may well turn out to be false); personal experiences and extrapolations from these, in combination with the "official" knowledge, and a broad mix of all these sources. Many fundamental beliefs and attitudes go back to our education and even heredity; but we leave this to others to discuss. However the clearer we know the sources of our convictions, the better we can deal with conflicting new information, in judging the relative reliability of both claims.

The background information comes in layers. Usually we take only the most obvious layer or belief for granted, and when the new information is in agreement with it, then this belief will only be somewhat reinforced. But when there is a contradiction between default background and new information, then we have to dig deeper and choose a less likely background as our updated background. Even then, we shall usually consider only interpretations of the next most likely layer (there may be several).

The idea of looking at the set of all possible interpretations of the world and then choosing the most likely one may sound good in philosophy and in pure mathematics, but this is not how we work in real life, nor in science. A physicist will not consider the set of all possible physical theories and then select the most plausible one, once he is forced by experiment to abandon an old theory; but rather he will only look at a "neighborhood" of the old theory and try to get along with as few (or as simple) changes as possible (simple changes may be radical, but only as much as needed by circumstances). If we wanted to consider ALL possibilities of what could happen when we leave the house (like the famous tile dropping from the roof; or being shot to death by mistake, as happened to a wellknown statistician in Mexico City), we would never set a foot in front of the door. This is

a question of efficiency of life. We normally act and think as if only the most likely, or "most plausible" assumptions would be true.

In addition, we may also look at the set of alternatives which are still "quite possible" (like an unexpected delay in something), just to be on the safe side, depending on how pessimistic we are or how strong the consequences would be. But if we observe a contradiction with the "most plausible" assumption, we fully switch to the set of "quite possible" alternatives and perhaps choose the most likely one among them. Only when an occurrence would have drastic consequences (as in cases of life and death), shall we look also at "unlikely" events (and perhaps write a testament or take out an insurance). We hardly ever (except in theory) shall consider "extremely unlikely" interpretations of the world (or even, for logicians and pure mathematicians, "impossible" ones).

These ordered categories: "most plausible", "quite possible", "unlikely", "extremely unlikely" (and perhaps "impossible"), of which we normally only use the first and the second one, to me seem to provide a sufficiently accurate, but also important and necessary valuation of aspects of reality, both in science and in real life. These valuations may differ according to personal experience and present circumstances (cf. the examples below). New experiences may change the category, but usually only to a neighboring one. Cf. also Hampel, 2007.

5 New information

The "new information" in general is not simple and unstructured either. It is connected with "everything that can be said about it", by considering it from its meaning, its sources, its context, its aims, its different possibilities of interpretation, and so on. When faced with new information (and the problem of reconciling it with the background information), and if "So what?" is not the most appropriate reaction (it often is!), the following questions may be helpful:

Who says so?

What is the purpose behind it?

What says the other side? (If controversial)

How does one know this?

What is lacking? (What was forgotten or concealed?)

What does this really mean?

The reliability of the source of information is clearly very important. (I once studied and compared two locally wellknown newspapers for a while. One had a surprisingly large number of – mostly small – inaccuracies. The other, supposed to be very reliable, was so most of the time, but sometimes it contained big blunders – the more misleading as they were unexpected.)

The purpose of news may be a "good story", the fame of a scientist (and the associated money), the need to publish something rather than perish, political influencing, but also neutral information, like the weather report. (Even the weather, and more so the climate, can be political, and even in leading Western countries sometimes scientists have been forbidden to publish their findings.)

The old Roman rule: "Audiatur et altera pars! Listen also to the other side!" is very important in all controversial issues. A comparison of the arguments, motives, backgrounds, reputations, etc. may well allow a decision for one side or the other. Often the truth is somewhere in between; sometimes it is even beyond the range of present opinions.

The question how the new information could have been obtained means going beyond the surface of the information to its possible origins. Sometimes these origins are very subjective and biased, or shaky in other ways.

As is wellknown among statisticians (and still not enough known among nonstatisticians), every statistical number should have with it at least an implicit rough indication of its statistical accuracy. But this is not enough. A good, objective information should also contain a discussion of possible systematic (and semisystematic) errors and their orders of magnitude, of likeliness and effects of gross errors, and of possible reinterpretations of the findings, which might show the results in a completely different light. And often we can only hope that no relevant information has been left out of the discussion. We are reminded of the (in)famous "oath of the statistician": "I swear to tell the truth – nothing but the truth – but not the whole truth." Contrary to deductive logic, conclusions in inductive logic can be changed completely by leaving skillfully out part of the premises.

Often it pays to go a step back and ask oneself: Is this information really what it is supposed to be? Or does it actually mean something noticeably different? Is it only suggestive, and perhaps even without real contents?

A delightful collection of arguments and examples in these directions can be found in the classic book "How to lie with statistics" (Huff 1954); there is also a number of more recent books along similar lines.

6 Some examples of interpretation of new information

A prototype situation is the following: We are living on, without much thinking, in our "most plausible" world, and then (if we are awake and attentive) we observe something strange and surprising (like a Zurich tram in the wrong street, cf. Hampel 2007), which forces us to consider other interpretations of our present surrounding reality (e.g., an accident) which may influence our plans (e.g., requiring us to take another route, or enforcing a delay). Thus, several "quite possible" interpretations are raised to the category "most plausible", until we have learned more. The event observed might even be a "non-event", like suddenly no cars coming from the opposite direction, or the not-barking of the dog in a story about Sherlock Holmes.

The following examples are often from my own experience, especially from ornithology: because I know them best, and because they are sufficiently "unimportant" to allow a neutral discussion. If we discussed "God and the World", which we formally could do equally well, we would soon end up in heated arguments about "God and the World" and not the logical structure of thinking.

6.1 Alarms

When I recently heard a siren wailing at home, I remembered that there were regular test alerts, but I did not remember when. So I looked at the watch; the "round" time (precisely a half-hour) seemed to confirm this interpretation, but to be more sure, I looked also out of the window to see the people on the street walking casually as usual. (The newspaper announcement of the trial alert, which I found later, was somewhat hidden.)

A real alarm under my circumstances fortunately was not very probable, but one never knows. For me, in the beginning both real and test alert were "most plausible" (though the test alert was much more "probable" in the subjectivistic Bayesian sense), and only the two indications (and later the proof) diminished the plausibility of a real alarm.

However, the year before, there was a real alarm in a nearby community because of a pollution of the drinking water. Since it was not too long after the test alert, many people did not pay any attention to it. (In addition, the alarm came only more than five hours after the pollution; and many people were sick for several days.)

Some years ago, during the wars in ex-Yougoslavia, a child from that region had come to Switzerland and went to a Swiss school. When an airplane flew low over the school building, this child immediately dove under a table; and the Swiss classmates had to learn how lucky they were not to be traumatized in this way and not carrying such a background experience with them.

6.2 The meaning of a phrase

The interpretation of a new information may depend strongly on the context. Thus (cf. Hampel 2007), when we ask someone whether a certain way leads to a certain place, in our Western culture a "yes" normally just means "yes" (unless there are or may be reasons that the person answering may want to lead us astray). But experienced world travellers have gained the background knowledge that a "yes" in a different cultural context can mean many different things, for example: 1. "Yes". 2. "Yes, I understand your question." (Perhaps the actual answer comes later.) 3. "Yes, I heard that you said something (without understanding it)." 4. "Yes – you seem to believe so, and I don't want to contradict you." 5. "Yes – I really don't know." 6. "Yes – any other answer would be impolite." (Cf. the story of the East Asian student in Berkeley who finally learnt to say "yes yes" or "yes no", depending on what he really meant, because he was obliged to say always "yes.")

(There is also the true story of the white man who spoke perfectly well Chinese and who asked two old Chinese men whether this was the way to the Ming graves. The two men just stared at him openmouthed; he asked again; the same reaction. Finally he gave up. When he was leaving, he heard one man say to the other: "This sounded just as if he asked whether this was the way to the Ming graves." These men certainly had a strong background conviction.)

More generally, let us assume we learn that a person makes a statement "A". This may mean: 1. A; 2. the opposite of A; 3. approximately A; 4. perhaps A; 5. something related to A; 6. A and B (A incomplete and misleading without B); 7. a polite phrase with no other meaning; 8. an attempt to conceal B, and to divert attention away from B (a frequent trick of tourist guides, if B would be embarassing); 9. an unsubstantiated claim (advertisement); 10. a misunderstanding or a mistake: what was meant was B; 11. A under a side condition forgotten to be mentioned; 12. A under an assumption obvious to the speaker, but unfortunately wrong; 13. A and a seemingly obvious conclusion B, which however is wrong; 14. A and the denial of a conclusion B which is considered too obvious to be true (as pure mathematicians sometimes think); 15. A, but not a fully obvious and

correct conclusion B (which would cause a judge to be called prejudiced and biased; this problem seems to be not uncommon in law), and so on.

(We are also reminded of the joke about the absent-minded professor who says A, writes B, thinks C, means D, and E would have been right.)

6.3 Prejudices I and overreactions

There is often a tendency to cling to old convictions and to defend them by exaggerated means. When I once in fall discovered a Citril Finch (Serinus citrinella) in the Harz mountains in northern Germany, far north of the nearest breeding range in the Black Forest which it hardly ever left, suddenly the Citril Finch was supposed to be a "rather common cage bird" (which it definitely was not, though it was entirely appropriate to consider the possibility of an escaped bird). But when some weeks later I discovered a whole flock of Citril Finches in the same area, opinions switched to the other extreme that some ornithologists believed the bird was even breeding in the Harz. (Compare also the extreme switch of opinions about redescending M-estimators in the Princeton Monte Carlo study, cf. Hampel 1997.)

A rather ridiculous attempt to defend a preconceived attitude by all means once happened in Zurich, when many people saw an "UFO" (a slowly descending chain of lights) in about 10 km distance in a very hazy night. Since some explanation had to be found (to dispel any chance of believing in the little green men), this was officially explained (and believed, even by hobby astronomers) as a chain of burning candles hanging below balloons! As I explained in my farewell lecture, it was nothing but a chain of car headlights descending from an (invisible) mountain lookout.

A very illuminating experiment concerning the strength of false imagination, but also the occurrence of rare exceptions, was once done by an astronomer on British TV (Hunt & Moore 1982, p. 32f). Near inferior conjunction of Venus, he showed the telescopic view of its crescent, whose visibility with the naked eye (under favorable circumstances) is a question of debate among astronomers, and asked the viewers to send in little sketches when they thought they saw the crescent. More than 200 sketches were received; all but two – both by surprised young people – showed the crescent in the inverted view of the telescope. Apparently only these two people genuinely saw the crescent of Venus (as is corroborated by a number of other well-documented observations). Thus, 99% of the claims were illusions, but 1% were proper.

6.4 Layers of questioning

A sceptic (who does not know about the other evidence) might still claim that the two young people could be cheating: they could have known about the inversion of astronomical telescopes and, to make it look more convincing, they might have claimed to be surprised about the right picture (which they did not see). In our case, this appears to be a very far-fetched argumentation, especially since the stake is very low; but in other situations, sceptical digging into deeper layers might well be appropriate.

During the period when I was collecting the bird observations in southernmost Lower Saxony, a young field ornithologist claimed to have seen a female Red-crested Pochard (Netta rufina) on a certain lake, which would have been only the second record of this (mostly very rare) duck for the whole area. I let him describe his observation in a neutral mood and asked him also whether he could see the little red spot at the bill of the female. "Oh yes", he said, "the sun was so bright that it looked as if the whole bill was red." Then I knew two things for certain: that it was a male Red-crested Pochard in eclipse plumage (which, as I knew, has an all-red bill and otherwise looks like a female, but which was not painted even in the best bird book of that time), and (as I had not doubted anyway) that the observation was not made up. (It was also to the credit of the observer that he was very aware of the dangers of light effects.) It can often pay to have some more knowledge or experience than the other person. And what was puzzling for him (the bright red bill), found an explanation and was a proof that the observation was basically correct (apart from the sex).

When I prepared a talk for the European Meeting of Statisticians 1987 in Thessaloniki, I also read something from Aristotle, the genius loci, and was surprised to find that what we consider his "logic" was only a small part of his discussions of a logic in a much broader sense. One of his examples was the story of a very strong man who was accused to have robbed another person during the night. His defense was that he would never have done it, because if he did, he knew that he immediately would be suspected and arrested (being the strongest man around). So it could not have been him.

We can iterate this argument: Since he had such a convincing (?) defence, he could have been the robber, after all. (Cf. the "Theorem" 2 in my talk in Thessaloniki: "The game is indefinite.") Where to put the limit and stop? In general, this may be a difficult problem, with no guidelines except insight into the situation and common sense. In the cases of the cres-

cent of Venus and of the Red-crested Pochard, there clearly was no reason to go further (also because I knew the observer personally in the second case), but when the stakes are high, the question becomes more delicate. The stake might, for example, be the fame of some sort, as in the case of the British ornithologist who shot birds in Asia, imported them frozen to England and layed them out in a small stretch of sea shore where he then obtained "first records" and other remarkable "rare records" of these species for Great Britain (even with "proofs", namely the dead bodies).

He was actually convicted by a statistical argument: there were far too many "rare birds" concentrated on that otherwise rather ordinary piece of sea shore, also compared with the wider surroundings. But even here one has to be careful. While on suitably located islands like Heligoland or the Scilly Islands many rare bird records can be expected, the number of special records around Hildesheim (Lower Saxony), in a very "ordinary" landscape, is at first really amazing. In this case it was due to the sheer fanatic ardor with which the Hildesheim group of ornithologists made their (well-documented) observations; and it showed how little we really know about our surroundings.

6.5 Subtle clues

It is a general experience of mine that unintentional, casual, neutral, often subtle observations or remarks often have the "ring of truth" (as long as this is not used against me on purpose, in the next round of argumentation as described above), while I mistrust all claims with a hidden (or even obvious) purpose behind them. An example of my beginner's time in ornithology is a flock of (very variable) Dunlins (Calidris alpina) in fall; as I counted them back and forth, each time my eye stayed longer with a particularly clean bird (which in a process of "Gestaltwahrnehmung" seemed more and more like a nearby outlier), until I flushed it and could safely identify it as a Curlew Sandpiper (Calidris ferruginea), an uncommon migrant from Siberia.

It may also be that something seems "to be the same and not the same", as when I twice in 3 days observed a Kentish Plover (Charadrius alexandrinus), which is very rare inland. At closer scrutiny one was a male and the other a (distinguishable) female.

Sometimes also "traces of memory" can be helpful for explaining a strange observation.

A very informative clue can be the "Gestalt" of a bird song. Once I woke up by a bird song I had never heard before; it was a Greenish Warbler (Phylloscopus trochiloides), one of the first records in Western Germany, later published (Hampel 1964; Hampel 1965) and corroborated by several other West German records during the same summer. Decades later I was thrilled to hear and recognize the same song again for the second time in the wintering area in India.

Another acoustic observation was more complicated. On May 31, 1985, just before leaving Poland, I heard a new song at Milicz railway station (Silesia) which according to the Swedish bird records appeared to be an Arctic Warbler (Phylloscopus borealis). But the scientists I contacted claimed that the Wood Warbler (Ph. sibilatrix) can have a very similar song. So I spent some summers to check the breadth of variation of Ph. sibilatrix songs. There was some variation, but I never came close to the song in question. (Of course, I cannot exclude that in some areas Ph. sibilatrix can sing almost like Ph. borealis, but I suspect that it was the same reaction as with Serinus citrinella suddenly being a "rather common cage bird.") Meanwhile, I got a record with Mongolian bird songs, including several songs by Ph. borealis, and one sounded exactly like the bird I had heard. My last personal doubts vanished when in the tropical jungle of southwestern China, amidst lots of new songs, I suddenly heard again the Milicz song (and briefly saw the bird).

I got some feedback on my observation of Oct. 20, 1962, of a possibly Phylloscopus schwarzi in Goettingen (Hampel 2007), asking why I did not put it from the category "extremely unlikely" to "possible" if not "plausible", but only to "unlikely" after hearing of the "invasion" in Europe. But in this case I had very little positive evidence for the species. I mainly knew that according to the call it was not Ph. collybita (nor Ph. trochilus), but I could not positively and safely identify the call, not knowing more about this and other similar-looking Asiatic accidentals (and about details of the "invasion"). Nevertheless, with additional information a new assessment might be possible.

6.6 Prejudices II and stability of opinion

It is a hard situation when people are so convinced of their "background knowledge" that they refuse to look at anything nonfitting (like the astronomer and the philosopher in Brecht's "Leben des Galilei" who refused to look through Galilei's telescope with the moons of Jupiter visible, only arguing whether such moons were "possible" and were "necessary"). I once had an experienced ornithologist with me who literally (for a long time) refused to look at a Crane (Grus grus) who was there at a very unusual time of year, because he "knew" it could not be. (Probably he thought I was pulling his leg.)

There is actually a fairly unknown variant of Bayes'

theorem, derived from general principles, with an exponent on one of the two factors. As the exponent varies between zero and infinity, we get all kinds of people from those who are completely stuck in their prejudices, to those who believe everything. (A fitting story is the Sufi story of two persons A and B arguing strongly; a third person C listens to A and says: "You are right." Then he listens to B and says: "You are right." When another person points out to C that A and B cannot be both right, he says: "You are also right.")

It is clear that some medium stability of opinion is needed in the flow of new informations, and science certainly should lean somewhat more to the conservative side. But when a reputable scientist observed, with good documentation, that Lichtenstein's Sandgrouse (Pterocles lichtensteinii), an extreme desert dweller, flew daily up to 80 km to the nearest waterhole, walked into the water until the belly feathers soaked up the water like a sponge, and then flew back to water the young birds in the nest with its belly, this was first ignored and then emphatically denied for 70 years, until it was confirmed also for several other Pterocles species (cf. Scott et al. 1974, p. 153).

Another such story (cf. Barth 1991): the similarity of orchid flowers of the genus Ophrys with several species of sand bees had long been noticed; but when a wellknown specialist observed an actual "copulation" attempt between bee and flower, he first kept it for himself; and when he later wanted to publish it, it was put down as "dirty fantasies of an old man." (By now, there are not only documentary movies, but also fascinating research about the female smell of unpolluted and polluted flowers, as well as a new systematics of the orchids based on the bees.)

6.7 Some tough situations

Very often we have to deal with half-truths ("there may be something to it ...") which are very hard to judge properly.

But one of the worst things that can happen to the pursuit of truth in science is when it is distorted and suppressed by political and religious, commercial and financial interests, as happened again and again. In evaluating new evidence (or even the lack of public evidence), we unfortunately have to take such interests and influences into account.

Acknowledgments: I owe to Werner Stahel, besides technical help, several valuable remarks. – One anonymous referee provided me with an additional reference.

References

Alchourron, C. F., Gardenfors, P. and Makinson, D. (1985). On the logic of theory change: partial meet contraction and revision functions, *The Journal of Symbolic Logic* **50**(2): 510–530.

Barth, F. G. (1991). *Insects and Flowers: The Biology of a Partnership*, Princeton University Press, Princeton, N.J.

Coolen, F. P. A. (1998). Low structure imprecise predictive inference for Bayes' problem, *Statistics and Probability Letters* **36**: 349–357.

Dubois, D. and Prade, H. (1988). *Theory of Possibility*, Plenum, London, UK. *Original Edition in French (1985) Masson, Paris.*

Hampel, F. (1964). Grüner Laubsänger (Phylloscopus trochiloides) in Göttingen, *Journal für Ornithologie* **105**(2): 199.

Hampel, F. (1965). Artenliste vom Seeburger See 1955-1964 (unter knapper Berücksichtigung des Raumes um Göttingen), Mimeographed manuscript, Göttingen, 23 pp (later reprinted).

Hampel, F. (1993a). Some thoughts about the foundations of statistics, *in* S. Morgenthaler, E. Ronchetti and W. A. Stahel (eds), *New Directions in Statistical Data Analysis and Robustness*, Birkhäuser Verlag, Basel, pp. 125–137.

Hampel, F. (1993b). Predictive inference and decisions: Successful bets and enforced fair bets, *Proc. 49th Session of the ISI, Contrib. Papers, Book 1. Firenze (Italy)*, pp. 541–542.

Hampel, F. (1997). Some additional notes on the "Princeton Robustness Year,", *in* D. R. Brillinger, L. T. Fernholz and S. Morgenthaler (eds), *The Practice of Data Analysis: Essays in Honor of John W. Tukey*, Princeton University Press, Princeton, pp. 133–153.

Hampel, F. (1998). On the foundations of statistics: A frequentist approach, *in* M. S. de Miranda and I. Pereira (eds), *Estatística: a diversidade na unidade*, Edições Salamandra, Lda., Lisboa, Portugal, pp. 77–97.

Hampel, F. (2001). An outline of a unifying statistical theory, *in* G. de Cooman, T. L. Fine and T. Seidenfeld (eds), *Proc. of the 2nd Internat. Symp. on Imprecise Probabilities and their Applications, ISIPTA'01, Cornell University, 26-29 June 2001*, Shaker Publishing Maastricht, 2000, pp. 205–212. *Also:* `ftp://ftp.stat.math.ethz.ch/Research-Reports/95.pdf`

Hampel, F. (2002). Some thoughts about classification, *in* H.-H. Bock, K. Jajuga and A. Sokołowski (eds), *Classification, Clustering, and Data Analysis. Recent Advances and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, July 16–19, 2002, Cracow, Poland, Invited keynote lecture, 8th Conference of the International Federation of Classification Societies, Springer, Berlin, pp. 5–26. *Also:* `ftp://ftp.stat.math.ethz.ch/Research-Reports/102.pdf`

Hampel, F. (2005). The proper fiducial argument. Extended abstract, *Electronic Notes in Discrete Mathematics* (Elsevier Science) **21**: 297–300.

Hampel, F. (2007). Upper and lower probabilities in real life, *CD-ROM containing the Proc. 56th Session of the ISI, Contrib. Papers, Lisboa, Portugal*. *Also:* `ftp://ftp.stat.math.ethz.ch/Research-Reports/145.pdf`.

Hampel, F. (2009). Nonadditive probabilities in statistics, *Journal of Statistical Theory and Practice* **3** (No.1: Special Issue on Imprecision): 11–23. *Also:* `ftp://ftp.stat.math.ethz.ch/Research-Reports/146.pdf`

Huff, D. (1954). *How to Lie With Statistics*, Lowe and Brydone, London, UK.

Hunt, G. E. and Moore, P. (1982). *The Planet Venus*, Faber and Faber, London.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*, University of Chicago Press.

Scott, P., Fry, C. H., Flegg, J. J. M., Vevers, G. and Pettingill Jr., O. S. (eds) (1974). *The World Atlas of Birds*, Crescent Books, New York.

Shafer, G. (1976). *A Mathematical Theory of Evidence*, Princeton Univ. Press, Princeton, N. J.

Smets, P. (1990). Constructing the pignistic probability function in a context of uncertainty, *in* M. Henrion, R. D. Shachter, L. N. Kanal and J. F. Lemmer (eds), *Uncertainty in Artificial Intelligence*, Vol. 5, Elsevier Sci. Publ., pp. 29–39.

Smets, P. (1991). The transferable belief model and other interpretations of Dempster-Shafer's model, *in* P. P. Bonissone, M. Henrion, L. N. Kanal and J. F. Lemmer (eds), *Uncertainty in Artificial Intelligence*, Vol. 6, Elsevier Science Publ., pp. 375–383.

Smets, P. (1993). No Dutch book can be built against the TBM even though update is not obtained by Bayes rule of conditioning, *in* R. Scozzafava (ed.), *Workshop on Probabilistic Expert Systems*, Soc. Italiana di Statistica, Roma, pp. 181–204.

Zadeh, L. A. (1965). Fuzzy sets, *Inform. Control* **8**: 338–353.

Dutch Books and Combinatorial Games

Peter Harremoës
Centrum Wiskunde & Informatica (CWI), The Netherlands
P.Harremoes@cwi.nl

Abstract

The theory of combinatorial games (like board games) and the theory of social games (where one looks for Nash equilibria) are normally considered two separate theories. Here we shall see what comes out of combining the ideas. J. Conway observed that there is a one-to-one correspondence between the real numbers and a special type of combinatorial games. Therefore the payoffs of a social games are combinatorial games. Probability theory should be considered a safety net that prevents inconsistent decisions via the Dutch Book Argument. This result can be extended to situations where the payoff function yields a more general game than a real number. The main difference between number-valued payoff and game-valued payoff is that the existence of a probability distribution that gives non-negative mean payoff does not ensure that the game will not be lost.

Keywords. Combinatorial game, Dutch Book Theorem, exchangable sequences, game theory, surreal number.

1 Introduction

The word game in mathematics has two different meanings. The first type of games are the *social games* where a number of agents at the same time have to make a choice and where the payoff to each agent is a function of all agents' choices. Each agent has his own payoff function. The question is how the agents should choose in order to maximize their own payoff. In general the players may benefit by making coalitions against each other. This kind of game theory has found important applications in social sciences and economy. A special class of these social games are the two-person zero-sum games where collaboration between the agents makes no sense.

The second type of games are the *combinatorial games*. These are mathematical models of board games. These games are the ones that people find interesting and amusing. Games that people play for amusement often involve an element of chance, generated by, for instance, dice, but the combinatorial games are by definition the ones that do not contain this element. Therefore they are sometimes called *games of no chance* [15]. Examples from this category are chess, nim, nine-mens-morris, and go. Combinatorial game theory has been particularly successful in the analysis of impartial games like nim [5] and has lead to a better understanding of endgames in go [3, 4, 15].

The Dutch Book Theorem is important in our understanding of imprecise probabilities. The Dutch Book Theorem was first formulated and proved by F. P. Ramsay in 1926 (reprinted in [16]) and later independently by B. de Finetti [8], who used it as an argument for a subjective interpretation of probabilities. Since the original formulation of the Dutch Book Theorem most of the research has been in the direction of more subjective versions. As it is normally formulated, the theorem relies on the concept of a *real-valued payoff* function. One may think of an outcome of the payoff function as money but the uniform mean of having £ 1.000.000 and having £ 0 is having £ 500.000. Most people have a very clear preference for having £ 500.000 rather than an unknown amount of money with mean £ 500.000. Instead one may think of the payoff as a more subjective notion of *value,* but this is also a highly debatable concept and one may actually consider money as our best attempt to quantify value. Savage showed that the concept of value and payoff function can be replaced by the concept of preference, so that a coherent set of preferences corresponds to the existence of a payoff function and a probability measure. This line of research has been followed up by many other researchers [6, 17]. All those studies involve some subjective notion of value or preference.

In order to better understand the Dutch Book Theorem it is desirable to see how the theory would look

in an environment where a subjective notion of value plays no role. In this study we replace the normal payoff functions by game-valued functions. There are several reasons why this is of interest:

- A real-valued payoff function is a special case of a game-valued payoff function.
- The theory of probability has its origin in the study of games involving chance.
- Social game theory and combinatorial game theory may mutually benefit from a closer interaction.
- One can often get insight into a special case by the study of its generalizations.

With a game-valued payoff function the players in a social game have to play a certain combinatorial game that depends on their decisions and/or on some random event. This setup may seem quite contrived, but many board games that involve chance are of this form.

Example 1 *In chess it is normally considered a slight advantage to play white. Therefore one normally randomly selects who should play white and who should play black.*

Example 2 *M. Ettinger has developed an interesting version of combinatorial game theory where after each move a coin is flipped to determine who is going to play next [9].*

Actually any board game involving chance may be considered as an example. It will be the subject of a future paper how to take advantage of a combined probabilistic and combinatorial game approach for some specific board games. In this short note we shall focus entirely on how we should formulate or reformulate the Dutch Book Theorem when the payoffs are combinatorial games.

Social games and combinatorial games are built on quite different ideas and many scientists only know one of the types of game theory. There have only been few attempts to combine the two types of game theory [9, 22]. In this exposition we will assume that the reader has basic knowledge about social games such as two-person zero-sum games. Nevertheless we have to repeat some of the elementary definitions from social game theory in order to fix notation and, in particular, to avoid confusion with similar but slightly different concepts from combinatorial game theory.

Our main result is that it is possible to formulate versions of the Dutch Book Theorem for game-valued payoff functions, but there will be some important modifications of the theorem. For instance our probability distributions will not always be real-valued. In our approach the focus is on order structure (induced by games) and its relation to decision theory. A somewhat orthogonal approach was taken in [13] where the probabilities were elements of a metric space with no order structure.

2 Combinatorial games

The theory of combinatorial games was developed by J. Conway as a tool to analyze board games [5, 7]. A short and more careful exposition can be found in [18]. In a board game the players *alternate* in making moves. Each move changes the configuration of the pieces on the board to some other configuration but only certain changes are allowed. It is convenient to call the two players *Left* and *Right*. We shall often consider different board configurations as different games. If G denotes a game, i.e. a certain configuration then the game is specified by the configurations G^L that Left is allow to move to and the configurations G^R that Right is allowed to move to, and we write $G = \left\{G^L \mid G^R\right\}$. Note that we have not told who is playing first, and therefore we have to describe it from both Left's and Right's perspective. Now the point is that G^L and G^R are sets of games, so a game is formally a specification of two sets of games. In a board game it is nice to have many options to choose among and bad if there are only few options. The worst case for Left is if there are *no options left* and in this case we say that Left has *lost* the game. So Left has lost the game if he is to move next and G^L is empty. Similar Right loses the game if it is Right to move and G^R is empty. The rules of many board games can be modelled in this way.

Example 3 (Games illustrated in Figure 1.) *The game* $\{\varnothing \mid \varnothing\}$ *is a boring one. The one to move first loses this game. This game is denoted* 0.

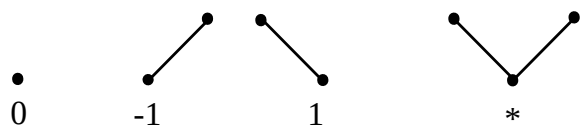


Figure 1: Games can be illustrated by game trees. Options for Left are illustrated by left slanting edges, and options for Right are illustrated by right slanting edges. Here are the simplest ones. In more complicted games there may be several left or right slanting edges from each node.

The game $\{\varnothing \mid 0\}$ *is lost by Left if Left has to move first. If Right goes first Right has to choose* 0*. Now it is Left to move but this is a losing position for the one who is going to move, so poor Left loses. Thus Right always wins the game* $\{\varnothing \mid 0\}$*. This game is denoted* -1.

The game $\{0 \mid \varnothing\}$ *is lost by Right if Right has to move first. If Left goes first Left has to choose* 0*. Now it is Right to move but this is a losing position for the one who is going to move, so now Left is happy again because he wins. Thus Left always wins the game* $\{0 \mid \varnothing\}$*. This game is denoted* 1.

Similarly we see that $\{0 \mid 0\}$ *is won by the player that moves first. This game is called* star *and is denoted* $*$*. In Japanese go terminology such a position is called* dame.

Here we shall use the following recursive definition of a game.

Definition 1 *A* game *is a pair* $\{G^L \mid G^R\}$ *where* G^L *and* G^R *are sets of already defined games.*

The *status* of a game G can be classified according to who wins if both players play optimally. We define

$$\begin{array}{ll} G = 0, & \text{if second player wins;} \\ G < 0, & \text{if Right wins whoever plays first;} \\ G > 0, & \text{if Left wins whoever plays first;} \\ G \parallel 0, & \text{if first player wins.} \end{array}$$

For a game G we can reverse the role of Left and Right and call this the *negative of the game*. Formally we use the following recursive definition.

$$-\{G^L \mid G^R\} = \{-G^R \mid -G^L\}.$$

Left and Right can play two games in parallel. In every round each player should make a move in one of the games of his own choice. Perhaps there are urgent moves to be made in both games so the players have to prioritize in which game it is most important to make the move. Several games played in parallel is called the *sum of the games*, and many positions in actual board games can be understood as sums of sub-games. Combinatorial game theory is essentially the theory of how to prioritize your moves in a board game that has the structure of a sum of independent sub-games. Formally the sum of the games G and H is defined recursively by

$$G + H = \left\{\left(G^L + H\right) \cup \left(G + H^L\right) \mid \left(G^R + H\right) \cup \left(G + H^R\right)\right\}.$$

The sum of games is normally illustrated by the disjoint union of the game trees of the individual games. The game $G - H$ is by definition the game $G + (-H)$.

Now, we are able to define what it should mean that two games are equal. We write $G = H$ if $G - H = 0$, i.e. second player wins $G - H$. One can define $G > H$, $G < H$, and $G \parallel H$ in the same way. We say that G and H are *confused* if $G \parallel H$. One can prove that $G = H$ if and only if $G + K$ and $H + K$ have the same status for any game K.

With these operations the class of games has the structure as a partially ordered Abelian group. Any Abelian group is a module over the ring of integers with multiplication defined as follows. If n is a natural number we define $n \cdot G$ by

$$\overbrace{G + G + \cdots + G}^{n \text{ times}}.$$

If $n = 0$ then $0 \cdot G$ is by definition equal to 0. If n is a negative integer we define $n \cdot G$ to be equal to $(-n) \cdot (-G)$.

The equation $2 \cdot G = 0$ has $G = 0$ as solution, but $G = *$ is also a solution. Therefore there is in general no unique way of defining multiplication of a game by $1/2$, and the same holds for other non-integers. From this point of view it is surprising that all dyadic fractions (rational numbers of the form $n/2^m$) can be identified with games. One way of doing it goes as follows.

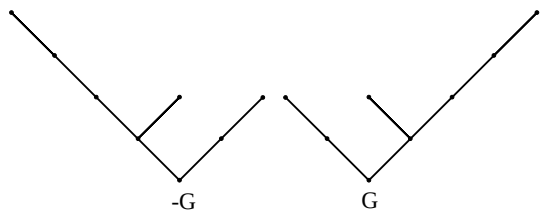


Figure 2: The game tree of $-G$ is simply the mirror image of the game tree of G.

3 Numbers may be identified with games

J. Conway discovered that all real numbers can be identified with games but his construction will lead to a larger class of numbers called the *surreal numbers* (or Conway numbers). The surreal numbers were first described in a mathematical novel by D. Knuth [14], and later in much detail by J. Conway [7]. For newer and more complete descriptions we refer to [1, 11].

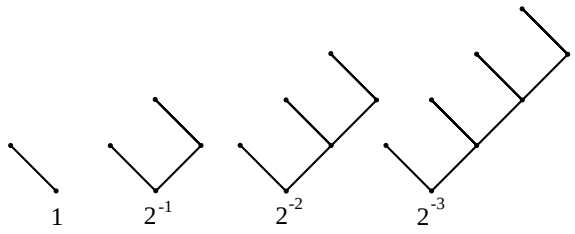


Figure 3: Some dyadic fractions.

We have already defined the game 1 so the integer n is identified with the game $n \cdot 1$. The game $\{0 \mid 1\}$ satisfies

$$2 \cdot \{0 \mid 1\} = 1.$$

Hence the 2^{-1} can be identified with the game $\{0 \mid 1\}$. In general the game $\{0 \mid 2^{-m}\}$ satisfies

$$2 \cdot \left\{0 \mid 2^{-m}\right\} = 2^{-m}$$

so the fraction $2^{-(m+1)}$ can be identified with the game $\{0 \mid 2^{-m}\}$ (see Figure 3). Thus the fraction $n/2^m$ can be identified with the game $n \cdot 2^{-m}$. In this way any dyadic fraction can be identified with a game.

A real number can be identified with a Dedekind section in the group of dyadic fractions. In other words, a real number r, can be identified with the partition of the dyadic fractions into the sets

$$A = \{n \cdot 1/2^m < r \mid m, n \in \mathbb{N}\},$$
$$B = \{n \cdot 1/2^m > r \mid m, n \in \mathbb{N}\}.$$

Now, A and B can be identified with sets of games and therefore $\{A \mid B\}$ is a game. When r is a real number that is not a dyadic fraction, it can be identified with the game $\{A \mid B\}$. At this step one has to check that the structure of the real numbers as an ordered group is preserved under the embedding but this turn out to be the case [7].

We have seen that real numbers may be identified with games, but combining the definition of a game with the idea of a Dedekind section leads to the much larger class of numbers called the *surreal numbers*. Formally a surreal number is a game of the form $\{A \mid B\}$ where A and B are sets of (already constructed) surreal numbers such that $a < b$ for $a \in A$ and $b \in B$. That means that a surreal number can always be played as a combinatorial game.

Example 4 *The first transfinite ordinal number ω is identified with the game $\{\mathbb{N} \mid \varnothing\}$. The equation $\omega - \omega = 0$ makes no sense in Cantor's arithmetic for transfinite ordinals or cardinals, but if we identify ω with a game the equation makes sense, because we have*

$$\omega - \omega = \{1, 2, 3, \cdots \mid \varnothing\} + \{\varnothing \mid -1, -2, -3, \cdots\}.$$

This game is essentially like "my father has more money than your father" and most children soon experience that one should not start in such a game. It is clear that ω should not be interpreted as an amount but is better understood as a huge set of options. Conway identified all Cantor's ordinal numbers with surreal numbers, but Cantor and Conway use different *additive structures so the identification is somewhat problematic. For instance Conway's addition is commutative but Cantor's addition of ordinal numbers is not. Here we shall use ω as a symbol for a game rather than an ordinal in Cantor's sense.*

Formally the surreal numbers are constructed by (transfinite) recursion. It starts with the number $0 = \{\varnothing \mid \varnothing\}$. In each recursion step one adds new surreal numbers to the ones already constructed. Addition and multiplication extend to surreal numbers and with these operations the surreal numbers are a maximal ordered field. Although the definition of surreal multiplication is relevant for the next two sections we cannot present the definition in this short note but have to refer to [7,18]. For most computations surreal numbers are not different from real numbers but the topology is different.

A game G is said to be *infinitesimal* if $-2^{-m} \leq G \leq 2^{-m}$ for all natural numbers m. The number $1/\omega$ is an example of an infinitesimal number that is positive. Between any two different real numbers there are more than continuously many surreal numbers, and the intersection of the intervals $[-2^{-m}; 2^{-m}]$ contains infinitely many *infinitesimal numbers*. Formally there are so many surreal numbers that they do not form a set but a class.

4 Surreal probabilities and payoffs

Here we will introduce a version of the *Dutch Book Theorem* for surreal payoff functions. Because of the somewhat different topology of the surreal numbers, we have to be a little careful in the formulation and proof of the Dutch Book Theorem. In particular some of the standard methods for proving these results like the Hahn-Banach theorem and the separation theorem for convex sets, do not hold in their normal formulation when we are using surreal numbers. Those used to to non-standard analysis may note that what we are doing is essentially to veify that our result may be formulated in first order language.

The setup is as follows. Alice wishes to make a bet on an outcome $a \in A$. A bookmaker $b \in B$ offers the surreal payoff $g(a, b)$ (positive or negative) if the outcome

of a random event is $a \in A$. Thus $(a,b) \to g(a,b)$ can be considered as a matrix when A and B are finite sets. Alice should reject to play with a bookmaker b if Alice thinks that the payoff function $a \to g(a,b)$ is not favorable. For simplicity we shall assume that Alice accepts the payoff functions offered by the bookmakers $b \in B$. We recall that a surreal number is a game so if the outcome is a and the bookmaker is b then Alice has to play the game $g(a,b)$ against the bookmaker with Alice playing Left and the bookmaker playing Right.

By a *portfolio* we shall mean a probability vector $Q = (q_b)_{b\in B}$ on B. In this section will allow the portfolio to have surreal values. Such a portfolio is described by the payoff function

$$a \to \sum_{b\in B} q_b \cdot g(a,b), \tag{1}$$

A *Dutch book* is a portfolio such that (1) is negative for all $a \in A$, i.e. the portfolio game will be lost by Alice for any value of $a \in A$.

We assume that one of the bookmakers b_0 offers a payoff function $g(a,b_0) = 0$ for all $a \in A$ (b_0 acts like a bank with interest rate 0). Let Q be a portfolio and assume that there exists a Dutch book Q'. If Q has B as support then $q_{\min} = \min_{b\in B} q_b > 0$ and the payoff is

$$\sum_{b\in B} q_b \cdot g(\cdot,b) =$$
$$\sum_{b\in B} (q_b - q_{\min}\cdot q'_b)\cdot g(\cdot,b) + q_{\min} q_b \sum_{b\in B} q'_b \cdot g(\cdot,b) <$$
$$\sum_{b\in B} (q_b - q_{\min}\cdot q'_b)\cdot g(\cdot,b) + \left(q_{\min}\sum_{b\in B} q'_b\right)\cdot g(\cdot,b_0).$$

Hence Alice should reject to play with at least one of the bookmakers. If no Dutch book exists the set of payoff functions is said to be *coherent*. The notion of convexity will be used, and in this section we allow surreal coefficients in convex combinations.

Theorem 1 *Let A and B denote finite sets and let $(a,b) \to g(a,b)$ denote a surreal valued payoff function. If the payoff function is coherent then there exists non-negative surreal numbers p_a such that $\sum p_a = 1$ and*

$$\sum_{a\in A} p_a \cdot g(a,b) \geq 0 \tag{2}$$

for all $b \in B$.

Proof. Assume that A has d elements. Then each function $g(\cdot,b)$ may be identified with a d-dimensional surreal vector. Let K be the convex hull of $\{g(\cdot,b) \mid b \in B\}$, and let L denote the strictly negative surreal functions on A. They are convex classes.

If K and L intersect then there exists non-negative surreal numbers q_b such that $\sum q_b = 1$ and such that (1) defines a strictly negative function.

Assume that K and L are disjoint. Then define $C = K - L$ as the class of vectors $\bar{x} - \bar{y}$ where $\bar{x}$ in K and $\bar{y}$ in L. This is convex and does not contain $\bar{0}$. Now, K is a polytope (convex hull of finitely many extreme points) and L is polyhedral (given by finitely many inequalities), so C is polyhedral. Hence, each of the faces of C is given by a linear inequality of the form $\sum_{a\in A} p_a \cdot g(a) \geq c$ for $g \in C$. The delta function δ_α is non-negative so if g is in C then $g - \ell\cdot\delta_\alpha$ is also in C for ℓ positive. In particular

$$\begin{aligned} c &\leq \sum_{a\in A} p_a \cdot (g - \ell\cdot\delta_\alpha)(a) \\ &= \sum_{a\in A} p_a \cdot g(a) - \sum_{a\in A} p_a \ell \delta_\alpha(a) \\ &= \sum_{a\in A} p_a \cdot g(a) - \ell \cdot p_\alpha \end{aligned}$$

for all positive ℓ. Hence $p_\alpha \geq 0$ for all $\alpha \in A$. Further we know that $\bar{0}$ is not in C so that $\sum_{a\in A} p_a \cdot 0 \geq c$ does not hold and therefore $c > 0$. In particular p_a cannot be 0 for all a. The result follows by replacing p_a by

$$\frac{p_a}{\sum_{a\in A} p_a}.$$

■

Note that our surreal valued version Dutch Book Theorem states there are *two exclusive* cases:

1. Dutch book.
2. Non-negative mean value.

The theorem leads to surreal probabilities $p_a \geq 0$. Due to the normalization we do not have infinite probabilities, but there is no problem in having infinitesimal probabilities. In general the probability distribution will not be uniquely determined, but will merely be located in a non-empty convex set (credal set). Therefore the Dutch Book Theorem suggests that uncertainty about some unknown event should be represented by a *convex set of surreal probability distributions* rather than a single real valued distribution. Real functions are special cases of surreal functions so even if the payoff functions are real valued one can model our uncertainty by a convex set of surreal probability distributions.

If either g is acceptable or $-g$ is acceptable then it is called a two-sided bet. In this case the convex set of

probability distributions reduces to a point. The term one-sided bet is taken from F. Hampel [12]. In general people will find it difficult to decide that either g or $-g$ is acceptable and thus the two-sided bet is not realistic. In De Finetti [8] only two-sided bets were considered. In our formulation of the Dutch Book Theorem we just have a one-sided bet with a set of acceptable payoff functions.

A special case that has been studied in great detail is when the functions $g(\cdot, b)$ only assume two different values, i.e. $g(\cdot, b)$ has the form

$$g(a,b) = \begin{cases} g_1(b), & \text{for } a \in A_b; \\ g_2(b), & \text{for } a \notin A_b. \end{cases}$$

Without loss of generality we may assume that $g_1(b) \geq 0 > g_2(b)$. Then the g is accepted when $P(A_b) g_1(b) + (1 - P(A_b)) g_2(b) \geq 0$ or equivalently

$$P(A_b) \geq \frac{-g_2(b)}{g_1(b) - g_2(b)}. \tag{3}$$

We then define the *lower provision function* [21] by

$$L(A) = \min P(A)$$

where the minimum is taken over all distributions P that satisfies (3) for all $b \in B$. One may form surreal lower provisions in the same way as ordinary lower provisions are formed.

In this section we have seen that uncertainty may be identified with a convex set of surreal-valued probability distribution, but often such convex sets contain a lot of real-valued distributions. One may therefore ask whether the surreal-valued distributions add anything to the theory. Are they of any use? This we will try to answer in the next section.

5 Two-person zero-sum games

The theory of two persons zero sum games was founded by J. von Neumann together with O. Morgenstern [20] and has been extended to social games with more players. The readers who are interested in a deeper understanding of the theory of social games should consult [19] for an easy introduction or [10] for a more detailed exposition.

A social game with 2 players, that we will call Alice and Bob, is described by 2 sets of *strategies* A, B such that Alice can choose a strategy from A and Bob can choose a strategy from B. If Alice choose a and Bob choose b then the payoff for Alice will be $g(a,b)$ and the payoff for Bob will be $-g(a,b)$, where g is a function from $A \times B$ to surreal numbers. Alice and Bob will never collaborate in a zero-sum game because what is good for one of the players is equally bad for the other.

A pair of strategies (a,b) is called a *Nash equilibrium* if no player will benefit by changing his own strategy if the other player leaves his strategy unchanged. If a game has a unique Nash pair and both players are *rational*, then both players should play according to the Nash equilibrium.

Assume that the players are allowed to use mixed strategies, i.e. choose independent probability distributions over the strategies. The probabilities are allowed to take surreal values. Let P be the mixed strategy of Alice and Q be a mixed strategy of Bob. Then the *mean payoff* for Alice is

$$g(P,Q) = \sum_{(a,b)} g(a,b) \cdot p_a q_b.$$

This number is considered as the payoff of the social game where mixed strategies are allowed.

Theorem 2 *Consider a game with surreal valued payoffs. If the players are allowed to use mixed strategies, then the game has a Nash equilibrium.*

There exists various different proofs of the existence of Nash equilibria for two-person zero-sum games [2, 10, 19, 20]. In this note we shall focus on its equivalence with the Dutch Book Theorem.

The minimax inequality

$$\max_{a \in A} \min_{b \in B} g(a,b) \leq \min_{b \in B} \max_{a \in A} g(a,b)$$

is proved in exactly the same way for surreal payoff functions as for real payoff functions. The game is said to be in *equilibrium* when these quantities are equal. The common value is the *value of the game*. For any mixed strategy P for Alice the minimum of $g(P,Q)$ over distributions Q is attained when Q is concentrated in a point, i.e. $Q = \delta_b$ for some pure strategy $b \in B$. Thus

$$\min_Q g(P,Q) = \min_b \sum_a g(a,b) \cdot p_a. \tag{4}$$

To maximize this over all surreal-valued distributions P is a linear programming problem and can be solved by the same methods as if the payoff functions were real valued. In particular there exists a surreal valued distribution that maximizes (4). Using this argument we see that minimax and maximin are obtained even for mixed strategies.

Proof of equivalence of Thm. 1 and Thm. 2. Assume that for a two person zero-sum game there exists a value λ with optimal strategies P and Q. Then

g	a_1	a_2
b_1	$1+1/\omega$	$-1-2/\omega$
b_2	-1	$1+1/\omega$

Table 1: Payoff for Alice.

g	a_1	a_2
b_1	$\omega+1$	$-\omega-2$
b_2	$-\omega$	$\omega+1$

Table 2: Payoff for Alice multiplied by ω.

$\lambda < 0$ leads to the existence of a Dutch book and $\lambda \geq 0$ leads to the existence of a distribution P satisfying (2).

Assume that the Dutch Book Theorem holds. Assume that there exist a surreal number λ such that

$$\max_P \min_Q g(P,Q) < \lambda < \min_Q \max_P g(P,Q)$$

Consider the payoff function $f(a,b) = g(a,b) - \lambda$. According to the Dutch Book Theorem there exists a probability distribution P on A

$$\sum_{a\in A} p_a \cdot f(a,b) \geq 0$$

for all $b \in B$; or there exists a probability distribution Q on B such that

$$\sum_{b\in B} q_b \cdot f(a,b) < 0$$

for all $a \in A$. Therefore there exists a probability distribution P on A such that

$$\sum_{a\in A} p_a \cdot g(a,b) \leq \lambda \tag{5}$$

for all $a \in A$ or there exists a probability distribution Q on B such that

$$\sum_{b\in B} q_b \cdot g(a,b) \geq \lambda \tag{6}$$

for all strategies $a \in A$. Inequality (5) contradicts that $\lambda < \min_Q \max_P g(P,Q)$ and Inequality (6) contradicts that $\max_P \min_Q g(P,Q) < \lambda$. Hence, $\max_P \min_Q g(P,Q) = \min_Q \max_P g(P,Q)$. ■

The importance of the proof that the Dutch Book Theorem is equivalent to the existence of a Nash equilibrium for two-person zero-sum games is that it means that the two results refer to the same type of rationality. The next example show that the use of using surreal probabilities may make the difference between winning and losing.

Example 5 *Consider the payoff function in Table 1. If Alice ignores infinitesimals her optimal strategy is the distribution* $(1/2, 1/2)$, *which gives a payoff function for Bob that is* $-1/2\omega$ *if* $b = b_1$ *and* $1/2\omega$ *if* $b = b_2$. *In this case Bob could win the game by choosing* $b = b_1$. *The minimax optimal strategy for Alice is the mixed strategy* $\left(1/2 + \frac{1}{4(\omega+1)}, 1/2 - \frac{1}{4(\omega+1)}\right)$. *If she choose this mixed strategy the payoff is always positive and she will win the game.*

One should note that playing this game is not very different from playing the game where we have scaled the payoff up by a factor ω *(see Table 2). We may also scale up Bob's optimal strategy by a factor* $4(\omega+1)$ *to obtain* $(2\omega+3, 2\omega+1)$. *Therefore an optimal strategy for Alice is to play the game* $4(\omega+1)$ *"times" in parallel in such a way that* a_1 *is "chosen* $2\omega+3$ *times" and* a_2 *is "chosen* $2\omega+1$ *times ".*

If a two-persons zero-sum game has a Nash equilibrium pair $\left(\tilde{a}, \tilde{b}\right)$, which is always the case if A and B are finite, then $\sup_{a\in A} g\left(a, \tilde{b}\right) = g\left(\tilde{a}, \tilde{b}\right)$ and therefore $\inf_{b\in B} \sup_{a\in A} g(a,b) \leq g\left(\tilde{a}, \tilde{b}\right)$. Similarly, $\sup_{a\in A} \inf_{b\in B} g(a,b) \geq g\left(\tilde{a}, \tilde{b}\right)$. Thus, the game is in equilibrium and the value of the game is $g\left(\tilde{a}, \tilde{b}\right)$. In particular all Nash equilibria have the same value. The same argument holds for mixed strategies.

6 Dutch books for short games

Surreal numbers are totally ordered and never confused with each other. Games that are not surreal number are confused with a small or large interval of surreal numbers. For instance $*$ is confused with 0 and the game $\{100 \mid -100\}$ is confused with any number between -100 and 100. Before formulating a Dutch Book Theorem for general combinatorial games we need to introduce the *mean value* $\mu(G)$ of a short game G. A game G is said to be *short* if it only has finitely many positions. Our recursive definition of games allows transfinite recursion and games that are not short, but for the definition of mean values we shall focus on the short games. Note that if a short game is a number then it is a dyadic fraction.

The mean value of a game G is a real number $\mu(G)$ that satisfies the following mean value theorem.

Theorem 3 ([7]) *If* G *is a short game then there exists a natural number* m *and a number* $\mu(G)$ *that satisfies*

$$n \cdot \mu(G) - m \leq n \cdot G \leq n \cdot \mu(G) + m$$

for all natural numbers n.

Mean values of short games can be calculated by the *thermographic method* described in [7] and using this method it is easy to see that the mean value of a short game is always a rational number. Mean values of games share some important properties with mean values of random variables. For instance we have

- $\mu(n \cdot G) = n \cdot \mu(G)$,
- $\mu(G + H) = \mu(G) + \mu(H)$,
- $G \geq 0 \Rightarrow \mu(G) \geq 0$,
- $\mu(1) = 1$.

Example 6 *The game $G = \{1 \mid \{0 \mid -2\}\}$ that is illustrated in Figure 2, satisfies $G > 0$. In the game $n \cdot G$ Right can only play in a sub-game where Left has not played and the response optimal for Left is always to answer a move of Right by a move in the same sub-game. From this one sees that $n \cdot G \leq 1$ and therefore that $\mu(G) = 0$. We see that Left may win a game* for sure *although the game has mean value zero!*

The setup is as before that each bookmaker $b \in B$ tells Alice which game he wants to play if a certain horse $a \in A$ wins. Alice is going to play Left and the bookmaker or the bookmakers are going to play Right. After certain bookmakers have been accepted the bookmakers choose natural numbers $n_b, b \in B$ and combine these into a super game $\sum_{b \in B} n_b \cdot G(a, b)$ that will depend on which horse wins. We say that we have a *Dutch book* if there exists natural numbers $n_1, n_2, \cdots, n_k$ such that Alice will lose the game

$$\sum_{b \in B} n_b \cdot G(a, b) \tag{7}$$

for any value of a. Otherwise the set of game valued payoff functions is said to be *coherent*. If all the games are short surreal numbers then this notion of coherence is equivalent to the definition of coherence given in Section 4.

Alice is allowed to choose that the game should be played a number of times in parallel. With this setup we get the following version of the Dutch Book Theorem.

Theorem 4 *If a payoff function $G(a, b), a \in A, b \in B$ with short games as values, is coherent then either exists a probability vector $a \to p_a$ and a natural number n such that $np_a \in \mathbb{N}$ and the game*

$$\sum_a (np_a) \cdot G(a, b) > 0, \text{ for all } b \in B, \tag{8}$$

or there exist natural numbers $n_1, n_2, \cdots, n_k$, a natural number n and a probability vector $a \to p_a$ such that both games (7) and (8) have mean value 0.

Proof. We apply the existence of an equilibrium in the two-person zero-sum game with payoff function $(a, b) \to \mu(G(a, b))$. If the value of the two-person zero-sum game is negative then the game (7) is negative if the coefficients $n_1, n_2, \cdots, n_k$ are large enough. If the value of the two-person zero-sum game is non-negative there exists a probability vector $a \to p_a$ such that

$$\sum_a p_a \cdot \mu(G(a, b)) \geq 0.$$

The mean value of a short game is a rational number. Therefore the probability vector $a \to p_a$ can be chosen with rational point probabilities. Hence, there exists a natural number m such that $m \cdot p_a$ is an integer for all $a \in A$. Therefore

$$\begin{aligned} 0 &\leq m \sum_a p_a \cdot \mu(G(a, b)) \\ &\leq \sum_a m p_a \cdot \mu(G(a, b)) \\ &= \mu\left(\sum_a m p_a \cdot G(a, b)\right). \end{aligned}$$

If

$$\mu\left(\sum_a m p_a \cdot G(a, b)\right) > 0$$

then there exists a natural number k such that

$$k \sum_a m p_a \cdot G(a, b) > 0$$

and the game defined in (8) is winning for Alice who plays as Left when $n = km$. Otherwise

$$\mu\left(\sum_a m p_a \cdot G(a, b)\right) = 0. \tag{9}$$

■

Here we should note that our short-game-valued Dutch Book Theorem stated there are *three* cases that are *not exclusive:*

1. Dutch book.
2. Positive mean.
3. Zero mean.

As we saw in Example 6 a game with mean zero may be positive or negative. Therefore a decision strategy

in which only games with positive means are acceptable will exclude some games that one will win for sure and a decision strategy where games with non-negative mean are acceptable will include some games that are lost for sure. The most reasonable solution to this problem seems to be to accept or reject according to the mean payoff with respect to some probability distribution, but leave the cases with mean zero undecided because a more detailed non-probabilistic analysis is needed for these cases.

7 More on infinitesimals

The Dutch Book Theorem for short games only used rational valued mean values. One may hope for a better Dutch Book Theorem if we allow also allow a mean value function with infinitesimal surreal numbers as mean values. For short games this will not solve the problem.

Definition 2 *A game G is said to be* strongly infinitesimal *if $-s \leq G \leq s$ for any surreal number $s > 0$.*

Example 7 *The game $\{0 \mid *\}$ is called* up *and denoted $\uparrow$. It is easy to check that $\uparrow > 0$. The game $\uparrow$ is infinitesimal (check how Left can win $2^{-s} - \uparrow$). One can prove that any infinitesimal short game is strongly infinitesimal [18].*

An interesting situation is when all games $G(a, b)$ are infinitesimal. In this case the Dutch Book Theorem for games as formulated in Theorem 4 tells exactly nothing because the mean value of strongly infinitesimal games would always be 0 even if surreal mean values are allowed. But if all games are infinitesimal one could shift to a different "mean value" concept. For short games one compares the game with $n \cdot 1$ and the game 1 can be considered as a unit in the theory. For infinitesimal short games one can compare with the infinitesimal game $\uparrow$ instead. It is possible to define an *atomic mean value* such that $\uparrow$ has mean 1, but the proofs are more involved. One can also prove a version of the Dutch Book Theorem for infinitesimally short games that involves three cases. The three cases are Dutch book, positive mean, and some games G that cannot be analyzed in the sense that their atomic mean value is zero. Although infinitesimal games can be treated with their own mean value concept this will not solve all problems because games that are not infinitesimal may sometimes be combined into strongly infinitesimal games. A simple example consist of the games 1 and $\uparrow - 1$ whose sum is the strongly infinitesimal game $\uparrow$.

8 Discussion

In any frequency interpretation of probability theory, probabilities should be interpreted as limits of frequencies. Obviously surreal probabilities cannot have such interpretations because a frequency interpretation cannot distinguish between surreal probabilities that have an infinitesimal difference. This leads us to the following conclusion: frequency probabilities are real numbers but uncertainty should in general be modelled by convex sets of surreal numbers.

In a *subjective Bayesian* approach to probability and statistics one will assign probabilities expressing the individual feeling of how probable or likely an event is. Many subjective Bayesians justify this point of view by reference to the Dutch Book Theorem. We note that unlike some of the modification by Savage et al. neither our formulation of the Dutch Book Theorem nor its original formulation of de Finetti has any reference to subjectivity. For short-game valued payoffs even the one-to-one correspondence between probability and coherent decisions breaks down. Experiments have demonstrated that most people have a bad intuition of probabilities and are unable to assign probabilities in a consistent manner. It should be even harder to make a consistent distinction between the probabilities $1/3$ and $1/3 + 1/\omega$ although the Dutch Book Theorem give the same type of justification for surreal probabilities as for real probabilities.

We have seen that from a mathematical point of view uncertainties may be modeled by a convex set of surreal probability vectors, but the reader may wonder why infinitesimals do normally not appear in probability theory. Actually there are many real numbers that never appear as probabilities. For instance all the numbers that *do* appear are *computable*, and there are only countably many computable numbers. Therefore, it seems that the use of surreal numbers is an idealization that is not worse than the use of real numbers as subjective probabilities. At the moment two-person zero-sum games like the ones described in Example 5 are the only known kind of calculations that gives surreal valued probabilities as results.

In this paper we used the operations $+$ and $\cdot$ to define Dutch books and coherence. These operations refer to ways of combining games into new games. It is an open question what kind of Dutch Book Theorem one would get if other ways of combining games were considered.

For social games with several players and surreal-valued payoff functions we have not been able to prove existence of a Nash equilibrium, because one cannot use the usual fixed-point results that rely heavily on

the topology of the real numbers. We shall not discuss it here as it has less interest for our understanding of what probabilities are.

Acknowledgements

Thanks to Wouter Koolen-Wijkstra, Peter Grünwald, and Mogens Esrom Larsen for many useful comments and discussions.

This work has was supported by the European Pascal Network of Excellence.

References

[1] N. L. Alling. *Foundations of Analysis over Surreal Number Fields*, volume 141 of *North-Holland Mathematicsl Studies*. Elsevier, Amsterdam, 1987.

[2] J. P. Aubin. *Optima and Equilibria. An Introduction to Nonlinear Analysis*. Springer, Berlin, 2nd edition edition, 1993.

[3] E. Berlekamp, M. Müller, and B. Spight. *Generalized Thermography: Algorithms, Implementation, and Application to Go Endgames*. International Computer Science Institute, Berkeley, CA, 1996. TR-96-030.

[4] E. R. Berlekamp. Introductory overview of mathematical go endgames. In R. K. Guy, editor, *Combinatorial Games*, volume 43 of *Proceedings in Applied Mathematics*, pages 73–100. American mathematical Society, 1991.

[5] E. R. Berlekamp, J. H Conway, and R. K. Guy. *Winning Ways of your mathematical plays*, volume 1: Games in General. Academic Press, 1982.

[6] F. C. Chu and J. Y. Halpern. Great expectations. part i: On the customizability of generalized expected utility. *Theory and Decision*, 64(1):1–36, 2008.

[7] J. H. Conway. *On Numbers and Games*. Academic press, London, 1976.

[8] B. de Finetti. Foresight: Its logical laws in subjective sources. In H. E. Kyburg and H. E. Smokler, editors, *Studies in Subjective Probability*, pages 93–158. Wiley, London, 1964. Reprint from 1937 edition.

[9] M. Ettinger. *Topics in Combinatorial Games*. Phd thesis, University of Wisconsin, Madison, 1996.

[10] R. Gibbons. *A Primer in Game Theory*. Harvester Wheatsheaf, 1992.

[11] H. Gonshor. *An Introduction to the Theory of Surreal Numbers*, volume 110 of *London Mathematical Society*. Cambridge University Press, 1986.

[12] F. Hampel. A sketch of a unifying statistical theory. Research Report 86, Eidgenössische Technische Hochschule (ETH), Zürich, Schweiz, Feb. 1999.

[13] A. Khrennikov. *p-Adic Valued Distributions in Mathematical Physics*. Mathematics and its Applications. Springer, 1994.

[14] D. Knuth. *Surreal Numbers: How Two Ex-Students Turned on to Pure Mathematics and Found Total Happiness*. Addison-Wesley Pub., 1974.

[15] R. J. Nowakowski, editor. *Games of No Chance*. Cambridge University Press, 1996.

[16] F. P. Ramsey. Truth and probability. In R. B. Braithwaite, editor, *Foundations of Mathematics and other Essays*. Routledge & P. Kegan, 1931. reprinted in Studies in Subjective Probability, H. E. Kyburg, Jr. and H. E. Smokler (eds.), 2nd ed., (R. E. Krieger Publishing Company, 1980), 2352; reprinted in F. P. Ramsey: Philosophical Papers, D. H. Mellor (ed.) (Cambridge: University Press, Cambridge, 1990).

[17] L. J. Savage. *The Foundations of Statistics*. Wiley & Sons, New York, 1954.

[18] D. Schleicher and M. Stoll. An introduction to Conway's games and numbers. *Mosc. Math. J.*, 6(2):359–388, 2006.

[19] P. D. Straffin. *Game Theory and Strategy*, volume 36 of *New Mathematical Library*. Mathematical Ass. of America, Washington, DC, 1993.

[20] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, 1947. 2nd. edition.

[21] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

[22] Ling Zhao and M. Müller. Solving probabilistic combinatorial games. In *Advances in Computer Games*, volume 4250 of *Lecture Notes in Computer Science*, pages 225–238. Springer Berlin / Heidelberg, 2006.

Characterizing Factuality in Normal Form Sequential Decision Making

Nathan Huntley
Durham University
Department of Mathematical Sciences
Durham, UK
nathan.huntley@durham.ac.uk

Matthias C. M. Troffaes
Durham University
Department of Mathematical Sciences
Durham, UK
matthias.troffaes@gmail.com

Abstract

We prove necessary and sufficient conditions on choice functions for factuality to hold in normal form sequential decision problems. We find that factuality is sufficient for backward induction to work. However, choice must be induced by a total preorder for factuality to hold. Hence, many of the optimality criteria used in imprecise probability theory (such as interval dominance, maximality, and E-admissibility) are counterfactual under normal form decision making.

1 Introduction

Consider the two-stage decision problem depicted in Fig. 1. In the first stage, the subject chooses between either taking scones, or proceeding to the second stage. In the second stage, the subject chooses between either cake or ice cream. A normal form solution to this problem consists of the subject specifying all his admissible choices, at all stages, beforehand. One possible normal form solution is

$$\{\text{scones}, \text{no scones and then ice cream}\}.$$

Imagine now that the subject already chose not to have scones. To resolve his choice between cake and ice cream, the subject can go back to the original problem that involved scones, and take the ice cream, but we might also imagine that he simply forgets about the scones and considers the simpler problem of choosing between cake and ice cream, as in Fig. 2.

If, faced with the simpler problem, the subject would now not state ice cream as his only admissible choice, we say that he is *counterfactual*: his choice between cake and ice cream depends on whether or not he had the choice of scones before. Perhaps, this seems an awkward property at first, but as we shall see, counterfactual choices are legion in many theories—a notable exception is maximizing expected utility.

So, when faced with a sequential decision problem, at any particular stage, one has two ways of looking at its normal form solution. Either, the problem can be thought of as part of a much larger problem (considering past choices one did not make and events that did not happen), or the problem can be thought of in its simplest form, not considering any past stages. Intuitively, a reasonable requirement is that the solution at any particular stage does not depend on the larger problem it is embedded in, i.e., that it is *factual*.

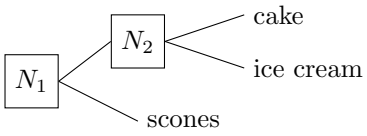


Figure 1: Two-stage problem.

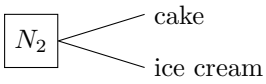


Figure 2: Second stage.

This papers studies necessary and sufficient conditions on choice functions for factuality in sequential decision problems when using normal form solutions, extending some results of Hammond [1] in his consequentialist theory. In doing so, factuality turns out to be sufficient for a backward induction scheme to work. We also find that choice must be induced by a total preorder for factuality to hold: for any choice function not induced by a total preorder, we can construct a counterfactual normal form example.

The relevance of this result for imprecise probability theory is that any criterion of optimality not induced by a total preorder (such as maximality, E-admissibility, and interval dominance) necessarily leads to counterfactuality. In other words, to satisfy

factuality, one *must reject* either (i) the normal form as a means of solving decision problems, or (ii) any criterion that is not induced by a total preorder.

A total preorder, however, is not sufficient to imply factuality. Indeed, many total preorders that have been proposed for choice are still counterfactual. When precise probabilities are used, Hammond showed that expected utility is factual, as is well known, as are several related criteria [1, Sec. 9]. We are not aware of any non-trivial factual criteria that do not rely on probability and expected utility, although they may exist. The representation of all factual optimality criteria is still an open problem.

The paper is structured as follows: Section 2 explains decision trees and introduces notation. Section 3 provides a careful definition of normal and extensive form solutions, and introduces the concept of gambles to more easily work with normal form solutions. Section 4 introduces choice functions and their relationship with normal form solutions. Section 5 defines factuality and contains the principal results.

2 Decision Trees

2.1 Definition and Example

A decision tree [6] consists of a rooted tree of decision nodes, chance nodes, and reward leaves, growing from left to right. The left hand side corresponds to what happens first, and the right hand side to what happens last.

Consider Fig. 3. Decision nodes are depicted by squares, and chance nodes by circles. From each node, branches emerge. For decision nodes, each branch matches a decision; for chance nodes, each branch matches an event. For each chance node, the events that emerge form a partition of the possibility space: exactly one of the events must obtain. Each path in a decision tree corresponds to a particular sequence of decisions and events. The payoff resulting from each such sequence is put at the right end of the tree.

2.2 Notation

Decision trees can be seen as combinations of smaller decision trees: for instance, in the example, one could draw the subtree corresponding to d_S, and also draw the subtree corresponding to $d_{\overline{S}}$. The full decision tree then joins these two subtrees at a decision node.

Hence, we can represent a decision tree by its subtrees and the type of its root node. Let $T_1, \ldots, T_n$ be decision trees and $E_1, \ldots, E_n$ be a partition of the possibility space. If T is rooted at a decision node, we write $T = \bigsqcup_{i=1}^{n} T_i$, and at a chance node, we write $T = \bigodot_{i=1}^{n} E_i T_i$. For instance, for the tree of Fig. 3,

$$(S_1(T_1 \sqcup T_2) \odot S_2(T_1 \sqcup T_2)) \sqcup (U_1 \sqcup U_2) \text{ with}$$
$$T_1 = E_1 9 \odot E_2 14 \quad U_1 = E_1 10 \odot E_2 15$$
$$T_2 = E_1 4 \odot E_2 19 \quad U_2 = E_1 5 \odot E_2 20$$

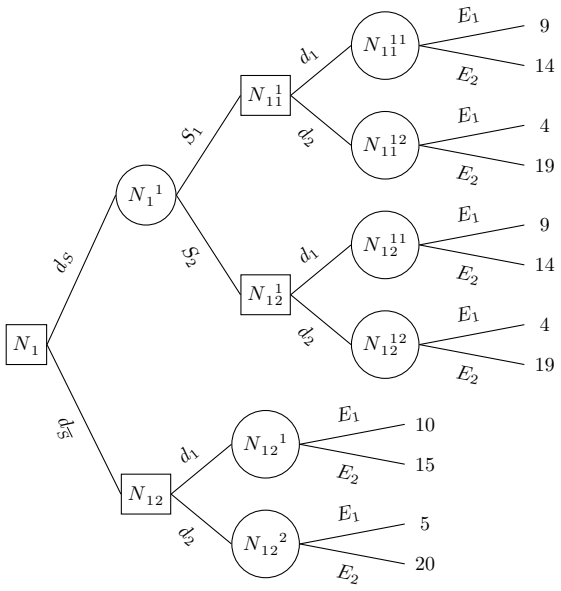


Figure 3: A decision tree.

Definition 1. *A subtree of a tree T obtained by removal of all non-descendants of a particular node N is called* the subtree of T at N *and is denoted by* $\mathrm{st}_N(T)$.

For any (sub)tree T, we summarize the events observed in the past as $\mathrm{ev}(T)$, which is the intersection of all the events on chance arcs that preceded T.

3 Solving Decision Trees

This paper deals with more general solutions of decision trees than are usually considered. Consequently, the standard definitions of extensive and normal forms, such as in Raiffa and Schlaifer [10], are insufficient for our purpose. Therefore, we first carefully define normal and extensive form solutions.

3.1 Extensive and Normal Form Solutions

An *extensive form solution* takes the decision tree and removes from each decision node some (possibly none), but not all, of the decision arcs. So, an extensive form solution is a subtree of the original decision tree, where at each decision node only a non-empty subset of arcs is retained. For instance, in the example, one of the extensive form solutions is: take $d_{\overline{S}}$,

and then either take d_1 or d_2. An extensive form solution can be used as follows: the subject, upon reaching a decision node, chooses one of the arcs in the extensive form solution, and follows it. The subject only needs to decide which arc to follow at a decision node when reaching that node.

Following Raiffa and Schlaifer [10], Luce and Raiffa [7], and many others, another way to describe solutions to decision trees goes as follows. First, an extensive form solution with just one arc out of each decision node, is called a *normal form decision*. Hence, once a normal form decision is specified, a subject's decisions are uniquely determined in every eventuality. For instance, in the example, one of the normal form decisions is: take d_S, followed by d_1 if S_1 obtains, and d_2 if S_2 obtains. We denote the set of all normal form decisions for a decision tree T by $\mathrm{nfd}(T)$.

The interpretation of a normal form decision is that, upon reaching a decision node, the subject chooses the arc specified in the normal form decision. Compare this with a more general extensive form solution, in which the subject, upon reaching a decision node, chooses one of a subset of the available arcs. The difference between the two is that, for a normal form decision, the subject's choice at every decision node is uniquely determined from the beginning. In the extensive form, the particular arc to follow does not need to be determined unless the subject actually reaches the decision node in question.

A *normal form solution* of a decision tree T is then simply a subset of $\mathrm{nfd}(T)$. The interpretation of this subset is that the subject simply picks one of the normal form decisions of the normal form solution, and then acts accordingly.

Of course, an extensive form solution can always be transformed into a normal form solution by taking every possible normal form decision that is compatible with it. However, there are usually more normal form solutions than there are extensive form solutions.

3.2 Extensive and Normal Form Operators

An *extensive form operator* is a function which maps each decision tree to an extensive form solution of that decision tree. Note that some definitions in the literature, such as Raiffa and Schlaifer [10], define extensive form solutions through backward induction. Our definition does not specify the method by which decision arcs are removed. There need be no relationship between extensive forms and recursive methods.

An *normal form operator* is a function which maps each decision tree to a normal form solution of that decision tree. Again, note that the method by which this subset is determined is not part of our definition.

These operators usually (but do not need to) have the interpretation of describing optimal solutions.

An example of an extensive form operator is the classical backward induction method. Moving from right to left in the tree, decision arcs are deleted unless they give the maximum expected utility over all available arcs at that node. The principal feature of the method is that, once an arc has been deleted, it is ignored in all future calculations at nodes further to the left in the tree. The corresponding normal form operator finds the expected utility of each normal form decision and then returns the set that maximizes expected utility.

While it is well documented that these two classical operators always give equivalent solutions, this relationship can fail for other criteria. Extensive form operators that recursively apply a criterion may give a solution that differs from the normal form operator that applies the same criterion to the set of all normal form decisions. Examples can be found in Seidenfeld [11], Machina [8], and Jaffray [4], among others.

3.3 Gambles

In this paper we are primarily investigating normal form solutions. To express normal form decisions and solutions efficiently, we first introduce some definitions and notation. Let Ω be the *possibility space*: the set of all possible states of the world. We only consider finite possibility spaces. Elements of Ω are denoted by ω. Subsets of Ω are called *events*. The arcs emerging from chance nodes in a decision tree correspond to events.

Let $\mathcal{R}$ be a set of rewards. Often, rewards are measured in utiles, and hence $\mathcal{R} = \mathbb{R}$, but this assumption is not necessary for our results.

A *gamble* is a function $X\colon \Omega \to \mathcal{R}$; in other words, gambles are Ω–$\mathcal{R}$ functions. Gambles are interpreted as uncertain rewards: should $\omega \in \Omega$ be the true state of the world, the gamble X will yield the reward $X(\omega)$. Note that no probabilities over Ω are assumed at all.

3.4 Normal Form Gambles

Recall that a normal form decision prescribes the subject's actions, so once one has been chosen, the reward is determined entirely by the events that obtain. In other words, a normal form decision has a corresponding gamble, which we call a *normal form gamble*. The set of all normal form gambles associated with a decision tree T is denoted by $\mathrm{gamb}(T)$, so gamb is an operator on trees which yields the set of all gambles induced by normal form decisions of the tree.

	ω_1	ω_2	ω_3	ω_4
$E_1 9 \oplus E_2 14$	9	9	14	14
$S_1\,(E_1 9 \oplus E_2 14)$ $\oplus S_2\,(E_1 4 \oplus E_2 19)$	9	4	14	19

Table 1: Example of normal form gambles.

Let us explain how to find the gamble corresponding to a normal form decision, using Fig. 3 as an example. Instead of looking at the full tree, for simplicity let us first consider the subtree with root at $N_1{}^1_1$. The only two normal form decisions in this subtree are simply d_1 and d_2. The former gives reward 9 utiles if $\omega \in E_1$ and 14 utiles if $\omega \in E_2$, which corresponds to a gamble

$$E_1 9 \oplus E_2 14. \tag{1}$$

In the above expression, the $\oplus$ operator combines partial maps defined on disjoint domains (i.e. the constant partial map $E_1 9$ defined on E_1, and the constant partial map $E_2 14$ defined on E_2).

Now consider the subtree with root at $N_1{}^1$, and in particular the normal form decision 'd_1 if S_1 and d_2 if S_2'. This gives reward 9 if $\omega \in S_1 \cap E_1$, reward 14 if $\omega \in S_1 \cap E_2$, and so on. The corresponding gamble is

$$(S_1 \cap E_1)9 \oplus (S_1 \cap E_2)14 \oplus (S_2 \cap E_1)4 \oplus (S_2 \cap E_2)19,$$

or briefly, if we omit '$\cap$' and employ distributivity,

$$S_1\,(E_1 9 \oplus E_2 14) \oplus S_2\,(E_1 4 \oplus E_2 19)\,,$$

where multiplication with an event is now understood to correspond to restriction, i.e., 9 is a constant map on Ω, $E_1 9$ is a constant map restricted to E_1, and $S_1(E_1 9)$ is obtained from $E_1 9$ by further restriction to $E_1 \cap S_1$. For illustration, we tabulate the values of some normal form gambles in Table 1, where $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, $E_1 = \{\omega_1, \omega_2\}$, and $S_1 = \{\omega_1, \omega_3\}$.

Observe that the above gamble includes the gamble in Eq. (1) from $N_1{}^1_1$. Relationships between sets of normal form gambles for different subtrees allows a very convenient recursive definition of the gamb operator, given next. First, we extend $\oplus$ to sets of gambles:

Definition 2. *For any events $E_1, \ldots, E_n$ which form a partition, and any finite family of sets of gambles $\mathcal{X}_1, \ldots, \mathcal{X}_n$, we define the following set of gambles:*

$$\bigoplus_{i=1}^{n} E_i \mathcal{X}_i = \left\{ \bigoplus_{i=1}^{n} E_i X_i \colon X_i \in \mathcal{X}_i \right\}$$

Definition 3. *With any decision tree T, we associate a set of gambles* $\mathrm{gamb}(T)$*, recursively defined through:*

- *If a tree T consists of only a leaf with reward $r \in \mathcal{R}$, then*

$$\mathrm{gamb}(T) = \{r\}. \tag{2a}$$

- *If a tree T has a chance node as root, that is, $T = \bigodot_{i=1}^{n} E_i T_i$, then*

$$\mathrm{gamb}\left(\bigodot_{i=1}^{n} E_i T_i \right) = \bigoplus_{i=1}^{n} E_i \,\mathrm{gamb}(T_i). \tag{2b}$$

- *If a tree T has a decision node as root, that is, if $T = \bigsqcup_{i=1}^{n} T_i$, then*

$$\mathrm{gamb}\left(\bigsqcup_{i=1}^{n} T_i \right) = \bigcup_{i=1}^{n} \mathrm{gamb}(T_i). \tag{2c}$$

Most decision problems can be modelled in more more than one way: there are usually multiple decision trees that model the same problem. This suggests the following definition (see for instance [8]):

Definition 4. *Two decision trees T_1 and T_2 are called* strategically equivalent *if* $\mathrm{gamb}(T_1) = \mathrm{gamb}(T_2)$.

4 Normal Form Solutions for Decision Trees

4.1 Choice Functions and Optimality

A normal form solution of a decision tree T is a subset of the set $\mathrm{nfd}(T)$ of all its normal form decisions. Ideally one would like to identify a single normal form decision that the subject considers the best, but there is no reason to suppose that this is always possible. The subject might, however, still be able to identify some normal form decisions that he would never consider choosing, and eliminate these. This leaves a subset of normal form decisions that the subject would be willing to choose from. We say that the subject considers elements of this subset to be *optimal*.

For instance, in classical decision theory, each normal form decision induces a random real-valued gain, and assuming that all probabilities are fully specified, a normal form decision is considered optimal if its expected gain is maximized. As another example, consider the situation where the probabilities are not precisely known, but a set $\mathcal{M}$ of plausible probability distributions can be specified. Then the subject might consider as optimal any of those normal form decisions whose expected gain is maximal under at least one probability distribution in $\mathcal{M}$. In other situations one might use a different optimality criterion.

In these two examples, optimal decisions are determined by comparison of gambles. This is a common approach, and one we follow here, since we have seen that normal form decisions have corresponding gambles, and gambles are easier to work with. We therefore suppose that the subject has some way of determining an optimal subset of any set of gambles,

conditional upon an event A (which corresponds to the $\mathrm{ev}(T)$ of the decision tree in question):

Definition 5. *A choice function* opt *is an operator that, for any non-empty event A, maps each non-empty finite set $\mathcal{X}$ of gambles to a non-empty subset of this set: $\emptyset \neq \mathrm{opt}(\mathcal{X}|A) \subseteq \mathcal{X}$.*

Note that common uses of choice functions in social choice theory, such as by Sen [12, p. 63, ll. 19–21] do not consider conditioning, and define choice functions for arbitrary sets of options (not for gambles only).

4.2 Normal Form Operator Induced by a Choice Function

We have seen that normal form decisions induce gambles, and have introduced choice functions, acting on sets of gambles, as a means to model optimality. Whence, we naturally arrive at a normal form operator $\mathrm{norm}_{\mathrm{opt}}$, simply by applying opt on the set of all gambles associated with the tree T and then finding the corresponding set of normal form decisions.

Definition 6. *Given any choice function* opt*, and any decision tree T with $\mathrm{ev}(T) \neq \emptyset$, we define*

$$\mathrm{norm}_{\mathrm{opt}}(T) = \{U \in \mathrm{nfd}(T): \mathrm{gamb}(U) \subseteq \mathrm{opt}(\mathrm{gamb}(T)|\mathrm{ev}(T))\}.$$

Of course, since U is always a normal form decision, $\mathrm{gamb}(U)$ is always a singleton in this definition. In particular, the following equality holds,

$$\mathrm{gamb}(\mathrm{norm}_{\mathrm{opt}}(T)) = \mathrm{opt}(\mathrm{gamb}(T)|\mathrm{ev}(T)). \qquad (3)$$

Note that, although $\mathrm{norm}_{\mathrm{opt}}$ is applied to trees, it really depends only on the set of normal form gambles associated with the tree. Hence, the operator $\mathrm{norm}_{\mathrm{opt}}$ will respect strategic equivalence:

Theorem 7. *If T_1 and T_2 are strategically equivalent, then* $\mathrm{gamb}(\mathrm{norm}_{\mathrm{opt}}(T_1)) = \mathrm{gamb}(\mathrm{norm}_{\mathrm{opt}}(T_2))$ *whenever* $\mathrm{ev}(T_1) = \mathrm{ev}(T_2) \neq \emptyset$.

If there are several strategically equivalent trees that are plausible representations of the same problem, the above theorem guarantees that our solution is independent of the particular representation we use.

When studying factuality, we consider $\mathrm{norm}_{\mathrm{opt}}$ for arbitrary subtrees of a given decision tree. To ensure that $\mathrm{norm}_{\mathrm{opt}}$ can be applied on each of such subtrees, the following condition is necessary:

Definition 8. *A decision tree T is called* consistent *if for every node N of T,* $\mathrm{ev}(\mathrm{st}_N(T)) \neq \emptyset$.

Clearly, if a decision tree T is consistent, then for any node N in T, $\mathrm{st}_N(T)$ is also consistent. We study only consistent decision trees because we consider $\mathrm{norm}_{\mathrm{opt}}(\mathrm{st}_N(T))$ for any node N in T, which is impossible when $\mathrm{ev}(\mathrm{st}_N(T)) = \emptyset$.

Usually, when constructing decision trees, one does not consider events which conflict with preceding events, hence consistency is satisfied. However, due to an oversight, some branch of a chance node might be connected to an event that cannot occur: such tree can always be made consistent by removing those nodes whose conditioning event is empty.

We sometimes need to know when a set of gambles can be represented by a consistent decision tree, conditional on some event. The following definition characterizes precisely those gambles:

Definition 9. *Let A be any non-empty event, and let $\mathcal{X}$ be a set of gambles. Then the following conditions are equivalent; if any (hence all) of them are satisfied, we say that $\mathcal{X}$ is* A-consistent.

(A) There is a consistent decision tree T with $\mathrm{ev}(T) = A$ and $\mathrm{gamb}(T) = \mathcal{X}$.

(B) For every $r \in \mathcal{R}$ and every $X \in \mathcal{X}$ such that $X^{-1}(r) \neq \emptyset$, it holds that $X^{-1}(r) \cap A \neq \emptyset$.

A gamble X is called A-consistent if $\{X\}$ is A-consistent.

5 Counterfactuals

We now give a discussion of issues arising from the use of operators, either normal form or extensive form, that use *counterfactual* reasoning, and find necessary and sufficient conditions on opt for $\mathrm{norm}_{\mathrm{opt}}$ to avoid counterfactuality. Counterfactual reasoning involves the consideration of events that did not occur or decisions that were not chosen. This is of interest because for many choice functions opt that have been suggested in the literature, $\mathrm{norm}_{\mathrm{opt}}$ is counterfactual.

5.1 Example and Definition

Counterfactuals are best illustrated by an example. Suppose we are applying an extensive form operator to the tree T in Fig. 3. This operator will delete some (possibly none) of the decision arcs at $N = N_{11}^1$. If the choice of arcs to delete is influenced only by $\mathrm{st}_N(T)$ (that is, the operator would delete the same arcs at N regardless of the larger tree in which $\mathrm{st}_N(T)$ is embedded) then the operator is called *factual*. If the operator does not have this property (for instance, if the solution were to depend on the possible consequences of $d_{\overline{S}}$ or S_2), then it is called *counterfactual*.

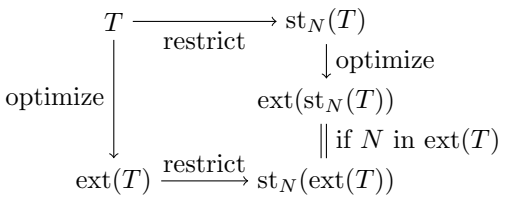


Figure 4: For a factual extensive form operator, optimization and restriction commute.

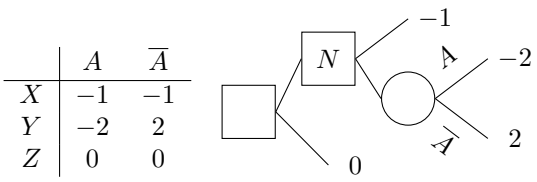


Figure 5: Decision tree for Example 12.

The definition of a counterfactual normal form operator requires the following extension to Definition 1.

Definition 10. *If $\mathcal{T}$ is a set of decision trees and N a node, then*

$$\mathrm{st}_N(\mathcal{T}) = \{\mathrm{st}_N(T) \colon T \in \mathcal{T} \text{ and } N \text{ in } T\}.$$

Definition 11. *An extensive form operator* ext *is called* factual *if for every consistent decision tree T and every node N such that N is in* $\mathrm{ext}(T)$*,*

$$\mathrm{st}_N(\mathrm{ext}(T)) = \mathrm{ext}(\mathrm{st}_N(T)),$$

otherwise, ext *is called* counterfactual*.*

An normal form operator norm *is called* factual *if for every consistent decision tree T and every node N such that N is in at least one element of* $\mathrm{norm}(T)$

$$\mathrm{st}_N(\mathrm{norm}(T)) = \mathrm{norm}(\mathrm{st}_N(T)),$$

otherwise, norm *is called* counterfactual*.*

In other words, for a factual operator, it does not matter whether we first restrict our attention to a subtree at a particular node N and then optimize this subtree, or first optimize, and only then look at the resulting subtree at a particular node N: roughly speaking, factuality means that optimization and restriction commute, as in Fig. 4 for an extensive form operator. For a counterfactual extensive form operator, $\mathrm{st}_N(\mathrm{ext}(T))$ can differ from $\mathrm{ext}(\mathrm{st}_N(T))$ for some decision trees T and nodes N in $\mathrm{ext}(T)$.

For example, the extensive form operator ext_P corresponding to the usual backward induction using expected utility is well known to be factual. Also, the usual normal form operator norm_P corresponding to maximizing expected utility over all normal form decisions is factual, because ext_P is equivalent to norm_P.

Before we examine factuality in more detail, we give an example of a counterfactual choice function.

Example 12. *Let T be the decision tree in Fig. 5, where X, Y, and Z are its normal form gambles. Under point-wise dominance, X and Y are incomparable, as are Y and Z. Hence,* $\mathrm{norm}(\mathrm{st}_N(T))$ *is $\{X, Y\}$ (where we conveniently identified normal form decisions with their normal form gambles). But* $\mathrm{norm}(T) = \mathrm{opt}(\{X, Y, Z\}) = \{Y, Z\}$ *as clearly Z dominates X. Restricting this solution to* $\mathrm{st}_N(T)$ *gives the normal form solution $\{Y\}$. Concluding,*

$$\{X, Y\} = \mathrm{norm}(\mathrm{st}_N(T)) \neq \mathrm{st}_N(\mathrm{norm}(T)) = \{Y\}$$

and therefore the normal form operator induced by point-wise dominance is counterfactual.

Even though point-wise dominance is counterfactual, it does satisfy $\mathrm{st}_N(\mathrm{norm}(T)) \subseteq \mathrm{norm}(\mathrm{st}_N(T))$, although this may not be true in general.

5.2 Necessary and Sufficient Conditions

In this section, we work extensively with normal form solutions, which are sets of trees. Therefore, it is convenient to extend gamb, $\odot$, and $\sqcup$, to sets of trees:

Definition 13. *For any set of decision trees $\mathcal{T}$,*

$$\mathrm{gamb}(\mathcal{T}) = \bigcup_{T \in \mathcal{T}} \mathrm{gamb}(T).$$

Definition 14. *For any sets of consistent decision trees $\mathcal{T}_1, \ldots, \mathcal{T}_n$, and any partition $E_1, \ldots, E_n$,*

$$\bigodot_{i=1}^{n} E_i \mathcal{T}_i = \left\{ \bigodot_{i=1}^{n} E_i T_i \colon T_i \in \mathcal{T}_i \right\}.$$

Definition 15. *For any sets of consistent decision trees $\mathcal{T}_1, \ldots, \mathcal{T}_n$,*

$$\bigsqcup_{i=1}^{n} \mathcal{T}_i = \left\{ \bigsqcup_{i=1}^{n} T_i \colon T_i \in \mathcal{T}_i \right\}.$$

For sets of trees, the gamb operator satisfies:

$$\mathrm{gamb}\left(\bigodot_{i=1}^{n} E_i \mathcal{T}_i \right) = \bigoplus_{i=1}^{n} E_i \,\mathrm{gamb}(\mathcal{T}_i),$$

$$\mathrm{gamb}\left(\bigsqcup_{i=1}^{n} \mathcal{T}_i \right) = \bigcup_{i=1}^{n} \mathrm{gamb}(\mathcal{T}_i).$$

$$\mathrm{gamb}(T) = \mathrm{gamb}(\mathrm{nfd}(T)).$$

The following three properties turn out to be necessary and sufficient for factuality of normal form operators induced by a choice function.

Property 1 (Conditioning Property)**.** *Let A be a non-empty event, and let $\mathcal{X}$ be a non-empty finite A-consistent set of gambles, with $\{X, Y\} \subseteq \mathcal{X}$ such that $AX = AY$. If $X \in \operatorname{opt}(\mathcal{X}|A)$, then $Y \in \operatorname{opt}(\mathcal{X}|A)$.*

Property 2 (Intersection property)**.** *For any event $A \neq \emptyset$ and any non-empty finite A-consistent sets of gambles $\mathcal{X}$ and $\mathcal{Y}$ such that $\mathcal{Y} \subseteq \mathcal{X}$ and $\operatorname{opt}(\mathcal{X}|A) \cap \mathcal{Y} \neq \emptyset$, it holds that $\operatorname{opt}(\mathcal{Y}|A) = \operatorname{opt}(\mathcal{X}|A) \cap \mathcal{Y}$.*

For the next property, we use the following notation: if A is a non-trivial event (non-empty and not Ω), then $A\mathcal{X} \oplus \overline{A}Z = \{AX \oplus \overline{A}Z \colon X \in \mathcal{X}\}$.

Property 3 (Mixture property)**.** *For any events A and B such that $A \cap B \neq \emptyset$ and $\overline{A} \cap B \neq \emptyset$, any $\overline{A} \cap B$-consistent gamble Z, and any non-empty finite $A \cap B$-consistent set of gambles $\mathcal{X}$,*

$$\operatorname{opt}(A\mathcal{X} \oplus \overline{A}Z|B) = A \operatorname{opt}(\mathcal{X}|A \cap B) \oplus \overline{A}Z.$$

Property 2 has a vast number of equivalent formulations, three of which we give next, yielding different interpretations to Property 2. These will be useful to discuss the implications of factuality later on.

Property 4 (Strong path independence)**.** *For any non-empty event A and any non-empty finite A-consistent sets of gambles $\mathcal{X}_1, \ldots, \mathcal{X}_n$, there is a non-empty $\mathcal{I} \subseteq \{1, \ldots, n\}$ such that*

$$\operatorname{opt}\left(\bigcup_{i=1}^{n} \mathcal{X}_i \middle| A\right) = \bigcup_{i \in \mathcal{I}} \operatorname{opt}(\mathcal{X}_i|A)$$

Property 5 (Very strong path independence)**.** *For any non-empty event A and any non-empty finite A-consistent sets of gambles $\mathcal{X}_1, \ldots, \mathcal{X}_n$,*

$$\operatorname{opt}\left(\bigcup_{i=1}^{n} \mathcal{X}_i \middle| A\right) = \bigcup_{\substack{i=1 \\ \mathcal{X}_i \cap \operatorname{opt}(\bigcup_{i=1}^{n} \mathcal{X}_i|A) \neq \emptyset}}^{n} \operatorname{opt}(\mathcal{X}_i|A)$$

Property 6 (Total preorder)**.** *For every event $A \neq \emptyset$, there is a total preorder $\succeq_A$ on A-consistent gambles such that for every non-empty finite set of A-consistent gambles $\mathcal{X}$,*

$$\operatorname{opt}(\mathcal{X}|A) = \{X \in \mathcal{X} \colon (\forall Y \in \mathcal{X})(X \succeq_A Y)\}$$

Lemma 16. *Properties 2, 4, 5 and 6 are equivalent.*

To show that Properties 1, 2 and 3 are necessary and sufficient for factuality of $\operatorname{norm}_{\operatorname{opt}}$, we require several lemmas (proofs are omitted due to space constraints).

We use this notation: for a decision tree T, $\operatorname{ch}(T)$ is the set of all child nodes of the root node of T.

Lemma 17. *Let* norm *be any normal form operator. Let T be a consistent decision tree. If,*

(i) for all nodes $K \in \operatorname{ch}(T)$ such that K is in at least one element of $\operatorname{norm}(T)$*,*

$$\operatorname{st}_K(\operatorname{norm}(T)) = \operatorname{norm}(\operatorname{st}_K(T)),$$

(ii) and, for all nodes $K \in \operatorname{ch}(T)$, and all nodes $L \in \operatorname{st}_K(T)$ such that L is in at least one element of $\operatorname{norm}(\operatorname{st}_K(T))$*,*

$$\operatorname{st}_L(\operatorname{norm}(\operatorname{st}_K(T))) = \operatorname{norm}(\operatorname{st}_L(\operatorname{st}_K(T))),$$

then, for all nodes N in T such that N is in at least one element of $\operatorname{norm}(T)$*,*

$$\operatorname{st}_N(\operatorname{norm}(T)) = \operatorname{norm}(\operatorname{st}_N(T)).$$

Lemma 18. *Let $A_1, \ldots, A_n$ be a finite partition of Ω, and let B be an event such that $A_i \cap B \neq \emptyset$ for all i. Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be a finite family of non-empty finite sets of gambles, where $\mathcal{X}_i$ is $A_i \cap B$-consistent. If a choice function* opt *satisfies Properties 2 and 3, then*

$$\operatorname{opt}\left(\bigoplus_{i=1}^{n} A_i\mathcal{X}_i \middle| B\right) = \bigoplus_{i=1}^{n} A_i \operatorname{opt}(\mathcal{X}_i|A_i \cap B).$$

Lemma 19. *Consider a consistent decision tree T whose root is a decision node, so $T = \bigsqcup_{i=1}^{n} T_i$, and any choice function* opt*. For each tree T_i, let N_i be its root. Then, N_i is in at least one element of* $\operatorname{norm}_{\operatorname{opt}}(T)$ *if and only if*

$$\operatorname{gamb}(T_i) \cap \operatorname{opt}(\operatorname{gamb}(T)|\operatorname{ev}(T)) \neq \emptyset.$$

Lemma 20. *For any consistent decision tree $T = \bigodot_{i=1}^{n} E_i T_i$, and any* opt *satisfying Property 1,*

$$\operatorname{gamb}(\operatorname{norm}_{\operatorname{opt}}(T)) = \bigoplus_{i=1}^{n} E_i \operatorname{gamb}(\operatorname{norm}_{\operatorname{opt}}(T_i))$$

implies

$$\operatorname{norm}_{\operatorname{opt}}(T) = \bigodot_{i=1}^{n} E_i \operatorname{norm}_{\operatorname{opt}}(T_i).$$

Lemma 21. *For any consistent decision tree $T = \bigsqcup_{i=1}^{n} T_i$ and any* opt *satisfying Property 2,*

$$\operatorname{gamb}(\operatorname{norm}_{\operatorname{opt}}(T)) = \bigcup_{i \in \mathcal{I}} \operatorname{gamb}(\operatorname{norm}_{\operatorname{opt}}(T_i)) \quad (4)$$

implies

$$\operatorname{norm}_{\operatorname{opt}}(T) = \bigsqcup_{i \in \mathcal{I}} \operatorname{norm}_{\operatorname{opt}}(T_i),$$

where $\mathcal{I} = \{i \colon \operatorname{gamb}(T_i) \cap \operatorname{opt}(\operatorname{gamb}(T)|\operatorname{ev}(T)) \neq \emptyset\}$.

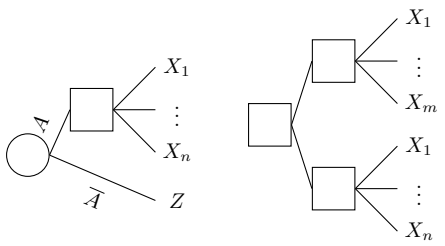


Figure 6: Decision trees for Theorem 22.

We are now ready to identify necessary and sufficient conditions for factuality.

Theorem 22. *A normal form operator* $\mathrm{norm}_{\mathrm{opt}}$ *is factual if and only if* opt *has Properties 1, 2 and 3.*

Proof. "only if". Omitting details, consider Fig. 6.

"if". We proceed by structural induction on all possible arguments of $\mathrm{norm}_{\mathrm{opt}}$, that is, on all consistent decision trees. In the base step, we prove the implication for trees consisting of only a single node. In the induction step, we prove that if the implication holds for the subtrees at every child of the root node, then the implication also holds for the whole tree.

First, if the decision tree T has only a single node, and hence, a reward at the root and no further children, then the condition for factuality is trivially satisfied.

Next, suppose that the consistent decision tree T has multiple nodes. Let $\{N_1, \ldots, N_n\} = \mathrm{ch}(T)$, and let $T_i = \mathrm{st}_{N_i}(T)$. The induction hypothesis says that factuality is satisfied for all subtrees at every child of the root node, that is, for all T_i. More precisely, for all $i \in \{1, \ldots, n\}$, and all nodes $L \in T_i$ such that L is in at least one element of $\mathrm{norm}_{\mathrm{opt}}(T_i)$

$$\mathrm{st}_L(\mathrm{norm}_{\mathrm{opt}}(T_i)) = \mathrm{norm}_{\mathrm{opt}}(\mathrm{st}_L(T_i)).$$

We must show that

$$\mathrm{st}_N(\mathrm{norm}_{\mathrm{opt}}(T)) = \mathrm{norm}_{\mathrm{opt}}(\mathrm{st}_N(T))$$

for all nodes N in T such that N is in at least one element of $\mathrm{norm}_{\mathrm{opt}}(T)$. By Lemma 17, and the induction hypothesis, it suffices to prove the above equality only for $N \in \mathrm{ch}(T)$, that is, it suffices to show that

$$\mathrm{st}_{N_i}(\mathrm{norm}_{\mathrm{opt}}(T)) = \mathrm{norm}_{\mathrm{opt}}(T_i) \tag{5}$$

for each $i \in \{1, \ldots, n\}$ such that N_i is in at least one element of $\mathrm{norm}_{\mathrm{opt}}(T)$.

If T has a chance node as its root, that is, $T = \bigodot_{i=1}^{n} E_i T_i$, then all N_i are actually in every element of $\mathrm{norm}_{\mathrm{opt}}(T)$, so we must simply establish Eq. (5) for all $i \in \{1, \ldots, n\}$. Equivalently, we must show that

$$\mathrm{norm}_{\mathrm{opt}}(T) = \bigodot_{i=1}^{n} E_i \, \mathrm{norm}_{\mathrm{opt}}(T_i) \tag{6}$$

Indeed, by Eq. (3),

$$\mathrm{gamb}(\mathrm{norm}_{\mathrm{opt}}(T)) = \mathrm{opt}(\mathrm{gamb}(T)|\mathrm{ev}(T))$$

and by the definition of the gamb operator, Eq. (2b) in particular,

$$= \mathrm{opt}\left(\bigoplus_{i=1}^{n} E_i \, \mathrm{gamb}(T_i) \middle| \mathrm{ev}(T) \right)$$

and so by Lemma 18,

$$= \bigoplus_{i=1}^{n} E_i \, \mathrm{opt}(\mathrm{gamb}(T_i)|\mathrm{ev}(T) \cap E_i)$$

so, since $\mathrm{ev}(T) \cap E_i = \mathrm{ev}(T_i)$, and again by Eq. (3),

$$= \bigoplus_{i=1}^{n} E_i \, \mathrm{gamb}(\mathrm{norm}_{\mathrm{opt}}(T_i))$$

Whence, Eq. (6) follows by Lemma 20.

Finally, assume that T has a decision node as its root, that is, $T = \bigsqcup_{i=1}^{n} T_i$. Let $\mathcal{I}$ be the subset of $\{1, \ldots, n\}$ such that $i \in \mathcal{I}$ if and only if N_i is in at least one element of $\mathrm{norm}_{\mathrm{opt}}(T)$. We must establish Eq. (5) for all $i \in \mathcal{I}$. Equivalently, we must show that

$$\mathrm{norm}_{\mathrm{opt}}(T) = \bigsqcup_{i \in \mathcal{I}} \mathrm{norm}_{\mathrm{opt}}(T_i). \tag{7}$$

Indeed, by Eq. (3),

$$\mathrm{gamb}(\mathrm{norm}_{\mathrm{opt}}(T)) = \mathrm{opt}(\mathrm{gamb}(T)|\mathrm{ev}(T))$$

and by the definition of the gamb operator, Eq. (2c),

$$= \mathrm{opt}\left(\bigcup_{i=1}^{n} \mathrm{gamb}(T_i) \middle| \mathrm{ev}(T) \right)$$

and so by Property 5,

$$= \bigcup_{i \in \mathcal{I}^*} \mathrm{opt}(\mathrm{gamb}(T_i)|\mathrm{ev}(T)),$$

where $\mathcal{I}^* = \{i\colon \mathrm{gamb}(T_i) \cap \mathrm{opt}(\mathrm{gamb}(T)|\mathrm{ev}(T)) \neq \emptyset\}$, and so because $\mathrm{ev}(T) = \mathrm{ev}(T_i)$, and again by Eq. (3),

$$= \bigcup_{i \in \mathcal{I}^*} \mathrm{gamb}(\mathrm{norm}_{\mathrm{opt}}(T_i)).$$

Hence, the conditions of Lemma 21 are satisfied, and $\mathcal{I}^* = \mathcal{I}$ by Lemma 19, so Eq. (7) is established. □

5.3 Backward Induction

A practical problem when solving decision trees using $\mathrm{norm}_{\mathrm{opt}}$, is that the set of normal form decisions of a tree T grows very fast with its size, and so $\mathrm{gamb}(T)$ may have many elements. For this reason, elsewhere [3, 2], we have suggested the following backward induction method, which generalizes classical backward induction to arbitrary choice functions. To express this most conveniently, we first extend the $\mathrm{norm}_{\mathrm{opt}}$ operator to act upon sets of decision trees.

Definition 23. *Given any set $\mathcal{T}$ of consistent decision trees, where $\mathrm{ev}(T) = A$ for all $T \in \mathcal{T}$,*

$$\mathrm{norm}_{\mathrm{opt}}(\mathcal{T}) = \{U \in \mathrm{nfd}(\mathcal{T}) : \mathrm{gamb}(U) \subseteq \mathrm{opt}(\mathrm{gamb}(\mathcal{T})|A)\}.$$

Definition 24. *The normal form operator $\mathrm{back}_{\mathrm{opt}}$ is defined for any consistent decision tree T through:*

- *If a tree T consists of only a leaf with reward $r \in \mathcal{R}$, then $\mathrm{back}_{\mathrm{opt}}(T) = \{T\}$.*
- *If a tree T has a chance node as root, that is, $T = \bigodot_{i=1}^{n} E_i T_i$, then*

$$\mathrm{back}_{\mathrm{opt}}(T) = \mathrm{norm}_{\mathrm{opt}}\left(\bigodot_{i=1}^{n} E_i \,\mathrm{back}_{\mathrm{opt}}(T_i)\right)$$

- *If a tree T has a decision node as root, that is, if $T = \bigsqcup_{i=1}^{n} T_i$, then*

$$\mathrm{back}_{\mathrm{opt}}(T) = \mathrm{norm}_{\mathrm{opt}}\left(\bigsqcup_{i=1}^{n} \mathrm{back}_{\mathrm{opt}}(T_i)\right).$$

If $\mathrm{back}_{\mathrm{opt}}$ always yields the same normal form solution as $\mathrm{norm}_{\mathrm{opt}}$, we can use the former as an efficient way of calculating the latter. In [2] we show that the following four properties are necessary and sufficient for $\mathrm{back}_{\mathrm{opt}}$ to coincide with $\mathrm{norm}_{\mathrm{opt}}$.

Property 7 (Backward conditioning property)**.** *Let A and B be events such that $A \cap B \neq \emptyset$ and $\overline{A} \cap B \neq \emptyset$, and let $\mathcal{X}$ be a non-empty finite $A \cap B$-consistent set of gambles, with $\{X, Y\} \subseteq \mathcal{X}$ such that $AX = AY$. Then $X \in \mathrm{opt}(\mathcal{X}|A \cap B)$ implies $Y \in \mathrm{opt}(\mathcal{X}|A \cap B)$ whenever there is a non-empty finite $\overline{A} \cap B$-consistent set of gambles $\mathcal{Z}$ such that, for at least one $Z \in \mathcal{Z}$,*

$$AX \oplus \overline{A}Z \in \mathrm{opt}(A\mathcal{X} \oplus \overline{A}\mathcal{Z}|B).$$

Property 8 (Insensitivity of optimality to the omission of non-optimal elements)**.** *For any event $A \neq \emptyset$, and any non-empty finite A-consistent sets of gambles $\mathcal{X}$ and $\mathcal{Y}$,*

$$\mathrm{opt}(\mathcal{X}|A) \subseteq \mathcal{Y} \subseteq \mathcal{X} \Rightarrow \mathrm{opt}(\mathcal{Y}|A) = \mathrm{opt}(\mathcal{X}|A).$$

Property 9 (Preservation of non-optimality under the addition of elements)**.** *For any event $A \neq \emptyset$, and any non-empty finite A-consistent sets of gambles $\mathcal{X}$ and $\mathcal{Y}$,*

$$\mathcal{Y} \subseteq \mathcal{X} \Rightarrow \mathrm{opt}(\mathcal{Y}|A) \supseteq \mathrm{opt}(\mathcal{X}|A) \cap \mathcal{Y}.$$

Property 10 (Backward mixture property)**.** *For any events A and B such that $B \cap A \neq \emptyset$ and $B \cap \overline{A} \neq \emptyset$, any $B \cap \overline{A}$-consistent gamble Z, and any non-empty finite $B \cap A$-consistent set of gambles $\mathcal{X}$,*

$$\mathrm{opt}\left(A\mathcal{X} \oplus \overline{A}Z|B\right) \subseteq A\,\mathrm{opt}(\mathcal{X}|A \cap B) \oplus \overline{A}Z.$$

Theorem 25 (Backward induction theorem [2])**.** *The following conditions are equivalent.*

(A) For any consistent decision tree T, it holds that $\mathrm{back}_{\mathrm{opt}}(T) = \mathrm{norm}_{\mathrm{opt}}(T)$.

(B) opt satisfies Properties 7, 8, 9, and 10.

Obviously, Property 1 implies Property 7, and Property 3 implies Property 10. Also,

Lemma 26. *Property 2 implies Properties 8 and 9.*

Hence, from Theorems 22 and 25, we conclude:

Corollary 27. *If $\mathrm{norm}_{\mathrm{opt}}$ is factual, then $\mathrm{norm}_{\mathrm{opt}} = \mathrm{back}_{\mathrm{opt}}$.*

Factuality is, however, not necessary for backward induction. For example, it is easy to see that point-wise dominance satisfies Properties 1, 8, 9, and 10, but as we saw in Example 12, it is counterfactual.

Backward induction does imply a weaker version of factuality: $\mathrm{st}_N(\mathrm{norm}(T)) \subseteq \mathrm{norm}(\mathrm{st}_N(T))$.

5.4 Total Preordering

From Theorem 22 and Lemma 16, we have:

Corollary 28. *If $\mathrm{norm}_{\mathrm{opt}}$ is factual then, for all $A \neq \emptyset$, $\mathrm{opt}(\cdot|A)$ is induced by a total preorder.*

This constitutes a strong restriction on opt. Indeed, without consideration of factuality, a choice function that is not a total preorder may be desirable in some circumstances. When one has limited information about the relative likelihood of the events or the relative values of the rewards, one may wish to use a choice function that allows no preference between gambles, but does not consider them equivalent.

For example, if one is working with coherent lower previsions, one may consider the choice functions E-admissibility, maximality, and interval dominance, but none of these corresponds to a total preorder.

	Property						
	1	2	3	7	8	9	10
E-admissibility	✓		✓	✓	✓	✓	✓
Maximality	✓		✓	✓	✓	✓	✓
Γ-maximin	✓	✓		✓	✓	✓	
Interval Dominance	✓			✓	✓	✓	

Table 2: Properties of various choice functions.

Anyone wishing to use these choice functions to solve sequential decision problems must either abandon factuality or seek an alternative operator to norm_{opt}.

Those who prefer their choice functions to give a total preorder, on the other hand, can use factuality to justify this preference. Indeed, without consideration of factuality and sequential decisions, it is much harder to justify a total preorder than it is to justify simpler conditions such as Properties 8 and 9: see for instance Luce and Raiffa [7, pp. 288-289], where Axioms 6, 7, and 7″ correspond to Properties 8, 9, and 2.

6 Conclusion

We defined factuality for extensive and normal forms. We found necessary and sufficient conditions for a choice function to induce a factual normal form operator. These turned out to be similar to, but stronger than, those for backward induction to work.

While many choice functions satisfy Property 1, Properties 2 and 3 are perhaps more restrictive than one would like. Is counterfactuality acceptable? We believe that factuality is a desirable property and one should think carefully before using a counterfactual operator. On the other hand, if one is attracted to the three properties for other reasons, then factuality gives them a strong justification.

Choice functions based on imprecise probability will typically violate at least one of Properties 2 and 3: Table 2 summarizes the properties satisfied by common choice functions. If one wishes to be factual in such cases, norm_{opt} cannot be used. Choice functions that induce factual extensive form operators are easier to find, particularly in the case of violations of Property 2 only: an example is sec_O in [11, p. 286]; also see Kikuti et al. [5]. Further investigation of factuality of extensive form operators, and in particular their relationships with backward induction and normal form operators, has been omitted due to lack of space.

Finally we mention that using counterfactuals is common in the field of causal inference [9]. This paper is quite different in character: we have not been concerned at all with causality and the use counterfactuals in causal inference. Instead, we have simply determined for what choice functions counterfactuals occur when solving decision trees.

Acknowledgements We thank the referees for useful feedback. EPSRC supports the first author.

References

[1] P. Hammond. Consequentialist foundations for expected utility. *Theory and Decision*, 25(1):25–78, Jul 1988.

[2] N. Huntley and M. C. M. Troffaes. Normal form backward induction for decision trees under arbitrary choice functions. Submitted.

[3] N. Huntley and M. C. M. Troffaes. An efficient normal form solution to decision trees with lower previsions. In *Soft Methods for Handling Variability and Imprecision*, Advances in Soft Computing, pages 419–426. Springer, Sep 2008.

[4] J. Jaffray. Rational decision making with imprecise probabilities. In *1st International Symposium on Imprecise Probabilities and Their Applications*, 1999.

[5] D. Kikuti, F. Cozman, and C. P. de Campos. Partially ordered preferences in decision trees: Computing strategies with imprecision in probabilities. In *IJCAI-05 Multidisciplinary Workshop on Advances in Preference Handling*, pages 118–123, 2005.

[6] D. V. Lindley. *Making Decisions*. Wiley, London, 2nd edition, 1985.

[7] R. D. Luce and H. Raiffa. *Games and Decisions: introduction and critical survery*. Wiley, 1957.

[8] M. J. Machina. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature*, 27(1622-1688), 1989.

[9] S. L. Morgan and C. Winship. *Counterfactuals and causal inference*. Cambridge University Press, Cambridge, 2007.

[10] H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory*. Harvard University Press, 1961.

[11] T. Seidenfeld. Decision theory without 'independence' or without 'ordering': What is the difference? *Economics and Philosophy*, 4:267–290, 1988.

[12] A. K. Sen. Social choice theory: A re-examination. *Econometrica*, 45(1):53–89, 1977.

Almost Bayesian Assignments and Conditional Independence (a contribution to Dempster-Shafer theory of evidence)

Radim Jiroušek
Faculty of Management, University of Economics
Jindřichův Hradec
and
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
radim@utia.cas.cz

Abstract

In the paper we introduce a family of almost Bayesian basic assignments, which slightly extends Bayesian basic assignments. This extension incorporates all the distributions that can be created from low-dimensional Bayesian basic assignments by application of the operator of composition, and simultaneously preserves the property of Bayesian basic assignments concerning the number of focal elements: it does not exceed cardinality of the frame of discernment. The other goal of the paper is to propagate a new way of definition of conditional independence relation in D-S theory. It follows ideas of P. P. Shenoy from [7], where the author defines the notion of conditional independence for valuation-based system based on his operation of "combination". Here we do the same but using the operator of "composition". The notion of independence we get in this way seems to meet better the general requirements on conditional independence relation for basic assignments.

Keywords. Dempster-Shafer theory of evidence, multidimensionality, operator of composition, conditional independence, semigraphoids.

1 Introduction

Regarding purely computational point of view, the greatest disadvantage of Dempster-Shafer theory of evidence (D-S) is that in contrast to probabilistic or possibilistic models, which can be described by the respective density functions (i.e. point functions), D-S models must be described by set functions. It means that while the number of necessary parameters of probabilistic or possibilistic models grows exponentially with the number of dimensions, for D-S models one needs a superexponential number of parameters.

It is known from theory of Bayesian networks (or graphical Markov models, in general) that the number of parameters can be drastically decreased by utilization of properties of conditional independence relations valid for the modelled situation. This was among the reasons why we designed an alternative approach for multidimensional probability distribution representation based on so called *operator of composition* [2]. The basic idea of these models is very simple: multidimensional models are assembled (composed) from a system of low-dimensional distributions by the operator of composition (in a specified order). Later on, Vejnarová introduced an analogous operator also for composition of possibility distributions and showed it manifested similar properties as its probabilistic counterpart [10, 11]. Recently we designed the operator of composition also for basic assignments in D-S theory of evidence [5] and proved that it met all the required properties necessary for multidimensional models representation [3, 4].

However, it is not the goal of this paper to publicize advantageous properties of the operator of composition for basic assignments. The goal of this contribution is twofold. The first one is to show that there exists a family of basic assignments, for specification of which one does not need more parameters than for probabilistic models and yet it enables modelling some type of ignorance (Section 4). The other goal is to show that if the conditional independence for basic assignments is defined with the help of the operator of composition (which was already done in [3]) one can prove semigraphoid axioms from a small number of operator's basic properties. This is done in Section 5.

2 Basic notion

Set notation

In the whole paper we shall deal with a finite number of variables $X_1, X_2, \ldots, X_n$ each of which is specified by a finite set $\mathbf{X}_i$ of its values. So, we will consider multidimensional space of discernment

$$\mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_n,$$

and its *subspaces*. For $K \subset N = \{1, 2, \ldots, n\}$, $\mathbf{X}_K$ denotes a Cartesian product of those $\mathbf{X}_i$, for which $i \in K$:

$$\mathbf{X}_K = \bigtimes_{i \in K} \mathbf{X}_i.$$

A *projection* of $x = (x_1, x_2, \ldots, x_n) \in \mathbf{X}_N$ into $\mathbf{X}_K$ will be denoted $x^{\downarrow K}$, i.e. for $K = \{i_1, i_2, \ldots, i_\ell\}$

$$x^{\downarrow K} = (x_{i_1}, x_{i_2}, \ldots, x_{i_\ell}) \in \mathbf{X}_K.$$

Analogously, for $K \subset L \subseteq N$ and $A \subset \mathbf{X}_L$, $A^{\downarrow K}$ will denote a *projection* of A into $\mathbf{X}_K$:

$$A^{\downarrow K} = \{y \in \mathbf{X}_K : \exists x \in A \;\; (y = x^{\downarrow K})\}.$$

Let us remark that we do not exclude situations when $K = \emptyset$. In this case $A^{\downarrow \emptyset} = \emptyset$.

Set $A \subseteq \mathbf{X}_K$ is said to be a *point-cylinder* if it can be expressed as a Cartesian product of a singleton and a subspace $\mathbf{X}_L$. More precisely: a point-cylinder is a set $A \subseteq \mathbf{X}_K$ for which there exists an index set (possibly empty) $L \subseteq K$ such that $|C^{\downarrow L}| \leq 1$ and

$$C = C^{\downarrow L} \times \mathbf{X}_{K \setminus L}.$$

Let us stress that if $L = \emptyset$ then $C = \mathbf{X}_K$ (it is the only situation when $|C^{\downarrow L}| < 1$), and when $L = K$ then $|C| = 1$.

In addition to the projection, in this text we will need also the opposite operation which will be called a join. By a *join* of two sets $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_L$ we will understand a set

$$A \otimes B = \{x \in \mathbf{X}_{K \cup L} : x^{\downarrow K} \in A \;\; \& \;\; x^{\downarrow L} \in B\}.$$

Notice that if K and L are disjoint then the join of the corresponding sets is just their Cartesian product

$$A \otimes B = A \times B.$$

For $K = L$, $A \otimes B = A \cap B$. If $K \cap L \neq \emptyset$ and $A^{\downarrow K \cap L} \cap B^{\downarrow K \cap L} = \emptyset$ then also $A \otimes B = \emptyset$.

In one of the following proofs we will need the following (rather technical) property of set joins.

Lemma 1. *Let $K_1 \cap K_2 \subseteq L \subseteq K_2 \subseteq N$. Then for any $C \subseteq \mathbf{X}_{K_1 \cup K_2}$ the following condition* (a) *holds if and only if both conditions* (b) *and* (c) *hold true.*

(a) $C = C^{\downarrow K_1} \otimes C^{\downarrow K_2}$;

(b) $C^{\downarrow K_1 \cup L} = C^{\downarrow K_1} \otimes C^{\downarrow L}$;

(c) $C = C^{\downarrow K_1 \cup L} \otimes C^{\downarrow K_2}$.

Proof. Let us prove the assertion in three steps. First, however, let us realize that

$$x \in C \implies \left(x^{\downarrow K_1} \in C^{\downarrow K_1} \;\&\; x^{\downarrow K_2} \in C^{\downarrow K_2}\right),$$

and therefore $C = C^{\downarrow K_1} \otimes C^{\downarrow K_2}$ is equivalent to

$$\forall x \in \mathbf{X}_{K_1 \cup K_2} \quad \left(x^{\downarrow K_1} \in C^{\downarrow K_1} \;\&\; x^{\downarrow K_2} \in C^{\downarrow K_2} \Longrightarrow x \in C\right).$$

(a) $\Longrightarrow$ (b).
Consider $x \in \mathbf{X}_{K_1 \cup L}$, such that $x^{\downarrow K_1} \in C^{\downarrow K_1}$ and $x^{\downarrow L} \in C^{\downarrow L}$. Since $x^{\downarrow L} \in C^{\downarrow L}$ there must exists (at least one) $y \in C^{\downarrow K_2}$, for which $y^{\downarrow L} = x^{\downarrow L}$. Now construct $z \in \mathbf{X}_{K_1 \cup K_2}$ for which $z^{\downarrow K_1} = x^{\downarrow K_1}$ and $z^{\downarrow K_2} = y$ (it is possible because $y^{\downarrow L} = x^{\downarrow L}$). From this construction we see that $z^{\downarrow K_1 \cup L} = x$. Therefore $z^{\downarrow K_1} = x^{\downarrow K_1} \in C^{\downarrow K_1}$ and $z^{\downarrow K_2} = y \in C^{\downarrow K_2}$ form which, because we assume that (a) holds, we get that $z \in C$, and therefore also $x = z^{\downarrow K_1 \cup L} \in C^{\downarrow K_1 \cup L}$.

(a) $\Longrightarrow$ (c).
Consider now $x \in \mathbf{X}_{K_1 \cup K_2}$, for which its projections $x^{\downarrow K_1 \cup L} \in C^{\downarrow K_1 \cup L}$ and $x^{\downarrow K_2} \in C^{\downarrow K_2}$. From $x^{\downarrow K_1 \cup L} \in C^{\downarrow K_1 \cup L}$ we immediately get that $x^{\downarrow K_1} \in C^{\downarrow K_1}$, which in combination with $x^{\downarrow K_2} \in C^{\downarrow K_2}$ (due to the assumption (a)) yields that $x \in C$.

(b) & (c) $\Longrightarrow$ (a).
Consider $x \in \mathbf{X}_{K_1 \cup K_2}$ such that $x^{\downarrow K_1} \in C^{\downarrow K_1}$ and $x^{\downarrow K_2} \in C^{\downarrow K_2}$. From the last property one gets also $x^{\downarrow L} \in C^{\downarrow L}$, which, in combination with $x^{\downarrow K_1} \in C^{\downarrow K_1}$ gives, because (b) holds true, that $x^{\downarrow K_1 \cup L} \in C^{\downarrow K_1 \cup L}$. And the last property in combination with $x^{\downarrow K_2} \in C^{\downarrow K_2}$ yields the required $x \in C$. □

Assignment notation

The role of a probability distribution from a probability theory is in Dempster-Shafer theory played by any of the set functions: belief function, plausibility function or basic (*probability or belief*) assignment. Knowing one of them, one can deduce the two remaining. In this paper we shall use exclusively basic assignments.

A *basic assignment* m on $\mathbf{X}_K$ ($K \subseteq N$) is a function

$$m : \mathcal{P}(\mathbf{X}_K) \longrightarrow [0, 1],$$

for which

$$\sum_{\emptyset \neq A \subseteq \mathbf{X}_N} m(A) = 1.$$

For the sake of this paper it is reasonable to consider only normalized basic assignments, for which $m(\emptyset)$ equals always 0. If $m(A) > 0$, then A is said to be a *focal element* of m.

Having a basic assignment m on $\mathbf{X}_K$ one can consider its *marginal assignment* on $\mathbf{X}_L$ (for $L \subseteq K$), which is defined (for each $\emptyset \neq B \subseteq \mathbf{X}_L$):

$$m^{\downarrow L}(B) = \sum_{A \subseteq \mathbf{X}_K : A^{\downarrow L} = B} m(A).$$

Basic assignment m is said to be *Bayesian* if all its focal elements are *singletons*, i.e.

$$m(A) > 0 \implies |A| = 1.$$

In this case, namely, both the other two functions, belief Bel and plausibility Pl which are defined by the following formulas (for all $A \subseteq \mathbf{X}_K$)

$$Bel(A) = \sum_{B \subseteq A} m(A),$$
$$Pl(A) = 1 - Bel(\bar{A}),$$

are normalized additive functions, and therefore probability distributions.

Another special case is represented by simple basic assignments. Basic assignments m on $\mathbf{X}_K$ is called *simple* if there exists A ($\emptyset \neq A \subset \mathbf{X}_K$) and a positive number a such that $m(A) = a$ and $m(\mathbf{X}_K) = 1 - a$.

3 Operator of composition

Originally, the operator of composition was designed in probability theory as a tool enabling creation of multidimensional probability distributions - multidimensional models - by successive composition of low-dimensional distributions. The basic idea of this operator was simple. It generalized the fact that one can construct a 3-dimensional probability distribution $P(X, Y, Z)$ from two 2-dimensional ones $Q(X, Y)$ and $R(Y, Z)$ just by assigning

$$P(X, Y, Z) = Q(X, Y) \cdot R(Z|Y).$$

In this case P reflects all the information contained in Q, because evidently $P(X, Y) = Q(X, Y)$, and some of the information contained in R ($P(Z|Y) = R(Z|Y)$). Moreover, P does not contain any additional information, because for this probability distribution variables X and Z are conditionally independent given variable Y.

Introduction of the probabilistic operator of composition opened a study of a new area called compositional models, which was an alternative to Bayesian networks, or to Graphical Markov models in general. Though it appeared that Bayesian networks and compositional models described exactly the same class of probability distributions, study of a new type of models appeared useful. First of all it offered new points of view to multidimensional probability distribution representation. In addition to this, compositional models were in some situations more advantageous from the computational point of view (some of the marginal distributions, computation of which may be algorithmically rather expensive, were in a compositional model expressed explicitly).

Later, the operator of composition was designed and studied in possibility theory by Vejnarová [10]. Being inspired by Didier Dubois, we introduced the operator of composition also for basic assignments [5]; this definition is presented below. In that paper we also showed that if the operator of composition is applied to Bayesian basic assignments it usually yields the Bayesian basic assignment, which corresponds to the probability distribution, which is constructed by the probabilistic operator of composition from the respective probability distributions. The only exception from this situation occurs when composing basic assignments corresponding to probability distributions, for which their probabilistic composition is not defined. In such a case, result of composition of such Bayesian basic assignments is not Bayesian. In the next section we will reveal the main characteristics of such basic assignments.

Definition 1. For two arbitrary basic assignments m_1 on $\mathbf{X}_K$ and m_2 on $\mathbf{X}_L$ ($K \neq \emptyset \neq L$) a *composition* $m_1 \triangleright m_2$ is defined for each $C \subseteq \mathbf{X}_{K \cup L}$ by one of the following expressions:

[a] if $m_2^{\downarrow K \cap L}(C^{\downarrow K \cap L}) > 0$ and $C = C^{\downarrow K} \otimes C^{\downarrow L}$ then

$$(m_1 \triangleright m_2)(C) = \frac{m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L})}{m_2^{\downarrow K \cap L}(C^{\downarrow K \cap L})};$$

[b] if $m_2^{\downarrow K \cap L}(C^{\downarrow K \cap L}) = 0$ and $C = C^{\downarrow K} \times \mathbf{X}_{L \setminus K}$ then

$$(m_1 \triangleright m_2)(C) = m_1(C^{\downarrow K});$$

[c] in all other cases $(m_1 \triangleright m_2)(C) = 0$.

Before illustrating the operator of composition on a simple example, let us remark that three expressions in Definition 1 correspond to three situations, which occur when one wants to define a basic assignments possessing those properties we highlighted when speaking about the probability distribution $P(X, Y, Z) = Q(X, Y) \cdot R(Z|Y)$. Point [a], in a way, directly corresponds to this well-known probabilistic formula. It disseminates the mass $m_1(C^{\downarrow K})$ into the respective subsets $C \subseteq \mathbf{X}_{K \cup L}$. The information describing the way how this mass is disseminated is taken over from m_2. Point [b] is applicable when

Table 1: 1-dimensional basic assignments m_1 and m_2.

$A \subseteq \mathbf{X}_1$	$m_1(A)$
$\{a\}$	0.5
$\{\bar{a}\}$	0.1
$\{a, \bar{a}\}$	0.4

$B \subseteq \mathbf{X}_2$	$m_2(B)$
$\{b\}$	0.5
$\{\bar{b}\}$	0.5

$m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) = 0$ and therefore m_2 does not determine the way how to disseminate the respective mass. Therefore the whole mass $m_1(C^{\downarrow K})$ is assigned to the least specific set: $C = C^{\downarrow K} \times \mathbf{X}_{L\setminus K}$ (expressing in this way maximal ignorance). Eventually, point [c] guarantees that no additional information is added to the resulting basic assignment $m_1 \triangleright m_2$. It assigns zero mass to all those subsets of $\mathbf{X}_{K\cup L}$, whose positive values would violate the notion of the required conditional independence (see e.g. [1]).

Example 1. Consider two 1-dimensional basic assignments[1] m_1, m_2 from Table 1, which are defined on $\mathbf{X}_1 = \{a, \bar{a}\}$ and $\mathbf{X}_2 = \{b, \bar{b}\}$, respectively.

Their composition $m_1 \triangleright m_2$ is in Table 2. Notice, that this composed basic assignment has only 6 focal elements, which means that for the remaining $(2^4 - 1) - 6 = 9$ subsets of $\mathbf{X}_1 \times \mathbf{X}_2$, values of $m_1 \triangleright m_2$ equal 0. It is the case of two groups of subsets. As for three subsets

$$\{ab, a\bar{b}\} = \{a\} \otimes \mathbf{X}_2,$$
$$\{\bar{a}b, \bar{a}\bar{b}\} = \{\bar{a}\} \otimes \mathbf{X}_2,$$
$$\{ab, a\bar{b}, \bar{a}b, \bar{a}\bar{b}\} = \mathbf{X}_1 \otimes \mathbf{X}_2,$$

their values of $m_1 \triangleright m_2$ are assigned by point [a] of Definition 1 and equal 0 because $m_2(\{b, \bar{b}\}) = 0$. On the other hand side, to the remaining six subsets

$$\{ab, \bar{a}\bar{b}\},$$
$$\{a\bar{b}, \bar{a}b\},$$
$$\{ab, a\bar{b}, \bar{a}b\},$$
$$\{ab, a\bar{b}, \bar{a}\bar{b}\},$$
$$\{ab, \bar{a}b, \bar{a}\bar{b}\},$$
$$\{a\bar{b}, \bar{a}b, \bar{a}\bar{b}\},$$

values of $m_1 \triangleright m_2$ are assigned by point [c] of Definition 1, because for these subsets it does not hold that $C = C^{\downarrow\{1\}} \otimes C^{\downarrow\{2\}}$. Assigning a positive value to any of these subsets we would, in a way, introduce a dependence of variables X_1 and X_2.

[1] In all examples in this paper we record in tables only focal elements. It means that for all subsets of space of discernment which are not included in the respective tables their respective basic assignment equals 0.

Table 2: Composed basic assignment $m_1 \triangleright m_2$.

$C \subseteq \mathbf{X}_1 \times \mathbf{X}_2$	$(m_1 \triangleright m_2)(C)$
$\{ab\}$	0.25
$\{a\bar{b}\}$	0.25
$\{\bar{a}b\}$	0.05
$\{\bar{a}\bar{b}\}$	0.05
$\{ab, \bar{a}b\}$	0.20
$\{a\bar{b}, \bar{a}\bar{b}\}$	0.20

Let us present the most important properties of the operator of composition for basic assignments.

Lemma 2. *Let $K, L \subseteq N$. For arbitrary basic assignments m_1, m_2 defined on $\mathbf{X}_K, \mathbf{X}_L$, respectively*

(i) *$m_1 \triangleright m_2$ is a basic assignment on $\mathbf{X}_{K\cup L}$;*

(ii) $(m_1 \triangleright m_2)^{\downarrow K} = m_1$*;*

(iii) $m_1 \triangleright m_2 = m_2 \triangleright m_1 \iff m_1^{\downarrow K\cap L} = m_2^{\downarrow K\cap L}$*;*

(iv) $L \supseteq M \supseteq (K \cap L)$
$\implies m_1 \triangleright m_2 = (m_1 \triangleright m_2^{\downarrow M}) \triangleright m_2$*;*

Proof. The first three properties were proved in [5]: properties (i)-(iii) are properties (i)-(iii) of Lemma 1. Thus, what has remained to be proved is just property (iv).

So, our goal is to show that for basic assignments m_1, m_2 and for any M such that $L \supseteq M \supseteq K \cap L$

$$(m_1 \triangleright m_2)(C) = ((m_1 \triangleright m_2^{\downarrow M}) \triangleright m_2)(C).$$

holds true for any $C \subseteq \mathbf{X}_{K\cup L}$.

The proof will be performed in three steps corresponding to cases [a], [b], [c] of Definition 1.

Ad [a]. Assume that $C = C^{\downarrow K} \otimes C^{\downarrow L}$ and $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) > 0$. From this we get from Lemma 1 that also $C^{\downarrow K\cup M} = C^{\downarrow K} \otimes C^{\downarrow M}$, and therefore (since $K \cap L = K \cap M$)

$$(m_1 \triangleright m_2^{\downarrow M})(C^{\downarrow K\cup M}) = \frac{m_1(C^{\downarrow K}) \cdot m_2^{\downarrow M}(C^{\downarrow M})}{m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L})}.$$

In the rest of this step we have to distinguish two situations depending whether $m_2^{\downarrow M}(C^{\downarrow M})$ equals 0 or not.

If $m_2^{\downarrow M}(C^{\downarrow M}) > 0$ (realize that in this case also

$m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) > 0$) then

$$\begin{aligned}((m_1 \triangleright m_2^{\downarrow M}) \triangleright m_2)(C) &= \frac{(m_1 \triangleright m_2^{\downarrow M})(C^{\downarrow K\cup M}) \cdot m_2(C^{\downarrow L})}{m_2^{\downarrow M}(C^{\downarrow M})} \\ &= \frac{\frac{m_1(C^{\downarrow K})\cdot m_2^{\downarrow M}(C^{\downarrow M})}{m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L})} \cdot m_2(C^{\downarrow L})}{m_2^{\downarrow M}(C^{\downarrow M})} \\ &= \frac{m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L})}{m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L})} = (m_1 \triangleright m_2)(C).\end{aligned}$$

If $m_2^{\downarrow M}(C^{\downarrow M}) = 0$ then, according to Definition 1, either

$$((m_1 \triangleright m_2^{\downarrow M}) \triangleright m_2)(C) = (m_1 \triangleright m_2^{\downarrow M})(C^{\downarrow K\cup M}),$$

in case that $C = C^{\downarrow K\cup M} \otimes \mathbf{X}_{L\setminus M}$, or

$$((m_1 \triangleright m_2^{\downarrow M}) \triangleright m_2)(C) = 0,$$

in opposite case. However, in this case also

$$(m_1 \triangleright m_2^{\downarrow M})(C^{\downarrow K\cup M}) = \frac{m_1(C^{\downarrow K}) \cdot m_2^{\downarrow M}(C^{\downarrow M})}{m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L})} = 0,$$

and therefore $((m_1 \triangleright m_2^{\downarrow M}) \triangleright m_2)(C) = 0$ regardless of the form of $C^{\downarrow L\setminus M}$ (i.e. for both situations: $C^{\downarrow L\setminus M} = \mathbf{X}_{L\setminus M}$ and $C^{\downarrow L\setminus M} \neq \mathbf{X}_{L\setminus M}$). Taking into consideration the fact that in the considered situation (i.e. $m_2^{\downarrow M}(C^{\downarrow M}) = 0$) also $m_2(C^{\downarrow L}) = 0$, and therefore also

$$(m_1 \triangleright m_2)(C) = \frac{m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L})}{m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L})} = 0,$$

we have finished the first step of the proof.

Ad [b]. Now we assume that $C = C^{\downarrow K} \otimes \mathbf{X}_{L\setminus K}$, and that $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) = 0$. In this case, naturally, also $m_2^{\downarrow M}(C^{\downarrow M}) = 0$ and $C = C^{\downarrow K} \otimes \mathbf{X}_{M\setminus K} \otimes \mathbf{X}_{L\setminus M}$. Therefore, according to case [b] of Definition 1,

$$(m_1 \triangleright m_2^{\downarrow M})(C^{\downarrow K\cup M}) = m_1(C^{\downarrow K}),$$

and because of the same reasons also

$$\begin{aligned}((m_1 \triangleright m_2^{\downarrow M}) \triangleright m_2)(C) &= (m_1 \triangleright m_2^{\downarrow M})(C^{\downarrow K\cup M}) \\ &= m_1(C^{\downarrow K}).\end{aligned}$$

In this case also $(m_1 \triangleright m_2)(C) = m_1(C^{\downarrow K})$, and we have finished the second step of the proof.

Ad [c]. The last step is trivial. In this case, as the reader can immediately see, both $((m_1 \triangleright m_2^{\downarrow M}) \triangleright m_2)(C)$ and $(m_1 \triangleright m_2)(C)$ equal 0 and therefore they equal to each other. □

Table 3: 2-dimensional basic assignments m_3 and m_4.

$A \subseteq \mathbf{X}_{\{1,2\}}$	$m_3(A)$	$B \subseteq \mathbf{X}_{\{2,3\}}$	$m_4(B)$
$\{a\bar{b}\}$	0.5	$\{bc\}$	0.5
$\{\bar{a}b\}$	0.1	$\{\bar{b}\bar{c}\}$	0.2
$\{ab, \bar{a}b\}$	0.4	$\{b\bar{c}, \bar{b}c\}$	0.3

Table 4: Basic assignments $m_3 \triangleright m_4$ and $m_4 \triangleright m_3$.

$C \subseteq \mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3$	$m_3 \triangleright m_4$	$m_4 \triangleright m_3$
$\{a\bar{b}\bar{c}\}$	0.5	0.2
$\{\bar{a}bc\}$	0.1	0.1
$\{abc, \bar{a}bc\}$	0.4	0.4
$\{ab\bar{c}, a\bar{b}c, \bar{a}b\bar{c}, \bar{a}\bar{b}c\}$		0.3

Example 2. Property (iii) of the previous lemma says that for consistent basic assignments the operator of composition is commutative. Since any couple of basic assignments defined on non-overlapping frames of discernment are consistent (because $m^{\downarrow\emptyset} = 1$), for basic assignments m_1 and m_2 from Table 1 $m_1 \triangleright m_2 = m_2 \triangleright m_1$. Therefore, if we want to illustrate non-commutativity of this operator we have to consider overlapping frames of discernment[2].

Consider basic assignments m_3, m_4 from Table 3. The reader can easily see that when computing $m_3 \triangleright m_4$, all the focal elements are computed according to case [a] of Definition 1. There are only three sets $C \subseteq \mathbf{X}_{\{1,2,3\}}$, for which $C = C^{\downarrow\{1,2\}} \otimes C^{\downarrow\{2,3\}}$, and for which both $m_3(C^{\downarrow\{1,2\}})$ and $m_3(C^{\downarrow\{2,3\}})$ are positive, namely

$$\begin{aligned}\{a\bar{b}\bar{c}\} &= \{a\bar{b}\} \otimes \{\bar{b}\bar{c}\}, \\ \{\bar{a}bc\} &= \{\bar{a}b\} \otimes \{bc\}, \\ \{abc, \bar{a}bc\} &= \{ab, \bar{a}b\} \otimes \{bc\}.\end{aligned}$$

On the other hand, when computing $m_4 \triangleright m_3$ there appears set $C = \{b\bar{c}, \bar{b}c\} \times \mathbf{X}_1$, for which $m_3(C^{\downarrow\{1,2\}}) = 0$ and therefore value $(m_4 \triangleright m_3)(C)$ is assigned by point [b] of Definition 1. The resulting basic assignment $m_4 \triangleright m_3$ is also recorded in Table 4.

Remark: In previous papers [5, 4] we showed a number of other properties of the operator of composition

[2]The simplest example of non-commutativity of the operator of composition can be got by considering two different assignments on the same frame of discernment. Then using property (i) of Lemma 2 we see that their composition is defined on the same frame of discernment as the considered original assignments and the non-commutativity of the operator $\triangleright$ immediately follows from property (ii) of Lemma 2.

for basic assignments, especially those useful for construction of multidimensional models. The four properties included in the previous lemma are those, which are sufficient to prove that conditional independence, if introduced with the help of the operator of composition (as done in Section 5), meets the semigraphoid axioms. In a way it is surprising that such a small group of elementary properties is sufficient. In connection with this fact a question arises whether the presented four properties are independent, whether some of them cannot be deduced from the remaining four.

Remark: Let us briefly answer a frequent question what is the relation of the introduced operator of composition and the famous Dempster's rule of combination[3]. Let us stress that the main difference emerges from the different purposes the operators where designed for. While Dempster's rule of combination was designed to have a tool enabling fusion of two basic assignments (the goal is to get a better information about the object than those contained in any of the original basic assignments), the operator of composition combines different descriptions of the object to comprehend all the information contained in original sources. This process corresponds to knowledge integration rather than knowledge fusion.

From the formal point of view this difference is reflected in property (ii) of Lemma 2, which holds for Dempster's rule of combination only in very specific (degenerated) situations. By the way, this difference is also the main reason why we consider the attempts to define a notion of conditional independence with the help of Dempster's rule of combination to be misleading.

4 Almost Bayesian basic assignments

One of the reasons (and from our point of view perhaps the most important) why D-S theory of evidence was designed and why it is in the center of attention of many researchers is the fact that probability theory has difficulties with representing some types of uncertainty; here we have in mind especially ignorance. For example, probability theory can hardly distinguish situation when an integer from $\{1, 2, \ldots, 6\}$ is determined by tossing a fair die, and when it is selected by a totally unknown mechanism (well, the second situation can be described by the set of all possible distributions, however it is rather inconvenient). On the other hand, D-S theory yields very complex models and the corresponding computational procedures are of extremely high algorithmic complexity. Now, we are about to specify a small family of basic assignments extending the set of Bayesian assignments but keeping the computational complexity on the level of probabilistic models. However, we have to admit that this new family, elements of which will be called almost Bayesian basic assignments, is very restrictive.

[3] Detailed study of formal similarities of these two operators will appear in [6].

Definition 2. Basic assignment m on $\mathbf{X}_K$ is called *cylindrical* if all its focal elements are point-cylinders.

Theorem 1. *Let $K, L \subseteq N$ and m_1, m_2 be basic assignments defined on $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively. If m_1, m_2 are cylindrical then $m_1 \triangleright m_2$ is also cylindrical.*

Proof. To prove this assertion we have to realize that a projection $A^{\downarrow K}$ of a point-cylinder A is a point-cylinder. Moreover, join $A \otimes B$ of two point-cylinders A and B is again a point-cylinder (recall that $\emptyset$ is a point-cylinder).

Values of focal elements of basic assignment are computed according to either point [a] or point [b] of Definition 1. In case [a], a positive value can be assigned only if $C = C^{\downarrow K} \otimes C^{\downarrow L}$ and both $C^{\downarrow K}$ and $C^{\downarrow L}$ are point-cylinders. Case [b] is applied only when $C = C^{\downarrow K} \times \mathbf{X}_{L \setminus K}$. So in both cases positive value can be assigned only to point-cylinders. □

Definition 3. Basic assignment m on $\mathbf{X}_K$ is *sparse* if all its focal elements are pairwise disjoint.

Theorem 2. *Let $K, L \subseteq N$ and m_1, m_2 be basic assignments defined on $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively. If m_1, m_2 are sparse then $m_1 \triangleright m_2$ is also sparse.*

Proof. Consider two non-disjoint focal elements C_1, C_2 of $m_1 \triangleright m_2$: $(m_1 \triangleright m_2)(C_1) > 0$ and $(m_1 \triangleright m_2)(C_2) > 0$. Since m_1 is marginal of $m_1 \triangleright m_2$, it is obvious that $C_1^{\downarrow K}$ and $C_2^{\downarrow K}$ are focal elements of m_1. Since we assume that C_1 and C_2 are non-disjoint the same must hold also for their projections

$$C_1^{\downarrow K} \cap C_2^{\downarrow K} \neq \emptyset$$

and therefore, because of our assumption that m_1 is sparse, $C_1^{\downarrow K} = C_2^{\downarrow K}$.

What are the focal elements C of $m_1 \triangleright m_2$, for which $C^{\downarrow K} = C_1^{\downarrow K}$? The answer to this question is offered by Definition 1 (realize that since we are considering focal elements C, values $(m_1 \triangleright m_2)(C)$ are defined by expressions in points [a] or [b]).

If $m_2^{\downarrow K \cap L}(C^{\downarrow K \cap L}) > 0$ then the considered focal elements can be expressed in the form

$$C = C^{\downarrow K} \otimes C^{\downarrow L} = C_1^{\downarrow K} \otimes D,$$

Table 5: Sparse basic assignment m on $\mathbf{X}_{\{1,2\}}$.

$A \subseteq \mathbf{X}_1 \times \mathbf{X}_2$	$m(A)$
$\{ab\}$	0.2
$\{\bar{a}b\}$	0.3
$\{a\bar{b}, \bar{a}\bar{b}\}$	0.5

Table 6: Marginal basic assignments $m^{\downarrow\{1\}}$, $m^{\downarrow\{2\}}$.

$A \subseteq \mathbf{X}_1$	$m^{\downarrow\{1\}}(A)$
$\{a\}$	0.2
$\{\bar{a}\}$	0.3
$\{a, \bar{a}\}$	0.5

$B \subseteq \mathbf{X}_2$	$m^{\downarrow\{2\}}(B)$
$\{b\}$	0.5
$\{\bar{b}\}$	0.5

where $D \subseteq \mathbf{X}_L$ is a focal element of m_2 and $D^{\downarrow K\cap L} = C_1^{\downarrow K\cap L}$. From this one can immediately see that $C_1 = C_1^{\downarrow K} \otimes C_1^{\downarrow L}$ and $C_2 = C_1^{\downarrow K} \otimes C_2^{\downarrow L}$ are disjoint if and only if also focal elements $C_1^{\downarrow L}$ and $C_2^{\downarrow L}$of m_2 are disjoint. In our case, because m_2 is sparse, and because we assume that $C_1 \cap C_2 \neq \emptyset$, it means that $C_1^{\downarrow L} = C_2^{\downarrow L}$, and therefore also $C_1 = C_2$.

In case that $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) = 0$ then the situation is even simpler because in this case there can be only one focal element $C = C_1^{\downarrow K\cap L} \times \mathbf{X}_{L\setminus K}$, which means again that $C_1 = C_2$. □

Remark: It is not difficult to show that a marginal basic assignment of a cylindrical assignment is again cylindrical. However, it is important to realize that, as we illustrate in the following simple example, an analogous property for sparse basic assignments does not hold. Nevertheless, the main advantage of sparse basic assignments is the fact that the number of their focal elements does not exceed the cardinality of the respective frame of discernment, i.e. the number of probabilities necessary to define a general probability distribution.

Example 3. Consider 2-dimensional case with $\mathbf{X}_1 = \{a, \bar{a}\}$ and $\mathbf{X}_2 = \{b, \bar{b}\}$ and basic assignment m in Table 5. From Table 6 one can immediately see that while marginal basic assignment $m^{\downarrow\{2\}}$ is sparse, the other marginal assignment $m^{\downarrow\{1\}}$ is not.

Remark: Now we are ready to answer the question raised at the beginning of the previous section: what are the basic assignments which are obtained from Bayesian basic assignments by a multiple application of the operator of composition? Since all Bayesian assignments are obviously sparse and cylindrical, Theorems 1 and 2 guarantee that the basic assignments corresponding to compositional models from Bayesian basic assignments are also cylindrical and sparse. This fact, somehow, justifies the following definition.

Definition 4. Basic assignment is called *almost Bayesian* if it is sparse and cylindrical.

As said at the beginning of this section, an expressive power of almost Bayesian basic assignments is not too strong. For example, even non-degenerated simple basic assignments are not almost Bayesian. Roughly speaking: Having a Bayesian basic assignment one knows a probability of each point of the frame of discernment. Having an almost Bayesian basic assignment and a fixed point of the frame of discernment one either knows its probability, or knows that it belongs to a cylindrical subset of the frame of discernment among whose elements one cannot make a difference; she knows only the probability of the whole subset. Nevertheless, let us stress once more that the importance of almost Bayesian assignments is in the fact that they describe compositional models constructed from an arbitrary system of low-dimensional probability distributions, which means that even in situations when probabilistic operator of composition is not defined. In this way we are getting a slight extension of probability theory.

5 Conditional independence

In this paper our attention is concentrated on properties of basic assignments which are, in a way, promising from the point of view of computational complexity. Last section was devoted to almost Bayesian basic assignment whose number of focal elements is not higher than the number of probabilities by which a general probability distribution must be specified.

It is well known that efficiency of Bayesian models is based on making the best of the dependence structure of the model, i.e. taking advantage of the knowledge of conditional independence relations [8, 9] holding for the multidimensional distribution in question. This is because the notion of conditional independence in probability theory is equivalent to the notion of *factorization*: for probability distribution P variables X and Z are conditionally independent given variable Y *iff* distribution $P(X, Y, Z)$ is uniquely determined by its marginals $P(X, Y)$ and $P(Y, Z)$. Unfortunately, as shown by Studený [8, 1], the notion of conditional non-interactivity (Shenoy's factorization [7], Studený conditional independence [8]) presented in [1] is *not consistent with marginalization*: there are situations when for two consistent basic assignments there does not exist their common extension with the respective conditional non-interactivity (for more precise explanation see footnote no. 6).

Therefore, in this paper we are going to eliminate this drawback using the definition of conditional independence for basic assignments introduced in [4], which is in fact based on the notion of factorization. Moreover we will present new proofs showing that for this concept all the semigraphoid axioms hold true. These proofs will be based on the fundamental properties of the operator of composition presented in Lemma 2. It should be stressed that the novelty of these proofs is mainly in application of property (iv) of Lemma 2, which seems to be surprisingly weak (and which, in a way, extends property (ii) of the same lemma).

Let us consider an arbitrary basic assignment. We will say that two groups of variables are conditionally independent given the third group of variables if the respective marginal basic assignment can be decomposed (factorized) in the way that it can be expressed as a composition of its respective smaller marginal assignments. Precisely this notion is introduced in the following definition.

Definition 5. Consider a basic assignment m on $\mathbf{X}_N$ and three disjoint index sets $K, L, M \subset N$, $K \neq \emptyset \neq L$. We say that groups of variables X_K and X_L are *conditionally independent given variables* X_M if

$$m^{\downarrow K\cup L\cup M} = m^{\downarrow K\cup M} \triangleright m^{\downarrow L\cup M}.$$

In symbol this fact will be recorded $K \perp\!\!\!\perp_m L \,|\, M$.

Example 4. Consider a basic assignment m on the same 3-dimensional binary frame of discernment as in previous examples: $\mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3$. If variables X_1 and X_2 are independent, i.e. $1 \perp\!\!\!\perp_m 2$, from Definition 1 one can immediately see that for all focal elements $C \subseteq \mathbf{X}_1 \times \mathbf{X}_2$ of the 2-dimensional marginal $m^{\downarrow\{1,2\}}$ it holds that $C = C^{\downarrow\{1\}} \otimes C^{\downarrow\{2\}}$. It means that from all 15 non-empty subsets of $\mathbf{X}_1 \times \mathbf{X}_2$ only 9 of them are potential focal elements (six subsets of $\mathbf{X}_1 \times \mathbf{X}_2$ that cannot be focal elements are listed in Example 1). Naturally, this condition on focal elements is only a necessary condition for the independence. This condition is not sufficient. For example, the reader can easily check that the two basic assignments $m_1 \triangleright m_2$ from Table 2 and m_3 from Table 3 (both defined on $\mathbf{X}_1 \times \mathbf{X}_2$) have the same marginal assignments: $((m_1 \triangleright m_2)^{\downarrow\{1\}} = m_3^{\downarrow\{1\}} = m_1$ and $(m_1 \triangleright m_2)^{\downarrow\{2\}} = m_3^{\downarrow\{2\}} = m_2)$. Moreover, for all of their focal elements the required property $C = C^{\downarrow\{1\}} \otimes C^{\downarrow\{2\}}$ holds true and simultaneously

$$1 \perp\!\!\!\perp_{m_1\triangleright m_2} 2 \quad \text{and} \quad 1 \not\perp\!\!\!\perp_{m_3} 2.$$

Analogously to what has just been said about (unconditional) independence, there is a necessary condition also on focal elements of basic assignments with conditional independence. Conditional independence

$$1 \perp\!\!\!\perp_m 3 \,|\, 2$$

means that all focal elements $C \subseteq \mathbf{X}_{\{1,2,3\}}$ of m must be of the form

$$C = C^{\downarrow\{1,2\}} \otimes C^{\downarrow\{2,3\}}.$$

It is not difficult to show that this property holds true only for 99 out of all possible 255 nonempty subsets of $\mathbf{X}_{\{1,2,3\}}$.

In the rest of this section we will show that the ternary relation $K \perp\!\!\!\perp_m L \,|\, M$ is a *semigraphoid*, i.e. it meets the four semigraphoid axioms listed below. For this, we will exclusively use the properties of the operator of composition presented in Lemma 2. In what follows, each axiom is reformulated into the language of composition and the corresponding theorem is proved.

Symmetry

$$I \perp\!\!\!\perp_m J \,|\, L \implies J \perp\!\!\!\perp_m I \,|\, L$$

Theorem 3. *If* $m^{\downarrow I\cup J\cup L} = m^{\downarrow I\cup L} \triangleright m^{\downarrow J\cup L}$ *then also* $m^{\downarrow I\cup J\cup L} = m^{\downarrow J\cup L} \triangleright m^{\downarrow I\cup L}$.

Proof. The assertion follows immediately from the fact that marginals $m^{\downarrow I\cup L}$ and $m^{\downarrow J\cup L}$ are consistent, and therefore property (iii) may be applied

$$m^{\downarrow I\cup L} \triangleright m^{\downarrow J\cup L} = m^{\downarrow J\cup L} \triangleright m^{\downarrow I\cup L}.$$

□

Decomposition

$$I \perp\!\!\!\perp_m J \cup K \,|\, L \implies I \perp\!\!\!\perp_m K \,|\, L$$

Theorem 4. *If* $m^{\downarrow I\cup J\cup K\cup L} = m^{\downarrow I\cup L} \triangleright m^{\downarrow J\cup K\cup L}$ *then also* $m^{\downarrow I\cup K\cup L} = m^{\downarrow I\cup L} \triangleright m^{\downarrow K\cup L}$.

Proof. The assertion will be obtained just by application of properties (iv) and (ii)

$$\begin{aligned} m^{\downarrow I\cup K\cup L} &= \left(m^{\downarrow I\cup J\cup K\cup L}\right)^{\downarrow I\cup K\cup L} \\ &= \left(m^{\downarrow I\cup L} \triangleright m^{\downarrow J\cup K\cup L}\right)^{\downarrow I\cup K\cup L} \\ &= \left((m^{\downarrow I\cup L} \triangleright m^{\downarrow K\cup L}) \triangleright m^{\downarrow J\cup K\cup L}\right)^{\downarrow I\cup K\cup L} \\ &= m^{\downarrow I\cup L} \triangleright m^{\downarrow K\cup L}. \end{aligned}$$

□

Weak Union

$I \perp\!\!\!\perp_m J \cup K \,|\, L \implies I \perp\!\!\!\perp_m J \,|\, K \cup L$

Theorem 5. *If* $m^{\downarrow I\cup J\cup K\cup L} = m^{\downarrow I\cup L} \triangleright m^{\downarrow J\cup K\cup L}$ *then also* $m^{\downarrow I\cup J\cup K\cup L} = m^{\downarrow I\cup K\cup L} \triangleright m^{\downarrow J\cup K\cup L}$.

Proof. To prove this assertion we have to realize that, due to property (iv),

$$m^{\downarrow I\cup L} \triangleright m^{\downarrow J\cup K\cup L} = \left(m^{\downarrow I\cup L} \triangleright m^{\downarrow K\cup L}\right) \triangleright m^{\downarrow J\cup K\cup L},$$

and that, because the assumptions of Theorem 4 are fulfilled, also

$$m^{\downarrow I\cup K\cup L} = m^{\downarrow I\cup L} \triangleright m^{\downarrow K\cup L}.$$

Using these two equalities we finish the proof in a simple way

$$\begin{aligned} m^{\downarrow I\cup J\cup K\cup L} &= m^{\downarrow I\cup L} \triangleright m^{\downarrow J\cup K\cup L} \\ &= \left(m^{\downarrow I\cup L} \triangleright m^{\downarrow K\cup L}\right) \triangleright m^{\downarrow J\cup K\cup L} \\ &= m^{\downarrow I\cup K\cup L} \triangleright m^{\downarrow J\cup K\cup L}. \end{aligned}$$

□

Contraction

$I \perp\!\!\!\perp_m K \,|\, L \;\&\; I \perp\!\!\!\perp_m J \,|\, K \cup L \implies I \perp\!\!\!\perp_m J \cup K \,|\, L$

Theorem 6. *If* $m^{\downarrow I\cup K\cup L} = m^{\downarrow I\cup L} \triangleright m^{\downarrow K\cup L}$, *and* $m^{\downarrow I\cup J\cup K\cup L} = m^{\downarrow I\cup K\cup L} \triangleright m^{\downarrow J\cup K\cup L}$, *then also* $m^{\downarrow I\cup J\cup K\cup L} = m^{\downarrow I\cup L} \triangleright m^{\downarrow J\cup K\cup L}$.

Proof. We will follow the same idea as in the preceding proof but in the reverse order. First, we will use property (iv) and then both assumptions of this assertion.

$$\begin{aligned} m^{\downarrow I\cup L} \triangleright m^{\downarrow J\cup K\cup L} &= \left(m^{\downarrow I\cup L} \triangleright m^{\downarrow K\cup L}\right) \triangleright m^{\downarrow J\cup K\cup L} \\ &= m^{\downarrow I\cup K\cup L} \triangleright m^{\downarrow J\cup K\cup L} \\ &= m^{\downarrow I\cup J\cup K\cup L}. \end{aligned}$$

□

6 Conclusions

In the paper we dealt with the two problems connected with computational complexity of Dempster-Shafer theory of evidence. Since full generality of the models leads to exponential grows of space and computational complexity we showed that focusing our attention only to models, which are constructed from Bayesian basic assignments by application of the operator of composition, one does not get beyond the boundaries of a rather limited class of models, which are called in the paper almost Bayesian. The most advantageous characteristics of these models is the fact that though they are able to describe a special type of an ignorance, they do not have a higher space requirements than classical probabilistic models.

The other goal of this paper was to show that when accepting the notion of conditional independence based on factorization corresponding to the operator of composition, one can easily prove validity of semigraphoid axioms just with the help of the four very elementary properties from Lemma 2. Since the same idea was employed by Prakash P. Shenoy in [7], a very natural question arises: what is the relation of composition introduced in this paper and the Shenoy's notion of *combination*?

Looking at Shenoy axioms[4] C1, C2 and C3 we see that Shenoy's axiom C1 *(Domain)* is equivalent to property (i) of Lemma 2 and therefore it holds also for our composition. However Shenoy's axioms C2 (Associative) and C3 *(Commutative)* hold for composition only under special conditions. The operator of composition is commutative only for consistent basic assignments; point (iii) of Lemma 2. In definition of conditional independence (Definition 5 of this paper) we consider only composition of consistent assignments (marginals of the considered basic assignment) and therefore we were able to prove axiom of Symmetry. Nevertheless, associativity holds for the operator of composition only under very specific conditions[5] and therefore the Shenoy's proofs cannot be used. Moreover, property (ii) of Lemma 2 does not hold for Shenoy's combination. So, one cannot be surprised that both of the definitions of conditional independence (i.e. the one proposed in this paper and Shenoy's conditional independence following from the definitions in Section 5 of [7]) are different from each other. They coincide only for unconditional independence and for conditional independence in case of Bayesian basic assignments. Moreover, as we showed in [4], our concept of conditional independence does not suffer from the drawback described in detail in [1], where the authors show that the notion of conditional independence used by Shenoy is not *consistent with marginalization*[6]. Therefore, we can conclude that our concept of conditional independence seems to meet better some of the intuitive requirements. Nevertheless, a question what is the relation of this notion and concepts of conditional basic assignments remains still open.

[4] We do not comment axiom C4 *(Zero)* because we consider only normalized basic assignments.

[5] For example, for basic assignments m_1, m_2, m_3 defined on $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}, \mathbf{X}_{K_3}$, respectively

$K_1 \supseteq (K_2 \cap K_3) \implies (m_1 \triangleright m_2) \triangleright m_3 = m_1 \triangleright (m_2 \triangleright m_3)$.

[6] Roughly speaking: one can find two consistent basic assignments m_1, m_2, on $\mathbf{X}_1 \times \mathbf{X}_2$ and $\mathbf{X}_2 \times \mathbf{X}_3$, respectively, for which there does not exist a 3-dimensional basic assignment m on $\mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3$ having m_1 and m_2 as its marginals, and for which $1 \perp\!\!\!\perp_m 3 \,|\, 2$.

Acknowledgements

The research was financially supported by GAČR under the grant no. ICC/08/E010, and 201/09/1891, and by Ministry of Education of the Czech Republic by grants no. 1M0572 and 2C06019.

References

[1] B. Ben Yaghlane, Ph. Smets, and K. Mellouli, "Belief Function Independence: II. The Conditional Case," *Int. J. of Approximate Reasoning*, vol. 31, no. (1-2), pp. 31–75, 2002.

[2] R. Jiroušek, "Composition of probability measures on finite spaces," *Proc. of the 13th Conf. Uncertainty in Artificial Intelligence UAI'97*, (D. Geiger and P. P. Shenoy, eds.). Morgan Kaufmann Publ., San Francisco, California, pp. 274–281, 1997.

[3] R. Jiroušek, "On a Conditional Irrelevance Relation for Belief Functions based on the Operator of Composition," *Dynamics of Knowledge and Belief* (Ch. Beierle, G. Kern-Isberner, eds.) Proceedings of the Workshop at the 30th Annual German Conference on Artificial Intelligence, Fern Universität in Hagen, Osnabrück, 2007, pp.28-41.

[4] R. Jiroušek, J. Vejnarová, "Compositional Models and Conditional Independence in Evidence Theory," submitted to *Int. J. of Approximate Reasoning.*

[5] R. Jiroušek, J. Vejnarová and M. Daniel, "Compositional models of belief functions," *Proc. of the 5th Symposium on Imprecise Probabilities and Their Applications* (G. de Cooman, J. Vejnarová and M. Zaffalon, eds.), Charles University Press, Praha, pp. 243–252, 2007.

[6] R. Jiroušek, J. Vejnarová, "There are Combinations and Compositions in Dempster-Shafer Theory of Evidence," submitted to *WUPES'09.*

[7] P. P. Shenoy, "Conditional independence in valuation-based systems," *Int. J. of Approximate Reasoning*, vol. 10, no. 3, pp. 203–234, 1994.

[8] M. Studený, "Formal properties of conditional independence in different calculi of AI," *Proceedings of European Conference on Symbolic and quantitative Approaches to Reasoning and Uncertainty ECSQARU'93*, (K. Clarke, R. Kruse and S. Moral, eds.). Springer-Verlag, 1993, pp. 341–351.

[9] M. Studený, "On stochastic conditional independence: the problems of characterization and description," *Annals of Mathematics and Artificial Intelligence,* vol. 35, p. 323-341, 2002.

[10] J. Vejnarová, "Composition of possibility measures on finite spaces: preliminary results," *Proceedings of 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems IPMU'98,* (B. Bouchon-Meunier, R.R. Yager, eds.). Editions E.D.K. Paris, 1998, pp. 25–30.

[11] J. Vejnarová, "Possibilistic independence and operators of composition of possibility measures," *Prague Stochastics'98,* (M. Hušková , J. Á. Víšek, P. Lachout, eds.) JČMF, 1998, pp. 575–580.

On the Behavior of the Robust Bayesian Combination Operator and the Significance of Discounting

Alexander Karlsson
Informatics Research Centre
University of Skövde, Sweden
alexander.karlsson@his.se

Ronnie Johansson
Informatics Research Centre
University of Skövde, Sweden
ronnie.johansson@his.se

Sten F. Andler
Informatics Research Centre
University of Skövde, Sweden
sten.f.andler@his.se

Abstract

We study the *combination problem* for credal sets via the *robust Bayesian combination operator*. We extend Walley's notion of *degree of imprecision* and introduce a measure for *degree of conflict* between two credal sets. Several examples are presented in order to explore the behavior of the robust Bayesian combination operator in terms of imprecision and conflict. We further propose a *discounting operator* that suppresses a source given an interval of *reliability weights*, and highlight the importance of using such weights whenever additional information about the reliability of a source is available.

Keywords. Imprecise probabilities, robust Bayesian combination, credal set, discounting, information fusion

1 Introduction

We define the *combination problem* as the problem of combining *evidences* regarding some reality of interest (cf., [9]). The problem has gained much attention in several different research fields, in particular *information fusion* (see, e.g., [2]) and *artificial intelligence* (see, e.g., [18]). We have here taken a "set-point-wise Bayesian", or *credal* [11, 5], approach to the combination problem via *the robust Bayesian combination operator*. One important advantage with such an approach is that it is easily adoptable for practitioners and researchers that already are familiar with (standard) Bayesian theory. It should be emphasized that the combination problem is different from the *aggregation problem* where the main goal is to find a common agreement among sources. If an *aggregation operator* [19, Section 1.1] is applied to identical operands, typically the result will also be the same, since it represents a "perfect agreement". If we consider the same scenario, using a *combination operator* instead, the result usually represents stronger evidence in comparison to any of the operands, since both sources agree on some hypotheses, i.e., the result is different from the operands. Several researchers have addressed the aggregation problem (see, e.g., [12, 13, 20]), however, the combination problem is an overlooked area in the case of general credal sets. Combination of evidences in the form of so-called *mass functions* (which can be transformed into a particular type of credal set [2]), have been thoroughly studied within *evidence theory* [16], mainly via some variant of *Dempster's rule*. However, it has been shown that Dempster's rule can yield disparate results in comparison to the robust Bayesian combination operator, in fact, the results can even be disjoint [2].

Our main concern in this paper is to characterize the behavior, interpretation, and implications of utilizing the robust Bayesian combination operator for the combination problem. Furthermore, we introduce a discounting operator which can be used whenever an interval of reliability weights are known for the sources involved in the combination.

The paper is organized as follows: in Section 2, we elaborate on *credal set theory*[1] and derive the robust Bayesian combination operator. In Section 3, we elaborate on *imprecision* and *conflict* with respect to credal sets. In Section 4, we present three examples and utilize imprecision and conflict in order to investigate the results. In Section 5, we introduce the discounting operator and revisit two of the mentioned examples. Lastly, in Section 6, we present a summary, our conclusions, and ideas for future work.

2 Preliminaries

We here present some background on credal set theory and derive the robust Bayesian combination operator via its precise counterpart, the Bayesian combination operator.

[1] Also known as *theory of credal sets*. We choose the term "credal set theory" since it is coherent with "Bayesian theory".

2.1 Credal Set Theory

Credal set theory [4, 5, 6, 11] is a generalization of Bayesian theory where one acknowledges that there might be more than one reasonable probability distribution for representing belief. As a consequence one is allowed to adopt a *closed convex set* of such distributions, commonly referred to as a *credal set*, as the fundamental representation of belief. In order to update such belief, one applies Bayes' theorem *pointwise* to a credal set of *priors* and a *convex set of likelihood functions*. As a last step one utilizes a convex hull operation. Note that in the special case of singleton sets, the theory reduces to standard Bayesian theory.

Let us denote a credal set by $\mathcal{P}_X$, containing probability distributions of the form $p(X)$, $\mathcal{P}_{X|y}$ for distributions in the conditional form $p(X|y)$, and $\mathcal{P}_{X,Y}$ for joint probability distributions $p(X,Y)$. Let $ext\,(\mathcal{P}_X)$ denote the set of *extreme points* (also known as *vertices*) of $\mathcal{P}_X$, i.e., distributions that cannot be expressed as a *convex combination*[2] of any other distributions in the set. We only consider credal sets that have a finite set of extreme points (also known as *polytopes*). Each credal set $\mathcal{P}_X$ can be described as the set of convex combinations of points in $ext\,(\mathcal{P}_X)$, in other words, it suffices to maintain a credal sets' extreme points in order to represent it. In a number of places throughout this paper we will use the credal set that contains all probability distribution for some random variable. Let us therefore formally define this credal set:

Definition 1. *Let $\mathcal{P}_X^*$ denote the set of all probability distributions for a random variable X with state space Ω_X, i.e., $\mathcal{P}_X^* \triangleq \{p : 0 \leq p(x_i) \leq 1,\ 1 \leq i \leq |\Omega_X|,\ \sum_{i=1}^{|\Omega_X|} p(x_i) = 1\}$*

One controversy in credal set theory is how one should define *independence* between variables (for an overview see [3]). We here adopt the most commonly used such definition, referred to as *strong independence* [6]:

Definition 2. *X and Y are strongly independent iff each $p_i \in ext\,(\mathcal{P}_{X,Y})$ can be expressed as $p_i = p_j p_k$, where $p_j \in \mathcal{P}_X$ and $p_k \in \mathcal{P}_Y$. X and Y are strongly conditionally independent given Z iff $p_i \in ext\left(\mathcal{P}_{X,Y|z}\right)$ can be expressed as $p_i = p_j p_k$, $\forall z \in \Omega_Z$, where $p_j \in \mathcal{P}_{X|z}$ and $p_k \in \mathcal{P}_{Y|z}$.*

[2] A convex combination of points $\{p_i : 1 \leq i \leq n\}$ is defined as $\sum_{i=1}^{n} \lambda_i p_i$, where $\sum_{i=1}^{n} \lambda_i = 1$, $\lambda_i \geq 0$

2.2 The Robust Bayesian Combination Operator

Let us first derive, via Bayes' theorem, the Bayesian combination operator, which we then generalize to operate on credal sets. The derivation is inspired by Arnborg [1, 2]. The derivation has previously been utilized in order to define *distinctness of evidences* in variants of evidence theory [17, Sect. 3.1]. Assume that two *sources* have made observations y_1 and y_2, respectively, related to a random variable X. If one wants to formulate one's belief regarding X, based on the observations made by the sources, one utilizes Bayes' theorem:

$$p(X|y_1,y_2) = \frac{p(y_1,y_2|X)p(X)}{\sum_{x\in\Omega_X} p(y_1,y_2|x)p(x)} \tag{1}$$

We see that the *posterior belief* $p(X|y_1,y_2)$ is affected by the observations through the *joint likelihood* $p(y_1,y_2|X)$. Hence, it is reasonable to interpret such likelihood as being *evidence* regarding X [9]. Now, if one's posterior belief $p(X|y_1,y_2)$ should be a representation of the available evidence solely, i.e., the posterior belief should be equal to the normalized joint likelihood function, then we need to set our prior belief $p(X)$ to the *uniform distribution* over Ω_X. If we also can assume that the sources have made conditionally independent observations given X, i.e.,:

$$p(y_1,y_2|X) = p(y_1|X)p(y_2|X) \tag{2}$$

and that both sources have adopted the uniform distribution as their prior belief $p(X)$, i.e., their belief is completely determined by likelihoods, then we get:

$$p(X|y_1,y_2) = \frac{p(y_1|X)p(y_2|X)p(X)}{\sum_{x\in\Omega_X} p(y_1|x)p(y_2|x)p(x)} \tag{3}$$

$$= \frac{\frac{p(X|y_1)p(y_1)}{p(X)}\frac{p(X|y_2)p(y_2)}{p(X)}}{\sum_{x\in\Omega_X}\frac{p(x|y_1)p(y_1)}{p(x)}\frac{p(x|y_2)p(y_2)}{p(x)}} \tag{4}$$

$$= \frac{p(X|y_1)p(X|y_2)}{\sum_{x\in\Omega_X} p(x|y_1)p(x|y_2)} \tag{5}$$

We know that:

$$\begin{aligned} p(X|y_i) &= \frac{p(y_i|X)p(X)}{\sum_{x\in\Omega_X} p(y_i|x)p(x)} \\ &= \frac{p(y_i|X)}{\sum_{x\in\Omega_X} p(y_i|x)}, \end{aligned} \tag{6}$$

$i \in \{1, 2\}$, since the sources have adopted the uniform distribution as prior belief. Hence, Eq. 5 constitutes an operator that takes two probability functions, interpreted as evidences, i.e., normalized likelihoods, as operands, and returns a new such function, representing the combined evidence, i.e., normalized joint likelihood. We are now ready to define the *Bayesian combination operator* [1, 2]:

Definition 3. *The Bayesian Combination (BC) Operator*[3] *is defined as:*

$$p_1(X) \otimes_{\mathcal{B}} p_2(X) \triangleq \frac{p_1(X)p_2(X)}{\sum_{x \in \Omega_X} p_1(x)p_2(x)},$$

where $p_1(X)$ and $p_2(X)$ are interpreted as conditionally independent evidences, i.e., normalized likelihoods that are conditionally independent given X (see Eq. 2). The operator is undefined when $\sum_{x \in \Omega_X} p_1(x)p_2(x) = 0$.

Let us first comment on the case when $\sum_{x \in \Omega_X} p_1(x)p_2(x) = 0$. The case implies that likelihoods are such that at least one of them is zero for every $x \in \Omega_X$, which is exceptional in any properly modeled system. The exact way of dealing with such an exceptional case is application dependent. One technique for resolving the case is to utilize discounting with reliability weights strictly smaller than one (see further Sect. 5).

Note that if the operands strongly agree on some $x \in \Omega_X$ as being the most probable, then the operator will reinforce such probability in the resulting posterior function. As mentioned in the introduction, such behavior is clearly different from what one would expect from an aggregation operator. The reason for why such behavior is reasonable is due to the assumption of conditionally independence between evidences given X, as described by Eq. 2. Let us demonstrate this behavior of the BC operator with a simple example:

Example 1. *Assume that two sources reports the following probability distributions as a representation of conditionally independent evidences regarding the random variable X with state space Ω_X:*

$$p_1(x_1) = 0.7,\ p_1(x_2) = 0.2,\ p_1(x_3) = 0.1$$
$$p_2(x_1) = 0.8,\ p_2(x_2) = 0.1,\ p_2(x_3) = 0.1,$$

Applying the BC operator to p_1 and p_2, i.e., $p_1 \otimes_{\mathcal{B}} p_2$, yields the following distribution:

$$p_{1,2}(x_1) \approx 0.95,\ p_{1,2}(x_2) \approx 0.03,\ p_{1,2}(x_3) \approx 0.02,$$

Hence, the result constitutes stronger evidence for x_1 than any of the operands.

Now if we want to define an operator that generalizes the BC operator, in the sense of "point-wise Bayesianism", then one can substitute the operand single distributions to credal sets and apply the BC operator point-wise on every pair of distributions within the sets. Indeed, such an operator exists under the name *robust Bayesian combination operator* [1, 2]:

Definition 4. *The Robust Bayesian Combination (RBC) Operator* [4]*:*

$$\mathcal{P}^1_X \otimes_{\mathcal{R}} \mathcal{P}^2_X \triangleq CH\Big\{ p_i(X) \otimes_{\mathcal{B}} p_j(X) : p_i \in \mathcal{P}^1_X,\ p_j \in \mathcal{P}^2_X \Big\},$$

where CH denotes the convex hull, $\mathcal{P}^1_X$ and $\mathcal{P}^2_X$ are interpreted as strongly conditionally independent evidences, i.e., convex sets of normalized likelihoods that are strongly conditionally independent given X (see Def. 2). The operator is undefined if there exists $p_i \in \mathcal{P}^1_X$ and $p_j \in \mathcal{P}^2_X$ such that $\sum_{x \in \Omega_X} p_i(x)p_j(x) = 0$.

The operator is both associative and commutative. Note that the case regarding division by zero is inherited from the BC operator (Def. 3). Discounting the operands (see further Sect. 5) using reliability weights strictly smaller than one, resolves such case (see further the discussion following Def. 3). Throughout the remainder of the paper we will assume that some technique, guaranteeing $\sum_{x \in \Omega_X} p_i(x)p_j(x) > 0$, for all $p_i \in \mathcal{P}^1_X$ and $p_j \in \mathcal{P}^2_X$, has been utilized (e.g., discounting).

The following theorem facilitates computation with the RBC operator (the theorem was implicitly mentioned in [2], with no proof, and explicitly stated in [1, Theorem 1], where only a "proof hint" was provided):

Theorem 1.

$$\mathcal{P}^1_X \otimes_{\mathcal{R}} \mathcal{P}^2_X = ext(\mathcal{P}^1_X) \otimes_{\mathcal{R}} ext(\mathcal{P}^2_X)$$

Proof. The proof is partly inspired by Noack et al. [14, Theorem 2]. First note that $ext(\mathcal{P}^1_X) \otimes_{\mathcal{R}} ext(\mathcal{P}^2_X) \subseteq \mathcal{P}^1_X \otimes_{\mathcal{R}} \mathcal{P}^2_X$ is trivial. Assume that $ext(\mathcal{P}^1_X) \otimes_{\mathcal{R}} ext(\mathcal{P}^2_X)$ is strictly smaller than $\mathcal{P}^1_X \otimes_{\mathcal{R}} \mathcal{P}^2_X$, i.e., $ext(\mathcal{P}^1_X) \otimes_{\mathcal{R}} ext(\mathcal{P}^2_X) \subset \mathcal{P}^1_X \otimes_{\mathcal{R}} \mathcal{P}^2_X$. Then there must exists at least one $u \in ext(\mathcal{P}^1_X \otimes_{\mathcal{R}} \mathcal{P}^2_X)$ such that $u \notin ext(\mathcal{P}^1_X) \otimes_{\mathcal{R}} ext(\mathcal{P}^2_X)$, where u has the following form: $u = p_1 p_2 / \sum_{x \in \Omega_X} p_1(x)p_2(x)$, $p_1 \in \mathcal{P}^1_X$ and

[3] Arnborg [2] referred to this operator as *Laplace's parallel composition*

[4] Arnborg [2] defined the operator without the inclusion of a convex-hull operator (however he mentioned in the discussion following his definition that such an operator should be utilized)

$p_2 \in \mathcal{P}_X^2$, where at least one of p_1 and p_2 is not an extreme point. We can express p_1 and p_2 as:

$$\begin{aligned} p_1 &= \sum_{i=1}^{m} \lambda_i v_i \\ p_2 &= \sum_{j=1}^{n} \alpha_j w_j, \end{aligned} \tag{7}$$

where $v_i \in ext(\mathcal{P}_X^1)$, $w_j \in ext(\mathcal{P}_X^2)$, $\lambda_i \geq 0$, $\alpha_j \geq 0$, $1 \leq i \leq m$, $1 \leq j \leq n$, $\sum_{i=1}^{m} \lambda_i = \sum_{j=1}^{n} \alpha_i = 1$. Therefore (remember that the denominator is assumed not to be equal to zero, see the discussion following Def. 3 and Def. 4):

$$u = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} \lambda_i \alpha_j v_i w_j}{\sum_{x \in \Omega_X} \left(\sum_{i=1}^{m}\sum_{j=1}^{n} \lambda_i \alpha_j v_i(x) w_j(x) \right)} \tag{8}$$

Let us introduce the following notation:

$$\gamma_{i,j} \triangleq \frac{\lambda_i \alpha_j \sum_{x \in \Omega_X} v_i(x) w_j(x)}{\sum_{x \in \Omega_X} \left(\sum_{i=1}^{m}\sum_{j=1}^{n} \lambda_i \alpha_j v_i(x) w_j(x) \right)} \tag{9}$$

We can now rephrase u as:

$$u = \sum_{i=1}^{m}\sum_{j=1}^{n} \gamma_{i,j} \frac{v_i w_j}{\sum_{x \in \Omega_X} v_i(x) w_j(x)} \tag{10}$$

Since:

$$\frac{v_i w_j}{\sum_{x \in \Omega_X} v_i(x) w_j(x)} \in ext(\mathcal{P}_X^1) \otimes_{\mathcal{R}} ext(\mathcal{P}_X^2), \tag{11}$$

and $\gamma_{i,j} \geq 0$, $\sum_{i=1}^{m}\sum_{j=1}^{n} \gamma_{i,j} = 1$, we get $u \in ext(\mathcal{P}_X^1) \otimes_{\mathcal{R}} ext(\mathcal{P}_X^2)$, which is a contradiction. Hence we must conclude that $\mathcal{P}_X^1 \otimes_{\mathcal{R}} \mathcal{P}_X^2 = ext(\mathcal{P}_X^1) \otimes_{\mathcal{R}} ext(\mathcal{P}_X^2)$. □

3 Imprecision and Conflict

We here define measures for degree of *imprecision* and *conflict*.

3.1 Degree of Imprecision

Obviously, since credal set theory belongs to the family of theories referred to as *imprecise probabilities* [23], *imprecision* is an important concept to define. Walley [21, Section 5.1.4] has introduced a measure which he refers to as *the degree of imprecision* for an event $x_i \in \Omega_X$:

$$\Delta(x_i) \triangleq \max_{p \in \mathcal{P}_X} p(x_i) - \min_{p \in \mathcal{P}_X} p(x_i) \tag{12}$$

However, the measure does not capture the imprecision of a credal set, since it only operates on single events. At first, one might be tempted to think of the imprecision of a credal set as its *volume*. However, the volume can be made arbitrarily small while a high degree of imprecision for some event is preserved, something that is counterintuitive. Let us therefore base our measure of degree of imprecision for a credal set on Walley's measure in the following way:

Definition 5. *Degree of Imprecision:*

$$\mathcal{I}(\mathcal{P}_X) \triangleq \frac{1}{n} \sum_{x \in \Omega_X} \Delta(x)$$

where $\mathcal{P}_X \subseteq \mathbb{R}^n$ *and* $n = |\Omega_X|$

The optimization problems involved in the definition of $\mathcal{I}$ are linear, hence, the solutions can be found by iterating through the extreme points.

3.2 Degree of Conflict

Assume that two sources report (strongly conditionally independent) evidence in the form of credal sets $\mathcal{P}_X^1$ and $\mathcal{P}_X^2$ and that one wants to formulate the combined evidence concerning X based on these sets. If both sources report exactly the same credal set, then they are willing to act according to any distribution within any of their sets. In other cases, i.e., when the credal sets are not equal, then there exists a distribution which not both sources are willing to act upon, i.e., a certain *degree of conflict* is present. Intuitively, the degree of conflict between $\mathcal{P}_X^1$ and $\mathcal{P}_X^2$ should be related to some distance between the sets. Indeed, there exists such distance measure, which goes under the name of *Hausdorff distance* [10]. Let us therefore define a *degree of conflict* between two credal sets in the following way:

Definition 6. *Degree of Conflict:*

$$\mathcal{K}(\mathcal{P}_X^1, \mathcal{P}_X^2) \triangleq \frac{\mathcal{H}(\mathcal{P}_X^1, \mathcal{P}_X^2)}{\sqrt{2}},$$

where the denominator is a constant constituting the diameter *of the set* $\mathcal{P}_X^*$ *(see Def. 1), i.e.,* $\max_{p_i \in \mathcal{P}_X^*} \left\{ \max_{p_j \in \mathcal{P}_X^*} d(p_i, p_j) \right\} = \sqrt{2}$ *(the diameter of a credal set is found in the set of distances between extreme points [7, Theorem 12]) where* d *denotes the Euclidean distance, and* $\mathcal{H}$ *is the Hausdorff distance defined by:*

$$\mathcal{H}(\mathcal{P}_X^1, \mathcal{P}_X^2) \triangleq \max \left\{ \overrightarrow{\mathcal{H}}(\mathcal{P}_X^1, \mathcal{P}_X^2), \overrightarrow{\mathcal{H}}(\mathcal{P}_X^2, \mathcal{P}_X^1) \right\},$$

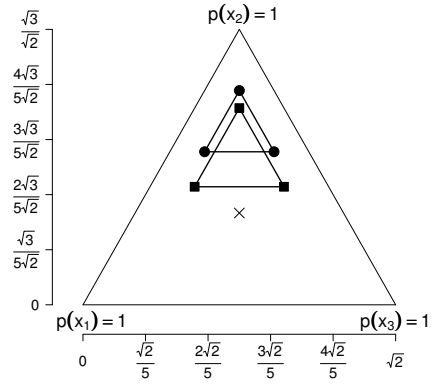


Figure 1: $\mathcal{P}_X^1$ (circles) and $\mathcal{P}_X^2$ (squares) projected on two-dimensional space. The triangle where extreme points $p(x_1) = p(x_2) = p(x_3) = 1$ have been, marked constitutes $\mathcal{P}_X^*$ (see Def. 1).

where $\overrightarrow{\mathcal{H}}$ *is the forward Hausdorff distance defined by:*

$$\overrightarrow{\mathcal{H}}(\mathcal{F}_1, \mathcal{F}_2) \triangleq \max_{f_i \in \mathcal{F}_1} \left\{ \min_{f_j \in \mathcal{F}_2} d(f_i, f_j) \right\},$$

where $\mathcal{F}_1$ *and* $\mathcal{F}_2$ *are general closed convex sets in* $\mathbb{R}^n$.

The forward Hausdorff-distance can be calculated in $O(|ext(\mathcal{F}_1)||fac(\mathcal{F}_2)|)$ [10], where fac denotes the set of faces. Let us demonstrate the conflict measure by a simple example:

Example 2. *Consider Fig. 1 where two credal sets,* $\mathcal{P}_X^1$ *and* $\mathcal{P}_X^2$*, for a random variable* X *with* $\Omega_X = \{x_1, x_2, x_3\}$*, has been plotted. From the figure, it is seen that* $\overrightarrow{\mathcal{H}}(\mathcal{P}_X^2, \mathcal{P}_X^1) > \overrightarrow{\mathcal{H}}(\mathcal{P}_X^1, \mathcal{P}_X^2)$*, since there exists at least one point in* $\mathcal{P}_X^2$ *(e.g., the lower right extreme point) from where the minimum distance to* $\mathcal{P}_X^1$ *is larger than the distance from any point in* $\mathcal{P}_X^1$ *to a point in* $\mathcal{P}_X^2$*. Hence, the Hausdorff distance* $\mathcal{H}(\mathcal{P}_X^1, \mathcal{P}_X^2)$ *must be equal to the forward Hausdorff distance* $\overrightarrow{\mathcal{H}}(\mathcal{P}_X^2, \mathcal{P}_X^1)$*, which is the maximum of the set of distances from the set of extreme points of* $\mathcal{P}_X^2$ *to* $\mathcal{P}_X^1$*'s faces [10]. In this example, the maximum such distance, approximately equal to* 0.16*, is found among the distances between the lower extreme points of* $\mathcal{P}_X^2$ *to the lower extreme points of* $\mathcal{P}_X^1$*, i.e.,* $\mathcal{H}(\mathcal{P}_X^1, \mathcal{P}_X^2) \approx \overrightarrow{\mathcal{H}}(\mathcal{P}_X^2, \mathcal{P}_X^1) \approx 0.16$*, yielding a degree of conflict* $\mathcal{K}(\mathcal{P}_X^1, \mathcal{P}_X^2) \approx 0.11$.

Notice that if $\mathcal{P}_X^1 = \mathcal{P}_X^2$ then $\mathcal{K}(\mathcal{P}_X^1, \mathcal{P}_X^1) = 0$. Also, if $ext(\mathcal{P}_X^1) \subseteq ext(\mathcal{P}_X^*)$ and $ext(\mathcal{P}_X^2) \subseteq ext(\mathcal{P}_X^*)$, and $ext(\mathcal{P}_X^1) \cap ext(\mathcal{P}_X^2) = \emptyset$ then $\mathcal{K}(\mathcal{P}_X^1, \mathcal{P}_X^1) = 1$ (since the distance between two different extreme points of the set $\mathcal{P}_X^*$ is $\sqrt{2}$).

4 Examples

We will here give some examples of utilizing the robust Bayesian combination (RBC) operator in scenarios where there are different degrees of conflict present. For simplicity, let us utilize the family of credal sets that can be obtained by the *imprecise Dirichlet model* (IDM) [22] for constructing the operand credal sets. Note that these sets stem from a credal set of priors (hence not from a set of likelihoods) and that we are only utilizing the IDM as a convenient way of constructing different geometrical shapes of credal sets for the examples. Consider a random variable X with state space $\Omega_X = \{x_1, x_2, x_3\}$. A credal set obtained from the IDM for this state space can be parameterized according to:

$$\begin{aligned} &\mathrm{IDM}(\boldsymbol{\alpha}, s) \triangleq \\ &\left\{ p : \frac{\alpha_i}{\sum_{i=1}^{3} \alpha_i + s} \leq p(x_i) \leq \frac{\alpha_i + s}{\sum_{i=1}^{3} \alpha_i + s}, \right. \\ &\qquad \left. 1 \leq i \leq 3, \ \sum_{i=1}^{3} p(x_i) = 1 \right\}, \end{aligned} \tag{13}$$

where α_i denotes the i^{th} component of $\boldsymbol{\alpha}$.

4.1 Low Conflict

Let us start with an example where there exists a low degree of conflict between the sources. We define the example by utilizing Eq. (13) on the following parameters:

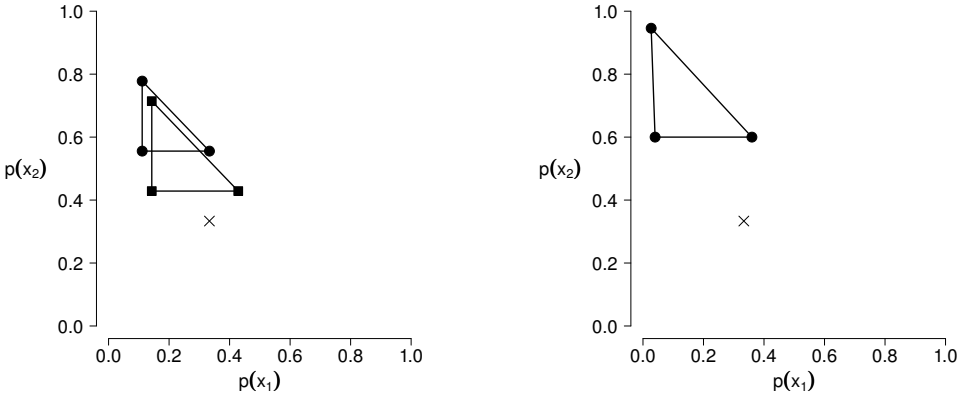
$$\begin{aligned} \mathcal{P}_X^1 &= \mathrm{IDM}((1, 5, 1), 2) \\ \mathcal{P}_X^2 &= \mathrm{IDM}((1, 3, 1), 2) \end{aligned} \tag{14}$$

The corresponding credal sets are shown in Fig. 2(a), where the sets have been projected on the components $p(x_1)$ and $p(x_2)$ (this enables one to see the probabilities directly from the plot). The line segment defined by coordinates $(0, 1)$ and $(1, 0)$ corresponds to the set of distributions where $p(x_3) = 0$. From the figure we see that there is only a slight conflict, $\mathcal{K}(\mathcal{P}_X^1, \mathcal{P}_X^2) \approx 0.11$, and that both sources essentially agree on "x_2" as being most probable. Therefore the result, denoted by $\mathcal{P}_X^{1,2}$ ($\mathcal{I}(\mathcal{P}_X^{1,2}) \approx 0.34$), is reinforced towards a high probability for "x_2", as is seen in Fig. 2(b).

4.2 Balanced Conflict

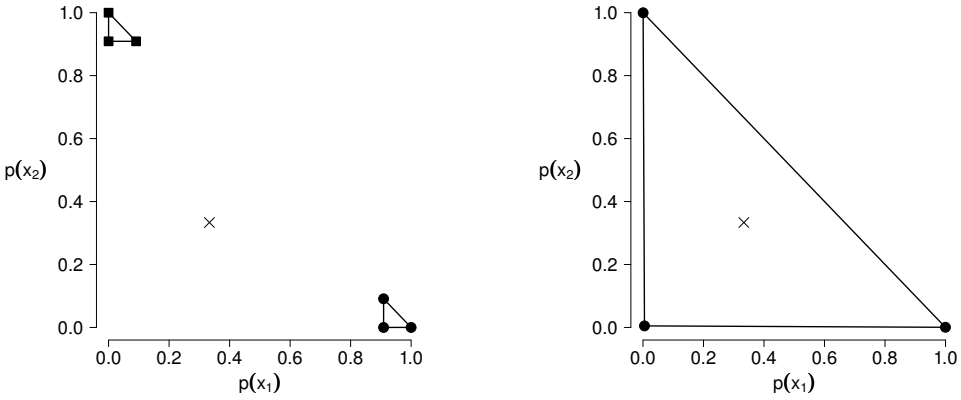
Consider an example where the evidences from the sources are strongly conflicting:

$$\begin{aligned} \mathcal{P}_X^1 &= \mathrm{IDM}((20, 10^{-3}, 10^{-3}), 2) \\ \mathcal{P}_X^2 &= \mathrm{IDM}((10^{-3}, 20, 10^{-3}), 2) \end{aligned} \tag{15}$$



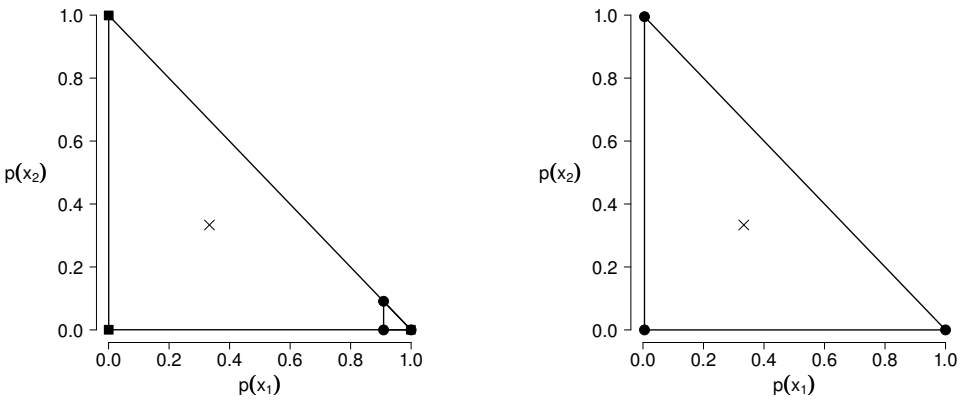
(a) $\mathcal{P}_X^1$ (circles) and $\mathcal{P}_X^2$ (squares) (b) $\mathcal{P}_X^{1,2}$

Figure 2: $\mathcal{P}_X^i$, $i \in \{1, 2\}$, and $\mathcal{P}_X^{1,2}$ for Example 1 – Low Conflict.



(a) $\mathcal{P}_X^1$ (circles) and $\mathcal{P}_X^2$ (squares) (b) $\mathcal{P}_X^{1,2}$

Figure 3: $\mathcal{P}_X^i$, $i \in \{1, 2\}$, and $\mathcal{P}_X^{1,2}$ for Example 2 – Balanced Conflict.



(a) $\mathcal{P}_X^1$ (circles) and $\mathcal{P}_X^2$ (squares) (b) $\mathcal{P}_X^{1,2}$

Figure 4: $\mathcal{P}_X^i$, $i \in \{1, 2\}$, and $\mathcal{P}_X^{1,2}$ for Example 3 – Unbalanced Conflict.

Since the sources expresses the same degree of imprecision, we refer to the conflict as *balanced*. The operand credal sets and result can be seen in Fig. 3. We see that there is a high degree of conflict, $\mathcal{K}(\mathcal{P}_X^1, \mathcal{P}_X^2) \approx 0.91$ and that the resulting credal set $\mathcal{P}_X^{1,2}$ has a high degree of imprecision, $\mathcal{I}(\mathcal{P}_X^{1,2}) \approx 1$. The main reason for this is due to that the "point-wise" combination of the lower left extreme points of $\mathcal{P}_X^1$ and $\mathcal{P}_X^2$ results in the lower left extreme point of $\mathcal{P}_X^{1,2}$; a case that is similar to the well-known *Zadeh's (counter) example* for Dempster's rule [24]. The reason for such behavior is due to that the extreme points component-wise suppress each other for events x_1 and x_2.

4.3 Unbalanced Conflict

Now consider an example where one of the operand credal set is highly imprecise while the other is not:

$$\begin{aligned} \mathcal{P}_X^1 &= \text{IDM}((20, 10^{-3}, 10^{-3}), 2) \\ \mathcal{P}_X^2 &= \text{IDM}((10^{-3}, 10^{-3}, 10^{-3}), 2) \end{aligned} \tag{16}$$

The corresponding credal sets can be seen in Fig. 4. We see that the resulting credal set $\mathcal{P}_X^{1,2}$ has been strongly affected by the second source since $\mathcal{I}(\mathcal{P}_X^{1,2}) \approx 1$. However, since there exist distributions in $\mathcal{P}_X^2$ that are positioned at a large distance from any distribution in $\mathcal{P}_X^1$, there is a strong conflict present: $\mathcal{K}(\mathcal{P}_X^1, \mathcal{P}_X^2) \approx 0.91$. Since the conflict in this case is due to differences in imprecision, we will refer to the conflict as *unbalanced*.

5 Discounting

Assume that one possesses information concerning the *reliability* of the sources and that one encodes this information via a convex set of *reliability weights* $\mathcal{W}$ [5], i.e., an interval. If one knows that some source is not fully reliable, e.g., a sensor of low quality, then one should suppress the statement from that source accordingly, i.e., the source should have less influence on the end result. Such procedure is commonly referred to as *discounting* in the literature [16]. If both the credal set and set of reliability weights are singleton, then discounting is achieved by transforming the single distribution, with respect to the weight, to a new distribution that is more similar to the uniform distribution. The reason for this is that the uniform distribution represents evidence that has no influence on the end result when combined with another distribution, i.e., the latter is always returned as result in such case.

[5] Imprecision in reliability weights was inspired by Troffaes [20]

Now, if we generalize the above approach to credal sets and set of reliability weights, preserving the idea of "point-wise Bayesianism", we obtain the following discounting operator:

Definition 7. *The RBC Discounting Operator:*

$$\mathcal{D}(\mathcal{P}_X, \mathcal{W}) \triangleq CH\left\{wp + (1-w)p_u : w \in \mathcal{W}, p \in \mathcal{P}_X\right\},$$

where $\mathcal{P}_X \subseteq \mathbb{R}^n$, $\mathcal{W} \subseteq [0,1]^2$ is an interval of reliability weights, and $p_u \in \mathbb{R}^n$, $n = |\Omega_X|$, is the uniform distribution over Ω_X.

The RBC discounting operator collapses a credal set "towards" the uniform distribution. Note that when the uniform distribution is combined, using the RBC operator, with any other credal set, the latter is obtained as result. Hence, by applying the RBC discounting operator on an operand, the end result will be less influenced by that operand, depending on $\mathcal{W}$ (the collapse towards the uniform distribution should therefore not be interpreted as a "bias" towards the uniform distribution as it would have been for an aggregation operator). The following theorem allows one to perform computation with the RBC discounting operator:

Theorem 2.

$$\mathcal{D}(\mathcal{P}_X, \mathcal{W}) = \mathcal{D}(ext\left(\mathcal{P}_X\right), ext\left(\mathcal{W}\right))$$

Proof. First note that $\mathcal{D}\left(ext\left(\mathcal{P}_X\right), ext\left(\mathcal{W}\right)\right) \subseteq \mathcal{D}\left(\mathcal{P}_X, \mathcal{W}\right)$ is trivial. Assume that $\mathcal{D}\left(ext\left(\mathcal{P}_X\right), ext\left(\mathcal{W}\right)\right)$ is strictly smaller than $\mathcal{D}\left(\mathcal{P}_X, \mathcal{W}\right)$, i.e., $\mathcal{D}\left(ext\left(\mathcal{P}_X\right), ext\left(\mathcal{W}\right)\right) \subset \mathcal{D}\left(\mathcal{P}_X, \mathcal{W}\right)$. Then there must exists at least one $u \in ext(\mathcal{D}\left(\mathcal{P}_X, \mathcal{W}\right))$ such that $u \notin \mathcal{D}\left(ext\left(\mathcal{P}_X\right), ext\left(\mathcal{W}\right)\right)$ where u has the following form: $u = wp + (1-w)p_u$, $w \in \mathcal{W}$, and $p \in \mathcal{P}_X$, where at least one of w and p is not an extreme point. There are three cases:

Case 1 – $p \in ext\left(\mathcal{P}_X\right)$, $w \notin ext\left(\mathcal{W}\right)$:
We know that w can be expressed as:

$$w = \lambda w_1 + (1-\lambda)w_2, \tag{17}$$

where $w_1 \neq w_2$, $w_1, w_2 \in ext\left(\mathcal{W}\right)$, $\lambda \in (0,1)$. We get:

$$\begin{aligned} u &= wp + (1-w)p_u \\ &= p_u + (\lambda w_1 + (1-\lambda)w_2)(p - p_u) \\ &= p_u + \lambda w_1(p - p_u) + (1-\lambda)w_2(p - p_u) \\ &\quad + \lambda p_u - \lambda p_u \\ &= \lambda(p_u + w_1(p - p_u)) + (1-\lambda)p_u \\ &\quad + (1-\lambda)w_2(p - p_u) \\ &= \lambda(p_u + w_1(p - p_u)) \\ &\quad + (1-\lambda)(p_u + w_2(p - p_u)) \\ &= \lambda(w_1 p + (1-w_1)p_u) \\ &\quad + (1-\lambda)(w_2 p + (1-w_2)p_u) \end{aligned} \tag{18}$$

Hence $u \in \mathcal{D}\left(ext\left(\mathcal{P}_X\right), ext\left(\mathcal{W}\right)\right)$, which is a contradiction.

Case 2 – $p \notin ext\left(\mathcal{P}_X\right)$, $w \in ext\left(\mathcal{W}\right)$:
We know that p can be expressed as:

$$p = \sum_{i=1}^{n} \alpha_i p_i, \tag{19}$$

where $p_i \in ext\left(\mathcal{P}_X\right)$, $\alpha_i \geq 0$, $\sum_{i=1}^{n} \alpha_i = 1$. Therefore:

$$\begin{aligned} u &= w\left(\sum_{i=1}^{n} \alpha_i p_i\right) + (1-w)p_u \\ &= \left(\sum_{i=1}^{n} w\alpha_i p_i\right) + (1-w)p_u \\ &\quad + \left(\sum_{i=1}^{n} \alpha_i (1-w)p_u\right) \\ &\quad - \left(\sum_{i=1}^{n} \alpha_i (1-w)p_u\right) \\ &= \left(\sum_{i=1}^{n} \alpha_i (wp_i + (1-w)p_u)\right) \\ &\quad + (1-w)p_u - \left(\sum_{i=1}^{n} \alpha_i (1-w)p_u\right) \\ &= \sum_{i=1}^{n} \alpha_i (wp_i + (1-w)p_u) \end{aligned} \tag{20}$$

Hence $u \in \mathcal{D}\left(ext\left(\mathcal{P}_X\right), ext\left(\mathcal{W}\right)\right)$, which is a contradiction.

Case 3 – $p \notin ext\left(\mathcal{P}_X\right)$, $w \notin ext\left(\mathcal{W}\right)$:
As is explained in case 1 and 2, we know that:

$$\begin{aligned} w &= \lambda w_1 + (1-\lambda) w_2 \\ p &= \sum_{i=1}^{n} \alpha_i p_i, \end{aligned} \tag{21}$$

We get:

$$\begin{aligned} u = (\lambda w_1 + (1-\lambda) w_2) \left(\sum_{i=1}^{n} \alpha_i p_i\right) \\ + (1 - (\lambda w_1 + (1-\lambda) w_2)) p_u \end{aligned} \tag{22}$$

From Case 1 we know that Eq. (22) is equivalent to:

$$\begin{aligned} u = \lambda \left(w_1 \left(\sum_{i=1}^{n} \alpha_i p_i\right) + (1-w_1)p_u\right) \\ + (1-\lambda)\left(w_2 \left(\sum_{i=1}^{n} \alpha_i p_i\right) + (1-w_2)p_u\right) \end{aligned} \tag{23}$$

From Case 2 we know that Eq. (23) is equivalent to:

$$\begin{aligned} u = \lambda \left(\sum_{i=1}^{n} \alpha_i (w_1 p_i + (1-w_1)p_u)\right) \\ + (1-\lambda)\left(\sum_{i=1}^{n} \alpha_i (w_2 p_i + (1-w_2)p_u)\right) \end{aligned} \tag{24}$$

Hence $u \in \mathcal{D}\left(ext\left(\mathcal{P}_X\right), ext\left(\mathcal{W}\right)\right)$, which is a contradiction.

Since all cases lead to contradictions we must conclude that $\mathcal{D}(\mathcal{P}_X, \mathcal{W}) = \mathcal{D}(ext\left(\mathcal{P}_X\right), ext\left(\mathcal{W}\right))$. □

Let us now revisit the previous presented examples where a strong conflict was present.

5.1 Balanced Conflict – Revisited

Assume that the following set of reliability weights regarding the sources is available:

$$\begin{aligned} \mathcal{W}_1 &= [0.80, 0.90] \\ \mathcal{W}_2 &= [0.90, 0.95] \end{aligned} \tag{25}$$

The result of applying the RBC discounting operator on the operands in Sect. 4.2, utilizing the above set of reliability weights, is seen in Fig. 5, where we denote the discounted resulting credal set as $\mathcal{P}_X^{1_d,2_d}$. We get $\mathcal{I}(\mathcal{P}_X^{1_d,2_d}) \approx 0.53$, hence, a significant difference compared to the non-discounted case in Fig. 3(b).

5.2 Unbalanced Conflict – Revisited

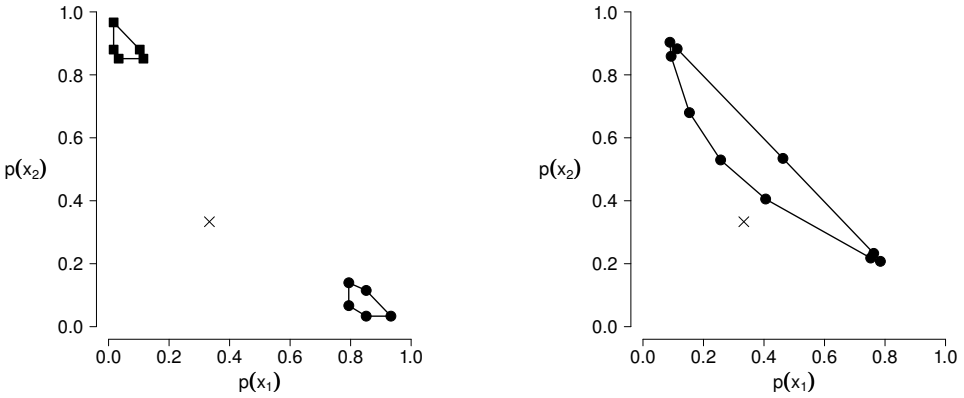
Assume that the following reliability weights regarding the sources are available:

$$\begin{aligned} \mathcal{W}_1 &= [1.00, 1.00] \\ \mathcal{W}_2 &= [0.75, 0.80], \end{aligned} \tag{26}$$

The result of applying the RBC discounting operator on the operands in Sect. 4.3, utilizing the above set of reliability weights, is seen in Fig. 6, where $\mathcal{I}(\mathcal{P}_X^{1_d,2_d}) \approx 0.56$. The lower bound of $\mathcal{W}_2$ will in this case not have any effect since $\mathcal{P}_X^2$ is centered on the uniform distribution.

6 Summary and Conclusions

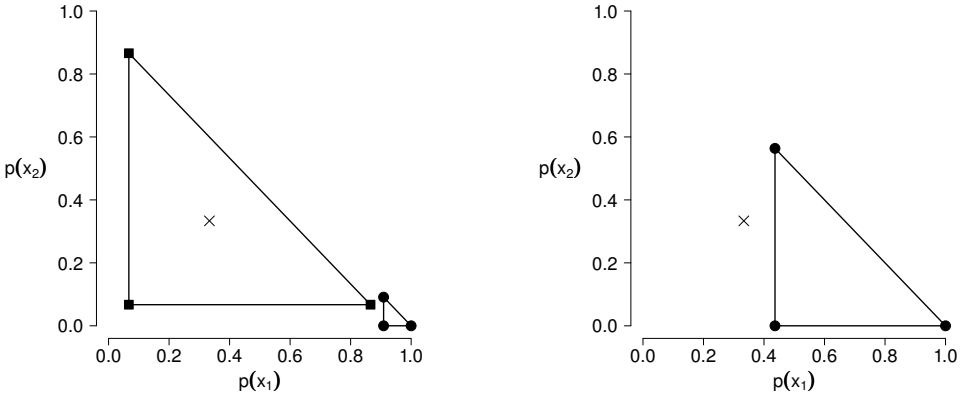
We have studied the combination problem for credal sets via the robust Bayesian combination operator. We extended Walley's notion of degree of imprecision and introduced a measure for degree of conflict between two credal sets. We investigated the behavior of the operator through a number of examples where different degrees of conflict between the operands were



(a) $\mathcal{D}\left(\mathcal{P}_X^1, \mathcal{W}_1\right)$ (circles) and $\mathcal{D}\left(\mathcal{P}_X^2, \mathcal{W}_2\right)$ (squares)

(b) $\mathcal{P}_X^{1_d,2_d}$

Figure 5: $\mathcal{D}\left(\mathcal{P}_X^i, \mathcal{W}_i\right)$, $i \in \{1, 2\}$, and $\mathcal{P}_X^{1_d,2_d}$ for Example 2 – Revisited.



(a) $\mathcal{D}\left(\mathcal{P}_X^1, \mathcal{W}_1\right)$ (circles) and $\mathcal{D}\left(\mathcal{P}_X^2, \mathcal{W}_2\right)$ (squares)

(b) $\mathcal{P}_X^{1_d,2_d}$

Figure 6: $\mathcal{D}\left(\mathcal{P}_X^i, \mathcal{W}_i\right)$, $i \in \{1, 2\}$, and $\mathcal{P}_X^{1_d,2_d}$ for Example 3 – Revisited.

present. We proposed the RBC discounting operator to be used with the combination operator when a set of reliability weights for the sources are available. We showed that the result of the operators can be computed by utilizing the extreme points of the operand sets. Both operators preserve the intuitive paradigm of "point-wise Bayesianism".

An important aspect to recognize when using the robust Bayesian combination operator is that a source, which reports a credal set that is highly imprecise, can considerably affect the result of the combination (see Fig. 4). If a strong conflict is present among the sources, then additional information about the reliability of the sources can be encoded as reliability weights to be used by the RBC discounting operator. If no such information is available, the conflict may be regarded as irrelevant, if a sufficient number of sources make strong statements that are not in conflict (i.e., the sources have made similar observations). For example, if a large number of credal sets similar to $\mathcal{P}_X^1$ in Fig. 4(a), are combined with $\mathcal{P}_X^2$ in the same figure, then the conflict can be sufficiently suppressed to be regarded as irrelevant

Our next step is to evaluate the robust Bayesian combination and discounting operators against other combination operators, e.g., the Bayesian combination operator and Dempster's rule. Such an evaluation must also concern different modeling strategies for obtaining the reliability weights. We are convinced that if credal set theory is going to be accepted by a broader

body of researchers and practitioners, it is necessary to thrust towards research where it can be shown that the theory yields measurable advantages in comparison to other broadly accepted theories, e.g., Bayesian theory.

Acknowledgements

Computations have been performed with *R* [15] with the package *rcdd* [8]. This work was supported by the Information Fusion Research Program (University of Skövde, Sweden) in partnership with the Swedish Knowledge Foundation under grant 2003/0104 (URL: http://www.infofusion.se).

References

[1] Stefan Arnborg. Robust Bayesianism: Imprecise and paradoxical reasoning. In *Proceedings of the 7th International Conference on Information fusion*, 2004.

[2] Stefan Arnborg. Robust Bayesianism: Relation to evidence theory. *Journal of Advances in Information Fusion*, 1(1):63–74, 2006.

[3] Inés Couso, Serafín Moral, and Peter Walley. A survey of concepts of independence for imprecise probabilities. *Risk Decision and Policy*, 5:165–181, 2000.

[4] Fabio Cozman. *Decision Making Based on Convex Sets of Probability Distributions: Quasi-Bayesian Networks and Outdoor Visual Position Estimation*. PhD thesis, The Robotics Institute, Carnegie Mellon University, 1997.

[5] Fabio G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.

[6] Fabio Gagliardi Cozman. Graphical models for imprecise probabilities. *International Journal of Approximate Reasoning*, 39:167–184, 2005.

[7] Harold Gordon Eggleston. *Convexity*. CUP Archive, 1958.

[8] Charles J. Geyer, Glen D. Meeden, and incorporates code from cddlib (ver 0.94f) written by Komei Fukuda. *rcdd: rcdd (Computational Geometry)*, 2008. R package version 1.1-1.

[9] Joseph Y. Halpern and Ronald Fagin. Two views of belief: Belief as generalized probability and belief as evidence. *Artificial Intelligence*, 54:275–317, 1992.

[10] Antonio Irpino and Valentino Tontodonato. Cluster reduced interval data using Hausdorff distance. *Computational Statistics*, 21:241–288, 2006.

[11] Isaac Levi. *The enterprise of knowledge*. The MIT press, 1983.

[12] Serafín Moral and José del Sagrado. Aggregation of imprecise probability. Technical report, Department of Computer Science and Artificial Intelligence, 1997.

[13] Robert F. Nau. The aggregation of imprecise probabilities. *Journal of Statistical Planning and Inference*, 105:265–282, 2002.

[14] Benjamin Noack, Vesa Klumpp, Dietrich Brunn, and Uwe D. Hanebeck. Nonlinear Bayesian estimation with convex sets of probability densities. In *Proceedings of the 11th International Conference on Information fusion*, 2008.

[15] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[16] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[17] Philippe Smets. Analyzing the combination of conflicting belief functions. *Information Fusion*, 8:387–412, 2006.

[18] Philippe Smets and Robert Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994.

[19] Vicenç Torra and Yasuo Narukawa. *Modeling Decisions*. Springer, 2007.

[20] Matthias C.M. Troffaes. Generalizing the conjunction rule for aggregating conflict expert opinions. *International Journal of Intelligent Systems*, 21:361–380, 2006.

[21] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

[22] Peter Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:3–57, 1996.

[23] Peter Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24:125–148, 2000.

[24] Lofti A. Zadeh. Review of books: A mathematical theory of evidence. *AI Magazine*, 5:81–83, 1984.

Affinity and Continuity of Credal Set Operator

Tomáš Kroupa
Institute of Information Theory and Automation of the ASCR
Pod Vodárenskou věží 4, 182 08 Prague, Czech Republic
kroupa@utia.cas.cz
and
Department of Mathematics, Faculty of Electrical Engineering
Czech Technical University in Prague, 166 27 Prague

Abstract

The credal set operator is studied as a set-valued mapping that assigns the set of dominating probabilities to a coherent lower prevision on some set of gambles. It is shown that this mapping is affine on certain classes of coherent lower previsions, which enables to find a decomposition of credal sets. Continuity of the credal set operator is investigated on finite universes with the aim of approximating credal sets.

Keywords. credal set, coherent lower prevision, superdifferential, Hausdorff metric

1 Introduction

The main purpose of this paper is to investigate the geometrical-topological relations between the two important classes of imprecise probability models of Walley [12]: coherent lower previsions and credal sets of linear previsions. The credal set operator is studied as a set-valued mapping that sends every coherent lower prevision to the nonempty, weak*-compact and convex set of dominating linear previsions. Since the set of all coherent lower previsions is a convex subset of a linear topological space, the basic question is whether the credal set operator acts as a morphism between the corresponding mathematical objects. Precisely, the question is if the credal set operator is

(i) an *affine mapping*, that is, convex combinations of coherent lower previsions are mapped to the corresponding "convex combinations" of the credal sets,
(ii) *homeomorphism*, provided a topology is introduced on the set of all credal sets.

In Section 2 we introduce basic notions and notations. The main tool used in this paper are the elements of subdifferential (superdifferential) calculus developed for continuous convex (concave) functions [10]. Theorem 2 in Section 3 shows that every credal set can be represented as the superdifferential. This idea goes back to the solution of coalition games by core and appears already in Aubin's work [1]. Further, it is proven that the credal set operator is an affine mapping on the class of all coherent lower probabilities defined on the set of all subsets of some universe (Theorem 3) and on the class of all coherent lower previsions defined on the set of all the gambles (Corollary 1). It is demonstrated in section 3.1 how the former result can be used to obtain a decomposition of credal sets of belief measures.

Section 4 is devoted to the topological properties. The exposition is confined to the case of finite universes. If the Hausdorff metric is introduced on the set of all nonempty compact convex subset of the set of all linear previsions, then the credal set operator is a homeomorphism (Theorem 7). The consequence of this result mentioned in section 4.1 makes possible to approximate an arbitrary credal set by a "simple" credal set in the Hausdorff metric. The study of continuity of credal set operator need not be limited to finite universes. The principal difficulty in the general non-metrizable case is how to define a topology on the set of all nonempty, weak*-compact convex subsets of the dual of the Banach space of all gambles considered in its weak* topology. Only a brief discussion of this issue would, however, lead to introducing quite complicated mathematical apparatus such as uniformities defined on spaces of credal sets (cf. [2, Chapter II]). Since such considerations go far beyond the intended scope of the paper, the general case is left for separate future investigations.

2 Basic Notions

In this section we introduce the notation and repeat the notions and basic results from Walley's theory of imprecise probabilities [12]. Let Ω be a nonempty set. A *gamble* is a bounded function $\Omega \to \mathbb{R}$. If $a \in \mathbb{R}$, then we use the same symbol a to denote a constant

gamble on Ω. By $\mathscr{L}$ we denote the Banach space of all gambles endowed with the supremum norm $\|.\|_\infty$, that is,

$$\|f\|_\infty = \sup\{|f(\omega)| \mid \omega \in \Omega\}, \quad f \in \mathscr{L}.$$

Let $\mathscr{K} \subseteq \mathscr{L}$. A *lower prevision* $\underline{P}$ is a real function defined on $\mathscr{K}$. If the set $\mathscr{K}$ contains only characteristic functions of subsets of Ω, then $\underline{P}$ is called a *lower probability.* The *conjugate upper prevision* $\overline{P}$ is defined on $-\mathscr{K} = \{f \mid -f \in \mathscr{K}\}$ by letting $\overline{P}(f) = -\underline{P}(-f)$ for every $f \in -\mathscr{K}$. A *coherent lower prevision* on $\mathscr{L}$ is a lower prevision $\underline{P}$ defined on $\mathscr{L}$ that satisfies the following conditions for every $f, g \in \mathscr{L}$:

(i) $\underline{P}(f) \geq \inf\{f(\omega) \mid \omega \in \Omega\}$,
(ii) $\underline{P}(\lambda f) = \lambda \underline{P}(f)$, for every $\lambda \geq 0$,
(iii) $\underline{P}(f+g) \geq \underline{P}(f) + \underline{P}(g)$.

In particular, every coherent lower prevision on $\mathscr{L}$ is a continuous concave function on the Banach space $\mathscr{L}$. If $\underline{P}$ is a lower prevision defined on $\mathscr{K}$, then $\underline{P}$ is called *coherent* provided there exists a coherent lower prevision defined on $\mathscr{L}$ and coinciding with $\underline{P}$ on $\mathscr{K}$.

A *linear prevision* P on $\mathscr{L}$ is a self-conjugate coherent lower prevision on $\mathscr{L}$, that is, $P(-f) = -P(f)$ for every $f \in \mathscr{L}$. Every linear prevision P is a positive linear functional on $\mathscr{L}$ with $P(1) = 1$. A real functional defined on $\mathscr{K}$ is called a *linear prevision on* $\mathscr{K}$ if it can be extended to a linear prevision on $\mathscr{L}$. By $\mathscr{L}^*$ we denote the dual Banach space of $\mathscr{L}$: the elements of $\mathscr{L}^*$ are precisely the continuous linear functionals $\mathscr{L} \to \mathbb{R}$. Every linear prevision belongs to $\mathscr{L}^*$.

The sets of linear previsions appearing in the theory of imprecise probabilities are usually not compact in the norm topology of $\mathscr{L}^*$. If the Banach space $\mathscr{L}^*$ is considered with the weak* topology, then the set $\mathcal{P}$ of all linear previsions on $\mathscr{L}$ becomes a weak*-compact subset of $\mathscr{L}^*$ [12, p.610]. Let $\underline{P}$ be a coherent lower prevision on $\mathscr{K}$. The *credal set of* $\underline{P}$ is the set

$$\mathcal{M}(\underline{P}) = \{P \in \mathcal{P} \mid P(f) \geq \underline{P}(f), f \in \mathscr{K}\}.$$

The terminology is not unified so $\mathcal{M}(\underline{P})$ is called a *core* or a *structure* by some authors. The credal set $\mathcal{M}(\underline{P})$ is a nonempty, convex and weak* compact subset of $\mathscr{L}^*$.

Given a coherent lower prevision $\underline{P}$ on $\mathscr{K}$, put

$$E_{\underline{P}}(f) = \inf\{P(f) \mid P \in \mathcal{M}(\underline{P})\}, \text{ for every } f \in \mathscr{L},$$

and call the function $E_{\underline{P}}$ the *natural extension of* $\underline{P}$. Every natural extension $E_{\underline{P}}$ is the (pointwise) smallest coherent lower prevision that extends $\underline{P}$ to the set $\mathscr{L}$.

3 Superdifferential of Coherent Lower Prevision

The notion of superdifferential of a continuous concave function is one of the generalizations of the classical concept of Gâteaux derivative of a differentiable function. In the next paragraph only basic definitions and results are needed. The reader is referred to [10] for details. Although the theory is developed for subdifferentials of convex functions in [10], the analogous definitions and theorems for superdifferentials of concave functions are derived straightforwardly.

Let X be a Banach space and E be a nonempty open convex subset of X. By X^* we denote the dual space of X. In this paragraph we always assume that φ is a *concave function* $E \to \mathbb{R}$: for every $x_1, x_2 \in E$ and every $\alpha \in [0,1]$, we have

$$\varphi(\alpha x_1 + (1-\alpha)x_2) \geq \alpha\varphi(x_1) + (1-\alpha)\varphi(x_2).$$

For every $x_0 \in E$ and $x \in X$, put

$$\mathrm{d}^+\varphi(x_0)(x) = \lim_{t \to 0_+} \frac{\varphi(x_0 + tx) - \varphi(x_0)}{t}$$

and call $\mathrm{d}^+\varphi(x_0)(x)$ the *right-hand directional derivative of* φ *at* x_0. If follows from [10, Lemma 1.2] that the limit defining $\mathrm{d}^+\varphi(x_0)(x)$ exists for every $x_0 \in E$ and every $x \in X$, and that $\mathrm{d}^+\varphi(x_0)$ is a positively homogeneous concave function on X. The function φ is said to be *Gâteaux differentiable at* x_0 if the functional $\mathrm{d}^+\varphi(x_0) : X \to \mathbb{R}$ is actually linear (not necessarily continuous). Equivalently, the function φ is Gâteaux differentiable at $x_0 \in E$ iff the limit

$$\mathrm{d}\varphi(x_0)(x) = \lim_{t \to 0} \frac{\varphi(x_0 + tx) - \varphi(x_0)}{t}$$

exists for each $x \in X$, and in this case $\mathrm{d}\varphi(x_0) = \mathrm{d}^+\varphi(x_0)$. The functional $\mathrm{d}\varphi(x_0)$ is the *Gâteaux derivative of* φ *at* x_0.

Definition 1. *Let* $x \in E$. *The* superdifferential of φ at x *is the set*

$$\partial\varphi(x) = \{\varphi^* \in X^* \mid \varphi^*(y) \geq \mathrm{d}^+\varphi(x)(y),\ y \in X\}.$$

The superdifferential of φ at x can be equivalently expressed as

$$\partial\varphi(x) = \{\varphi^* \in X^* \mid \varphi^*(y-x) \geq \varphi(y) - \varphi(x),\ y \in E\}. \tag{1}$$

The elements of $\partial\varphi(x)$ are called *supergradients of* φ *at* x. Each supergradient $\varphi^* \in X^*$ is viewed as a plausible candidate for a derivative of φ at x. The following existence result is well-known ([10, Proposition 1.11]).

Theorem 1. *Let X be a Banach space and E be a nonempty open convex subset of X. If the concave function φ is continuous at $x \in E$, then $\partial\varphi(x)$ is a nonempty, convex and weak*-compact subset of X^*.*

For example, let $X = E = \mathbb{R}^2$ and $\varphi(x) = \varphi(x_1, x_2) = -|x_1| - |x_2|$, for every $x = (x_1, x_2) \in \mathbb{R}^2$. Since φ is continuous and concave, the superdifferential of φ at 0 exists. The direct calculation shows that $\partial\varphi(0)$ equals the convex hull of the set of vectors $\{(1,1),(1,-1),(-1,1),(-1,-1)\}$.

The next theorem enables to identify the credal set of $\underline{P}$ with the set of all supergradients at 1 of the natural extension of $\underline{P}$.

Theorem 2. *Let $\mathscr{K} \subseteq \mathscr{L}$. If $\underline{P}$ is a coherent lower prevision on $\mathscr{K}$ and $E_{\underline{P}}$ is the corresponding natural extension, then*

$$\mathcal{M}(\underline{P}) = \partial E_{\underline{P}}(1).$$

Moreover, if $E_{\underline{P}}$ is Gâteaux differentiable at 1, then $\underline{P}$ is a linear prevision on $\mathscr{K}$.

Proof. Let $P \in \mathcal{M}(\underline{P})$. Then $P \geq E_{\underline{P}}$ and $P(1) = 1 = E_{\underline{P}}(1)$, which implies for every gamble f that

$$P(f) - P(1) \geq E_{\underline{P}}(f) - E_{\underline{P}}(1).$$

Since every linear prevision is a norm continuous linear functional, the inequality above means that P is a supergradient of $E_{\underline{P}}$ at 1 by (1).

Suppose, on the other hand, that $P^* \in \partial E_{\underline{P}}(1)$. The equation (1) gives that for every gamble $f \in \mathscr{L}$ we have

$$P^*(f - 1) \geq E_{\underline{P}}(f) - 1. \tag{2}$$

Hence for every real $\alpha > 0$,

$$P^*(\alpha f - 1) \geq E_{\underline{P}}(\alpha f) - 1,$$

and after dividing by α,

$$P^*(f) - \frac{P^*(1)}{\alpha} \geq E_{\underline{P}}(f) - \frac{1}{\alpha}.$$

Letting $\alpha \to 0$ leads to $P^*(f) \geq E_{\underline{P}}(f)$. If $f = 0$, then $P^*(-1) \geq -1$ from (2) so that $P^*(1) = 1$. The functional P^* is a linear prevision as P^* is self-conjugate and satisfies

$$P^*(f) \geq E_{\underline{P}}(f) \geq \inf\{f(\omega) \mid \omega \in \Omega\}$$

for every $f \in \mathscr{L}$. Since $E_{\underline{P}}(f) = \underline{P}(f)$ whenever $f \in \mathscr{K}$, we get $P^* \in \mathcal{M}(\underline{P})$.

To prove the second assertion, assume that $E_{\underline{P}}$ is Gâteaux differentiable at 1. It follows from [10, Proposition 1.8] that this is equivalent to

$$\partial E_{\underline{P}}(1) = \{\mathrm{d}E_{\underline{P}}(1)\}.$$

Since $\mathcal{M}(E_{\underline{P}}) = \partial E_{\underline{P}}(1)$, this means that the continuous concave function $E_{\underline{P}}$ is dominated by the unique continuous linear functional $\mathrm{d}E_{\underline{P}}(1)$. The Hahn-Banach theorem then implies that $E_{\underline{P}}$ itself must be linear and hence, a fortiori, $\underline{P}$ must be a linear prevision. □

The second assertion of the previous theorem can not be reversed: if $\underline{P}$ is a linear prevision on $\mathscr{K}$, then the natural extension $E_{\underline{P}}$ is not in general Gâteaux differentiable at 1.

On the set $2^{\mathscr{L}^*}$ of all subsets of $\mathscr{L}^*$ we consider the multiplication of a set $\mathcal{A} \subseteq \mathscr{L}^*$ by a real number α and the (Minkowski) sum of sets $\mathcal{A}_1 \subseteq \mathscr{L}^*$ and $\mathcal{A}_2 \subseteq \mathscr{L}^*$:

$$\alpha\mathcal{A} = \{\alpha P^* \mid P^* \in \mathcal{A}\},$$
$$\mathcal{A}_1 \oplus \mathcal{A}_2 = \{P_1^* + P_2^* \mid P_1^* \in \mathcal{A}_1, P_2^* \in \mathcal{A}_2\}.$$

Let K_1, K_2 be convex subsets of linear spaces X_1, X_2, respectively. A mapping $a : K_1 \to K_2$ is *affine*, whenever for every convex combination $\sum_{i=1}^n \alpha_i x_i$ of elements $x_1, \ldots, x_n \in K_1$, we have

$$a\left(\sum_{i=1}^n \alpha_i x_i\right) = \sum_{i=1}^n \alpha_i a(x_i).$$

Let 2^Ω be the set of all subsets of Ω. A lower probability $\underline{P}$ on 2^Ω is *supermodular* if

$$\underline{P}(A \cup B) + \underline{P}(A \cap B) \geq \underline{P}(A) + \underline{P}(B)$$

for every $A, B \in 2^\Omega$.

Theorem 3. *If $\underline{P}^1, \ldots, \underline{P}^n$ are supermodular coherent lower probabilities on 2^Ω and $\alpha_i \in [0,1], i = 1, \ldots, n$, are such that $\sum_{i=1}^n \alpha_i = 1$, then*

$$\mathcal{M}\left(\sum_{i=1}^n \alpha_i \underline{P}^i\right) = \bigoplus_{i=1}^n \alpha_i \mathcal{M}(\underline{P}^i). \tag{3}$$

Proof. The lower probability $\sum_{i=1}^n \alpha_i \underline{P}^i$ is coherent [12, Theorem 2.6.4]. The coherent lower probability $\sum_{i=1}^n \alpha_i \underline{P}^i$ is supermodular since each $\underline{P}^i$ is supermodular, so the set of all supermodular coherent lower probabilities on 2^Ω is a convex subset of $\mathbb{R}^{2^\Omega}$. It follows from [8, Theorem 5.2] that the natural extension $E_{\underline{P}}$ of any supermodulat coherent lower probability $\underline{P}$ on 2^Ω coincides with the asymmetric Choquet integral $I^a_{\underline{P}} : \mathscr{L} \to \mathbb{R}$, where

$$I^a_{\underline{P}}(f) = \int_{-\infty}^0 \underline{P}(f^{-1}((t,\infty))) - \underline{P}(\Omega)\,\mathrm{d}t + \int_0^\infty \underline{P}(f^{-1}((t,\infty)))\,\mathrm{d}t,$$

for every $f \in \mathscr{L}$. A routine verification shows that the mapping sending each supermodular coherent lower probability $\underline{P}$ to $I^a_{\underline{P}}$ is affine, hence

$$E_{\sum_{i=1}^n \alpha_i \underline{P}^i} = I^a_{\sum_{i=1}^n \alpha_i \underline{P}^i} = \sum_{i=1}^n \alpha_i I^a_{\underline{P}^i} = \sum_{i=1}^n \alpha_i E_{\underline{P}^i}$$

Theorem 2 together with the preceding equality give

$$\mathcal{M}\left(\sum_{i=1}^n \alpha_i \underline{P}^i\right) = \partial\left(E_{\sum_{i=1}^n \alpha_i \underline{P}^i}\right)(1) = \\ = \partial\left(\sum_{i=1}^n \alpha_i E_{\underline{P}^i}\right)(1).$$

It follows directly from the definition of superdifferential that for every $i = 1, \ldots, n$,

$$\partial(\alpha_i E_{\underline{P}^i})(1) = \alpha_i \partial(E_{\underline{P}^i})(1). \tag{4}$$

By the Moreau-Rockafellar theorem [10, Theorem 3.23], the equality (4) and Theorem 2, we obtain

$$\partial\left(\sum_{i=1}^n \alpha_i E_{\underline{P}^i}\right)(1) = \bigoplus_{i=1}^n \partial(\alpha_i E_{\underline{P}^i})(1) = \\ = \bigoplus_{i=1}^n \alpha_i \partial(E_{\underline{P}^i})(1) = \bigoplus_{i=1}^n \alpha_i \mathcal{M}(\underline{P}^i).$$

□

One of key ingredients in the above proof is the affinity of the natural extension operator $\underline{P} \mapsto E_{\underline{P}}$ derived from the representation of the natural extension by the assymetric Choquet integral [8, Theorem 5.2]. This suggests the following general result.

Theorem 4. *Let $\mathscr{K}$ be a set of gambles and $\underline{\mathcal{C}}_{\mathscr{K}}$ be the convex set of all coherent lower probabilities on $\mathscr{K}$. If the mapping*

$$\underline{P} \in \underline{\mathcal{C}}_{\mathscr{K}} \mapsto E_{\underline{P}}$$

is affine, then the equality (3) *holds true for every $\underline{P}^1, \ldots, \underline{P}^n \in \underline{\mathcal{C}}_{\mathscr{K}}$.*

Proof. Let $\underline{P}^1, \ldots, \underline{P}^n \in \underline{\mathcal{C}}_{\mathscr{K}}$ and $\alpha_i \in [0,1], i = 1, \ldots, n$, be such that $\sum_{i=1}^n \alpha_i = 1$. Then

$$E_{\sum_{i=1}^n \alpha_i \underline{P}^i} = \sum_{i=1}^n \alpha_i E_{\underline{P}^i},$$

so

$$\mathcal{M}\left(\sum_{i=1}^n \alpha_i \underline{P}^i\right) = \partial\left(\sum_{i=1}^n \alpha_i E_{\underline{P}^i}\right)(1),$$

and the equality (3) again follows from the Moreau-Rockafellar theorem [10, Theorem 3.23] together with (4) and Theorem 2. □

Let S be the set of all nonempty weak*-compact convex subsets of $\mathcal{P}$. In the sequel we will study the properties of the set-valued mapping

$$\mathcal{M}(.) : \underline{P} \mapsto \mathcal{M}(\underline{P})$$

that sends a coherent lower probability on some set of gambles $\mathscr{K}$ to a credal set from S. A superficial look at the equality (3) would then suggest that this mapping is affine on the class of coherent lower probabilities mentioned in Theorem 3. A necessary condition is that the codomain S of $\mathcal{M}$ is a convex set. But this notion of convexity is not even defined in the present framework since the set $2^{\mathscr{L}^*}$ endowed with the Minkowski sum of sets and the scalar multiplication of a set is not a linear space. The main difficulty is that the algebra $(2^{\mathscr{L}^*}, \oplus)$ is not a group but only a commutative monoid. The properties of the Minkowski sum and the scalar multiplication of sets defined above can be summarized as follows.

Proposition 1. *The set $2^{\mathscr{L}^*}$ together with the Minkowski sum $\oplus$ is a* real semilinear space*, that is:*

(i) the algebra $(2^{\mathscr{L}^*}, \oplus)$ is a commutative monoid with the neutral element $\{0\}$,
(ii) $\alpha(\beta\mathcal{A}) = (\alpha\beta)\mathcal{A}$, for every $\alpha, \beta \in \mathbb{R}$ and every $\mathcal{A} \in 2^{\mathscr{L}^*}$,
(iii) $1\mathcal{A} = \mathcal{A}$,
(iv) $0\mathcal{A} = \{0\}$,
(v) $\alpha(\mathcal{A}_1 \oplus \mathcal{A}_2) = (\alpha\mathcal{A}_1) \oplus (\alpha\mathcal{A}_2)$, for every $\mathcal{A}_1, \mathcal{A}_2 \in 2^{\mathscr{L}^*}$.

Semilinear spaces, which generalize linear spaces, are algebraic structures close to semirings [5]. The definitions of convexity and affine maps can be directly carried over to a more general framework of semilinear spaces. In that follows these generalized definitions are tacitly assumed. Thus we will say that S is *convex* (as a subset of $2^{\mathscr{L}^*}$) if

$$\alpha\mathcal{A}_1 \oplus (1-\alpha)\mathcal{A}_2 \in \mathsf{S}$$

holds true for every $\mathcal{A}_1, \mathcal{A}_2 \in \mathsf{S}$ and every $\alpha \in [0,1]$.

Proposition 2. *The set S is a convex subset of the real semilinear space $2^{\mathscr{L}^*}$.*

Proof. Consider any $\mathcal{A}_1, \mathcal{A}_2 \in \mathsf{S}$ and a real number $\alpha \in [0,1]$. Put $\mathcal{A} = \alpha\mathcal{A}_1 \oplus (1-\alpha)\mathcal{A}_2$. Then $\mathcal{A}$ is a nonempty convex subset of $\mathcal{P}$ since both $\mathcal{A}_1, \mathcal{A}_2$ are nonempty and convex. As both $\alpha\mathcal{A}_1$ and $(1-\alpha)\mathcal{A}_2$ are weak*-closed, their Minkowski sum $\mathcal{A}$ is a weak*-closed subset of $\mathcal{P}$, and thus weak*-compact. □

With these facts in mind, it is safe to interpret the conclusions of Theorem 3 and 4 as expressing the fact that "the mapping $\mathcal{M}$ is affine". We will show that

the mapping $\mathcal{M}$ is an affine isomorphism[1] from the convex set $\underline{\mathcal{C}}$ of all coherent lower previsions on $\mathscr{L}$ to S. The essential result is the following theorem [12, Theorem 3.6.1].

Theorem 5 (Walley)**.** *The mapping $\mathcal{M}$ is a bijection from $\underline{\mathcal{C}}$ to S. The inverse mapping $\mathcal{M}^{-1}$ sends $\mathcal{A} \in \mathsf{S}$ to the coherent lower prevision*

$$\mathcal{M}^{-1}(\mathcal{A})(f) = \inf\{P(f) \mid P \in \mathcal{A}\}, \quad f \in \mathscr{L}.$$

Corollary 1. *The mapping $\mathcal{M}$ is an affine isomorphism of $\underline{\mathcal{C}}$ and S.*

Proof. The mapping $\mathcal{M}$ is one-to-one by Theorem 5. It suffices to show that $\underline{P} \in \underline{\mathcal{C}} \mapsto E_{\underline{P}}$ is affine since this gives the affinity of $\mathcal{M}$ by Theorem 4. However, this is trivial as $\underline{P} = E_{\underline{P}}$ for every $\underline{P} \in \underline{\mathcal{C}}$. □

Hence the mutual correspondence between the two different models of imprecise probabilities (coherent lower previsions and credal sets) introduced by Walley is retained also on the geometric level.

3.1 Decomposition of credal sets

Theorem 3 can be useful in situations in which a coherent lower probability $\underline{P}$ on 2^Ω is a convex combination of the coherent lower probabilities whose credal sets have a special shape (such as simplices). In this case, the credal set of $\underline{P}$ is decomposed into the convex combination of the respective "basic" credal sets. In particular, Theorem 3 is an infinite-dimensional generalization of Corollary 4 from [3], where a similar result is achieved for finite Ω and totally monotone set functions investigated in the framework of cooperative games. We will explicitly show how Theorem 3 can be applied to the credal sets of belief measures [11] by reformulating [3, Corollary 4] as a consequence of Theorem 3 in this paper.

Theorem 6. *Let Ω be finite, $\underline{P}$ be a belief measure on 2^Ω and $\mu^{\underline{P}}$ its Möbius transform. Then*

$$\mathcal{M}(\underline{P}) = \bigoplus_{A \subseteq \Omega} \mu^{\underline{P}}(A) \mathcal{S}_A,$$

where $\mathcal{S}_A$ is the simplex of probabilities on 2^Ω supported by A, that is, $\mathcal{S}_A = \{P \in \mathcal{P} \mid P(A) = 1\}$.

Proof. The set $\mathcal{S}_A$ is a simplex since it is a face of the simplex of all probabilities on 2^Ω. A belief measure $\underline{P}$ is a supermodular coherent lower probability on 2^Ω, so Theorem 3 can be employed. Since $\sum_{A \subseteq \Omega} \mu^{\underline{P}}(A) = 1$, where $\mu^{\underline{P}}(A) \geq 0$ for each $A \subseteq \Omega$, and

$$\underline{P} = \sum_{A \subseteq \Omega} \mu^{\underline{P}}(A) \underline{P}_A,$$

where the set functions

$$\underline{P}_A(B) = \begin{cases} 1, & A \subseteq B, \\ 0, & \text{otherwise}, \end{cases}$$

are belief functions, it suffices to realize that $\mathcal{M}(\underline{P}_A) = \mathcal{S}_A$. □

[1] An *affine isomorphism* is a bijective affine mapping between two convex subsets of real semilinear spaces. Its inverse is then necessarily an affine mapping too.

4 Continuity of Credal Set Mapping

The main purpose of this section is to study the topological properties of the credal set operator. We will confine the investigations to the case of finite Ω. The first necessary step is an introduction of topologies on both $\underline{\mathcal{C}}$ and S.

If $\Omega = \{1, \ldots, n\}$, then the set of all gambles $\mathscr{L}$ can be identified with the Euclidean space $\mathbb{R}^n$. A gamble is then viewed as an n-dimensional vector $f = (f_1, \ldots, f_n) \in \mathbb{R}^n$. The dual space $\mathscr{L}^*$ is identified with $\mathbb{R}^n$. If $\langle ., . \rangle$ denotes the usual scalar product on $\mathbb{R}^n$, then every linear prevision P on $\mathscr{L}$ canonically corresponds to the vector of reals $p = (p_1, \ldots, p_n)$ such that $\langle p, 1 \rangle = 1$ and $p_i \geq 0$ for each $i = 1, \ldots, n$. We have

$$P(f) = \langle p, f \rangle, \quad f \in \mathscr{L}. \tag{5}$$

The pointwise limit of coherent lower previsions on any set of gambles $\mathscr{K}$ is a coherent lower prevision on $\mathscr{K}$. Consequently, the set $\underline{\mathcal{C}}$ is a closed convex subset of the locally convex space $\mathbb{R}^{\mathscr{L}}$. Let $\|.\|$ be the Euclidean norm on $\mathbb{R}^n$. The topology of pointwise convergence on $\underline{\mathcal{C}}$ is described by the metric

$$\Delta(\underline{P}^1, \underline{P}^2) = \max\{|\underline{P}^1(f) - \underline{P}^2(f)| \mid \|f\| \leq 1\}.$$

Precisely, the sequence $(\underline{P}_n)$ in $\underline{\mathcal{C}}$ pointwise converges to $\underline{P} \in \underline{\mathcal{C}}$ iff $\Delta(\underline{P}_n, \underline{P}) \to 0$ (see [7, Theorem 1.3.5,p.133]).

The set S contains all the nonempty compact convex subsets of

$$\mathcal{P} = \{p \in \mathbb{R}^n \mid \langle p, 1 \rangle = 1, p_i \geq 0,\ i = 1 \ldots, n\}.$$

The topology on S can be introduced by the Hausdorff metric [2, Chapter II]. For every $\mathcal{A} \in \mathsf{S}$ and every $p \in \mathcal{P}$, define

$$d_{\mathcal{A}}(p) = \min\{\|p - p'\| \mid p' \in \mathcal{A}\}. \tag{6}$$

If $\mathcal{A}_1, \mathcal{A}_2 \in \mathsf{S}$, then denote

$$e_H(\mathcal{A}_1, \mathcal{A}_2) = \sup\{d_{\mathcal{A}_2}(p_1) \mid p_1 \in \mathcal{A}_1\}.$$

The *Hausdorff metric* Δ_H on S is defined as

$$\Delta_H(\mathcal{A}_1, \mathcal{A}_2) = \max\{e_H(\mathcal{A}_1, \mathcal{A}_2), e_H(\mathcal{A}_2, \mathcal{A}_1)\},$$

for every $\mathcal{A}_1, \mathcal{A}_2 \in \mathsf{S}$.

The topology corresponding to the metric Δ_H is called the *Hausdorff metric topology.* The Hausdorff metric topology depends only on the topology of $\mathcal{P}$: if any metric equivalent to the Euclidean metric is used in place of $\|.\|$ in (6), the resulting metric topology on S would coincide with the Hausdorff metric topology. Indeed, it follows from [2, Theorem II-6] that the Hausdorff metric topology on the family K of nonempty compact subsets of $\mathcal{P}$ is generated by the sets

$$\{K \in \mathsf{K} \mid K \subseteq U,\ U \text{ open in } \mathcal{P}\}$$

and

$$\{K \in \mathsf{K} \mid K \cap V \neq \emptyset,\ V \text{ open in } \mathcal{P}\}.$$

The Hausdorff metric topology of S arises as a subspace topology from K. Hence it is immaterial if the Euclidean norm or the supremum norm originally defined on the space of gambles is used.

Theorem 7. *Let $\Omega = \{1, \ldots, n\}$. If S is endowed with the Hausdorff metric topology, then the mapping $\mathcal{M} : \underline{\mathcal{C}} \to \mathsf{S}$ is an affine isomorphism and homeomorphism.*

Proof. The mapping $\mathcal{M}$ is an affine isomorphism by Corollary 1 so that it remains to prove the continuity in both directions. To this end, we use the following convergence result, which can be easily deduced from [7, Corollary 3.3.8, p.156]: if $(\mathcal{A}_n)$ is a sequence in S and $\mathcal{A} \in \mathsf{S}$, then $\mathcal{A}_n \to \mathcal{A}$ in the Hausdorff metric iff the sequence of functions

$$((f \in \mathbb{R}^n \mapsto \inf\{\langle p, f\rangle \mid p \in \mathcal{A}_n\})_n)$$

pointwise converges to the function

$$f \in \mathbb{R}^n \mapsto \inf\{\langle p, f\rangle \mid p \in \mathcal{A}\}.$$

To show that the mapping $\mathcal{M}$ is continuous, consider a sequence $(\underline{P}_n)$ converging to $\underline{P}$ in $\underline{\mathcal{C}}$. Theorem 5 and (5) together yield

$$\underline{P}_n(f) = \mathcal{M}^{-1}(\mathcal{M}(\underline{P}_n))(f) = \inf\{\langle p, f\rangle \mid p \in \mathcal{M}(\underline{P}_n)\}$$

and

$$\underline{P}(f) = \mathcal{M}^{-1}(\mathcal{M}(\underline{P}))(f) = \inf\{\langle p, f\rangle \mid p \in \mathcal{M}(\underline{P})\}.$$

This implies $\mathcal{M}(\underline{P}_n) \to \mathcal{M}(\underline{P})$ in the Hausdorff metric. Continuity of the inverse mapping $\mathcal{M}^{-1}$ is shown similarly. □

4.1 Approximation of credal sets

By Theorem 5 of Walley every nonempty compact convex subset $\mathcal{A} \in \mathsf{S}$ is a credal set of the coherent lower prevision $\mathcal{M}^{-1}(\mathcal{A})$. Although every credal set is characterized by the Krein-Milman theorem as the closed convex hull of its vertices, it can be convenient to find a class of subsets of S whose members have a particular geometric structure and which is sufficient for an approximation of every credal set. The polytopes from S are natural candidates for such a task. A *polytope* is the convex hull of finitely-many points in $\mathbb{R}^n$. For our purposes it will be even enough to focus on so-called simple polytopes. A polytope is called *simple* if each of its vertices is contained in the same number of facets. For example, a cube or a simplex are simple polytopes, an Egyptian pyramid is not a simple polytope. It was proven in [9] that the credal set of every possibility measure is a simple polytope. The class of simple polytopes is considered to be computationally tractable: see [13] or the discussion in [9, p.243-244] and the references therein.

Theorem 8. *Let $\Omega = \{1, \ldots, n\}$ and S be endowed with the Hausdorff metric topology. If $\underline{P}$ is any coherent lower prevision on a set of gambles $\mathscr{K} \subseteq \mathbb{R}^n$, then there exists a sequence $(\mathcal{S}_n)$ of simple polytopes in S such that*

(i) $\mathcal{S}_n \to \mathcal{M}(\underline{P})$ *in the Hausdorff metric,*
(ii) $\mathcal{M}^{-1}(\mathcal{S}_n) \to \underline{P}$ *pointwise on* $\mathscr{K}$*,*
(iii) $\mathcal{M}^{-1}(\mathcal{S}_n) \to \underline{P}$ *uniformly on each compact subset of* $\mathscr{K}$*.*

Proof. (i) is basically Theorem 2.8 in [4], which says that simple polytopes form a dense set in S. The assertion (ii) follows from (i) in conjunction with Theorem 7: the sequence of coherent lower previsions $\mathcal{M}^{-1}(\mathcal{S}_n)$ pointwise converges to $\underline{P}$ on $\mathscr{K}$ as $\mathcal{M}^{-1}$ is continuous. The last assertion (iii) is a well-known property of the convergence of concave functions $\mathbb{R}^n \to \mathbb{R}$ (see [7, Theorem B.3.1.4], for instance). □

The proof of [4, Theorem 2.8] is based on a strong compactness argument: given any open cover of a polytope K by balls with a given diameter and with centers in the extreme boundary of K, there exists a finite refinement of this cover. The idea is analogous to inscribing a polygon into a circle. Hence the theorem does not give an algorithm for finding the convergent sequence of simple polytope. Nevertheless, at least in case that $\mathcal{M}(\underline{P})$ is a polytope, it is possible to explicitly find a simple polytope "sufficiently close to $\mathcal{M}(\underline{P})$" [13].

5 Conclusions

In this contribution we identified two main cases in which the credal set mapping is affine (Theorem 3 and Corollary 1). Yet none of them covers the whole variety of coherent lower previsions since "completeness" of their domains is required: the set of gambles is required to be the set of all events or the set of all the gambles. Theorem 4 then gives a sufficient condition enabling to get rid of those assumptions: it is the affinity of the natural extension operator. In general, the natural extension operator is not stable under the usual operations with imprecise probabilities: it need not preserve neither convex combinations nor limits of convergent sequences of coherent lower previsions [6, Section 5]. In future investigations our aim will be to single out the sets of gambles satisfying the assumption of Theorem 4 and to extend the material presented in Section 4 to infinite universes.

Acknowledgements

The work of the author was supported by the grant GA ČR 201/09/1891 and by the grant No.1M0572 of the Ministry of Education, Youth and Sports of the Czech Republic.

References

[1] J.-P. Aubin. Coeur et valeur des jeux flous à paiements latéraux. *C. R. Acad. Sci. Paris Sér. A*, 279:891–894, 1974.

[2] C. Castaing and M. Valadier. *Convex analysis and measurable multifunctions*. Springer-Verlag, Berlin, 1977. Lecture Notes in Mathematics, Vol. 580.

[3] V.I. Danilov and G.A. Koshevoy. Cores of cooperative games, superdifferentials of functions, and the Minkowski difference of sets. *J. Math. Anal. Appl.*, 247(1):1–14, 2000.

[4] G. Ewald. *Combinatorial convexity and algebraic geometry*, volume 168 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1996.

[5] J.S. Golan. *Semirings and their applications*. Kluwer Academic Publishers, Dordrecht, 1999.

[6] R. Hable. *Data-based decisions under complex uncertainty*. PhD thesis, Ludwig-Maximilians-Universität (LMU) Munich, 2009.

[7] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of convex analysis*. Grundlehren Text Editions. Springer-Verlag, Berlin, 2001.

[8] V. Krätschmer. Coherent lower previsions and Choquet integrals. *Fuzzy Sets and Systems*, 138(3):469–484, 2003.

[9] T. Kroupa. Geometry of possibility measures on finite sets. *Internat. J. Approx. Reason.*, 48(1):237–245, 2008.

[10] R. R. Phelps. *Convex functions, monotone operators and differentiability*, volume 1364 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1989.

[11] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press. Princeton, NJ, 1976.

[12] P. Walley. *Statistical reasoning with imprecise probabilities*, volume 42 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London, 1991.

[13] G. Ziegler. *Lectures on polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995.

Imprecise Probabilities from Imprecise Descriptions of Real Numbers

Jonathan Lawry
Dept of Engineering Maths,
University of Bristol,
Bristol, BS8 1TR, UK
j.lawry@bris.ac.uk

Inés González-Rodríguez
Dept. of Maths, Stats and Compt.
University of Cantabria
Santander, Spain
ines.gonzalez@unican.es

Yongchuan Tang
College of Computer Science,
Zhejiang University,
Hangzhou 310027, PR China
tyongchuan@gmail.com

Abstract

A prototype theory interpretation of the label semantics framework is proposed as a possible model of imprecise descriptions of real numbers. It is shown that within this framework conditioning given imprecise descriptions of a real variable naturally results in imprecise probabilities. An inference method is proposed from data in the form of a set of imprecise descriptions, which naturally suggests an algorithm for estimating lower and upper probabilities given imprecise data values.

Keywords. Label Semantics, Prototype Theory, Random Sets, Lower and Upper Distributions, Second Order Distributions

1 Introduction

The label semantics framework [3], [4] is an epistemic theory of the uncertainty associated with vague or imprecise descriptions of an object or value. In label semantics the focus is on the decision making process an intelligent agent must go through in order to identify which labels or expressions can actually be used to describe an object or value. In other words, in order to make an assertion describing an object in terms of some set of linguistic labels, an agent must first identify which of these labels are appropriate or assertible in this context. Given the way that individuals learn language through an ongoing process of interaction with the other communicating agents and with the environment, then we can expect there to be considerable uncertainty associated with any decisions of this kind. Furthermore, there is a subtle assumption central to the label semantic model, that such decisions regarding appropriateness or assertibility are meaningful. For instance, the fuzzy logic view is that vague descriptions like 'John is *tall*' are generally only partially true and hence it is not meaningful to consider which of a set of given labels can truthfully be used to described John's height. However, we contest that the efficacy of natural language as a means of conveying information between members of a population lies in shared conventions governing the appropriate use of words which are, at least loosely, adhered to by individuals within the population.

In our everyday use of language we are continually faced with decisions about the best way to describe objects and instances in order to convey the information we intend. For example, suppose you are witness to a robbery, how should you describe the robber so that police on patrol in the streets will have the best chance of spotting him? You will have certain labels that can be applied, for example *tall, short, medium, fat, thin, blonde, etc*, some of which you may view as inappropriate for the robber, others perhaps you think are definitely appropriate while for some labels you are uncertain whether they are appropriate or not. On the other hand, perhaps you have some ordered preferences between labels so that *tall* is more appropriate than *medium* which is in turn more appropriate than *short*. Your choice of words to describe the robber should surely then be based on these judgments about the appropriateness of labels. Yet where does this knowledge come from and more fundamentally what does it actually mean to say that a label is or is not appropriate? Label semantics proposes an interpretation of vague description labels based on a particular notion of appropriateness and suggests a measure of subjective uncertainty resulting from an agent's partial knowledge about what labels are appropriate to assert. Furthermore, it is suggested that the vagueness of these description labels lies fundamentally in the uncertainty about if and when they are appropriate as governed by the rules and conventions of language use.

The above argument brings us very close to the epistemic view of vagueness as expounded by Timothy Williamson [12]. Williamson assumes that for the extensions of a vague concept there is a precise but unknown dividing boundary between it and the ex-

tension of the negation of that concept. However, while there are marked similarities between the epistemic theory and the label semantics view, there are also some subtle differences. For instance, the epistemic view would seem to assume the existence of some objectively correct, but unknown, definition of a vague concept. Instead of this we argue that individuals when faced with decision problems regarding assertions find it useful as part of a decision making strategy to assume that there is a clear dividing line between those labels which are and those which are not appropriate to describe a given instance. We refer to this strategic assumption across a population of communicating agents as the *epistemic stance* [5], a concise statement of which is as follows:

> *Each individual agent in the population assumes the existence of a set of labeling conventions, valid across the whole population, governing what linguistic labels and expressions can be appropriately used to describe particular instances.*

In practice these rules and conventions underlying the appropriate use of labels would not be imposed by some outside authority. In fact, they may not exist at all in a formal sense. Rather they are represented as a distributed body of knowledge concerning the assertability of predicates in various cases, shared across a population of agents, and emerging as the result of interactions and communications between individual agents all adopting the epistemic stance. The idea is that the learning processes of individual agents, all sharing the fundamental aim of understanding how words can be appropriately used to communicate information, will eventually converge to some degree on a set of shared conventions. The very process of convergence then to some extent vindicates the epistemic stance from the perspective of individual agents. Of course, this is not to suggest complete or even extensive agreement between individuals as to these appropriateness conventions. However, the overlap between agents should be sufficient to ensure the effective transfer of useful information.

In this paper we consider the application of label semantics to model the description of real numbers using vague or imprecise labels. In particular, given a real valued variable x and a label L for real numbers we attempt to understand the nature of the information provided by assertions of the form 'x is L'. Indeed we will argue that from an epistemic perspective such assertions naturally result in imprecise probabilities. The model we propose will be based on a new interpretation of label semantics linking random set theory and Rosch's [9] prototype theory of concepts.

2 The Prototype Interpretation of Label Semantics

Label semantics proposes two fundamental and interrelated measures of the appropriateness of labels as descriptions of an object or value. Given a finite set of labels LA a set of compound expressions LE can then be generated through recursive applications of logical connectives. The labels $L_i \in LA$ are intended to represent words such as adjectives and nouns which can be used to describe elements from the underlying universe Ω. In other words, L_i correspond to description labels for which the expression 'x is L_i' is meaningful for any $x \in \Omega$. For example, if Ω is the set of all possible *rgb* values then LA could consist of the basic colour labels such as *red*, *yellow*, *green*, *orange* etc. In this case LE then contains those compound expression such as *red & yellow*, *not blue nor orange* etc. The measure of appropriateness of an expression $\theta \in LE$ as a description of instance x is denoted by $\mu_\theta(x)$ and quantifies the agent's subjective belief that θ can be used to describe x based on his/her (partial) knowledge of the current labeling conventions of the population. From an alternative perspective, when faced with an object to describe, an agent may consider each label in LA and attempt to identify the subset of labels that are appropriate to use. Let this set be denoted by $\mathcal{D}_x$. In the face of their uncertainty regarding labeling conventions the agent will also be uncertain as to the composition of $\mathcal{D}_x$, and in label semantics this is quantified by a probability mass function $m_x : 2^{LA} \to [0,1]$ on subsets of labels. The relationship between these two measures will be described below.

Definition 1. *Label Expressions*
Given a finite set of labels LA the corresponding set of label expressions LE is defined recursively as follows:

- *If $L \in LA$ then $L \in LE$*
- *If $\theta, \varphi \in LE$ then $\neg\theta, \theta \wedge \varphi, \theta \vee \varphi \in LE$*

The mass function m_x on sets of labels then quantifies the agent's belief that any particular subset of labels contains all and only the labels with which it is appropriate to describe x.

Definition 2. *Mass Function on Labels*
$\forall x \in \Omega$ a mass function on labels is a function $m_x : 2^{LA} \to [0,1]$ such that $\sum_{F \subseteq LA} m_x(F) = 1$

The appropriateness measure, $\mu_\theta(x)$, and the mass function m_x are then related to each other on the basis that asserting 'x is θ' provides direct constraints on $\mathcal{D}_x$. For example, asserting 'x is $L_1 \wedge L_2$', for labels $L_1, L_2 \in LA$ is taken as conveying the infor-

mation that both L_1 and L_2 are appropriate to describe x so that $\{L_1, L_2\} \subseteq \mathcal{D}_x$. Similarly, '$x$ is $\neg L$' implies that L is not appropriate to describe x so $L \notin \mathcal{D}_x$. In general we can recursively define a mapping $\lambda : LE \to 2^{2^{LA}}$ from expressions to sets of subsets of labels, such that the assertion 'x is θ' directly implies the constraint $\mathcal{D}_x \in \lambda(\theta)$ and where $\lambda(\theta)$ is dependent on the logical structure of θ.

Definition 3. *λ-mapping*
$\lambda : LE \to 2^{2^{LA}}$ is defined recursively as follows: $\forall L_i \in LA$, $\forall \theta,\ \varphi \in LE$

- $\lambda(L_i) = \{F \subseteq LA : L_i \in F\}$
- $\lambda(\theta \wedge \varphi) = \lambda(\theta) \cap \lambda(\varphi)$
- $\lambda(\theta \vee \varphi) = \lambda(\theta) \cup \lambda(\varphi)$
- $\lambda(\neg\theta) = \lambda(\theta)^c$

Based on the λ mapping we then define $\mu_\theta(x)$ as the sum of m_x over those sets of labels in $\lambda(\theta)$.

Definition 4. *Appropriateness Measure*
The appropriateness measure defined by mass function m_x is a function $\mu : LA \times \Omega \to [0,1]$ satisfying

$$\forall \theta \in LE,\ \forall x \in \Omega\ \mu_\theta(x) = \sum_{F \in \lambda(\theta)} m_x(F)$$

where $\mu_\theta(x)$ is used as shorthand notation for $\mu(\theta, x)$.

Prototype theory and imprecise probabilities have already been linked by Walley and de Cooman [11] who identified labels based on prototypes as a special case of monotonic predicates which they argue naturally induce possibility distributions. A prototype theory interpretation of Label Semantics has recently been proposed [6], [7], [10] in which the basic labels LA correspond to natural categories each with an associated set of prototypes. A label L_i is then deemed to be an appropriate description of an element $x \in \Omega$ provided x is *sufficiently similar* to the prototypes of L_i. The requirement of being 'sufficiently similar' is clearly imprecise and is modelled here by introducing an uncertain threshold on distance from prototypes.

A distance function d is defined on Ω such that $d : \Omega^2 \to [0, \infty)$ and satisfies $d(x,x) = 0$ and $d(x,y) = d(y,x)$ for all elements $x, y \in \Omega$. This function is then extended to sets of elements such that for $S, T \subseteq \Omega$, $d(S,T) = \inf\{d(x,y) : x \in S \text{ and } y \in T\}$. For each label $L_i \in LA$ let there be a set $P_i \subseteq \Omega$ corresponding to prototypical elements for which L_i is certainly an appropriate description. Within this framework L_i is deemed to be appropriate to describe an element $x \in \Omega$ provided x is sufficiently close or similar to a prototypical element in P_i. This is formalized by the requirement that x is within a maximal distance threshold ϵ of P_i. i.e. L_i is appropriate to describe x if $d(x, P_i) \leq \epsilon$ where $\epsilon \geq 0$. From this perspective an agent's uncertainty regarding the appropriateness of a label to describe a value x is characterised by his or her uncertainty regarding the distance threshold ϵ. Here we assume that ϵ is a random variable and that the uncertainty is represented by a probability density function δ for ϵ defined on $[0, \infty)$. Within this interpretation a natural definition of the complete description of an element $\mathcal{D}_x$ and the associated mass function m_x can be given as follows:

Definition 5. *Prototype Interpretations of $\mathcal{D}_x$ and m_x*
For $\epsilon \in [0, \infty)$ $\mathcal{D}_x^\epsilon = \{L_i \in LA : d(x, P_i) \leq \epsilon\}$ and $\forall F \subseteq LA$ $m_x(F) = \delta(\{\epsilon : \mathcal{D}_x^\epsilon = F\})$[1]

Appropriateness measures can then be evaluated according to definition 4. Alternatively, we can define a random set neighbourhood for each expression $\theta \in LE$ corresponding to those elements of Ω which can be appropriately described as θ, and then define $\mu_\theta(x)$ as the single point coverage function of this random set as follows:

Definition 6. *Random Set Neighbourhood of an Expression*
For $\theta \in LE$ and $\epsilon \in [0, \infty)$, $\mathcal{N}_\theta^\epsilon \subseteq \Omega$ is defined recursively as follows: $\forall L_i \in LA$, $\forall \theta, \varphi \in LE$

- $\mathcal{N}_{L_i}^\epsilon = \{x \in \Omega : d(x, P_i) \leq \epsilon\}$
- $\mathcal{N}_{\theta \wedge \varphi}^\epsilon = \mathcal{N}_\theta^\epsilon \cap \mathcal{N}_\varphi^\epsilon$
- $\mathcal{N}_{\theta \vee \varphi}^\epsilon = \mathcal{N}_\theta^\epsilon \cup \mathcal{N}_\varphi^\epsilon$
- $\mathcal{N}_{\neg\theta}^\epsilon = (\mathcal{N}_\theta^\epsilon)^c$

Theorem 1. *Random Neighbourhood Representation Theorem [7]*

$$\forall \theta \in LE,\ \forall x \in \Omega\ \mu_\theta(x) = \delta(\{\epsilon : x \in \mathcal{N}_\theta^\epsilon\})$$

Proof. Initially we show by induction that $\forall \theta \in LE$, $\forall \epsilon \geq 0$ $\mathcal{N}_\theta^\epsilon = \{x : \mathcal{D}_x^\epsilon \in \lambda(\theta)\}$. Let $LE^{(1)} = LA$ and for $k > 1$ $LE^{(k)} = LE^{(k-1)} \cup \{\theta \wedge \varphi, \theta \vee \varphi, \neg\theta : \theta, \varphi \in LE^{(k-1)}\}$. We now proceed by induction on k.
Limit Case: $k = 1$ For $L_i \in LA$ we have by definition 6 that $\mathcal{N}_{L_i}^\epsilon = \{x : d(x, P_i) \leq \epsilon\} = \{x : L_i \in \mathcal{D}_x^\epsilon\} = \{x : \mathcal{D}_x^\epsilon \in \lambda(L_i)\}$ by definition 3.
Inductive Step: Assume true for k For $\Psi \in$

[1] For Lesbegue measurable set $I \subseteq [0, \infty)$, we denote $\delta(I) = \int_I \delta(\epsilon) d\epsilon$ i.e. we also use δ to denote the probability measure induced by density function δ.

$LE^{(k+1)}$ either $\Psi \in LE^{(k)}$, in which case the result holds trivially by the inductive hypothesis, or one of the following holds for $\theta, \varphi \in LE^{(k)}$:

- $\Psi = \theta \wedge \varphi$ so that $\mathcal{N}^{\epsilon}_{\Psi} = \mathcal{N}^{\epsilon}_{\theta\wedge\varphi} = \mathcal{N}^{\epsilon}_{\theta} \cap \mathcal{N}^{\epsilon}_{\varphi}$ (by definition 6) $= \{x : \mathcal{D}^{\epsilon}_x \in \lambda(\theta)\} \cap \{x : \mathcal{D}^{\epsilon}_x \in \lambda(\varphi)\}$ (by the inductive hypothesis) $= \{x : \mathcal{D}^{\epsilon}_x \in \lambda(\theta) \cap \lambda(\varphi)\} = \{x : \mathcal{D}^{\epsilon}_x \in \lambda(\theta \wedge \varphi)\}$ (by definition 3).
- $\Psi = \theta \vee \varphi$ so that $\mathcal{N}^{\epsilon}_{\Psi} = \mathcal{N}^{\epsilon}_{\theta\vee\varphi} = \mathcal{N}^{\epsilon}_{\theta} \cup \mathcal{N}^{\epsilon}_{\varphi}$ (by definition 6) $= \{x : \mathcal{D}^{\epsilon}_x \in \lambda(\theta)\} \cup \{x : \mathcal{D}^{\epsilon}_x \in \lambda(\varphi)\}$ (by the inductive hypothesis) $= \{x : \mathcal{D}^{\epsilon}_x \in \lambda(\theta) \cup \lambda(\varphi)\} = \{x : \mathcal{D}^{\epsilon}_x \in \lambda(\theta \vee \varphi)\}$ (by definition 3).
- $\Psi = \neg\theta$ so that $\mathcal{N}^{\epsilon}_{\Psi} = \mathcal{N}^{\epsilon}_{\neg\theta} = (\mathcal{N}^{\epsilon}_{\theta})^c$ (by definition 6) $= \{x : \mathcal{D}^{\epsilon}_x \in \lambda(\theta)\}^c$ (by the inductive hypothesis) $= \{x : \mathcal{D}^{\epsilon}_x \notin \lambda(\theta)\} = \{x : \mathcal{D}^{\epsilon}_x \in \lambda(\theta)^c\} = \{x : \mathcal{D}^{\epsilon}_x \in \lambda(\neg\theta)\}$ (by definition 3).

Now by definition 4 we have that $\forall\theta \in LE\ \mu_\theta(x) = \sum_{F\in\lambda(\theta)} m_x(F) = \sum_{F\in\lambda(\theta)} \delta(\{\epsilon : \mathcal{D}^{\epsilon}_x = F\})$ (by definition 5) $= \delta(\{\epsilon : \mathcal{D}^{\epsilon}_x \in \lambda(\theta)\}) = \delta(\{\epsilon : x \in \mathcal{N}^{\epsilon}_{\theta}\})$ (by above). □

For example, for $L_i \in LA\ \mathcal{N}^{\epsilon}_{L_i} = \{x : d(x, P_i) \leq \epsilon\}$. Hence, $\mu_{L_i}(x) = \Delta(d(x, P_i))$ where $\Delta(\epsilon) = \delta([\epsilon, \infty))$.

Theorem 1 shows a clear link between appropriateness measures and Goodman and Nguyen's characterisation of fuzzy set membership functions as single point coverage functions of random sets [1], [2], [8].

Theorem 2. *Restricted Consonance [7]*
Let $LE^{\wedge,\vee}$ be those expressions in LE which can be generated from LA using only the connectives $\wedge$ and $\vee$. Then $\forall\theta \in LE^{\wedge,\vee}$, $\forall 0 \leq \epsilon \leq \epsilon'\ \mathcal{N}^{\epsilon}_{\theta} \subseteq \mathcal{N}^{\epsilon'}_{\theta}$

Proof. Let $LE^{\wedge,\vee,(1)} = LA$ and for $k > 1$ let $LE^{\wedge,\vee,(k)} = LE^{\wedge,\vee,(k-1)} \cup \{\theta \wedge \varphi, \theta \vee \varphi : \theta, \varphi \in LE^{\wedge,\vee,(k-1)}\}$. We now proceed by induction on k.
Limit Case: $k = 1$ For $L_i \in LA$, since $\epsilon' \geq \epsilon$ then trivially $\mathcal{N}^{\epsilon}_{L_i} = \{x : d(x, P_i) \leq \epsilon\} \subseteq \{x : d(x, P_i) \leq \epsilon'\} = \mathcal{N}^{\epsilon'}_{L_i}$
Inductive Step: Assume true for k For $\Psi \in LE^{(k+1)}$ either $\Psi \in LE^{(k)}$, in which case the result holds trivially by the inductive hypothesis, or one of the following holds for $\theta, \varphi \in LE^{(k)}$:

- $\Psi = \theta \wedge \varphi$: In this case $\mathcal{N}^{\epsilon}_{\Psi} = \mathcal{N}^{\epsilon}_{\theta\wedge\varphi} = \mathcal{N}^{\epsilon}_{\theta} \cap \mathcal{N}^{\epsilon}_{\varphi}$ (by definition 6) $\subseteq \mathcal{N}^{\epsilon'}_{\theta} \cap \mathcal{N}^{\epsilon'}_{\varphi}$ (by the inductive hypothesis) $= \mathcal{N}^{\epsilon'}_{\Psi}$ (by definition 6).
- $\Psi = \theta \vee \varphi$: In this case $\mathcal{N}^{\epsilon}_{\Psi} = \mathcal{N}^{\epsilon}_{\theta\vee\varphi} = \mathcal{N}^{\epsilon}_{\theta} \cup \mathcal{N}^{\epsilon}_{\varphi}$ (by definition 6) $\subseteq \mathcal{N}^{\epsilon'}_{\theta} \cup \mathcal{N}^{\epsilon'}_{\varphi}$ (by the inductive hypothesis) $= \mathcal{N}^{\epsilon'}_{\Psi}$ (by definition 6).

□

3 Imprecise Descriptions of Real Numbers

In this section we apply label semantics and prototype theory to model inference from imprecise descriptions of real numbers. Adopting random set neighbourhoods to represent extensions of concepts we will consider what imprecise probabilities result from conditioning given linguistic descriptions of a real variable. This approach is grounded in a clear interpretation of vague linguistic descriptions, in contrast to fuzzy methods in which membership functions and consequently probabilities of fuzzy events have no clear operational semantics [4].

Here we take $\Omega = \mathbb{R}$ and $d(x, y) = \|x - y\|$ and we consider descriptions based on number labels of the following form:

Definition 7. *Number Labels*
We consider a set LA of number labels L_i describing $\mathbb{R}$ with prototype sets P_i each corresponding to an interval of $\mathbb{R}$

The appropriateness measure for a number expressions $\theta \in LE$ (generated as in definition 1) is defined directly as the single point coverage function of $\mathcal{N}^{\epsilon}_{\theta}$ as in theorem 1. This allows us to relax the requirement in label semantics that LA is finite.

Here we particularly consider appropriateness measures generated by two types of density δ; normal distributions and uniform distributions.

Let $f(c, \sigma, \epsilon)$ denote the normal density function with mean c and standard deviation σ so that:

$$f(c, \sigma, \epsilon) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\epsilon-c)^2}{2\sigma^2}}$$

From this we can define a density δ as a normalised normal density of the form:

$$\delta(c, \sigma, \epsilon) = \frac{f(c, \sigma, \epsilon)}{1 - k} \text{ where } k = \int_{-\infty}^{0} f(c, \sigma, \epsilon) d\epsilon$$

From this we also have that:

$$\Delta(c, \sigma, \epsilon) = \frac{erfc(\frac{\epsilon-c}{\sigma\sqrt{2}})}{erfc(\frac{-c}{\sigma\sqrt{2}})}$$

where $erfc$ is the complementary error function

$$erfc(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt$$

Now let $\mu^{(c,\sigma)}_{\theta}(x)$ denote the appropriateness measure for θ generated by a normalised normal distribution δ with mean c and standard deviation σ. Figure 1 shows the appropriateness measure for a number label with prototypes $P_i = [5, 7]$ based on a normalised normal distribution with $c = 2$ and $\sigma = 1$.

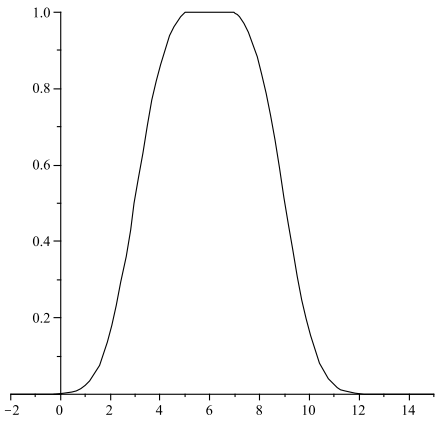


Figure 1: Appropriateness for label with prototypes $[5, 7]$ generated by a normalised normal distribution with mean $c = 2$ and standard deviation $\sigma = 1$

Theorem 3. *Let $c \leq c'$ then for $L_i \in LA$ it holds that:*

$$\forall x \in \mathbb{R}, \forall \sigma \in \mathbb{R}\ \mu_{L_i}^{(c,\sigma)}(x) \leq \mu_{L_i}^{(c',\sigma)}(x)$$

Proof. It is sufficient to show that $\Delta(c, \sigma, \epsilon)$ is an increasing function of c. Let $t = \frac{c}{\sigma\sqrt{2}}$ and $s = \frac{\epsilon}{\sigma\sqrt{2}}$ then:

$$\Delta(c, \sigma, \epsilon) = h(t, s) = \frac{erfc(s - t)}{erfc(-t)}$$

Hence it is sufficient to show that h is an increasing function of t.

$$\frac{\partial h}{\partial t} = \frac{2e^{-(s-t)^2}}{\sqrt{2}erfc(t)} + \frac{2erfc(s - t)e^{-t^2}}{erfc(t)^2\sqrt{\pi}} \geq 0$$

as required. □

Another interesting case is where δ is the uniform distribution on an interval $[k, r]$ for $r > k \geq 0$. This results in trapezoidal (or triangular) appropriateness measures. In this case we have:

$$\delta(k, r, \epsilon) = \begin{cases} 0 : \epsilon < k \\ \frac{1}{r-k} : \epsilon \in [k, r] \\ 0 : \epsilon > r \end{cases}$$

$$\text{and } \Delta(k, r, \epsilon) = \begin{cases} 1 : \epsilon < k \\ \frac{r-\epsilon}{r-k} : \epsilon \in [k, r] \\ 0 : \epsilon > r \end{cases}$$

Now let $\mu_\theta^{(k,r)}(x)$ denote the appropriateness measure for θ generated by a uniform distribution δ on $[k, r]$. Figure 2 shows the appropriateness for a number label with prototypes $[a, b]$ based on a uniform δ.

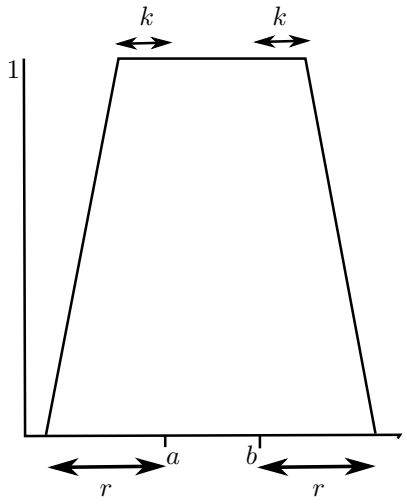


Figure 2: Appropriateness for label with prototypes $[a, b]$ generated by a uniform distribution δ on $[k, r]$

Theorem 4. *Let $0 \leq k < r$, $0 \leq k' < r'$, $k \leq k'$, and $r \leq r'$ then for $L_i \in LA$ it holds that:*

$$\forall x \in \mathbb{R}\ \mu_{L_i}^{(k,r)}(x) \leq \mu_{L_i}^{(k',r')}(x)$$

Proof. Trivially from the above it holds that $\forall \epsilon \geq 0$ $\Delta(k, r, \epsilon) \leq \Delta(k', r', \epsilon)$ and hence $\Delta(k, r, d(x, P_i)) \leq \Delta(k', r', d(x, P_i))$ as required. □

4 Information from Imprecise Descriptions

In this section we discuss the issue of conditioning given information in the form of imprecise descriptions of a real valued variable x. In other words, suppose we learn that 'x is *high*' or 'x is *high* $\wedge$ $\neg$*very high*' or more generally 'x is θ', what can we infer from such information about the value of x? To answer this question it is necessary to have a clear operational interpretation of statements 'x is θ'. For example, Zadeh [14] proposes that such statements define a possibility distribution on x when imprecise descriptions are represented by fuzzy sets. However, such a claim remains unconvincing while there is no clear operational meaning for fuzzy set membership functions. For the prototype model proposed in this paper a statement 'x is θ' is clearly interpreted as $x \in \mathcal{N}_\theta^\epsilon$. In other words, an imprecise description of x

restricts x to a random set neighbourhood generated by that description. Consequently, given the information 'x is θ' the remaining uncertainty concerning the value of x has two distinct sources. Firstly, for a specific value of ϵ, $\mathcal{N}_\theta^\epsilon$ is imprecise in the sense that typically it is a union of intervals of $\mathbb{R}$ rather than a precise value. Secondly, the value of the threshold ϵ is uncertain resulting in uncertainty about the definition of $\mathcal{N}_\theta^\epsilon$. Here, we shall argue that in the absence of any further information about x these two sources of uncertainty naturally result in lower and upper cumulative distributions. Furthermore, in the presence of a known prior probability distribution on x, conditioning on 'x is θ' results in a second order probability distribution on the cumulative probabilities for x.

Definition 8. *Upper and Lower Distributions*
Given a real valued random variable x for which we know only that 'x is θ' for some $\theta \in LE$ we define upper and lower cumulative distribution functions for the probability that $x \leq y$ as follows:

$$\underline{F}(y|\theta) = \delta_\theta(\{\epsilon : \mathcal{N}_\theta^\epsilon \subseteq (-\infty, y]\}) \text{ and}$$
$$\overline{F}(y|\theta) = \delta_\theta(\{\epsilon : \mathcal{N}_\theta^\epsilon \cap (-\infty, y] \neq \emptyset\}) \text{ and where}$$
$$\delta_\theta(\epsilon) = \begin{cases} \frac{\delta(\epsilon)}{\int_{\epsilon:\mathcal{N}_\theta^\epsilon \neq \emptyset} \delta(\epsilon)d\epsilon} : \mathcal{N}_\theta^\epsilon \neq \emptyset \\ 0 : \text{otherwise} \end{cases}$$

In definition 8 δ_θ is the posterior density on ϵ resulting from updating δ based on the information that $\mathcal{N}_\theta^\epsilon \neq \emptyset$. A possible justification for this normalisation of δ is that if we learn 'x is θ' this would intuitively imply that $\mathcal{N}_\theta^\epsilon \neq \emptyset$ since otherwise our information would be contradictory. In other words, accepting the assertion 'x is θ' implicitly implies accepting that the threshold ϵ must be such that $\mathcal{N}_\theta^\epsilon \neq \emptyset$. Clearly such conditioning is only possible if $\delta(\{\epsilon : \mathcal{N}_\theta^\epsilon \neq \emptyset\}) > 0$ otherwise the lower and upper probabilities given in definition 8 are undefined.

Theorem 5. *For $\theta \in LE^{\wedge,\vee}$ then $\forall y \in \mathbb{R}$*

$$\overline{F}(y|\theta) = \sup\{w\mu_\theta(x) : x \leq y\}$$
$$\underline{F}(y|\theta) = 1 - \sup\{w\mu_\theta(x) : x > y\}$$
$$\text{where } w = \frac{1}{\int_{\epsilon:\mathcal{N}_\theta^\epsilon \neq \emptyset} \delta(\epsilon)d\epsilon}$$

Proof. Straightforward from theorem 2 and definition 8 □

Theorem 5 shows that for θ not involving negation $\underline{F}(y|\theta)$ and $\overline{F}(y|\theta)$ are necessity and possibility measures respectively generated by the normalised possibility distribution $w\mu_\theta(x)$.

Corollary 1. *Let $\overline{F}^{(c,\sigma)}$ and $\underline{F}^{(c,\sigma)}$ be the upper and lower cumulative distributions as given in definition 8 and where δ is the normalised normal distribution with parameters c and σ. Then $\forall L_i \in LA$, $\forall c \leq c'$, $\forall \sigma \in \mathbb{R}$, $\forall y \in \mathbb{R}$*

$$\underline{F}^{(c',\sigma)}(y|L_i) \leq \underline{F}^{(c,\sigma)}(y|L_i) \text{ and}$$
$$\overline{F}^{(c,\sigma)}(y|L_i) \leq \overline{F}^{(c',\sigma)}(y|L_i)$$

Proof. Straightforward from theorems 5 and 3 □

Corollary 2. *Let $\overline{F}^{(k,r)}$ and $\underline{F}^{(k,r)}$ be the upper and lower cumulative distributions as given in definition 8 and where δ is the a uniform distribution on $[k, r]$. Then $\forall L_i \in LA$, $0 \leq k < r$, $0 \leq k' < r'$, $k \leq k'$, and $r \leq r'$, $\forall y \in \mathbb{R}$*

$$\underline{F}^{(k',r')}(y|L_i) \leq \underline{F}^{(k,r)}(y|L_i) \text{ and}$$
$$\overline{F}^{(k,r)}(y|L_i) \leq \overline{F}^{(k',r')}(y|L_i)$$

Proof. Straightforward from theorems 5 and 4 □

Now suppose we have prior information that x is distributed according to density function $p(x)$. In this case if we learn 'x is θ' then we should generate a posterior distribution by updating $p(x)$ given the new constraint that $x \in \mathcal{N}_\theta^\epsilon$. Let $F(y|\mathcal{N}_\theta^\epsilon)$ denote the corresponding updated cumulative distribution. However, the values of $F(y|\mathcal{N}_\theta^\epsilon)$ are uncertain given the remaining uncertainty about the value of the threshold ϵ. Hence, updating a prior distribution on x given an imprecise description of x results in a second order probability distribution as follows:

Definition 9. *Second Order Distribution*
Given a prior density $p(x)$ for x we define a second order cumulative distribution on the cumulative probability that $x \leq y$ as follows: $\forall p \in [0, 1]$

$$\tilde{F}_{y,\theta}(p) = \delta_\theta(\{\epsilon : F(y|N_\theta^\epsilon) \leq p\}) \text{ where}$$
$$F(y|N_\theta^\epsilon) = \int_{-\infty}^{y} p(x|\mathcal{N}_\theta^\epsilon)dx \text{ and where}$$
$$p(x|\mathcal{N}_\theta^\epsilon) = \begin{cases} \frac{p(x)}{\int_{\mathcal{N}_\theta^\epsilon} p(x)dx} : x \in \mathcal{N}_\theta^\epsilon \\ 0 : \text{otherwise} \end{cases}$$

If a precise posterior distribution is required conditional on θ, then one possibility is to take the expected value of posterior distributions given $\mathcal{N}_\theta^\epsilon$, as ϵ varies.

Definition 10. *Expected Density*
Given prior density $p(x)$ for x we can define an expected density for x conditional on θ by taking the

expected value of $p(x|\mathcal{N}_\theta^\epsilon)$ *as* ϵ *varies:*

$$p(x|\theta) = E_{\delta_\theta}(p(x|\mathcal{N}_\theta^\epsilon))$$

Notice that the above is a clearly motivated definition of conditional probability given imprecise linguistic information, consistent with a random set and prototype theory view of vague concepts. This is a distinct advantage over earlier work on the probability of fuzzy events [13], in which definitions do not appear to be linked to any underlying interpretation of fuzziness.

The following theorem shows that the expected cumulative distribution obtained from definition 10 is consistent with the lower and upper distributions given in definition 8.

Theorem 6. *For* $y \in \mathbb{R}$ *and* $\theta \in LE$, $\underline{F}(y|\theta) \leq F(y|\theta) \leq \overline{F}(y|\theta)$ *where* $F(y|\theta) = \int_{-\infty}^{y} p(x|\theta)dx = E_{\delta_\theta}(F(y|\mathcal{N}_\theta^\epsilon))$.

Proof.

$$\begin{aligned} F(y|\theta) = \int_{-\infty}^{y} p(x|\theta)dx &= \int_{-\infty}^{y}\int_{0}^{\infty} p(x|\mathcal{N}_\theta^\epsilon)\delta_\theta(\epsilon)d\epsilon dx \\ &= \int_0^\infty F(y|\mathcal{N}_\theta^\epsilon)\delta_\theta(\epsilon)d\epsilon \\ &= \int_{\epsilon:\mathcal{N}_\theta^\epsilon \cap (-\infty,y]\neq\emptyset} F(y|\mathcal{N}_\theta^\epsilon)\delta_\theta(\epsilon)d\epsilon \\ &\leq \int_{\epsilon:\mathcal{N}_\theta^\epsilon \cap (-\infty,y]\neq\emptyset} \delta_\theta(\epsilon)d\epsilon = \overline{F}(y|\theta) \end{aligned}$$

Alternatively

$$\begin{aligned} F(y|\theta) &= \int_{\epsilon:\mathcal{N}_\theta^\epsilon \subseteq (-\infty,y]} F(y|\mathcal{N}_\theta^\epsilon)\delta_\theta(\epsilon)d\epsilon + \\ &\quad \int_{\epsilon:\mathcal{N}_\theta^\epsilon \not\subseteq (-\infty,y]} F(y|\mathcal{N}_\theta^\epsilon)\delta_\theta(\epsilon)d\epsilon \\ &= \int_{\epsilon:\mathcal{N}_\theta^\epsilon \subseteq (-\infty,y]} \delta_\theta(\epsilon)d\epsilon + \int_{\epsilon:\mathcal{N}_\theta^\epsilon \not\subseteq (-\infty,y]} F(y|\mathcal{N}_\theta^\epsilon)\delta_\theta(\epsilon)d\epsilon \\ &\geq \int_{\epsilon:\mathcal{N}_\theta^\epsilon \subseteq (-\infty,y]} \delta_\theta(\epsilon)d\epsilon = \underline{F}(y|\theta) \end{aligned}$$

□

Example 1. *Consider the number label* L_i = *about* 2 *for which* $P_i = \{2\}$. *Let* $\delta(\epsilon) = \begin{cases} 1 : \epsilon \in [0,1] \\ 0 : otherwise \end{cases}$ *then the lower and upper cumulative distributions given the information 'x is about 2' are as follows:*

$$\underline{F}(y|L_i) = \begin{cases} 0 : y \leq 2 \\ 1 - \mu_{L_i}(y) : y > 2 \end{cases} \quad and$$

$$\overline{F}(y|L_i) = \begin{cases} 0 : y \leq 1 \\ \mu_{L_i}(y) : 1 < y \leq 2 \\ 1 : y > 1 \end{cases} \quad and\ where$$

$$\mu_{L_i}(y) = \begin{cases} 0 : x < 1 \\ x - 1 : x \in [1,2] \\ 3 - x : x \in (2,3] \\ 0 : x > 3 \end{cases}$$

Suppose we now further learn that x *is distributed according to a uniform distribution on* $[0, 10]$ *then we can infer a second order distribution the probability that* $x \leq y$ *as follows: Initially note that* $\mathcal{N}_{L_i}^\epsilon = [2 - \epsilon, 2 + \epsilon]$ *so that for* $\epsilon \leq 1$

$$p(x|\mathcal{N}_{L_i}^\epsilon) = \begin{cases} \frac{1}{2\epsilon} : x \in [2-\epsilon, 2+\epsilon] \\ 0 : otherwise \end{cases} \quad and\ hence$$

$$F(y|\mathcal{N}_{L_i}^\epsilon) = \begin{cases} 1 : y > 2 + \epsilon \\ \frac{y+\epsilon-2}{2\epsilon} : y \in [2-\epsilon, 2+\epsilon] \\ 0 : y < 2 - \epsilon \end{cases}$$

From this we obtain four cases of $\tilde{F}_{y,L_i}$ *as follows:*

For $y < 1$

$$\forall p \in [0,1]\ \tilde{F}_{y,L_i}(p) = 1$$

For $1 \leq y \leq 2$ *(see figure 3)*

$$\tilde{F}_{y,L_i}(p) = \begin{cases} 1 : p > \frac{y-1}{2} \\ \frac{2-y}{1-2p} : p \leq \frac{y-1}{2} \end{cases}$$

For $2 < y \leq 3$ *(see figure 4)*

$$\tilde{F}_{y,L_i}(p) = \begin{cases} 0 : p < \frac{y-1}{2} \\ \frac{2p-y+1}{2p-1} : \frac{y-1}{2} \leq p < 1 \\ 1 : p = 1 \end{cases}$$

For $y > 3$

$$\tilde{F}_{y,L_i}(p) = \begin{cases} 0 : p < 1 \\ 1 : p = 1 \end{cases}$$

The expected density $p(x|L_i)$ *is given by (figure 5):*

$$p(x|L_i) = \begin{cases} 0 : x < 1 \\ -\frac{1}{2}\ln(2-x) : 1 \leq x < 2 \\ -\frac{1}{2}\ln(x-2) : 2 < x \leq 3 \\ 0 : x > 3 \end{cases}$$

Figure 6 shows the upper and lower cumulative distributions given 'x is about 2' together with the expected cumulative distribution assuming that x *is distributed according to a uniform distribution on* $[0, 10]$.

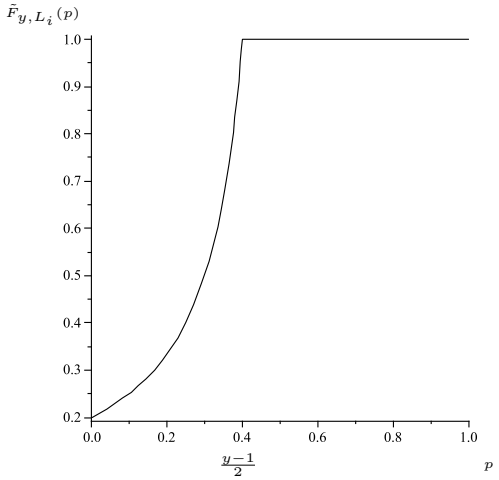


Figure 3: $\tilde{F}_{y,L_i}$ for $1 < y \leq 2$

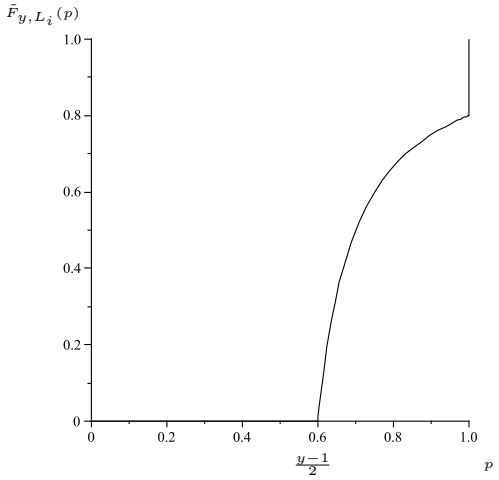


Figure 4: $\tilde{F}_{y,L_i}$ for $2 < y \leq 3$

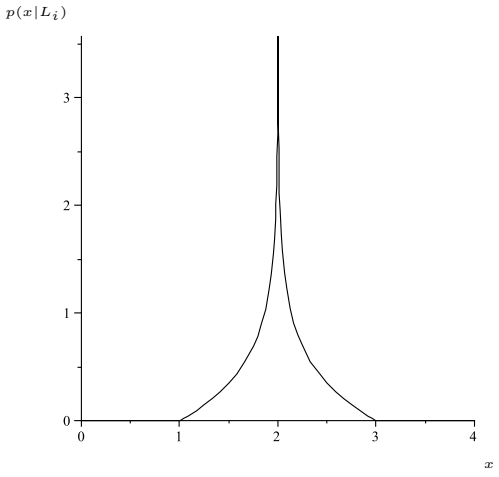


Figure 5: Expected density for x conditional on L_i

5 Information from Imprecise Data

In this section we consider inference on the basis of data taking the form of imprecise descriptions of real values. Let x be a real valued random variable, and let $DB = \{\theta_1, \ldots, \theta_N\}$ where $\theta_i \in LE$ be a set of independently generated descriptions of x. Given DB we define lower and upper cumulative distributions for x as follows:

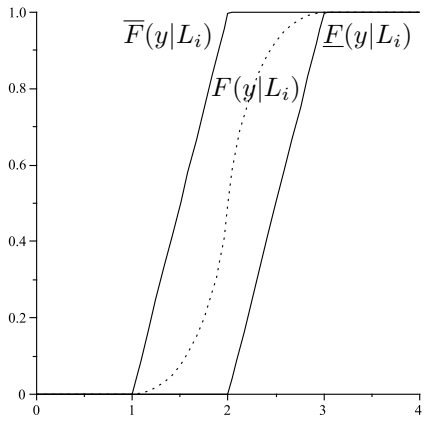


Figure 6: Lower and upper cumulative distribution $\underline{F}(y|L_i)$, $\overline{F}(y|L_i)$ together with expected cumulative distribution $F(y|L_i)$ assuming a uniform prior on x

Definition 11.

$$\forall y \in \mathbb{R}\ \underline{F}(y|DB) = \frac{1}{N}\sum_{i=1}^{N} \underline{F}(y|\theta_i)$$

$$\overline{F}(y|DB) == \frac{1}{N}\sum_{i=1}^{N} \overline{F}(y|\theta_i)$$

Definition 12. *Given a density $p(x)$ we can define a expected density cumulative distribution conditional on DB according to:*

$$\forall x \in \mathbb{R}\ p(x|DB) = \frac{1}{N}\sum_{i=1}^{N} p(x|\theta_i) \text{ and}$$

$$\forall y \in \mathbb{R}\ F(y|DB) = \frac{1}{N}\sum_{i=1}^{N} F(y|\theta_i)$$

The underlying intuition behind these definitions is as follows: In order to estimate x one approach would be to randomly select a description θ_i from DB and then condition on the information 'x is θ_i'. Assuming that each element of DB is equally likely to be selected (i.e. has equal weighting) then the expected information we would learn about x is as given in definitions 11 and 12.

One natural example of this approach is where we have an independent sample $\{x_1, \ldots, x_N\}$ of values of x for which we are assuming there is an associated

uncertain error ϵ with density δ, so that each x_i effectively identifies a random set interval $[x_i - \epsilon, x_i + \epsilon]$. In this case we define $DB = \{L_1, \ldots, L_N\}$ where L_i is a number label with prototype $P_i = \{x_i\}$ (i.e. $L_i = about\ x_i$).

Example 2. *A sample of* 100 *values was drawn at random from the normal mixture distribution* $g = \frac{N(2,3)+N(8,0.5)}{2}$. *DB was then taken to correspond to the set of labels* L_i *with prototype* $P_i = \{x_i\}$ *for each value* x_i *in the sample.* δ *was assumed to be a uniform distribution on* $[k, r]$ *where* k *and* r *are effectively treated as parameters in the estimating of distributions from DB.*

To compare the upper and lower cumulative distributions obtained from DB with that of the generating distribution g we introduce two measure as follows:

$$IE := \frac{1}{N}\sum_{i=1}^{N} \chi_{[\underline{F}(x_i|DB),\overline{F}(x_i|DB)]}(G(x_i))$$

where $\chi_{[\underline{F}(x_i|DB),\overline{F}(x_i|DB)]}$ *is the characteristic function for the interval* $[\underline{F}(x_i|DB), \overline{F}(x_i|DB)]$ *and* G *is the cumulative distribution function for density* g. *Hence,* IE *provides a measure of the extent to which the generating cumulative density* G *is contained within the estimated upper and lower envelope across the original sample.*

We also evaluate the average range of the upper and lower distribution envelope according to:

$$Range = \frac{1}{N}\sum_{i=1}^{N}(\overline{F}(x_i|DB) - \underline{F}(x_i|DB))$$

Table 1 shows the IE *and Range values for a number of different* k, r *values. Notice that by corollary 2 it follows immediately that as* k *and* r *increase the* IE *values decrease. Figure 7 shows the upper and lower envelope together with* G *for* $k = 0$ *and* $r = 2.2$, *these corresponding to the values in table 1 for which* IE *is* 0 *and Range is minimal.*

Table 2 compares $p(x|DB)$ *with* $g(x)$ *according to* MSE *defined as follows:*

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(p(x_i|DB) - g(x_i))^2$$

Figure 8 shows $p(x|DB)$ *and* $g(x)$ *for* $k = 0.4$, $r = 0.5$ *these corresponding to the values in table 2 with lowest* MSE.

6 Summary and Conclusions

The prototype theory interpretation of label semantics has been introduced as a possible model for imprecise descriptions of real numbers. Based on this interpretation it has been shown that conditioning given information in the form 'x is θ', for $\theta \in LE$, naturally results in imprecise probabilities. Also, within this framework, we have proposed a possible approach to inference from data in the form of imprecise descrip-

Table 1: Table showing IE and $Range$ for different values of k and r

k	r	IE	$Range$
0.7	1.1	0.1	0.2926
0.8	1.1	0.09	0.3036
0.9	1.1	0.07	0.31324
1	1.1	0.03	0.3221
0.8	1.2	0.04	0.3122
0.9	1.2	0.02	0.3213
1	1.2	0.01	0.3298
1.1	1.2	0	0.3584
0.9	1.3	0.01	0.3286
1	1.3	0	0.3367
0.9	1.4	0.01	0.3286
0.7	1.4	0.02	0.3171
0.8	1.4	0	0.3245
0.6	1.5	0.02	0.31413
0.7	1.5	0	0.3245
0.6	1.6	0	0.3212
0.5	1.7	0	0.3176
0.4	1.8	0	0.3135
0.3	1.9	0	0.3087
0.2	2	0	0.3034
0.1	2.1	0	0.2979
0	2.2	0	0.2920

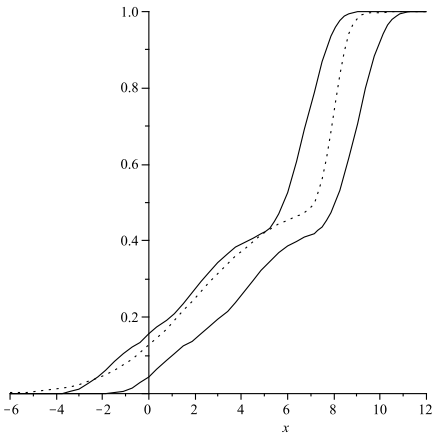


Figure 7: Upper and lower cumulative distributions based on uniform δ with $k = 0$ and $r = 2.2$, compared with cumulative distribution for the generating distribution g (dashed line)

Table 2: Table showing MSE for different values of k and r

k	r	MSE
0.1	0.2	0.0039
0.1	0.3	0.00229
0.2	0.3	0.001689
0.1	0.4	0.001596
0.2	0.4	0.001224
0.3	0.4	0.001005
0.1	0.5	0.00119
0.2	0.5	0.000929
0.3	0.5	0.00077
0.4	0.5	0.000698
0.1	0.6	0.000968
0.2	0.6	0.000817
0.3	0.6	0.000733
0.4	0.6	0.0007535
0.5	0.6	0.000925

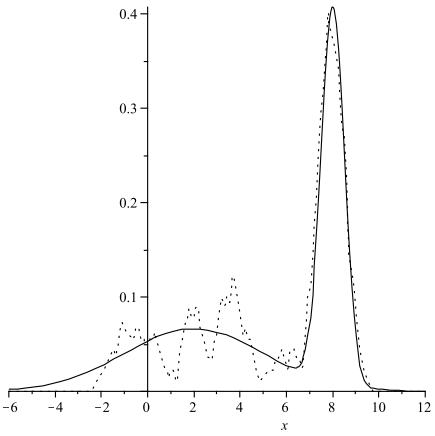


Figure 8: Density estimate based on uniform δ with $k = 0.4$ and $r = 0.5$ (dashed line), compared with generating distribution g

tions of a real variable. This naturally suggests an algorithm of estimating distributions given imprecise data values.

Acknowledgements

Inés González-Rodríguez is supported by MEC-Programa José Castillejo Grant JC2007-00152 and MEC-FEDER Grant TIN2007-67466-C02-01.
Yongchuan Tang is funded by the National Natural Science Foundation of China (NSFC) under Grant 60604034, the joint funding of NSFC and MSRA under Grant 60776798, and the Science and Technology Program of Zhejiang Province under Grant 2007C223061

References

[1] I.R. Goodman, H.T. Nguyen, (1985), *Uncertainty Model for Knowledge Based Systems*, North Holland.

[2] I.R. Goodman, (1982), 'Fuzzy Sets as Equivalence Classes of Random Sets', in *Fuzzy Set and Possibility Theory* (ed. R. Yager), pp327-342

[3] J. Lawry, (2004), 'A Framework for Linguistic Modelling', *Artificial Intelligence*, Vol. 155, pp1-39.

[4] J. Lawry, (2006), *Modelling and Reasoning with Vague Concepts*, Springer

[5] J. Lawry, (2008), 'Appropriateness Measures: An Uncertainty Model for Vague Concepts', *Synthese*, Vol. 161, No.2, pp255-269

[6] J. Lawry, Y. Tang, (2008), 'Relating Prototype Theory and Label Semantics', in *Soft Methods for Handling Variability and Imprecision*, (Eds. D. Dubois, M.A. Lubiano, H. Prade, M.A. Gil, P. Grzegorzewski, O. Hryniewicz), pp35-42

[7] J. Lawry, Y. Tang, (2009), 'Uncertainty Modelling for Vague Concepts: A Prototype Theory Approach', *submitted*

[8] H.T. Nguyen, (1984), 'On Modeling of Linguistic Information using Random Sets', *Information Science*, Vol. 34, pp265-274

[9] E.H. Rosch, (1975), 'Cognitive Representation of Semantic Categories', *Journal of Experimental Psychology: General*, Vol. 104, pp192-233

[10] Y. Tang, J. Lawry, (2009), 'Linguistic Modelling and Information Coarsening based on Prototype Theory and Label Semantics', *International Journal of Approximate Reasoning*, in Press

[11] P. Walley, G. de Cooman, (2001), 'A Behavioural Model of Linguistic Uncertainty', *Information Sciences*, Vol. 134, pp1-37

[12] T. Williamson, (1994), *Vagueness*, Routledge.

[13] L.A. Zadeh, (1968), 'Probability Measures of Fuzzy Events' *Journal of Mathematical Analysis and Applications* Vol. 23, pp421-427

[14] L.A. Zadeh, (1978), 'Fuzzy Sets as a Basis for a Theory of Possibility', *Fuzzy Sets and Systems*, Vol. 1, pp3-28

Reasoning with imprecise probabilistic knowledge on enzymes for rapid screening of potential substrates or inhibitor structures

Weiru Liu, Anbu Yue
School of Electronics, Electrical Engineering
and Computer Science,
Queen's University Belfast,
Belfast BT7 1NN, UK
{w.liu, a.yue}@qub.ac.uk

David J Timson
School of Biological Science
Queen's University Belfast,
Belfast BT9 7BL, UK
d.timson@qub.ac.uk

Abstract

In many applications, there is a need to model and reason with imprecise probabilistic knowledge. In this paper, we discuss how to model imprecise probabilistic knowledge obtained from experiments in biological sciences on enzymes for rapid screening of potential substrate or inhibitor structures. Each imprecise probabilistic knowledge base is modelled as a probabilistic logic program (PLP). To predict a meaningful substrate structure, we have developed a framework (and a tool) in which a user (bioscientist) can query against a PLP (or a collection of PLPs), can examine how relevant a PLP is for answering a query, and can select a query result that is more satisfactory. This framework is implemented by integrating an optimizer in MatLab to solve the optimization problems subject to linear constraints. A preliminary version of the tool was demonstrated in the ECAI08 Demo session. Experimental results on evaluating the tool with probabilistic knowledge on enzymes for rapid screening of potential substrates or inhibitor structures demonstrate that this tool has a great potential to be used in many similar areas for the initial screening of compound structures in drug discovery.

Keywords. Imprecise probabilistic knowledge, prediction, substrate structure, enzymes, rapid screening.

1 Introduction

Most of the knowledge that is used, for example, for advanced knowledge base systems or for cognitive modeling is uncertain, incomplete, imprecise and subject to changes. Very often, this uncertainty and incompleteness is characterized by probabilities, especially, when the knowledge concerned is elicited from experiments. Therefore, there is a need to develop adequate theories and frameworks to model and reason with such probabilistic knowledge.

Probabilistic logic programming is a framework to represent and reason with imprecise (conditional) probabilistic knowledge. An agent's knowledge is represented by a *probabilistic logic program* (PLP) which is a set of (conditional) logical formulae with probability intervals. The impreciseness of the agent's knowledge is explicitly represented by assigning a probability interval (or a single probability) to every logical formula indicating that the probability of the formula shall be in the given interval. Probabilistic logic programming has been used to represent and reason with probabilistic knowledge in many real world applications, e.g., [2, 5, 9]. Among various types of probabilistic logic programming, conditional probabilistic logic programming (PLP for short) [7, 8] is particularly tailored to represent conditional events with probabilities of the form $(C|A_1 \wedge ... \wedge A_n)[l, u]$ where A_is are conditions, C is a conclusion. $(C|A_1 \wedge ... \wedge A_n)[l, u]$ is interpreted as the probability of conditional event $C|A_1 \wedge ... \wedge A_n$ falling in interval $[l, u]$.

To illustrate the use of conditional probabilistic logic programming, let us consider medical treatments for certain medical conditions (diseases), such as a patient is diagnosed with liver cancer. There are various types of treatments for cancers, such as, (A) surgery only to remove the organ; (B) surgery plus Radiotherapy; (C) Radiotherapy only, depending on the stage of the cancer, the health condition of the patient and possibly other factors. Then statistical summaries from clinical trials studied on the relationship between mortality and treatments can be represented as conditional events shown below.

$$Mortality(X, Year10)|LiverCancer(X, Year0) \wedge CancerStage(X, early) \wedge Surgery(X, Year0) \wedge RT(X; Year0)[0.223; 0.225]$$

This piece of imprecise probabilistic knowledge says that from this trial (or the meta-analysis of many trials), the probability of a patient's 10-year mortality, given that the patient is in his/her early liver cancer stage, undergoing a surgery plus Radiotherapy, is in between 0.223 to 0.225.

Conditional events like above cannot always be simply interpreted as cause-effect relationships. For the above example, it is not that the surgery and RT caused a patient to die in 10-years, rather, it says that if those actions are taken place (given that the patient's liver cancer stage is early), then what the probability of this patient being dead in 10-years could be. Of course, if no action was taken

place, the probability of the patient being dead in 10-years would be much greater than 22.5%. Therefore, conditional probabilistic logic programming offers a much more suitable framework for capturing imprecise probabilistic scientific knowledge of this kind than other approaches. Given a PLP and a query against the PLP, traditionally, a probability interval is returned as the answer. This interval implies that the true probability of the query shall be within the given interval. However, when this interval is too wide, it provides no useful information. For instance, if a PLP contains knowledge $\{(fly(X)|bird(X)[0.98, 1], (bird(X)|magpie(X))[1, 1]\}$, then the answer to the query *Can a magpie fly?* (i.e., $?(fly(t)|magpie(t))$) is a trivial bound $[0, 1]$.

One way to enhance the reasoning power of a PLP is to apply the maximum entropy principle [6]. From this principle, a single probability distribution is selected and it is assumed to be the most acceptable one for the query among all possible probability distributions. As a consequence, a precise probability is given for a query even when the agent's original knowledge is imprecise. In the above example, by applying the maximum entropy principle, 0.98 is returned as the answer for the query. Intuitively, accepting such a precise probability from (a prior) imprecise knowledge can be risky. When an agent's knowledge is rich enough then a single probability could be reliable, however, when an agent's knowledge is (very) imprecise, an interval is more appropriate than a single probability.

Therefore, how useful a probabilistic logic program (PLP) is to answering a given query? This question is important in two fold: first, it helps to analyze if a PLP is adequate to answer a query and second, if a PLP is sufficiently relevant to a query, then shall a single probability be obtained or shall a probability interval be more suitable? To answer these questions, in [18], we proposed two concepts, *the measure of ignorance* and *the measure of degree of satisfaction*, w.r.t. a PLP and a query. The former analyzes the impreciseness of the PLP w.r.t. the query, and the latter measures which (tighter) interval is sufficiently informative to answer the query.

In this paper, we present our investigation about how to use PLPs to represent and reason with imprecise probabilistic knowledge obtained from experiments, especially on substrates prediction in biomedical sciences. We first discuss the importance of evaluating the relevance of a knowledge base w.r.t a query, focusing on how reliable a query result returned from querying a PLP could be, knowing that the knowledge contained in the PLP is imprecise. To quantitatively measure the reliability of a query result, we introduce our formal analysis of ignorance and degree of satisfaction about a query result obtained from the PLP [18]. We then present our implementation of a probabilistic querying system which takes PLPs as input knowledge bases and produces probabilistic results for queries (against a chosen PLP). The results are either in the form of pure probabilistic terms (an interval or a maximum entropy), or the maximum entropy plus its ignorance, or an interval plus its degree of satisfaction. The first form of output is the traditional type of output from probabilistic logic programming, whilst the latter two are our extensions – adding extra information about a query result to tell a user how reliable this result could be when using this particular knowledge base.

We apply our theory and system to enzymes for rapid prediction of potential substrate or inhibitor structures. We conducted two sets of experiments, one is on the human enzyme galactokinase, which uses galactose as a substrate, and the other is on substrate prediction for NQO1. The experimental results demonstrate that using imprecise probabilistic knowledge as a first step in screening for substrates can be very useful and significant in many similar applications, since this initial prediction could allow bioscientists to selectively experiment on more hopeful candidates, saving both time and money in the whole process.

This paper is organized as follows. In Section 2, we briefly review probabilistic logic programming. In Section 3, we describe how to analyze the quality of knowledge in a PLP and in Section 4 we introduce a general theory and an instantiation on measuring the ignorance and the degree of satisfaction w.r.t. a PLP and a query. In Section 5, we describe our system architecture and efficient implementation. In Section 6, we illustrate our framework with two case studies in bioscience. Finally, we conclude the paper in Section 7.

2 Preliminary

We briefly review conditional probabilistic logic programming here [7, 8].

We use Φ to denote the finite set of predicate symbols, $\mathcal{V}$ to denote the set of *object variables*, and $\mathcal{B}$ to denote the set of *bound constants* which describe the bound of probabilities, and bound constants are in [0,1]. We use $a, b, \ldots$ to denote constants from Φ and $X, Y \ldots$ to denote object variables from $\mathcal{V}$. An *object term* t is a constant from Φ or an object variable from $\mathcal{V}$. An *atom* is of the form $p(t_1, \ldots, t_k)$, where p is a predicate symbol and $t_1, \ldots, t_k$ are object terms. We use Greek letters $\phi, \varphi, \psi, \ldots$ to denote *events* (or *formulae*) which are obtained from atoms by logic connectives $\wedge, \vee, \neg$ as usual. A *conditional event* is of the form $(\psi|\phi)$ where ψ and ϕ are events, and φ is called the *antecedent* and ψ is called the *consequent*. A *probabilistic formula*, denoted as $(\psi|\varphi)[l, u]$, means that the probability of conditional event $\psi|\varphi$ is between l and u, where l, u are bound constants. A set of probabilistic formulae is called a *conditional probabilistic logic program (PLP)*, a PLP is denoted as P in the rest of the paper.

A *ground* term, (resp. event, conditional event, probabilis-

tic formula, or PLP) is a term, (resp. event, conditional event, probabilistic formula, or PLP) that does not contain any object variables in $\mathcal{V}$.

All the constants in Φ form the Herbrand universe, denoted as HU_Φ, and the Herbrand base, denoted as HB_Φ, is the finite nonempty set of all events constructed from the predicate symbols in Φ and constants in HU_Φ. A subset I of HB_Φ is called a *possible world* and $\mathcal{I}_\Phi$ is used to denote the set of all possible worlds over Φ. A function σ that maps each object variable to a constant is called an *assignment*. It is extended to object terms by $\sigma(c) = c$ for all constant symbols from Φ. An event φ satisfied by I under σ, denoted by $I \models_\sigma \varphi$, is defined inductively as:

- $I \models_\sigma p(t_1, \ldots, t_n)$ iff $p(\sigma(t_1), \ldots, \sigma(t_n)) \in I$;
- $I \models_\sigma \phi_1 \wedge \phi_2$ iff $I \models_\sigma \phi_1$ and $I \models_\sigma \phi_2$;
- $I \models_\sigma \phi_1 \vee \phi_2$ iff $I \models_\sigma \phi_1$ or $I \models_\sigma \phi_2$;
- $I \models_\sigma \neg\phi$ iff $I \not\models_\sigma \phi$

An event φ is satisfied by a possible world I, denoted by $I \models_{cl} \varphi$, iff $I \models_\sigma \varphi$ for all assignments σ. An event φ is a *logical consequence* of event ϕ, denoted as $\phi \models_{cl} \varphi$, iff all possible worlds that satisfy ϕ also satisfy φ.

In this paper, we use $\top$ to represent (ground) tautology, and we have that $I \models_{cl} \top$ for all I and all assignments σ. And we use $\bot$ to denote $\neg\top$.

If Pr is a function (or distribution) on $\mathcal{I}_\Phi$ (i.e., as $\mathcal{I}_\Phi$ is finite, Pr is a mapping from $\mathcal{I}_\Phi$ to the unit interval [0,1] such that $\sum_{I\in\mathcal{I}_\Phi} Pr(I) = 1$), then Pr is called a *probabilistic interpretation*. For an assignment σ, the probability assigned to an event φ by Pr, is denoted as $Pr_\sigma(\varphi)$ where $Pr_\sigma(\varphi) = \sum_{I\in\mathcal{I}_\Phi, I\models_\sigma\varphi} Pr(I)$. When φ is ground, we simply written it as $Pr(\varphi)$. When $Pr_\sigma(\phi) > 0$, the conditional probability, $Pr_\sigma(\psi|\phi)$, is defined as $Pr_\sigma(\psi|\phi) = Pr_\sigma(\psi \wedge \phi)/Pr_\sigma(\phi)$. When $Pr_\sigma(\phi) = 0$, $Pr_\sigma(\psi|\phi)$ is undefined. Also, when $(\psi|\phi)$ is ground, we simply write $Pr(\psi|\phi)$.

A probabilistic interpretation Pr satisfies or is a *probabilistic model* of a probabilistic formula $(\psi|\phi)[l,u]$ under assignment σ, denoted as $Pr \models_\sigma (\psi|\phi)[l,u]$, iff $l \leq Pr_\sigma(\psi|\phi) \leq u$ or $Pr_\sigma(\phi) = 0$. A probabilistic interpretation Pr satisfies or is a *probabilistic model* of a probabilistic formula $(\psi|\phi)[l,u]$ iff Pr satisfies $(\psi|\phi)[l,u]$ under all assignments. A probabilistic interpretation Pr satisfies or is a *probabilistic model* of a PLP P iff for all assignment σ, $\forall(\psi|\phi)[l,u] \in P, Pr \models_\sigma (\psi|\phi)[l,u]$. A probabilistic formula $(\psi|\varphi)[l,u]$ is a *consequence* of PLP P, denoted by $P \models (\psi|\varphi)[l,u]$, iff all probabilistic models of P satisfy $(\psi|\varphi)[l,u]$. A probabilistic formula $(\psi|\varphi)[l,u]$ is a *tight consequence* of P, denoted by $P \models_{tight} (\psi|\varphi)[l,u]$, iff $P \models (\psi|\varphi)[l,u]$, $P \not\models (\psi|\varphi)[l,u']$, $P \not\models (\psi|\varphi)[l',u]$ for all $l' > l$ and $u' < u$ ($l', u' \in [0,1]$). It is worth noting that if $P \models (\phi|\top)[0,0]$ then $P \models_{tight} (\psi|\phi)[1,0]$ where $[1,0]$ stand for the empty set.

A query is of the form $?(\psi|\phi)$ or $?(\psi|\phi)[l,u]$, where ψ and ϕ are ground events and $l, u \in [0,1]$. For query $?(\psi|\phi)$, by the tight consequence relation, a bound $[l,u]$ is given as the answer, such that $P \models_{tight} (\psi|\phi)[l,u]$. For query $?(\psi|\phi)[l,u]$, a bound $[l,u]$ is given by the user. A PLP returns *True (or Yes)* if $P \models (\psi|\phi)[l,u]$ and *False (or No)* if $P \not\models (\psi|\phi)[l,u]$ [8].

The principle of maximum entropy is a well known techniques to represent probabilistic knowledge. Entropy quantifies the indeterminateness inherent to a distribution Pr by $H(Pr) = -\Sigma_{I\in\mathcal{I}_\Phi} Pr(I)\log Pr(I)$. Given a logic program P, the *principle of maximum entropy model* (or *me-model*), denoted by $me[P]$, is defined as: $H(me[P]) = \max H(Pr) = \max_{Pr\models P} -\Sigma_{I\in\mathcal{I}_\Phi} Pr(I)\log Pr(I)$

$me[P]$ is the unique probabilistic interpretation Pr that is a probabilistic model of P and that has the greatest entropy among all the probabilistic models of P.

Let P be a ground PLP, we say that $(\psi|\varphi)[l,u]$ is a *me-consequence* of P, denoted by $P \models^{me} (\psi|\varphi)[l,u]$, iff P is unsatisfiable, or $me[P] \models (\psi|\varphi)[l,u]$.

We say that $(\psi|\varphi)[l,u]$ is a *tight me-consequence* of P, denoted by $P \models^{me}_{tight} (\psi|\varphi)[l,u]$, iff either P is unsatisfiable, $l = 1$, $u = 0$, or $P \models \bot \leftarrow \varphi$, $l = 1$, $u = 0$, or $me[P](\varphi) > 0$ and $me[P](\psi|\varphi) = l = u$.

3 A Formal Analysis of PLPs

In information theory, the information entropy is a measure of the uncertainty associated with a random variable. Entropy quantifies information in a piece of data. Informally, $-\log p(X = x_i)$ means the degree of surprise[1] when one observes that the random variable turns out to be x_i. In another word, $-\log p(X = x_i)$ reflects the information one receives from the observation. The entropy is an expectation of the information one may receive from a random domain by observing random events. Inspired by this, we define a knowledge entropy, which reflects how much an agent knows the truth value of ψ given ϕ prior any observations. Informally, more surprised an agent is by the observation, more knowledge it learns from the observation, and thus, less knowledge it has about ψ given ϕ before observing ψ or $\neg\psi$ given ϕ.

Definition 1 *Let P be a PLP, and $(\psi|\phi)$ be a conditional event. Suppose that Pr is a probabilistic model for P, then the* knowledge entropy *of inferring ψ from ϕ under Pr, denoted as $K_{Pr}(\psi|\phi)$, is defined as $K_{Pr}(\psi|\phi)$*

$$= 1 + \frac{1}{2}(Pr(\psi|\phi)\log Pr(\psi|\phi) + Pr(\neg\psi|\phi)\log Pr(\neg\psi|\phi))$$

It is obvious that $K_{Pr}(\psi|\phi) = K_{Pr}(\neg\psi|\phi)$ and $K_{Pr}(\psi|\phi) \in [0,1]$. Trivially, $K_{Pr}(\phi|\phi) = 1$ and $K_{Pr}(\neg\phi|\phi) = 1$, since from Pr, the truth values of an event and its negation are known, when the event is given.

[1] http://en.wikipedia.org/wiki/Self-information

By extending the above definition, we can define a knowledge measurement for a PLP.

Definition 2 *Let P be a PLP, and $(\psi|\phi)$ be a conditional event. Suppose that Pr is a probabilistic model for P and $Pr(\phi) > 0$, then the* knowledge measurement $K_P(\psi|\phi)$ *is defined by:*

$$\begin{array}{lcl} minK_P(\psi|\phi) & = & \min_{Pr\models P} K_{Pr}(\psi|\phi) \\ maxK_P(\psi|\phi) & = & \max_{Pr\models P} K_{Pr}(\psi|\phi) \\ K_P(\psi|\phi) & = & [minK_P(\psi|\phi), maxK_P(\psi|\phi)] \end{array}$$

The measurement $K_P(\psi|\phi)$ is used to characterize the usefulness of knowledge contained in PLP P for inferring ψ when knowing or observing ϕ. When ψ or $\neg\psi$ can be inferred from ϕ under P, P contains all the necessary knowledge for inferring ψ given ϕ, and so we have $minK_P(\psi|\phi) = 1$. When knowledge in P excludes the possibility that the probability of ψ (or $\neg\psi$) may be 1 given ϕ, i.e., $P \cup \{(\psi|\phi)[1,1]\}$ (or $P \cup \{(\psi|\phi)[0,0]\}$) is unsatisfiable, then the knowledge contained in P can not fully support ψ given ϕ, so $maxK_P(\psi|\phi) < 1$. Specifically, if it can not be inferred that ψ is more (or less) likely to be true than $\neg\psi$ (i.e. the probability of ψ given ϕ is bigger (or smaller) than $\neg\psi$ given ϕ), then $minK_P(\psi|\phi) = 0$.

We can define a partial order $\preceq$ over the set $\{[x,y]|x,y \in [0,1]\}$ as $[a,b] \preceq [c,d]$ iff $a \geq c, b \leq d$, and $[a,b] \prec [c,d]$ iff $[a,b] \preceq [c,d]$ and $a > c$ or $b < d$. We say a PLP P is more precise than P' about $\psi|\phi$, if $K_P(\psi|\phi) \preceq K_{P'}(\psi|\phi)$, denoted as $P \preceq^k_{(\psi|\phi)} P'$.

If $minK_P(\psi|\phi) \neq maxK_P(\psi|\phi)$ given P, then the knowledge contained in P is not sufficient to decide the probability of ψ given ϕ, that is, the knowledge contained in P about inferring ψ given ϕ is imprecise. In order to infer the actual probability of ψ given ϕ under P, we need more knowledge.

Proposition 1 *Let P and P' be two PLPs. If $P \models P'$ then $P \preceq^k_{(\psi|\phi)} P'$ for any conditional event $(\psi|\phi)$.*

This proposition suggests that the consequence relation $\models$ considers all statements in the PLP while the knowledge measurement focuses only on the knowledge about ψ given ϕ.

In the view of knowledge entropy, reasoning under the maximum entropy principle implicitly introduces some extra knowledge to enhance the reasoning power of PLP. We should be aware that although this assumption seems intuitive, it may be wrong, as shown below.

Example 1 *Let* $P_1 = \{(headUp(X)|toss(X))[0.5,0.5]\}$, $P_2 = \{(headUp(X)|toss(X))\ [0,1]\}$ *be two PLPs. Here, P_1 states that tossing a fair coin may result in head up with probability 0.5, however, in P_2, we do not know whether the coin is fair.*

In this example, the knowledge in P_1 is richer than that in P_2 since from P_1 we know the coin is fair. With the maximum entropy principle, we get that $P_1 \models^{me} (headUp(coin)|toss(coin))[0.5,0.5]$, $P_2 \models^{me} (headUp(coin)|toss(coin))[0.5,0.5]$. This result suggests that the difference between P_1 and P_2 is ignored under the maximum entropy reasoning. By calculating the knowledge entropy of P_1 and P_2, we have $K_{P_1}(headUp(coin)|toss(coin)) = [0,0]$ and $K_{P_2}(headUp(coin)|toss(coin)) = [0,1]$. Thus we know that P_1 is more precise than P_2. Obviously, the conclusion $(headUp(coin)|toss(coin))[0.5,0.5]$ is more acceptable under P_1 than under P_2.

4 Ignorance and Degree of Satisfaction

The knowledge measurement defined above is not sufficient either. Intuitively, the knowledge measurement $K_P(\psi|\phi)$ indicates the ignorance about the conditional event $(\psi|\phi)$. Unfortunately, such an interval cannot sufficiently reflex the ignorance about $(\psi|\phi)$. This is not surprising, since $K_P(\psi|\phi)$ is determined only by the tight probability bound of the conditional event $(\psi|\phi)$, and other knowledge is not considered in $K_P(\psi|\phi)$.

Example 2 *Let a PLP P be defined as*

$$P = \left\{ \begin{array}{l} (fly(X)|bird(X))[0.9,1], \\ (bird(X)|magpie(X))[1,1] \\ (sickMagpie(X)|magpie(X))[0,0.1], \\ (magpie(X)|sickMagpie(X))[1,1] \end{array} \right\}$$

From P, we can infer that

$$\begin{array}{l} P \models_{tight} (fly(t)|magpie(t))[0,1], \\ P \models_{tight} (fly(t)|sickmagpie(t))[0,1], \\ P \models^{me}_{tight} (fly(t)|magpie(t))[0.9,0.9], \\ P \models^{me}_{tight} (fly(t)|sickMagpie(t))[0.9,0.9]. \end{array}$$

Here, we have $K_P(fly(t)|sickmagpie(t)) = K_P(fly(t)|magpie(t))$. However, since the proportion of sick magpies in birds is smaller than the proportion of magpies in birds, the knowledge about birds can fly should be cautiously applied to sick magpies than magpies. In another word, *more than 90% birds can fly* is more about magpies than sick magpies. Therefore, accepting that 90% magpies can fly is more rational than accepting that 90% sick magpies can fly. However, knowledge measurement cannot differentiate this. Below, we introduce two measures to overcome this weakness. These two measures are the instantiated measures from the general framework for analyzing and reasoning with imprecise PLPs proposed in [18].

In probabilistic theory and information theory, how to measure the *distance* between probability distributions is a major topic. One of the most common measures for comparing probability distributions is the KL-divergence.

Definition 3 *Let Pr and Pr' be two probability distributions over the same set $\mathcal{I}_\Phi$. The KL-divergence between Pr and Pr' is defined as:*

$$KL(Pr\|Pr') = -\Sigma_{I\in\mathcal{I}_\Phi} Pr(I)\log\frac{Pr'(I)}{Pr(I)}$$

It is worth noting that KL-divergence is asymmetric. KL-divergence is also called *relative entropy*.[2] An important conclusion is that $H(Pr) = KL(Pr||Pr_{unif})$, where Pr_{unif} is the uniform distribution.

From the KL-divergence, we can measure the amounts of the information that should be received to believing lower and upper bounds for $(\psi|\phi)$ other than the probability given by maximum entropy.

$$\begin{aligned}\nu^{pos}_{P,(\psi|\phi)}(v) &= \min_{Pr\models P, Pr(\psi|\phi)=v} KL(Pr||me[P]),\\ &\quad \text{where } v \geq me[P]\\ \nu^{neg}_{P,(\psi|\phi)}(v) &= \min_{Pr\models P, Pr(\psi|\phi)=v} KL(Pr||me[P]),\\ &\quad \text{where } v \leq me[P]\end{aligned}$$

$$\begin{aligned}dis^{pos}_{P,(\psi|\phi)}(u,v) &= |\nu^{pos}_{P,(\psi|\phi)}(u) - \nu^{pos}_{P,(\psi|\phi)}(v)|\\ dis^{neg}_{P,(\psi|\phi)}(u,v) &= |\nu^{neg}_{P,(\psi|\phi)}(u) - \nu^{neg}_{P,(\psi|\phi)}(v)|\end{aligned}$$

Definition 4 *Let P be a PLP and $(\psi|\phi)$ be a conditional event. Suppose that $P \models_{tight} (\psi|\phi)[l,u]$ and $P \models^{me}_{tight} (\psi|\phi)[p_{me}, p_{me}]$, then we have that:*

$$\mathsf{SAT}^{KL}_P((\psi|\phi)[a,b]) = \begin{cases} 0.5 * (\frac{dis^{pos}_{P,(\psi|\phi)}(p_{me},\min(u,b))}{dis^{pos}_{P,(\psi|\phi)}(p_{me},u)} + \frac{dis^{neg}_{P,(\psi|\phi)}(p_{me},\max(a,l))}{dis^{neg}_{P,(\psi|\phi)}(p_{me},l)}), \\ \quad \textit{if } p_{me}\in[a,b] \\ 0, \quad \textit{otherwise}\end{cases}$$

It is proved in [18] that $\mathsf{SAT}^{KL}_P(\psi|\phi)[a,b]$ can be interpreted as the *second order* probability that the actual probability of $(\psi|\phi)$ falls in the interval $[a,b]$.

Example 3 *Let P be a PLP:*

$$P = \left\{\begin{array}{l}(fly(X)|bird(X))[0.9,1]\\(bird(X)|magpie(X))[1,1]\\(magpie(X)|sickmagpie(X))[1,1]\end{array}\right\}$$

Let two queries be $?(fly(t)|magpie(t))$ and $?(fly(t)|sickmagpie(t))$. The we have (c.f. [18])

$P \models_{tight} (fly(t)|magpie(t))[0,1]$,
$P \models^{me}_{tight} (fly(t)|magpie(t))[0.9,0.9]$ *and*
$P \models_{tight} (fly(t)|sickmagpie(t))[0,1]$,
$P \models^{me}_{tight} (fly(t)|sickMagpie(t))[0.9,0.9]$.

So, we cannot differentiate magpies from sick magpies in their ability of flying, although sick magpies are more a special kind of magpies, and therefore they are less likely to be able to fly than magpies. In contract, in our framework, we have $\mathsf{SAT}^{KL}_P((fly(t)|magpie(t))[0.8,1]) = 0.58$, and $\mathsf{SAT}^{KL}_P((fly(t)|sickmagpie(t))[0.8,1]) = 0.53$ for two queries $?(fly(t)|magpie(t))[0.8,1]$ and $?(fly(t)|sickmagpie(t))[0.8,1]$. By comparing their KL degrees of satisfaction, it is clear that magpies are more likely able to fly than sick magpies.

[2] It should be noted that $KL(Pr\|Pr')$ is undefined if $Pr'(I) = 0$ and $Pr(I) \neq 0$. This means that Pr has to be absolutely continuous w.r.t. Pr' for $KL(Pr\|Pr')$ to be defined.

5 A System for Answering Queries

5.1 Efficient implementation

To efficiently return a query result given a PLP, we implemented the efficient algorithms proposed in [6, 8]. Using these algorithms, a PLP can be translated into a liner or nonlinear optimization problem. We implemented these algorithms in Java and solved the underlying optimization problem using a component in Matlab. In addition, we also implemented the calculation of ignorance and degree of satisfaction with the algorithms given below. These algorithms rely on the algorithms provided in [6, 8] as well as the software Matlab to optimize a PLP.

Algorithm 1 (KLIgnorance)
Input: *PLP P and a ground query $Q = ?(\psi|\phi)$*
Output: *Ignorance value for Q*

1. **IF** *P is unsatisfiable* **THEN return** 1
2. **IF** $P \models_{tight} (\phi|\top)[0,0]$ **THEN return** 1
3. *Compute the tight bound $[l,u]$ for $(\psi|\phi)$ by Algorithm Tight_0_Consequence in Fig. 5. in [6].*
4. *Compute the simplified PLP D index sets R and associate numbers a_r and optimal solution $y^\star_r$ ($r \in R$) by Algorithm Tight_me_Consequence in Fig. 7. in [6].*
5. *Compute the optimal value ig_{neg} of the optimization problem:*

$$ig_{neg} = \max\left(-\sum_{r\in R} y^l_r(\log y^l_r - \log a_r)\right)$$

subject to: y^l_r satisfies $LC(\top, D^l, R)$, where $D^l = D \cup \{(\psi|\phi)[l,l]\}$

6. *Compute the optimal value ig_{pos} of the optimization problem:*

$$ig_{pos} = \max\left(-\sum_{r\in R} y^u_r(\log y^u_r - \log a_r)\right)$$

subject to: y^u_r satisfies $LC(\top, D^u, R)$, where $D^u = D \cup \{(\psi|\phi)[u,u]\}$.

7. *Compute optimal solution y'_r ($r \in R$) for $P' = \emptyset$ by Algorithm Tight_me_Consequence in Fig. 7. in [6]. $p_{me} := me[P'](\psi|\phi)$.*
8. *Compute the optimal value ig'_{neg} of the optimization problem:*

$$ig'_{neg} = \max\left(-\sum_{r\in R} y^l_r(\log y^l_r - \log a_r)\right)$$

subject to: y^l_r satisfies $LC(\top, D^l_0, R)$, where $D^l_0 = \{(\psi|\phi)[l,l]\}$

9. *Compute the optimal value ig'_{pos} of the optimization problem:*

$$ig'_{pos} = \max\left(-\sum_{r\in R} y_r^u(\log y_r^u - \log a_r)\right)$$

subject to: y_r^u satisfies $LC(\top, D_0^u, R)$, where $D_0^u = \{(\psi|\phi)[u,u]\}$.

10. **IF** $p_{me} < u$ **THEN** $s_1 := 1$ **ELSE** $s_1 := -1$
11. **IF** $p_{me} > l$ **THEN** $s_2 := 1$ **ELSE** $s_2 := -1$
12. $ig := (s_1 * ig_{pos} + s_2 * ig_{neg})/(ig'_{pos} + ig'_{neg})$
13. **RETURN** ig

Algorithm 2 (KLDivergence)
Input: *PLP P, $me[P]$, a conditional event $(\psi|\phi)$, and a probability value v.*
Output: $kl = \min_{Pr\models P, Pr(\psi|\phi)=v} KL(Pr||me[P])$

1. *$me[P]$ is obtained from Algorithm 1 and is represented as y^{me}.*
2. *Compute the tight bound $[l', u']$ for $(\psi|\phi)$ by Algorithm Tight_0_Consequence in Fig. 5. in [6].*
3. **IF** $v \notin [l', u']$ **THEN return** *ERROR.*
4. *Compute the optimal value kl of the optimization problem:*

$$kl = \min\left(\sum_{r\in R} y_r \log y_r - \sum_{r\in R} y^r \log y^{me}\right)$$

subject to: y_r satisfies $LC(\top, D^V, R)$, where $D^V = D \cup \{(\psi|\phi)[v,v]\}$.

5. **return** *kl*

Algorithm 3 (KLSatisfaction)
Input: *PLP P and a ground query $Q = ?(\psi|\phi)[l,u]$*
Output: *KL degree of satisfaction for Q*

1. **IF** $P \models_{tight} (\phi|\top)[0,0]$ **THEN return** 1.
2. **IF** $l \geq u$**THEN return** *0.*
3. *Compute the tight bound $[l', u']$ for $(\psi|\phi)$ by Algorithm Tight_0_Consequence in Fig. 5. in [6].*
4. **IF** $l < l'$ **THEN** $l := l'$.
5. **IF** $u > u'$ **THEN** $u := u'$.
6. *Compute $s_p = \nu^{pos}_{P,(\psi|\phi)}(u')$ by Algorithm 2.*
7. *Compute $s_n = \nu^{neg}_{P,(\psi|\phi)}(u')$ by Algorithm 2.*
8. *Compute $s'_p = \nu^{pos}_{P,(\psi|\phi)}(l)$ by Algorithm 2.*
9. *Compute $s'_n = \nu^{neg}_{P,(\psi|\phi)}(u)$ by Algorithm 2.*
10. $sat := 0.5 * (s'_p/s_p + s'_n/s_n)$
11. **return** sat

In our querying system, shown in Figure 1, we have observations, background knowledge, as well as the knowledge obtained from sources (e.g. experts). Background knowledge and the knowledge from different sources are merged to obtain a PLP. Observations are used when constructing queries. Each PLP can be analyzed with the measures defined/introduced previously. The details on other components (like merging and revision) are omitted here due to space limitation.

5.2 Additional information used in querying PLPs

Observation vs. a priori facts: In PLPs, ground formulae of the form $(\phi(t)|\top)[1,1]$ are used to state a priori facts from statistics, i.e., something must be true (statistically) is regarded as a fact. From $(\phi(t)|\top)[1,1]$, we know that object t must possess property ϕ even before we observe it. This is different from observing t having property ϕ. Observing an event (such as a test result) about an individual does not infer that the event would happen for sure (for another individual). So, observations cannot be represented as formulae of the form $(\psi(a)|\top)[1,1]$ in a PLP. Doing so implies that we know $\psi(a)$ being true even before it is observed. In another word, taking $\psi(a)$ as a probabilistic event, we cannot predict if $\psi(a)$ is true or false before we observe it. In our System, all observations are stored in a separate database (named OBS). When querying $(\psi|\phi)[l,u]$ on PLP P, this observation database OBS is automatically called, so querying $(\psi|\phi)[l,u]$ is equivalent to querying $(\psi|\phi \wedge \bigwedge OBS)[l,u]$ on P.

Background knowledge: In practice, source knowledge bases (PLPs) can be obtained from experts, from some experiments, or are elicited from data in published papers. However, given an application, there is richer knowledge that is normally not included in a paper or stated in an experiment, but this knowledge may have been implicitly used. When such knowledge is present, we include it in a PLP when appropriate. For example, knowledge about some general population statistics should be treated as background knowledge, while the effectiveness of a new drug should be treated as specialized knowledge.

6 Application to Substrates Prediction

Considerable investment has been made into the in silico prediction of substrates, and especially, inhibitors of enzymes. This investment has been driven by a fundamental desire to understand more about how biomolecules recognize their ligands and by the commercial imperative to develop new drugs. Almost all pharmaceutical companies include an element of target-based approaches in their drug discovery programmes. The aim of our analyzing/querying system is to provide a very rapid screening for likely ligands (either substrates or inhibitors, depending on the context). It will be particularly useful in situations where a number of similar compounds have been screened experimentally, but information is not available for all possible members of that group of compounds. By providing a simple means to encode existing experimental knowledge and return results within minutes we see this as a valuable addition to initial computational screening approaches.

6.1 Case study I: Rapid sugar kinase enzymes prediction

Our first example is from biochemistry on the human enzyme galactokinase, which uses galactose as a substrate.

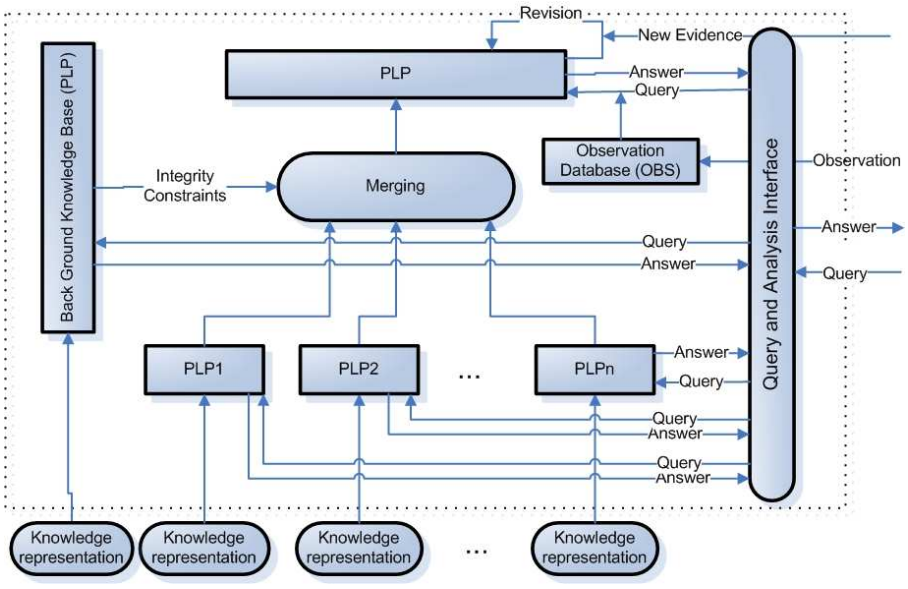


Figure 1: System Architecture

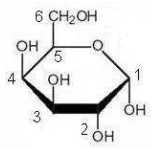


Figure 2: The α-D-Galactose molecule

Galactose has the molecular formula $C_6H_{12}O_6$, but other compounds have the same or similar formula. Since not all possible substrates for the enzyme have been tested, the information regarding this enzyme and its substrates is incomplete, can we then predict which will be the substrates for the human enzyme galactokinase based on incomplete and imperfect information? Many factors lead to the information being imperfect including different research laboratories using different criteria for scoring a compound as a substrate and some information is based on galactokinases from other species, so we cannot be certain that substrate specificity is conserved for humans.

Each galactose molecule is arranged as a hexagonal ring (e.g., the α-D-Galactose molecule in Figure 2). There are six carbon atoms in a galactose molecule and one oxygen atom. These six carbon atoms are numbered from 1 to 6 with the right-most carbon atom numbered 1, and then the remaining carbons are numbered clockwise round the ring. The oxygen atom is not numbered. The other atoms can be regarded as coming off these carbon atoms. The first four of the carbon atoms each has an OH molecule attached to it, and the fifth one has the sixth carbon atom attached to it from outside the ring, forming a CH_2OH group. The OH group can either be up or down (i.e. they are chiral). The combination of ups and downs gives a specific form of the molecule (in effect, each form of the molecule is a different compound), and the actual combination can significantly affect the biochemical behavior of the molecule. Therefore, for the OH groups attached to these atoms, we need to know if they are *up, down* or *absent*. The sixth carbon is not chiral, and so the OH is neither up nor down. Hence, the OH for the sixth carbon is marked as either *present* or *absent*. Current experimental results published in the literature provide a set of conditional events with probabilities suggesting how likely a particular structure is a substrate for the enzyme. Table 1 contains this knowledge collected from papers [14, 15, 16, 17], which is then translated into a PLP. For instance, the 5th row of the table (Talose) defines a probabilistic formula as

$$(sub(X)|c1(X,d) \wedge c2(X,u) \wedge c3(X,u) \\ \wedge c4(X,u) \wedge c5(X,u) \wedge c6(X,p))[0.4, 0.6]$$

Initially probabilities were estimated using experimental data and an element of intuition. Where a particular substrate had been demonstrated experimentally to be a substrate of human galactokinase it was assigned a probability of 1.0. Where there was experimental data indicating that a substrate was not phosphorylated by human galactokinase, a value of 0 was assigned. Compounds which had been shown to be substrates of galactokinase from other species were assigned probabilities between 0 and 1. However, not all substrates are equally good. Therefore a second measure, the *product* was calculated. To calculate this value, the specificity constant k_{cat}/K_m was used [4], scaled such that the product value with galactose (which is expected to be the best substrate) was equal to 1.0.

Therefore, in Table 1, we have a column representing their probabilities (or intervals) and another column representing their products of the corresponding compounds to be

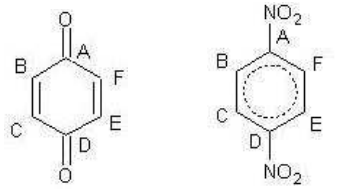
Sugar	C1 -OH	C2 -OH	C3 -OH	C4 -OH	C5 $-CH_2OH$	C6 -OH	P(substrate)	Product	Source
Galactose	D	D	U	U	U	P	1.0	1	[15]
Glucose	D	D	U	D	U	P	0.0	0	[15]
2-Deoxygalactose	D	A	U	U	U	P	1.0	0.47	[15]
Fucose	D	D	U	U	U	A	0.0	0	[15]
Talose	D	U	U	U	U	P	[0.4, 0.6]	[0.056,0.084]	[17, 14]
4-deoxygluocse	D	D	U	A	U	P	[0, 0.5]	[0,0.021]	[16]
3-deoxygalactose	D	D	A	D	U	P	[0.6, 0.9]	[0.036,0.054]	[16]

Table 1: The compounds and their probabilities and products to be substrates, obtained from published papers.

(good) substrates. Column *Source* indicates from which published paper this knowledge is obtained. Based on the probabilistic knowledge in the probabilistic logic program, we can predict the probability for any combination of these six carbons. Twenty-six queries detailed in Table 2 were executed against this PLP and the query results are presented in Table 2. Below we analysis these query results.

By querying the tight bounds for the twenty-six queries detailed in Table 6.1, we can only obtain a trivial interval $[0, 1]$ for all of these queries. This trivial interval indicates that we know nothing about the probability of a compound being a substrate. Reasoning under the maximum entropy principle, we can get probabilities as listed in Table 6.1. One question is that how reliable these values are to guide us finding most possible substrate from these possible compounds. With the analysis of knowledge entropy in Section 3, we have $K_P(sub(s)|\phi_s) = [0, 1]$ for any compound s where the compound structure is stated by ϕ_s. So this measurement does not provide us with useful information either. Now we look into the ignorance measures of these query results. The ignorance values of the query results of the twenty-six compounds w.r.t. this PLP are around 0.005, a very low value. These ignorance values suggest that the probabilities obtained by applying the maximum entropy principle are acceptable and can serve as good indicators about how likely a compound can be a substrate. Since in substrate prediction, the comparisons of probabilities are much more important than actual probability values, we do not need to calculate the degrees of satisfaction for these queries with intervals.

Overall, the predictions appear to over-estimate the probabilities for each possible substrate. For example, given that the *Fucose* (which has the OH group attached to the sixth carbon atom absent) has been shown experimentally not to be a substrate, it is surprising to see compounds which also lack this OH group predicted as having high probabilities as substrates. Of course in compiling the data in Table 1, all the information was weighted equally - for example the presence or absence of the OH group at position 6 was considered of equal worth to the information about the OH at position 2. In fact it is likely that some positions are more important than others in determining substrate specificity. However, in implementing screens such as these, the amount of knowledge to be included will always be a balance between including enough to enable valid predictions, but not so much that the initial knowledge collection and tabulation becomes unreasonably time consuming.



(a) (b)

Figure 3: Examples of NAD(H)-quinone oxidoreductase 1 (NQO1) substrates.

Despite these limitations, the predictions do appear to have some value in that the ranking of the compounds in terms of their probability of being a substrate seems mostly reasonable and in line with chemical intuition. Ultimately for such a system to be useful to bioscientists, it is this ranking which must be reliable. The most likely use of such a system is to act as a preliminary screen for potential substrates or inhibitors followed by experimental testing of those compounds. Time and expense can be saved if those compounds most likely to be good substrates (or inhibitors) appear at the top of the list and are, therefore, prioritized in the experimental work. Thus the absolute values of the predicted probabilities are less important than the rank order of the compounds.

6.2 Case Study II: Substrate prediction for NQO1

NAD(H)-quinone oxidoreductase 1 (NQO1) is a broad specificity enzyme which catalyses the reduction of a range of aromatic compounds. It was chosen for the second case study as a large variety of different compounds (including quinones, nitroaromatics and benzimidazoles) have been tested as substrates. In contrast to Case study I, the chemical diversity of the known substrates is wider leading to a greater number of variables to consider.

Two of the many compounds which have been tested experimentally as substrates for NQO1 are a quinone com-

Sugar	$C1$ -OH	$C2$ -OH	$C3$ -OH	$C4$ -OH	$C5$ -CH_2OH	$C6$ -OH	P(substrate)	Product
2dAll	D	A	D	D	U	P	0.6529	0.4611
2dGlc	D	A	U	D	U	P	0.6154	0.3939
2dGul	D	A	D	U	U	P	0.6694	0.5000
I	D	A	A	D	U	P	0.5869	0.4083
II	D	A	A	U	U	P	0.6676	0.5376
2,3,4d	D	A	A	A	U	P	0.5509	0.4721
3dAll	D	D	A	D	U	P	0.6003	0.1138
3dMan	D	U	A	D	U	P	0.5539	0.5000
3dTal	D	U	A	U	U	P	0.5636	0.4282
III	D	D	A	A	U	P	0.5321	0.3503
IV	D	U	A	A	U	P	0.5134	0.4785
4dAll	D	D	D	A	U	P	0.5314	0.4611
4dMan	D	U	U	A	U	P	0.4706	0.4282
V	D	A	D	D	U	A	0.5463	0.4811
VI	D	A	U	D	U	A	0.5481	0.4514
VII	D	A	D	U	U	A	0.5481	0.5000
VIII	D	A	A	D	U	A	0.5703	0.4572
IX	D	A	A	U	U	A	0.5682	0.5020
X	D	A	A	A	U	A	0.5233	0.4814
XI	D	D	A	D	U	A	0.5451	0.3518
XII	D	U	A	D	U	A	0.5234	0.5000
XIII	D	U	A	U	U	A	0.5278	0.4670
XIV	D	D	A	A	U	A	0.5146	0.4179
XV	D	U	A	A	U	A	0.5064	0.4895
XVI	D	D	D	A	U	A	0.5144	0.4811
XVIII	D	U	U	A	U	A	0.4879	0.4670

Table 2: The probabilities and products of some compounds being a substrate by querying on the PLP.

pound, benzo-1,4-quinone (Figure 3(a)) and a nitroaromatic compound 1,4-dinitrobenzene (Figure 3(b)). Representing these compounds in tabular form required assigning each position in the six-membered ring a letter descriptor from A to F. For each molecule, the most oxidised substituent was placed at the top of the structural representation and designated A. Positions B through F were then defined by moving round the ring sequentially in an anticlockwise fashion. In these initial studies we concentrated on six membered rings substituted with ketone, methyl and nitro groups.

A	B	C	D	E	F	Probability
NO2	H	H	H	H	H	[0,0]
NO2	H	NO2	H	H	H	[0,0]
NO2	H	H	CHO	H	H	[0,0]
NO2	NO2	H	H	H	H	[0,0]
NO2	H	H	NO2	H	H	[0,0]
O	H	H	O	H	H	[0.20,0.28]
O	CH3	H	O	H	H	[0.17,0.31]
O	CH3	H	O	CH3	H	[0.19,0.33]
O	CH3	CH3	O	CH3	H	[0.20,0.28]

Table 3: The compounds and their probability intervals, obtained from published papers [1, 3].

In this initial case study, knowledge was collected from a limited number of papers [1, 3] which described the activity of the enzyme towards a number of structurally related compounds (Table 3). Probabilities were derived from published data in these papers on specificity constants in which the error in the experimental determination was used to define the range of values. When used to make predictions about unknown compounds (Table 4), the results were broadly similar to those seen in Case Study I. Table 4 gives the summary of sixteen queries based on the probabilistic knowledge given in Table 3. There appeared to be a tendency to over-estimate probabilities (especially for compounds closely related in structure to those with low, or zero, experimentally determined activity). Nevertheless, if these compounds are excluded the rank order of the remaining ones appears sensible.

7 Related Work and Conclusion

Some systems are provided for modeling and querying on probabilistic knowledge, for example, SPIRIT [12] and PIT [13]. These two systems work on propositional probabilistic logics while our system works on PLPs. The main advantage of our framework is its ability to analyze the knowledge contained in PLPs, especially w.r.t. queries. For analyzing probabilistic knowledge bases, in [11, 10], the authors provided a second order uncertainty to measure the reliability of accepting the precise prob-

A	B	C	D	E	F	Probability
NO2	H	H	H	NO2	H	0.0000
NO2	H	H	NO2	CH3	H	0.3194
NO2	H	H	CHO	CH3	H	0.3194
NO2	H	H	O	CH3	H	0.3294
NO2	H	NO2	H	CH3	H	0.3217
NO2	H	NO2	NO2	H	H	0.1949
NO2	H	NO2	O	H	H	0.2235
NO2	NO2	H	O	H	H	0.2172
O	H	H	H	NO2	H	0.2949
O	H	H	NO2	CH3	H	0.3917
O	H	H	CHO	CH3	H	0.3197
O	H	H	O	CH3	H	0.3629
O	H	NO2	H	CH3	H	0.4000
O	H	NO2	NO2	H	H	0.3612
O	H	NO2	O	H	H	0.3477
O	NO2	H	O	H	H	0.3338

Table 4: The Predictions for some compounds.

ability obtained by applying maximum entropy principle as the answer to a query in propositional probabilistic logic. Their second order uncertainty is directly computed from the probability interval of the query inferred from P, and therefore is independent of the knowledge base. In contrast, our ignorance provides more information about the underlying knowledge base and is more accurate in terms of reflecting the knowledge in a PLP. In Example 3, the second order probabilities of $(fly(t)|magpie(t))$ and $(fly(t)|sickMagpie(t))$ are the same. However the ignorance values for the two queries are different.

In this paper, we provided a framework and a tool for reasoning with imprecise probabilistic logic programs. In our framework, background knowledge and application-specific knowledge are combined to create a PLP (or maybe multiple PLPs), and observations are represented in a separate set. In this tool, a user has an option to analyze the quality of a PLP by retrieving the ignorance values with respect to application-specific queries. Also, the reasoning power is enhanced because reliable informative bounds can be extracted for any query. Two case studies are deployed to demonstrate how this framework and the tool can be used in real world applications. Our system can also perform merging when multiple PLPs concerning the same application are available, and perform revision (of a PLP) when some sure new evidence is collected.

References

[1] Z Anusevicius, J Sarlauskas, and N Cenas. Two-electron reduction of quinones by rat liver nad(p)h:quinone oxidoreductase: quantitative structure-activity relationships. *Arch Biochem Biophys*, 404(2):254–262, 2002.

[2] C Baral and M Hunsaker. Using the probabilistic logic programming language p-log for causal and counterfactual reasoning and non-naive conditioning. In *Proc. of IJCAI*:243-249, 2007.

[3] N Cenas, A Nemeikaite-Ceniene, E Sergediene, H Nivinskas, Z Anusevicius, and J Sarlauskas. Quantitative structure-activity relationships in enzymatic single-electron reduction of nitroaromatic explosives: implications for their cytotoxicity. *Biochim Biophys Acta*, 1528(1):31–38, 2001.

[4] A Cornish-Bowden. *Fundamentals of Enzyme Kinetics (3rd Edition)*. Portland Press, London, UK, 2004.

[5] N Fuhr. Probabilistic datalog: Implementing logical information retrieval for advanced applications. *JASIS*, 51(2):95–110, 2000.

[6] G Kern-Isberner and T Lukasiewicz. Combining probabilistic logic programming with the power of maximum entropy. *Artificial Intelligence*, 157(1-2):139–202, 2004.

[7] T Lukasiewicz. Probabilistic logic programming. In *Proc. of ECAI*:388–392, 1998.

[8] T Lukasiewicz. Probabilistic logic programming with conditional constraints. *ACM Trans. Comput. Log.*, 2(3):289–339, 2001.

[9] L De Raedt, A Kimmig, and H Toivonen. Problog: A probabilistic prolog and its application in link discovery. In *Proc. of IJCAI*:2462–2467, 2007.

[10] W Rödder. On the measurability of knowledge acquisition, query processing. *Int. J. Approx. Reasoning*, 33(2):203–218, 2003.

[11] W Rödder and G Kern-Isberner. From information to probability: An axiomatic approach - inference isinformation processing. *Int. J. Intell. Syst.*, 18(4):383–403, 2003.

[12] W Rödder, E Reucher, and F Kulmann. Features of the expert-system-shell spirit. *Logic Journal of the IGPL*, 14(3):483–500, 2006.

[13] M Schramm and V Fischer. Probabilistic reasoning with maximum entropy - the system pit (system description). In *WLP*, 1997.

[14] C Sellick and R Reece. Contribution of Amino Acid Side Chains to Sugar Binding Specificity in a Galactokinase, Gal1p, and a Transcriptional Inducer, Gal3p. *J. Biol. Chem.*, 281(25):17150–17155, 2006.

[15] DJ Timson and RJ Reece. Sugar recognition by human galactokinase. *BMC Biochemistry*, 4:16, 2003.

[16] J Yang, X Fu, Q Jia, J Shen, J Biggins, J Jiang, J Zhao, J Schmidt, P Wang, and J Thorson. Studies on the substrate specificity of escherichia coli galactokinase. *Organic Letters*, 5(13):2223–2226, 2003.

[17] J Yang, L Liu, and J Thorson. Structure-basedenhancement of the first anomeric glucokinase. *ChemBioChem*, 5(7):992–996, 2004.

[18] A Yue, W Liu, and A Hunter. Measuring the ignorance and degree of satisfaction for answering queries in imprecise probabilistic logic programs. In *Proc. of SUM*:386-400, 2008.

Noise quantization via possibilistic filtering

Kevin Loquin
IRIT
Université Paul Sabatier
118 Route de Narbonne
F-31062 Toulouse Cedex 9
Kevin.Loquin@irit.fr

Olivier Strauss
LIRMM
Université Montpellier II
161, rue Ada
F-34392 Montpellier Cedex 5
Olivier.Strauss@lirmm.fr

Abstract

In this paper, we propose a novel approach for quantifying the noise level at each location of a digital signal. This method is based on replacing the conventional kernel-based approach extensively used in signal filtering by an approach involving another kind of kernel: a possibility distribution. Such an approach leads to interval-valued resulting methods instead of point-valued ones. We show, on real and artificial data sets, that the length of the obtained interval and the local noise level are highly correlated. This method is non-parametric and advantageous over other methods since no assumption about the nature of the noise has to be made, except its local ergodicity.

Keywords. Signal processing, kernel methods, possibility distribution, noise quantization, Choquet integral.

1 Introduction

The reliability of a great number of signal processing methods inherently relies on the possibility of adjusting their parameters to account for noise level over the input signal. Examples of such procedures are image restoration, edge detection [18], motion estimation [1], denoising [26, 27], super-resolution [14], shape-from-shading [34], sensor fusion [3, 29] and feature extraction or segmentation [22].

Noise in a signal is usually referred to random variations of the measured signal. These variations can be produced by several factors including thermal effect, saturation, sampling, quantization and transmission. Since repeating the acquisition process is usually not possible, the noise level has to be estimated by means of a single signal occurrence.

Noise is generally considered as being independent from the signal level and added to it. One of the most widely encountered model assumes this random noise as being centered and normally distributed. However, phenomena like film grain, speckle, impulse noise, sampling effect, quantization or saturation induce a fluctuation of signal's value that cannot be modelled by a Gaussian zero mean process. For example, in medical images produced by a gamma camera, the noise is rather described by a Poisson process (i.e. the noise level depends on the signal level).

In early approaches (see e.g [25]), noise estimation consisted in assuming stationarity of the random variations of the signal. The computation of the standard deviation of the noise were performed by analyzing the signal obtained by high-pass filtering of the original signal. The main challenge in these estimations is to be able to tell whether a signal variation is due to the noise or to the signal itself, which can involve significant variations.

In more recent papers, some authors propose to abandon either stationarity or additivity of the noise. Rangayyan et al. [28] consider an adaptive neighbourhood approach that is able to account for an additive non-stationary noise. Corner et al. show that analyzing the Laplacian of the signal allows to deal with both additive and multiplicative noise [5].

Unfortunately, neither additive nor multiplicative random noise are good models for real signal contamination, even for instance, for conventional CCD sensor [18]. Therefore, many approaches [18, 16, 23] propose to model the acquisition noise as being Poisson distributed.

In these model-based approaches, the noise is assumed to follow a hypothetically known distribution and noise level estimation consists in estimating the different parameters on which the variance of the assumed distribution depends. Moreover, any model-based method assumes the type of the acquisition machine to be known.

If nothing can be assumed about the nature of the noise, except its local ergodicity, only a very local approach has to be considered to estimate the noise level for each location or, at least, for each user-selected homogeneous region of the signal. Moreover, since signal processing mainly consists of extracting or estimating some physically meaningful characteristics from intensity values of the signal, it should be important to understand how the uncertainty due to random perturbation propagates through any algorithm step.

A wide range of those signal processing methods relies on a kernel-based approach [20] for direct or iterative, linear or non-linear algorithms and for filtering (stochastic, band pass, anti-aliasing, ...), geometrical transformations (rescaling, rotations, homographies, anamorphosis, ...), sampling rate conversion, fusion, for enhancing or removing details, etc. The kernels usually encountered are probability distributions: they are positive functions whose total weight (their integral in the infinite domain and their sum in the finite domain) sums to 1. The main difficulty in these kernel-based methods is that the nature of both signal and perturbation can change during the complete analysis, from step to step.

By switching from probability theory to possibility theory, we propose new methods that take into account a lack of knowledge on the proper kernel to be used [21]. Indeed, a possibility distribution represents a convex hull of probability distributions and hence of kernels. In this adaptation of the usual kernel methods, the conventional Lebesgue integral operator is replaced by a pair of Choquet integrals according to the possibility measure and the necessity measure associated with the chosen possibility distribution. The resulting interval (and more precisely its length) reflects the lack of knowledge of the modeller on the most adequate kernel to use.

As an example, the use of the interval-valued gradient estimation of an image, proposed in [17], leads to a threshold-free robust edge detector. This robustness is due to the fact that the length of the interval-valued estimation is highly correlated with the input image random noise. The information (about the noise) contained in the resulting interval is properly taken into account in the edge detector, thus enabling an automatic rejection of the "false" edges due to noise.

In this paper, we propose to study the link between the length of the interval-valued result of a possibilistic filtering on a signal and the input signal random noise. Actually, we discuss the fact that this approach is, to our opinion, in its spirit, better founded than the usual noise level estimators. Furthermore, we propose to highlight the empirical correlation between the length of the output of the interval-valued filtering and the input signal random noise on repeated acquisitions of real and synthetic images.

The paper is organized as follows. In section 2, we present how the digital filtering is performed by means of convolution kernels with unitary gains. We present the possibility distribution-based filtering, which is theoretically justified by Theorem 1. In section 3, we describe our method for estimating the noise level at each sample location of a signal. In section 4, we compare our method to three other usual noise level estimates on synthetic and real noisy images, before concluding in section 5.

2 A possibilistic extension of convolution kernel-based linear filtering

2.1 Convolution kernel-based signal filtering

Let $S = (S_i)_{i=1,...,N}$ be a digital signal defined on N locations $\{\omega_1, ..., \omega_N\}$ of an underlying infinite domain Ω. Note that the locations $\{\omega_1, ..., \omega_N\}$ can be identified with their indices $\{1, ..., N\}$. Processing S by a filter, defined by its impulse response κ, mathematically corresponds to the discrete convolution of S by κ. This is why κ can also be called a convolution kernel. The value $\hat{S}_n$ of the filtered signal at the n^{th} location of $\{1, ..., N\}$ is thus obtained by:

$$\hat{S}_n = \sum_{i=1}^{N} S_i \kappa_{n-i}.$$

$\kappa_{n-\bullet} = (\kappa_{n-i})_{i=1,...,N}$ is the convolution kernel κ shifted to the location n of $\{1, ..., N\}$. We propose to denote this particular shifted kernel by $\kappa^n = (\kappa_i^n)_{i=1,...,N}$. $\hat{S}_n$ is thus obtained by:

$$\hat{S}_n = \sum_{i=1}^{N} S_i \kappa_i^n. \tag{1}$$

In many applications like low-pass filtering, the used convolution kernels are positive and have a unitary gain, i.e.

$$\sum_{i=1}^{N} \kappa_i = 1.$$

In that case, the convolution kernel can be seen as a probability distribution that induces a discrete probability measure P_κ, computed in this way:

$$\forall A \subseteq \Theta,\ P_\kappa(A) = \sum_{i \in A} \kappa_i.$$

For each location n, its associated shifted convolution kernel κ^n is still a probability distribution. Thus, expression (1) is equivalent to computing the expected value $\hat{S}_n$ of the signal S at the location n, considering the probability measure P_{κ^n} on $\{1, ..., N\}$, i.e.:

$$\hat{S}_n = \mathbb{E}_{P_{\kappa^n}}(S). \tag{2}$$

In that case, the filtered value of the signal can be interpreted as the expected value of the signal, knowing that the uncertainty concerning the location is modelled by P_{κ^n}. This interpretation is not very relevant because the aim of the filtering is not to try to evaluate the real value of a signal under uncertainty. The aim of the filtering is to modify the input signal according to the practitioner's needs. The only reason why we propose to rewrite the linear filtering with the expectation operator is that it enables us to deal with a family of convolution kernels by switching from the usual probability theory to imprecise probability theory.

2.2 Extension of signal filtering to possibility theory

By writing the linear filtering with a unitary gain filter as an expectation according to a probability measure, we open new perspectives to this approach by repositioning it in the field of new uncertainty theories. Instead of using an additive measure for each neighbourhood of a sample location, i.e. a probability measure, we propose to use the simple non-additive confidence measure called a possibility measure [9]. We propose to use this theory among others because of its computational simplicity. First, the possibility distribution is a tool that can be simply modelled by just a set of N weights on the locations $\{1, ..., N\}$, whereas most of the other imprecise probability theories [33] require more assessments. Besides, we propose to use the Choquet integral, that extends the usual linear expectation operator, by extending the convolution operator to possibility measures in place of probability measures. This tool is well know and very simply computed.

This section presents and interprets this new filtering approach, based on possibility measures and Choquet integrals, that enables a signal to be filtered by means of a family of convolution kernels.

2.2.1 A possibility distribution is a family of filters

A possibility measure is non-additive and possesses a dual confidence measure, called a necessity measure, denoted by N and computed in this way:

$$\forall A \subseteq \Theta,\ N(A) = 1 - \Pi(A^c). \tag{3}$$

The two measures, Π and N, encode a family of probability measures, denoted by $\mathcal{M}(\Pi)$, and defined by:

$$\mathcal{M}(\Pi) = \{P \mid \forall A \subseteq \Theta, N(A) \leq P(A) \leq \Pi(A)\}.$$

This encoding property is due to the sensitivity analysis interpretation [32] of possibility theory.

A possibility measure can be defined from a possibility distribution π^n. Such a distribution is normalized in the sense that

$$\max_{i \in \Theta} \pi_i^n = 1.$$

Its associated possibility measure is obtained by:

$$\forall A \subseteq \Theta,\ \Pi_{\pi^n}(A) = \max_{i \in A} \pi_i^n.$$

Thus a unique possibility distribution π^n can encode a whole family of convolution kernels κ^n with unitary gain, denoted by $\mathcal{M}(\pi^n)$ and defined by:

$$\mathcal{M}(\pi^n) = \{\kappa^n \mid \forall A \subseteq \Theta, N_{\pi^n}(A) \leq P_{\kappa^n}(A) \leq \Pi_{\pi^n}(A)\}.$$

This family of convolution kernels being defined, the extension of the convolution (or expectation) operator has to be studied.

2.2.2 The possibilistic extension of the linear filtering

Since a possibility measure is non-additive, the conventional expectation operator cannot be used for filtering. The expectation operator must be replaced by its generalization, called the Choquet integral [6]. Using a Choquet integral and a possibility distribution leads to an interval-valued expectation, instead of a single value, whose upper and lower bounds are given by:

$$\overline{S}_n = \mathbb{C}_{\Pi_{\pi^n}}(S), \tag{4}$$

$$\underline{S}_n = \mathbb{C}_{N_{\pi^n}}(S). \tag{5}$$

The Choquet integral can be considered as a generalization of the conventional expectation operator since, when the used confidence measure is a probability measure, expressions (4) and (5) coincide and are equal to the conventional expectation operator (2).

The key point of this approach is that the interval-valued expectation obtained by means of a possibility distribution is the set of all the single-valued expectations obtained by using all the convolution kernels encoded by the considered possibility distribution.

As a preliminary to the theorem (and its proof) justifying this assertion, some notations are necessary. Let us denote by $\mathcal{L}(\{1, ..., N\})$ the set of bounded sequences of weights on $\{1, ..., N\}$, i.e. $\forall I = (I_i)_{i=1,...,N} \in \mathcal{L}(\{1, ..., N\})$, $\max_{i=1,...,N} |I_i| < \infty$. In [32], this set is called the set of bounded gambles on $\{1, ..., N\}$. Denote $\mathcal{B}(\{1, ..., N\})$, the set of binary (i.e. $\{0, 1\}$-valued) sequences of weights on $\{1, ..., N\}$. Obviously, $\mathcal{B}(\{1, ..., N\}) \subset \mathcal{L}(\{1, ..., N\})$. $\mathcal{B}(\{1, ..., N\})$ can be seen as the set of events on $\{1, ..., N\}$.

Theorem 1. *Let π^n be a possibility distribution. $\forall S \in \mathcal{L}(\{1, ..., N\})$, $\forall \kappa^n \in \mathcal{M}(\pi^n)$,*

$$\mathbb{C}_{N_{\pi^n}}(S) \leq \mathbb{E}_{P_{\kappa^n}}(S) \leq \mathbb{C}_{\Pi_{\pi^n}}(S). \tag{6}$$

Moreover, the bounds are reached: $\forall S \in \mathcal{L}(\{1, ..., N\})$, $\exists \kappa_1^n, \kappa_2^n \in \mathcal{M}(\pi^n)$, such that

$$\mathbb{C}_{N_{\pi^n}}(S) = \mathbb{E}_{P_{\kappa_1^n}}(S),$$
$$\mathbb{C}_{\Pi_{\pi^n}}(S) = \mathbb{E}_{P_{\kappa_2^n}}(S).$$

Proof. The natural extension principle [32] is required to prove Theorem 1. Note that the natural extension of a probability measure P, defined for all the events A of $\mathcal{B}(\{1, ..., N\})$, is the expectation according to P, defined for all S of $\mathcal{L}(\{1, ..., N\})$. Similarly, the natural extension of a possibility measure Π, defined for all the events A of $\mathcal{B}(\{1, ..., N\})$, is the Choquet integral with respect to Π, defined for all S of $\mathcal{L}(\{1, ..., N\})$[1].

[1] This remark is true for the more general belief functions

The natural extension, as defined by Walley, is conservative concerning the imprecision of a possibility measure. The family of natural extensions of the probability measures of the family $\mathcal{M}(\pi^n)$, noted $E(\mathcal{M}(\pi^n))$, is the same as the family of expectations dominated by the Choquet integral according to π^n, noted $\mathcal{M}(\mathbb{C}_{\Pi_{\pi^n}})$. This property of the natural extension can be found in Walley's book [32] for an upper prevision $\underline{P}$ and its associated set of linear previsions $\mathcal{M}(\underline{P})$. It is enough to conclude that $\forall S \in \mathcal{L}(\{1,...,N\}), \forall \kappa^n \in \mathcal{M}(\pi^n)$,

$$\mathbb{C}_{N_{\pi^n}}(S) = \min\{\mathbb{E}_{P_{\kappa^n}}(S) : \kappa^n \in \mathcal{M}(\pi^n)\},$$
$$\mathbb{C}_{\Pi_{\pi^n}}(S) = \max\{\mathbb{E}_{P_{\kappa^n}}(S) : \kappa^n \in \mathcal{M}(\pi^n)\}.$$

□

This theorem is also valid for infinite domains. The proof is derived from domination theorems proved by Denneberg [7], proposition 10.3 and Schmeidler [30], proposition 3.

This propagation of the imprecision in the choice of the possibility distribution representing a family of kernels to the result of this new possibilistic filtering operation is very interesting. Using a possibility distribution allows the modelling of a lack of knowledge on the proper convolution kernel to be used. Using the generalized expectation operator (4) and (5) directly impacts this ill-knowledge on the output.

Note that in the case of a positive signal S (which is the case of the images that will be processed in section 4), the Choquet integrals, forming the upper and lower expectations, can be explicitely computed by :

$$\overline{S}_n = \mathbb{C}_{\Pi_{\pi^n}}(S) = \sum_{i=1}^{N} \Pi_{\pi^n}(A_{(i)})(S_{(i)} - S_{(i-1)}), \quad (7)$$

$$\underline{S}_n = \mathbb{C}_{N_{\pi^n}}(S) = \sum_{i=1}^{N} N_{\pi^n}(A_{(i)})(S_{(i)} - S_{(i-1)}). \quad (8)$$

The index notation $(.)$ indicates a permutation that sorts the sample locations such that $S_{(1)} \leq S_{(2)} \leq ... \leq S_{(N)}$ and $A_{(i)}$ is a set of samples locations whose value is greater than $S_{(i)}$, i.e. $A_{(i)} = \{j \in \{1,...,N\} | S_j > S_{(i)}\}$. By convention, $S_{(0)} = 0$.

2.2.3 How to choose the possibilistic filter?

The use of a possibility distribution as a family of linear filters is new in signal processing. This approach does not offer clues (especially to possibility theory novices) for choosing the possibility distribution that matches the practitioner's knowledge on the proper convolution kernel to be used. Two hints for helping him to choose a possibility distribution are explored and provided in this paragraph.

First, we propose to use the triangular possibility distribution since it encodes (among others) all the symmetric convolution kernels with the same support [13]. Indeed, many algorithms (for example low-pass filtering) extensively use symmetric convolution kernels.

Second, probability/possibility transformations studied by Dubois et al. [13] can be used, when the practitioner has a vague idea of the convolution kernel to be used. The possibility distributions obtained by these transformations form families of convolution kernels including the kernel to approximate [10, 8]. The objective transformation results in the smallest family containing the original kernel and the subjective transformation [11, 12] results in a larger family of convolution kernels. The latter transformation should be preferred in case of little confidence in the choice of the original convolution kernel.

3 Noise estimation

3.1 Nuggets effect and local estimation by neighbourhood

Geostatistic is the branch of applied statistics that concentrates on the description of spatial patterns [4, 24, 15]. The central tool of geostatistic is the random function which describes the uncertainty of a given spatial characteristic over a domain. The structural assumption underlying most of the geostatistical methods is based on the intuitive idea that, the closer are the regions of interest, the more similar are their associated characteristic values.

However, this intuitive idea is no more so obvious when looking at the closest pairs of sample locations of a spatial data set. Indeed, in general, when plotting the empirical increment of a particular observed property, function of the distance, between different sample locations, this increment[2] does not seem to vanish when the distance tends to 0. This discontinuity, which is supposedly due to geostatical noise, is called the "nuggets effect". This denomination comes from the fact that in gold deposits, gold commonly occurs as nuggets of pure gold that are much smaller than the size of a sample.

When translating this concept from geostatistics to signal processing, the nuggets effect can be illustrated as follows: the variability of a subset A of the signal domain is supposed to reflect the co-occurrence of the intrinsic local variability of the supposed continuous signal underlying the samples and a measurement error. This measurement error sums up the systematic error due to the impulse response of the sensor, the imprecision due to sampling and quantization of the signal and a random variability due to noise. Typically, the variability due to signal increases with the radius Δ of the subset A. On the contrary, the

[2] Generally, the curve of the halve squared increments is plotted. This curve is called the sample variogram [4]

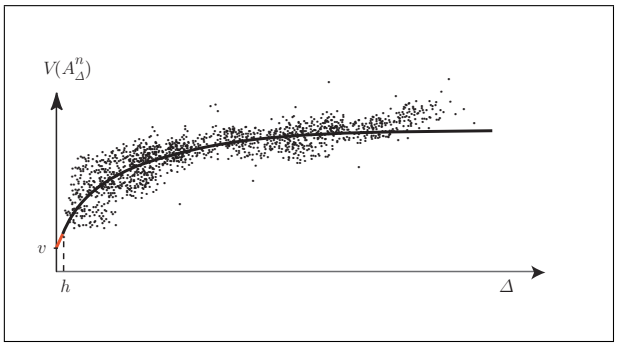


Figure 1: Qualitative example of variogram.

variability due to the measurement error is usually supposed not to depend on Δ. This assumption is reasonable when the sampling is regular and the random noise is supposed to be locally stationary. Thus, if A^n_Δ is a neighbourhood of radius Δ of the n^{th} location ω_n, $V(A^n_\Delta)$, the variability of A^n_Δ is such that :

$$\lim_{\Delta \to 0} V(A^n_\Delta) = v_n, \tag{9}$$

with v_n being the variability due to measurement error at location n. This limit is known as the nuggets effect in the geostatistic field [15]. However, due to sampling, v_n cannot be computed because the local variability cannot be estimated for a scale smaller than the sampling distance h.

A standard technique for catching this variability is to plot a variogram, i.e. to plot the variability of all the sampling locations of $n \in \{1, ..., N\}$, $V(A^n_\Delta)$, as a function of Δ. A manual fitting[3] is generally performed to provide an estimation of the nuggets effect, which is the value of the regression equation for a radius $\Delta = 0$. This estimation is denoted by v.

However, this method presupposes that the error is stationary all over the signal. Moreover, the choice to be made for a particular variogram equation is not generally justified in signal processing. The expert's knowledge is generally not available in this scientific domain to evaluate local dependencies, whereas in geostatistic, the expert, according to the physical nature of the studied area, can provide such information.

A more pointwise estimation of these measurement errors can be obtained by means of a small neighbourhood around each sampled location. This approach is based on assuming local ergodicity. Local ergodicity states that the local variability of the signal in a small neighbourhood of a sampling point reflects the statistical variations of the signal at this location, due to measurement errors. The neighbourhood commonly used is a probability distribution defined over the set of pixels by $\kappa^n = (\kappa^n_i)_{i=1,...,N}$.

[3]Sometimes, automatic fitting procedures (which are not recommanded by geostatisticians), as regression analysis, are performed

The local variability computation leads to a weighted sum due to the additivity of the probability measure. Estimations of the nuggets effect are given by:

$$v_n = \sqrt{\sum_{k=1}^{N} (S_k - \hat{S}_n)^2 \kappa^n_k}, \tag{10}$$

if variability is measured by the standard deviation. And:

$$v_n = \sum_{k=1}^{N} |S_k - \hat{S}_n| \kappa^n_k, \tag{11}$$

if variability is measured by the mean error.

Most of the kernels used to perform this estimation are unimodal, centered and symmetric around the sample location n.

3.2 Noise quantization via possibilistic filtering

Our approach is also based on the assumption of local ergodicity. On top of that, it exploits the domination properties presented in section 2, i.e. of the fact that a possibility distribution can be seen as a family of convolution kernels.

Suppose you want to low-pass filter a signal with two different filters having the same cutoff frequency f_c. Such a filter eliminates from the input signal its component with a frequency higher than the cutoff frequency f_c (this is the explanation for the origin of the denomination "low-pass filter"). Suppose that the maximal frequency of the input signal is lower than f_c. Then the two output signals will be approximately equal. Now, suppose that we apply this same filtering procedure to an input signal having frequencies beyond f_c. Then, generally, the output signals will be different, depending on the shape of the convolution kernel.

Now, consider the same procedure with a family of low-pass filters (instead of just two). The previous remark still holds. Moreover, the dispersion in the outputs of this family of low-pass filters is a direct consequence of the high frequency level of the input signal. If we now suppose that the high frequencies of the input signal are only due to noise[4], then the dispersion in the outputs of this family of low-pass filters can be considered as a marker of the variability of the input signal.

As mentioned before, the impulse responses of the usual linear low-pass filters are convolution kernels (uniform, Gaussian filters...). Since a possibility distribution is equivalent to a family of convolution kernels, we propose to replace the usual low-pass filtering based on a convolution kernel by a possibility distribution-based low-pass filtering procedure.

[4]This is the hypothesis underlying the low-pass filters

The imprecision or the dispersion in the result of a possibility distribution-based filtering is quantified by the length of the interval $[\underline{S}_n, \overline{S}_n]$, as defined by expressions (8) and (7).

Therefore, under the assumption of local ergodicity, we propose to estimate the noise level by :

$$\lambda_n = \overline{S}_n - \underline{S}_n. \tag{12}$$

As the most usual low-pass filters have impulse responses, which are unimodal and symmetric convolution kernels around n, the triangular possibility distribution plays a central role in possibility-distribution-based filtering. Indeed, as already mentionned, the triangular possibility distribution is the most specific possibility distribution that dominates the class of all unimodal symmetric convolution kernels with the same mode and support.

In the case of image processing, i.e. with a 2D signal, the used triangular neighbourhood of each pixel can be simply represented by the possibilistic 3×3 matrix:

$$\pi_{3\times 3} = \begin{pmatrix} 0.25 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 0.25 \end{pmatrix} \tag{13}$$

In the case of 1D signal processing, the used triangular neighbourhood of each sample location can be simply represented by the vector:

$$\pi_3 = \begin{pmatrix} 0.5 \\ 1 \\ 0.5 \end{pmatrix} \tag{14}$$

In order to weaken the influence of the signal variations on the noise level estimator that we propose, we have to choose the smallest possible neighbourhood. Under a π_3 or a $\pi_{3\times 3}$ possibilistic neighbourhood, is only the Kronecker possibility distribution that is equal to 1 on the estimation's location that would have led to a canonical estimation of S_n on the location n. This is why we propose to use π_3 or $\pi_{3\times 3}$ to estimate the noise level.

We conjecture that the length of the interval-valued estimate $[\underline{S}_n, \overline{S}_n]$ obtained with $\pi_{3\times 3}$ or with π_3 is an estimate of the noise level at the location n. This conjecture is illustrated by the experiments in section 4.

4 Experiments

4.1 Simulated noise experiment

For this first experiment, we synthesized a set of noisy images from the benchmark image Lena. A Gaussian noise is simulated for standard deviations ranging from 0 to 60 and added to the original Lena image. With this set of noisy images, we can directly compare the noise level estimates presented in this paper (10), (11) and (12) with the simulated added noise.

Figure 2: images of Lena with simulated Gaussian noise with standard deviations of 0, 30 and 60.

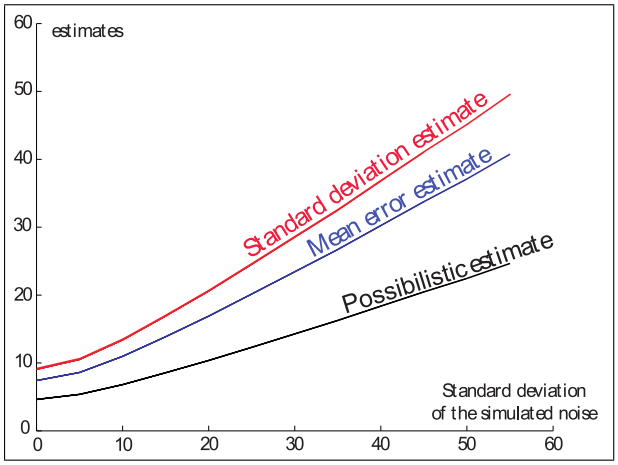


Figure 3: Usual and possibilistic local estimates of the noise level.

This experiment attempts to show the ability of the possibility distribution based approach, presented in subsection 3.2, to quantify the noise level on an image when the noise is supposed to be locally ergodic. The noise level is known and represented by the standard deviation of the added Gaussian noise.

The average over all the pixels of the noisy images of the noise level estimates (10), (11) and (12) is plotted on Figure 3 versus the level of the simulated added noise. The highest curve corresponds to the standard deviation estimate, i.e. expression (10) with a 3×3 convolution kernel, the curve in the middle, corresponds to the mean error estimate, i.e. expression (11) with a 3×3 convolution kernel and the lowest curve corresponds to the possibility distribution-based noise level estimate, i.e. expression (12).

As can be seen on Figure 3, all these estimators are good markers of the noise level, since the three plotted curves are linear functions of the noise level. The part of the curves with small simulated noise levels (i.e. with standard deviation lower than 5) is not fully in agreement with this remark. This is due to the fact that for low noise levels,

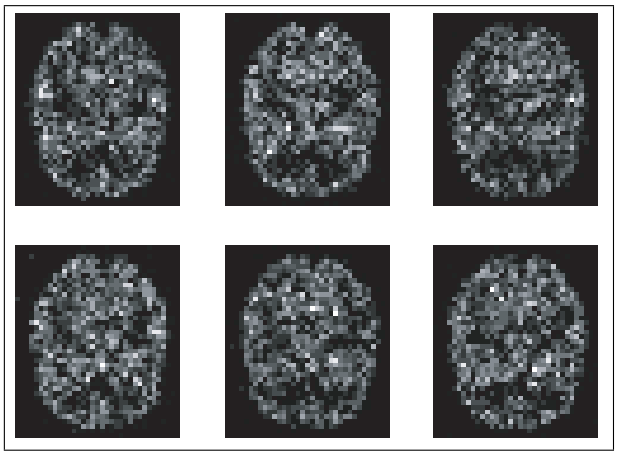


Figure 4: six images of the 1000 HBP direct acquisitions.

the signal to noise ratio is high and the observed variations of the noisy image are mainly due to the image, and not to the noise.

From this experiment, we can not pretend that our estimator is better than the other existing local estimators to quantify the noise level, since the three curves are very similar. However, put in a more general context, our approach looks more appropriate to handle the noise in further processing. In any usual method, an additional step is necessary to handle the noise in the processing. The advantage of the possibilistic approach is that noise level quantization is part of the processing (in that case the filtering) of the data without any additional computation.

4.2 Real noise experiment

A Hoffman $2D$ brain phantom (Data Spectrum Corporation) was filled with a $99m$ technetium solution ($148MBq/L$) and placed in front of one of the detectors of a dual-head gamma camera using a low-energy high-resolution parallel-hole collimator (INFINIA, General Electric Healthcare). A dynamic study was performed to provide 1000 planar acquisitions (acquisition time: 1 second; average count per image 1.5 kcounts, 128×128 images to satisfy the Shannon condition), representing 1000 measures of a random $2D$ image supposedly ruled by a Poisson process.

The acquisition time being very short, the images are very noisy, i.e. the signal to noise ratio is very low. More precisely, the average pixel value in the brain corresponds to a coefficient of variation of the Poisson noise of 69%. $I_{n,p}$ is the measured activity of the n^{th} pixel within the p^{th} acquired image. Note that Figure 4 only shows the 40×35 central parts of the images that contains the HBP projection.

This experiment attempts to show that the possibility distribution-based noise level estimator (12) is more correlated to the statistical variations of the image than the standard deviation noise estimation approach.

The randomness of the radioactive decay being statistically described by the Poisson probability, it cannot really be assumed to be stationary all over the image. Since the signal to noise ratio is very low, the local variation of the activity level, in the neighbourhood of each pixel, is still highly correlated with the statistical variations due to acquisition noise.

On the one hand, the statistical variation of the activity of the n^{th} pixel can be estimated by its standard deviation σ_n all over its different realizations:

$$\sigma_n = \sqrt{\frac{1}{999}\sum_{p=1}^{1000}(I_{n,p} - m_n)^2}, \tag{15}$$

with m_n, the weighted mean of the image at the n^{th} pixel:

$$m_n = \frac{1}{1000}\sum_{p=1}^{1000} I_{n,p}. \tag{16}$$

On the other hand, the local variation of the measurement in the neighbourhood of the n^{th} pixel within the p^{th} image can be estimated by computing the standard deviation via the expression (10) with a highly specific kernel (the same experiment made with expression (11) led to similar results). In this experiment, we propose two estimates of this standard deviation: $\gamma_{n,p}$ is computed by using a 3×3 uniform neighbourhood, and $\delta_{n,p}$ is computed by using a Gaussian kernel with a standard deviation equal to 1.6, i.e. a kernel whose bandwidth has been adapted to equal the bandwidth of the uniform kernel [20, 31].

In the meantime, we compute, for each image, an interval valued activity $[\underline{I}_{n,p}, \overline{I}_{n,p}]$ by using the possibility distribution based method described in subsection 3.2. The local variation in the neighbourhood of the n^{th} pixel within the p^{th} image is estimated by the length $\lambda_{n,p}$ of each interval:

$$\lambda_{n,p} = \overline{I}_{n,p} - \underline{I}_{n,p}. \tag{17}$$

We aim at testing whether the distribution of the estimated standard deviation σ_n is correlated or not with $\gamma_{n,p}$, $\delta_{n,p}$ and $\lambda_{n,p}$. To provide a clear illustration, we compute, for each n, the mean of the distributions of the deviation measures: $\tilde{\gamma}_n = \frac{1}{1000}\sum_{p=1}^{1000}\gamma_{n,p}$, $\tilde{\delta}_n = \frac{1}{1000}\sum_{p=1}^{1000}\delta_{n,p}$ and $\tilde{\lambda}_n = \frac{1}{1000}\sum_{p=1}^{1000}\lambda_{n,p}$.

Figure 5 plots $\tilde{\gamma}_n$ versus σ_n, as well as the straight line of equation $\sigma_n = \tilde{\gamma}_n$, figure 6 plots $\tilde{\delta}_n$ versus σ_n, as well as the straight line of equation $\sigma_n = \tilde{\delta}_n$ and figure 7 plots $\tilde{\lambda}_n$ versus σ_n, as well as the straight line of equation $\sigma_n = \tilde{\lambda}_n$.

These figures clearly show that all these estimations are, on average, correlated with σ_n. The choice of the value

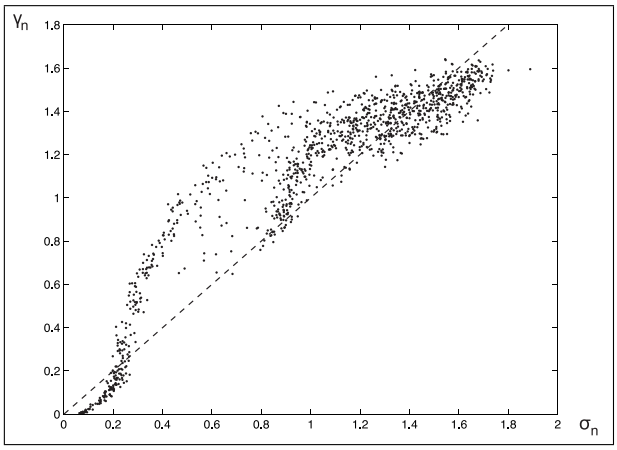


Figure 5: local variation measured by using a 3×3 uniform kernel versus the statistical variation.

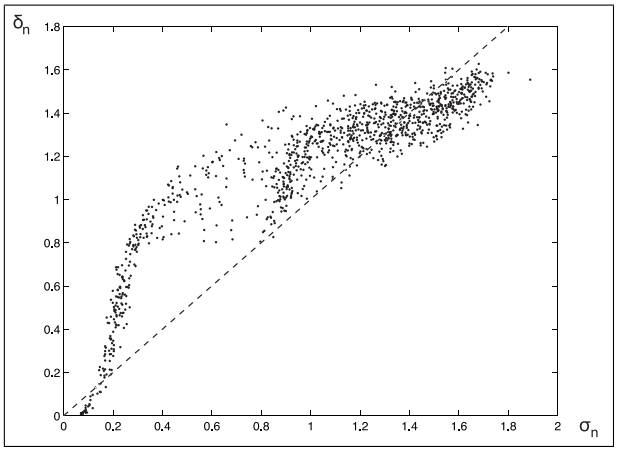


Figure 6: local variation measured by using a Gaussian kernel with a 1.6 standard deviation versus the statistical variation.

1.6 for the Gaussian kernel is appropriate since the estimated local standard deviations $\tilde{\delta}_n$ are in the same range as the statistical standard deviations σ_n. Indeed, the points $(\sigma_n, \tilde{\delta}_n)$ are close to the straight line $\sigma_n = \tilde{\delta}_n$. Actually for values smaller than 1.6, nothing is caught by the Gaussian neighbourhood for this estimation, whereas for greater values, the estimation depends more on the signal than on the variability. The same remarks can be made about the choice of the size of the uniform kernel that seems to be appropriate. When comparing Figure 7 with both Figure 5 and 6, it can be seen that the range of $\tilde{\lambda}_n$ is slightly higher than the range of $\tilde{\gamma}_n$ and $\tilde{\delta}_n$. This is due to the fact that the measure $\tilde{\lambda}_n$ is just correlated to the noise level and is not an estimation of the standard deviation.

To objectively compare those three dispersion measures, we compute three correlation coefficients: Pearson, Spearman and Kendall. As can be seen in Table 1, the three averaged variability measures $\tilde{\gamma}_n$, $\tilde{\delta}_n$ and $\tilde{\lambda}_n$ are highly

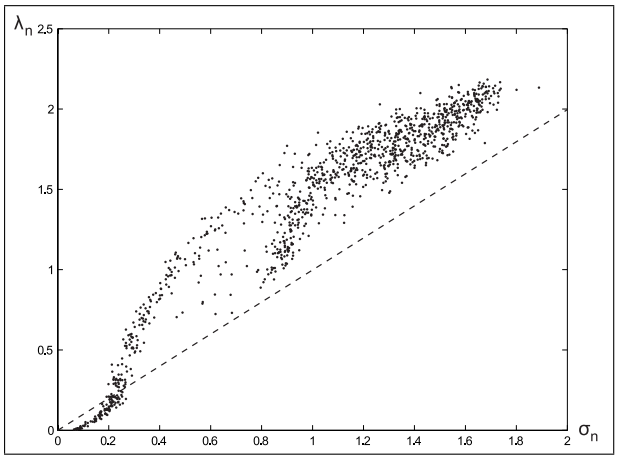


Figure 7: local variation measured by the length of the interval provided by the possibility distribution based method versus the statistical variation.

	$\gamma_{n,p}$	$\tilde{\gamma}_n$	$\delta_{n,p}$	$\tilde{\delta}_n$	$\lambda_{n,p}$	$\tilde{\lambda}_n$
Pearson	0.70	0.93	0.64	0.90	0.71	0.96
Spearman	0.64	0.92	0.63	0.90	0.67	0.95
Kendall	0.47	0.77	0.47	0.75	0.51	0.81

Table 1: Correlation coefficients between the statistical standard deviation and the different measures of dispersion.

correlated with σ_n. The correlations between σ_n and the variability measures $\gamma_{n,p}$, $\delta_{n,p}$ and $\lambda_{n,p}$ are lower but are sufficient to show a dependency between these measures and the statistical variations of the set of images. We can notice that $\lambda_{n,p}$ is always more correlated with σ_n than the other variability measures $\gamma_{n,p}$ and $\delta_{n,p}$. The same remark is also true for $\tilde{\gamma}_n$, $\tilde{\delta}_n$ and $\tilde{\lambda}_n$. We can conclude that, in this experiment, the possibilistic approach that we propose seems to better quantify the noise level than the usual local approach.

As we conjecture that $\lambda_{n,p}$ could be regarded as a spread factor measuring the local noise level, we expect that two intervals $[\underline{I}_{n,p}, \overline{I}_{n,p}]$ and $[\underline{I}_{n,q}, \overline{I}_{n,q}]$ intersect for most of pairs (p, q) of images. We propose to complete this experimentation, by computing, for each pixel n, the ratio of the intervals that intersect versus the total number of tested intervals. We compute the same ratio using $\gamma_{n,p}$ and $\delta_{n,p}$ considered as spread factors measuring statistical standard deviations: we then test each couple of intervals $[I_{n,p}-3\gamma_{n,p}, I_{n,p}+3\gamma_{n,p}]$ and $[I_{n,p}-3\delta_{n,p}, I_{n,p}+3\delta_{n,p}]$. Since the 3σ interval is usually assumed to be the 99% confidence interval, one can expect a high rate of overlapping. Table 2 presents the average ratio for all the pixels of the image and for only the pixels with a value greater than three.

	with all pixels	only with pixels such that $I_n > 3$
Uniform kernel	0.11	0.88
gaussian kernel	0.13	0.89
possibility distribution	0.98	0.92

Table 2: Ratio of intersecting confidence intervals.

As can be seen easily on Table 2, the possibility distribution based confidence interval fulfils a 98% intersecting intervals while the usual probabilistic based confidence intervals are far from this 99% ratio. The bad ratio of the other methods is mainly due to the fact that the spread factor is underestimated by these methods for low values (as it can be easily seen on Figures 5 to 7). In fact, selecting only the pixels whose level always exceeds a certain level over the different realizations increases the score of the probabilistic based methods. In fact, by assuming that the measured values are Poisson distributed, a local Gaussian approximation can be valid except for small values of the illumination signal.

5 Conclusion

In this article, we have presented a method for quantifying the noise level at each sample location of a signal. This method is based on replacing the conventional probabilistic by a possibilistic filtering approach. One of the main advantage of this method is the fact that nothing has to be assumed on the nature of the noise except its local ergodicity. Moreover, when a possibilistic approach is used in signal processing, the noise estimation is propagated all along the different steps of the algorithm by the model itself, which is an advantage compared to usual kernel based approaches, where the noise estimation requires a parallel computation.

6 Acknowledgment

The authors would like to thank Dr Mariano-Goulard for providing them the data used in the experiments.

References

[1] S. Baker and I. Matthews, Lucas-Kanade 20 years on: a unifying framework, *International Journal on Computer Vision*, vol. 56(3), 2004, pp. 221255.

[2] I. Bloch and H. Maitre, Fuzzy mathematical morphologies: A comparative study, *Pattern Recognition*, vol. 28, 1995, pp. 1341-1387.

[3] I. Bloch and H. Maitre, Fusion in image processing, In: *Information Fusion in Signal and Image Processing, I. Bloch (Ed.), Wiley*, 2008.

[4] J.P. Chilès et P. Delfiner, Geostatistics (Modeling Spatial Uncertainty), *Wiley, New-York, U.S.A.*, 1999.

[5] B.R. Corner, R.M. Narayanan, and S.E. Reichenbach, Noise estimation in remote sensing imagery using data masking, *International Journal of Remote Sensing*, vol. 24, 2003, pp. 689-702.

[6] G. de Cooman, Integration and conditioning in numerical possibility theory, *Annals of Mathematics and Artificial Intelligence*, vol. 32, 2001, pp. 87-123.

[7] D. Denneberg, Non-Additive Measure and Integral, *Kluwer Academic Publishers*, 1994.

[8] D. Dubois, Possibility theory and statistical reasoning, *Computational Statistics and Data Analysis*, vol. 51, 2006, pp. 47-69.

[9] D. Dubois and H. Prade, Possibility theory: an approach to computerized processing of uncertainty, *Plenum Press*, 1988.

[10] D. Dubois, H. Prade, and S. Sandri, On possibility/probability transformations, *Proceedings of Fourth IFSA Conference, Kluwer Academic Publ*, 1993, pp. 103-112.

[11] D. Dubois, H. Prade, and P. Smets, New semantics for quantitative possibility theory, *ISIPTA01, 2nd International Symposium on Imprecise Probabilities and Their Applications, Ithaca, New York, USA, June*, 2001.

[12] D. Dubois, H. Prade, and P. Smets, A definition of subjective possibility, *International Journal of Approximate Reasoning*, vol. 48, 2008, pp. 352-364.

[13] D. Dubois, L. Foulloy, G. Mauris, and H. Prade, Probability-Possibility Transformations, Triangular Fuzzy Sets, and Probabilistic Inequalities, *Reliable Computing*, vol. 10, 2004, pp. 273-297.

[14] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, Learning low-level vision, *International Journal on Computer Vision*, vol. 40(1), 2000, pp. 2547.

[15] P. Goovaerts, Geostatistics for natural resources evaluation, *Oxford University Press New York*, 1997.

[16] G.E. Healey and R. Kondepudy, Radiometric CCD camera calibration and noise estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, 1994, pp. 267-276.

[17] F. Jacquey, K. Loquin, F. Comby, and O. Strauss, Nonadditive approach for gradient-based edge detection, *ICIP07, Int. Conf. on Image Processing, San Antonio, Texas, USA, September*, 2007, pp. 16-19.

[18] C. Liu, W.T. Freeman, R. Szeliski, and S.B. Kang, Noise Estimation from a Single Image, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[19] K. Loquin, De l'utilisation des noyaux maxitifs en traitement de l'information, *PhD report, LIRMM, Université Montpellier II*, 2008

[20] K. Loquin and O. Strauss, On the granularity of summative kernels, *Fuzzy Sets and Systems*, vol. 159, 2008, pp. 1952-1972.

[21] K. Loquin, and O. Strauss, Imprecise functional estimation: the cumulative distribution case, *SMPS08, Soft Methods in Probability and Statistic*, Toulouse, 2008.

[22] D. Lowe, Object recognition from local scale-invariant features, *In Proc. IEEE Intl Conf. Computer Vision*, 1999, pp. 11501157.

[23] C. Manders and S. Mann, Digital Camera Sensor Noise Estimation from Different Illuminations of Identical Subject Matter, *Proceedings of the Fifth International Conference on Information, Communications and Signal Processing*, 2005, pp. 1292-1296.

[24] B.P. Marchant and R.M. Lark, Robust estimation of the variogram by residual maximum likelihood, *Geoderma*, vol. 140, 2007, pp. 62-72.

[25] S.I. Olsen, Estimation of noise in images: an evaluation, *CVGIP: Graphical Models and Image Processing*, vol. 55, 1993, pp. 319-323.

[26] P. Perona and J. Malik, Scale-space and edge detection using anisotropic diffusion, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12(7), 1990, pp. 629639.

[27] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, Image denoising using scale mixtures of Gaussians in the wavelet domain, *IEEE Transactions on Image Processing*, vol. 12(11), 2003, pp. 13381351.

[28] R.M. Rangayyan, M. Ciuc, and F. Faghih, Adaptive-neighborhood filtering of images corrupted by signal-dependent noise, *Applied Optics*, vol. 37, 1998, pp. 4477-4487.

[29] M. Rombaut, Fusion in robotics, In: *Information Fusion in Signal and Image Processing, I. Bloch (Ed.), Wiley*, 2008.

[30] D. Schmeidler, Integral representation without additivity, *Proceedings of the American Mathematical Society*, vol. 97, 1986, pp. 255261.

[31] J.S. Simonoff, Smoothing Methods in Statistics, *Springer*, 1996.

[32] P. Walley, Statistical reasoning with imprecise probabilities, *Chapman and Hall*, 1991.

[33] P. Walley, Towards a unified theory of imprecise probability, *International Journal of Approximate Reasoning*, vol. 24, 2000, pp. 125-148.

[34] R. Zhang, P. Tsai, J. Cryer, and M. Shah, Shape from shading: A survey, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21(8), 1999, pp.690706.

Nonparametric Predictive Multiple Comparisons with Censored Data and Competing Risks

Tahani A. Maturi
Dept of Mathematical Sciences,
Durham University, UK
tahani.maturi@durham.ac.uk

Pauline Coolen-Schrijner
Dept of Mathematical Sciences,
Durham University, UK

Frank P.A. Coolen
Dept of Mathematical Sciences,
Durham University, UK
frank.coolen@durham.ac.uk

Abstract

This paper provides an overview of nonparametric predictive inference for comparison of multiple groups of data including right-censored observations. Different right-censoring schemes discussed are early termination of an experiment, progressive censoring and competing risks. Theoretical results are briefly stated, detailed justifications are presented elsewhere. The methods are illustrated and discussed via examples with data from the literature.

Keywords. Competing risks, early termination, nonparametric predictive inference, precedence testing, progressive censoring, right-censored data.

1 Introduction

This paper presents a brief overview of recent results on nonparametric predictive inference (NPI) for multiple comparisons in situations with right-censored observations. Such data typically occur in reliability or survival analysis, due to several reasons. For example, when interest is in a specific failure mode for a technical unit, it may fail due to a different failure cause. If multiple failure modes are of interest, and failure will be due to only a single failure mode, then this situation is known as "competing risks", where an observed failure time is actually a right-censoring time with regard to all failure modes that did not cause the failure. Another reason for right-censoring may be removal of units from a lifetime experiment, normally to save time or reduce cost, but this also occurs if, at some point, one wishes to study units which have not yet failed in an experiment in more detail. If right-censoring is due to an experiment being terminated before all units have failed, multiple comparisons of different groups of units based on such data is known as "precedence testing". If non-failing units are removed from the experiment at several possible stages it is known as "progressive censoring". Recently, we have developed NPI for multiple comparisons for precedence testing, progressive censoring, and competing risks, and these results are briefly presented here and illustrated and discussed via examples. Detailed justifications of the results are presented elsewhere. It should be emphasized that, throughout the paper, unspecified reasons for right-censoring are assumed to be based on processes that are independent of the residual lifetimes of the censored units.

NPI is a statistical method that aims at using relatively few modelling assumptions, it uses lower and upper probabilities to quantify uncertainty. Some basic applications of NPI in reliability were summarized by Coolen, *et al* [12], recently a variety of further applications in this area have been presented, including probabilistic safety assessment if zero failures have been observed [7], prediction of not-yet occurred failure modes [8], comparison of success-failure data [17], and system reliability with optimal redundancy allocation [18]. NPI has also been developed for replacement problems, with specific attention to age replacement of technical units [19, 21]. Imprecise probabilistic methods are attractive in reliability, as their flexibility for dealing with limited information is a particular advantage for dealing with practical aspects of many reliability situations. Utkin and Coolen [35] present an extensive overview of the literature, for a concise overview see [13].

In Section 2 of this paper NPI is briefly introduced, followed in Section 3 by explanation of the way in which NPI deals with right-censored data. Recent developments of NPI for multiple comparisons with the different right-censoring schemes discussed above are presented in Sections 4, 5 and 6, and illustrated and discussed in examples in Section 7. The same notation is used for different quantities in Sections 4-6, but in the general NPI approach to multiple comparisons they all relate to similar concepts, just the interpretations in the specific applications are different per section.

2 Nonparametric predictive inference

Nonparametric predictive inference (NPI) is based on Hill's assumption $A_{(n)}$ [25], which implies direct (lower and upper) probabilities for a future observable random quantity, based on observed values of n related random quantities [6]. Suppose that $X_1, \ldots, X_n, X_{n+1}$ are positive, continuous and exchangeable random quantities representing lifetimes. Let the ordered observed values of $X_1, \ldots, X_n$ be denoted by $x_1 < x_2 < \ldots < x_n < \infty$, and let $x_0 = 0$ and $x_{n+1} = \infty$ for ease of notation, note that the latter is not considered to be an observation for X_{n+1}. We assume that no ties occur, our results can be generalised to allow ties [26]. For positive X_{n+1}, representing a future observation, based on n observations, $A_{(n)}$ assigns $P(X_{n+1} \in (x_i, x_{i+1})) = 1/(n+1)$ for $i = 0, 1, \ldots, n$. $A_{(n)}$ does not assume anything else, and is a post-data assumption related to exchangeability [22]. Hill [24] discusses $A_{(n)}$ in detail, and he also provided a Bayesian justification for $A_{(n)}$ under finite additivity [26]. Inferences based on $A_{(n)}$ can be considered suitable if there is hardly any knowledge about the random quantity of interest, other than the n observations, or if one does not want to use such information. $A_{(n)}$ is not sufficient to derive precise probabilities for many events of interest, but it provides bounds for probabilities via the 'fundamental theorem of probability' [22], which are lower and upper probabilities in interval probability theory [36, 37]. NPI has strong consistency properties within the theory of interval probability [1], attractive frequentist properties, and compares favourably to objective Bayesian methods [6, 24].

3 NPI for right-censored data

Coolen and Yan [16] presented rc-$A_{(n)}$ as a generalization of $A_{(n)}$ for right-censored data, using the additional assumption that, at a moment of censoring, the residual lifetime of a right-censored unit is exchangeable with the residual lifetimes of all other units that have not yet failed or been censored.

Suppose that there are n observations consisting of u event times, $x_1 < x_2 < \ldots < x_u$, and $v(= n - u)$ right-censored observations, $c_1 < c_2 < \ldots < c_v$. Let $x_0 = 0$ and $x_{u+1} = \infty$, and suppose that there are s_i right-censored observations in the interval (x_i, x_{i+1}) at times $c_1^i < c_2^i < \ldots < c_{s_i}^i$, where $\sum_{i=0}^{u} s_i = v$. These data can also be denoted by pairs (t_i, δ_i) for $i = 1, \ldots, n$, where $t_i = x_i$ (so a failure time, or time of other actual event of interest) if $\delta_i = 1$ and $t_i = c_i$ (a right-censored observation) if $\delta_i = 0$. For ease of notation, let $(t_0, \delta_0) = (0, 1)$ and $x_{n+1} = \infty$. The assumption rc-$A_{(n)}$ partially specifies the probability distribution for X_{n+1} by the following M-functions [16], for $i = 1, \ldots, n$:

$$M_{X_{n+1}}(t_i, x_{i+1}) = \frac{1}{n+1}(\tilde{n}_{t_i})^{\delta_i - 1} \prod_{\{r: c_r < t_i\}} \frac{\tilde{n}_{c_r} + 1}{\tilde{n}_{c_r}} \quad (1)$$

where $\tilde{n}_{c_r}$ and $\tilde{n}_{t_i}$ are the numbers of units in the risk set (i.e. that have not yet failed or been censored) just prior to time c_r and t_i, respectively. These M-functions are basic probability assignments in the sense of Shafer [33], and lead to the following precise probabilities for X_{n+1} to be between two consecutive observed failure times x_i and x_{i+1},

$$P(X_{n+1} \in (x_i, x_{i+1})) = \frac{1}{n+1} \prod_{\{r: c_r < x_{i+1}\}} \frac{\tilde{n}_{c_r} + 1}{\tilde{n}_{c_r}} \quad (2)$$

Coolen and Yan [15] developed NPI for comparison of two groups of lifetime data including right-censored observations. By applying the appropriate rc-$A_{(n)}$ assumption for each group, their method is based on comparing the next observation from each group, say X_{n_x+1} and Y_{n_y+1}. The NPI lower and upper probabilities for the event that $X_{n_x+1} < Y_{n_y+1}$ are

$$\underline{P} = \sum_{i=0}^{u_x} \sum_{j=0}^{n_y} 1(x_{i+1} < t_{y,j}) P^X(x_i, x_{i+1}) M^Y(t_{y,j}, y_{j+1})$$

$$\overline{P} = \sum_{i=0}^{n_x} \sum_{j=0}^{u_y} 1(t_{x,i} < y_{j+1}) P^Y(y_j, y_{j+1}) M^X(t_{x,i}, x_{i+1})$$

where $M^X(t_{x,i}, x_{i+1})$, $M^Y(t_{y,j}, y_{j+1})$, $P^X(x_i, x_{i+1})$ and $P^Y(y_j, y_{j+1})$ are as given by (1) and (2), and $1(A)$ is the indicator function that equals 1 if A is true and 0 else. Coolen and Yan [15] did not consider situations with more than two groups, nor the effect of early termination of the lifetime experiment or the specific features of progressive censoring and competing risks. NPI for multiple comparisons for real-valued data without right-censored observations was presented in [14], and NPI multiple comparisons for Bernoulli data in [11].

4 Early termination of experiment

In some circumstances, mostly in order to save costs or time, an experiment to compare lifetimes of units in different groups may be terminated before all units have failed. We assume that all units are placed simultaneously on a lifetime experiment which is terminated at a certain specified time, which may also be the moment a specified number of failures have occurred. The situation where for all units failing before the moment of termination of the experiment the lifetimes are observed, is also known as *precedence testing*

in the literature [3]. Coolen-Schrijner *et al* [20] presented NPI for comparison of two groups of lifetime data with early termination of the experiment, say at time T_0, and they illustrated the effect of varying T_0. The resulting data set contains, for each of the two groups in the experiment, failure times prior to T_0 and right-censored observations at T_0 for all units that do not fail before T_0. Maturi *et al* [28] extend this to more than two groups, with a variety of inferential goals for the multiple comparisons in line with different goals as presented in the statistical selection literature [4]. Maturi *et al* [30] present further generalized results, which also generalize the results by Coolen and Yan [15], by developing NPI for comparison of multiple groups of lifetime data including right-censored observations, and with possible early termination of the experiment.

Consider an experiment to compare lifetimes of units from $k \geq 2$ groups, which are assumed to be fully independent, with the experiment starting on all units at time 0. The experiment can be terminated before all units have failed, say at time T_0. This T_0 can be fixed or random, but it is essential that it is assumed not to hold any information on residual time-to-failure for units that have not yet failed. We also allow non-informative right-censoring to occur for some units before the experiment is stopped. For group j, $j = 1, \ldots, k$, n_j units are in the experiment, of which u_j units fail before (or at) T_0, with ordered failure times $0 < x_{j,1} < x_{j,2} < \ldots < x_{j,u_j} \leq T_0$, and with right-censoring times $c_{j,1} < c_{j,2} < \ldots < c_{j,v_j} < T_0$. Let $x_{j,0} = 0$ and $x_{j,u_j+1} = \infty$ $(j = 1, \ldots, k)$, and let s_{j,i_j} be the number of right-censored observations in the interval (x_{j,i_j}, x_{j,i_j+1}), with $x_{j,i_j} < c_{j,1}^{i_j} < c_{j,2}^{i_j} < \ldots < c_{j,s_{j,i_j}}^{i_j} < x_{j,i_j+1}$ and $\sum_{i_j=0}^{u_j} s_{j,i_j} = v_j$, so $n_j - (u_j + v_j)$ units from group j are right-censored at T_0.

For NPI with data containing right-censored observations, and with early termination of the experiment at time T_0, the assumption rc-$A_{(n_j)}$ implies that the following M-function values apply for a nonnegative random quantity X_{j,n_j+1}, on the basis of data consisting of u_j failure times and $(n_j - u_j)$ right-censored observations:

$$M_{i_j}^j = M_{X_{j,n_j+1}}(x_{j,i_j}, x_{j,i_j+1}) = \frac{1}{n_j+1} \prod_{\{r:c_r<x_{j,i_j}\}} \frac{\tilde{n}_{j,c_r}+1}{\tilde{n}_{j,c_r}}$$

$$M_{i_j,a_j}^j = M_{X_{j,n_j+1}}(c_{j,a_j}^{i_j}, x_{j,i_j+1}) = \frac{(\tilde{n}_{j,c_{j,a_j}^{i_j}})^{-1}}{n_j+1} \prod_{\{r:c_r<c_{j,a_j}^{i_j}\}} \frac{\tilde{n}_{j,c_r}+1}{\tilde{n}_{j,c_r}}$$

$$M_{T_0}^j = M_{X_{j,n_j+1}}(T_0, \infty) = \frac{n_j - (u_j+v_j)}{n_j+1} \prod_{\{r:c_r<T_0\}} \frac{\tilde{n}_{j,c_r}+1}{\tilde{n}_{j,c_r}}$$

where $i_j = 0, \ldots, u_j$, $a_j = 1, \ldots, s_{j,i_j}$, and $\tilde{n}_{j,c_r}$ and $\tilde{n}_{j,c_{j,a_j}^{i_j}}$ are the number of units from group j in the risk set just prior to time c_r and $c_{j,a_j}^{i_j}$, respectively. Also

$$P_{i_j}^j = P(X_{j,n_j+1} \in (x_{j,i_j}, x_{j,i_j+1})) = \frac{1}{n_j+1} \prod_{\{r:c_r<x_{j,i_j+1}\}} \frac{\tilde{n}_{j,c_r}+1}{\tilde{n}_{j,c_r}}$$

$$P_{T_0}^j = P(X_{j,n_j+1} \in (T_0, \infty)) = M_{X_{j,n_j+1}}(T_0, \infty) = M_{T_0}^j$$

The NPI lower and upper probabilities for the event that the next observed lifetime from group l is the maximum of all next observed lifetimes for the k groups in the experiment, i.e. $X_{l,n_l+1} = \max_{1 \leq j \leq k} X_{j,n_j+1}$, are

$$\begin{aligned} \underline{P}^{(l)} &= \sum_{i_l=0}^{u_l} \left\{ \prod_{\substack{j=1 \\ j \neq l}}^{k} \left[\sum_{i_j=0}^{u_j} 1(x_{j,i_j+1} < x_{l,i_l}) P_{i_j}^j \right] M_{i_l}^l \right. \\ &+ \left. \sum_{a_l=1}^{s_{l,i_l}} \prod_{\substack{j=1 \\ j \neq l}}^{k} \left[\sum_{i_j=0}^{u_j} 1(x_{j,i_j+1} < c_{l,a_l}^{i_l}) P_{i_j}^j \right] M_{i_l,a_l}^l \right\} \\ &+ M_{T_0}^l \prod_{\substack{j=1 \\ j \neq l}}^{k} \sum_{i_j=0}^{u_j} 1(x_{j,i_j+1} < T_0) P_{i_j}^j \end{aligned} \quad (3)$$

$$\begin{aligned} \overline{P}^{(l)} &= \sum_{i_l=0}^{u_l} P_{i_l}^l \prod_{\substack{j=1 \\ j \neq l}}^{k} \left\{ \sum_{i_j=0}^{u_j} 1\left(x_{j,i_j} < x_{l,i_l+1}\right) M_{i_j}^j \right. \\ &+ \sum_{i_j=0}^{u_j} \sum_{a_j=1}^{s_{j,i_j}} 1(c_{j,a_j}^{i_j} < x_{l,i_l+1}) M_{i_j,a_j}^j \\ &+ \left. 1\left(T_0 < x_{l,i_l+1}\right) M_{T_0}^j \right\} + P_{T_0}^l \end{aligned} \quad (4)$$

If the experiment is not terminated before the event times of all units have been observed, so for each unit either the failure time or a right-censoring time not due to the experiment ending, then the terms including T_0 in formulae (3) and (4) disappear, and we get a generalization of the results by Coolen and Yan [15], who only considered NPI for comparison of two groups of lifetime data including right-censored observations. Another special case occurs if there are no right-censored observations before T_0. In this case our method generalizes the results by Coolen-Schrijner *et al* [20], who considered NPI for comparison of two groups with early termination of the experiment, but without earlier right-censoring.

At any value of T_0, we can state that the data provide a strong indication that group l is the best if

$\underline{P}^{(l)} > \overline{P}^{(j)}$ for all $j \neq l$. It might seem attractive to state that, if $\underline{P}^{(l)} > \underline{P}^{(j)}$ and $\overline{P}^{(l)} > \overline{P}^{(j)}$ for all $j \neq l$, there would be a weak indication that group l is the best. The difference between the upper and lower probabilities reflects the amount of information available, it decreases if more relevant information becomes available. A typical feature of NPI for these methods with the experiment terminated at T_0 is that, if T_0 is increased, the upper (lower) probability never increases (decreases), while its value can only change at observed event times.

5 Progressive censoring

Maturi *et al* [29] considered the comparison of two groups, say X and Y, in which progressive censoring schemes are applied for one or both groups. They allow several such censoring schemes, known in the literature as progressive Type-I censoring, progressive Type-II censoring and Type-II progressively hybrid censoring scheme [2]. The main characteristic of progressive censoring is that, at several stages some units are randomly removed from the experiment. For NPI for a progressive Type-II censoring scheme with $R = (R_1, R_2, \ldots, R_r)$, where R_i is the number of units that are removed from the experiment at the ith failure, the assumption rc-$A_{(n)}$ implies that the probability distribution for a nonnegative random quantity X_{n+1} on the basis of data including r real and $n-r$ progressively censored observations, is partially specified by the following M-function values, for $i = 0, 1, \ldots, r$,

$$M^X(x_i, x_{i+1}) = \frac{1}{n+1} \prod_{k=1}^{i-1} \frac{n-k-\sum_{l=1}^{k-1} R_l + 1}{n-k-\sum_{l=1}^{k} R_l + 1} \quad (5)$$

$$M^X(x_i^+, x_{i+1}) = \frac{R_i}{n-i-\sum_{l=1}^{i} R_l + 1} \, M^X(x_i, x_{i+1}) \quad (6)$$

where x_i^+ represents the lower bound for the interval that contains the set of censored units at x_i, $x_0 = 0$ and $x_{r+1} = \infty$. The corresponding NPI probabilities for X_{n+1} to be in (x_i, x_{i+1}) are

$$P^X(x_i, x_{i+1}) = \frac{1}{n+1} \prod_{k=1}^{i} \frac{n-k-\sum_{l=1}^{k-1} R_l + 1}{n-k-\sum_{l=1}^{k} R_l + 1} \quad (7)$$

Suppose that we have two independent groups, X and Y, for which n_x and n_y units, respectively, are placed on a lifetime experiment. Both groups are progressively Type-II censored with the schemes $R^x = (R_1^x, R_2^x, \ldots, R_{r_x}^x)$ and $R^y = (R_1^y, R_2^y, \ldots, R_{r_y}^y)$. Given the data, R^x, R^y, and the assumptions rc-$A_{(n_x)}$ and rc-$A_{(n_y)}$, the NPI lower and the upper probabilities that the next observation from group Y is greater than the next observation from group X, are

$$\underline{P} = \sum_{j=0}^{r_y} \sum_{i=0}^{r_x} 1(x_{i+1} < y_j) P^X(x_i, x_{i+1}) P^Y(y_j, y_{j+1}) \quad (8)$$

$$\overline{P} = \sum_{j=0}^{r_y} \sum_{i=0}^{r_x} 1(x_i < y_{j+1}) P^X(x_i, x_{i+1}) P^Y(y_j, y_{j+1}) \quad (9)$$

We refer to [29] for NPI comparisons in case of progressive Type-I and Type-II progressively hybrid censoring. It should be emphasized that, in classical frequentist methods for such comparisons [2], via hypothesis tests of assumed equality of underlying lifetime distributions, the details of the exact applied censoring scheme are relevant, as they influence the counter-factuals, outcomes of the experiment that were possible but did not occur. In NPI such counter-factuals play no role, as the comparison is directly based on random quantities representing lifetimes of one future unit per group. The different censoring schemes affect the M-function values, but the corresponding derivations of the lower and upper probabilities of interest is similar in all cases.

6 Competing risks

In competing risks, a unit is subject to failure from one of k distinct failure modes. Throughout we assume that these failure modes are independent. Tsiatis [34] showed that competing risks data as considered here do not hold information about dependence of failure modes. We assume that the unit fails due to the first occurrence of a failure caused by one of the possible failure modes, at which moment it is withdrawn from further use. We suppose that such failure observations are obtained for n units, and that failure modes causing failures are known with certainty. As is common in study of failure data under competing risks, we consider for each unit k random quantities, say T_i for $i = 1, \ldots, k$, where T_i represents the unit's time to failure under the condition that failure occurs due to failure mode i. We assume that these T_i are independent continuous random quantities, and the failure time of the unit is $T = \min(T_1, \ldots, T_k)$. Therefore, for each unit considered we can have one failure time and we will know, with certainty, the failure mode that caused the failure. Hence, for the T_i corresponding to the other failure modes, which did not cause the failure of the unit, the unit's observed failure time is a right-censoring time.

For the NPI approach, let the failure time of a future item be denoted by X_{n+1}, and let the corresponding notation for the failure time including indication of

the actual failure mode, say failure mode j, be $X_{j,n+1}$ (so X_{n+1} corresponds to an observation T for unit $n+1$, and $X_{j,n+1}$ to T_j, according to the notation in the previous paragraph). As we assume independence between the different failure modes, our competing risk data per failure mode consist of (possibly) a number of observed failure times for failures caused by the specific failure mode considered, and right-censoring times for failures caused by other failure modes. Hence we can apply rc-$A_{(n)}$ per failure mode j, for inference on $X_{j,n+1}$. Let the number of failures caused by failure mode j be u_j and let $v_j (= n - u_j)$ be the number of the right-censored observations corresponding to failure mode j. It should be emphasized that we do not assume that each unit considered must actually fail, if a unit does not fail then there will be a right-censored observation recorded for this unit for each failure mode, as we assume that the unit will then be withdrawn from the study, or the study ends, at some point. The random quantity representing the failure time of the next unit, with all k failure modes considered, is $X_{n+1} = \min\limits_{1 \leq j \leq k} X_{j,n+1}$.

For failure mode j, $j = 1, \ldots, k$, we have as data n pairs $(t_{j,i_j}, \delta_{j,i_j})$, for $i_j = 1, \ldots, n$, where $\delta_{j,i_j} = 1$ if a failure at time $t_{j,i_j} (= x_{j,i_j})$ was caused by failure mode j and where $\delta_{j,i_j} = 0$ denotes that the event at the corresponding time $t_{j,i_j} (= c_{j,i_j})$ is, for as far as this specific failure mode j is concerned, a right-censored observation.

We can specify the NPI M-functions for $X_{j,n+1}$ ($j = 1, \ldots, k$), similar to (1), as

$$M^j_{t_{j,i_j}} = M^j(t_{j,i_j}, x_{j,i_j+1}) = \frac{(\tilde{n}_{t_{j,i_j}})^{\delta_{j,i_j}-1}}{(n+1)} \prod_{\{r: c_r < t_{j,i_j}\}} \frac{\tilde{n}_{c_r}+1}{\tilde{n}_{c_r}} \tag{10}$$

with $\tilde{n}_{c_r}$ and $\tilde{n}_{t_{j,i_j}}$ the numbers of units in the risk set just prior to times c_r and t_{j,i_j}, respectively. The corresponding NPI probabilities, similar to (2), are

$$P^j = P^j(x_{j,i_j}, x_{j,i_j+1}) = \frac{1}{n+1} \prod_{\{r: c_r < x_{j,i_j+1}\}} \frac{\tilde{n}_{c_r}+1}{\tilde{n}_{c_r}} \tag{11}$$

where x_{j,i_j} and x_{j,i_j+1} are two consecutive observed failure times caused by failure mode j (and $x_{j,0} = 0$, $x_{j,n+1} = \infty$).

The event of interest is that a single future unit, which we call the 'next unit', undergoing the same test or process as the n units for which failure data are available, fails due to a specific failure mode, say mode l. The NPI lower and upper probabilities for the event $X_{l,n+1} = \min\limits_{1 \leq j \leq k} X_{j,n+1}$, for $l = 1, \ldots, k$, are

$$\underline{P}^{(l)} = \sum_{\substack{i_j=0 \\ j \neq l}}^{n} \left[\sum_{i_l=0}^{u_l} 1(x_{l,i_l+1} < \min_{\substack{1 \leq j \leq k \\ j \neq l}} \{t_{j,i_j}\}) P^l \right] \prod_{\substack{j=1 \\ j \neq l}}^{k} M^j_{t_{j,i_j}} \tag{12}$$

$$\overline{P}^{(l)} = \sum_{\substack{i_j=0 \\ j \neq l}}^{u_j} \left[\sum_{i_l=0}^{n} 1(t_{l,i_l} < \min_{\substack{1 \leq j \leq k \\ j \neq l}} \{x_{j,i_j+1}\}) M^l_{t_{l,i_l}} \right] \prod_{\substack{j=1 \\ j \neq l}}^{k} P^j \tag{13}$$

where the first summation signs denote the sums over all i_j from 0 to n or u_j for $j = 1, \ldots, k$ but not including $j = l$. The derivation of these NPI lower and upper probabilities is given in [31].

We briefly consider the special case of the general competing risks problem in which there are only two failure modes (so $k = 2$), 1 and 2, also denoted by FM1 and FM2, and with all n units considered actually failing due to one of these two failure modes. Therefore, any unit which fails due to FM1 leads to a right-censored observation for FM2, and vice versa. In this case, the number of failures due to FM1 (FM2) is equal to the number of right-censored observations for FM2 (FM1), so $v_1 = u_2$ and $v_2 = u_1$. The NPI lower and upper probabilities for the event that the next unit will fail due to FM1 are

$$\underline{P}^{(1)} = \sum_{i_2=0}^{n} \left\{ \sum_{i_1=0}^{u_1} 1(x_{1,i_1+1} < t_{2,i_2}) P^1 \right\} M^2_{t_{2,i_2}} \tag{14}$$

$$\overline{P}^{(1)} = \sum_{i_2=0}^{u_2} \left\{ \sum_{i_1=0}^{n} 1(t_{1,i_1} < x_{2,i_2+1}) M^1_{t_{1,i_1}} \right\} P^2 \tag{15}$$

This special case enables us to illustrate some interesting features of the NPI approach in this setting. We consider two specific scenarios in detail:

(A) all failures due to FM2 come first, followed by all failures due to FM1, meaning that the u_2 failure times of failures due to FM2 are all smaller than the u_1 failure times of failures due to FM1. In this case, the NPI lower and upper probabilities for the event that the next unit will fail due to FM1 are

$$\underline{P}^{(1),\mathrm{A}} = \frac{1}{u_1+1} \sum_{i_2=1}^{v_2} i_2 \; M^2(c_{2,i_2}, \infty)$$

$$\overline{P}^{(1),\mathrm{A}} = \frac{1}{n+1} \left[v_2 + 1 + \frac{u_2}{n+1} + \sum_{i_1=1}^{v_1-1} (v_1 - i_1) \; M^1(c_{1,i_1}, x_{1,1}) \right]$$

(B) all failures due to FM1 come first, followed by all failures due to FM2, in which case the NPI lower and

upper probabilities for the event that the next unit will fail due to FM1 are

$$\underline{P}^{(1),\mathrm{B}} = \frac{1}{n+1}\left[\frac{u_1\, u_2}{u_2+1} + \sum_{i_1=1}^{v_2} i_1\, M^2(c_{2,i_2}, x_{2,1})\right]$$

$$\overline{P}^{(1),\mathrm{B}} = \frac{1}{u_2+1}\left[1+\frac{u_2(u_1+1)}{n+1}+\sum_{i_1=1}^{v_1-1}(v_1-i_1)M^1(c_{1,i_1},\infty)\right]$$

These NPI lower and upper probabilities follow straightforwardly from the general expressions given before. The main reason for highlighting these two special cases is an interesting observation in our study of NPI for competing risks data, namely that case (A) always seems to give the minimal NPI lower and upper probabilities, when all possible orderings of u_2 failures due to FM2 and u_1 failures due to FM1 are considered, while case (B) always seems to give the maximal NPI lower and upper probabilities. For now, we propose this property as a conjecture, which we strongly believe to hold and hope to prove generally in the near future.

7 Examples

7.1 Example I: Early termination

Desu and Raghavarao [23] present recorded times (months) until promotion at a large company, for 19 employees in $k = 3$ departments. The data are: Dept 1: $15, 20+, 36, 45, 58, 60$ ($n_1 = 6$); Dept 2: $12, 25+, 28, 30+, 30+, 36, 40, 45, 48$ ($n_2 = 9$); Dept 3: $30+, 40, 48, 50$ ($n_3 = 4$), where " + " indicates that the employee left the company at that length of service before getting promotion, this is considered to be a right-censored observation (one could argue about whether or not this right-censoring process is independent of the promotion process, but as we only use this data set for illustration, and have no further circumstantial information, we do not address this in more detail). We consider at which department the data suggest that one needs to work the longest to get a promotion. This data set contains tied observations, in NPI these are dealt with by assuming that they differ by a very small amount, in such a way that the lower (or upper) probability of interest is smallest (largest) over all possible ways to break the ties.

To illustrate NPI for multiple comparisons with early termination, as summarized in Section 4, assume that all these employees started working at this company at the same time, and that one considers the data after T_0 months, so all larger observations in the data above are treated as being right-censored at T_0. For several values of T_0, the lower and upper probabilities for the event that one has to work the longest in department

T_0	$\underline{P}^{(1)}$	$\overline{P}^{(1)}$	$\underline{P}^{(2)}$	$\overline{P}^{(2)}$	$\underline{P}^{(3)}$	$\overline{P}^{(3)}$
11	0	1	0	1	0	1
14	0	1	0	0.903	0	1
17	0	0.863	0	0.903	0.011	1
27	0	0.863	0	0.903	0.011	1
33	0	0.863	0	0.797	0.024	1
38	0	0.714	0	0.659	0.089	1
42	0.068	0.714	0.025	0.540	0.114	0.833
47	0.081	0.615	0.032	0.434	0.197	0.833
49	0.167	0.615	0.032	0.354	0.216	0.748
52	0.239	0.615	0.032	0.354	0.216	0.662
59	0.239	0.615	0.032	0.354	0.216	0.662
61	0.239	0.615	0.032	0.354	0.216	0.662

Table 1: Lower and upper probabilities, Example I

l, $\underline{P}^{(l)}$ and $\overline{P}^{(l)}$, for $l = 1, 2, 3$, are presented in Table 1. There is no value of T_0 for which the corresponding data would strongly indicate that one of the departments leads to longest time to promotion, according to the formulation of such indications as explained in Section 4. For several T_0, for example $T_0 = 17$, both the lower and upper probabilities for department 3 are greater than the lower and upper probabilities, respectively, for department 1 and for department 2. As discussed in Section 4, one could argue that this provides a weak indication that department 3 leads to the longest times until promotion. However, the large imprecision in these lower and upper probabilities indicates that the evidence for such a claim is weak, so care must be taken when formulating any conclusion along these lines. For larger values of T_0, department 3 has most imprecision remaining, which reflects that there are only few observations for this department.

7.2 Example II: Progressive censoring

In this example, we illustrate the above presented NPI approach for comparison of two groups of lifetime data under several progressive censoring schemes. We use a subset of Nelson's data [32] on breakdown times (in minutes) of an insulating fluid that is subject to high voltage stress. The data are given below, 10 units per group involved in the experiment, so $n_x = n_y = 10$.
$X : 0.49, 0.64, 0.82, 0.93, 1.08, 1.99, 2.06, 2.15, 2.57, 4.75$
$Y : 1.34, 1.49, 1.56, 2.10, 2.12, 3.83, 3.97, 5.13, 7.21, 8.71$
We present the NPI lower and upper probabilities that group Y is better than group X, by comparing single next future observations from both groups, X_{11} and Y_{11}. The appropriate assumptions rc-$A_{(n)}$ are again made per group, and it is assumed that the groups are fully independent.

Suppose that progressive Type-II censoring is applied to group Y, with three units withdrawn from the experiment at the first observed breakdown time

for group Y (at $y_1 = 1.34$), and two units for this group withdrawn at the last observed breakdown time, $y_5 = 5.13$, so with $R^y = (3, 0, 0, 0, 2)$. It is also assumed that all breakdown times for the units from group X are observed. Assume that, with y^c denoting a right-censored observation at time y, the data actually observed in this case are
$X : 0.49, 0.64, 0.82, 0.93, 1.08, 1.99, 2.06, 2.15, 2.57, 4.75$
$Y : 1.34, 1.34^c, 1.34^c, 1.34^c, 1.49, 1.56, 2.12, 5.13, 5.13^c, 5.13^c$
The NPI lower and upper probabilities are $\underline{P}(Y_{11} > X_{11}) = 0.6139$ and $\overline{P}(Y_{11} > X_{11}) = 0.8052$.

Now suppose that the progressive Type-II censoring scheme is applied to both groups X and Y, with $R^x = (2, 1, 0, 1, 0, 0)$ and $R^y = (1, 2, 0, 3)$ and resulting in the following data, X : $0.49, 0.49^c, 0.49^c, 0.64, 0.64^c, 0.93, 1.99, 1.99^c, 2.06, 4.75$
$Y : 1.34, 1.34^c, 1.49, 1.49^c, 1.49^c, 2.10, 2.12, 2.12^c, 2.12^c, 2.12^c$
These data lead to NPI lower and upper probabilities $\underline{P}(Y_{11} > X_{11}) = 0.5148$ and $\overline{P}(Y_{11} > X_{11}) = 0.8506$.

Precedence testing can be considered as a special case of progressive censoring. Suppose that the experiment is terminated as soon as the fifth breakdown from group Y is observed, i.e. at time $y_5 = 2.12$. Then the breakdown times of five units from group Y are right-censored at that time, together with three units from group X, resulting in data
$X : 0.49, 0.64, 0.82, 0.93, 1.08, 1.99, 2.06, 2.12^c, 2.12^c, 2.12^c$
$Y : 1.34, 1.49, 1.56, 2.10, 2.12, 2.12^c, 2.12^c, 2.12^c, 2.12^c, 2.12^c$
For these data, NPI gives $\underline{P}(Y_{11} > X_{11}) = 0.5289$ and $\overline{P}(Y_{11} > X_{11}) = 0.8264$. Coolen-Schrijner *et al* [20] present several results for NPI precedence testing, including the attractive fact that, if one increases the end-time of the experiment, such an NPI lower (upper) probability for comparison of two groups never decreases (increases).

7.3 Example III: Competing risks

In this example, a well-known data set from the literature [27] is used to illustrate some aspects of the NPI method for dealing with competing risks. The data contain information about 36 units of a new model of a small electrical appliance which were tested, and where the lifetime observation per unit consists of the number of completed cycles of use until the unit failed. These data are presented in Table 2, which also includes the specific failure mode (FM) that caused the unit to fail. In the study, there were 18 different ways in which an appliance could fail, so 18 failure modes, but to illustrate the NPI method we will first reduce this to two failure modes, thereafter we consider grouping into three failure modes. Three units in the test did not fail before the end of the experiment, so for these units we have right-censored observations (2565, 6367 and 13403) for all failure modes considered, indicated by '-' for the failure mode in Table 2.

# cycles	FM	# cycles	FM	# cycles	FM
11	1	1990	9	3034	9
35	15	2223	9	3034	9
49	15	2327	6	3059	6
170	6	2400	9	3112	9
329	6	2451	5	3214	9
381	6	2471	9	3478	9
708	6	2551	9	3504	9
958	10	2565	-	4329	9
1062	5	2568	9	6367	-
1167	9	2702	10	6976	9
1594	2	2761	6	7846	9
1925	9	2831	2	13403	-

Table 2: Failure data for electrical appliance test

The two most frequently occurring failure modes in these data are FM9, which caused 17 units to fail, and FM6 which caused 7 failures. We consider how likely it is that the next unit, say unit 37, would fail due to FM9, assuming it would undergo the same test and its number of completed cycles would be exchangeable with these numbers for the 36 units reported. Let us first group all failure modes other than FM9 together, and consider these jointly as a failure mode, so we consider the NPI approach with 2 failure modes, FM9 and, say, 'other failure mode' (OFM). There are still three units that do not fail and for which we only have right-censored observations (RC). The data corresponding to this definition of failure modes are presented in Table 3.

FM9	1167	1925	1990	2223	2400	2471
	2551	2568	3034	3034	3112	3214
	3478	3504	4329	6976	7846	
OFM	11	35	49	170	329	381
	708	958	1062	1594	2327	2451
	2702	2761	2831	3059		
RC	2565	6367	13403			

Table 3: Failure data for FM9, OFM and RC

In this case there are tied observations, as two units have failed due to FM9 after 3034 completed cycles. To deal with this, we assume a small difference between these values, such that their ordering does not change with regard to observations of units in other groups, so, we assume that one of these two units actually failed after 3035 completed cycles. If such a tie would occur among different groups, then one can break it similarly in two ways, different for upper and lower probabilities in such a way that these are maximal and minimal, respectively, over the possible ways of breaking such ties, without changing the order of

these observations with respect to all other observations. For competing risks data, a failure time observation caused by one failure mode is simultaneously a right-censored observation for all other failure modes. This situation is dealt with in the NPI approach, as is common in many statistical approaches, by assuming that the right-censoring time is just beyond the failure time. For the three right-censored observations for units that were not observed to fail, we also have tied observations for the two failure modes considered (FM9 and OFM), so for both these right-censoring times coincide. We deal with this again by assuming that for one of the failure modes this event occurred fractionally later than for the other, and then we calculate the lower and upper probabilities for the event of interest by considering the maximum and minimum of the upper and lower probabilities, respectively, corresponding to the different possible orderings of these 'un-tied' right-censoring times.

The NPI lower and upper probabilities for the event that unit 37 will fail due to FM9 are

$$\underline{P}(X_{37}^{FM9} < X_{37}^{OFM}) = 0.4358,$$

$$\overline{P}(X_{37}^{FM9} < X_{37}^{OFM}) = 0.5804$$

while the corresponding NPI lower and upper probabilities for unit 37 to fail due to OFM are

$$\underline{P}(X_{37}^{OFM} < X_{37}^{FM9}) = 0.4196,$$

$$\overline{P}(X_{37}^{OFM} < X_{37}^{FM9}) = 0.5642$$

These lower and upper probabilities satisfy the conjugacy property as, implicit in our method, it is assumed that the experiment on unit 37 would actually continue until it fails, and this is assumed to happen with certainty. NPI can be generalized to take the possibility of 'non-failure' of the next unit by a certain time into account, but we have not developed this further. On the basis of these NPI lower and upper probabilities, one could interpret the data as containing weak evidence that the event that unit 37 will fail due to FM9 is (a bit) more likely than for it to fail due to another failure mode, with all the other failure modes grouped together as done in this case.

Let us now group the failure modes differently, by considering FM9 and FM6 separately, causing 17 and 7 units to fail, respectively. We group all the other failure modes together into OFM. The data used here are given in Table 4. The NPI lower and upper probabilities for the event that unit 37 will fail due to FM9, due to FM6 or due to OFM, are

$$\underline{P}\big(X_{37}^{FM9}<\min\{X_{37}^{FM6}, X_{37}^{OFM}\}\big)=0.3915\ ,$$

FM9	1167	1925	1990	2223	2400	2471
	2551	2568	3034	3034	3112	3214
	3478	3504	4329	6976	7846	
FM6	170	329	381	708	2327	2761
	3059					
OFM	11	35	49	958	1062	1594
	2451	2702	2831			
RC	2565	6367	13403			

Table 4: Failure data for FM9, FM6, OFM and RC

$$\overline{P}\big(X_{37}^{FM9}<\min\{X_{37}^{FM6}, X_{37}^{OFM}\}\big)=0.5804$$
$$\underline{P}\big(X_{37}^{FM6}<\min\{X_{37}^{FM9}, X_{37}^{OFM}\}\big)=0.1749\ ,$$
$$\overline{P}\big(X_{37}^{FM6}<\min\{X_{37}^{FM9}, X_{37}^{OFM}\}\big)=0.3279$$
$$\underline{P}\big(X_{37}^{OFM}<\min\{X_{37}^{FM6}, X_{37}^{FM9}\}\big)=0.2265\ ,$$
$$\overline{P}\big(X_{37}^{OFM}<\min\{X_{37}^{FM6}, X_{37}^{FM9}\}\big)=0.3808$$

Since

$$\underline{P}\left(X_{37}^{FM9} < \min\left\{X_{37}^{FM6}, X_{37}^{OFM}\right\}\right) > \overline{P}\left(X_{37}^{FM6} < \min\left\{X_{37}^{FM9}, X_{37}^{OFM}\right\}\right)$$

one could interpret the data as providing strong evidence that unit 37 is more likely to fail due to FM9 than due to FM6, in this setting with all other failure modes grouped into OFM. If one adopts a subjective interpretation of lower and upper probabilities in terms of prices for desirable gambles, in line with Walley [36], then these lower and upper probabilities would imply that, for any price between 0.3279 and 0.3915, one would be willing both to buy the gamble which pays 1 if unit 37 fails due to FM9 and to sell the gamble which pays 1 if unit 37 fails due to FM6. If one has a quick look at the data, one may be surprised that FM6 is not the more likely one to lead to failure, as it has caused relatively many early failures. However, it only caused failure of 7 out of the 36 units tested, the comparisons would be different if the data were not competing risks data on the same units but failure times for independent groups without the important aspect of a failure due to one failure mode providing a right-censored observation for all other failure modes. Similarly, strong evidence that unit 37 is more likely to fail due to FM9 than due to OFM can be claimed because

$$\underline{P}\left(X_{37}^{FM9} < \min\left\{X_{37}^{FM6}, X_{37}^{OFM}\right\}\right) > \overline{P}\left(X_{37}^{OFM} < \min\left\{X_{37}^{FM6}, X_{37}^{FM9}\right\}\right)$$

Comparison of these two cases illustrates some features that are different in statistics using lower and upper probabilities when compared to methods using

precise probabilities. The lower and upper probabilities for unit 37 to fail due to FM9 are $[0.4358, 0.5804]$ in the first case, with all other failure modes grouped together, and $[0.3915, 0.5804]$ in the second case, with FM6 also taken separately. In the latter case, there is more imprecision in these upper and lower probabilities, while data are represented in more detail. This increase in imprecision, actually the fact that these upper and lower probabilities are nested with more imprecision if data are represented in more detail, is in line with a fundamental principle of NPI proposed and discussed by Coolen and Augustin [9, 10] in the context of multinomial data. This leads to the conjecture that, for such competing risks data, if more failure modes are treated separately instead of being grouped together, then lower and upper probabilities for an event that the next unit's failure is caused by a specific failure mode are nested, with imprecision increasing with the number of failure modes used. We hope to prove this conjecture in the near future.

The two NPI upper probabilities for the event that unit 37 will fail due to FM9, for the cases with all other failure modes grouped together (first case) and with FM6 separated (second case), are both equal to 0.5804. This is a consequence of the fact that this upper probability is realized with the extreme assignments of probability masses in the intervals created by the data in accordance to the lower survival function for FM9 and the upper survival function for the other failure modes. With all failure modes assumed to be independent, the upper survival function for the other failure modes combined is actually the same, whether or not FM6 is considered separately, this was discussed by Coolen *et al* [12], who presented individual NPI lower and upper survival functions and also considered the data used in this example, but they did not develop the NPI method for multiple comparisons that underlies the NPI method for competing risks presented here.

Acknowledgments

This work forms part of the research for my PhD degree, with Pauline Coolen-Schrijner as main supervisor. To my great sadness, Pauline passed away on 23 April 2008 due to her illness. She was a very supportive and nice person, who helped me very much during my study and she has a strong influence on my future career. I am grateful to you Pauline. *Tahani*

We are grateful to two anonymous referees for suggesting improvements to this paper.

References

[1] T. Augustin and F.P.A. Coolen. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124: 251–272, 2004.

[2] N. Balakrishnan and R. Aggarwala. *Progressive Censoring : Theory, Methods, and Applications*. Birkha, 2000.

[3] N. Balakrishnan and H.K.T. Ng. *Precedence-Type Tests and Applications*. Wiley, 2006.

[4] R.E. Bechhofer, T.J. Santner and D. Goldsman. *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons*. Wiley, 1995.

[5] F.P.A. Coolen. Comparing two populations based on low stochastic structure assumptions. *Statistics and Probability Letters*, 29: 297–305, 1996.

[6] F.P.A. Coolen. On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, 15: 21–47, 2006.

[7] F.P.A. Coolen. On probabilistic safety assessment in case of zero failures. *Journal of Risk and Reliability*, 220: 105–114, 2006.

[8] F.P.A. Coolen. Nonparametric prediction of unobserved failure modes. *Journal of Risk and Reliability*, 221: 207–216, 2007.

[9] F.P.A. Coolen and T. Augustin. Learning from multinomial data: a nonparametric predictive alternative to the imprecise dirichlet model. In *ISIPTA'05: Proceedings of the Fourth International Symposium on Imprecise Probability Theory and Applications*, Cozman *et al* (eds), pp 125–134, 2005.

[10] F.P.A. Coolen and T. Augustin. A nonparametric predictive alternative to the imprecise dirichlet model: the case of a known number of categories. *International Journal of Approximate Reasoning*, 50: 217–230, 2009.

[11] F. P. A. Coolen and P. Coolen-Schrijner. Nonparametric predictive comparison of proportions. *Journal of Statistical Planning and Inference*, 137: 23–33, 2007.

[12] F.P.A. Coolen, P. Coolen-Schrijner and K.J. Yan. Nonparametric predictive inference in reliability. *Reliability Engineering and System Safety*, 78: 185–193, 2002.

[13] F.P.A. Coolen and L.V. Utkin. Imprecise reliability. In *Wiley Encyclopedia of Quantitative Risk Analysis and Assessment*, Melnick and Everitt (eds), pp 875–881. Wiley, 2008.

[14] F.P.A. Coolen and P. van der Laan. Imprecise predictive selection based on low structure assumptions. *Journal of Statistical Planning and Inference*, 98: 259–277, 2001.

[15] F.P.A. Coolen and K.J. Yan. Nonparametric predictive comparison of two groups of lifetime data. In *ISIPTA'03: Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications*, Bernard *et al* (eds), pp 148–161, 2003.

[16] F.P.A. Coolen and K.J. Yan. Nonparametric predictive inference with right-censored data. *Journal of Statistical Planning and Inference*, 126: 25–54, 2004.

[17] P. Coolen-Schrijner and F.P.A. Coolen. Nonparametric predictive comparison of success-failure data in reliability. *Journal of Risk and Reliability*, 221: 319–327, 2007.

[18] P. Coolen-Schrijner, F.P.A. Coolen and I.M. MacPhee. Nonparametric predictive inference for systems reliability with redundancy allocation. *Journal of Risk and Reliability*, 222: 463–476, 2008.

[19] P. Coolen-Schrijner, F.P.A. Coolen and S.C. Shaw. Nonparametric adaptive opportunity-based age replacement strategies. *Journal of the Operational Research Society*, 57: 63–81, 2006.

[20] P. Coolen-Schrijner, T.A. Maturi and F.P.A. Coolen. Nonparametric predictive precedence testing for two groups. *Journal of Statistical Theory and Practice,*, to appear.

[21] P. Coolen-Schrijner, S.C. Shaw and F.P.A. Coolen. Opportunity-based age replacement with a one-cycle criterion. *Journal of the Operational Research Society*, to appear.

[22] B. De Finetti. *Theory of Probability*. Wiley, 1974.

[23] M.M. Desu and D. Raghavarao. *Nonparametric Statistical Methods for Complete and Censored Data*. Chapman and Hall, 2004.

[24] B.M. Hill. De Finetti's theorem, induction, and A_n, or Bayesian nonparametric predictive inference (with discussion). In *Bayesian Statistics 3*, Lindley *et al* (eds), pp 211–241, 1988.

[25] B.M. Hill. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63: 677–691, 1968.

[26] B.M. Hill. Parametric models for $A_{(n)}$: Splitting processes and mixtures. *Journal of the Royal Statistical Society B*, 55: 423–433, 1993.

[27] J.F. Lawless. *Statistical Models and Methods for Lifetime Data*. Wiley, 2003.

[28] T. A. Maturi, P. Coolen-Schrijner, and F. P. A. Coolen. Nonparametric predictive selection with early experiment termination. In preparation.

[29] T.A. Maturi, P. Coolen-Schrijner and F.P.A. Coolen. Nonparametric predictive comparison of lifetime data under progressive censoring. In submission.

[30] T.A. Maturi, P. Coolen-Schrijner and F.P.A. Coolen. Nonparametric predictive comparison of lifetime data with early termination of experiments. In submission.

[31] T.A. Maturi, P. Coolen-Schrijner and F.P.A. Coolen. On nonparametric predictive inference for competing risks. In *Proceedings of the 18th Advances in Risk and Reliability Technology Symposium*, Lisa Bartless (ed), Loughborough University, pp 196–211, 2009.

[32] W. Nelson. *Applied Life Data Analysis*. Wiley, 2004.

[33] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[34] A. Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences of the United States of America*, 72: 20–22, 1975.

[35] L.V. Utkin and F.P.A. Coolen. Imprecise reliability: an introductory overview. In *Computational Intelligence in Reliability Engineering, Volume 2: New Metaheuristics, Neural and Fuzzy Techniques in Reliability*, Levitin (ed), pp 261–306. Springer, 2007.

[36] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

[37] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervalwahrscheinlichkeit als umfassendes Konzept*. Physika, 2001.

Object association in the TBM framework, application to vehicle driving aid

David Mercier
EA 3926 LGI2A
Université d'Artois, France
david.mercier@univ-artois.fr

Eric Lefevre
EA 3926 LGI2A
Université d'Artois, France
eric.lefevre@univ-artois.fr

Daniel Jolly
EA 3926 LGI2A
Université d'Artois, France
daniel.jolly@univ-artois.fr

Abstract

The problem tackled in this paper deals with obstacle tracking in the context of vehicle driving aid, especially the association step, which consists in associating perceived objects with known objects detected at a previous time. A contribution in the modeling of this association problem in the belief function framework is introduced. By interpreting belief functions as weighted opinions according to the Transferable Belief Model semantics, pieces of information regarding the association of known objects and perceived objects can be expressed in a common global space of association to be combined by the conjunctive rule of combination, and a decision making process using the pignistic transformation can be made. This approach is validated on real data.

Keywords. Obstacle tracking, association step, belief functions, Transferable Belief Model.

1 Introduction

In obstacle tracking, the association step consists in establishing a correlation between tracks (known objects) and targets (perceived objects) from information usually provided by different sensors or captors. Such a mapping can be even more complex depending on the number of targets and tracks, as well as the quality of the provided information. Introduced by Dempster [5] and Shafer [21], belief functions constitute a suitable framework for the representation and manipulation of imperfect information. Thus, next to architectures based on Bayesian probabilistic framework [2, 3], Rombaut [18, 19] develops a first modeling based on belief functions. In this model, information regarding the association of couples (known objects, perceived objects) is represented by belief functions, which are combined using, for simplicity reasons, an adapted combination introduced by Rombaut. In [12] this latter model is developed by using a decision-making system based on belief matrices and the application of a coupling algorithm.

In this paper, a modeling of this association step problem is introduced in the Smets' semantic approach of belief functions: the Transferable Belief Model (TBM) [24], a subjectivist and non-probabilistic interpretation of the Dempster-Shafer theory of belief function. In particular, it is shown that TBM classical tool like the conjunctive combination rule and the pignistic decision-making can be implemented and tested in a real time application, these experimental results demonstrating the effectiveness of this approach as compared to Rombaut's combination rule.

Let us note that the works presented here reexpress and extend in the Transferable Belief Model a former model presented by some of the authors in [13]. Likewise, the association problem described here has many similarities with the works undertaken by Ristic and Smets in [17].

This paper is organized as follows. The TBM basic concepts we need are recalled in Section 2. An association algorithm based on belief functions is introduced in Section 3 and discussed in particular with the other approaches in Section 4. Then, experimental results on real data are presented in Section 5. Finally, Section 6 concludes this paper.

2 Transferable Belief Model (TBM): basic concepts

The Transferable Belief Model (TBM) is a model of uncertain reasoning and decision-making based on two levels [10, 24]:

- the credal level, where available pieces of information are represented by belief functions, and manipulated;
- the pignistic or decision level, where belief functions are transformed into probability measures

when a decision has to be made, and the expected utility is maximized.

2.1 Representing information with belief functions

2.1.1 Belief functions

The knowledge held by an agent is represented by the allocation of a finite mass of belief to subsets of the universe of discourse.

Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$, called the frame of discernment, be a finite set composed of all possible answers to a given question Q of interest. The beliefs held by a rational agent Ag regarding the answer to question Q can be quantified by a *belief mass function* or *basic belief assignment (BBA)* $m^{\Omega}_{Ag} : 2^{\Omega} \rightarrow [0, 1]$ s.t.:

$$\sum_{A \subseteq \Omega} m^{\Omega}_{Ag}(A) = 1 \ . \tag{1}$$

The quantity $m^{\Omega}_{Ag}(A)$ represents the part of the unit mass allocated to the hypothesis that the answer to question Q is in the subset A of Ω. When there is no ambiguity, the notation m^{Ω}_{Ag} will be simplified as follows m^{Ω} or m.

- A subset A of Ω such that $m(A) > 0$ is called a *focal set* of m.
- A BBA m with only one focal set A is called a *categorical BBA* and is denoted m_A; then $m_A(A) = 1$.
- Total ignorance is represented by the BBA m_{Ω} called the *vacuous BBA*.
- A *normal BBA* m satisfies the condition $m(\emptyset) = 0$.
- Let A be a subset of Ω, the cardinality of A, denoted $|A|$, is the number of elements of Ω in A; if $|A| = 1$, A is said to be a *singleton*.

The belief and plausibility functions associated with a BBA m are defined, respectively, as:

$$bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B) \qquad \forall A \subseteq \Omega \ , \tag{2}$$

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B) \qquad \forall A \subseteq \Omega \ . \tag{3}$$

Functions m, bel and pl are in one-to-one correspondence, and thus constitute different forms of the same information.

2.1.2 Refinements and Coarsenings

When applying the TBM to a real-world application, the determination of the frame of discernment Ω, which defines the set of states on which beliefs will be expressed, is a crucial step. As noticed by Shafer [21, chapter 6], the degree of granularity of Ω is always, to some extent, a matter of convention, as any element of Ω representing a given state can always be split into several alternatives. Hence, it is fundamental to examine how a belief function defined on a frame may be expressed in a finer or, conversely, in a coarser frame. The concepts of refinement and coarsening can be defined as follows.

Let Θ and Ω denote two frames of discernment. A mapping $\rho : 2^{\Theta} \rightarrow 2^{\Omega}$ is called a *refining* of Θ (Figure 5) if it verifies the following properties:

1. The set $\{\rho(\{\theta\}), \theta \in \Theta\} \subseteq 2^{\Omega}$ is a partition of Ω, and
2. For all $A \subseteq \Theta$:

$$\rho(A) = \bigcup_{\theta \in A} \rho(\{\theta\}). \tag{4}$$

Θ is then called a coarsening of Ω, and Ω is called a refinement of Θ.

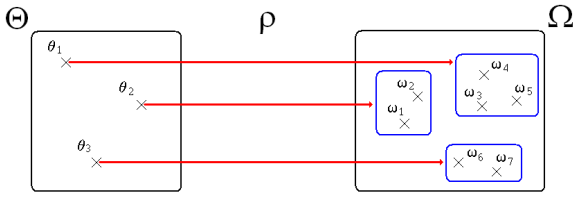


Figure 1: Illustration of a coarsening Θ of Ω associated with a refining ρ of Θ.

2.2 Manipulating information with belief functions

2.2.1 Vacuous Extension

The *vacuous extension* operation allows one to convey a belief mass function m^{Θ}, expressing a state of belief on Θ, to a finer frame Ω, a refinement of Θ. Stemming from the *least committed principle* [22], this operation is denoted with an arrow pointing up, and is defined by:

$$m^{\Theta \uparrow \Omega}(\rho(A)) = m^{\Theta}(A), \quad \forall A \subseteq \Theta \ , \tag{5}$$

where ρ is the refining of Θ in Ω.

2.2.2 Combining beliefs

Two BBAs m_1 and m_2, induced by distinct and reliable sources of information, can be combined using the *conjunctive rule of combination* (CRC), also called *unnormalized Dempster's rule of combination*, defined for all $A \subseteq \Omega$ by:

$$m_1 \textcircled{\cap} m_2(A) = \sum_{B \cap C = A} m_1(B) m_2(C) \ . \tag{6}$$

The normalization hypothesis ($m(\emptyset) = 0$) can be recovered with the following normalization step:

$$m_1 \oplus m_2(A) = \begin{cases} \frac{m_1 \textcircled{\cap}_2(A)}{1 - m_1 \textcircled{\cap}_2(\emptyset)} & \text{if } \emptyset \neq A \subseteq \Omega \ , \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

This latter rule of combination is called *Dempster's rule of combination*.

2.3 Decision-making level

When a decision has to be made regarding the answer to question Q, some *rationally principles* [4] justify the strategy consisting in choosing the decision d among a set of possible decisions $\mathcal{D}$, which minimizes the *expected risk* defined by:

$$R(d) = \sum_{\omega \in \Omega} c(d, \omega) P^{\Omega}(\{\omega\}), \tag{8}$$

where $P^{\Omega} : 2^{\Omega} \rightarrow [0, 1]$ is a probability measure and $c : \mathcal{D} \times \Omega \rightarrow \mathbb{R}$ a cost function, $c(d, \omega)$ representing the cost to decide d while the truth is ω.

At this level, the mass function m^{Ω} representing the available information regarding the answer to question Q belonging to Ω (resulting in practice from a fusion process) has then to be transformed in a probability measure. A solution [7] consists in computing the *pignistic probability* [23] defined by:

$$BetP^{\Omega}(\{\omega\}) = \sum_{\{A \subseteq \Omega, \omega \in A\}} \frac{m(A)}{|A| \ (1 - m(\emptyset))}, \quad \forall \omega \in \Omega. \tag{9}$$

The chosen decision is then the one that minimizes the pignistic risk defined by:

$$R_{Bet}(d) = \sum_{\omega \in \Omega} c(d, \omega) BetP^{\Omega}(\{\omega\}) \ . \tag{10}$$

In the case of 0-1 costs with $\mathcal{D} = \Omega$, which means that $c(\omega_i, \omega_j) = 1$ if $i = j$, 0 otherwise, choosing the decision d which minimises the pignistic risk (10) is equivalent to choose the decision d which maximizes the pignistic probability (9).

An other case consists in choosing 0-1 costs with $\mathcal{D} = \Omega \cup \{d_0\}$, where d_0, called *rejection decision* [7], consists in refusing to make a decision belonging to $\mathcal{D} \setminus \{d_0\}$ when the risk is judged too high. By denoting $c_0 = c(d_0, \omega_i) \ \forall i \in \{1, \ldots, N\}$, minimizing the pignistic risk (10) is equivalent to choose the decision:

- d_0 if $\max\limits_{i=1,\ldots,N} BetP(\{\omega_i\}) < 1 - c_0$,
- ω_j if $BetP(\{\omega_j\}) = \max\limits_{i=1,\ldots,N} BetP(\{\omega_i\}) \geq 1 - c_0$.

The cost c_0 is called the *rejection cost*.

3 Object association algorithm

3.1 Representing information with belief functions

The first step when building belief functions is to define the universe of discourse.

Let us consider the following notations:

- X_i: designate a perceived object at time t, $i \in I = \{1, \ldots, N\}$, N being the number of perceived objects at time t;
- Y_j: designate a known object at previous time $t-1$, $j \in J = \{1, \ldots, M\}$, M being the number of known objects at time $t-1$;
- *: a proposition meaning "no object".

The association process objective consists in finding the best possible association between a set of perceived objects $\{X_1, X_2, \ldots, X_N, ^{\star}\}$ and a set of known objects $\{Y_1, Y_2, \ldots, Y_M, ^{\star}\}$, under the following constraints:

- each perceived object X_i is associated with at most one known object;
- each known object Y_j is associated with at most one perceived object;
- proposition * can be associated to any objects.

Frames of discernment involved in this application are then the followings:

- $\Omega_{i,j} = \{y_{i,j}, n_{i,j}\}$, containing the two possible answers (yes or no) to the question $Q_{i,j}$: "Is the perceived object X_i associated with the known object Y_j?";

- $\Omega_{X_i} = \{Y_1, Y_2, \ldots, Y_M, \star\}$, containing the set of possible answers to the question Q_{X_i} : "Who is associated with the perceived object X_i?", proposition $\star$ meaning that X_i has appeared;
- $\Omega_{Y_j} = \{X_1, X_2, \ldots, X_N, \star\}$, containing the set of possible answers to the question Q_{Y_j} : "Who is associated with the known object Y_j?", proposition $\star$ meaning that Y_j has disappeared or is hidden.

Let us remark that $\Omega_{Y_j} = \Omega_{Y_k}$, for all $j, k \in J$, and $\Omega_{X_i} = \Omega_{X_\ell}$, for all $i, \ell \in I$. Thus, Ω_{X_i} (respectively Ω_{Y_j}) can be denoted Ω_X $\forall i$ (respectively Ω_Y $\forall j$). At last, when there is no ambiguity, the frames elements will be simplified as follows :

- $\Omega_{X_i} = J \cup \{\star\} = \{1, \ldots, M, \star\}$,
- $\Omega_{Y_j} = I \cup \{\star\} = \{1, \ldots, N, \star\}$.

In the domain of intelligent vehicles, sensors or measures generally provide information regarding the association between each perceived object X_i and each known object Y_j [18, 19, 12, 11]. More precisely, initial information is represented by belief mass functions $m^{\Omega_{i,j}}$ on frames $\Omega_{i,j}$, $i \in I$, $j \in J$:

- the mass allocated to $\{y_{i,j}\}$ expresses information on the fact that X_i is associated with Y_j;
- the mass allocated to $\{n_{i,j}\}$ expresses information on the fact that X_i is not associated with Y_j;
- the mass allocated to $\Omega_{i,j} = \{y_{i,j}, n_{i,j}\}$ expresses the ignorance regarding the association of X_i and Y_j.

$N \times M$ belief mass functions $m^{\Omega_{i,j}}$ have been defined regarding the association of each object (perceived objects X_i, known objects Y_j). These pieces of information have then to be fused to determine:

- Where do perceived objects X_i come from?
- What are known objects Y_j become?

3.2 Expressing pieces of information in a common frame

To answer these questions, the $N \times M$ belief mass functions can be combined when expressed on two possible common frames: Ω_X and Ω_Y. Frames Ω_{X_i} and Ω_{Y_j} being refinements of $\Omega_{i,j}$, each information $m^{\Omega_{i,j}}$ can be expressed either on Ω_{X_i} or on Ω_{Y_j} by the vacuous extension operation (5):

$$m^{\Omega_{i,j}\uparrow\Omega_{X_i}}(\rho_{i,j}(A)) = m^{\Omega_{i,j}}(A), \quad \forall A \subseteq \Omega_{i,j} \ , \quad (11)$$

where $\rho_{i,j}$ is the refining of $\Omega_{i,j}$ on Ω_{X_i} illustrated in Figure 2, and defined by $\rho_{i,j}(\{o_{i,j}\}) = \{j\}$ and $\rho_{i,j}(\{n_{i,j}\}) = \overline{\{j\}}$.

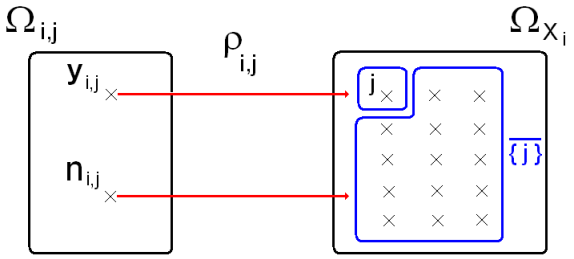


Figure 2: Refining $\rho_{i,j}$ allowing one to transport the information $m^{\Omega_{i,j}}$ on Ω_{X_i}.

Thus, for all $(i, j) \in I \times J$:

$$\begin{cases} m^{\Omega_{i,j}\uparrow\Omega_{X_i}}(\{j\}) & = m^{\Omega_{i,j}}(\{y_{i,j}\}) \\ m^{\Omega_{i,j}\uparrow\Omega_{X_i}}(\overline{\{j\}}) & = m^{\Omega_{i,j}}(\{n_{i,j}\}) \\ m^{\Omega_{i,j}\uparrow\Omega_{X_i}}(\Omega_{X_i}) & = m^{\Omega_{i,j}}(\Omega_{i,j}) \end{cases} \quad (12)$$

In the same manner, it is also possible to vacuously extend $m^{\Omega_{i,j}}$ on Ω_{Y_j}:

$$\begin{cases} m^{\Omega_{i,j}\uparrow\Omega_{Y_j}}(\{i\}) & = m^{\Omega_{i,j}}(\{y_{i,j}\}) \\ m^{\Omega_{i,j}\uparrow\Omega_{Y_j}}(\overline{\{i\}}) & = m^{\Omega_{i,j}}(\{n_{i,j}\}) \\ m^{\Omega_{i,j}\uparrow\Omega_{Y_j}}(\Omega_{Y_j}) & = m^{\Omega_{i,j}}(\Omega_{i,j}) \end{cases} \quad (13)$$

In the following of this paper, $m^{\Omega_{i,j}\uparrow\Omega_{X_i}}$ (respectively $m^{\Omega_{i,j}\uparrow\Omega_{Y_j}}$) is denoted $m_j^{\Omega_{X_i}}$ (respectively $m_i^{\Omega_{Y_j}}$).

3.3 Combining belief mass functions

At this level:

- for each $i \in I = \{1, \ldots, N\}$, M belief mass functions $m_j^{\Omega_{X_i}}$ have been created regarding the association of each object X_i toward the Y_j, the focal elements of each one being $\{j\}$, $\overline{\{j\}}$, and Ω_{X_i}.
- for each $j \in J = \{1, \ldots, M\}$, N belief mass functions $m_i^{\Omega_{Y_j}}$ have been created regarding the association of each object Y_j toward the X_i, the focal elements of each one being $\{i\}$, $\overline{\{i\}}$, et Ω_{Y_j}.

The M belief mass functions $m_j^{\Omega_{X_i}}$, considered as distinct and reliable, are combined using the conjunctive rule of combination (6).

Let us denote $m^{\Omega_{X_i}}$ the resulting mass function:

$$m^{\Omega_{X_i}} = \textcircled{\cap}_{j\in J}\, m_j^{\Omega_{X_i}} \ . \tag{14}$$

For all $k \in J$:

$$m^{\Omega_{X_i}}(\{k\}) = \sum_{\cap A_j = \{k\}} \prod_{j\in J} m_j^{\Omega_{X_i}}(A_j) \ , \tag{15}$$

where, for all $j \in J$, $A_j = \{j\}$, $\overline{\{j\}}$, or Ω_{X_i}.

But:

$$\begin{aligned} \cap_{j\in J} A_j = \{k\} \quad &\Leftrightarrow \quad A_k = \{k\} \text{ and } (A_j = \overline{\{j\}} \text{ or} \\ &\qquad A_j = \Omega_{X_i},\ \forall j \in J \setminus \{k\}), \\ &\Leftrightarrow \quad A_k = \{k\} \text{ and} \\ &\qquad A_j \neq \{j\},\ \forall j \in J \setminus \{k\} \ . \end{aligned}$$

Thus, for all $k \in J$:

$$m^{\Omega_{X_i}}(\{k\}) = m_k^{\Omega_{X_i}}(\{k\}) \prod_{\substack{j=1 \\ j\neq k}}^{M} (1 - m_j^{\Omega_{X_i}}(\{j\})) \ . \tag{16}$$

Similarly, for all $K \subseteq J$:

$$\begin{aligned} m^{\Omega_{X_i}}(\overline{K}) &= \sum_{\cap A_j = \overline{K}} \prod_{j\in J} m_j^{\Omega_{X_i}}(A_j) \ , \\ &= \prod_{j\in K} m_j^{\Omega_{X_i}}(\overline{\{j\}}) \prod_{j\in\overline{K}} m_j^{\Omega_{X_i}}(\Omega_{X_i}) \ . \end{aligned}$$

In particular:

$$\begin{aligned} m^{\Omega_{X_i}}(\{\star\}) &= m^{\Omega_{X_i}}(\overline{J}) = \prod_{j\in J} m_j^{\Omega_{X_i}}(\overline{\{j\}}) \ , \\ m^{\Omega_{X_i}}(\Omega_{X_i}) &= m^{\Omega_{X_i}}(\overline{\emptyset}) = \prod_{j\in J} m_j^{\Omega_{X_i}}(\Omega_{X_i}) \ . \end{aligned}$$

At last:

$$\begin{aligned} m^{\Omega_{X_i}}(\emptyset) &= \sum_{\cap A_j = \emptyset} \prod_{j\in J} m_j^{\Omega_{X_i}}(A_j) \ , &(17)\\ &= \sum_{\substack{j,k\in J \\ j\neq k}} m_j^{\Omega_{X_i}}(\{j\})\, m_k^{\Omega_{X_i}}(\{k\}). &(18) \end{aligned}$$

In the same manner, the N belief mass functions $m_i^{\Omega_{Y_j}}$ can also be conjunctively combined to result in a mass function $m^{\Omega_{Y_j}}$.

Example 1 *Let us consider one perceived object X_1 and two known objects Y_1 and Y_2 s.t.:*

$$\left\{\begin{array}{ll} m^{\Omega_{1,1}}(\{o_{1,1}\}) &= .2 \\ m^{\Omega_{1,1}}(\{n_{1,1}\}) &= .45 \\ m^{\Omega_{1,1}}(\Omega_{1,1}) &= .35 \end{array}\right. \quad \left\{\begin{array}{ll} m^{\Omega_{1,2}}(\{o_{1,2}\}) &= .45 \\ m^{\Omega_{1,2}}(\{n_{1,2}\}) &= .15 \\ m^{\Omega_{1,2}}(\Omega_{1,2}) &= .4 \end{array}\right. \tag{19}$$

By expressing this information on Ω_{X_1} (X_1 point of view: with which known object, the perceived object X_1 is associated? In other words: Where does X_1 come from?), it is obtained:

$$\left\{\begin{array}{ll} m_1^{\Omega_{X_1}}(\{1\}) &= .2 \\ m_1^{\Omega_{X_1}}(\overline{\{1\}}) &= .45 \\ m_1^{\Omega_{X_1}}(\Omega_{X_1}) &= .35 \end{array}\right. \quad \left\{\begin{array}{ll} m_2^{\Omega_{X_1}}(\{2\}) &= .45 \\ m_2^{\Omega_{X_1}}(\overline{\{2\}}) &= .15 \\ m_2^{\Omega_{X_1}}(\Omega_{X_1}) &= .4 \end{array}\right. \tag{20}$$

The conjunctive combination of $m_1^{\Omega_{X_1}}$ and $m_2^{\Omega_{X_1}}$ provides the following result:

$$\begin{array}{lll} m^{\Omega_{X_1}}(\{1\}) & = .2 \times (1 - .45) & = .2 \times .55 = .11 \\ m^{\Omega_{X_1}}(\{2\}) & = .45 \times (1 - .2) & = .45 \times .8 = .36 \\ m^{\Omega_{X_1}}(\overline{\{1\}}) & = m^{\Omega_{X_1}}(\{2,\star\}) & = .45 \times .4 = .18 \\ m^{\Omega_{X_1}}(\overline{\{2\}}) & = m^{\Omega_{X_1}}(\{1,\star\}) & = .15 \times .35 = .05 \\ m^{\Omega_{X_1}}(\overline{\{1,2\}}) & = m^{\Omega_{X_1}}(\{\star\}) & = .45 \times .15 = .07 \\ m^{\Omega_{X_1}}(\Omega_{X_1}) & = m^{\Omega_{X_1}}(\{1,2,\star\}) & = .35 \times .4 = .14 \\ m^{\Omega_{X_1}}(\emptyset) & = .2 \times .45 & = .09 \ . \end{array} \tag{21}$$

3.4 Decision-making

The pignistic probability $BetP^{\Omega_{X_i}}$ (9) computed from $m^{\Omega_{X_i}}$ is defined for all $\omega \in \Omega_{X_i}$ by:

$$BetP^{\Omega_{X_i}}(\{\omega\}) = \sum_{\{A\subseteq\Omega_{X_i}, \omega\in A\}} \frac{m^{\Omega_{X_i}}(A)}{|A|\ (1 - m^{\Omega_{X_i}}(\emptyset))}. \tag{22}$$

Then, for all $k \in J$:

$$\begin{aligned} BetP^{\Omega_{X_i}}(\{k\}) = \mathcal{K}_1 \Bigg[& m_k^{\Omega_{X_i}}(\{k\}) \prod_{\substack{j=1 \\ j\neq k}}^{M} (1 - m_j^{\Omega_{X_i}}(\{j\})) \\ & + \sum_{\substack{k\in\overline{K} \\ K\subseteq J}} \frac{1}{|\overline{K}|} \prod_{j\in K} m_j^{\Omega_{X_i}}(\overline{\{j\}}) \prod_{j\in\overline{K}} m_j^{\Omega_{X_i}}(\Omega_{X_i}) \Bigg] \ , \end{aligned} \tag{23}$$

where

$$\mathcal{K}_1 = \frac{1}{1 - m^{\Omega_{X_i}}(\emptyset)} = \frac{1}{1 - \sum_{\substack{j,k\in J \\ j\neq k}} m_j^{\Omega_{X_i}}(\{j\})\, m_k^{\Omega_{X_i}}(\{k\})} \ . \tag{24}$$

At last:

$$\begin{aligned} &BetP^{\Omega_{X_i}}(\{\star\}) = \\ &\mathcal{K}_1 \sum_{K\subseteq J} \frac{1}{|\overline{K}|} \prod_{j\in K} m_j^{\Omega_{X_i}}(\overline{\{j\}}) \prod_{j\in\overline{K}} m_j^{\Omega_{X_i}}(\Omega_{X_i}) \ . \end{aligned} \tag{25}$$

Once the pignistic probabilities $BetP^{\Omega_{X_i}}$ computed for each $i \in I$, the chosen decision is the one that maximizes the pignistic probability associated to the joint

law $BetP^{\Omega_{X_1}\times\ldots\times\Omega_{X_N}}$ which verifies the constraints expressed in Section 3.1.

Similarly, an equivalently justified solution consists in computing the decision from the Y_j points of view, by maximizing the pignistic probability $BetP^{\Omega_{Y_1}\times\ldots\times\Omega_{Y_M}}$.

Example 2 (Example 1 continued) *Let us consider again one perceived object X_1 and two known objects Y_1 and Y_2 with the information represented by the BBAs $m^{\Omega_{1,1}}$ and $m^{\Omega_{1,2}}$ defined by Equation 19.*

From X_1 point of view, the conjunctive combination of $m_1^{\Omega_{X_1}}$ and $m_2^{\Omega_{X_1}}$ has been detailed in Example 1. The pignistic probability $BetP^{\Omega_{X_1}}$ regarding the association of X_1 is then given by:

A	$\emptyset$	$\{1\}$	$\{2\}$	$\{\star\}$	$\{1,\star\}$	$\{2,\star\}$	$\{1,2,\star\}$
$m^{\Omega_{X_1}}(A)$	.09	.11	.36	.07	.05	.18	.14
$BetP^{\Omega_{X_1}}(A)$		.20	.55	.25	.45	.80	1

Conclusion from X_1 point of view:

1. *The singleton maximizing $BetP^{\Omega_{X_1}}$ is $\{2\}$, so X_1 is associated with Y_2;*
2. *knowing that Y_1 is not associated, Y_1 has disappeared (or is hidden).*

On the other hand, it is also possible to express the available information on Ω_{Y_1} and Ω_{Y_2}:

$$\begin{cases} m_1^{\Omega_{Y_1}}(\{1\}) & = .2 \\ m_1^{\Omega_{Y_1}}(\overline{\{1\}}) & = .45 \\ m_1^{\Omega_{Y_1}}(\Omega_{Y_1}) & = .35 \end{cases} \quad \begin{cases} m_1^{\Omega_{Y_2}}(\{1\}) & = .45 \\ m_1^{\Omega_{Y_2}}(\overline{\{1\}}) & = .15 \\ m_1^{\Omega_{Y_2}}(\Omega_{Y_2}) & = .4 \end{cases}$$

As there is only one perceived object X_1, no combination is necessary:

A	$\emptyset$	$\{1\}$	$\{\star\}$	$\{1,\star\}$
$m^{\Omega_{Y_1}}(A)$		.2	.45	.35
$BetP^{\Omega_{Y_1}}(A)$		.375	.625	1
$m^{\Omega_{Y_2}}(A)$		.45	.15	.4
$BetP^{\Omega_{Y_2}}(A)$		.65	.35	1

From the association constraints (Section 3.1), the known objects (Y_1, Y_2) can be associated to $(1,\star)$, $(\star,1)$, or $(\star,\star)$. As:

- $BetP^{\Omega_{Y_1}\times\Omega_{Y_2}}(\{1,\star\}) = .375 \times .35 = .131$;
- $BetP^{\Omega_{Y_1}\times\Omega_{Y_2}}(\{\star,1\}) = .625 \times .65 = .406$;
- $BetP^{\Omega_{Y_1}\times\Omega_{Y_2}}(\{\star,\star\}) = .625 \times .35 = .219$,

then $BetP^{\Omega_{Y_1}\times\Omega_{Y_2}}$ reaches its "valid" maximum in $\{\star,1\}$, so (Y_1,Y_2) is associated with $(\star,1)$; in other words, Y_1 has disappeared and Y_2 is associated with X_1.

In the previous example, the decision coming from X_1 and the decision coming from the Y_j are the same. Unfortunately, as illustrated by the following example, the decision providing by the criteria of maximizing the joint pignistic probability can be different depending on which point of view (perceived objects X_i or known objects Y_j) it is computed.

Let us also remark that the introduction of a rejection decision, as presented in Section 2.3, can also imply a different decision according to the X_i or Y_j points of view. For instance, by choosing c_0 equal to 0.5 in the previous Example 2, from X_1 the same decision is made as $BetP^{\Omega_{X_1}}(\{2\}) \geq 1 - c_0$, however as $BetP^{\Omega_{Y_1}\times\Omega_{Y_2}}(\{\star,1\}) < 1 - c_0$, the decision made according to the Y_j is d_0 (a rejection).

Example 3 *Let us considered one perceived object X_1, and two known objects Y_1 and Y_2, s.t.:*

$$\begin{cases} m^{\Omega_{1,1}}(\{o_{1,1}\}) & = .5 \\ m^{\Omega_{1,1}}(\{n_{1,1}\}) & = 0 \\ m^{\Omega_{1,1}}(\Omega_{1,1}) & = .5 \end{cases} \quad \begin{cases} m^{\Omega_{1,2}}(\{o_{1,2}\}) & = .7 \\ m^{\Omega_{1,2}}(\{n_{1,2}\}) & = .3 \\ m^{\Omega_{1,2}}(\Omega_{1,2}) & = 0 \end{cases}.$$

By expressing the beliefs on the frames Ω_{X_i}:

$$\begin{cases} m_1^{\Omega_{X_1}}(\{1\}) & = .5 \\ m_1^{\Omega_{X_1}}(\overline{\{1\}}) & = 0 \\ m_1^{\Omega_{X_1}}(\Omega_{X_1}) & = .5 \end{cases} \quad \begin{cases} m_2^{\Omega_{X_1}}(\{2\}) & = .7 \\ m_2^{\Omega_{X_1}}(\overline{\{2\}}) & = .3 \\ m_2^{\Omega_{X_1}}(\Omega_{X_1}) & = 0 \end{cases},$$

the following results are obtained:

A	$\emptyset$	$\{1\}$	$\{2\}$	$\{\star\}$
$m^{\Omega_{X_1}}(A)$	.35	.15	.35	0
$BetP^{\Omega_{X_1}}(A)$		.35	.54	.11

A	$\{1,\star\}$	$\{2,\star\}$	$\{1,2,\star\}$
$m^{\Omega_{X_1}}(A)$	.15	0	0
$BetP^{\Omega_{X_1}}(A)$	.46	.65	1

Then from object X_1 point of view:

- *X_1 is associated with Y_2,*
- *Y_1 has disappeared.*

From Y_1 and Y_2 points of view:

$$\begin{cases} m_1^{\Omega_{Y_1}}(\{1\}) & = .5 \\ m_1^{\Omega_{Y_1}}(\overline{\{1\}}) & = 0 \\ m_1^{\Omega_{Y_1}}(\Omega_{Y_1}) & = .5 \end{cases} \quad \begin{cases} m_1^{\Omega_{Y_2}}(\{1\}) & = .7 \\ m_1^{\Omega_{Y_2}}(\overline{\{1\}}) & = .3 \\ m_1^{\Omega_{Y_2}}(\Omega_{Y_2}) & = 0 \end{cases}. \quad (26)$$

So:

A	$\{1\}$	$\{\star\}$
$BetP^{\Omega_{Y_1}}$	.75	.25
$BetP^{\Omega_{Y_2}}$	.70	.30

(27)

As $.75 \times .3 > .7 \times .25$*,* $BetP^{\Omega_{Y_1} \times \Omega_{Y_2}}$ *reaches its valid maximum in* $\{1, \star\}$*, which implies that:*

- Y_1 *is associated with* X_1*,*
- Y_2 *has disappeared.*

This decision is then different from the previous one.

Works are currently undertaken by the authors to investigate properties input BBAs $m^{\Omega_{i,j}}$ should verify to not encounter this problem. A conjecture to be proved, is that if BBAs $m^{\Omega_{i,j}}$ are simple BBAs, which means BBAs that have two focal elements: the universe $\Omega_{i,j}$ and an other one element, then no conflicting decision appears. In other words, BBAs $m^{\Omega_{i,j}}$ should not assign masses to both propositions $\{y_{i,j}\}$ and $\{n_{i,j}\}$.

Until something better turns up, a practical solution consists in choosing a decision by favoring either the perceived objects or the known objects. However, to relativize this problem, it is shown on a particular application described in Section 5, that conflicting decisions can happen in very few cases, less than 1% in this instance.

4 Discussion

4.1 What's new in comparison to Rombaut and Gruyer's approaches?

The approach presented in this paper differs mainly from Rombaut and Gruyer's approaches [18, 12] by regarding two points:

1. the combination of BBAs $m_j^{\Omega_{X_i}} = m^{\Omega_{i,j}\uparrow\Omega_{X_i}}$ and $m_i^{\Omega_{Y_j}} = m^{\Omega_{i,j}\uparrow\Omega_{Y_j}}$;
2. the decision-making process.

In both Rombaut's approach [18] and Gruyer's approach [12], BBAs $m_j^{\Omega_{X_i}}$ and $m_i^{\Omega_{Y_j}}$ are not classically conjunctively combined with (14). To simplify the combination and to make it computationally efficient, it is proposed to allocate masses only on singletons and the universe. Thus the following mergers are proposed, $\forall i \in I$:

$$\begin{aligned}
m^{\Omega_{X_i}}_{Rombaut}(\{\emptyset\}) &= m^{\Omega_{X_i}}(\{\emptyset\}) \\
m^{\Omega_{X_i}}_{Rombaut}(\{k\}) &= m^{\Omega_{X_i}}(\{k\}), \quad \forall k \in J, \\
m^{\Omega_{X_i}}_{Rombaut}(\{\star\}) &= m^{\Omega_{X_i}}(\{\star\}) \\
m^{\Omega_{X_i}}_{Rombaut}(\Omega_{X_i}) &= \prod_{j\in J} (m_j^{\Omega_{X_i}}(\Omega_{X_i}) + m_j^{\Omega_{X_i}}(\overline{\{j\}})) \\
&\quad - \prod_{j\in J} m_j^{\Omega_{X_i}}(\overline{\{j\}}).
\end{aligned} \tag{28}$$

In [12], the authors suggest a decision-making system based on BBAs $m^{\Omega_{X_i}}$ and $m^{\Omega_{Y_j}}$ whose focal elements, thanks to Rombaut's combination, are either a singleton or the universe. In outline:

- An association matrix $N \times M$ is built such that each of its elements (i, j) is equal to the product $m^{\Omega_{X_i}}(\{j\}) \times m^{\Omega_{Y_j}}(\{i\})$. Each row i is then associated with a perceived object X_i, and each column j is associated with a known object Y_j.
- If necessary, fictive objects are added to make the latter matrix squared.
- A coupling algorithm, the Hungarian algorithm, is then applied to this matrix, this latter algorithm providing an optimal decision regarding the sum of the beliefs.
- A final treatment deals with the objects appearance.

In the examples presented in [18] and [12], the model presented in this paper and Gruyer's approach lead to the same results.

4.2 About Ristic and Smets' approach

The problem tackled by Ristic and Smets in [17] is somewhat different from the association problem described in this paper. Ristic and Smets consider a given volume of interest containing an unknown number of objects. While sensors we consider give information regarding the associations of each object detected at a time step t, with previous objects detected at a previous time step $t-1$, Ristic and Smets's sensors provide information regarding the class of each object they have detected in the scene, for instance helicopter, airplane, ... The "association problem" they try to solve consists then in determining the number of objects as well as the class of each one. Besides, the appearance and disappearance of objects do not take directly part of their problem. The application of Ristic and Smets' works to our problem is consequently not straightforward.

However, some technical points of this model should be taken into account and investigated.

Following [8], the authors remark that the mass given to the empty set, after conjunctively combining two BBAs expressing themselves on the class of two different objects is equal to the belief that these two objects do not belong to the same class, an idea already present in [1] (multi-sensor fusion for submarine detection) and in [20] (intelligence clustering).

At last, the criteria the authors maximize is based on the plausibility of each possible associations. As justified in [23], the pignistic transformation has been chosen to make the decision in this paper. A first investigation in the direction of the plausibility consists in using the plausibility-probability transformation [16].

5 Results on real data

In this section, the approach presented in this paper (Section 3) is compared to the approach of Rombaut and Gruyer on real data coming from a DV camera placed behind the windshield of a car. This DV camera has a CCD sensor, a 720×576 pixels resolution, an angle ranging from -0.5 to $+0.5$ radians (i.e. approximately $\pm 30°$), and works at 25 images per second ($\Delta_t = .04s$), a filmed image example being presented in Figure 3.

Figure 3: Four vehicles in a selected filmed image.

The video sequence allowing one to compare the two approaches includes 3250 images corresponding to a 130-second playing time. Images contain 1 to 8 objects. During the sequence, 75 distinct objects were manually identified as illustrated in Figure 3, the number of associations to realize being equal to 6800. The ground truth is known, which allows one to compute the good recognition rate of each approach during this sequence.

Distance and angle criteria allow the creation of two belief functions denoted $m^{\Omega_{i,j}}_{distance}$ and $m^{\Omega_{i,j}}_{angle}$, regarding the association between each perceived object X_i and known object Y_j.

The distance was estimated as a function of the height and the width in pixels of the object observed in the scene thanks to an interpolation method illustrated in Figure 4.

On the other hand, the angle between two objects is computed from the gravity center of the perceived object in the image (Figure 3).

The measurements provided are very noisy. For instance, there can be a variation of $20m$ for the same object from an image to the next one. Likewise, angle variations can be as high as 100%, from $0.01rd$ to $0.02rd$ for two consecutive measurements of the same object.

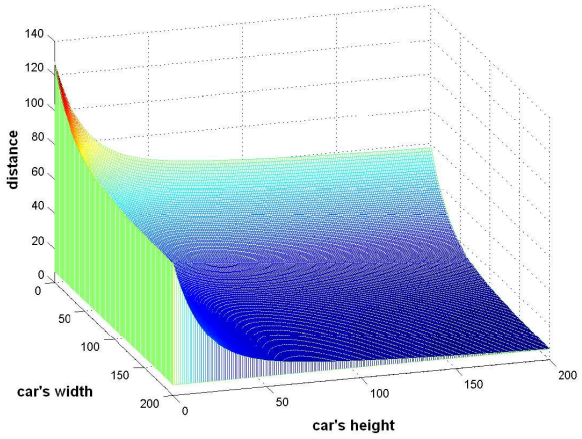


Figure 4: Interpolation function giving the distance in meter depending on the height and the width in pixels of the object in the scene.

In this application, masses are fixed in the following way:

$$\begin{cases} m^{\Omega_{i,j}}(\{y_{i,j}\}) & = \beta\ \phi_{i,j}(e_{i,j}) \\ m^{\Omega_{i,j}}(\{n_{i,j}\}) & = \beta\ (1 - \phi_{i,j}(e_{i,j})) \\ m^{\Omega_{i,j}}(\Omega_{i,j}) & = 1 - \beta \end{cases} \quad (29)$$

where:

- $0 < \beta < 1$ is a constant representing the degree of reliability of the source of information (cf the discounting operation [21, page 252], and [14, 15] for other correction mechanisms).

- $\phi_{i,j}(.)$ is a monotone decreasing function s.t. $\phi_i(0) = 1$ and $\lim_{e\to\infty} \phi_i(e) = 0$;
- $e_{i,j}$ is the dissimilarity measure between the perceived object X_i and the known object Y_j, which means the difference of distance and the difference of angle in this application.

The function $\phi_{i,j}$ is chosen as follows [6]:

$$\phi_{i,j}(e_{i,j}) = exp^{(-(e_{i,j})^2)}. \quad (30)$$

Constant β being fixed at 0.9, these two belief mass functions are combined thanks to the Dempster's rule of combination to obtain a mass function $m^{\Omega_{i,j}}$:

$$m^{\Omega_{i,j}} = m^{\Omega_{i,j}}_{distance} \oplus m^{\Omega_{i,j}}_{angle} \quad \forall i \in I, \forall j \in J\ . \quad (31)$$

The association model presented in Section 3 only need one BBA expressing the information regarding the association between object X_i and object Y_j. In this application, we are lucky enough to have two information sources. Thus these two pieces of information are firstly combined using a well justified rule for the combination of two distinct sources. The choice to combine theses sources at this step, and the choice of the rule have been left for further study.

In Figure 5, the good recognition rate of the two approaches presented in this paper obtained in this video sequence is represented as a function of the rejection cost (Section 2.3). It can be observed that as soon as the rejection cost becomes greater than 0, the good recognition rates obtained with the conjunctive combination are greater than those obtained with Rombaut's combination, which is recalled to be also used in Gruyer's approach.

Let us note that the decisions have been computed on the basis of the perceived objects. As mentioned in Section 3.4, these decisions are not necessary identical with those computed from the known objects point of view. However, as illustrated in Figure 6, this conflicting decision rate remains very low in this application (from 0% to less than 1% depending on the rejection cost). Let us also recall that, as illustrated at the end of Example 2, the introduction of a rejection cost enhances the appearance of conflicting decisions.

6 Conclusion and prospects

In this paper, a modeling of the association step problem in obstacle tracking in the belief function framework has been presented. In particular, it has been shown how tools from the theory of belief functions such as the vacuous extension, the conjunctive combination rule and the pignistic transformation can

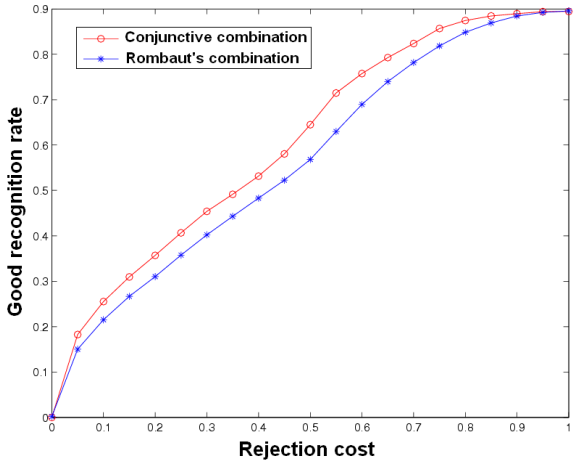


Figure 5: Good recognition rate in function of the rejection cost.

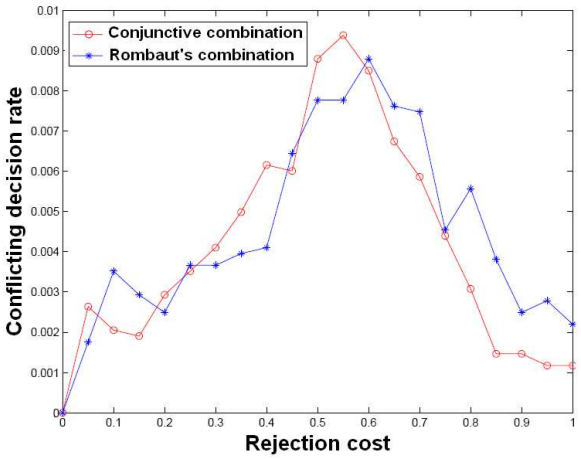


Figure 6: Conflicting decision rate in function of the rejection cost.

be applied. Validated on real data, this approach can perform better good recognition rates than Rombaut's initial approach as soon as a rejection cost is introduced.

Concerning the prospects, even if it concerns a reduce number of cases, a more convincing solution has to be brought regarding the resolution of the possible conflicting decisions between the perceived and known objects points view. This points is currently under investigation.

The decomposition of the BBAs [9] expressing the beliefs regarding the associations between known objects and perceived objects could also be studied in order to use a more adapted rule.

Subsequently, this approach should be enhanced by introducing information coming from the tracking of

vehicles at time steps preceding the current analysis.

Acknowledgements

The authors are very grateful to the anonymous reviewers for their helpful and constructive comments. These works have been financed by the French region Nord-Pas de Calais under the project CISIT (Campus International pour la Sécurité et l'Intermodalité des Transports).

References

[1] A. Ayoun, Ph. Smets. Data association in multitarget detection using the transferable belief model. *International Journal of Intelligent Systems*, 16(10):1167–1182, 2001.

[2] Y. Bar-Shalom. *Multitarget/Multisensor Tracking: Applications and Advances vol III*, Artech House, 2000.

[3] S. Blackman and R. Popoli. *Design and Analysis of Modern Tracking Systems*, Artech House, 1999.

[4] M.H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.

[5] A. Dempster. A generalization of Bayesian inference. *Journal of Royal Statistical Society*, B 30:205-247, 1968.

[6] T. Denœux. A k-nearest neighbour classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, 25(5):804-813, 1995.

[7] T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7):1095–1107, 1997.

[8] T. Denoeux and M. Masson. EVCLUS: Evidential Clustering of Proximity Data. *IEEE Transactions on Systems, Man and Cybernetics B*, 34(1):95–109, 2004.

[9] T. Denœux. Conjunctive and Disjunctive Combination of Belief Functions Induced by Non Distinct Bodies of Evidence. *Artificial Intelligence*, 172:234-264, 2008.

[10] D. Dubois, H. Prade, and Ph. Smets. Representing partial ignorance. *IEEE Transactions on Systems, Man and Cybernetics*, 26(3):361–377, 1996.

[11] M. El Najjar and P. Bonnifait. A road-matching method for precise vehicle localization using belief theory and Kalman filtering. *Autonomous Robots*, 19(2):173-191, 2005.

[12] D. Gruyer, C. Royère, R. Labayrade and D. Aubert. Credibilistic multi sensor fusion for real time application, application to obstacle detection and tracking. *IEEE Int. Conf. on Advanced Robotics, ICAR'2003*, paper P366, 2003.

[13] Y. Lemeret, E. Lefevre and D. Jolly. Improvement of an association algorithm for obstacle tracking. *Information Fusion*, 9(2):234-245, 2008.

[14] D. Mercier, B. Quost and T. Denœux. Refined modeling of sensor reliability in the belief function framework using contextual discounting. *Information Fusion*, 9(2):246–258, 2008.

[15] D. Mercier, T. Denœux and M.-H. Masson. Belief function correction mechanisms. *Studies in Fuzziness and Soft Computing*. To appear.

[16] B.R. Cobb, P.P. Shenoy. On the plausibility transformation method for translating belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3):314–330, 2006.

[17] B. Ristic and Ph. Smets. Global cost of assignment in the TBM framework for association of uncertain ID reports. *Aerospace Science and Technology*, 11(4):303–309, 2007.

[18] M. Rombaut. Decision in multi-obstacle matching process using the theory of belief. *Advances in Vehicle Control and Safety, AVCS98*, pp. 63-68, 1998.

[19] M. Rombaut and V. Cherfaoui. Decision making in data fusion using Dempster-Shafer's theory. *Symposium on Intelligent Components and Instrumentation for Control Applications*, 1997.

[20] J. Schubert. Managing inconsistent intelligence. In *Proceedings of the 3rd International Conference on Information Fusion, FUSION'2000*, pp. TuB4/10–16, Paris, France, 2000.

[21] G. Shafer. *A Mathematical Theory Of Evidence*. Princeton University Press, Princeton, N.J., 1976.

[22] Ph. Smets. Belief functions: the disjunctive rule of combination and the generalized bayesian theorem. *International Journal of Approximate Reasoning*, 9:1–35, 1993.

[23] Ph. Smets. Decision making in the TBM: the necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, 38(2):133–147, 2005.

[24] Ph. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–243, 1994.

Natural extension as a limit of regular extensions

Enrique Miranda
University of Oviedo (Spain)
mirandaenrique@uniovi.es

Marco Zaffalon
IDSIA, Lugano (Switzerland)
zaffalon@idsia.ch

Abstract

This paper is devoted to the extension of conditional assessments that satisfy some consistency criteria, such as weak or strong coherence, to further domains. In particular, we characterise the natural extension of a number of conditional lower previsions on finite spaces, by showing that it can be calculated as the limit of a sequence of conditional lower previsions defined by regular extension. Our results are valid for conditional lower previsions with non-linear domains, and allow us to give an equivalent formulation of the notion of coherence in terms of credal sets.

Keywords. Coherent lower previsions, weak and strong coherence, natural extension, regular extension, desirable gambles.

1 Introduction

A distinctive feature of subjective (or personal) probability is its being founded on a notion of self-consistency, which is often called *coherence*. Loosely speaking, coherence requires that the logical implications of any part of the assessments made cannot force a change in the remaining assessments. Since de Finetti [4], coherence is at the heart of precise personal probability, such as the Bayesian theory; later work by Williams [18] and Walley [14] has made of it the central notion also for imprecisely specified probabilities. Nowadays coherence is largely used in imprecise probability to guide research in *coherent lower previsions*.

A coherent lower prevision formalises a subject's beliefs about *gambles*, which represent uncertain rewards. In this it implements a 'direct' approach to belief assessment. The more traditional approach made of probability measures, can be regarded as dual to the former: in fact, a coherent lower prevision is a model equivalent to a closed convex set of probability measures, also called *credal set* after Levi [8].

Despite this equivalence, coherence is used almost exclusively together with coherent lower previsions instead of with sets of probability measures. The reason is that coherence has been, somewhat naturally, formulated only in terms of gambles and lower previsions. This is unfortunate as it prevents coherent modelling to be easily carried over to traditional probability, which is the framework much more commonly used and understood.

With this paper we make a step in the direction of expressing coherence in a dual form. We focus in particular on Walley's notion of *strong* (or *joint*) coherence [14, Section 7.1.4]. We work with variables $X_1, \ldots, X_n$ that are assumed to take finitely many values, and furthermore assume to be given m coherent lower previsions $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ that express beliefs about them.

What we show, loosely speaking, is that $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ are jointly coherent if and only if there is a sequence of unconditional lower previsions $\underline{P}_\epsilon(X_1, \ldots, X_n)$, $\epsilon \in \mathbb{R}^+$, such that $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ are the limit, when ϵ goes to zero, of conditional assessments derived from $\underline{P}_\epsilon(X_1, \ldots, X_n)$. This means that by applying Bayes' rule whenever possible to the mass functions in the set equivalent to $\underline{P}_\epsilon(X_1, \ldots, X_n)$, we recover the original conditional lower previsions in the limit.

This result relates coherence to the existence of a sequence of joint unconditional credal sets for $X_1, \ldots, X_n$. This is interesting because traditionally in precise probability self-consistency is often intended as the existence of a global model: a joint mass function for $X_1, \ldots, X_n$. In a sense our results confirm that having a global model is essential for coherence, but also that we need more than that. This is related to the existence of events which are assigned lower probability zero through the original assessments: in fact, a single global model cannot detect in general the inconsistencies that may arise on top of zero probabil-

ities (see [12, Theorem 1], [10]); the sequence, instead, can.

But the sequence does more than that: any least-committal coherent inference that logically follows from $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ alone, can be equivalently done again applying Bayes' rule to the elements of the sequence: in other words, the so-called *natural extension* of the original assessments to a new lower prevision $\underline{P}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$ is nothing else but the application of Bayes' rule to $\underline{P}_\epsilon(X_1, \ldots, X_n)$ with $\epsilon \to 0$. This appears to give coherent inference a very accessible interpretation from the dual perspective of traditional probability.

We should mention that ours is not the first work in this direction. A very interesting paper by Walley, Pelessoni and Vicig [17] has introduced the same ideas we consider in Section 5 while restricting the attention to events (rather than gambles) and therefore to finitely many probabilistic assessments. Our work builds upon those ideas, while generalising them so that the only actual restriction now is the finiteness of the spaces.

In particular, we are not limited to what Walley calls *finitely generated* sets of gambles [14, Section 4.2]: we consider also credal sets that cannot be summarised by any finite set of mass functions, or equivalently, that have infinitely many *extreme points* (remember that credal sets are convex). This infinitary dimension has required us to use technical tools other than those in [17], and this has made the technical development somewhat more involved.

We begin by recalling some introductory notions about coherent lower previsions in Section 2. In Section 3 we give new characterisations of avoiding uniform and partial loss, while in Section 4 we deal with weak coherence. In this case, we focus on extending weakly coherent lower previsions $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ to new ones, and an interesting result here is that this extension can be made through conditioning the smallest unconditional prevision $\underline{P}(X_1, \ldots, X_n)$ that is weakly coherent with them. In doing so, we give a number of side results that generalise previous work to domains made of arbitrary sets of gambles. In Section 5 we are finally able to address the main problems described above. Moreover, we relate the need of the sequence $\underline{P}_\epsilon(X_1, \ldots, X_n)$, $\epsilon \in \mathbb{R}^+$, to the existence of events of lower probability equal to zero. This shows also that the natural extension of a number of strongly coherent lower previsions cannot be done, as in the case of weak coherence, through the smallest unconditional joint lower prevision that is coherent with them.

2 Coherence notions on finite spaces

2.1 The behavioural interpretation

Let us give a short introduction to the concepts and results from the behavioural theory of imprecise probabilities that we shall use in the rest of the paper. We refer to [14] for an in-depth study of these and other properties, and to [9] for a brief survey.

Given a possibility space Ω, a *gamble* is a bounded real-valued function on Ω. This function represents a random reward $f(\omega)$, which depends on the a priori unknown value ω of Ω. We shall denote by $\mathcal{L}(\Omega)$ the set of all gambles on Ω. A *lower prevision* $\underline{P}$ is a real functional defined on some set of gambles $\mathcal{K} \subseteq \mathcal{L}(\Omega)$. It is used to represent a subject's supremum acceptable buying prices for these gambles, in the sense that for all $\epsilon > 0$ and all f in $\mathcal{K}$ the subject is disposed to accept the uncertain reward $f - \underline{P}(f) + \epsilon$.

From any lower prevision $\underline{P}$ we can define an upper prevision $\overline{P}$ using conjugacy: $\overline{P}(f) = -\underline{P}(-f)$ for any gamble f. $\overline{P}(f)$ can be interpreted as the infimum acceptable selling price for the gamble f. Because of this relationship, it will suffice for the purposes of this paper to concentrate on lower previsions.

Consider variables $X_1, \ldots, X_n$, taking values in respective *finite* sets $\mathcal{X}_1, \ldots, \mathcal{X}_n$. For any non-empty subset $J \subseteq \{1, \ldots, n\}$ we shall denote by X_J the (new) variable $X_J := (X_j)_{j \in J}$, which takes values in the product space $\mathcal{X}_J := \times_{j \in J} \mathcal{X}_j$. This means that X_J is made of variables that are *logically independent*. We shall also use the notation $\mathcal{X}^n$ for $\mathcal{X}_{\{1,\ldots,n\}}$. In the current formulation made by variables, $\mathcal{X}^n$ is just the definition of the possibility space Ω.

Definition 1. Let J be a subset of $\{1, \ldots, n\}$, and let $\pi_J : \mathcal{X}^n \to \mathcal{X}_J$ be the so-called *projection operator*, i.e., the operator that drops the elements of a vector in $\mathcal{X}^n$ that do not correspond to indexes in J. A gamble f on $\mathcal{X}^n$ is called $\mathcal{X}_J$*-measurable* when for all $x, y \in \mathcal{X}^n$, $\pi_J(x) = \pi_J(y)$ implies that $f(x) = f(y)$.

There is a one-to-one correspondence between the gambles on $\mathcal{X}^n$ that are $\mathcal{X}_J$-measurable and the gambles on $\mathcal{X}_J$. We shall denote by $\mathcal{K}_J$ the set of $\mathcal{X}_J$-measurable gambles.

Consider two disjoint[1] subsets O, I of $\{1, \ldots, n\}$, with $O \neq \emptyset$. $\underline{P}(X_O|X_I)$ represents a subject's behavioural dispositions about the gambles that depend on the outcome of the variables $\{X_j, j \in O\}$, after coming to know the outcome of the variables $\{X_j, j \in I\}$. As such, it is defined at most on gambles that depend on the values of the variables in $O \cup I$ only, i.e., on

[1] That they are taken disjoint is not restrictive. This can be shown using *separate coherence*, given in Definition 2.

the set $\mathcal{K}_{O\cup I}$ of the $\mathcal{X}_{O\cup I}$-measurable gambles on $\mathcal{X}^n$. Given such a gamble f and $z \in \mathcal{X}_I$, $\underline{P}(f|X_I = z)$ represents a subject's supremum acceptable buying price for the gamble f, provided he later comes to know that the variable X_I took the value z (and nothing else). When there is no possible confusion about the variables involved in the lower prevision, we shall use the notation $\underline{P}(f|z)$ for $\underline{P}(f|X_I = z)$. We can define the gamble $\underline{P}(f|X_I)$, which takes the value $\underline{P}(f|z)$ on the elements of $\pi_I^{-1}(z)$ for every $z \in \mathcal{X}_I$. This is a *conditional lower prevision*.

We shall also use the notations

$$G(f|z) := \pi_I^{-1}(z)(f - \underline{P}(f|z))$$
$$G(f|X_I) := \sum_{z\in\mathcal{X}_I} G(f|z) = f - \underline{P}(f|X_I)$$

for all $f \in \mathcal{K}_{O\cup I}$ and all $z \in \mathcal{X}_I$. In the case of an unconditional lower prevision $\underline{P}$, we shall denote $G(f) := f - \underline{P}(f)$ for any gamble f in its domain. Here, and in the rest of the paper, we shall use A to denote both a set A and its indicator function.

The gambles $G(f|z)$ and $G(f|X_I)$ are *almost-desirable*, in the sense that for every $\epsilon > 0$, the gambles $G(f|z) + \epsilon\pi_I^{-1}(z)$ and $G(f|X_I) + \epsilon$ should be desirable for our subject.

2.2 Consistency notions

These assessments can be made for any disjoint subsets O, I of $\{1, \ldots, n\}$, and therefore it is not uncommon to model a subject's beliefs using a finite number of different conditional previsions. We should verify then that all the assessments modelled by these conditional previsions are coherent with each other. The first requirement we make is that for any disjoint $O, I \subseteq \{1, \ldots, n\}$, the conditional lower prevision $\underline{P}(X_O|X_I)$ defined on a subset $\mathcal{H}_{O\cup I}$ of $\mathcal{K}_{O\cup I}$ should be separately coherent.

Definition 2. A conditional lower prevision $\underline{P}(X_O|X_I)$ with domain $\mathcal{H}_{O\cup I}$ is *separately coherent* if for every $z \in \mathcal{X}_I$, the gamble $\pi_I^{-1}(z)$ belongs to $\mathcal{H}_{O\cup I}$ and $\underline{P}(\pi_I^{-1}(z)|z) = 1$, and moreover

$$\max_{x\in\pi_I^{-1}(z)} \left[\sum_{j=1}^{n} \lambda_j G(f_j|z) - G(f_0|z)\right](x) \geq 0$$

for every $n \in \mathbb{N}, f_j \in \mathcal{H}_{O\cup I}, \lambda_j \geq 0, j = 1, \ldots, n, f_0 \in \mathcal{H}_{O\cup I}$.

It is also useful for this paper to consider the particular case where $I = \emptyset$, that is, when we have unconditional information about the variables X_O. We have then an *(unconditional) lower prevision* $\underline{P}(X_O)$ on a subset $\mathcal{H}_O$ of the set $\mathcal{K}_O$ of $\mathcal{X}_O$-measurable gambles. Separate coherence is called then simply *coherence*, and it holds if and only if

$$\max_{x\in\mathcal{X}^n} \left[\sum_{j=1}^{n} \lambda_j G(f_j) - G(f_0)\right](x) \geq 0 \qquad (1)$$

for every $n \in \mathbb{N}, f_0, f_1, \ldots, f_n \in \mathcal{H}_O, \lambda_1, \ldots, \lambda_n \geq 0$.

Consider now separately coherent conditional lower previsions $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ with respective domains $\mathcal{H}^1, \ldots, \mathcal{H}^m \subseteq \mathcal{L}(\mathcal{X}^n)$, where $\mathcal{H}^j$ is a subset of the set $\mathcal{K}^j$ of $\mathcal{X}_{O_j\cup I_j}$-measurable gambles,[2] for $j = 1, \ldots, m$. There are different ways in which we can guarantee their consistency.

Definition 3. $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ *avoid uniform sure loss* if for every $f_j^k \in \mathcal{H}^j$ and every $\lambda_j^k \geq 0, j = 1, \ldots, m, k = 1, \ldots, n_j$,

$$\max_{x\in\mathcal{X}^n} \left[\sum_{j=1}^{m}\sum_{k=1}^{n_j} \lambda_j^k G_j(f_j^k|X_{I_j})\right](x) \geq 0.$$

A slightly stronger notion is called avoiding partial loss. For this, we define the $\mathcal{X}_I$*-support* $S(f)$ of a gamble f in $\mathcal{K}_{O\cup I}$ as

$$S(f) := \{\pi_I^{-1}(z) : z \in \mathcal{X}_I, f\pi_I^{-1}(z) \neq 0\};$$

i.e., it is the set of conditioning events for which the restriction of f is not identically zero.

Definition 4. $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ *avoid partial loss* if for every $f_j^k \in \mathcal{H}^j$ and every $\lambda_j^k \geq 0$, $j = 1, \ldots, m, k = 1, \ldots, n_j$ such that not all the $\lambda_j^k f_j^k$ are zero gambles,

$$\max_{x\in\bigcup\cup_{j=1}^{m}\cup_{k=1}^{n_j} S_j(\lambda_j^k f_j^k)} \left[\sum_{j=1}^{m}\sum_{k=1}^{n_j} \lambda_j^k G_j(f_j^k|X_{I_j})\right](x) \geq 0,$$

where by $\bigcup\cup_{j=1}^{m}\cup_{k=1}^{n_j} S_j(\lambda_j^k f_j^k)$ we mean the set of elements that belong to some set in $S_j(\lambda_j^k f_j^k)$ for some $j \in \{1, \ldots, m\}, k \in \{1, \ldots, n_j\}$.

The idea behind this notion is that a combination of transactions that are acceptable for our subject should not make him lose utiles. It is based on the rationality requirement that a gamble $f \leq 0$ such that $f < 0$ on some set A should not be desirable.

We next give two notions that generalise the concept of coherence in Eq. (1) to the conditional case:

Definition 5. $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ are *weakly coherent* if for every $f_j^k \in \mathcal{H}^j, \lambda_j^k \geq$

[2] We use $\mathcal{K}^j$ instead of $\mathcal{K}_{O_j\cup I_j}$ in order to alleviate the notation when no confusion is possible about the variables involved.

$0, j = 1, \ldots, m, k = 1, \ldots, n_j$, and for every $j_0 \in \{1, \ldots, m\}, f_0 \in \mathcal{H}^{j_0}, z_{j_0} \in \mathcal{X}_{I_{j_0}}$,

$$\max_{x \in \mathcal{X}^n} \left[\sum_{j=1}^{m} \sum_{k=1}^{n_j} \lambda_j^k G_j(f_j^k | X_{I_j}) - G_{j_0}(f_0 | z_{j_0}) \right](x) \geq 0.$$

With this condition we require that our subject should not be able to raise his supremum acceptable buying price $\underline{P}_{j_0}(f_{j_0}|z_{j_0})$ for a gamble f_{j_0} contingent on z_{j_0} by taking into account other conditional assessments. However, a number of weakly coherent conditional lower previsions can still present some forms of inconsistency with each other. See [14, Chapter 7], [10] and [17] for some discussion and [14, Sect. 7.3.5] and [10, Examples 4 and 7] for examples of weakly coherent conditionals. On the other hand, weak coherence neither implies nor is implied by the notion of avoiding partial loss. Because of these two facts, we consider another notion which is stronger than both, and which is called (*joint* or *strong*) *coherence*:[3]

Definition 6. $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ are *coherent* when for every $f_j^k \in \mathcal{H}^j, \lambda_j^k \geq 0, j = 1, \ldots, m, k = 1, \ldots, n_j$, and for every $j_0 \in \{1, \ldots, m\}, f_{j_0} \in \mathcal{H}^{j_0}, z_{j_0} \in \mathcal{X}_{I_{j_0}}$,

$$\left[\sum_{j=1}^{m} \sum_{k=1}^{n_j} \lambda_j^k G_j(f_j^k | X_{I_j}) - G_{j_0}(f_{j_0} | z_{j_0}) \right](x) \geq 0$$

for some $x \in \bigcup \pi_{I_{j_0}}^{-1}(z_{j_0}) \cup \bigcup_{j=1}^{n} \bigcup_{k=1}^{n_j} S_j(\lambda_j^k f_j^k)$.

Because we are dealing with finite spaces, this notion coincides in the case of linear domains with the one given by Williams in [18]. The coherence of a collection of conditional lower previsions implies their weak coherence; although the converse does not hold in general, it does in the particular case when we only have a conditional and an unconditional lower prevision $\underline{P}(X_O|X_I), \underline{P}$ with domains $\mathcal{H}_{O \cup I}, \mathcal{H}$. If in particular $\mathcal{H}_{O \cup I} = \mathcal{K}_{O \cup I}$ and $\mathcal{H} = \mathcal{L}(\mathcal{X}^n)$, coherence holds if and only if, for all $\mathcal{X}_{O \cup I}$-measurable f and all $z \in \mathcal{X}_I$,

$$\underline{P}(G(f|z)) = 0. \qquad \text{(GBR)}$$

This is called the Generalised Bayes Rule (GBR). When $\underline{P}(z) > 0$, GBR can be used to determine the value $\underline{P}(f|z)$: it is then the *unique* value for which $\underline{P}(G(f|z)) = \underline{P}(\pi_I^{-1}(z)(f - \underline{P}(f|z))) = 0$ holds.

2.3 Linear previsions and envelope theorems

We say that a conditional lower prevision $\underline{P}(X_O|X_I)$ on the set $\mathcal{K}_{O \cup I}$[4] is *linear* if and only if it is separately coherent and moreover $\underline{P}(f + g|z) = \underline{P}(f|z) + \underline{P}(g|z)$ for all $z \in \mathcal{X}_I$ and $f, g \in \mathcal{K}_{O \cup I}$. Conditional linear previsions correspond to the case where a subject's supremum acceptable buying price (lower prevision) coincides with his infimum acceptable selling price (or upper prevision) for any gamble on the domain. When a separately coherent conditional lower prevision $\underline{P}(X_O|X_I)$ is linear we shall denote it by $P(X_O|X_I)$; in the unconditional case, we shall denote it by P and assume that its domain is the set $\mathcal{L}(\mathcal{X}^n)$ of all gambles. The definition of linear prevision implies that in the unconditional case it is just a coherent prevision in de Finetti's sense. In the conditional case, this still holds but it is required that in addition $P(\pi_I^{-1}(z)|z) = 1$ for all $z \in \mathcal{X}_I$. In other words, conditional linear previsions correspond to conditional expectations with respect to a probability. In particular, an unconditional linear prevision P is the expectation with respect to the probability which is the restriction of P to events.

A number of conditional linear previsions are coherent if and only if they avoid partial loss. They are weakly coherent if and only if they avoid uniform sure loss.

Given an unconditional lower prevision $\underline{P}$ with domain $\mathcal{H}$, we shall denote the set of *dominating* linear previsions by $\mathcal{M}(\underline{P}) := \{P : P(f) \geq \underline{P}(f) \ \forall f \in \mathcal{H}\}$. Similarly, for a conditional lower prevision $\underline{P}(X_O|X_I)$ with domain $\mathcal{H}_{O \cup I}$, we define $\mathcal{M}(\underline{P}(X_O|X_I))$ as the set of linear previsions $P(X_O|X_I)$ such that

$$P(f|z) \geq \underline{P}(f|z) \ \forall f \in \mathcal{H}_{O \cup I}, z \in \mathcal{X}_I.$$

Then $\underline{P}$ is coherent if and only if it is the lower envelope of $\mathcal{M}(\underline{P})$, and $\underline{P}(X_O|X_I)$ is separately coherent if and only if it is the lower envelope of $\mathcal{M}(\underline{P}(X_O|X_I))$.

The situation is more complicated when we have more than one conditional lower prevision, as the previous results essentially hold for finite spaces. In [14] Walley proved that when the referential spaces are finite and the domains are linear spaces, coherent $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ are always the envelope of a set $\{P_1^\lambda(X_{O_1}|X_{I_1}), \ldots, P_m^\lambda(X_{O_m}|X_{I_m}) : \lambda \in \Lambda\}$ of dominating coherent conditional linear previsions. In [10], a similar property was established for weak coherence. In Section 4 we shall generalise this second property to arbitrary domains.

2.4 Extensions to further domains

Let $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ be separately coherent conditional lower previsions with domains $\mathcal{H}^i \subseteq \mathcal{K}^i$ for $i = 1, \ldots, m$ and avoiding partial loss.

Their *natural extensions* to the sets $\mathcal{K}^1, \ldots, \mathcal{K}^m$ are

[3] The distinction with the unconditional notion of coherence mentioned above will always be clear from the context.

[4] We shall always assume in this paper that the domain of a conditional linear prevision $P(X_O|X_I)$ is the whole set $\mathcal{K}_{O \cup I}$ of $\mathcal{X}_{O \cup I}$-measurable gambles.

defined, [5] for every $f \in \mathcal{K}^j$ and every $z_j \in \mathcal{X}_{I_j}$, by

$$\underline{E}_j(f|z_j) = \sup\{\alpha : \exists f_j^k \in \mathcal{H}^j, \lambda_j^k \geq 0, \text{ s.t.}$$
$$[\sum_{j=1}^{m}\sum_{k=1}^{n_j} \lambda_j^k G_j(f_j^k|X_{I_j}) - \pi_{I_j}^{-1}(z_j)(f-\alpha)] < 0$$
$$\text{on } \bigcup \cup_{j=1}^{m} \cup_{k=1}^{n_j} S_j(\lambda_j^k f_j^k) \cup \pi_{I_j}^{-1}(z_j)\}. \quad (2)$$

In the context of this paper, where all the conditioning spaces are finite, the natural extensions are the smallest conditional lower previsions which are coherent and dominate $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$. Moreover, they coincide with the initial assessments if and only if $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ are themselves coherent. Otherwise, they 'correct' the initial assessments taking into account the implications of the notions of coherence [11, Prop. 11]. In the rest of the paper we shall consider at some point also the natural extension $\underline{E}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$, for arbitrary disjoint subsets O_{m+1}, I_{m+1} of $\{1, \ldots, n\}$. Doing so amounts to implicitly include in the original set of lower previsions, an additional one $\underline{P}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$ defined on a trivial domain (such as the constant gambles), and then to take the natural extension.

In this paper, we shall also define conditional lower previsions coherently by using the *regular extension*. Given a credal set $\mathcal{M}$ and disjoint O, I, the regular extension $\underline{R}(X_O|X_I)$ is given by

$$\underline{R}(f|z) := \inf\left\{\frac{P(f\pi_I^{-1}(z))}{P(z)} : P \in \mathcal{M}, P(z) > 0\right\}$$

for every $z \in \mathcal{X}_I, f \in \mathcal{K}_{O\cup I}$. This amounts to applying Bayes' rule to the linear previsions in $\mathcal{M}$ whenever possible. The regular extension has been proposed and used a number of times in the literature as an updating rule [2, 3, 5, 6, 14, 15]. See [10] for a comparison with natural extension in the finite case.

3 Characterising avoiding uniform sure loss and avoiding partial loss

Let $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ be separately coherent conditional lower previsions with respective domains $\mathcal{H}^1, \ldots, \mathcal{H}^m$, where $\mathcal{H}^j$ is a (not necessarily linear) subset of the class $\mathcal{K}^j$ of $\mathcal{X}_{O_j\cup I_j}$-measurable gambles.

Our first result is an extension of [12, Prop. 5] to arbitrary domains. It uses the following lemma:

[5] We do not extend $\underline{P}_j(X_{O_j}|X_{I_j})$ beyond the set $\mathcal{K}^j$ of $\mathcal{X}_{O_j\cup I_j}$-measurable gambles as that would not be compatible with the interpretation we have given of $\underline{P}_j(X_{O_j}|X_{I_j})$; yet, it is possible to extend it to $\mathcal{L}(\mathcal{X}^n)$ by considering $\underline{P}(X_{I^c}|X_I)$ instead of $\underline{P}(X_O|X_I)$, and with the same initial domain.

Lemma 1. *Let $\underline{P}, \underline{P}(X_O|X_I)$ be coherent lower previsions with respective domains $\mathcal{L}(\mathcal{X}^n), \mathcal{H}_{O\cup I}$. For every $P \in \mathcal{M}(\underline{P})$ there is some conditional linear prevision $P(X_O|X_I)$ in $\mathcal{M}(\underline{P}(X_O|X_I))$ such that $P, P(X_O|X_I)$ are coherent. Moreover,*

$$\underline{P}(G(f|z)) = 0, \ \underline{P}(G(f|X_I)) \geq 0$$

for every gamble $f \in \mathcal{H}_{O\cup I}$ and every $z \in \mathcal{X}_I$.

Proposition 1. *$\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ avoid uniform sure loss if and only if there are dominating weakly coherent conditional linear previsions with domains $\mathcal{K}^1, \ldots, \mathcal{K}^m$.*

This result will be interesting in Section 4 when we study the smallest dominating weakly coherent lower previsions. It follows that avoiding uniform sure loss is a necessary and sufficient condition for the existence of such lower previsions. Since moreover we shall prove in Theorem 1 that when all the referential spaces are finite weak coherence is preserved by taking lower envelopes, we deduce that a way of computing the smallest dominating weakly coherent lower previsions is to take the lower envelopes of the (non-empty) sets of weakly coherent dominating conditional linear previsions.

On the other hand, it follows from [14, Sec. 8.1] that when all the referential spaces are finite and the domains are linear spaces, the notion of avoiding partial loss is equivalent to the existence of dominating coherent linear conditional previsions. We here generalise the result to non-linear domains.

Lemma 2. *Assume that the conditional lower previsions $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ avoid partial loss, and let $\underline{E}_1(X_{O_1}|X_{I_1}), \ldots, \underline{E}_m(X_{O_m}|X_{I_m})$ be their natural extensions to $\mathcal{K}^1, \ldots, \mathcal{K}^m$. Then $\underline{E}_1(X_{O_1}|X_{I_1}), \ldots, \underline{E}_m(X_{O_m}|X_{I_m})$ are coherent.*

Proposition 2. *$\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ avoid partial loss if and only if there are dominating coherent conditional linear previsions with domains $\mathcal{K}^1, \ldots, \mathcal{K}^m$.*

The notions of avoiding partial and uniform sure loss constitute a generalisation, to conditional assessments, of a consistency notion for unconditional lower previsions, called *avoiding sure loss*. It is established in [14, Thm. 3.3.3] that avoiding sure loss is equivalent to the existence of a dominating coherent linear prevision, and therefore can be seen as a minimal consistency requirement.

When we move towards conditional lower previsions, we have seen in Section 2 that there are two ways of extending the notion of coherence of lower previsions, called weak and (strong) coherence. What we have proved by means of Propositions 1 and 2 is that

avoiding uniform sure and partial loss are the respective counterparts of avoiding sure loss for each of these two extensions.

We conclude the section with another characterisation of avoiding partial loss, where we can find some of the ideas we shall use in our approximation of the natural extension in Section 5.

Proposition 3. *$\underline{P}_1(X_{O_1}|X_{I_1}),\dots,\underline{P}_m(X_{O_m}|X_{I_m})$ avoid partial loss if and only if for all $\epsilon>0$, $f_j^k\in\mathcal{H}^j$, $\lambda_j^k\geq 0$, $j=1,\dots,m$, $k=1,\dots,n_j$ such that not all products $\lambda_j^k f_j^k$ are zero gambles, it holds that*

$$\max_{x\in\mathcal{X}^n}\left[\sum_{j=1}^m\sum_{k=1}^{n_j}\lambda_j^k(G_j(f_j^k|X_{I_j})+\epsilon S_j(f_j^k))\right](x)>0.$$

Hence, by introducing these ϵ-terms, we can replace the maximum on the union of the supports with a maximum on $\mathcal{X}^n$. We shall relate this later to the weak coherence of some approximations of our conditional lower previsions.

4 Extensions of weakly coherent conditionals

We focus next on the notion of weak coherence of a number of conditional lower previsions. We begin by giving a characterisation of weak coherence and determining the smallest (unconditional) coherent lower prevision which is weakly coherent with a number of conditionals. This extends [10, Thms. 2 and 3] to arbitrary domains:

Theorem 1. *Let $\underline{P}_1(X_{O_1}|X_{I_1}),\dots,\underline{P}_m(X_{O_m}|X_{I_m})$ be separately coherent conditional lower previsions with domains $\mathcal{H}^1,\dots,\mathcal{H}^m$. The following are equivalent:*

(WC1) They are weakly coherent.

(WC2) They are the lower envelopes of a class of weakly coherent conditional linear previsions, $\{P_1^\lambda(X_{O_1}|X_{I_1}),\dots,P_m^\lambda(X_{O_m}|X_{I_m}):\lambda\in\Lambda\}$.

(WC3) There is a coherent lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X}^n)$ which is weakly coherent with them.

(WC4) There is a coherent lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X}^n)$ which is pairwise coherent with them.

Moreover, the smallest coherent lower prevision in (WC3) and (WC4) is given, for any gamble f on $\mathcal{X}^n$, by

$$\underline{P}(f)=\sup\{\alpha:\exists f_j^k\in\mathcal{H}^j,\lambda_j^k\geq 0,\ s.t.\ \max_{x\in\mathcal{X}^n}[\sum_{j=1}^m\sum_{k=1}^{n_j}\lambda_j^k G(f_j^k|X_{I_j})-(f-\alpha)](x)<0\}. \quad (3)$$

We summarise the relationships between the different consistency conditions when all the referential spaces are finite in the following figure.

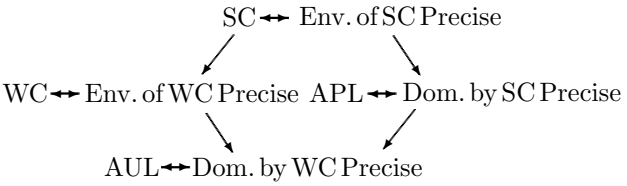


Figure 1: Equivalences and implications between consistency concepts analysed in the paper. Keys: SC = strongly coherent; WC = weakly coherent; AUL = avoiding uniform sure loss; APL = avoiding partial loss; Env. = envelope; Dom. = dominated.

Under some conditions, the functional we just defined is also the natural extension of a number of conditional lower previsions:

Corollary 1. *$\underline{P}$ is the smallest coherent lower prevision which is coherent with $\underline{P}_1(X_{O_1}|X_{I_1}),\dots,\underline{P}_m(X_{O_m}|X_{I_m})$ if and only if these conditional previsions are coherent.*

It is useful at this point to compare the functional $\underline{P}$ defined in Eq. (3) with the unconditional natural extension $\underline{E}$ that we should define using Eq. (2). In order to do this, we should consider $O_{m+1}=\{1,\dots,n\}, I_{m+1}=\emptyset$ and add $\underline{P}(X_{O_{m+1}})$ to our set of gambles with the trivial domain given by the constant gambles. For this discussion to make sense, we are going to assume also that $\underline{P}_1(X_{O_1}|X_{I_1}),\dots,\underline{P}_m(X_{O_m}|X_{I_m})$ avoid partial loss and are weakly coherent.

We see from [11, Theorem 12] that in that case the functionals $\underline{P}$ and $\underline{E}$ coincide. Hence, the unconditional natural extension $\underline{E}$ is the smallest unconditional lower prevision which is weakly coherent with $\underline{P}_1(X_{O_1}|X_{I_1}),\dots,\underline{P}_m(X_{O_m}|X_{I_m})$; and as we have proven in Corollary 1, it is coherent with them if and only if the initial assessments are coherent. A sufficient condition for the coherence of $\underline{P}_1(X_{O_1}|X_{I_1}),\dots,\underline{P}_m(X_{O_m}|X_{I_m})$ when their domains are $\mathcal{K}^1,\dots,\mathcal{K}^m$ is that $\underline{P}(z_j)>0$ for all $z_j\in\mathcal{X}_{I_j}$ and for all $j=1,\dots,m$ [10, Thm. 11]. On the other hand, in [10, Example 2] we can find an example of assessments which avoid partial loss and are weakly coherent, but are not coherent.

Assume now that we have weakly coherent $\underline{P}_1(X_{O_1}|X_{I_1}),\ldots,\underline{P}_m(X_{O_m}|X_{I_m})$, and that given disjoint O_{m+1}, I_{m+1}, we want to determine the smallest conditional lower prevision $\underline{P}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$ which is weakly coherent with the rest. Our next result shows that it suffices to go through the unconditional lower prevision $\underline{P}$ given by Eq. (3):

Theorem 2. *The smallest conditional lower prevision $\underline{P}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$ with domain $\mathcal{K}^{m+1}$ which is weakly coherent with $\underline{P}_1(X_{O_1}|X_{I_1}),\ldots,\underline{P}_m(X_{O_m}|X_{I_m})$ is given, for every $f\in\mathcal{K}^{m+1}, z_{m+1}\in\mathcal{X}_{I_{m+1}}$, by $\underline{P}_{m+1}(f|z_{m+1}):=$*

$$\begin{cases}\min_{x\in\pi^{-1}_{I_{m+1}}(z_{m+1})} f(x) & \text{if } \underline{P}(z_{m+1})=0\\ \min\{P(f|z_{m+1}): P\geq\underline{P}\} & \text{otherwise,}\end{cases} \tag{4}$$

where $\underline{P}$ is given by Eq. (3).

This stresses once more the fact that the informative content of a number of weakly coherent lower previsions is preserved by summarising them with an unconditional lower prevision. Note moreover that if $\underline{P}_{m+1}(z_{m+1})>0$ the conditional lower prevision $\underline{P}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$ is uniquely determined from $\underline{P}$ by the Generalised Bayes Rule.

Our final result in this section shows that we can use the definition of natural extension to obtain a conditional lower prevision which is weakly coherent with a number of assessments.

Proposition 4. *Consider weakly coherent $\underline{P}_1(X_{O_1}|X_{I_1}),\ldots,\underline{P}_m(X_{O_m}|X_{I_m})$ with domains $\mathcal{H}^1,\ldots,\mathcal{H}^m$, and let $\underline{E}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$ be defined on $\mathcal{K}^{m+1}$ by Eq.* (2). *Then $\underline{E}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$ is weakly coherent with $\underline{P}_1(X_{O_1}|X_{I_1}),\ldots,\underline{P}_m(X_{O_m}|X_{I_m})$.*

In particular, if $\underline{P}_1(X_{O_1}|X_{I_1}),\ldots,\underline{P}_m(X_{O_m}|X_{I_m})$ are coherent, it follows from the results in [11] that $\underline{E}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$ is the smallest conditional lower prevision that is coherent with them. It may be strictly greater than the conditional lower prevision $\underline{P}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$ derived in Eq. (4). An instance of such a situation can be found in [16, Example 8]; it can be checked that the smallest weakly coherent conditional lower prevision derived from the assessments in the example is vacuous.

This shows on the one hand that the notion of weak coherence is indeed too weak to fully capture the behavioural implications of our assessments, and on the other that the natural extension cannot be derived in general from the unconditional lower prevision $\underline{P}$. In the following section, we get around this problem by showing: (i) that we can instead derive it using a *sequence* of unconditional lower previsions that converges to $\underline{P}$ and (ii) that in some cases it coincides with the weakly coherent natural extension.

5 Natural extension as a limit of regular extensions

Let $\underline{P}_1(X_{O_1}|X_{I_1}),\ldots,\underline{P}_m(X_{O_m}|X_{I_m})$ now be separately coherent conditional lower previsions with domains $\mathcal{H}^j\subseteq\mathcal{K}^j$ for $j=1,\ldots,m$. We shall assume that they are weakly coherent and avoid partial loss, but they are not necessarily coherent. Our goal in this section is to characterise their natural extension $\underline{E}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$ given by Eq. (2).

Although in general we shall assume that the index $m+1$ does not belong to $\{1,\ldots,m\}$ (and then we have to include among the original assessments a conditional lower prevision $\underline{P}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$ defined on the set of constant gambles), the results are still valid if what we study is the natural extension of one of our assessments $\underline{P}_j(X_{O_j}|X_{I_j})$ to $\mathcal{K}^j$.

We shall prove later (in Theorem 3) that this natural extension can be computed as a limit of regular extensions. In order to do this, we are going to consider a sequence of credal sets which are compatible with conditional lower previsions which converge point-wise to $\underline{P}_1(X_{O_1}|X_{I_1}),\ldots,\underline{P}_m(X_{O_m}|X_{I_m})$. For every $\epsilon>0$, let $\mathcal{M}(\epsilon)$ be the set of linear previsions satisfying

$$P(f_j\pi^{-1}_{I_j}(z_j))\geq P(z_j)(\underline{P}_j(f_j|z_j)-\epsilon R(f_j)) \tag{5}$$

for every $f_j\in\mathcal{H}^j, z_j\in\mathcal{X}_{I_j}, j=1,\ldots,m$, where $R(f_j)=\max f_j-\min f_j$ is the range of the gamble f_j. Let us also consider the set of gambles

$$\mathcal{V}_\epsilon:=\{f\geq\sum_{j=1}^m\sum_{k=1}^{n_j}\lambda^k_j(G_j(f^k_j|X_{I_j})+\epsilon R(f^k_j)S_j(f^k_j)) \text{ for some } f^k_j\in\mathcal{H}^j,\lambda^k_j\geq 0\}, \tag{6}$$

where, with a certain abuse of notation, $S_j(f^k_j)$ is used to denote the indicator function of the set of elements which belong to some set in $S_j(f^k_j)$.

For $\epsilon=0$ we obtain the set $\mathcal{M}(0)$ of linear previsions P such that

$$P(f_j\pi^{-1}_{I_j}(z_j))\geq P(z_j)\underline{P}_j(f_j|z_j) \tag{7}$$

for all $f_j\in\mathcal{H}^j, z_j\in\mathcal{X}_{I_j}, j=1,\ldots,m$, and the set of gambles

$$\mathcal{V}:=\{f\geq\sum_{j=1}^m\sum_{k=1}^{n_j}\lambda^k_jG_j(f^k_j|X_{I_j}) \text{ for some } f^k_j\in\mathcal{H}^j,\lambda^k_j\geq 0\}. \tag{8}$$

It follows from their definition that $\mathcal{V}_\epsilon \subseteq \mathcal{V}$ and $\mathcal{M}(0) \subseteq \mathcal{M}(\epsilon)$ for any $\epsilon > 0$. Since the gamble constant on 0 belongs to $\mathcal{V}_\epsilon$ for all $\epsilon \geq 0$, we deduce that these sets of gambles are non-empty. On the other hand, it follows that $\mathcal{M}(\epsilon)$ are convex sets of linear previsions for all $\epsilon > 0$. $\mathcal{M}(0)$ (and therefore also $\mathcal{M}(\epsilon)$ for all $\epsilon > 0$) is non-empty because the conditional lower previsions $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ are weakly coherent. This follows from the following proposition. Let $\underline{P}_\epsilon$ denote the lower envelope of the credal set $\mathcal{M}(\epsilon)$, and $\underline{P}_0$ the lower envelope of $\mathcal{M}(0)$.

Proposition 5. *Consider weakly coherent $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ with respective domains $\mathcal{H}^1, \ldots, \mathcal{H}^m$ and avoiding partial loss.*

1. *For any $\epsilon \geq 0$, $\mathcal{M}(\epsilon) = \{P : P(f) \geq 0 \ \forall f \in \mathcal{V}_\epsilon\}$, and $\{f : P(f) \geq 0 \ \forall P \in \mathcal{M}(\epsilon)\} = \overline{\mathcal{V}}_\epsilon$, where the closure is taken in the topology of uniform convergence.*
2. *$\mathcal{M}(0) = \cap_{\epsilon>0}\mathcal{M}(\epsilon) = \mathcal{M}(\underline{P})$, where $\underline{P}$ is the coherent lower prevision given by Eq. (3).*
3. *$\underline{P}_0 = \sup_{\epsilon>0} \underline{P}_\epsilon = \underline{P}$.*

In the particular case of precise assessments (i.e., conditional linear previsions) we can go a bit further. In this case, and in analogy with the situation in the unconditional case, we can show that events provide all the information we need. Note also that in the linear case the notion of avoiding partial loss is equivalent to coherence (and implies therefore weak coherence).

Proposition 6. *Consider coherent $P_1(X_{O_1}|X_{I_1}), \ldots, P_m(X_{O_m}|X_{I_m})$ with domains $\mathcal{K}^1, \ldots, \mathcal{K}^m$. Let $\mathcal{V}_\epsilon$ be the set of gambles given by Eq. (6), and $\mathcal{M}(\epsilon)$ be the credal set given by Eq. (5). Let us denote moreover by $\mathcal{V}_\epsilon^A, \mathcal{M}_\epsilon^A$ the corresponding sets determined by the restrictions to events of $P_1(X_{O_1}|X_{I_1}), \ldots, P_m(X_{O_m}|X_{I_m})$.*

1. *For every $\epsilon > 0$, $\overline{\mathcal{V}}_\epsilon \subseteq \mathcal{V}_{\epsilon_1}^A$, where $\epsilon_1 = \frac{\epsilon}{\max_j |\mathcal{X}_{O_j}|}$, and as a consequence $\cup_\epsilon \mathcal{V}_\epsilon = \cup_\epsilon \mathcal{V}_\epsilon^A = \cup_\epsilon \overline{\mathcal{V}}_\epsilon$.*
2. *$\mathcal{M}(\epsilon) \supseteq \mathcal{M}_{\epsilon_1}^A$, whence $\cap_\epsilon \mathcal{M}(\epsilon) = \cap_\epsilon \mathcal{M}_\epsilon^A$.*

This result will be very useful for us because it allows us to connect our results with the ones established in [17] for the particular case of conditional lower previsions defined on events. The case of events is also interesting because the sets of desirable gambles we use are finitely generated, and this makes it easier to apply separation results.

Now that we have clarified a bit the structure of the sets $\mathcal{M}(\epsilon), \mathcal{V}_\epsilon$, we explore how they can be used to characterise the conditional natural extension.

Proposition 7. *Consider $f \in \mathcal{K}^{m+1}$ and $z_{m+1} \in \mathcal{X}_{I_{m+1}}$. Then $\sup\{\mu : \pi^{-1}_{I_{m+1}}(z_{m+1})(f - \mu) \in \cup_\epsilon \mathcal{V}_\epsilon\} = \underline{E}_{m+1}(f|z_{m+1}) \leq \sup\{\mu : \pi^{-1}_{I_{m+1}}(z_{m+1})(f - \mu) \in \mathcal{V}\}$, where $\mathcal{V}$ is given by Eq. (8).*

For every $\epsilon > 0$, let us define $\underline{R}^\epsilon_{m+1}(f|z_{m+1})$ from $\mathcal{M}(\epsilon)$ by regular extension, i.e., let it be given by

$$\inf\{P(f|z_{m+1}) : P \in \mathcal{M}(\epsilon), P(z_{m+1}) > 0\}. \quad (9)$$

The first thing we have to prove is that this definition makes sense.

Proposition 8. *For every $z_{m+1} \in \mathcal{X}_{I_{m+1}}$ and every $\epsilon > 0$, there is some $P \in \mathcal{M}(\epsilon)$ s.t. $P(z_{m+1}) > 0$.*

Since the credal set $\mathcal{M}(\epsilon)$ does not increase as ϵ converges to zero, we deduce that the conditional lower previsions $\underline{R}^\epsilon_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$ given by Eq. (9) do not decrease as ϵ goes to zero. We can thus consider

$$\underline{F}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}}) := \lim_{\epsilon\to 0} \underline{R}^\epsilon_{m+1}(X_{O_{m+1}}|X_{I_{m+1}}),$$

the limit of these conditional lower previsions.

In analogy with Proposition 7, we can characterise $\underline{F}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$ in terms of desirable gambles:

Lemma 3. *For every $f \in \mathcal{K}^{m+1}, z_{m+1} \in \mathcal{X}_{I_{m+1}}$, $\underline{F}_{m+1}(f|z_{m+1}) = \sup\{\mu : \pi^{-1}_{I_{m+1}}(z_{m+1})(f - \mu) \in \cup_\epsilon \overline{\mathcal{V}}_\epsilon\}$. As a consequence, $\underline{F}(f|z_{m+1}) \geq \underline{E}(f|z_{m+1})$.*

Since the sets $\mathcal{V}_\epsilon$ are not necessarily closed, we may wonder if the functional $\underline{F}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$ defined as a limit of regular extensions is actually more precise that the natural extension $\underline{E}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$. In our next result, we show that this is not the case. The proof is based on using Proposition 6 to obtain the result for linear previsions, and then apply envelope results. It is a generalisation of a result established in [17] for conditional lower probabilities:

Theorem 3. *Assume that $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ are weakly coherent and avoid partial loss. Then $\underline{E}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}}) = \underline{F}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$.*

Of course, the result is valid in particular if $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ are coherent. We can also determine, as a corollary, that the conditional lower prevision derived from an unconditional by natural extension is also the limit of conditional lower previsions obtained by regular extension. Note that in this particular case $\mathcal{M}(\epsilon), \mathcal{M}(0)$ would be

$$\mathcal{M}(\epsilon) = \{P : P(f) \geq \underline{P}(f) - \epsilon R(f) \ \forall f \in \mathcal{H}\}, \quad (10)$$

and $\mathcal{M}(0) = \mathcal{M}(\underline{P})$. Another interesting point is that in this particular case where we have a conditional and an unconditional lower prevision only, weak and strong coherence are equivalent:

Corollary 2. *Let $\underline{P}$ be a coherent lower prevision with domain $\mathcal{H}$, and consider disjoint O, I. For every $\epsilon > 0$, let $\underline{R}^\epsilon(X_O|X_I)$ be the conditional lower prevision defined from $\mathcal{M}(\epsilon)$ using regular extension, where $\mathcal{M}(\epsilon)$ is given by Eq.* (10)*. Then $\lim_{\epsilon\to 0}\underline{R}^\epsilon(X_O|X_I)$ coincides with the conditional natural extension $\underline{E}(X_O|X_I)$.*

At this point we may still be wondering if going through the sets $\mathcal{M}(\epsilon)$ is really necessary, or if we could have applied regular extension on the credal set $\mathcal{M}(0)$ given by Eq. (7) and use it to approximate $\underline{E}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$. This is not possible in general, because Proposition 8 does not necessarily hold for $\epsilon = 0$, i.e., there may not be any $P \in \mathcal{M}(0)$ such that $P(z_{m+1}) > 0$, and therefore we may not be able to use the regular extension in that case; this is easy to see with precise assessments. Moreover, even if we can apply regular extension to $\mathcal{M}(0)$, we do not necessarily have the equality $\underline{E}_{m+1}(f|z_{m+1}) = \inf\{P(f|z_{m+1}) : P \in \mathcal{M}(0), P(z_{m+1}) > 0\}$. This is discussed for the particular case of lower probabilities in [17, Sects. 3.7,3.8], and some illustrative examples are provided.

Hence, the inequality given in Proposition 7 is not necessarily an equality. In the following result, we show that a sufficient condition for the equality to hold is that the lower probability of the conditioning event is positive; see also [14, Thm. 8.1.4]:

Proposition 9. *Consider weakly coherent $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ that avoid partial loss. Let $\underline{P}$ be their unconditional natural extension, given by Eq.* (3)*, and let $\underline{P}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}})$ be given by Eq.* (4)*. If $\underline{P}(z_{m+1}) > 0$, then for all $f \in \mathcal{K}^{m+1}$, $\underline{E}_{m+1}(f|z_{m+1}) = \underline{P}_{m+1}(f|z_{m+1}) = \sup\{\mu : \pi^{-1}_{I_{m+1}}(z_{m+1})(f-\mu) \in \mathcal{V}\}$.*

Hence, we also show that in this case the natural extension is also the smallest conditional lower prevision that is weakly coherent with the initial assessments. In particular, if $\underline{P}(z_{m+1}) > 0$ for all $z_{m+1} \in \mathcal{X}_{I_{m+1}}$, we should deduce that

$$\underline{E}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}}) = \underline{P}_{m+1}(X_{O_{m+1}}|X_{I_{m+1}}).$$

The intuition here is that in that case $\underline{R}^\epsilon(z_{m+1}) > 0$ for all $z_{m+1} \in \mathcal{X}_{I_{m+1}}$ and for ϵ small enough, and then the regular extension from $\mathcal{M}(\epsilon)$ coincides with the natural extension. From here it suffices then to apply a limit result.

Finally, we are going to show that our results allow to derive a characterisation of the notion of coherence for conditional lower previsions on finite spaces.

Lemma 4. *Consider a sequence of conditional lower previsions $\{\underline{P}^k_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}^k_m(X_{O_m}|X_{I_m})\}_{k\in\mathbb{N}}$ with respective domains $\mathcal{H}^1, \ldots, \mathcal{H}^m$. Assume their point-wise limits $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ exist. If $\underline{P}^k_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}^k_m(X_{O_m}|X_{I_m})$ are weakly coherent (resp., coherent) for all k, then so are $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$.*

Using this lemma, we can derive the following:

Theorem 4. *Let $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ be separately coherent conditional lower previsions. They are coherent if and only if they are the point-wise limits of a sequence of coherent conditional lower previsions defined by regular extension.*

Hence, in the case of finite spaces the notion of coherence, which, as we have argued, is the central (and in a way the unique) consistency notion in Walley's theory, is equivalent to the approximation by means of regular extensions.

6 Conclusions

In this paper we have focused on providing a dual view of Walley's strong coherence and natural extension in the case of finite spaces. Our main result shows that there is an equivalent model made of a sequence of unconditional credal sets. By this sequence we can recreate the original conditional lower previsions using Bayes' rule; moreover, we can use this rule to compute any natural extension. This shows, in a sense, that the essence of coherence within finite spaces is just Bayes' rule. But it also suggests that the basic modelling unit in a traditional theory of (coherent) probability, even a precise one, should be a sequence of unconditional credal sets rather than a single unconditional model. This might give a new perspective on probabilistic modelling; and it might make coherence and natural extension accessible and usable concepts without notions of coherent lower previsions.

In developing the main results we have given a number of new results more strictly related to coherent lower previsions. We have given new characterisations of the notions of avoiding partial and uniform sure loss. We have shown that there is an extension of weakly coherent lower previsions that we could call *weak natural extension* and that it can be characterised through conditioning the smallest unconditional lower prevision that is weakly coherent with the former ones. Finally, we have discussed some key differences between the weak natural extension and the natural extension. All of this seems to be interesting in its own as it shows, for example, that what some applications of credal sets do is to make weakly coherent inferences rather than computing natural extensions, and therefore points to possible improvements of those approaches.

With respect to future work, we should like to point out three avenues: one is the obvious possibility to try to extend the results presented here to the case of infinite spaces. We envisage that most of them will not be immediately extendable because in our proofs we have used a number of separation theorems and envelope results that do not apply directly to the infinite case. Another aspect worth investigating is whether the equivalence mentioned initially between conditional lower previsions and the sequence remains valid also when structural judgments are introduced in a model. Finally, the idea of using a certain sequence to check coherence and compute extensions is present also in other works [1, 13] which have a common root in the work of Krauss [7]. The relationship between the sequences used here and those used in the mentioned works should also be investigated.

Acknowledgements

Work partially supported by the projects TIN2008-06796-C04-01, MTM2007-61193 and by the Swiss NSF grants n. 200020-116674/1 and 200020-121785/1.

References

[1] G. Coletti and R. Scozzafava. *Probabilistic logic in a coherent setting*. Kluwer, 2002.

[2] L. M. de Campos, M. T. Lamata, and S. Moral. The concept of conditional fuzzy measures. *International Journal of Intelligent Systems*, 5:237–246, 1990.

[3] G. de Cooman and M. Zaffalon. Updating beliefs with incomplete observations. *Artificial Intelligence*, 159:75–125, 2004.

[4] B. de Finetti. *Theory of Probability*. John Wiley & Sons, Chichester, 1974–1975. English Translation of *Teoria delle Probabilità* (Einaudi, Turin, 1970), two volumes.

[5] R. Fagin and J. Y. Halpern. A new approach to updating beliefs. In P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, volume 6, pages 347–374. North-Holland, Amsterdam, 1991.

[6] J.-Y. Jaffray. Bayesian updating and belief functions. *IEEE Transactions on Systems, Man and Cybernetics*, 22:1144–1152, 1992.

[7] P. H. Krauss. Representation of conditional probability measures on boolean algebras. *Acta Math. Acad. Sci. Hungar*, 19:229–241, 1968.

[8] I. Levi. Potential surprise: its role in inference and decision making. In L. J. Cohen and M. Hesse, editors, *Applications of Inductive Logic*, pages 1–27. Clarendon Press, Oxford, 1980.

[9] E. Miranda. A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 48(2):628–658, 2008.

[10] E. Miranda. Updating coherent previsions on finite spaces. *Fuzzy Sets and Systems*, 160(9):1286–1307, 2009.

[11] E. Miranda and G. de Cooman. Coherence and independence in non-linear spaces. Technical report, 2005. Downloadable at `http://bellman.ciencias.uniovi.es/~emiranda/wp05-10.pdf`.

[12] E. Miranda and M. Zaffalon. Coherence graphs. *Artificial Intelligence*, 173(1):104–144, 2009.

[13] R. Pelessoni and P. Vicig. A consistency problem for imprecise conditional probability assessments. In *Proceedings of the Seventh International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 1998)*, pages 1478–1485. EDK, Paris, 1998.

[14] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[15] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996. With discussion.

[16] P. Walley, R. Pelessoni, and P. Vicig. Direct algorithms for checking coherence and making inferences for conditional probability assessments. Technical report, University of Trieste, 1999. Downloadable at `http://www2.units.it/~renatop/`.

[17] P. Walley, R. Pelessoni, and P. Vicig. Direct algorithms for checking consistecy and making inferences for conditional probability assessments. *Journal of Statistical Planning and Inference*, 126:119–151, 2004.

[18] P. M. Williams. Notes on conditional previsions. Technical report, School of Mathematical and Physical Science, University of Sussex, UK, 1975. Reprinted in *International Journal of Approximate Reasoning*, 44(3):366–383, 2007.

Duality Between Maximization of Expected Utility and Minimization of Relative Entropy When Probabilities are Imprecise

Robert F. Nau
The Fuqua School of Business
Duke University
Durham, NC, USA
robert.nau@duke.edu

Victor Richmond R. Jose
The McDonough School of Business
Georgetown University
Washington, DC, USA

Robert L. Winkler
The Fuqua School of Business
Duke University
Durham, NC, USA

Abstract

In this paper we model the problem faced by a risk-averse decision maker with a precise subjective probability distribution who bets against a risk-neutral opponent or invests in a financial market where the beliefs of the opponent or the representative agent in the market are described by a convex set of imprecise probabilities. The problem of finding the portfolio of bets or investments that maximizes the decision maker's expected utility is shown to be the dual of the problem of finding the distribution within the set that minimizes a measure of divergence, i.e., relative entropy, with respect to the decision maker's distribution. In particular, when the decision maker's utility function is drawn from the commonly used exponential/logarithmic/power family, the solutions of two generic utility maximization problems are shown to correspond exactly to the minimization of divergences drawn from two commonly-used parametric families that both generalize the Kullback-Leibler divergence. We also introduce a new parameterization of the exponential/logarithmic/power utility functions that allows the power parameter to vary continuously over all real numbers and which is a natural and convenient parameterization for modeling utility gains relative to a non-zero status quo wealth position.

Keywords. decision theory, decision analysis, relative entropy, utility theory, imprecise probabilities, portfolio optimization

1 Introduction

There are many situations in which it is of interest to measure the distance between two probability distributions – say, **p** and **q** – but the appropriate metric may depend on the field of application. In statistics the relevant metric might be the loss that results from basing an inference or decision on **q** when the true distribution is **p**. In information processing the metric might be the channel capacity that is wasted by using an encoding scheme based on **q** when **p** is the true distribution of a stream of independent signals to be transmitted. In decision analysis the metric might be the value of information that results in the updating of a prior subjective probability distribution **q** to a posterior distribution **p** prior to making a choice. In probability forecasting the metric might be a scoring rule that is used to provide an incentive for a forecaster to report **p** rather than **q** as her prediction if she believes **p** is correct. In finance the metric might be the gain in expected utility that can be achieved by an investor in a market under uncertainty when her personal distribution for future asset prices is **p** and she has the opportunity to trade with a "representative agent" whose probability distribution is **q**. If one of the distributions – say, **q** – is imprecise, then the quantity of interest to be measured may be the distance from **p** to the nearest or farthest of the possible values of **q**.

In this paper[1] we consider the problem of measuring the distance between probability distributions in the case where one is imprecise, and we focus especially on the case of expected-utility gains in a financial market, although we also discuss how all of the applications mentioned above are linked to each other by duality relationships in which an information-theoretic measure of distance – known as a *relative entropy* or *divergence* – can be identified with a loss function or a utility function in a decision or inference problem. The best-known relative entropy measure is the *Kullback-Leibler divergence*, but it has a number of generalizations. We show that two well-known parametric families of generalized divergence, namely the *power* and *pseudospherical* families, have a one-to-one correspondence with the two most commonly used parametric families of scoring rules, and they also have a one-to-one correspondence with the solutions of two canonical investment problems involving the most commonly

[1] This paper is adapted from Jose et al. 2008 with some new material. An earlier, incomplete version, Nau et al. 2007, was presented at ISIPTA '07.

used parametric family of utility functions, namely the *generalized power family* that includes the exponential and logarithmic utility functions as special cases. We also introduce a new parameterization of this family of utility functions that allows the power parameter to vary continuously over all real numbers and which is the most natural and convenient parameterization for modeling utility gains relative to a nonzero status quo wealth position. This parameterization turns out to have the property that it yields an exact agreement between the utility scale and the scales that are conventionally used for the generalized divergences.

Imprecise probabilities naturally arise in the analysis of financial markets under uncertainty wherever those markets are incomplete, which is to say, virtually everywhere. A market is incomplete if some assets have distinct bid and ask prices (or are not priced at all) because of caution or lack of information on the part of buyers and sellers and/or because of transaction costs. The simplest case of a market under uncertainty is one in which assets are purchased at time 0 and sold at time 1, and the uncertainty about asset prices at time 1 is modeled by a finite set of states. Any financial asset in such a market can be constructed from a portfolio of "Arrow securities," where an Arrow security is an asset whose payoff is \$1 in a given state and zero otherwise. The bid and ask prices for a state-i Arrow security can viewed as lower and upper probabilities assigned to state i by the representative agent. Bid and ask prices for more complex assets (which may yield arbitrary payoffs in different states) establish other linear inequality constraints on the probability distribution of the representative agent, so that in general the imprecise beliefs of the representative agent are described by a convex polytope of distributions that is the intersection of all the constraints. This set is non-empty if and only if there are no arbitrage opportunities in the market, a result that is known as the "fundamental theorem of asset pricing" but which was introduced much earlier by de Finetti as the "fundamental theorem of subjective probability." The problem we consider is that of an investor whose (precise) subjective probability distribution is $\mathbf{p}$ and who invests optimally in a market where the imprecise probabilities of the representative agent are described by a convex set Q that is disjoint from $\mathbf{p}$.

2 Generalized measures of entropy and divergence

The *entropy* of a probability distribution, as defined by Shannon (1948), is a measure of the amount of information conveyed by the observation of an event drawn from that distribution. Shannon proved that under the most efficient encoding scheme the average number of bits (binary digits) needed to report the occurrence of an event whose relative frequency is p is proportional to $\ln(1/p) = -\ln(p)$, so the expected number of bits per event to encode events drawn from a distribution $\mathbf{p}$ is proportional to $H(\mathbf{p}) \equiv -\sum_i p_i \ln(p_i)$.[2] This quantity is known as the entropy of the distribution $\mathbf{p}$, because up to a multiplicative constant (namely Boltzmann's constant) it coincides exactly with the definition of the Gibbs entropy of a physical system whose distribution of internal states is $\mathbf{p}$, which in turn is the microscopic interpretation of the macroscopic concept of entropy from classical thermodynamics. If an engineer who had optimized the encoding scheme on the assumption that the distribution was $\mathbf{q}$ subsequently learns or decides (via Bayesian updating or some other method of discovery) that it is actually some other distribution $\mathbf{p}$, then the encoding scheme based on $\mathbf{q}$ is revealed to be suboptimal, and $H(\mathbf{q})$ underestimates the average number of bits per event that are actually being transmitted. A practical measure of the amount of information gained in updating $\mathbf{q}$ to $\mathbf{p}$ is the reduction in the expected number of bits needed to encode an event by re-optimizing for the distribution now believed to be correct, which is known as the Kullback-Leibler (KL) divergence of $\mathbf{p}$ with respect to $\mathbf{q}$:

$$D_{KL}(\mathbf{p}\|\mathbf{q}) \equiv \sum_i p_i(\ln(1/q_i) - \ln(1/p_i)) = \mathbf{E}_{\mathbf{p}}[\ln(\mathbf{p}/\mathbf{q})]. \quad (1)$$

The KL divergence has several very convenient and appealing properties that are often cited as reasons for adopting it as a universal measure of information gain. First, it is naturally *additive* with respect to independent experiments. Suppose that A and B are statistically independent partitions of the state space whose prior marginal probability distributions are $\mathbf{q}_A$ and $\mathbf{q}_B$, so that their prior joint distribution is $\mathbf{q}_A \times \mathbf{q}_B$. Now suppose that independent experiments are performed, which result in the updating of $\mathbf{q}_A$ and $\mathbf{q}_B$ to $\mathbf{p}_A$ and $\mathbf{p}_B$, respectively, so that the posterior joint distribution is $\mathbf{p}_A \times \mathbf{p}_B$. Then the total information gain of the two experiments is the sum of their separate KL divergences:

$$D_{KL}(\mathbf{p}_A \times \mathbf{p}_B\|\mathbf{q}_A \times \mathbf{q}_B) = D_{KL}(\mathbf{p}_A\|\mathbf{q}_A)+D_{KL}(\mathbf{p}_B\|\mathbf{q}_B). \quad (2)$$

Second, and even stronger, the KL divergence has the property of *recursivity* with respect to the splitting of events. Suppose that information is transmitted

[2] Throughout the paper, upper-case functions such as $H(\mathbf{p})$, $D_{KL}(\mathbf{p}\|\mathbf{q})$, $S(\mathbf{r},\mathbf{p})$, etc., are scalar-valued functions of vector arguments, whereas lower-case functions such as $f(\mathbf{x})$, $\ln(\mathbf{x})$, $u(\mathbf{x})$, etc., are vector-valued functions in which a univariate function is applied elementwise to a vector argument, i.e., $f(\mathbf{x}) = (f(x_1), ..., f(x_n))$.

in a 2-step process, in which two out of n possible states - say, states 1 and 2 - are not distinguished on the first step. If the realized state is neither 1 or 2, the process stops there, but otherwise a second signal is sent to report which of those two has occurred. The probabilities of states 1 and 2 are aggregated in the first step, so the information gain on that step is $D_{KL}(p_1+p_2, p_3, ..., p_n \| q_1+q_2, q_3, ..., q_n)$. On the second step, which occurs with probability (p_1+p_2), the additional gain is $D_{KL}\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2} \| \frac{q_1}{q_1+q_2}, \frac{q_2}{q_1+q_2}\right)$. The recursivity property of the KL divergence requires the expected total information gain of the two-step process to be the same as that of a one-step process:

$$D_{KL}(\mathbf{p}\|\mathbf{q}) = D_{KL}(p_1+p_2, p_3, ..., p_n \| q_1+q_2, q_3, ..., q_n) + (p_1+p_2) D_{KL}\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2} \| \frac{q_1}{q_1+q_2}, \frac{q_2}{q_1+q_2}\right). \quad (3)$$

The KL divergence is the only distance measure that satisfies both additivity and recursivity, hence it is the divergence that is naturally obtained if those properties are embraced as axioms that an information measure should satisfy. However, in situations other than signal transmission, where the objective may be something other than economizing on bandwidth, these axioms may be unduly restrictive. In applications involving imprecise probabilities, it may be of interest to find the member of a convex set of distributions that is nearest to or farthest from some reference distribution, and desiderata of a distance measure may depend on the inference or decision problem to be solved.

One measure of distance between probability distributions that generalizes the Kullback-Leibler divergence is known as a *Brègman divergence* (Brègman 1967). Any strictly convex function F defines a Brègman divergence $B_F(\mathbf{p}\|\mathbf{r})$ as follows:

$$B_F(\mathbf{p}\|\mathbf{r}) \equiv F(\mathbf{p}) - F(\mathbf{r}) - \nabla F(\mathbf{r}) \cdot (\mathbf{p}-\mathbf{r}). \quad (4)$$

The decision-theoretic significance of a Brègman divergence is that it uniquely determines a *strictly proper scoring rule*, which is a reward function for truthfully eliciting subjective probabilities. As noted by McCarthy (1956) and further elaborated by Hendrickson and Buehler (1971) and Savage (1971), any strictly convex function F can be used to generate a strictly proper scoring rule S as follows:

$$S(\mathbf{r}, \mathbf{p}) \equiv F(\mathbf{r}) + \nabla F(\mathbf{r}) \cdot (\mathbf{p}-\mathbf{r}), \quad (5)$$

where $\nabla F(\mathbf{r})$ denotes the gradient (or more generally a subgradient) of F evaluated at $\mathbf{r}$, and conversely F can be recovered from S according to $F(\mathbf{p}) = S(\mathbf{p}, \mathbf{p})$. The function $S(\mathbf{r}, \mathbf{p})$ is used to "score" a probability forecast in the following way. A forecaster who reports $\mathbf{r}$ to be her probability distribution over the states is given a reward equal to $S(\mathbf{r}, \mathbf{e}_i)$ if state i occurs, where $\mathbf{e}_i$ denotes the probability distribution that assigns probability 1 to state i and zero to all other states, i.e., the indicator vector for state i. Because S is linear in $\mathbf{p}$, we have $S(\mathbf{r}, \mathbf{p}) = \sum_i p_i S(\mathbf{r}, \mathbf{e}_i)$, so the function $S(\mathbf{r}, \mathbf{p})$ represents the forecaster's *expected* score if her distribution is $\mathbf{p}$ and she reports distribution $\mathbf{r}$. If $F(\mathbf{p})$ is strictly convex, it follows from the subgradient inequality that $S(\mathbf{r}, \mathbf{p})$ is uniquely maximized when $\mathbf{r} = \mathbf{p}$, i.e., when the forecaster honestly reports her probability distribution, which is the defining property of a strictly proper scoring rule.

By construction, the function $F(\mathbf{p}) - S(\mathbf{r}, \mathbf{p})$, which represents the forecaster's expected *loss* for reporting $\mathbf{r}$ when her distribution is $\mathbf{p}$, is the Brègman divergence $B_F(\mathbf{p}\|\mathbf{r})$. A Brègman divergence is therefore a decision-theoretic measure of the "information deficit" that is faced by a decision maker who acts on the basis of the distribution $\mathbf{r}$ when the distribution is $\mathbf{p}$. In this capacity, Brègman divergences (and their corresponding strictly proper scoring rules) provide a potentially rich class of loss functions that can be used for robust Bayesian inference, as discussed by Grünwald and Dawid (2004), Dawid (2006), and Gneiting and Raftery (2007). A problem of this kind can be framed as a game against nature in which nature chooses a distribution $\mathbf{p}$ from some convex set $\mathcal{P}$, such as the set of distributions satisfying a mean value constraint. The robust Bayes problem for the decision maker is to determine the distribution $\mathbf{r}$ that minimizes her maximum expected loss over all $\mathbf{p} \in \mathcal{P}$, where the expected loss (in our terms) is the negative expected score $-S(\mathbf{r}, \mathbf{p})$. Grünwald and Dawid show that the optimal-expected-loss function, $-F(\mathbf{p})$, is interpretable as a generalized entropy, and minimizing the maximum expected loss is equivalent to maximizing this entropy on the set $\mathcal{P}$. The distribution $\mathbf{r}$ that solves this problem is the one that minimizes $B_F(\mathbf{p}\|\mathbf{r})$ with respect to an uninformative reference distribution $\mathbf{p}_0$ at which the entropy $-F(\mathbf{p})$ is maximized.

3 The pseudospherical and power divergences

In this paper, we will consider a different kind of game and a correspondingly different decision-theoretic measure of information, namely, we will suppose that a risk-averse decision maker with personal probability distribution $\mathbf{p}$ has the opportunity to bet against a non-strategic less-well-informed opponent whose distribution $\mathbf{q}$ is known to lie in some set Q that is disjoint from $\mathbf{p}$, which enables the decision maker to place bets that are profitable in the sense of increasing her expected utility relative to the status quo.

The "information surplus" enjoyed by this decision maker will be shown to be measured by the minimum of a generalized divergence between $\mathbf{p}$ and all $\mathbf{q} \in Q$, but it is generally not a Brègman divergence. The solution of this problem gives rise to families of "weighted" strictly proper scoring rules, in which $\mathbf{q}$ plays the role of a baseline distribution with respect to which the value of the forecaster's information is measured, and they generalize the well-known quadratic, logarithmic, and pseudospherical scoring rules – details are given in Jose et al. (2008).

There are various functional forms that could be used to define a divergence of $\mathbf{p}$ with respect to $\mathbf{q}$, and the one we that we find most compelling, for both practical and theoretical reasons, is that for a given p_i the divergence should depend on q_i only through the ratio p_i/q_i, which is the marginal value of a bet on state i: a \$1 bet yields a payoff of $\$1/q_i$ when that state occurs and zero otherwise, because this is a fair payoff from the perspective of the opponent, and its expected value for the decision maker is $\$p_i/q_i$. More generally, whenever low-probability states are explicitly distinguished in the setup of a decision model, it is usually because they have large consequences, in which case relative rather than absolute errors in probability estimation are what matter. Another rationale is illustrated by the following example: suppose that the state space consists of 4 states formed by the Cartesian product of two binary events E and F, and suppose it happens that the decision maker and her opponent both agree on the probability of F and they also agree that E and F are statistically independent. Then it seems reasonable that the marginal value of a bet on any state should depend only on the extent of disagreement about the probability of E, and this requires it to depend only on the ratio of the two agents' probabilities for that state, which divides out the common probability of F.

The measurement of distance between two probability distributions in terms of ratios has a long history in statistics and information theory, and it is the basis of another kind of generalized divergence known as an *f-divergence* (Csiszár 1967). If f is a strictly convex function, the corresponding f-divergence is defined as

$$D_f(\mathbf{p}\|\mathbf{q}) \equiv E_{\mathbf{p}}[f(\mathbf{p}/\mathbf{q})]. \tag{6}$$

Divergences of this general form have been widely used in statistics for many years as (seemingly) utility-free measures of the value of the information - e.g., Goel (1983) uses f-divergence to define a "conditional amount of sample information" for measuring prior-to-posterior information gains in Bayesian hierarchical models. More recently it has been recognized that f-divergences are interpretable as measures of expected utility gains that are available to decision makers who have opportunities to bet against less-well-informed opponents or to invest in financial markets, as will be more fully discussed in later sections of this paper.

As noted above, the KL divergence is the only distance measure that satisfies the axioms of both additivity and recursivity. However, it has been discovered that weakenings of these axioms lead to several interesting parametric families of f-divergences (or transformations thereof) which have their own merits and their own applications. Havrda and Chavrát (1967) defined a quantity that they called the *directed divergence of order β between* $\mathbf{p}$ *and* $\mathbf{q}$, and variants of this divergence, which are equivalent up to a scale factor, were discussed by Rathie and Kannappan (1972), Cressie and Read (1984), and Haussler and Opper (1997). Cressie and Read referred to this quantity as the *power divergence*, and that term will be adopted here. The power divergence (as originally introduced by Havrda and Chavrát) is defined for all $\beta \in \mathbb{R}$ by:

$$D_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{q}) \equiv \frac{E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}] - 1}{\beta(\beta-1)}, \tag{7}$$

which is an f-divergence based on the normalized power function $f_\beta(x) = (x^{\beta-1} - 1)/(\beta(\beta-1))$.[3] The cases of $\beta = -1$, 0, $\frac{1}{2}$, 1, and 2 are of special interest. At $\beta = 1$, the power divergence between $\mathbf{p}$ and $\mathbf{q}$ is equal to the KL divergence $D_{KL}(\mathbf{p}\|\mathbf{q})$, and at $\beta = 0$ it is the reverse KL divergence $D_{KL}(\mathbf{q}\|\mathbf{p})$. In fact, $D_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{q})$ is antisymmetric around $\beta = \frac{1}{2}$ in the sense that $D_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{q}) = D_{1-\beta}^{\mathbf{P}}(\mathbf{q}\|\mathbf{p})$, i.e., the reverse divergence is obtained by replacing β with $1-\beta$ for any value of β. The case $\beta = \frac{1}{2}$ has perfect symmetry, i.e., $D_{1/2}^{\mathbf{P}}(\mathbf{p}\|\mathbf{q}) = D_{1/2}^{\mathbf{P}}(\mathbf{q}\|\mathbf{p})$, and it reduces to

$$D_{1/2}^{\mathbf{P}}(\mathbf{p}\|\mathbf{q}) = 4\left(1 - \sum_{j=1}^{n} \sqrt{p_j q_j}\right), \tag{8}$$

which is proportional to the *squared Hellinger distance* between $\mathbf{p}$ and $\mathbf{q}$, as noted by Haussler and Opper (1997). The Hellinger distance $D_H(\mathbf{p}\|\mathbf{q})$ is widely used in statistics and is defined by

$$D_H(\mathbf{p}\|\mathbf{q}) \equiv \left(\sum_{j=1}^{n} \left(\sqrt{p_j} - \sqrt{q_j}\right)^2\right)^{1/2}, \tag{9}$$

whence

$$D_{1/2}^{P}(\mathbf{p}\|\mathbf{q}) = 2D_H(\mathbf{p}\|\mathbf{q})^2. \tag{10}$$

[3] $f_\beta(x)$ converges to $\ln(x)$ as $\beta \to 1$, but it goes to $\pm\infty$ as β approaches zero from above or below. Nevertheless, (7) is a continuous function of β at $\beta = 0$ by virtue of the special nature of the argument of f_β and its behavior inside the expectation: the individual terms go to $\pm\infty$, but their expectation converges. Note also that f_β is antisymmetric around $\beta = 1/2$ in the following way $f_\beta(x^\beta) = f_{1-\beta}(x^{1-\beta})$, which parallels a similar property of the divergences and utility functions discussed here.

At $\beta = 2$ the power divergence reduces to (a multiple of) another well-known divergence, the *Chi-square divergence* (Pearson 1900):

$$D_2^{\mathbf{P}}(\mathbf{p}\|\mathbf{q}) = \frac{1}{2}\left(E_{\mathbf{p}}[\mathbf{p}/\mathbf{q}] - 1\right) = \frac{1}{2}\chi^2(\mathbf{p}\|\mathbf{q}), \tag{11}$$

while at $\beta = -1$ it is the reverse Chi-square divergence $\frac{1}{2}\chi^2(\mathbf{q}\|\mathbf{p})$.

The power divergence is generally neither additive nor recursive, but it satisfies two slightly weaker properties for all values of β. First, it satisfies the following *pseudoadditivity* property with respect to independent partitions A and B:

$$D_\beta^{\mathbf{P}}(\mathbf{p}_A \times \mathbf{p}_B\|\mathbf{q}_A \times \mathbf{q}_B) = D_\beta^{\mathbf{P}}(\mathbf{p}_A\|\mathbf{q}_A) + D_\beta^{\mathbf{P}}(\mathbf{p}_B\|\mathbf{q}_B) + \beta(\beta-1)D_\beta^{\mathbf{P}}(\mathbf{p}_A\|\mathbf{q}_A)D_\beta^{\mathbf{P}}(\mathbf{p}_B\|\mathbf{q}_B). \tag{12}$$

Second, it satisfies the following *pseudorecursivity* property with respect to the splitting of events (Rathie and Kannappan 1972, Cressie and Read 1984):

$$\begin{aligned} D_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{q}) = {} & D_\beta^{\mathbf{P}}(p_1+p_2, p_3, ..., p_n\|q_1+q_2, q_3, ..., q_n) \\ & + (p_1+p_2)\left(\frac{p_1+p_2}{q_1+q_2}\right)^{\beta-1} \\ & \times D_\beta^{\mathbf{P}}\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\|\frac{q_1}{q_1+q_2}, \frac{q_2}{q_1+q_2}\right). \end{aligned} \tag{13}$$

Pseudoadditivity reduces to additivity in both of the special cases $\beta = 0$ and $\beta = 1$ (both the KL divergence and the reverse KL divergence are additive), while pseudorecursivity reduces to recursivity only in the special case $\beta = 1$. Also note that for $\beta \in (0, 1)$ the power divergence is *subadditive*, i.e., $D_\beta^{\mathbf{P}}(\mathbf{p}_A \times \mathbf{p}_B\|\mathbf{q}_A \times \mathbf{q}_B) \leq D_\beta^{\mathbf{P}}(\mathbf{p}_A\|\mathbf{q}_A) + D_\beta^{\mathbf{P}}(\mathbf{p}_B\|\mathbf{q}_B)$, while for $\beta < 0$ or $\beta > 1$ it is *superadditive*, i.e., $D_\beta^{\mathbf{P}}(\mathbf{p}_A \times \mathbf{p}_B\|\mathbf{q}_A \times \mathbf{q}_B) \geq D_\beta^{\mathbf{P}}(\mathbf{p}_A\|\mathbf{q}_A) + D_\beta^{\mathbf{P}}(\mathbf{p}_B\|\mathbf{q}_B)$.

A different form of generalized entropy was introduced by Arimoto (1971) and further elaborated by Sharma and Mittal (1975), Boekee and Van der Lubbe (1980) and Lavenda and Dunning-Davies (2003). Arimoto's generalized entropy of order β is defined for $\beta > 0$ as follows:

$$\frac{\beta}{\beta-1}\left(E_{\mathbf{p}}[\mathbf{p}^{\beta-1}]^{1/\beta} - 1\right). \tag{14}$$

(Here β corresponds to the term $1/\beta$ in Arimoto's original presentation and to the term R in Boekee and Van der Lubbe's presentation.) The factor of β in the numerator plays no essential role when β is restricted to be positive, and without it the measure is actually valid for all real β and closely related to the pseudospherical scoring rule (Jose et al. 2008). The corresponding relative entropy measure, which we will henceforth call the *pseudospherical divergence of order* β between $\mathbf{p}$ and $\mathbf{q}$, is obtained by introducing a reference distribution $\mathbf{q}$ and dividing out the unnecessary factor of β,

	$D_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{q})$	$D_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{q})$
$\beta = -1$	$\frac{1}{2}\chi^2(\mathbf{q}\|\mathbf{p})$	$\frac{1}{2}\left(1 - (\chi^2(\mathbf{q}\|\mathbf{p}) + 1)^{-1}\right)$
$\beta = 0$	$D_{KL}(\mathbf{q}\|\mathbf{p})$	$1 - \exp(-D_{KL}(\mathbf{q}\|\mathbf{p}))$
$\beta = \frac{1}{2}$	$2D_H(\mathbf{p}\|\mathbf{q})^2$ $= 2D_H(\mathbf{q}\|\mathbf{p})^2$	$2\left(1 - \left(1 - \frac{1}{2}D_H(\mathbf{p}\|\mathbf{q})^2\right)^2\right)$
$\beta = 1$	$D_{KL}(\mathbf{p}\|\mathbf{q})$	$D_{KL}(\mathbf{p}\|\mathbf{q})$
$\beta = 2$	$\frac{1}{2}\chi^2(\mathbf{p}\|\mathbf{q})$	$\sqrt{\chi^2(\mathbf{p}\|\mathbf{q}) + 1} - 1$

Table 1: Special cases of power and pseudospherical divergences

$$D_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{q}) \equiv \frac{\left(E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}]\right)^{1/\beta} - 1}{\beta - 1}. \tag{15}$$

This is a nonlinear transformation of the power divergence, hence it can also be expressed as a function of other well-known divergences for special cases of β, as summarized in Table 1, which highlights the antisymmetry of the power divergence around $\beta = \frac{1}{2}$.

Like the power divergence, the pseudospherical divergence satisfies a pseudoadditivity property:

$$D_\beta^{\mathbf{S}}(\mathbf{p}_A \times \mathbf{p}_B\|\mathbf{q}_A \times \mathbf{q}_B) = D_\beta^{\mathbf{S}}(\mathbf{p}_A\|\mathbf{q}_A) + D_\beta^{\mathbf{S}}(\mathbf{p}_B\|\,\mathbf{q}_B) + (\beta-1)D_\beta^{\mathbf{S}}(\mathbf{p}_A\|\mathbf{q}_A)D_\beta^{\mathbf{S}}(\mathbf{p}_B\|\mathbf{q}_B). \tag{16}$$

The coefficient of the cross-term in this case is $\beta - 1$, not $\beta(\beta - 1)$, and hence $D_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{q})$ is subadditive for $\beta < 1$ and superadditive for $\beta > 1$. However, the pseudospherical divergence is generally not pseudorecursive, and it is not an f-divergence, although it is monotonically related to one.

4 The family of normalized linear-risk-tolerance utility functions

In the optimization problems to be discussed in the following section of the paper, the decision maker's utility function will be assumed to be drawn from the most commonly used parametric family of utility functions, namely the generalized power family that includes the exponential and logarithmic utility functions as limiting cases. The utility functions from this family will be parameterized here as:

$$\begin{aligned} u_\beta(x) &\equiv \frac{1}{\beta-1}((1+\beta x)^{(\beta-1)/\beta} - 1) \text{ if } \beta x > -1 \\ u_\beta(x) &\equiv -\infty \text{ otherwise,} \end{aligned}$$

for all $\beta \in \mathbb{R}$. This parameterization, which was introduced by Jose et al. (2008), has two key properties. First, $u_\beta(0) = 0$ and $u'_\beta(0) = 1$, so that for every β the graph of u_β passes through the origin and has a slope of unity there. Second, the corresponding *risk tolerance function* $\tau_\beta(x)$, which is the reciprocal of the Pratt-Arrow risk aversion measure, is a linear function of wealth with slope equal to β and intercept

$\beta = -1$	quadratic utility	$u_{-1}(x) = -\frac{1}{2}((1-x)^2 - 1)$
$\beta = 0$	exponential utility	$u_0(x) = 1 - \exp(-x)$
$\beta = \frac{1}{2}$	reciprocal utility	$u_{1/2}(x) = 2\left(1 - \frac{1}{1+x/2}\right)$
$\beta = 1$	logarithmic utility	$u_1(x) = \ln(1+x)$
$\beta = 2$	square-root utility	$u_2(x) = \sqrt{1+2x} - 1$

Table 2: Examples of normalized linear-risk-tolerance utility functions

equal to 1: $\tau_\beta(x) \equiv -u'_\beta(x)/u''_\beta(x) = 1 + \beta x$.[4] Thus, risk tolerance as well as marginal utility is normalized to a value of 1 at $x = 0$. This amounts to choosing the unit of money to be the status quo risk tolerance (which is without loss of generality when there is a single risk-averse agent) and then choosing the unit of utility to be the status quo marginal utility of money (which is also without loss of generality and which yields money-utile parity at the status quo). Henceforth we will refer to u_β as a *normalized linear-risk-tolerance* (normalized LRT) utility function. The advantages of this normalization are that (a) it is a natural one for modeling utility gains and losses relative to the status quo rather than relative to some hypothetical zero-point of wealth at which utility goes to minus-infinity, and (b) for fixed $x, u_\beta(x)$ is a continuous function of β on the entire real line, so that it sweeps out the widest possible spectrum of local risk attitudes. (Utility functions with the property of linear risk tolerance but without this useful normalization are known as hyperbolic-absolute-risk-aversion (HARA) utility functions in the literature of financial economics, and they typically use different parameterizations for different ranges of the power parameter.) Some important special cases of $u_\beta(x)$ are given in Table 2.

The utility functions $\{u_\beta\}$ exhibit their own form of anti-symmetry around $\beta = \frac{1}{2}$, namely that $u_{1-\beta}(x) = -u_\beta(-x)$, or equivalently $u_\beta(-u_{1-\beta}(-x)) = x$. In other words, the graph of $u_{1-\beta}(x)$ is obtained from the graph of $u_\beta(x)$ by reflecting it around the line $y = -x$. The power (exponent) in u_β is the term $(\beta-1)/\beta$, which has the property that $((\beta-1)/\beta)^{-1} = ((1-\beta)-1)/(1-\beta)$, so that swapping β for $1-\beta$ results in another power utility function whose power is the reciprocal of the original. Thus, under this parameterization, the reciprocal utility function ($\beta = \frac{1}{2}$) is its own reflection around the line $y = -x$, the exponential and log utility functions ($\beta = 0$ and $\beta = 1$) are reflections of each other, as illustrated in Figure 1, and the power utility function with exponent δ is the reflection of the power utility function with exponent $1/\delta$ for any positive or negative δ other than 0 or 1.

[4] The decision maker's risk tolerance is the parameter that determines the mean-variance tradeoffs she is willing to make on the margin. To a second-order approximation, the amount that she is willing to pay for a risky asset whose payoff distribution has mean μ and variance σ^2 is equal to $\mu - \sigma^2/2\tau$, where τ is her risk tolerance. In other words, her *risk premium* for such an asset, which is the amount by which she devalues it relative to its expected value, is $\sigma^2/2\tau$. In general a decision maker's risk tolerance may be expected to change as her wealth changes, and with this utility function her risk tolerance is a linear function of wealth with slope coefficient β.

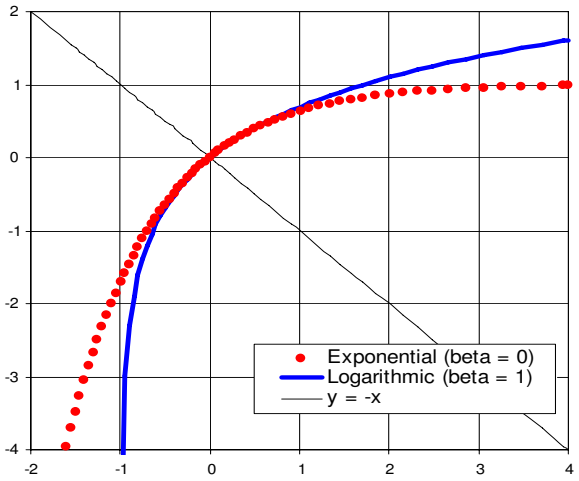


Figure 1: Reflection property of normalized LRT utility functions around $y = -x$

5 Duality between maximization of expected utility and minimization of relative entropy in incomplete markets

We now consider two generic optimization problems in which a risk averse decision maker with probability distribution $\mathbf{p}$ invests in an incomplete financial market where bid-ask spreads in asset prices are determined by a convex set Q of imprecise probabilities representing the beliefs of a risk-neutral representative agent, as noted in the introduction. The problem of expected-utility maximization in incomplete markets has been widely studied in the mathematical finance literature in recent years, and it has been shown that there is a duality relationship between maximization of expected utility and minimization of an appropriate divergence (e.g., Frittelli 2000, Rouge and El Karoui 2000, Goll and Rüschendorf 2001, Delbaen et al. 2002, Slomczyński and Zastawniak 2004, Ilhan et al. 2004, Samperi 2005). Most of this literature has focused on the case of exponential utility, for which the dual problem is the minimization of the reverse KL divergence $D_{KL}(\mathbf{q}\|\mathbf{p})$, as well as on issues that arise in multi-period or continuous-time markets. In

this section we will show that in a single-period or two-period market, there is a duality relation between the pseudospherical or power divergence and the solution of an expected-utility-maximization problem in which the utility function is drawn from the normalized linear-risk-tolerance family.

Let $\mathbf{x} \in \mathbb{R}^n$ denote the vector of monetary payoffs to the decision maker, and let $u_\beta(\mathbf{x}) \equiv (u_\beta(x_1), ..., u_\beta(x_n))$ denote the vector of utilities that the function u_β yields when applied to $\mathbf{x}$. An incomplete, single-period market can either be parameterized in terms of an $m \times n$ matrix $\mathbf{A}$ whose rows are the (net) payoff vectors of available assets, i.e., $\mathbf{A} = \{a_{ij}\}$ where a_{ij} is the net payoff to the decision maker of one unit of the i^{th}asset in state j, or else in terms of a $k \times n$ matrix $\mathbf{Q}$ whose rows are risk neutral probability distributions that support the asset prices, i.e., $\mathbf{Q} = \{q_{ij}\}$ where q_{ij} is the probability of state j under the i^{th} risk neutral distribution. The rows of $\mathbf{Q}$ are the extremal risk-neutral probability distributions assigning non-positive expectation to all the rows of $\mathbf{A}$, i.e., the rows of $-\mathbf{Q}$ are the dual cone of the rows of $\mathbf{A}$. The parameterization in terms of $\mathbf{Q}$ will be adopted here. Let $\mathbf{x}$ denote an arbitrary n-vector of monetary payoffs to the decision maker (an element of $\mathbb{R}^n$), and let $\mathbf{z}$ denote an arbitrary k-vector of non-negative weights summing to one (an element of $\triangle^k$, the unit simplex in $\mathbb{R}^k$). As before, let $\mathbf{p}$ denote the decision maker's subjective probability distribution, and henceforth let $\mathbf{q}$ denote one of many possible probability distributions attributable to a risk-neutral trading opponent: the representative agent.

In the first generic decision problem (**"S"**), there is a single time period in which consumption occurs, the decision maker has a single-attribute LRT utility function $u_\beta(x)$, and her objective is to find the payoff vector $\mathbf{x}$ that maximizes her subjective expected utility subject to the self-financing constraint $E_\mathbf{q}[\mathbf{x}] \leq 0$. The decision maker's optimal expected utility, denoted $U^\mathbf{S}(\mathbf{p}\|\mathbf{q})$, is determined by solving:

Primal Problem S:

$$U^\mathbf{S}_\beta(\mathbf{p}\|\mathbf{Q}) \equiv \max_{\mathbf{x}\in\mathbb{R}^n} E_\mathbf{p}[u_\beta(\mathbf{x})] \text{ subject to } \mathbf{Qx} \leq \mathbf{0}$$

Note that $-\mathbf{Qx}$ is the k-vector of the opponent's expected values for payoff vector $\mathbf{x}$ under all the extremal risk neutral distributions, hence the condition $\mathbf{Qx} \leq \mathbf{0}$ means that $\mathbf{x}$ yields non-negative expected value to the opponent under all those distributions.

In the second problem (**"P"**), there are two periods in which consumption occurs and the decision maker with probability distribution $\mathbf{p}$ has a quasilinear utility function $u_\beta(a, b) = a + u_\beta(b)$ where a is money consumed at time 0 and b is money consumed at time 1. Under the normalized LRT family of utility functions, the marginal rate of substitution between time-0 consumption and time-1 consumption is equal to unity at $x = 0$ in this problem, as though in the status quo the decision maker is indifferent between consuming the next dollar at time 0 or time 1. The decision maker's objective is to choose a vector $\mathbf{x}$ of time-1 payoffs to be purchased from time-0 funds at market prices so as to maximize the total expected utility of consumption in both periods. The time-0 cost of purchasing $\mathbf{x}$ is $E_\mathbf{q}[\mathbf{x}]$, so the optimal expected utility, denoted $U^\mathbf{P}(\mathbf{p}\|\mathbf{q})$, is the solution of:

Primal Problem P:

$$U^\mathbf{P}_\beta(\mathbf{p}\|\mathbf{Q}) \equiv \max_{y\in\mathbb{R},\mathbf{x}\in\mathbb{R}^n} E_\mathbf{p}[u_\beta(\mathbf{x})] - y \text{ subject to } \mathbf{Qx} \leq y\mathbf{1}$$

Henceforth, let $\mathbf{x}^\mathbf{S}_\beta(\mathbf{p}\|\mathbf{q})$ and $\mathbf{x}^\mathbf{P}_\beta(\mathbf{p}\|\mathbf{q})$ denote the solutions of Problems **S** and **P**, with i^{th} elements $x^\mathbf{S}_{\beta,i}(\mathbf{p}\|\mathbf{q})$ and $x^\mathbf{P}_{\beta,i}(\mathbf{p}\|\mathbf{q})$, respectively. Let $\mathbf{z} \in \triangle^k$ denote a vector of weights, so that $\mathbf{z}^T\mathbf{Q}$ is a mixture of the rows of $\mathbf{Q}$, which is an element of the convex polytope Q of risk neutral distributions. Our main result is that the utility gains to the decision maker under problems **S** and **P** are, respectively, the minima of the pseudospherical and power divergences between $\mathbf{p}$ and all $\mathbf{q} \in Q$ for the same β.

THEOREM (Jose et al. 2008):
(a) In an incomplete, single-period market, maximization of expected linear-risk-tolerance utility with risk tolerance coefficient β (Primal Problem **S**) is dual to minimization of the pseudospherical divergence of order β between the decision maker's subjective distribution $\mathbf{p}$ and a risk neutral distribution $\mathbf{q}$ consistent with asset prices. That is, the corresponding dual problem is:

$$\textbf{Dual Problem S: } D^\mathbf{S}_\beta(\mathbf{p}\|\mathbf{Q}) \equiv \min_{\mathbf{z}\in\triangle^k} D^\mathbf{S}_\beta(\mathbf{p}\|\mathbf{z}^T\mathbf{Q}).$$

Their optimal objective values are the same and the optimal values of the decision variables in one problem are equal to the normalized optimal values of the Lagrange multipliers in the other.

(b) In an incomplete, two-period market, maximization of expected quasilinear linear-risk-tolerance utility with second-period risk tolerance coefficient β (Primal Problem **P**) is equivalent to minimization of

the power divergence of order β between the decision maker's subjective distribution $\mathbf{p}$ and a risk neutral distribution $\mathbf{q}$ consistent with asset prices (Dual Problem $\mathbf{P}$). Their optimal objective values are the same and the optimal values of the decision variables in one problem are equal to the normalized optimal values of the Lagrange multipliers in the other. That is, the corresponding dual problem is:

$$\textbf{Dual Problem P:} D_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{Q}) \equiv \min_{\mathbf{z}\in\triangle^k} D_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{z}^T\mathbf{Q}).$$

Proof: For part (a), Lagrangian relaxation is applicable because the primal problem has a strictly concave, continuously differentiable objective function and linear constraints. Let $\boldsymbol{\lambda}$ denote the vector of Lagrange multipliers associated with the constraints $\mathbf{Qx} \le \mathbf{0}$. The Lagrangian relaxation of Primal Problem $\mathbf{S}$ is then $\min_{\boldsymbol{\lambda}\in\mathbb{R}^{k+}} L(\boldsymbol{\lambda})$ where

$$L(\boldsymbol{\lambda}) = \max_{\mathbf{x}\in\mathbb{R}^n} E_{\mathbf{p}}[u_\beta(\mathbf{x})] - \boldsymbol{\lambda}^T\mathbf{Qx}. \tag{17}$$

The Lagrangian $L(\boldsymbol{\lambda})$ is an unconstrained maximum of a continuously differentiable concave function, so it can be solved for $\mathbf{x}$ in terms of $\boldsymbol{\lambda}$ by setting $\nabla(E_{\mathbf{p}}[u_\beta(x)] - \boldsymbol{\lambda}^T\mathbf{Q}x) = \mathbf{0}$, which yields

$$\mathbf{x} = \frac{1}{\beta}\left(\left(\frac{\mathbf{p}}{\boldsymbol{\lambda}^T\mathbf{Q}}\right)^\beta - 1\right), \tag{18}$$

whence

$$\begin{aligned} L(\boldsymbol{\lambda}) &= E_{\mathbf{p}}\left[\frac{1}{\beta-1}\left[\left(\frac{\mathbf{p}}{\boldsymbol{\lambda}^T\mathbf{Q}}\right)^{\beta-1} - 1\right]\right] && (19)\\ &\quad -\boldsymbol{\lambda}^T\mathbf{Q}\left[\frac{1}{\beta}\left[\left(\frac{\mathbf{p}}{\boldsymbol{\lambda}^T\mathbf{Q}}\right)^{\beta} - 1\right]\right] \\ &= \frac{1}{\beta-1}\left[E_{\mathbf{p}}\left[\left(\frac{\mathbf{p}}{\boldsymbol{\lambda}^T\mathbf{Q}}\right)^{\beta-1}\right] - 1\right] && (20)\\ &\quad -\frac{1}{\beta}\left[E_{\mathbf{p}}\left[\left(\frac{\mathbf{p}}{\boldsymbol{\lambda}^T\mathbf{Q}}\right)^{\beta-1}\right] - \mathbf{1}^T(\boldsymbol{\lambda}^T\mathbf{Q})\right]. && (21)\end{aligned}$$

In the optimal solution $\boldsymbol{\lambda}^*$, where the constraints are satisfied, the second term will be zero, which implies

$$\mathbf{1}^T(\boldsymbol{\lambda}^{*T}\mathbf{Q}) = E_{\mathbf{p}}\left[\left(\frac{\mathbf{p}}{\boldsymbol{\lambda}^{*T}\mathbf{Q}}\right)^{\beta-1}\right] \tag{22}$$

and consequently

$$L(\boldsymbol{\lambda}^*) = \frac{1}{\beta-1}\left(E_{\mathbf{p}}\left[\left(\frac{\mathbf{p}}{\boldsymbol{\lambda}^{*T}\mathbf{Q}}\right)^{\beta-1}\right] - 1\right). \tag{23}$$

Now let $\mathbf{z}^* = \boldsymbol{\lambda}^*/\mathbf{1}^T\boldsymbol{\lambda}^*$ be the probability distribution that is obtained by normalization of the optimal Lagrange multipliers $\boldsymbol{\lambda}^*$. Then it follows from (21) that:

$$\mathbf{z}^{*T}\mathbf{Q} = \frac{\boldsymbol{\lambda}^{*T}\mathbf{Q}}{E_{\mathbf{p}}[(\mathbf{p}/\boldsymbol{\lambda}^{*T}\mathbf{Q})^{\beta-1}]}. \tag{24}$$

The pseudospherical divergence between $\mathbf{p}$ and $\mathbf{z}^{*T}\mathbf{Q}$ can therefore be expressed in terms of $\boldsymbol{\lambda}^*$ as:

$$\begin{aligned} &D_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{z}^{*T}\mathbf{Q}) \\ &= \frac{(E_{\mathbf{p}}[(\mathbf{p}/\mathbf{z}^{*T}\mathbf{Q})^{\beta-1}])^{1/\beta} - 1}{\beta-1} \\ &= \frac{(E_{\mathbf{p}}[(E_{\mathbf{p}}[(\mathbf{p}/\boldsymbol{\lambda}^{*T}\mathbf{Q})^{\beta-1}](\mathbf{p}/\boldsymbol{\lambda}^{*T}\mathbf{Q}))^{\beta-1}])^{1/\beta} - 1}{\beta-1} \\ &= \frac{(E_{\mathbf{p}}[(\mathbf{p}/\boldsymbol{\lambda}^{*T}\mathbf{Q})^{\beta-1}])^{1-1/\beta}(E_{\mathbf{p}}[(\mathbf{p}/\boldsymbol{\lambda}^{*T}\mathbf{Q})^{\beta-1}])^{1/\beta} - 1}{\beta-1} \\ &= \frac{1}{\beta-1}\left(E_{\mathbf{p}}\left[\left(\frac{\mathbf{p}}{\boldsymbol{\lambda}^{*T}\mathbf{Q}}\right)^{\beta-1}\right] - 1\right) \\ &= L(\boldsymbol{\lambda}^*), \end{aligned} \tag{25}$$

which is the optimal objective value of the primal problem. Furthermore $\mathbf{z}^* = \boldsymbol{\lambda}^*/\mathbf{1}^T\boldsymbol{\lambda}^*$ must also minimize $D_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{z}^T\mathbf{Q})$ over all $\mathbf{z} \in \triangle^k$, because if there were some other $\mathbf{z}^{**} \in \triangle^k$ such that $D_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{z}^{**T}\mathbf{Q}) < D_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{z}^{*T}\mathbf{Q})$, then it would be possible to find some $\boldsymbol{\lambda}^{**} \in \mathbb{R}^{k+}$ proportional to $\mathbf{z}^{**}$ such that $\mathbf{z}^{**T}\mathbf{Q} = \boldsymbol{\lambda}^{**T}\mathbf{Q}/(\mathbf{E}_{\mathbf{p}}[(\mathbf{p}/(\boldsymbol{\lambda}^{**T}\mathbf{Q}))^{\boldsymbol{\beta-1}}])$. By construction this $\boldsymbol{\lambda}^{**}$ would satisfy $E_{\mathbf{p}}[(\mathbf{p}/\boldsymbol{\lambda}^{**T}\mathbf{Q})^{\beta-1}] - \mathbf{1}^T(\boldsymbol{\lambda}^{**T}\mathbf{Q}) = 0$, implying $L(\boldsymbol{\lambda}^{**}) = D_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{z}^{**T}\mathbf{Q})$, and it would follow that $L(\boldsymbol{\lambda}^{**}) < L(\boldsymbol{\lambda}^*)$, contradicting the assumption that $\boldsymbol{\lambda}^*$ was optimal.

For part (b), the problem of finding the feasible risk neutral distribution that minimizes the power divergence of order β:

$$\min_{\mathbf{z}\in\triangle^k} D_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{z}^T\mathbf{Q}), \tag{26}$$

is equivalent to the Lagrangian problem $\min_{\boldsymbol{\lambda}\in\triangle^k} L(\boldsymbol{\lambda})$, where $L(\boldsymbol{\lambda}) = \max_{\mathbf{x}\in\mathbb{R}^n} E_{\mathbf{p}}[u_\beta(x)] - \boldsymbol{\lambda}^T\mathbf{Q}x$ is the same Lagrangian that was used in the proof of part (a) to minimize the pseudospherical divergence, except that here $\boldsymbol{\lambda}$ is constrained to be in the simplex, not just the non-negative orthant ($\boldsymbol{\lambda} \in \triangle^k$ rather than $\boldsymbol{\lambda} \in \mathbb{R}^{k+}$), which requires a Lagrange multiplier for the constraint $\mathbf{1}^T\mathbf{q} = 1$ in addition to the m Lagrange multipliers for the constraints $\mathbf{Aq} \ge \mathbf{0}$. The latter divided by the former are equal to the optimal values of the decision variables in Primal Problem $\mathbf{P}$ multiplied by $-\beta$. The power divergence is minimized by the same risk neutral distribution $\mathbf{q}^* = \mathbf{z}^{*T}\mathbf{Q}$ that minimizes the pseudospherical divergence (for the same $\mathbf{p}$, β and $\mathbf{Q}$), because they are both monotonic functions of $E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}]$. The optimal value of $\boldsymbol{\lambda}$ is a unit vector selecting the largest element of $\mathbf{Qx}$. Let z denote this largest element. Then $\min_{\boldsymbol{\lambda}\in\triangle^k} \max_{\mathbf{x}\in\mathbb{R}^n} E_{\mathbf{p}}[u_\beta(x)] - \boldsymbol{\lambda}^T\mathbf{Q}x$ is equivalent to $\max_{\mathbf{x}\in\mathbb{R}^n} E_{\mathbf{p}}[u_\beta(\mathbf{x})] - z$ subject to $\mathbf{Qx} \le z\mathbf{1}$.■

The special case $\beta = 1$ corresponds to log utility in the primal problem and KL divergence in the dual problem, while $\beta = 0$ corresponds to exponential utility in

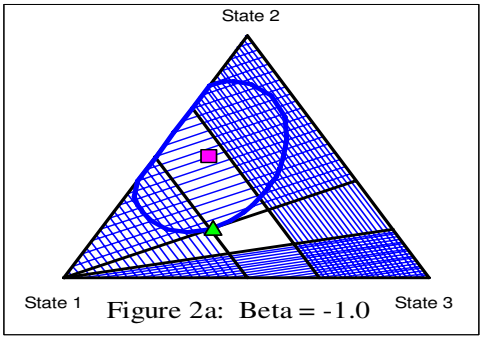


Figure 2a: Beta = -1.0

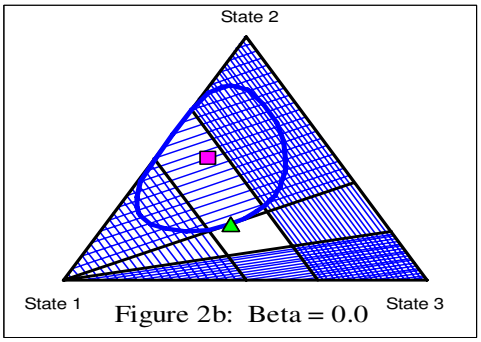


Figure 2b: Beta = 0.0

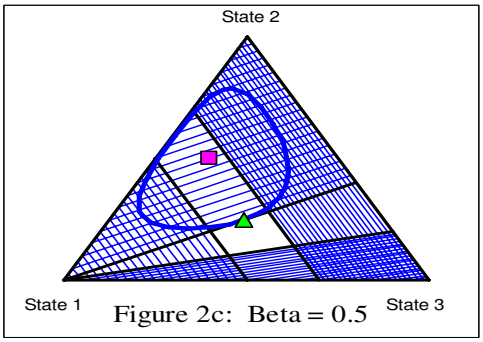


Figure 2c: Beta = 0.5

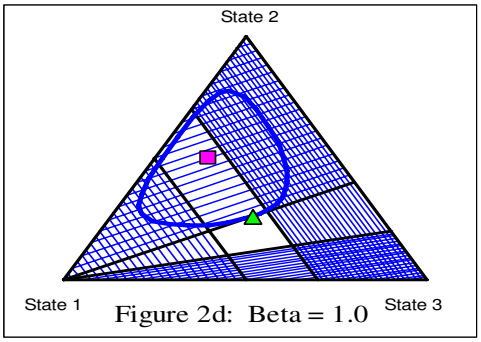


Figure 2d: Beta = 1.0

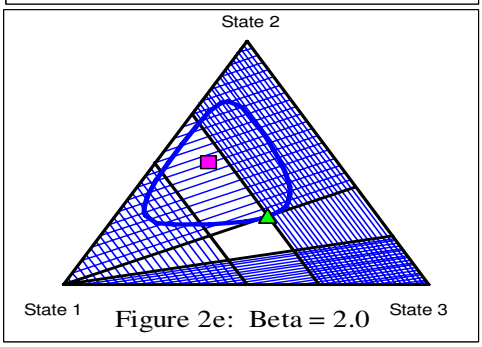


Figure 2e: Beta = 2.0

Figures 2a-e: Minimum-divergence solution for the power divergence with $\mathbf{p} = (0.35, 0.5, 0.15)$ for $\beta = -1$, 0, 0.5, 1, and 2.

the primal problem and reverse KL divergence in the dual problem, and the cases $\beta = 1/2$ and $\beta = 2$ are related to the squared Hellinger distance and the Chi-square divergence as shown in the right-hand column of Table 1. Because the pseudospherical divergence is a monotonic transformation of the power divergence, the distribution $\mathbf{q}$ ($= \mathbf{z}^T\mathbf{Q}$) that solves Dual Problem **S** is the same one that solves Dual Problem **P**, although the objective values and the primal payoff vectors are generally different. The power divergence is always strictly greater than the pseudospherical divergence ($D_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{q}) > D_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{q})$) except at $\beta = 1$, as pointed out earlier, but this inequality is further illuminated by a comparison of the corresponding Lagrangian relaxation problems: the minimization of $L(\boldsymbol{\lambda})$ over $\boldsymbol{\lambda} \in \triangle^k$ must yield a result greater than or equal to its minimization over the larger set $\boldsymbol{\lambda} \in \mathbb{R}^{k+}$, whether or not the market is complete.

Versions of the same duality theorem have been discussed in the mathematical finance literature, as noted above, although the full spectrum of LRT utility and its closed-form solution have not previously been characterized. The details of the correspondence between our results and those of Goll and Rüschendorf (2001) are given in Jose et al. (2008).

6 Illustration of the geometry of the divergence-minimization problem

To visualize the preceding results, consider a simple example in which there are three states and (only) lower and upper bounds of 0.3 and 0.5 are given for the probability of state 1 and lower and upper bounds of 0.6 and 0.8 are given for the conditional probability of state 3 given not-state-1. The set Q of probability distributions that satisfies these constraints is the unshaded quadrilateral in the lower center of the simplex in Figures 2a-e. Let the reference distribution be $\mathbf{p} = (0.35, 0.5, 0.15)$, which is the square dot in the upper left. Figures 2a-e show the solution of the dual problem of finding the element of Q that minimizes the pseudospherical or power divergence between itself and $\mathbf{p}$ for $\beta = -1$, 0, 0.5, 1, and 2. The triangular dot is the minimum-divergence solution, and the contour (level curve) that passes through it is also shown. In this case, the solution moves from the left to the right of the upper edge of the quadrilateral as β increases from -1 to 2. Also, the contours become more triangular in shape as β increases, flattening more near the edges of the simplex, because as $\mathbf{q}$ approaches an edge of the simplex, q_i goes to zero for some i, and the term $(p_i/q_i)^{\beta-1}$ in the divergence calculation blows up faster for larger values of β as that edge is approached.

7 Discussion

A financial market under uncertainty provides one of the purest and most economically important examples of a situation in which subjective beliefs – in this case those of a risk neutral representative agent with whom individual investors may trade – are represented by imprecise probabilities that are subject to direct measurement. The measurement process, which consists of setting bid and ask prices for portfolios of Arrow securities, is essentially the same operational method of eliciting subjective probabilities that was introduced by de Finetti, and it naturally leads to a representation of beliefs in the form of a convex polytope of probability distributions. In this paper we have considered the decision problem faced by a risk-averse investor in such a market when her risk preferences are represented by a utility function drawn from the generalized power family, which is the family most commonly used in finance theory and applied decision analysis. Under a natural (but novel) parameterization of the generalized power utility function, the investor's optimal expected utility is equal to the minimum of a generalized divergence between her own distribution and the nearest element of the polytope that characterizes the imprecise beliefs of the representative agent, where the generalized divergence is drawn from a parametric family that generalizes the Kullback-Lielber divergence. We have also pointed out connections with recent developments in the use of generalized divergences in robust Bayesian statistics. These results highlight the interconnections among information theory, Bayesian statistics, decision analysis, and finance theory with respect to the program of modeling imprecise probabilities.

References

[1] Arimoto, S. 1971. Information-theoretical considerations on estimation problems. *Infor. Contr.* **19**:181-194.

[2] Boekee, D. E., J. C. A. van der Lubbe. 1980. The R-norm information measure. *Infor. Contr.* **45**:136-155.

[3] Brègman, L. 1967. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. and Math. Phys.*, **7**:200-217.

[4] Cressie, N., T. R. C. Read. 1984. Multinomial goodness of fit. *J. Royal Stat. Soc. B* **46**(3):440-464.

[5] Csiszár, I. 1967 Information type measures of differences of probability distribution and indirect observations. *Studia Math. Hungarica* **2**:299-318

[6] Dawid, A.P. 2006. The geometry of proper scoring rules. Research Report No. 268, Department of Statistical Science, University College, London. (April 2006)

[7] Delbaen, F., P. Grandits, T. Rheinländer, D. Samperi, M. Schweizer, C. Stricker. 2002. Exponential hedging and entropy penalties. *Math. Finance.* **12:**99-123.

[8] Frittelli, M. 2000. The minimal entropy martingale measure and the valuation problem in incomplete markets. *Math. Finance.* **10:**39-52.

[9] Gneiting, T., A. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**:359-378

[10] Goel, P. 1983. Information measures and Bayesian hierarchical models. *J. Am. Stat. Assoc.* **78**:408-410.

[11] Goll, T., L. Rüschendorf. 2001. Minimax and minimal distance martingale measures and their relationship to portfolio optimization. *Finance Stoch.* **5**:557-581

[12] Grünwald, P.D., A.P. Dawid. 2004. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Ann. Statist.* **32**: 1367-1433.

[13] Havrda, J., F. Chavrát. 1967. Quantification method of classification processes: the concept of structural α-entropy. *Kybernetika* **3**:30-35.

[14] Haussler, D., M. Opper. 1997. Mutual information, metric entropy, and cumulative relative entropy risk. *Ann. Statist.* **25**: 2451-2492.

[15] Hendrickson, A. D., R. J. Buehler 1971. Proper scores for probability forecasters. *Ann. Math. Stat.* **42**:1916-1921.

[16] Ilhan, A., M. Jonsson, R. Sircar. 2004. Portfolio optimization with derivatives and indifference pricing. Working Paper, Princeton University.

[17] Jose, V.R.R., R. F. Nau, R. L. Winkler. 2008. Scoring rules, generalized entropy, and utility maximization. *Oper. Res.* **56**:1146-1147

[18] Lavenda, B. H., J. Dunning-Davies. 2003. Qualms concerning Tsallis's condition of pseudo-additivity as a definition of non-extensivity. http://arxiv.org/abs/cond-mat/0312132

[19] McCarthy, J. 1956. Measures of the value of information. *Proc. Nat. Acad. Sci. USA* **42**:654-655.

[20] Nau, R. F., V. R. R. Jose, R. L. Winkler. 2007. Scoring Rules, entropy and imprecise probabilities. *Proceedings of the 5th ISIPTA Conference*, 307-315.

[21] Pearson, K. 1900. On a criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *London Edinburgh Dublin Philos. Mag. J. Sci. Series 5* **50**:157-175.

[22] Rathie, P. N., P. Kannappan. 1972. A directed-divergence function of type β. *Infor. Contr.* **20**:38-45.

[23] Rouge, R., N. El Karoui. 2000. Pricing via utility maximization and entropy. *Math. Finance.* **10:**259-276.

[24] Samperi, D. 2005. Model selection using entropy and geometry: complements to the six-author paper. Working paper, Decision Synergy.

[25] Savage, L. J. 1971. Elicitation of personal probabilities and expectations. *J. Am. Stat. Assoc.* **66**:783-801.

[26] Selten, R. 1998. Axiomatic characterization of the quadratic scoring rule. *Exper. Econ.* **1**:43-62.

[27] Shannon, C. E. 1948. A mathematical theory of communication. *Bell Sys. Tech. J.* **27**:379-423.

[28] Sharma, B. D., D. P. Mittal. 1975. New non-additive measures of entropy for discrete probability distributions *J. Math. Sci.* **10**:28-40.

[29] Slomczyński, W., T Zastawniak. 2004. Utility maximizing entropy and the second law of thermodynamics. *Ann. Prob.* **32**:2261-2285.

The Pari-Mutuel Model

Renato Pelessoni
University of Trieste, Italy
renato.pelessoni@econ.units.it

Paolo Vicig
University of Trieste, Italy
paolo.vicig@econ.units.it

Marco Zaffalon
IDSIA, Switzerland
zaffalon@idsia.ch

Abstract

We explore generalizations of the pari-mutuel model (PMM), a formalization of an intuitive way of assessing an upper probability from a precise one. We discuss a naive extension of the PMM considered in insurance and generalize the natural extension of the PMM introduced by P. Walley and other related formulae. The results are subsequently given a risk measurement interpretation: in particular it is shown that a known risk measure, Tail Value at Risk (TVaR), is derived from the PMM, and a coherent risk measure more general than TVaR from its imprecise version. We analyze further the conditions for coherence of a related risk measure, Conditional Tail Expectation. Explicit formulae for conditioning the PMM and conditions for dilation or imprecision increase are also supplied and discussed.

Keywords. Pari-mutuel model, risk measures, natural extension, dilation, 2-monotonicity.

1 Introduction

The *pari-mutuel model* (PMM) formalizes a very intuitive and therefore widely used method of assigning an upper probability starting from a precise probability. To introduce it, consider, following [2], a probability P for event A as a *fair* price for a bet which returns 1 unit to the bettor if A is true, 0 units if A is false, i.e. returns the indicator I_A of A. The bettor's gain is $G = I_A - P(A)$, while that of his opponent, House, is $-G = G_H = P(A) - I_A$.

In most real-world betting schemes House is unwilling to accept such a fair game (the expectation $E(G_H)$ is 0), but asks for a positive gain expectation. It is so when House is a bookmaker, an insurance company, the organizer of a lottery, and so on. A way to achieve this goal is to raise the bettor's price, without altering his reward, and a *naive method* multiplies P by a constant greater than 1, say $1+\delta$, where $\delta > 0$ is a loading constant. The bettor pays $\overline{P}(A) = (1+\delta)P(A)$, while the gain for House is now $\overline{G}_H = (1+\delta)P(A) - I_A$. Alternatively, House might ask the same price to pay a reduced reward $(1-\tau)I_A$, where $0 < \tau < 1$ is interpreted as a commission, or also a taxation. This originates a gain $\overline{G}^*_H = P(A) - (1-\tau)I_A = (1-\tau)(\frac{P(A)}{1-\tau} - I_A) = (1-\tau)\overline{G}_H$ iff $\frac{1}{1-\tau} = 1+\delta$, i.e. iff $\tau = \frac{\delta}{1+\delta}$. Thus, up to a scaling factor, the two methods are equivalent if $\tau = \frac{\delta}{1+\delta}$; the latter is formally more adherent to common betting systems, called in fact *pari-mutuel systems*.

In the theory of imprecise probabilities, $\overline{P}$ is an upper probability, but a slight adjustment to $\overline{P}$ is necessary to achieve coherence. In fact, Walley [11] terms pari-mutuel model the upper probability

$$\overline{P}(A) = \min\{(1+\delta)P(A), 1\}. \tag{1}$$

Intuitively, the correction should be needed: when $P(A) > \frac{1}{1+\delta}$, it is $\overline{G}_H > 0$ in the naive method, i.e. a bettor suffers from a sure loss no matter whether A is true or false.

This paper investigates further the pari-mutuel model, extending the analysis in [11]. Preliminary issues are recalled in Section 2, very concisely in general, more extensively as for 2-monotone and 2-alternating previsions, since the upper probability $\overline{P}$ in (1) is 2-alternating. In Section 3 we discuss extensions of the PMM. First, we consider alternative expressions for the natural extension $\overline{E}(X)$ of $\overline{P}$, defined on a field $\mathcal{A}$, to any $\mathcal{A}$-measurable gamble X. These expressions were stated in [11], but we make a more detailed analysis of the conditions ensuring that $\overline{E}(X)$ is equal to a certain conditional prevision $(P(X|X > x_\tau))$, which has a risk measurement interpretation. In Section 3.1 we restrict to non-negative gambles and compare the natural extension $\overline{E}$ with the naive extension $\overline{P}_N(X) = \min\{(1+\delta)P(X), \sup X\}$, showing that quite often $\overline{P}_N$ is not coherent, or it sometimes coincides with $\overline{E}$. The motivation for this work is that $\overline{P}_N$ is a premium in insurance, although with different

premises: the starting point is not the PMM but a set of non-negative gambles. In Section 3.2 we generalize Walley's approach, obtaining a formula for $\overline{E}(X)$ when the PMM is given on a lattice of events and X is not necessarily measurable.

These results have an interesting and, to the best of our knowledge, so far not considered interpretation in the realm of risk measurement. This is the main topic of Section 4, where the natural extension of the PMM defined on a field is shown to correspond to a coherent risk measure, called Tail Value-at-Risk or TVaR (in [4]; other authors may use a different terminology). When the PMM is defined on a lattice, we obtain a generalization of TVaR (not discussed in the risk literature), which replaces precise with imprecise uncertainty measures; we name it ITVaR. Thus the PMM supplies a motivation for introducing 'imprecise' risk measures: one of them, ITVaR, is the natural extension of a PMM assigned on a lattice. Conditioning the PMM defined on a field is discussed in Section 5. We specialize general formulae for the natural extension of 2-alternating and 2-monotone probabilities to the case of the PMM and discuss the effect on them of *dilation* and of a weaker phenomenon, *imprecision increase*. We obtain a number of conditions for dilation or imprecision increase, and discuss in detail the operationally most relevant cases (when the commission τ is not "too high" and event A is either "common" or "rare"). Section 6 concludes the paper.

2 Preliminaries

Upper ($\overline{P}$) and lower ($\underline{P}$) probabilities are customarily related by the conjugacy relation $\overline{P}(A) = 1 - \underline{P}(A^c)$, which lets one refer to either $\overline{P}$ or $\underline{P}$ only. Applying it to (1), the lower probability in the PMM is [11]

$$\underline{P}(A) = \max\{(1+\delta)P(A) - \delta, 0\}. \tag{2}$$

As noted in the Introduction, the parameter $\tau \in]0;1[$ can, and later will, alternatively describe $\overline{P}$, $\underline{P}$ in the PMM. The relationship between τ and δ is:

$$\tau = \frac{\delta}{1+\delta} \quad ; \quad \delta = \frac{\tau}{1-\tau}. \tag{3}$$

An upper probability $\overline{P}$ defined by (1) for any A in an *arbitrary* set of events $\mathcal{D}$ (or $\underline{P}$ defined by (2)) is *coherent* on $\mathcal{D}$, and probably the simplest way to see it is to apply the later Proposition 2. In general, an *upper prevision* $\overline{P}$ is a mapping from a set $\mathcal{D}$ of *gambles* (bounded random variables) into the real line, and an *upper probability* is its special case that the domain $\mathcal{D}$ is made of (indicators of) events only. The upper prevision $\overline{P}$ is *coherent* on $\mathcal{D}$ iff, $\forall n \in \mathbb{N}$, $\forall s_0, s_1, \ldots, s_n \geq 0$, $\forall X_0, X_1, \ldots, X_n \in \mathcal{D}$, defining $\overline{G} = \sum_{i=1}^{n} s_i(\overline{P}(X_i) - X_i) - s_0(\overline{P}(X_0) - X_0)$, it holds that $\sup \overline{G} \geq 0$.

There are several necessary conditions for coherence, in particular: *internality*, $\inf X \leq \overline{P}(X) \leq \sup X$, and *subadditivity*, $\overline{P}(X+Y) \leq \overline{P}(X) + \overline{P}(Y)$.

We refer to [11] for a thorough presentation of the theory of coherent upper/lower previsions. One of its most important notions is that of *natural extension* [11, Section 3].

In our framework, the natural extension $\overline{E}$ on $\mathcal{D}'$ of a coherent upper prevision (or probability) $\overline{P}$ defined on $\mathcal{D} \subset \mathcal{D}'$ is the *least-committal* coherent extension of $\overline{P}$ on $\mathcal{D}'$, i.e. $\overline{E}(X) = \overline{P}(X)$, $\forall X \in \mathcal{D}$, and for any coherent $\overline{P}^*$ such that $\overline{P}^* = \overline{P}$ on $\mathcal{D}$, $\overline{E}(X) \geq \overline{P}^*(X)$, $\forall X \in \mathcal{D}'$, i.e. $\overline{E}$ *dominates* $\overline{P}^*$. It can be shown that $\overline{E}$ always exists. Symmetrically, the natural extension $\underline{E}$ on $\mathcal{D}'_L$ of a coherent lower prevision $\underline{P}$ on $\mathcal{D}_L$ is such that $\underline{E} = \underline{P}$ (on $\mathcal{D}_L$), and every coherent extension $\underline{P}^*$ of $\underline{P}$ dominates $\underline{E}$ on $\mathcal{D}'_L$.

If condition '$\forall s_0, s_1, \ldots, s_n \geq 0$' is replaced by '$\forall s_0, s_1, \ldots, s_n \in \mathbb{R}$' in the definition of coherent upper prevision, we obtain de Finetti's notion of *dF-coherent* (precise) prevision [2]. A dF-coherent prevision P is coherent both as an upper and as a lower prevision. The precise previsions or probabilities in the sequel are meant to be dF-coherent.

Although the domain of an upper prevision may be arbitrary, it will have a special structure in most of the paper, to exploit results on 2-alternating previsions.

More specifically, a set of events $\mathcal{A}$ is a *field* when $\emptyset \in \mathcal{A}$ and $A \vee B, A^c \in \mathcal{A}, \forall A, B \in \mathcal{A}$. If $\mathcal{A}$ is a field, a gamble X is $\mathcal{A}$-*measurable* when the events $X > x$ and $X < x$ are in $\mathcal{A}$, $\forall x \in \mathbb{R}$.

A set of gambles S is a *lattice* if X, $Y \in S$ implies $\max(X,Y) \in S$ and $\min(X,Y) \in S$.

An upper prevision $\overline{P}$ defined on a lattice S is 2-*alternating* iff $\overline{P}(\max(X,Y)) \leq \overline{P}(X) + \overline{P}(Y) - \overline{P}(\min(X,Y))$, $\forall X, Y \in S$. A lower prevision $\underline{P}$ on S is 2-*monotone* iff $\underline{P}(\max(X,Y)) \geq \underline{P}(X) + \underline{P}(Y) - \underline{P}(\min(X,Y))$, $\forall X, Y \in S$.

Results stated for 2-monotone previsions are easily reworded for 2-alternating ones (and vice versa), since the conjugate $\overline{P}(X) = -\underline{P}(-X)$ of a 2-monotone lower prevision is 2-alternating (and vice versa).

When S is a set of (indicators of) events and $\overline{P}$ is therefore an upper probability, S is a lattice iff A, $B \in S$ implies $A \vee B \in S$, $A \wedge B \in S$, and $\overline{P}$ is 2-alternating iff $\overline{P}(A \vee B) \leq \overline{P}(A) + \overline{P}(B) - \overline{P}(A \wedge B)$, $\forall A, B \in S$. With a mild additional condition, 2-alternating upper probabilities are coherent [1]:

Proposition 1. *Let $\overline{P}$ be a 2-alternating upper probability on a lattice S containing $\emptyset$ and Ω. Then $\overline{P}$ is coherent iff $\overline{P}(\emptyset) = 0$ and $\overline{P}(\Omega) = 1$.*

Notation Let S^+ be a lattice of events containing $\emptyset$ and Ω.

One way to obtain coherent 2-alternating upper probabilities defines $\overline{P}$ as a special *distorted probability*, by the following result, adapted from [3], Example 2.1.

Proposition 2. *Let P be a dF-coherent probability on S^+ and ϕ a (weakly) increasing concave function defined on $[0;1]$ with $\phi(0) = 0$, $\phi(1) = 1$. Then the* distorted probability $\overline{P}(\cdot) = \phi(P(\cdot))$ *is a 2-alternating and coherent upper probability.*

Proposition 2 ensures that $\overline{P}$ in (1) is 2-alternating and coherent (put $\phi(x) = \min((1+\delta)x, 1)$), hence its conjugate $\underline{P}$ is 2-monotone and coherent.

To deal with the natural extension of the PMM in Section 3, the following Proposition 3 will be exploited.

Notation The natural extension of interest is that of $\overline{P}$ from S^+ to the set $\mathcal{L} = \mathcal{L}(\mathbb{P}_u)$ of all gambles defined on a "universal" partition $\mathbb{P}_u$ (termed Ω in [11]). That is, $\mathbb{P}_u$ is a set of pairwise disjoint events, whose sum is the sure event Ω, and such that its powerset $2^{\mathbb{P}_u}$ contains all the events of interest. In particular $S^+ \subseteq 2^{\mathbb{P}_u}$. Given $\overline{P} : S^+ \to \mathbb{R}$, its *outer (set) function* $\overline{P}^*$ is defined on $2^{\mathbb{P}_u}$ by $\overline{P}^*(B) = \inf\{\overline{P}(A) : A \in S^+, B \Rightarrow A\}$, $\forall B \in 2^{\mathbb{P}_u}$.

Proposition 3. *[1] Let $\overline{P} : S^+ \to \mathbb{R}$ be a coherent 2-alternating upper probability. Its natural extension $\overline{E}$ on $\mathcal{L}$ is given by*

$$\overline{E}(X) = \inf X + \int_{\inf X}^{\sup X} \overline{P}^*(X > x)dx \tag{4}$$

and is 2-alternating too. Further,

(a) The restriction of $\overline{E}$ on $2^{\mathbb{P}_u}$ coincides with the outer function $\overline{P}^*$.

(b) If $S^+ = 2^{\mathbb{P}_u}$, $\overline{E}$ is the only *2-alternating coherent extension of $\overline{P}$ on $\mathcal{L}$.*

In Section 5 we shall be concerned with natural extensions on conditional events, like $\overline{E}(A|B)$ or $\underline{E}(A|B)$, while precise conditional previsions, like $P(X|X > x_\tau)$, appear in Section 3. Although the paper presentation does not focus on coherence concepts in a conditional environment, our approach employs formally *Williams coherence* or *W-coherence*, in the version presented in [7], Definition 4, which unlike Walley's coherence in [11, Section 7.1.4 (b)] imposes no structure constraints on the domain $\mathcal{D}$ of the upper or lower previsions. However, when finitely many conditioning events are involved in $\mathcal{D}$ (as is always the case in the paper), Williams and Walley's coherence are equivalent (after extending the given W-coherent prevision on a suitable set $\mathcal{D}'$, which can be always done keeping W-coherence, cf. [7]). Thus the results in the paper hold also in terms of Walley's coherence.

Several necessary conditions hold for W-coherence, whenever they are well-defined. Recall *internality*: $\inf(X|B) \le \overline{P}(X|B) \le \sup(X|B)$, where for instance $\sup(X|B) = \sup\{X(\omega)|\omega \Rightarrow B\}$, and the *Generalized Bayes Rule* (GBR) $\overline{P}(I_A(X - \overline{P}(X|A))) = 0$, which in the case of precise previsions specialises to

$$P(XI_A) = P(X|A)P(A). \tag{5}$$

3 Extending the pari-mutuel model

The natural extension $\overline{E}$ of $\overline{P}(A) = \min\{(1+\delta)P(A), 1\}$ from a field $\mathcal{A}$ to any $\mathcal{A}$-measurable gamble X was shown in [11] to be

$$\overline{E}(X) = x_\tau + (1+\delta)P((X - x_\tau)^+), \tag{6}$$

where $(X - x_\tau)^+ = \max\{X - x_\tau, 0\}$ and the (upper) *quantile* x_τ is defined as

$$x_\tau = \sup\{x \in \mathbb{R} : P(X \le x) \le \tau\}. \tag{7}$$

An alternative expression for $\overline{E}(X)$ is:[1]

$$\begin{gathered}\overline{E}(X) = (1-\varepsilon)P(X|X > x_\tau) + \varepsilon x_\tau, \\ \varepsilon = 1 - (1+\delta)P(X > x_\tau).\end{gathered} \tag{8}$$

It is also stated in [11] that $\overline{E}(X) = P(X|X > x_\tau)$ if X has a continuous *distribution function* $F_X(x) \overset{\text{def}}{=} P(X \le x)$.

We shall now explore more thoroughly the relationship between $\overline{E}(X)$ and $P(X|X > x_\tau)$. The results will be exploited also in Section 4, where they will be reinterpreted in a risk measurement perspective.

To begin with, we gather some known or anyway elementary, but useful facts in the following proposition.

Proposition 4. *Let X be $\mathcal{A}$-measurable and for $\tau \in]0;1[$ define: x_τ by (7), $F_X(x_\tau^+) = \lim_{x \to x_\tau^+} F_X(x)$, $F_X(x_\tau^-) = \lim_{x \to x_\tau^-} F_X(x)$.*

a) $\tau \in [F_X(x_\tau^-); F_X(x_\tau^+)]$; besides, all values of τ in $[F_X(x_\tau^-); F_X(x_\tau^+)[$ originate by (7) the same (upper) quantile x_τ.

b) $\inf X \le x_\tau \le \sup X$.

[1] Equation (8) is stated without proof in [11], Note 3 to Section 3.2. A proof may follow from the later Proposition 9.

c) $(X > x_\tau) = \emptyset$ *iff* $x_\tau = \sup X$*; if* $(X \leq x_\tau) = \emptyset$ *then* $x_\tau = \inf X$.

d) It holds for ε *in (8) that* $\varepsilon \lesseqgtr 0$ *iff* $\tau \gtreqless F_X(x_\tau)$.[2]

Corollary 1. *If* $(X > x_\tau) = \emptyset$, $\overline{E}(X) = \sup X$.

Proof. Substitute (by Proposition 4, c) $x_\tau = \sup X$ in (6), noting that $P((X - x_\tau)^+) = P(0) = 0$. □

Remark 1. When P is σ-additive, $F_X(x_\tau^+) = F_X(x_\tau)$, i.e. F_X is right-continuous. But an often neglected issue broadens the number of possible alternatives in comparing $\overline{E}(X)$ with $P(X|X > x_\tau)$ (and with another extension in the next Section 3.1): since F_X is originated by a not necessarily σ-additive probability P, there may exist non-zero *adherent probabilities* at x_τ (cf. [2], Section 6.4.11). Precisely,

$$F_X(x_\tau^+) - F_X(x_\tau^-) = P_{x_\tau}^- + P_{x_\tau}^+ + P(X = x_\tau),$$

where $P_{x_\tau}^- = F_X(x_\tau) - F_X(x_\tau^-) - P(X = x_\tau)$ is the *left adherent* probability at x_τ, $P_{x_\tau}^+ = F_X(x_\tau^+) - F_X(x_\tau)$ is the *right adherent* probability at x_τ. Hence,

$$F_X(x_\tau) = F_X(x_\tau^-) + P_{x_\tau}^- + P(X = x_\tau). \tag{9}$$

While $P_{x_\tau}^+$ is zero iff F_X is right-continuous at x_τ (always if P is σ-additive), from (9), F_X may be left-discontinuous in x_τ also when $P_{x_\tau}^- = 0$, if $P(X = x_\tau) > 0$ (σ-additivity of P implies $P_{x_\tau}^- = 0$). □

Proposition 5. *a) If* $P(X|X > x_\tau) = x_\tau$*, then* $\overline{E}(X) = P(X|X > x_\tau)$.

b) If $P(X|X > x_\tau) > x_\tau$*, then* $\overline{E}(X) \lesseqgtr P(X|X > x_\tau)$ *iff* $\tau \lesseqgtr F_X(x_\tau)$.

Proof. Using (8), $\overline{E}(X) \lesseqgtr P(X|X > x_\tau)$ iff $\varepsilon(x_\tau - P(X|X > x_\tau)) \lesseqgtr 0$, from which a) follows immediately, b) using also Proposition 4, d). □

Proposition 5, a) considers a really extreme situation. Assuming from now that $P(X|X > x_\tau) > x_\tau$, Proposition 5, b) reduces the comparison between $\overline{E}(X)$ and $P(X|X > x_\tau)$ to comparing τ and $F_X(x_\tau)$ in the further subcases that can be identified. The most notable instances are:

i) F_X is continuous at x_τ. This implies $\tau = F_X(x_\tau)$, and $\overline{E}(X) = P(X|X > x_\tau)$.

ii) F_X is right-continuous, but not continuous at x_τ, and $\tau \neq F_X(x_\tau)$. This implies $F_X(x_\tau) = F_X(x_\tau^+) > \tau$, and $P(X|X > x_\tau) > \overline{E}(X)$.

[2] We write $\lesseqgtr$ or $\gtreqless$ to summarize three conditions, here $\varepsilon < 0$ iff $\tau > F_X(x_\tau)$, $\varepsilon = 0$ iff $\tau = F_X(x_\tau)$, $\varepsilon > 0$ iff $\tau < F_X(x_\tau)$.

Case ii) is the most obvious instance that ensures $P(X|X > x_\tau) > \overline{E}(X)$, but not the only one. By Proposition 4, a), it can be $\tau < F_X(x_\tau)$ also when F_X is not right-continuous (while being left-discontinuous). Similarly, there are other cases when $P(X|X > x_\tau) = \overline{E}(X)$, because $\tau = F_X(x_\tau)$, apart from case i), which remains the most important one. And it is also possible that

iii) $P(X|X > x_\tau) < \overline{E}(X)$.

Obviously, case iii) cannot occur when P is σ-additive, since it is equivalent to $\tau > F_X(x_\tau)$, hence $\tau \in]F_X(x_\tau); F_X(x_\tau^+)] = I^>$ and $I^> \neq \varnothing$ iff $P_{x_\tau}^+ > 0$.

When $P(X|X > x_\tau) > \overline{E}(X)$, then $P(X|X > x_\tau)$ is clearly not a coherent extension to X of $\overline{P}$ in the PMM, while it is so when it coincides with $\overline{E}(X)$.

3.1 Comparison with a naive extension

In actuarial applications the upper probability $\overline{P}(A)$ in (1) is the price, determined by increasing P by a loading $\delta > 0$, of an insurance policy which pays 1 unit if and only if event A occurs. In analogy with (1), one could set the price of an insurance policy which refunds x units iff the loss $X = x$ occurs, to $(1+\delta)P(X)$, up to a maximum of $\sup X$. Here $P(X)$ is the expectation of X computed from P. This procedure defines the *naive extension*:

$$\overline{P}_N(X) = \min\{(1 + \delta)P(X), \sup X\}.$$

This extension, without the upper bound $\sup X$ (which is however necessary for $\overline{P}_N$ to be coherent), is referred to as *expected value principle* in risk theory literature [5, p. 67]. To fix the framework, suppose (throughout this section only) that P is defined on the field $2^{\mathbb{P}_u}$, and that we are interested in extending it to some set $\mathcal{D}$ strictly contained in the cone $\mathcal{L}^+(\mathbb{P}_u)$ of the non-negative gambles in $\mathcal{L}(\mathbb{P}_u)$. The gambles in $\mathcal{D}$ are non-negative, being refunds to the insured: hence $\inf X \geq 0$, $\forall X \in \mathcal{D}$.

The inclusion $\mathcal{D} \subsetneqq \mathcal{L}^+(\mathbb{P}_u)$ is strict because $\overline{P}_N$ cannot in general be coherent on a set $\mathcal{D}$ containing X, $X + k$, when $k \in \mathbb{R}^+$ is large enough. For instance, if $\overline{P}_N(X) = (1 + \delta)P(X) < \sup X$, then $\overline{P}_N(X + k) = \sup X + k > \overline{P}_N(X) + k$ for $k \geq \frac{\sup X - (1+\delta)P(X)}{\delta}$, violating property (c) in [11], Section 2.6.1, which is a necessary condition for coherence.

But even when $\mathcal{D} = \{X\}$, $\overline{P}_N$ may be incoherent with the PMM:

Example 1. Take $\mathbb{P}_u = \{e_0, e_1, e_2, e_3\}$, and let $X(e_i) = i, i = 0, \ldots, 3$, $P(X = 0) = 0$, $P(X = 1) = 0.1$, $P(X = 2) = 0.5$, $P(X = 3) = 0.4$ and $\delta = 1/10$.

Then $P(X) = 2.3$ and hence $\overline{P}_N(X) = 2.53$. Let us now compute the natural extension in X. We have that $\tau = \frac{\delta}{1+\delta} = 1/11$, hence $x_\tau = 1$, as can be checked using F_X. Applying (6), $\overline{E}(X) = 1 + \frac{11}{10}P(\max\{X-1, 0\}) = 1 + \frac{11}{10}1.3 = 2.43$. □

In Example 1, $\overline{P}_N(X) > \overline{E}(X)$. This is interesting because the natural extension is shown to lead to a price smaller than would be expected from the intuition at the basis of the PMM and also because $\overline{P}_N$ is incoherent with the PMM, being larger than $\overline{E}$.

The dominance relationship between $\overline{P}_N$ and $\overline{E}$ is the object of the following proposition.

Proposition 6. *a) If $(X > x_\tau) = \emptyset$ then $\overline{P}_N(X) \leq \overline{E}(X)$.*

b) If either $(X \leq x_\tau) = \emptyset$ or $\overline{P}_N(X) = \sup X$, then $\overline{P}_N(X) \geq \overline{E}(X)$.

Suppose now $(X > x_\tau) \neq \emptyset$, $(X \leq x_\tau) \neq \emptyset$, $\overline{P}_N(X) < \sup X$, and let ε be given by (8).

c) If $\varepsilon = 0$, then $\overline{P}_N(X) \geq \overline{E}(X)$ and $\overline{P}_N(X) = \overline{E}(X)$ iff $P(X|X \leq x_\tau) = 0$.

d) If $\varepsilon \neq 0$ and $x_\tau = 0$, then $\overline{P}_N(X) = \overline{E}(X)$.

e) If $\varepsilon \neq 0$ and $x_\tau > 0$, condition $F_X(x_\tau) < \tau$ implies $\overline{P}_N(X) > \overline{E}(X)$, while condition $F_X(x_\tau) > \tau$ is necessary, but not sufficient, to ensure $\overline{P}_N(X) \leq \overline{E}(X)$.

Proof. a) follows from Corollary 1. To prove the non-trivial implication in b), put (Proposition 4, c)) $x_\tau = \inf X$ in (6), to get $\overline{E}(X) = \inf X + (1+\delta)P(X - \inf X) = (1+\delta)P(X) - \delta \inf X \leq \min\{(1+\delta)P(X), \sup X\} = \overline{P}_N(X)$.

To prove c), write $(1+\delta)P(X) = P(X|X > x_\tau)(1+\delta)P(X > x_\tau) + (1+\delta)P(X|X \leq x_\tau)P(X \leq x_\tau) = P(X|X > x_\tau)(1-\varepsilon) + (1+\delta)P(X|X \leq x_\tau)(1 - P(X > x_\tau)) = P(X|X > x_\tau)(1-\varepsilon) + (1+\delta)P(X|X \leq x_\tau) - P(X|X \leq x_\tau)(1-\varepsilon) = (1-\varepsilon)P(X|X > x_\tau) + (\delta + \varepsilon)P(X|X \leq x_\tau)$. From here

$$\overline{P}_N(X) = \min\{(1-\varepsilon)P(X|X > x_\tau) + (\delta+\varepsilon)P(X|X \leq x_\tau), \sup X\}.$$

Comparing this equality and (8),

$$\overline{P}_N(X) \gtreqless \overline{E}(X) \text{ iff } (\delta+\varepsilon)P(X|X \leq x_\tau) \gtreqless \varepsilon x_\tau. \quad (10)$$

When $\varepsilon = 0$, c) follows directly from (10).

To prove the remaining cases, we write the right-hand side inequality in (10) in a different form. Since $\delta + \varepsilon = (1+\delta)P(X \leq x_\tau)$ and $\varepsilon x_\tau = ((1+\delta) - (1+\delta)P(X > x_\tau) - \delta)x_\tau = ((1+\delta)P(X \leq x_\tau) - \delta)x_\tau$, using also (5) we get $\overline{P}_N(X) \gtreqless \overline{E}(X)$ iff $P(XI_{X \leq x_\tau}) \gtreqless (P(X \leq x_\tau) - \frac{\delta}{1+\delta})x_\tau$, or equivalently

$$\overline{P}_N(X) \gtreqless \overline{E}(X) \text{ iff } P(XI_{X \leq x_\tau}) \gtreqless (F_X(x_\tau) - \tau)x_\tau.$$

From here and Proposition 4 d), parts d) and e) follow at once (for d), recall that $x_\tau = 0$ implies $P(XI_{X \leq x_\tau}) = 0$). □

It appears from Proposition 6 that $\overline{P}_N$ is only occasionally equal to $\overline{E}$, and may easily be incoherent. Cases a), b), d) treat really extreme situations, while in the common case that F_X is continuous at x_τ, c) ensures that $\overline{P}_N$ is incoherent, unless the limiting evaluation $P(X|X \leq x_\tau) = 0$ applies. Case e) shows that $\overline{P}_N$ can possibly be coherent when $F_X(x_\tau) > \tau$. The most important practical case concerns discrete gambles (with finitely many possible values). However, it should be checked even then whether $\overline{P}_N \leq \overline{E}$, and this makes the use of $\overline{P}_N$ less convenient. For instance, $\overline{P}_N > \overline{E}$ in Example 1.

3.2 A generalization

We shall derive here $\overline{E}$ in the more general framework of Proposition 3, that $\overline{P}$ is defined by the PMM on S^+ and $\overline{E}$ on $\mathcal{L}(I\!P_u)$. We first obtain an expression for $\overline{E}(B)$, for any event B in $2^{I\!P_u}$.

Proposition 7. *In the PMM, the natural extension of $\overline{P} : S^+ \to \mathbb{R}$ on $2^{I\!P_u}$ is*

$$\overline{E}(B) = \min\{(1+\delta)\tilde{P}^*(B), 1\}, \quad (11)$$

where the upper probability $\tilde{P}^*(B) = \inf\{P(A) : A \in S^+, B \Rightarrow A\}$ is the outer function of P.

Proof. By Proposition 3 (a), $\overline{E}(B) = \overline{P}^*(B) = \inf\{\min\{(1+\delta)P(A), 1\} : A \in S^+, B \Rightarrow A\}$. Defining $L_B = \{A \in S^+ : B \Rightarrow A, (1+\delta)P(A) < 1\}$, $L_B = \varnothing$ iff $(1+\delta)\tilde{P}^*(B) \geq 1$.

Two cases may occur: if $L_B = \varnothing$, that is if $(1+\delta)\tilde{P}^*(B) \geq 1$, then $\overline{E}(B) = 1$; if $L_B \neq \varnothing$, that is if $(1+\delta)\tilde{P}^*(B) < 1$, $\overline{E}(B) = \inf\{(1+\delta)P(A) : A \in L_B\} = (1+\delta)\inf\{P(A) : A \in L_B\} = (1+\delta)\tilde{P}^*(B)$. In summary, equation (11) holds. □

We emphasize that $\tilde{P}^*$ in (11) is generally not a precise, but an upper probability. In fact, by Proposition 3 (a), it coincides with the natural extension $\overline{E}_P$ on $2^{I\!P_u}$ of the probability P, when P is interpreted as a special upper probability.

Proposition 8. *In the PMM, the natural extension of $\overline{P} : S^+ \to \mathbb{R}$ on $\mathcal{L}(I\!P_u)$ is:*

$$\overline{E}(X) = x_\tau^u + (1+\delta)\overline{E}_P((X - x_\tau)^+) \quad (12)$$

where $\overline{E}_P$ is the natural extension of P (also of $\tilde{P}^*$) on $\mathcal{L}$, and x_τ^u is the (upper) quantile relative to $\tilde{P}^*$

$$x_\tau^u = \sup\{x \in \mathbb{R} : \tilde{P}^*(X \leq x) \leq \tau\}. \quad (13)$$

Proof. Apply (4) and Proposition 3, (a) substituting $\overline{P}^* = \overline{E}$ with its expression in equation (11), getting

$$\overline{E}(X) = \inf X + \int_{\inf X}^{\sup X} \min\{(1+\delta)\tilde{P}^*(X > x), 1\}dx.$$

From here, the derivation of (12) is identical to that sketched in [11], Section 3.2.5, to obtain (6). In fact, $\tilde{P}^*$ is defined on the field $2^{I\!P_u}$, and every $X \in \mathcal{L}$ is measurable with respect to such a field. □

Clearly, (12) generalizes (6). We might summarize the difference between the natural extension in (12) and that in (6) as follows: computing the natural extension of $\overline{P}$ on gambles which are not necessarily measurable with respect to the domain of $\overline{P}$ introduces imprecision by transforming the precise prevision $P((X - x_\tau)^+)$ in (6) into the upper prevision $\overline{E}_P((X - x_\tau^u)^+)$ in (12). Also the quantile x_τ refers to probability P in (7), while x_τ^u employs the upper probability $\tilde{P}^*$ in (13).

But there is another attractive interpretation: $\overline{E}(B)$ in (11) can be viewed as a kind of *imprecise PMM*, defined via natural extension on $2^{I\!P_u}$ starting from a (precise) PMM on a narrower set S^+: then (12) describes the natural extension of this imprecise model.

Some properties of the natural extension of the PMM generalize to the natural extension of the imprecise PMM. The following proposition relaxes (8):

Proposition 9. *If $(X > x_\tau^u) \neq \emptyset$, it holds for the natural extension $\overline{E}$ on $\mathcal{L}(I\!P_u)$ of $\overline{P} : S^+ \to \mathbb{R}$ that*

$$\overline{E}(X) \leq \varepsilon^u x_\tau^u + (1 - \varepsilon^u)\overline{E}_P(X|X > x_\tau^u) \qquad (14)$$

where $\varepsilon^u \stackrel{\text{def}}{=} 1 - (1+\delta)\overline{E}_P(X > x_\tau^u)$.

Proof. Noting that $(X - x_\tau^u)^+ = (X - x_\tau^u)I_{X>x_\tau^u}$ and by subadditivity of coherent upper previsions and, at the second equality, the GBR,[3] $\overline{E}_P((X - x_\tau^u)^+) = \overline{E}_P((X - x_\tau^u)I_{X>x_\tau^u}) \leq \overline{E}_P(I_{X>x_\tau^u}(X - \overline{E}_P(X|X > x_\tau^u))) + \overline{E}_P(I_{X>x_\tau^u}(\overline{E}_P(X|X > x_\tau^u) - x_\tau^u)) = \overline{E}_P(I_{X>x_\tau^u}(\overline{E}_P(X|X > x_\tau^u) - x_\tau^u)) \stackrel{\text{def}}{=} \lambda$.

Using also the definition of ε^u and λ, we get further $x_\tau^u + (1+\delta)\lambda = x_\tau^u(1 - (1+\delta)\overline{E}_P(X > x_\tau^u)) + (1+\delta)(\lambda + x_\tau^u\overline{E}_P(X > x_\tau^u)) = \varepsilon^u x_\tau^u + (1+\delta)(\overline{E}_P(X > x_\tau^u)(\overline{E}_P(X|X > x_\tau^u) - x_\tau^u) + x_\tau^u\overline{E}_P(X > x_\tau^u)) = \varepsilon^u x_\tau^u + (1 - \varepsilon^u)\overline{E}_P(X|X > x_\tau^u)$.

Finally, by (12) and the expressions above, $\overline{E}(X) = x_\tau^u + (1+\delta)\overline{E}_P((X - x_\tau^u)^+) \leq x_\tau^u + (1+\delta)\lambda = \varepsilon^u x_\tau^u + (1 - \varepsilon^u)\overline{E}_P(X|X > x_\tau^u)$. □

Although the inequality in (14) can be strict (we omit proving this), when $\overline{P}$ is defined on $2^{I\!P_u}$ then $\overline{E}_P$ is equal to P (or to its extension using (5)), and x_τ^u, ε^u to x_τ, ε respectively. Thus (14) reduces to (8).

The statement corresponding to Proposition 4 d) is $\varepsilon^u \gtreqless 0$ iff $\overline{E}_P(X > x_\tau^u) \lesseqgtr \frac{1}{1+\delta}$, or also $\varepsilon^u \gtreqless 0$ iff $\underline{E}_P(X \leq x_\tau^u) \gtreqless \tau$.

We know that $\varepsilon = 0$ when F_X is continuous at x_τ. When $\overline{F}_X(x) = \overline{E}_P(X \leq x)$ is continuous at x_τ^u, then $\overline{F}_X(x_\tau^u) = \tau$. Hence $\underline{F}_X(x_\tau^u) = \underline{E}_P(X \leq x_\tau^u) \leq \overline{F}_X(x_\tau^u) = \tau$. In terms of ε^u, as seen above, this means that $\varepsilon^u \leq 0$, with $\varepsilon^u = 0$ only when $\underline{F}_X(x_\tau^u) = \overline{F}_X(x_\tau^u)$, a condition obviously warranted when $\underline{F}_X = \overline{F}_X = F_X$. Thus continuity at x_τ^u of $\overline{F}_X$ implies $\varepsilon^u \leq 0$, typically $\varepsilon^u < 0$.

4 Risk measurement interpretations

If Y is a gamble, it is known [6] that $\overline{P}(-Y)$ may be interpreted as a *risk measure* for Y, i.e. a number measuring how risky Y is, or also the amount of money to be reserved to cover potential losses from Y. Several risk measures were introduced in the literature, and there is often no unanimity on the terminology. To ensure comparisons with [4], we shall refer the risk measure to $X = -Y$; this corresponds, when $Y \leq 0$, to thinking in terms of losses and is frequently done in insurance, where X represents the amount to be paid for insurance claims (however, X is not necessarily non-negative in what follows).[4] Thus the upper previsions $\overline{E}(X)$ in (6), (8) and (12) may be seen as risk measures for X, and there is a strong correspondence with measures studied in the literature.

Consider equation (6): x_τ is the *Value-at-Risk* of X at level τ, $VaR_\tau(X)$, while $P((X - x_\tau)^+)$ is the *expected shortfall* $ES_\tau(X)$ (whenever P is replaced by or thought of as an expectation) [4]. In fact, $(X - x_\tau)^+$ measures the shortfall, i.e. the residual loss in absolute value of an agent who reserves an amount of money equal to $VaR_\tau(X) = x_\tau$ to cover losses from X. Also $P(X|X > x_\tau)$ corresponds to a well-known risk measure (when P is an expectation), termed *Conditional Tail Expectation* (CTE_τ) in [4].

Equation (6) corresponds to (2.7) in [4], which defines another measure of risk, $TailVaR_\tau(X)$ or $TVaR_\tau(X)$. This equation is identical to (6), after replacing $\overline{E}$, x_τ, $P((X - x_\tau)^+)$ with, respectively, $TVaR_\tau(X)$, $VaR_\tau(X)$, $ES_\tau(X)$:

$$TVaR_\tau(X) = VaR_\tau(X) + (1+\delta)ES_\tau(X).$$

[3] Recall also that the natural extension $\overline{E}_P$ always exists with W-coherence, cf. [7].

[4] While ensuring compatibility with the prevailing literature and the formulae in [11], the convention of referring to losses modifies the range of the typical values for τ. In this section τ should be fairly close to 1, representing the probability that the loss is not too high, while in the rest of the paper should rather be close to 0, being a taxation or commission.

Analogously, equation (8) corresponds to

$$TVaR_\tau(X) = (1-\varepsilon)CTE_\tau(X) + \varepsilon VaR_\tau(X). \quad (15)$$

The novel fact in our approach (apart from using previsions instead of expectations) is that $TVaR_\tau$ is derived as the natural extension of the PMM, while the starting point in the literature for defining this or other measures is usually a set of random variables, often a linear space equipped with a σ-additive probability measure, using which the various expectations are computed. Recalling also Proposition 3, we deduce the following properties for $TVaR_\tau$:

Proposition 10. *$TVaR_\tau(X)$ is the natural extension on $\mathcal{L}(\mathbb{P}_u)$ of the PMM defined on $2^{\mathbb{P}_u}$. Hence, it is the least-committal risk measure extending the PMM which is coherent. Actually, it is its* only *coherent extension which is 2-alternating.*

CTE_τ complements VaR_τ, in the sense that VaR_τ, unlike CTE_τ, is nearly uninformative about what are the losses, should the threshold x_τ be exceeded. Unfortunately, neither VaR_τ nor CTE_τ is generally coherent, even though their linear combination in (15) originates a coherent risk measure. Conditions for coherence of CTE_τ are discussed in Section 3, and are commoner in practice than those ensuring coherence of VaR_τ.[5] In the classical risk measurement approach using a σ-additive probability, the comparison between CTE_τ and $TVaR_\tau$ is limited to cases i), ii) in Section 3 which, as we pointed out there, are not exhaustive in general.

The generalization in Section 3.2 forms a basis for further considerations on the risk measurement side. This time, $\overline{E}(X)$ in (12) is the natural extension of the PMM defined on $S^+(\subset 2^{\mathbb{P}_u})$, and may again be interpreted as a risk measure, let us name it *Imprecise TailVar* or $ITVaR_\tau$. Using Proposition 3, $ITVaR_\tau$ is coherent and also 2-alternating. However, $ITVaR_\tau$ has no analogue in the risk measurement literature. The reason lies in the standard way of defining risk measures from an underlying precise probability, which rules out potentially interesting risk measures which are functions of imprecise measures. And looking at (12), we notice that $ITVaR_\tau$ is a linear combination of other two measures which are imprecise versions of VaR_τ and ES_τ: x^u_τ is defined in (13) as a function of the upper probability $\tilde{P}^*$, the shortfall $(X - x^u_\tau)^+$ is evaluated by the upper prevision $\overline{E}_P$. We may conclude that the PMM provides a formal justification for the existence of a new, and still largely not investigated, kind of risk measures, those defined in terms of imprecise uncertainty measures.

[5] For VaR_τ, see the discussion in [6].

5 Conditioning the pari-mutuel model

Reconsider the basic PMM, with $\overline{P}(A)$, $\underline{P}(A)$ given by (1), (2), $A \in \mathcal{D}$, and $\mathcal{D}$ is now a *field* of events. We shall compute the natural extensions $\overline{E}(A|B)$, $\underline{E}(A|B)$ of $\underline{P}$ and $\overline{P}$ on $A|B$, with $B \in \mathcal{D}$, $B \neq \emptyset$. Since $\overline{P}$ and $\underline{P}$ are, respectively, 2-alternating and 2-monotone, from a well-known result ([10], Thm. 7.2; see also [8]), when $\underline{P}(B) > 0$:

$$\begin{aligned} \overline{E}(A|B) &= \frac{\overline{P}(A \wedge B)}{\overline{P}(A \wedge B) + \underline{P}(A^c \wedge B)}, \\ \underline{E}(A|B) &= \frac{\underline{P}(A \wedge B)}{\underline{P}(A \wedge B) + \overline{P}(A^c \wedge B)}. \end{aligned} \quad (16)$$

When $\underline{P}(B) = 0$, equations (16) do not apply, but it can be shown (directly, using Williams coherence, or alternatively from results in [11]) that

Lemma 1. *Given a coherent lower probability $\underline{P}$ on a set $\mathcal{D}$ of (unconditional) events, let $B \in \mathcal{D}$, $\underline{P}(B) = 0$. The natural extension $\underline{E}$ of $\underline{P}$ on $\mathcal{D} \cup \{A_1|B, \ldots, A_n|B\}$ is $\underline{E}(A_i|B) = 1$ if $B \Rightarrow A_i$, $\underline{E}(A_i|B) = 0$ otherwise, for $i = 1, \ldots, n$.*

Applying Lemma 1 for $n = 2$, $A_1 = A$, $A_2 = A^c$ and using conjugacy, it follows that, when $\underline{P}(B) = 0$ in the PMM, then $\underline{E}(A|B) = 0$, $\forall A$ such that $B \not\Rightarrow A$, and $\overline{E}(A|B) = 1$, $\forall A$ such that $A \wedge B \neq \emptyset$.

We assume in the sequel $\underline{P}(B) > 0$; note that by (2) $\underline{P}(B) > 0$ iff $P(B) > \frac{\delta}{\delta+1} = \tau$. Further, $\underline{P}(B) > 0$ ensures that the denominators in (16) are non-zero. Take $\overline{E}(A|B)$: using property 2.7.4 (d) in [11], $\overline{P}(A \wedge B) + \underline{P}(A^c \wedge B) \geq \underline{P}(B) > 0$. Similarly for $\underline{E}(A|B)$.

To derive $\overline{E}(A|B)$, from (16), two alternatives occur:

a) $\underline{P}(A^c \wedge B) = \max\{(1+\delta)P(A^c \wedge B) - \delta, 0\} = 0$. Hence $\overline{E}(A|B) = 1$.

b) $\max\{(1+\delta)P(A^c \wedge B) - \delta, 0\} > 0$. This happens iff $P(A^c \wedge B) > \frac{\delta}{1+\delta} = \tau$ and implies $\min\{(1+\delta)P(A \wedge B), 1\} < 1$ (otherwise $P(A \wedge B) \geq \frac{1}{1+\delta}$ and $P(B) > \frac{\delta}{\delta+1} + \frac{1}{1+\delta} = 1$). Hence $\overline{E}(A|B) = \frac{(1+\delta)P(A \wedge B)}{(1+\delta)(P(A \wedge B) + P(A^c \wedge B)) - \delta} = \frac{P(A \wedge B)}{P(B) - \tau}$.

The derivation of $\underline{E}(A|B)$ is analogous:

a) If $\underline{P}(A \wedge B) = \max\{(1+\delta)P(A \wedge B) - \delta, 0\} = 0$, $\underline{E}(A|B) = 0$.

b) If $\max\{(1+\delta)P(A \wedge B) - \delta, 0\} > 0$, this implies $\tau < P(A \wedge B)$ and $\min\{(1+\delta)P(A^c \wedge B), 1\} < 1$; then $\underline{E}(A|B) = \frac{P(A \wedge B) - \tau}{P(B) - \tau}$.

$$\overline{P}(A) = \begin{cases} \frac{P(A)}{1-\tau} & \text{if } \tau < P(A^c) \\ 1 & \text{if } \tau \geq P(A^c) \end{cases}$$

$$\underline{P}(A) = \begin{cases} \frac{P(A)-\tau}{1-\tau} & \text{if } \tau < P(A) \\ 0 & \text{if } \tau \geq P(A) \end{cases}$$

$$\overline{E}(A|B) = \begin{cases} \frac{P(A\wedge B)}{P(B)-\tau} & \text{if } \tau < P(A^c \wedge B) \\ 1 & \text{if } \tau \geq P(A^c \wedge B) \end{cases}$$

$$\underline{E}(A|B) = \begin{cases} \frac{P(A\wedge B)-\tau}{P(B)-\tau} & \text{if } \tau < P(A \wedge B) \\ 0 & \text{if } \tau \geq P(A \wedge B) \end{cases}$$

Table 1: Values of $\overline{P}(A)$, $\underline{P}(A)$, $\overline{E}(A|B)$, $\underline{E}(A|B)$.

Table 1 lists the values of $\overline{P}(A)$, $\underline{P}(A)$, $\overline{E}(A|B)$, $\underline{E}(A|B)$. They are written as functions of τ, to simplify the inequalities in the 'if' clauses (referring to δ, the clauses involve ratios of probabilities instead of probabilities). Note that the expressions for $\overline{E}(A|B)$, $\underline{E}(A|B)$ reduce to those for $\overline{P}(A)$, $\underline{P}(A)$ when $B = \Omega$.

5.1 Dilation and imprecision increase

How does imprecision in the evaluations vary when conditioning in the PMM model? To supply some answers, we first recall two concepts.

Definition 1. Given a partition of non-impossible events $\mathbb{P}$, we say that *(weak) dilation* occurs (with respect to A and $\mathbb{P}$) when

$$\underline{P}(A|B) \leq \underline{P}(A) \leq \overline{P}(A) \leq \overline{P}(A|B), \forall B \in \mathbb{P}, \quad (17)$$

while there is an *imprecision increase* when

$$\overline{P}(A) - \underline{P}(A) \leq \overline{P}(A|B) - \underline{P}(A|B), \forall B \in \mathbb{P}. \quad (18)$$

Dilation is a so far little investigated phenomenon (see [9]), which implies that our a posteriori opinions on A will be *vaguer* and hence also *more imprecise* (at least in a weak sense, if the first or last weak inequalities in (17) are equalities) than the a priori ones, *no matter* which $B \in \mathbb{P}$ is true. Even though dilation is $\mathbb{P}$-dependent (so that we may hope that a well-chosen partition $\mathbb{P}$ avoids dilation), it is a puzzling phenomenon. Clearly, dilation implies the weaker concept of imprecision increase, which captures one of the two basic features of dilation, the growth in the degree of imprecision.

To discuss the occurrence of dilation or of imprecision increase in the PMM, we assume that $\mathbb{P} = \{B, B^c\}$ and the conditional probabilities are the natural extensions. The formulas for $\overline{E}(A|B^c)$, $\underline{E}(A|B^c)$ are obtained from those for $\overline{E}(A|B)$, $\underline{E}(A|B)$ in Table 1 (when $\tau < P(B^c)$) replacing B with B^c.

We present now a number of results, whose operational relevance is discussed in Section 5.2.

Notation We write A' to denote, indifferently, either A or A^c. For instance, $\min\{P(A'\wedge B')\}$ is a short form for $\min\{P(A\wedge B), P(A^c\wedge B), P(A\wedge B^c), P(A^c\wedge B^c)\}$.

Proposition 11. *Each of the following conditions is necessary for dilation (of A, relative to $\{B, B^c\}$), whenever the denominator is positive:*

$$\tau < P(A \wedge B) \quad \Rightarrow \tau \geq \frac{P(A\wedge B)-P(A)P(B)}{P(A^c\wedge B^c)} \quad (19)$$

$$\tau < P(A^c \wedge B) \quad \Rightarrow \tau \geq \frac{P(A)P(B)-P(A\wedge B)}{P(A\wedge B^c)} \quad (20)$$

$$\tau < P(A \wedge B^c) \quad \Rightarrow \tau \geq \frac{P(A\wedge B^c)-P(A)P(B^c)}{P(A^c\wedge B)} \quad (21)$$

$$\tau < P(A^c \wedge B^c) \quad \Rightarrow \tau \geq \frac{P(A)P(B^c)-P(A\wedge B^c)}{P(A\wedge B)} \quad (22)$$

Proof. Impose either $\underline{E}(A|B') \leq \underline{P}(A)$ or $\overline{E}(A|B') \geq \overline{P}(A)$ in (17), and use Table 1 to choose the appropriate values of $\underline{E}$, $\overline{E}$, $\underline{P}$, $\overline{P}$.

To exemplify, Equation (19) implements the condition $\underline{E}(A|B) \leq \underline{P}(A)$, which is written as $\frac{P(A\wedge B)-\tau}{P(B)-\tau} \leq \frac{P(A)-\tau}{1-\tau}$. Multiply by $(P(B)-\tau)(1-\tau) > 0$ and solve the ensuing linear inequality in τ to get (19). □

Proposition 12. *Define* $m = \min\{P(A' \wedge B')\}$, $M = \max\{P(A' \wedge B')\}$, $M_\tau = \max\{(P(A \wedge B) - P(A)P(B))/P(A^c \wedge B^c), (P(A)P(B) - P(A \wedge B))/P(A \wedge B^c), (P(A \wedge B^c) - P(A)P(B^c))/P(A^c \wedge B), (P(A)P(B^c) - P(A \wedge B^c))/P(A \wedge B)\}$

(a) If $\tau < m$, dilation occurs if and only if $\tau \geq M_\tau$.

(b) The condition $\tau \geq M$ is sufficient for dilation.

Proof. (a): when $\tau < m = \min\{P(A' \wedge B')\}$, (17) holds iff τ satisfies the weak inequalities in (19÷22) i.e. iff $\tau \geq M_\tau$.

(b): when $\tau \geq M$, $\underline{E}(A|B') = 0$ and $\overline{E}(A|B') = 1$,[6] so dilation occurs no matter what are $\underline{P}(A)$, $\overline{P}(A)$. □

Remark 2. At most two of the four weak inequalities in (19÷22) need to be checked. In fact, let A and B be *positively correlated under* P, hence A and B^c are negatively correlated and $P(A)P(B) - P(A \wedge B) < 0$, $P(A \wedge B^c) - P(A)P(B^c) < 0$. Thus, (20) and (21) trivially hold ($\tau > 0$) and $M_\tau = \max\left\{\frac{P(A\wedge B)-P(A)P(B)}{P(A^c\wedge B^c)}, \frac{P(A)P(B^c)-P(A\wedge B^c)}{P(A\wedge B)}\right\}$. Similarly, (19) and (22) trivially hold when A and B are *negatively correlated* under P. □

Let us point out some special instances of dilation.

[6] This ensues from Table 1 when $\tau \leq \min P(B')$, if not use also Lemma 1.

Corollary 2. *Dilation occurs if:*

(a) $P(A' \wedge B') = P(A')P(B')$ *and* $\tau < m$.

(b) P *is* uniform *on* $\mathbb{P}_{A,B} = \{A \wedge B, A \wedge B^c, A^c \wedge B, A^c \wedge B^c\}$, $\forall \tau \in]0,1[$.

Proof. Condition (a) ensures dilation, as it implies $M_\tau = 0$ and hence (a) of Proposition 12. As for (b), it implies $P(A' \wedge B') = P(A')P(B')$ and $m = M = 0.25$: hence dilation occurs by (a) when $\tau < M = m$, by Proposition 12, (b) when $\tau \geq M$. □

Concerning imprecision increase, it holds that

Proposition 13. *Imprecision increases, i.e., Equation* (18) *holds, if the following system holds for* τ*:*

$$\begin{cases} (\tau - P(A \wedge B))(\tau - P(A^c \wedge B)) > 0 \\ (\tau - P(A \wedge B^c))(\tau - P(A^c \wedge B^c)) > 0 \end{cases} \tag{23}$$

Proof. Check that (18) holds, using Table 1. □

Remark 3. Note that (23) holds in particular when $\tau < m = \min\{P(A' \wedge B')\}$. Therefore imprecision *always increases* in this case. □

5.2 Imprecision variation in practice

As a general remark, the existence and relevance of dilation and imprecision increase in the PMM should be investigated distinguishing more cases, according to the relative ordering of $P(A' \wedge B')$, $P(A')$, and τ in $[0,1]$. However, the importance of each case varies greatly in the applications. We present in detail the most significant ones, while the remaining may be analyzed using Table 1 and the preceding results to check (17) and (18), as demonstrated in Example 3.

Case i) $\tau < m = \min\{P(A' \wedge B')\}$;

Case ii) $P(A) \leq \tau < \min\{P(A^c \wedge B')\}$.

Case i) is probably the most important: τ will often be rather low, recalling that it has the meaning of a commission or taxation (this happens for instance with Internet betting). In such circumstances case i) applies if none among $P(A' \wedge B')$ is too low.

Case i) is completely solved by the results in Section 5.1: dilation occurs iff $\tau \geq M_\tau$ (Proposition 12, (a)), imprecision always increases (Remark 3).

We do not necessarily meet case i) when A is a rare event, or $P(A)$ is anyway smaller than the commission τ in favour of House or of an insurer (these cases are relatively frequent in non-life insurance). If τ is also smaller than $\min\{P(A^c \wedge B')\}$, *case ii)* occurs. We discuss it in the next example.

Example 2. When $P(A) \leq \tau < \min\{P(A^c \wedge B')\}$, then (see Table 1) $\overline{P}(A) = P(A)/(1-\tau)$, $\overline{E}(A|B) = P(A \wedge B)/(P(B) - \tau)$, $\overline{E}(A|B^c) = P(A \wedge B^c)/(P(B^c) - \tau)$, $\underline{E}(A|B') = \underline{P}(A) = 0$. Imposing either (17) or (18) originates the same system of inequalities, i.e. *in this case* there is dilation iff there is imprecision increase. The system is

$$\begin{cases} \frac{P(A \wedge B)}{P(B) - \tau} \geq \frac{P(A)}{1 - \tau} \\ \frac{P(A \wedge B^c)}{P(B^c) - \tau} \geq \frac{P(A)}{1 - \tau} \end{cases} \tag{24}$$

and its inequalities are easily seen to be equivalent to (20) and (22). Thus dilation arises iff both (20) and (22) hold (and the lower bound they supply for τ is not greater than $\min\{P(A^c \wedge B')\}$). In practice, only one of them (at most) has to be checked, depending on the correlation of A and B, by Remark 2. For instance, if $P(A) = 0.02$, $P(A \wedge B) = 0.005$, $P(A \wedge B^c) = 0.015$, $P(B) = 0.4$, then $P(A|B) = 0.0125 < P(A)$ and (20) gives the bound $\tau \geq 0.2$. Since $\min\{P(A^c \wedge B')\} = 0.395 > 0.2$, the bound is effective: there is dilation (and imprecision increase) for $\tau \in [0.2; 0.395]$, none of them for $\tau \in [0.02; 0.2[$. □

Discussion We point out that dilation occurs in both case i) and ii) when A' and B' are judged *stochastically independent* or at least *not correlated* by P (as follows from Corollary 2 (a) and Example 2).

Further, dilation occurs when τ is too "large": Proposition 12 (b) ensures it when $\tau \geq M = \max\{P(A' \wedge B')\}$. This happens merely because $\overline{E}$, $\underline{E}$ are then vague, but dilation may occur also when $\tau < M$, as in the next example. □

Example 3. Assign P on $\mathbb{P}_{A,B}$ as follows: $P(A \wedge B) = \frac{1}{10}$, $P(A \wedge B^c) = P(A^c \wedge B^c) = \frac{2}{10}$,$P(A^c \wedge B) = \frac{1}{2}$. Consequently $P(A) = \frac{3}{10}$, $P(B) = \frac{6}{10}$, $P(A|B) = \frac{1}{6}$, $P(A|B^c) = \frac{1}{2}$.

Dilation occurs when $\tau \geq \frac{1}{2}$, by Proposition 12, (b). When $\tau < \frac{1}{2} = P(A^c \wedge B)$, use the necessary condition in (20), which requires that $\tau \geq \frac{4}{10}$, to rule out dilation for $\tau \in]0; \frac{4}{10}[$. If $\tau \in [\frac{4}{10}; \frac{1}{2}[$, (20) ensures that $\overline{E}(A|B) \geq \overline{P}(A)$, and the other inequalities in (17) hold too, because $\overline{E}(A|B^c) = 1$ and $\underline{E}(A|B') = 0$. Thus there is dilation for $\tau \in [\frac{4}{10}, \frac{1}{2}[$ too.

As for imprecision increase, it is ensured by Proposition 13 (Remark 3) when $\tau < \frac{1}{10}$. For $\tau \in [\frac{1}{10}; \frac{4}{10}[$, we have to check whether the inequalities (18) hold, distinguishing more subcases according to the different expressions for $\overline{P}$, $\underline{P}$, $\overline{E}(A|B')$, $\underline{E}(A|B')$. Conditioning on B^c, we should check whether

$$\overline{E}(A|B^c) - \underline{E}(A|B^c) \geq \overline{P}(A) - \underline{P}(A). \tag{25}$$

Now, $\overline{E}(A|B^c) - \underline{E}(A|B^c) = 1$ and (25) therefore holds if $\tau \in [\frac{2}{10}; \frac{4}{10}[$, while (25) specialises into $\frac{\tau}{P(B^c) - \tau} \geq$

$\frac{\tau}{1-\tau}$ when $\tau \in [\frac{1}{10}; \frac{2}{10}[$, and this inequality is true. Therefore (25) is verified for $\tau \in [\frac{1}{10}; \frac{4}{10}[$, and imprecision increase in this interval depends only on whether the inequality $\overline{E}(A|B) - \underline{E}(A|B) \geq \overline{P}(A) - \underline{P}(A)$ holds. Noting that $\overline{E}(A|B) - \underline{E}(A|B) = \frac{P(A \wedge B)}{P(B)-\tau} = \frac{1}{6-10\tau}$, $\forall \tau \in [\frac{1}{10}; \frac{4}{10}[$, we have to check whether:

$$\begin{array}{lll} \frac{1}{6-10\tau} \geq \frac{P(A)}{1-\tau} = \frac{3}{10(1-\tau)} & & \text{if } \tau \in [\frac{3}{10}; \frac{4}{10}[\\ \frac{1}{6-10\tau} \geq \frac{\tau}{1-\tau} & & \text{if } \tau \in [\frac{1}{10}; \frac{3}{10}[. \end{array}$$

The former inequality has no solution in $[\frac{3}{10}; \frac{4}{10}[$, the latter is true for $\tau \in [\frac{1}{10}; \frac{2}{10}]$. Conclusions: dilation occurs iff $\tau \in [\frac{4}{10}; 1[$, imprecision increase (without dilation) iff $\tau \in]0; \frac{2}{10}]$, neither of them iff $\tau \in]\frac{2}{10}; \frac{4}{10}[$. □

Limiting dilation or imprecision increase in the PMM is not straightforward. This may be achieved by an appropriate choice of τ in some, but not all cases (for instance, $\tau \in [\frac{2}{10}; \frac{4}{10}[$ might be too high a percentage in Example 3). More generally, choosing a coherent extension other than the natural extension often shrinks imprecision, by the dominance properties of the natural extension, but finding a computationally simple such extension may be not so easy in practice.

6 Conclusions

The pari-mutuel model represents a simple and natural way of eliciting upper/lower probabilities, and can be extended in more directions, thanks to the availability of standard procedures for 2-monotone and 2-alternating previsions. We computed explicitly its natural extension $\overline{E}$ starting from a PMM assignment on a lattice of events, generalizing the approach in [11], which is anyway discussed, focusing on comparing the different formulae available for $\overline{E}$. While a naive extension, considered in insurance premium pricing, does not seem to be a valuable alternative to the natural extension, being generally not coherent, the various formulae for the natural extension have a notable meaning in risk measurement. In fact, they correspond to known measures of risk or generalize them. We discussed also how to use the natural extension when conditioning, delimiting the influence of dilation and imprecision increase for the PMM.

A tempting new direction would, in a sense, merge our analysis in the conditional and unconditional framework, studying the natural extension to conditional gambles. Here a difficulty arises: available generalisations of equations (16), studied in [8], are lower/upper bounds for the natural extension and might not be reached, even when $\overline{P}$ is 2-alternating. In other words, the available procedures seem to give weaker results. This and the considerations at the end of Section 5.2 on how to limit dilation or imprecision increase might motivate investigating coherent extensions of the PMM alternative to the natural extension.

Acknowledgements

We wish to thank the referees for their helpful comments. Renato Pelessoni and Paolo Vicig acknowledge financial support from the PRIN Project 'Metodi di valutazione di portafogli assicurativi danni per il controllo della solvibilità', Marco Zaffalon from the Swiss NSF Grant n. 200020 − 116674/1.

References

[1] G. de Cooman, M. C. M. Troffaes and E. Miranda. *n*-Monotone lower previsions. *Journal of Intelligent & Fuzzy Systems*, 16: 253–263, 2005.

[2] B. de Finetti. *Theory of Probability*, volume 1, Wiley, 1974.

[3] D. Denneberg. *Non-Additive Measure and Integral*. Kluwer, 1994.

[4] M. Denuit, J. Dhaene, M. Goovaerts and R. Kaas. *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. Wiley, 1994.

[5] H.U. Gerber. *An Introduction to Mathematical Risk Theory*. Huebner Foundation, 1979.

[6] R. Pelessoni and P. Vicig. Imprecise Previsions for Risk Measurement. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 11: 393–412, 2003.

[7] R. Pelessoni and P. Vicig. Williams coherence and beyond. *International Journal of Approximate Reasoning*, 50(4):612–626, 2009.

[8] R. Pelessoni and P. Vicig. Bayes' theorem bounds for convex lower previsions. *Journal of Statistical Theory and Practice*, 3(1):85–101, 2009.

[9] T. Seidenfeld and L. Wasserman. Dilation for sets of probabilities. *The Annals of Statistics*, 21(3):1139–1154, 1993.

[10] P. Walley. Coherent lower (and upper) probabilities. *Research Report*, University of Warwick, Coventry, 1981.

[11] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

Interpretation and Computation of α-Junctions for Combining Belief Functions

Frédéric Pichon
Thales Research and Technology,
RD 128, F-91767 Palaiseau cedex, France.
Frederic.Pichon@thalesgroup.com

Thierry Denœux
UMR CNRS 6599 Heudiasyc,
Université de Technologie de Compiègne,
BP 20529, F-60205 Compiègne Cedex, France.
Thierry.Denoeux@utc.fr

Abstract

The α-junctions are the associative, commutative and linear operators for belief functions with a neutral element. This family of rules includes as particular cases the unnormalized Dempster's rule and the disjunctive rule. Until now, the α-junctions suffered from two main limitations. First, they did not have an interpretation in the general case. Second, it was difficult to compute a combination by an α-junction. In this paper, an interpretation for these rules is proposed. It is shown that the α-junctions correspond to a particular form of knowledge about the truthfulness of the sources providing the belief functions to be combined. Simple means to compute a combination by an α-junction are also laid bare. These means are based on generalizations of mechanisms that exist to compute the combination by the unnormalized Dempster's rule.

Keywords. Transferable Belief Model, Dempster-Shafer Theory, Belief Functions, Information Fusion, Uncertain reasoning.

1 Introduction

The Transferable Belief Model (TBM) [16, 12] is a model for quantifying beliefs using belief functions [8]. An essential part of the TBM is the aggregation of belief functions, which is done using so-called combination rules. To accodomate for various information fusion problems, many combination rules have been proposed (see, e.g., [15] for a recent survey) and, in particular, the unnormalized version of Dempster's rule [1], referred to as the conjunctive rule in this paper, the disjunctive rule [3, 9], the exclusive disjunctive rule and its negation [3, 11].

The use of the conjunctive rule is appropriate when one can assume that all sources providing the belief functions to be combined, tell the truth [11]. On the other hand, the disjunctive rule should be used when it is known that at least one of the sources tells the truth, but it is not known which one [11]. The uses of the exclusive disjunctive rule and its negation are also conditioned by knowledge on the truthfulness of the sources of information: the former fits with the case where exactly one of the sources is known to tell the truth, but it is not known which one, whereas the latter corresponds to a situation where either all or none of the sources are known to tell the truth [11]. Furthermore, all of these four rules assume that the sources are independent, meaning that those sources are assumed to provide distinct pieces of evidence.

In [11], Smets introduced an infinite family of combination rules, which he called α-junctions. This family basically represents the set of associative, commutative and linear operators for belief functions with a neutral element. It includes as special cases the four rules mentioned above. The behavior of an α-junction is determined by a parameter α and by the neutral element. The four special cases are recovered for particular values of α. For other values of this parameter, the α-junctions did not have an interpretation.

To our knowledge, this family of rules has never been exploited. This can be explained, at least in part, by the fact that these rules suffered from two main limitations until now. First, those operators did not have an interpretation in the general case. Second, it was difficult to compute a combination by an α-junction using the methods proposed in [11], as already remarked by Smets [13].

In this paper, this theoretical contribution of Smets is carefully reexamined: some new light on the meaning of the α-junctions is shed and their mathematics are simplified to make their computation easier. More precisely, it is first shown that these operators correspond to a particular form of knowledge, determined by the parameter α, on the truthfulness of the sources. The α-junctions become thus suitable as flexible combination rules that allow us to take into account some particular knowledge about the sources.

Several efficient and simple ways of computing a combination by an α-junction are then presented, making the practical use of the α-junctions in applications possible. These new means are based on generalizations of mechanisms that can be used to compute the combination by the conjunctive rule.

The rest of this paper is organized as follows. Necessary concepts of the TBM are first recalled in Section 2. In Section 3, basic notions on α-junctions are given. An interpretation for the α-junctions is proposed in Section 4. Several simple means to compute a combination by an α-junction are then unveiled in Section 5. Section 6 concludes the paper.

Note that due to lack of space, the proofs of the theorems and propositions presented in this paper, are not provided. They can be found in [7].

2 Fundamental Concepts of the TBM

2.1 Representation of Beliefs

In this paper, the TBM [16, 12] is accepted as a model to quantify uncertainties based on belief functions [8]. Let $\Omega = \{\omega_1, ..., \omega_K\}$ denote a finite set of possible values of a variable $\boldsymbol{\omega}$; Ω is called the frame of discernment of $\boldsymbol{\omega}$. In the TBM, the beliefs held by a rational agent Ag regarding the actual value ω_0 taken by $\boldsymbol{\omega}$ is represented by a basic belief assignment (BBA) m defined as a mapping from 2^Ω to $[0,1]$ verifying $\sum_{A\subseteq\Omega} m(A) = 1$. Subsets A of Ω such that $m(A) > 0$ are called focal sets of m. A BBA m is said to be: *vacuous* if Ω is the only focal set, this BBA is denoted by m_Ω; *categorical* if it has only one focal set; *simple* if if has at most two focal sets and, if it has two, Ω is one of those. A simple BBA (SBBA) m such that $m(A) = 1-\alpha$ for some $A \neq \Omega$ and $m(\Omega) = \alpha$, can be written A^α. This notation for SBBAs is useful in this paper to shorten some expressions.

A BBA m can equivalently be represented by its associated belief, plausibility and commonality functions defined, respectively, as:

$$bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B),$$

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B),$$

and

$$q(A) = \sum_{B \supseteq A} m(B), \tag{1}$$

for all $A \subseteq \Omega$. The BBA m can be recovered from any of these functions. For instance:

$$m(A) = \sum_{B\supseteq A} (-1)^{|B|-|A|} q(B), \quad \forall A \subseteq \Omega,$$

where $|A|$ denotes the cardinality of A.

The negation (or complement) $\overline{m}$ of a BBA m is defined as the BBA verifying $\overline{m}(A) = m(\overline{A})$, $\forall A \subseteq \Omega$, where $\overline{A}$ denotes the complement of A [3]. $\overline{m}$ represents the BBA that would be induced if the agent knows that the source providing a BBA m is not telling the truth, i.e., is lying [11].

Another important concept of the TBM is the least commitment principle (LCP) [9]. This principle postulates that, given a set of BBAs compatible with a set of constraints, the most appropriate BBA is the least informative. The LCP becomes operational through the definition of partial orderings allowing the informational comparison of BBAs. Such orderings were proposed in [17] and [3]. For instance, the q-ordering is defined as follows: a BBA m_1 is said to be at least as q-committed, or at least as q-informed, than a BBA m_2 iff we have $q_1(A) \leq q_2(A)$, for all $A \subseteq \Omega$.

2.2 Combination of Beliefs

The beliefs represented by BBAs can be aggregated using appropriate operators, called combination rules. In this section, the definitions of some of these combination rules are provided. Some notions related to these rules, which will be generalized in later parts of this paper, are also given.

The conjunctive rule is denoted by $\textcircled{\cap}$. It is defined as follows. Let m_1 and m_2 be two BBAs, and let $m_{1\textcircled{\cap}2}$ be the result of their combination by $\textcircled{\cap}$. We have, for all $A \subseteq \Omega$:

$$m_{1\textcircled{\cap}2}(A) = \sum_{B\cap C = A} m_1(B)\, m_2(C). \tag{2}$$

This rule is appropriate when the sources that have induced m_1 and m_2, are known to tell the truth and to be independent. Furthermore, this rule is commutative, associative and admits a unique neutral element: the vacuous BBA m_Ω. Of interest is that this rule has a simple expression in terms of commonality functions. We have:

$$q_{1\textcircled{\cap}2}(A) = q_1(A) \cdot q_2(A), \quad \forall A \subseteq \Omega.$$

In the TBM, conditioning by $B \subseteq \Omega$ is equivalent to conjunctive combination with a categorical BBA m_B focused on B, i.e., $m_B(B) = 1$. The result is denoted by $m[B]$, with $m[B] = m\textcircled{\cap}m_B$. The conditional BBA $m[B]$ quantifies our belief on Ω, in a context where B holds. This operation is called the unnormalized Dempster's rule of conditioning. The combination by the conjunctive rule $\textcircled{\cap}$ admits a simple expression using the unnormalized Dempster's rule of conditioning. Indeed, let m_1 and m_2 be two BBAs. We have, for all

$A \subseteq \Omega$

$$m_{1 \textcircled{\cap} 2}(A) = \sum_{B \subseteq \Omega} m_1[B](A)\, m_2(B). \quad (3)$$

When it cannot be assumed that all the sources tell the truth, it may be assumed that at least one of them tells the truth, without knowing which one. In such a situation, and provided that the sources are independent, the disjunctive rule [3, 9] is appropriate. The disjunctive rule is denoted by $\textcircled{\cup}$. Let m_1 and m_2 be two distinct BBAs, and let $m_{1\textcircled{\cup}2}$ be the result of their combination by $\textcircled{\cup}$. We have:

$$m_{1 \textcircled{\cup} 2}(A) = \sum_{B \cup C = A} m_1(B)\, m_2(C), \quad \forall A \subseteq \Omega.$$

The disjunctive rule is commutative, associative and admits a unique neutral element: the BBA which assigns the total mass of belief to the empty set, i.e., $m(\emptyset) = 1$. This BBA, which we denote by $m_\emptyset$, is the negation of the neutral BBA m_Ω of the conjunctive rule and is sometimes called the or-vacuous BBA [11]. The dual nature of $\textcircled{\cap}$ and $\textcircled{\cup}$ becomes apparent when one notices that these operators are linked by De Morgan's laws [3]:

$$\begin{aligned} \overline{m_1 \textcircled{\cup} m_2} &= \overline{m_1} \textcircled{\cap} \overline{m_2} \\ \overline{m_1 \textcircled{\cap} m_2} &= \overline{m_1} \textcircled{\cup} \overline{m_2}. \end{aligned}$$

Of interest for this paper are two other rules: the exclusive disjunctive rule denoted by $\textcircled{\underline{\cup}}$ and its negation denoted by $\textcircled{\underline{\cap}}$ [11], which are defined as follows. We have, for all $A \subseteq \Omega$:

$$m_{1 \textcircled{\underline{\cup}} 2}(A) = \sum_{A = B \underline{\cup} C} m_1(B)\, m_2(C),$$

where $\underline{\cup}$ is the exclusive OR (XOR), i.e., $B \underline{\cup} C = \left(B \cap \overline{C}\right) \cup \left(\overline{B} \cap C\right)$ for all $B, C \subseteq \Omega$, and

$$m_{1 \textcircled{\underline{\cap}} 2}(A) = \sum_{A = B \underline{\cap} C} m_1(B)\, m_2(C),$$

where $\underline{\cap}$ denotes logical equality, i.e., $B \underline{\cap} C = (B \cap C) \cup (\overline{B} \cap \overline{C})$ for all $B, C \subseteq \Omega$.

The rules $\textcircled{\cup}$ and $\textcircled{\cap}$ are commutative, associative and admit a unique neutral element: $m_\emptyset$ and m_Ω, respectively. Furthermore, they are linked by De Morgan's laws. The rule $\textcircled{\cup}$ corresponds to the situation where it is known that exactly one of the sources of information tells the truth, but it is not known which one [11]. The rule $\textcircled{\cap}$ corresponds to the situation where it is known that either all or none of the sources of information tell the truth [11].

2.3 Operations on Product Spaces

In Section 4 of this paper, some operations that allow the manipulation of BBAs defined on product spaces, are needed. They are succinctly presented here. Let $m^{\Omega \times \Theta}$ denote a BBA defined on the Cartesian product $\Omega \times \Theta$ of the frames of two variables $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$. The marginal BBA $m^{\Omega \times \Theta \downarrow \Omega}$ is defined, for all $A \subseteq \Omega$, as

$$m^{\Omega \times \Theta \downarrow \Omega}(A) = \sum_{\{B \subseteq \Omega \times \Theta, (B \downarrow \Omega) = A\}} m^{\Omega \times \Theta}(B),$$

where $(B \downarrow \Omega)$ denotes the projection of B onto Ω, defined as

$$(B \downarrow \Omega) = \{\omega \in \Omega | \exists \theta \in \Theta, (\omega, \theta) \in B\}.$$

Conversely, let m^Ω be a BBA defined on Ω. Its vacuous extension on $\Omega \times \Theta$ is defined as:

$$m^{\Omega \uparrow \Omega \times \Theta}(B) = \begin{cases} m^\Omega(A) & \text{if } B = A \times \Theta, \\ & \text{for some } A \subseteq \Omega, \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

Given two BBAs m_1^Ω and m_2^Θ, their conjunctive combination on $\Omega \times \Theta$ can be obtained by combining their vacuous extensions on $\Omega \times \Theta$ using (4). Formally:

$$m_1^\Omega \textcircled{\cap} m_2^\Theta = m_1^{\Omega \uparrow \Omega \times \Theta} \textcircled{\cap} m_2^{\Theta \uparrow \Omega \times \Theta}.$$

Two other operations that have been defined for BBAs on product spaces are the *conditioning* operation, and its inverse operation called the *ballooning extension*. They are defined as follows. Let $m^{\Omega \times \Theta}$ denote a BBA on $\Omega \times \Theta$, and $m_B^{\Omega \times \Theta}$ the BBA on $\Omega \times \Theta$ with single focal set $\Omega \times B$ with $B \subseteq \Theta$, i.e., $m_B^{\Omega \times \Theta}(\Omega \times B) = 1$. The conditional BBA on Ω given $\boldsymbol{\theta} \in B$ is defined as:

$$m^\Omega[B] = \left(m^{\Omega \times \Theta} \textcircled{\cap} m_B^{\Omega \times \Theta}\right)^{\downarrow \Omega}.$$

Now, let $m^\Omega[B]$ denote the conditional BBA on Ω, given $\boldsymbol{\theta} \in B \subseteq \Theta$. The ballooning extension of $m^\Omega[B]$ on $\Omega \times \Theta$ is the least committed BBA, whose conditioning on B yields $m^\Omega[B]$ [9]. It is obtained as:

$$m^\Omega[B]^{\Uparrow \Omega \times \Theta}(C) = m^\Omega[B](A),$$

if $C = (A \times B) \cup (\Omega \times (\Theta \backslash B))$, for some $A \subseteq \Omega$, and $m^\Omega[B]^{\Uparrow \Omega \times \Theta}(C) = 0$ otherwise. Example 1 illustrates the ballooning extension.

Example 1. *Consider two frames $\Omega = \{\omega_1, \omega_2\}$ and $\Theta = \{\theta_1, \theta_2\}$. Further, let $m^\Omega[\theta_2]$ be a conditional BBA defined by $m^\Omega[\theta_2](\{\omega_1\}) = 0.6$ and $m^\Omega[\theta_2](\Omega) = 0.4$. The ballooning extension of $m^\Omega[\theta_2]$ is:*

$$\begin{aligned} m^\Omega[\theta_2]^{\Uparrow \Omega \times \Theta}(\{(\omega_1, \theta_2), (\omega_1, \theta_1), (\omega_2, \theta_1)\}) &= 0.6, \\ m^\Omega[\theta_2]^{\Uparrow \Omega \times \Theta}(\Omega \times \Theta) &= 0.4. \end{aligned}$$

2.4 Matrix Notation

The matrix notation can be used to greatly simplify the mathematics of belief function theory. In [13], Smets proposed a review of the application of the matrix calculus to belief functions. This section is devoted to a summary of parts of [13] that are relevant to this paper.

Belief functions as column vectors

A BBA m (and its associated functions *bel*, *pl* and q) defined on 2^Ω can be seen as a column vector of size $2^{|\Omega|}$. The elements of m can be ordered arbitrarily but the so-called binary order is particularly convenient. The binary order means that the first element of m is related to the empty set, the next to $\{a\}$, the next to $\{b\}$, the next to $\{a,b\}$, etc. More generally, the ith element of the vector $\mathbf{m}$ corresponds to the set with elements indicated by 1 in the binary representation of $i-1$. For instance, let $\Omega=\{a,b,c,d\}$. The first element ($i=1$) of the vector $\mathbf{m}$ corresponds to the emptyset since the binary representation of $1-1$ is 0000. The twelfth element ($i=12$) corresponds to $\{a,b,d\}$ since the binary representation of $12-1$ is 1011.

We use the following conventions. By default, the length of vectors and matrices are $2^{|\Omega|}$, and vectors are column vectors. Matrices and vectors are written in bold type, and their elements in normal type, e.g., a matrix is noted $\mathbf{M}$ and the element on its ith row and jth column is noted $M(i,j)$. Sometimes a matrix will be defined by its general term, in this case we write $\mathbf{M}=[M(i,j)]$. For instance, if $M(i,j)$ is defined by $M(i,j)=0,\forall i,j$, then $\mathbf{M}$ is a matrix, whose elements are zeros. Finally, $\mathbf{I}$ denotes the unit matrix and $\mathbf{Kron}(\mathbf{A},\mathbf{B})$ denotes the $mp\times nq$ matrix resulting from the Kronecker product of a $m\times n$ matrix $\mathbf{A}$ with a $p\times q$ matrix $\mathbf{B}$. The matrix $\mathbf{Kron}(\mathbf{A},\mathbf{B})$ is defined by:

$$\mathbf{Kron}(\mathbf{A},\mathbf{B})=\begin{bmatrix} A(1,1)\mathbf{B} & \cdots & A(1,n)\mathbf{B} \\ \vdots & \ddots & \vdots \\ A(m,1)\mathbf{B} & \cdots & A(m,n)\mathbf{B} \end{bmatrix}.$$

The transformation (1) of a BBA m into its associated commonality function q can be represented using the matrix notation. We have

$$q(A) \;=\; \sum_{B\subseteq\Omega} Q(A,B)m(B),$$

where $Q(A,B)=1$ iff $B\supseteq A$ and 0 otherwise. Letting $\mathbf{Q}=[Q(A,B)]$, $A,B\subseteq\Omega$, we have $\mathbf{q}=\mathbf{Q}\cdot\mathbf{m}$ and $\mathbf{m}=\mathbf{Q}^{-1}\cdot\mathbf{q}$ [13]. The matrix $\mathbf{Q}$ may be obtained in a very simple way using Kronecker multiplication. Indeed, we have:

$$\mathbf{Q}^{i+1} \;=\; \mathbf{Kron}\left(\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix},\quad \mathbf{Q}^i\right),\mathbf{Q}^1=1,$$

where $\mathbf{Q}^i$ denotes the matrix $\mathbf{Q}$ when $|\Omega|=i$.

Transformations of BBA into BBA

In this paragraph, we present how the transformation of a BBA into another BBA, given a piece of evidence, can be expressed using the matrix notation.

Definition 1. *A stochastic matrix* $\mathbf{M}=[M(i,j)]$ *is a square matrix with* $M(i,j)\geq 0$ *and* $\sum_i M(i,j)=1,\forall j$.

Let $\mathcal{M}^\Omega$ be the set of BBAs defined on Ω. As shown by [13, Theorem 6.1], the set of matrices that map every element of $\mathcal{M}^\Omega$ into an element of $\mathcal{M}^\Omega$ is the set of stochastic matrices.

The *revision* of a BBA m_1 by a piece of evidence Ev can be represented by a stochastic matrix $\mathbf{M}_{Ev,m_1}$ that transforms m_1 into $m_1[Ev]$:

$$\mathbf{m}_1[Ev]=\mathbf{M}_{Ev,m_1}\cdot\mathbf{m}_1.$$

If the value of the matrix depends only on Ev and not on m_1 (in which case the pieces of evidence that induced m_1 and Ev are said 'distinct' [13]), we can write:

$$\mathbf{m}_1[Ev]=\mathbf{M}_{Ev}\cdot\mathbf{m}_1.$$

The combinations by the rules ⓝ, ⓤ, ⓤ and ⓝ are particular cases of revision. For instance, the conjunctive revision of a BBA m_1 by a distinct piece of evidence inducing a BBA m_2 is achieved by a special kind of matrix, called a Dempsterian specialization matrix [5] and denoted by $\mathbf{S}_{m_2}$. This matrix is defined as a function of m_2: its general term is $S_{m_2}(A,B)=m_2[B](A)$, $A,B\subseteq\Omega$. We have $\mathbf{m}_2ⓝ\mathbf{m}_1=\mathbf{S}_{m_2}\cdot\mathbf{m}_1$.

3 α-Junctions: Basic Notions

In [11], Smets studies the set of possible associative, commutative and linear combination rules with a neutral element. Smets calls this set the α-junctions because they cover the conjunction, the disjunction and the exclusive disjunction. We report in this section the summary of [11] given in [13].

Let $\mathbf{m}_1$ and $\mathbf{m}_2$ be two BBAs on Ω. Suppose we want to build a BBA $\mathbf{m}_{12}$ such that $\mathbf{m}_{12}=f(\mathbf{m}_1,\mathbf{m}_2)$, i.e., $\mathbf{m}_{12}$ depends only on $\mathbf{m}_1$ and $\mathbf{m}_2$. Smets [11] determines the operators that map $\mathcal{M}^\Omega\times\mathcal{M}^\Omega$ to $\mathcal{M}^\Omega$ and that satisfy the following requirements (the origins of those requirements are summarized in [13, p.25]).

- Linearity[1]: $f(\mathbf{m}, p\mathbf{m}_1 + q\mathbf{m}_2) = pf(\mathbf{m}, \mathbf{m}_1) + qf(\mathbf{m}, \mathbf{m}_2)$, $p \in [0,1]$, $q = 1-p$.
- Commutativity: $f(\mathbf{m}_1, \mathbf{m}_2) = f(\mathbf{m}_2, \mathbf{m}_1)$.
- Associativity: $f(f(\mathbf{m}_1, \mathbf{m}_2), \mathbf{m}_3) = f(\mathbf{m}_1, f(\mathbf{m}_2, \mathbf{m}_3))$.
- Neutral element: existence of a belief function $\mathbf{m}_{vac}$ such that $f(\mathbf{m}, \mathbf{m}_{vac}) = \mathbf{m}$ for any $\mathbf{m}$.
- Anonymity: relabeling the elements of Ω does not affect the results.
- Context preservation: if $pl_1(X) = 0$ and $pl_2(X) = 0$ for some $X \subseteq \Omega$, then $pl_{12}(X) = 0$.

It is shown in [11] that the solutions are stochastic matrices. We have:

$$\mathbf{m}_{12} = \mathbf{K}_{m_1} \cdot \mathbf{m}_2, \tag{5}$$

where

$$\mathbf{K}_{m_1} = \sum_{X \subseteq \Omega} m_1(X) \cdot \mathbf{K}_X. \tag{6}$$

Smets [11] proves that the $2^{|\Omega|} \times 2^{|\Omega|}$ matrices $\mathbf{K}_X$ depend only on $\mathbf{m}_{vac}$ and one parameter $\alpha \in [0,1]$. Furthermore, he shows that there are only two solutions for $\mathbf{m}_{vac}$: either $\mathbf{m}_{vac} = \mathbf{m}_\Omega$ or $\mathbf{m}_{vac} = \mathbf{m}_\emptyset$. Hence, there are only two sets of solutions, which are presented now.

3.1 Case $\mathbf{m}_{vac} = \mathbf{m}_\Omega$

The definition of the matrices $\mathbf{K}_X$ that satisfy the above requirements when $\mathbf{m}_{vac} = \mathbf{m}_\Omega$ is the following.

$$\begin{aligned} \mathbf{K}_\Omega &= \mathbf{I}, \\ \mathbf{K}_X &= \prod_{x \notin X} \mathbf{K}_{\{\overline{x}\}}, \quad \forall X \subset \Omega, \end{aligned}$$

where

$$\mathbf{K}_{\{\overline{x}\}} = \left[k_{\overline{x}}(A,B)\right], \quad \forall x \in \Omega,$$

with

$$k_{\overline{x}}(A,B) = \begin{cases} 1 & \text{if } x \notin A, \quad B = A \cup \{x\}, \\ \alpha & \text{if } x \notin B, \quad B = A, \\ 1-\alpha & \text{if } x \notin B, \quad A = B \cup \{x\}, \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha \in [0,1]$ and is constant for all $\mathbf{K}_X$. Example 2 illustrates the various matrices $\mathbf{K}_X$ when $\Omega = \{a, b\}$ and $\mathbf{m}_{vac} = \mathbf{m}_\Omega$.

[1] Smets referred to this property as "linearity". However, it is not real linearity, as it is only valid for convex combinations. We have kept the same terminology for lack of a more appropriate term.

Example 2. *From (5) and (6), we have (in the matrices below, dots replace zeros and $\overline{\alpha} = 1-\alpha$)*

$$\begin{aligned} \mathbf{m}_{12} &= (m_1(\emptyset) \cdot \mathbf{K}_\emptyset + m_1(a) \cdot \mathbf{K}_a \\ &\quad + m_1(b) \cdot \mathbf{K}_b + m_1(\Omega) \cdot \mathbf{K}_\Omega) \cdot \mathbf{m}_2 \\ &= (m_1(\emptyset) \cdot \mathbf{K}_{\{\overline{a}\}} \cdot \mathbf{K}_{\{\overline{b}\}} + m_1(a) \cdot \mathbf{K}_{\{\overline{b}\}} \\ &\quad + m_1(b) \cdot \mathbf{K}_{\{\overline{a}\}} + m_1(\Omega) \cdot \mathbf{I}) \cdot \mathbf{m}_2 \\ &= (m_1(\emptyset) \cdot \begin{bmatrix} \alpha & 1 & . & . \\ \overline{\alpha} & . & . & . \\ . & . & \alpha & 1 \\ . & . & \overline{\alpha} & . \end{bmatrix} \cdot \begin{bmatrix} \alpha & . & 1 & . \\ . & \alpha & . & 1 \\ \overline{\alpha} & . & . & . \\ . & \overline{\alpha} & . & . \end{bmatrix} \\ &\quad + m_1(a) \cdot \begin{bmatrix} \alpha & . & 1 & . \\ . & \alpha & . & 1 \\ \overline{\alpha} & . & . & . \\ . & \overline{\alpha} & . & . \end{bmatrix} \\ &\quad + m_1(b) \cdot \begin{bmatrix} \alpha & 1 & . & . \\ \overline{\alpha} & . & . & . \\ . & . & \alpha & 1 \\ . & . & \overline{\alpha} & . \end{bmatrix} \\ &\quad + m_1(\Omega) \cdot \mathbf{I}) \cdot \mathbf{m}_2. \end{aligned}$$

When $\mathbf{m}_{vac} = \mathbf{m}_\Omega$ and $\alpha = 1$, the matrix $\mathbf{K}_{m_1}$ computed using (6) becomes the Dempsterian specialization matrix and we have $\mathbf{K}_{m_1} \cdot \mathbf{m}_2 = \mathbf{m}_{1 \textcircled{\cap} 2}$ [13]. The case $\alpha = 0$ corresponds to the rule $\textcircled{\cap}$. When $\mathbf{m}_{vac} = \mathbf{m}_\Omega$, an α-junction is referred to as an α-conjunction by Smets since $\mathbf{m}_\Omega$ is the neutral element of the conjunction [11]. The result of the α-conjunction of two BBAs m_1 and m_2 is written $m_1 \textcircled{\cap}^\alpha m_2$. Let us remark that despite what the appelation "α-conjunction" might lead one to think, an α-conjunction do not necessarily exhibit a conjunctive behavior. For instance, consider a frame $\Omega = \{\omega_1, \omega_2\}$ and two precise BBAs m_1 and m_2 such that $m_1(\{\omega_1\}) = m_2(\{\omega_1\}) = 1$. We have $m_{1 \textcircled{\cap} 2}(\Omega) = 1$, which is the most imprecise BBA.

3.2 Case $\mathbf{m}_{vac} = \mathbf{m}_\emptyset$

The definition of the matrices $\mathbf{K}_X$ that satisfy the above requirements when $\mathbf{m}_{vac} = \mathbf{m}_\emptyset$ is the following.

$$\begin{aligned} \mathbf{K}_\emptyset &= \mathbf{I}, \\ \mathbf{K}_X &= \prod_{x \in X} \mathbf{K}_{\{x\}}, \quad \forall X \in 2^\Omega \backslash \{\emptyset\}, \end{aligned}$$

where

$$\mathbf{K}_{\{x\}} = \left[k_x(A,B)\right], \quad \forall x \in \Omega,$$

with

$$k_x(A,B) = \begin{cases} 1 & \text{if } x \notin B, \quad A = B \cup \{x\}, \\ \alpha & \text{if } x \in B, \quad B = A, \\ 1-\alpha & \text{if } x \notin A, \quad B = A \cup \{x\}, \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha \in [0,1]$ and is constant for all $\mathbf{K}_X$.

When $\mathbf{m}_{vac} = \mathbf{m}_\emptyset$, an α-junction is referred to as an α-disjunction since $\mathbf{m}_\emptyset$ is the neutral element of the disjunction [11]; we denote an α-disjunctive rule by $\textcircled{\cup}^\alpha$. Furthermore, when $\mathbf{m}_{vac} = \mathbf{m}_\emptyset$ and $\alpha = 1$, we have $\mathbf{K}_{m_1} \cdot \mathbf{m}_2 = \mathbf{m}_{1\textcircled{\cup}2}$. The case $\alpha = 0$ corresponds to the rule $\textcircled{\cup}$.

Finally, we have, for any $\alpha \in [0,1]$ [13, Theorem 12.2]:

$$\begin{aligned} \overline{m_1 \textcircled{\cup}^\alpha m_2} &= \overline{m_1} \textcircled{\cap}^\alpha \overline{m_2}, \\ \overline{m_1 \textcircled{\cap}^\alpha m_2} &= \overline{m_1} \textcircled{\cup}^\alpha \overline{m_2}, \end{aligned} \tag{7}$$

i.e., α-conjunctive rules and α-disjunctive rules are linked by De Morgan laws. In particular, the De Morgan duality between the conjunctive and disjunctive rules is recovered by setting $\alpha = 1$ in (7).

4 Interpretation

In this section, an interpretation for the α-junctions is proposed. This interpretation relies on the concept of the truthfulness of the sources of information.

4.1 Truthfulness of the Sources

Let $\boldsymbol{\omega}$ be a variable, which takes its values in a frame Ω. Suppose an agent who does not know anything about the actual value ω_0 taken by $\boldsymbol{\omega}$. Suppose a source S_1 that tells the agent that the actual value ω_0 is in $A \subseteq \Omega$, i.e., $\omega_0 \in A$. If the source tells the truth or, equivalently, is truthful, then the agent believes $\omega_0 \in A$. If the source does not tell the truth, then the agent believes $\omega_0 \in \overline{A}$.

Let $\boldsymbol{\tau}$ be a variable taking its values in a frame $T = \{t, f\}$. We use $\boldsymbol{\tau}$ to denote the truthfulness of the source. The information $\omega_0 \in A$ provided by S_1 can be modeled by a BBA m_1^Ω such that $m_1^\Omega(A) = 1$. The information *when the source tells the truth, ω_0 must be in A, and when the source does not tell the truth, ω_0 must be in $\overline{A}$*, may be modeled by a BBA noted $m_{1'}^{\Omega\times T}$ and defined on the product space $\Omega \times T$ by

$$m_{1'}^{\Omega\times T}(A \times \{t\} \cup \overline{A} \times \{f\}) = 1. \tag{8}$$

Note that we use the index $1'$ in $m_{1'}^{\Omega\times T}$, i.e., the source number followed by the prime symbol, to highlight that the BBA $m_{1'}^{\Omega\times T}$ is obtained from the source S_1, as is the case of the BBA m_1^Ω, but that it conveys a different information from the BBA m_1^Ω.

One verifies that the BBA $m_{1'}^{\Omega\times T}$ is appropriate to model the information available in this scenario since

- combining $m_{1'}^{\Omega\times T}$ with a BBA m_t^T defined on T by $m_t^T(t) = 1$, and then marginalizing on Ω, i.e., performing
$$(m_{1'}^{\Omega\times T} \textcircled{\cap} m_t^T)^{\downarrow\Omega}, \tag{9}$$
yields a BBA m_{Ag}^Ω such that $m_{Ag}^\Omega = m_1^\Omega$, which means that the agent's beliefs are equated to the source's beliefs if the agent believes that the source tells the truth;
- combining $m_{1'}^{\Omega\times T}$ with a BBA m_f^T defined on T by $m_f^T(f) = 1$, and then marginalizing on Ω, i.e., performing
$$(m_{1'}^{\Omega\times T} \textcircled{\cap} m_f^T)^{\downarrow\Omega}, \tag{10}$$
yields a BBA m_{Ag}^Ω such that $m_{Ag}^\Omega = \overline{m_1^\Omega}$, which is sound since $\overline{m}$ represents the BBA that would be induced if the agent knows that a source providing a BBA m is not telling the truth [11], as mentioned in Section 2.1.

This reasoning may be generalized when the source produces an information in the form of a BBA rather than a set, in which case the BBA $m_{1'}^{\Omega\times T}$ is such that

$$m_{1'}^{\Omega\times T}(A \times \{t\} \cup \overline{A} \times \{f\}) = m_1^\Omega(A), \quad \forall A \subseteq \Omega. \tag{11}$$

Here again, if we perform (9) and (10), we find $m_{Ag}^\Omega = m_1^\Omega$ and $m_{Ag}^\Omega = \overline{m_1^\Omega}$, respectively, which means that, as expected, the agent's beliefs are equated to what the source says if the source tells the truth, and the agent's beliefs are equal to the negation of what the source says if the source does not tell the truth.

Using the BBA $m_{1'}^{\Omega\times T}$, as defined by (11), to represent the agent's beliefs when it receives a BBA m_1^Ω from a source S_1, we may now derive an interpretation for the α-junctions.

4.2 Interpretation of the α-Conjunctions

Suppose two distinct sources S_1 and S_2 that induce two BBAs m_1^Ω and m_2^Ω on Ω. Let $T_1 = \{t_1, f_1\}$ and $T_2 = \{t_2, f_2\}$; these two frames will be used to model beliefs on the truthfulness of S_1 and S_2, respectively. Suppose we want to quantify the agent's beliefs on Ω given m_1^Ω, m_2^Ω and the following distinct pieces of evidence.

- A piece of evidence stating that both or none of the sources tell the truth. This piece of evidence may be modeled by a BBA $m_{xand}^{T_1\times T_2}$ defined by
$$m_{xand}^{T_1\times T_2}(\{(t_1,t_2),(f_1,f_2)\}) = 1.$$
- Distinct items of evidence for all $x \in \Omega$ of the form
$$pl^{T_1\times T_2}[x](\{(f_1,f_2)\}) = 1 - \alpha, \tag{12}$$

indicating that if $\omega_0 = x$, then it is plausible with strength $1-\alpha$ that none of the sources tell the truth.

To compute the agent's beliefs on Ω given these distinct pieces of evidence, the items of evidence of the form given by (12), must be transformed into BBAs. In the TBM, this may be done using the LCP. The least committed BBA $m^{T_1\times T_2}[x]$ corresponding to (12) is the SBBA $m^{T_1\times T_2}[x] = \{(t_1,t_2),(f_1,t_2),(t_1,f_2)\}^{1-\alpha}$. Using all these distinct items of evidence, the agent's belief m^{Ω}_{Ag} on Ω is then equal to

$$\begin{aligned} m^{\Omega}_{Ag} &= (m^{\Omega\times T_1}_{1'} \textcircled{\cap} m^{\Omega\times T_2}_{2'} \textcircled{\cap} m^{T_1\times T_2}_{xand} \\ &\quad \textcircled{\cap}(\textcircled{\cap}_{x\in\Omega} m^{T_1\times T_2}[x]^{\Uparrow\Omega\times T_1\times T_2}))^{\downarrow\Omega}, \quad (13) \end{aligned}$$

with, for $i=1$ and $i=2$ and all $A\subseteq\Omega$

$$m^{\Omega\times T_i}_{i'}(A\times\{t_i\}\cup\overline{A}\times\{f_i\}) = m^{\Omega}_i(A), \quad (14)$$

and, for all $x\in\Omega$

$$m^{T_1\times T_2}[x] = \{(t_1,t_2),(f_1,t_2),(t_1,f_2)\}^{1-\alpha},$$

and

$$m^{T_1\times T_2}_{xand}(\{(t_1,t_2),(f_1,f_2)\}) = 1.$$

Theorem 1. *Let m^{Ω}_1 and m^{Ω}_2 be two BBAs. The BBA m^{Ω}_{Ag} defined by (13) verifies*

$$m^{\Omega}_{Ag} = m^{\Omega}_1 \textcircled{\cap}^{\alpha} m^{\Omega}_2.$$

This theorem may be illustrated with a simple valuation network [6] (see Figure 1), which is a graphical display of a set of BBAs, where variables are represented by circular nodes and BBAs are represented by square nodes.

As shown by Theorem 1, an α-conjunction is equivalent to the pooling by the conjunctive rule of some simple pieces of evidence, which can all be interpreted and that are, moreover, all related to the truthfulness of the sources. In particular, the parameter α involved in the α-conjunctions can be interpreted in terms of the plausibility, given $\omega_0 = x$, that the sources lie, since this plausibility is equal to $1-\alpha$. Note that since the BBA m_{xand} excludes the fact that one and only one source tells the truth, we clearly see, from the interpretation given to α, that we pass from the conjunctive rule to the rule $\textcircled{\cap}$ as α varies from 1 to 0. Finally, we may remark that, since (12) is logically equivalent to

$$bel^{T_1\times T_2}[x](\{(t_1,t_2),(f_1,t_2),(t_1,f_2)\}) = \alpha,$$

then the parameter α involved in the α-conjunctions is equal to the belief, given $\omega_0 = x$, that at least one of the sources tells the truth.

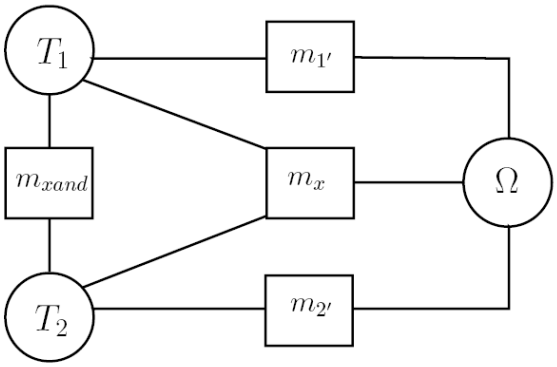


Figure 1: Valuation network for the α-conjunction of two BBAs m_1 and m_2. In the network, the term $(\textcircled{\cap}_{x\in\Omega} m^{T_1\times T_2}[x]^{\Uparrow\Omega\times T_1\times T_2})$ appearing in (13), is replaced by a BBA m_x defined on $\Omega\times T_1\times T_2$.

Let us eventually remark that Theorem 1 does not extend to more than two sources. Indeed, let m_1, m_2 and m_3 be three BBAs. The combination $m^{\Omega}_1 \textcircled{\cap}^{\alpha} m^{\Omega}_2 \textcircled{\cap}^{\alpha} m^{\Omega}_3$ is in general not equal to

$$\begin{aligned} &(m^{\Omega\times T_1}_{1'} \textcircled{\cap} m^{\Omega\times T_2}_{2'} \textcircled{\cap} m^{\Omega\times T_3}_{3'} \textcircled{\cap} m^{T_1\times T_2\times T_3}_{xand} \\ &\textcircled{\cap}(\textcircled{\cap}_{x\in\Omega} m^{T_1\times T_2\times T_3}[x]^{\Uparrow\Omega\times T_1\times T_2\times T_3}))^{\downarrow\Omega}, \end{aligned}$$

with $m^{\Omega\times T_i}_{i'}$, $i=1,2,3$, defined by (14), and where $m^{T_1\times T_2\times T_3}[x]$ is the least committed BBA corresponding to $pl^{T_1\times T_2\times T_3}[x](\{(f_1,f_2,f_3)\}) = 1-\alpha$, and with $m^{T_1\times T_2\times T_3}_{xand}(\{(t_1,t_2,t_3),(f_1,f_2,f_3)\}) = 1$.

4.3 Interpretation of the α-Disjunctions

The α-disjunctions can be interpreted in a similar way. Suppose two distinct sources S_1 and S_2 that induce two BBAs m^{Ω}_1 and m^{Ω}_2 on Ω. Suppose we want to compute the agent's beliefs on Ω given m^{Ω}_1, m^{Ω}_2 and the following distinct pieces of evidence.

- A piece of evidence stating that the sources do not lie simultaneously. This piece of evidence may be modeled by a BBA $m^{T_1\times T_2}_{or}$ defined by

$$m^{T_1\times T_2}_{or}(\{(t_1,t_2),(t_1,f_2),(f_1,t_2)\}) = 1.$$

- Distinct items of evidence for all $x\in\Omega$ of the form

$$pl^{T_1\times T_2}[x](\{(t_1,t_2)\}) = \alpha, \quad (15)$$

indicating that if $\omega_0 = x$, then it is plausible with strength α that both sources tell the truth.

The least committed BBA $m^{T_1\times T_2}[x]$ corresponding to (15) is the SBBA $m^{T_1\times T_2}[x] =$

$\{(f_1,t_2),(t_1,f_2),(f_1,f_2)\}^\alpha$. Using all these distinct items of evidence, the agent's belief m_{Ag}^Ω on Ω is then equal to

$$m_{Ag}^\Omega = (m_{1'}^{\Omega\times T_1} \textcircled{\cap} m_{2'}^{\Omega\times T_2} \textcircled{\cap} m_{or}^{T_1\times T_2} \textcircled{\cap} (\textcircled{\cap}_{x\in\Omega} m^{T_1\times T_2}[x]^{\Uparrow\Omega\times T_1\times T_2}))^{\downarrow\Omega}, \quad (16)$$

with $m_{i'}^{\Omega\times T_i}$, $i=1,2$, defined by (14), and where $m^{T_1\times T_2}[x] = \{(f_1,t_2),(t_1,f_2),(f_1,f_2)\}^\alpha$ for all $x\in\Omega$, and with $m_{or}^{T_1\times T_2}(\{(t_1,t_2),(t_1,f_2),(f_1,t_2)\})=1$.

Theorem 2. *Let m_1^Ω and m_2^Ω be two BBAs. The BBA m_{Ag}^Ω defined by (16) verifies*

$$m_{Ag}^\Omega = m_1^\Omega \textcircled{\cup}^\alpha m_2^\Omega.$$

As shown by Theorem 2, an α-disjunction is equivalent to the pooling by the conjunctive rule of some simple pieces of evidence. In particular, the parameter α involved in the α-disjunctions is equal to the plausibility that the sources tell the truth given $\omega_0 = x$. Note that since the BBA m_{or} excludes the fact that both sources lie, we clearly see, from the interpretation given to α, that we pass from the disjunctive rule to the exclusive disjunctive rule as α varies from 1 to 0.

To complete this section on the interpretation of the α-junctions, we may note that the idea of recovering the disjunctive rule and the exclusive disjunctive rule through the use of the conjunctive rule and BBAs defined on product spaces was investigated by Haenni in [4]. The difference between Haenni's approach and ours is that Haenni used the notion of the reliability of the sources, rather than their truthfulness. The main difference between a reliable source and a truthful source is the following. Suppose a source tells $\omega_0\in A$. If the source is lying, then the agent believes $\omega_0\in\overline{A}$, whereas when the source is unreliable, the agent believes $\omega_0\in\Omega$. As stated in [11] and as may easily be shown using the degenerate case $\alpha=0$ in Theorem 2, the exclusive disjunctive rule corresponds to the situation where exactly one of the sources tells the truth, without knowing which one. However, as shown in [7], this rule does not correspond to the situation where exactly one of the sources is reliable, without knowing which one, as wrongly claimed without proof by Theorem 3.3 of [4]. As a matter of fact, it can even be shown that it is actually the disjunctive rule that corresponds to that particular situation [7].

5 Computation

In addition to lacking an interpretation, the α-junctions suffered in [11] from another limitation: they were hard to compute. Indeed, the definitions of the matrices underlying the α-junctions are "quite laborious" [13] and thus using an α-junctive rule looks like a complicated task. It seems thus interesting to have simpler mechanisms to perform a combination by an α-junctive rule. As shown by Theorem 1, it is possible to compute the combination by an α-conjunctive rule using the conjunctive rule and BBAs defined on product spaces. In this section, several other new and simple means are provided to compute the combination by an α-conjunction. These new methods are based on generalizations of mechanisms that can be used to compute a combination by the conjunctive rule. Note that, although not provided in this paper, similar new means exist for the computation of the combination by an α-disjunction.

5.1 α-Conditioning Operation

Definition 2 below introduces a new notion, called α-conditioning, which will be useful to uncover a simple expression for the α-conjunctions.

Definition 2. *The α-conditioning of a BBA by a subset $B\subseteq\Omega$ is equal to the α-conjunction of this BBA with a categorical BBA focused on B.*

The result of the α-conditioning operation on a BBA m given a subset $B\subseteq\Omega$, i.e., the result of $m\textcircled{\cap}^\alpha m_B$ with m_B the categorical BBA focused on B, is denoted by $m[B]^\alpha$. We use the term "α-conditioning" because $m[B]^\alpha = m[B]$ when $\alpha=1$.

The following proposition provides an expression for the α-conditioning operation.

Proposition 1. *Let $B\subseteq\Omega$. We have, for all $X\subseteq\Omega$,*

$$m[B]^\alpha(X) = \sum_{(A\cap B)\cup(\overline{A}\cap\overline{B}\cap C)=X} m(A)\, m_\alpha(C),$$

where m_α is a BBA defined by, for all $A\subseteq\Omega$, $m_\alpha(A) = \alpha^{|\overline{A}|}(1-\alpha)^{|A|}$.

The following proposition introduces a new way to compute a combination by an α-conjunction, through the use of the α-conditioning operation.

Proposition 2. *Let m_1 and m_2 be two BBAs. We have, for all $A\subseteq\Omega$*

$$m_1\textcircled{\cap}^\alpha m_2(A) = \sum_{B\subseteq\Omega} m_1[B]^\alpha(A)\, m_2(B). \quad (17)$$

Note that, when $\alpha=1$, Equation (17) becomes equivalent to (3). Hence, Equation (17) may be seen as a generalization of (3).

5.2 "Classical" Expression

Using Propositions 1 and 2, it may be shown that the following proposition holds.

Proposition 3. *Let m_1 and m_2 be two BBAs. Let $m_{1\textcircled{\cap}^\alpha 2}$ denote $m_1 \textcircled{\cap}^\alpha m_2$. We have, for all $X \subseteq \Omega$,*

$$m_{1\textcircled{\cap}^\alpha 2}(X) = \sum_{(A\cap B)\cup(\overline{A}\cap\overline{B}\cap C)=X} m_1(A)\, m_2(B)\, m_\alpha(C), \tag{18}$$

where $m_\alpha(A) = \alpha^{|\overline{A}|}(1-\alpha)^{|A|}$, for all $A \subseteq \Omega$.

This proposition gives us yet another new expression for the α-conjunctions. We call (18) the "classical" expression for the α-conjunction since (18) is a generalization of the classical, or most often encountered, definition of the conjunctive rule given by Equation (2). Indeed, if $\alpha = 1$, then the BBA m_α of Proposition 3 is such that $m_\alpha(\emptyset) = 1$ and thus the term on the right side of (18) reduces to

$$\begin{aligned} & \sum_{(A\cap B)\cup(\overline{A}\cap\overline{B}\cap\emptyset)=X} m_1(A)\, m_2(B)\, m_\alpha(\emptyset) \\ = & \sum_{(A\cap B)=X} m_1(A)\, m_2(B) \\ = & \; m_1\textcircled{\cap}m_2(X), \end{aligned}$$

as expected. Similarly, if $\alpha = 0$, then $m_\alpha(\Omega) = 1$, and thus the term on the right side of (18) reduces to $m_1\textcircled{\cap}m_2(X)$, as expected.

5.3 α-Commonality Function

Using the eigendecomposition of $\mathbf{K}_m$ when $\mathbf{m}_{vac} = \mathbf{m}_\Omega$, Smets [11] showed that we have

$$g_{1\textcircled{\cap}^\alpha 2} = g_1 \cdot g_2 \tag{19}$$

with

$$\mathbf{g}_{1\textcircled{\cap}^\alpha 2} = \mathbf{G} \cdot \mathbf{m}_{1\textcircled{\cap}^\alpha 2}, \tag{20}$$

and $\mathbf{g}_1 = \mathbf{G} \cdot \mathbf{m}_1$ and $\mathbf{g}_2 = \mathbf{G} \cdot \mathbf{m}_2$, where $\mathbf{G}$ is a matrix of eigenvectors of $\mathbf{K}_m$ (due to lack of space, we refer the reader to [13, p. 26] for the definition of $\mathbf{G}$). From (19) and (20), we obtain

$$\mathbf{m}_{1\textcircled{\cap}^\alpha 2} = \mathbf{G}^{-1} \cdot \mathbf{Diag}(\mathbf{g}_1) \cdot \mathbf{g}_2, \tag{21}$$

where $\mathbf{Diag}(\mathbf{g}_1)$ denotes the diagonal matrix, whose diagonal elements are the elements of the vector $\mathbf{g}_1$. As shown by (21), the combination of two BBAs m_1 and m_2 by an α-conjunctive rule can be simply expressed as the pointwise product of the functions g_1 and g_2 associated, respectively, to m_1 and m_2. This is a first step in the simplification of the computation by an α-conjunction. However, the definition of the matrix $\mathbf{G}$ is as tedious as the definition of the matrix $\mathbf{K}_m$. Fortunately, Theorem 3 shows that it is possible to obtain the matrix $\mathbf{G}$ in a simple manner.

Theorem 3. *The matrix $\mathbf{G}$ may be obtained using Kronecker multiplication. We have:*

$$\mathbf{G}^{i+1} = \mathbf{Kron}\left(\begin{bmatrix} 1 & 1 \\ \alpha-1 & 1 \end{bmatrix}, \mathbf{G}^i\right), \mathbf{G}^1 = 1,$$

where $\mathbf{G}^i$ denotes the matrix $\mathbf{G}$ when $|\Omega| = i$.

We now have a very simple way to compute an α-conjunction, i.e., pointwise product of functions g which may themselves be obtained by a simple Kronecker product. Furthermore, it may now easily be seen that the $\mathbf{G}$ matrix generalizes the $\mathbf{Q}$ matrix in that we have $\mathbf{G} = \mathbf{Q}$ when $\alpha = 1$ and thus $g = q$ in this case. The fact that the function g generalizes the commonality function can be used to call g the *α-commonality function* associated to a BBA m.

5.4 Comparison of the Computation Methods

In this section, the various new means proposed for the computation of the combination by an α-conjunctive rule, are briefly compared.

We have laid bare four new ways of performing such a combination: (1) using the α-conditioning operation (see Proposition 2), (2) using a "classical" expression (see Proposition 3), (3) using the conjunctive rule and BBAs defined on product spaces (see Theorem 1) and (4) using the α-commonality function obtained from a Kronecker product (see (21) and Theorem 3).

Each of these techniques has some advantages and some drawbacks. Method 4 is arguably the simplest one to implement. However, it may rapidly become impossible to use if the frame of discernment Ω is too big, since this method requires computing matrices $\mathbf{G}$ of size $2^{|\Omega|} \times 2^{|\Omega|}$, which are, in addition, not sparse, and it requires performing the pointwise product of vectors g of size $2^{|\Omega|}$. Method 3 is also rather simple to implement, since we merely need to perform combinations by the conjunctive rule. However, it requires working in the space $\Omega \times T_1 \times T_2$. Method 1 and 2 share the same characteristics: they are more efficient than method 4 when the frame is big, since they do not require to work with vectors of size $2^{|\Omega|}$ as m_1 and m_2 may have only a few focal sets, but they are harder to implement.

6 Conclusion

The α-junctions represent the set of associative, commutative and linear combination operators for belief

functions with a neutral element. They include as particular cases familiar combination rules such as the conjunctive and disjunctive rules. They have never been used in the literature due, most certainly, to two limiting factors: in the original article of Smets [11], they lacked (1) an interpretation and (2) simple means to compute them. This paper has proposed solutions to these two issues.

It was first shown that the α-junctions correspond to some particular form of knowledge about the truthfulness of the sources, making the α-junctions interesting for applications where such kind of knowledge may be available. This might for instance be the case when dealing with automatic deceiving agents [14]. Then, it was shown that various notions that can be used to perform the computation by the conjunctive rule can be generalized to the α-junctions. This allowed us to uncover simple methods to perform a combination by an α-junctive rule. The α-junctions become thus more usable in practice and potentially useful, irrespective of their meaning, for, e.g., classification applications, as demonstrated in [7].

To conclude, let us mention that, as suggested in [13] and shown in [7], it is possible to obtain α-junctive canonical decompositions of a belief function, generalizing the conjunctive and disjunctive canonical decompositions [10, 2].

References

[1] Dempster, A. P.: Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statistics* 38, 325–339 (1967)

[2] Denœux, T.: Conjunctive and Disjunctive Combination of Belief Functions Induced by Non Distinct Bodies of Evidence. *Artificial Intelligence* 172, 234–264 (2008)

[3] Dubois, D., Prade, H.: A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets. *Int. J. Gen Syst* 12(3), 193–226 (1986)

[4] Haenni, R.: Uncover Dempster's Rule Where It Is Hidden. In: *Proceedings of the 9th International Conference on Information Fusion* (FUSION'2006), Florence, Italy, July (2006)

[5] Klawonn, F., Smets, Ph.: The dynamic of belief in the transferable belief model and specialization generalization matrices. In: Dubois, D., Wellman, M.P., D'Ambrosio, B., Smets, Ph. (eds.) *Proceedings of the 8th conference on Uncertainty in Artificial Intelligence*, pp 130–137. Morgan Kaufmann, San Meteo, CA (1992)

[6] Kohlas, J., Shenoy, P. P.: Computation in Valuation Algebras. In: Gabbay, D. M., Smets, Ph. (eds.) *Handbook of Defeasible Reasoning and Uncertainty Management Systems: Algorithms for Uncertainty and Defeasible Reasoning*, vol. 5, pp 5–39. Kluwer, Dordrecht (2000)

[7] Pichon, F.: Belief functions: canonical decompositions and combination rules. PhD thesis, Université de Technologie de Compiègne, March (2009) http://www.hds.utc.fr/~tdenoeux/perso/doku.php?id=en:students

[8] Shafer, G.: *A mathematical theory of evidence*. Princeton Univ. Press, Princeton, N.J. (1976)

[9] Smets, Ph.: Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *Int. J. Approx. Reasoning* 9, 1–35 (1993)

[10] Smets, Ph.: The canonical decomposition of a weighted belief. In: *Proceedings of the 14th Int. Joint Conf. on Artificial Intelligence*, San Mateo, USA, pp 1896–1901. Morgan Kaufmann (1995)

[11] Smets, Ph.: The α-junctions: combination operators applicable to belief functions. In: Gabbay, D.M., Kruse, R., Nonnengart, A., Ohlbach, H.J. (eds.), *Proc. of the 1st Int. Joint Conf. on Qualitative and Quantitative Practical Reasoning*, Bad Honnef, Germany. Lecture Notes in Computer Science, vol. 1244, pp 131–153. Springer (1997)

[12] Smets, Ph.: The Transferable Belief Model for quantified belief representation. In: Gabbay, D. M., Smets, Ph. (eds.), *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, vol. 1, pp 267–301. Kluwer (1998)

[13] Smets, Ph.: The application of the matrix calculus to belief functions. *Int. J. Approximate Reasoning* 31(1–2), 1–30 (2002)

[14] Smets, Ph.: Managing deceitful reports with the transferable belief model. In: Proc. of the Eighth International Conference on Information Fusion, IEEE, pp. C8–3/1–7, Piscataway, NJ (2005)

[15] Smets, Ph.: Analyzing the combination of conflicting belief functions. *Information Fusion* 8(4), 387–412 (2007)

[16] Smets, Ph., Kennes, R.: The Transferable Belief Model. *Artificial Intelligence* 66, 191–243 (1994)

[17] Yager, R. R.: The entailment principle for Dempster-Shafer granules. *Int. J. Intell. Syst.* 1, 247–262 (1986)

On solutions of stochastic differential equations with parameters modelled by random sets

Bernhard Schmelzer
Unit for Engineering Mathematics,
University of Innsbruck, Austria
bernhard.schmelzer@uibk.ac.at

Abstract

We consider ordinary stochastic differential equations whose coefficients depend on parameters. Conditions are given under which modelling the parameter uncertainty by compact-valued random sets leads to set-valued stochastic processes. Finally, we define analogues of first entrance times for set-valued processes.

Keywords. Stochastic differential equation, random set, set-valued stochastic process, first entrance time.

1 Introduction

Stochastic differential equations of the form

$$dx_t = f(t, x_t)dt + G(t, x_t)dw_t \tag{1}$$

or the equivalent integral form

$$x_t = x_{t_0} + \int_{t_0}^{t} f(s, x_s)ds + \int_{t_0}^{t} G(s, x_s)dw_s \tag{2}$$

with initial value x_{t_0}, coefficients $f : [t_0, \overline{t}] \times \mathbb{R}^d \to \mathbb{R}^d$, $G : [t_0, \overline{t}] \times \mathbb{R}^d \to \mathbb{R}^{d\times m}$ and $\{w_t\}_{t\in[t_0,\overline{t}]}$ being an m-dimensional Wiener process (Brownian motion) are used in many applications to model classical problems in physics and engineering under random disturbances. The theory of such equations and their solutions being stochastic processes can be found in [1] or [12], for example.

The motivation for this work is the desire for ultimately investigating mechanical systems under stochastic excitations depending on parameters. The purpose of this article is thus to consider SDEs whose initial value x_{t_0} and coefficients f and G depend on parameters. The uncertainty of these parameters can be modelled by random variables which requires the assumption of certain probability distributions. But in practice, there may only be scarce information available like a small sample size or estimates on the mean value and the variance. Hence, the classical probabilistic approach might involve tacit assumptions that cannot be verified and the need for alternative uncertainty models may arise (for a general discussion see for example [24]). Among those alternative models are random sets which can be interpreted as imprecise observations of random variables, that is, instead of a single value one assigns a set which is supposed to include the actual value to each of the elements of the underlying probability space. It has been demonstrated in [23, 25, 26] how random intervals constructed from Tchebycheff's inequality can serve as a non-parametric model of the variability of a parameter, given its mean value and variance as sole information.

We will start in Section 2 with a rather detailed review of the basic theory of stochastic processes and measurability of random sets which is necessary to understand the definitions and propositions of Section 3 where conditions will be given under which solution processes continuously depend on the parameters contained in x_{t_0}, f and G. We will show that this continuity together with using random compact sets for modelling parameter uncertainty leads to set-valued processes with compact values which are continuous with respect to the Hausdorff metric. Section 4 discusses possible definitions of analogues of first entrance times for set-valued processes and their representability by first entrance times of selections. In Section 5 an example is given to illustrate the theoretical concept developed in the foregoing sections.

We point out that this article addresses the case where f and G are $\mathbb{R}^d$-valued coefficient functions depending on random set parameters. This is in contrast with the case where f and G are functions taking values in the space of (closed) subsets of $\mathbb{R}^d$ which is discussed in [15, 18, 19, 20, 27, 28]. Note that the latter

approach could also be applied to the case of single-valued coefficients involving set-valued (even time dependent) parameters. But one substantial restriction is that a set-valued coefficient G in the noise term can lead to unbounded random sets in the solution process (even in very simple examples - see [27], Theorem 1) whereas using the method proposed in this paper leads to compact values when random compact sets are used to model parameter uncertainty. Of course, instead of random sets we could use fuzzy sets. But since each fuzzy set can be interpreted as a consonant random set on the interval $[0, 1]$ as underlying probability space, dealing with random sets is more general.

2 Preliminaries

2.1 Stochastic Processes

Throughout this section let (Ω, Σ, P) denote a probability space with σ-algebra Σ and probability measure P and let (T, r) and $(\mathbb{E}, \rho)$ be metric spaces. A stochastic process is a map

$$x : T \times \Omega \to \mathbb{E}, \omega \mapsto x_t(\omega) = x(t, \omega)$$

such that for each $t \in T$ the map

$$x_t : \Omega \to \mathbb{E}, \omega \mapsto x_t(\omega)$$

is a random variable, that is, it is measurable. For fixed $\omega \in \Omega$ the map

$$x.(\omega) : T \to \mathbb{E}, t \mapsto x_t(\omega)$$

is called sample function. Very often properties of stochastic processes cannot be verified for all $\omega \in \Omega$ but only for almost all ω, that is, for some subset of Ω whose probability is 1. That is why the term version is frequently used. Two stochastic processes x and $\tilde{x}$ are called versions of each other (or stochastically equivalent) if for all $t \in T$ it holds that

$$P(\{\omega : x_t(\omega) = \tilde{x}_t(\omega)\}) = 1.$$

The first property that should be mentioned here is separability.

Definition 1. ([5, 11]) Suppose that (T, r) is separable. A stochastic process $x : T \times \Omega \to \mathbb{E}$ is said to be separable if there exists a dense countable subset D of T and a set $N \in \Sigma$ of measure 0 such that for each open subset $G \subseteq T$ and every closed subset $F \subseteq \mathbb{E}$ the two sets

$$\{\omega : \forall t \in G \cap D : x_t(\omega) \in F\}$$
$$\{\omega : \forall t \in G : x_t(\omega) \in F\}$$

differ at most in N.

Hence, one could say that separability means that considering x for countably many $t \in T$ is enough to observe the behavior of the whole process. The following theorem whose proof can for example be found in [5] or [11] is fundamental for the theory of stochastic processes.

Theorem 1. ([5, 11]) Suppose that T is separable and $\mathbb{E}$ is compact. Then for any stochastic process $x : T \times \Omega \to \mathbb{E}$ there is a separable version.

Note that if $\mathbb{E}$ is only locally compact (which is the case if $\mathbb{E} = \mathbb{R}^d$) then one can always find a separable version in some compactification of $\mathbb{E}$ and its values are still in $\mathbb{E}$ with probability 1 for each $t \in T$.

Definition 2. A stochastic process is called (almost surely) continuous if (almost) all sample functions are continuous.

Recall that a probability space (Ω, Σ, P) is said to be complete if all subsets of sets $N \in \Sigma$ with $P(N) = 0$ are measurable, that is, lie in Σ. The completion of a probability space (Ω, Σ, P) is denoted $(\Omega, \overline{\Sigma}^P, \overline{P})$.

Proposition 1. ([10]) Suppose that (T, r) is separable and (Ω, Σ, P) is complete. Then a separable stochastic process which has an almost surely continuous version is almost surely continuous itself.

The next theorem states the so-called Kolmogorov-Chentsov criterion for almost sure continuity of sample functions.

Theorem 2. ([16]) Let $T = \mathbb{R}^p$, let $(\mathbb{E}, \rho)$ be a complete metric space. Suppose that a process $x : T \times \Omega \to \mathbb{E}$ satisfies for some positive constants α, β, γ the following condition

$$\mathrm{E}(\rho(x_s, x_t)^\alpha) \leq \gamma \, \|s - t\|^{p+\beta} \qquad \forall s, t \in T = \mathbb{R}^p. \quad (3)$$

Then x has an almost surely continuous version.

In the situation of the above Theorem 2, separability of x implies almost sure continuity of x if (Ω, Σ, P) is complete.

Definition 3. A stochastic process $x : T \times \Omega \to \mathbb{E}$ is called measurable if x is a measurable function with respect to the product-σ-algebra $\mathcal{B}(T) \otimes \Sigma$ where $\mathcal{B}(T)$ denotes the Borel-σ-algebra of (T, r).

Theorem 3. ([13]) Suppose that T is separable. Then a continuous process $x : T \times \Omega \to \mathbb{E}$ is measurable.

In the case where it is only known that almost all sample functions are continuous one can construct a version possessing only continuous sample functions by choosing a continuous sample path and replacing all discontinuous sample functions with this path.

2.2 Random Sets

A random set is a random variable whose values are sets. It is usual to consider random closed sets, that is, random variables whose values are closed subsets of some topological space $\mathbb{E}$. The Borel-σ-algebra on $\mathbb{E}$ is denoted by $\mathcal{B}(\mathbb{E})$ while $\mathcal{G}(\mathbb{E})$, $\mathcal{F}(\mathbb{E})$ and $\mathcal{K}(\mathbb{E})$ denote, respectively, the family of open, closed and compact subsets of $\mathbb{E}$. By $\mathcal{F}'(\mathbb{E})$ and $\mathcal{K}'(\mathbb{E})$ we mean $\mathcal{F}(\mathbb{E})\backslash\{\emptyset\}$ and $\mathcal{K}(\mathbb{E})\backslash\{\emptyset\}$, respectively.

Again let (Ω, Σ, P) be a probability space. As with random variables a random closed set $A : \Omega \to \mathcal{F}(\mathbb{E})$ has to fulfill some measurability condition. We shall demand that

$$A^-(B) = \{\omega : A(\omega) \cap B \neq \emptyset\} \in \Sigma, \quad \forall B \in \mathcal{B}(\mathbb{E}). \tag{4}$$

For other measurability definitions for set-valued maps we refer to [2, 13], for example. Furthermore, we call A a random compact set if Condition (4) is satisfied and for all $\omega \in \Omega$ it holds that $A(\omega) \in \mathcal{K}(\mathbb{E})$.

One can view a random set A as a collection of random variables that fit inside A. Such single-valued measurable functions $\alpha : \Omega \to \mathbb{E}$ fulfilling

$$\alpha(\omega) \in A(\omega), \quad \forall \omega \in \Omega$$

are called selections of A. Let $\mathcal{S}(A)$ denote the set of all measurable selections of A. The following theorem which is referred to as the Fundamental Measurability Theorem gives conditions for the measurability of random closed sets and the existence of measurable selections. For its proof and related results see [2] and [13].

Theorem 4. ([2, 13]) Suppose that $(\mathbb{E}, \rho)$ is a complete separable metric space. Let $A : \Omega \to \mathcal{F}'(\mathbb{E})$ be a set-valued mapping with non-empty values. Consider the following properties:

(i) For all $B \in \mathcal{B}(\mathbb{E})$ it holds that $A^-(B) \in \Sigma$,

(ii) for all $F \in \mathcal{F}(\mathbb{E})$ it holds that $A^-(F) \in \Sigma$,

(iii) for all $G \in \mathcal{G}(\mathbb{E})$ it holds that $A^-(G) \in \Sigma$,

(iv) there is a Castaing representation of A, that is, a sequence $\{\alpha_n\}_{n\in\mathbb{N}}$ of measurable selections such that for all $\omega \in \Omega$

$$A(\omega) = \mathrm{cl}(\{\alpha_n(\omega)\}_{n\in\mathbb{N}})$$

where cl denotes the closure in $\mathbb{E}$,

(v) for all $x \in \mathbb{E}$ the function $\omega \mapsto \inf_{y\in A(\omega)} \rho(x, y)$ is measurable,

(vi) the graph of A belongs to $\Sigma \otimes \mathcal{B}(\mathbb{E})$.

Then the following implications hold:

$$\text{(i)}\Rightarrow\text{(ii)}\Rightarrow\text{(iii)}\Leftrightarrow\text{(iv)}\Leftrightarrow\text{(v)}\Rightarrow\text{(vi)}$$

If (Ω, Σ, P) is a complete probability space then all properties are equivalent.

Note that in the literature (for example in [22]) one can also find "almost all" versions of the above theorem and definitions. For further background information on random sets see [21, 22, 29].

3 Stochastic Differential Equations with Random Set Parameters

3.1 Deterministic parameters

Let us consider stochastic differential equations of the form (2) whose initial value x_{t_0} and coefficients f and G depend on some vector $a = (a_1, \ldots, a_p) \in \mathbb{A}$ of parameters where $\mathbb{A} \subseteq \mathbb{R}^p$ denotes the set of possible parameter values, that is, we consider differential equations of the form

$$x_{t,a} = x_{t_0,a} + \int_{t_0}^{t} f(s, a, x_{s,a})ds + \int_{t_0}^{t} G(s, a, x_{s,a})dw_s \tag{5}$$

where $t_0 \leq t \leq \bar{t} < \infty$, $a \in \mathbb{A}$, w_t denotes an m-dimensional Wiener process on a probability space (Ω, Σ, P) and

$$\begin{array}{rll} x_{t_0}: & \mathbb{A}\times\Omega \to \mathbb{R}^d, & (a,\omega) \mapsto x_{t_0,a}(\omega), \\ f: & [t_0,\bar{t}]\times\mathbb{A}\times\mathbb{R}^d \to \mathbb{R}^d, & (t,a,x) \mapsto f(t,a,x), \\ G: & [t_0,\bar{t}]\times\mathbb{A}\times\mathbb{R}^d \to \mathbb{R}^{d\times m}, & (t,a,x) \mapsto G(t,a,x). \end{array}$$

Assume that for each $a \in \mathbb{A}$ the partial maps $f(\cdot, a, \cdot)$ and $G(\cdot, a, \cdot)$ are measurable functions and the usual conditions for the existence of a solution process ([1, 12]) are fulfilled, that is,

(IV) $x_{t_0,a}$ is a random variable independent of the increments $w_t - w_{t_0}$ for $t \geq t_0$.

(Lip) Lipschitz condition: There is a constant $L > 0$ such that for all $t \in [t_0, \bar{t}]$ and all $x, y \in \mathbb{R}^d$ it holds that

$$\begin{aligned} &\|f(t,a,x) - f(t,a,y)\| \\ &\quad + \|G(t,a,x) - G(t,a,y)\| \leq L\|x - y\|. \end{aligned}$$

(RG) Restriction on growth: There is a constant $K > 0$ such that for all $t \in [t_0, \bar{t}]$ and all $x \in \mathbb{R}^d$ it holds that

$$\|f(t,a,x)\|^2 + \|G(t,a,x)\|^2 \leq K(1 + \|x\|^2).$$

Note that the constants L and K can depend on a. If the above conditions are fulfilled we get for each $a \in \mathbb{A}$ a solution process $\{x_t\}_{t\in[t_0,\overline{t}]} = \{x_{t,a}\}_{t\in[t_0,\overline{t}]}$, which leads to a map of the form

$$x : [t_0, \overline{t}] \times \mathbb{A} \times \Omega \to \mathbb{R}^d, (t, a, \omega) \mapsto x_{t,a}(\omega). \quad (6)$$

Since for each $a \in \mathbb{A}$ and each $t \in [t_0, \overline{t}]$ the partial map $x_{t,a} = x(t, a, \cdot) : \Omega \to \mathbb{R}^d$ is measurable, (6) can be interpreted as a stochastic process on $[t_0, \overline{t}] \times \mathbb{A}$ which is a metric space. Hence, according to Theorem 1, we can assume x to be separable.

Looking at the process x defined by Equation (6) the question arises if it is continuous in (t, a). From Itô's theory it is well-known that for fixed $a \in \mathbb{A}$ the solution process $\{x_{t,a}\}_{t\in[t_0,\overline{t}]}$ is continuous in t. Furthermore, it fulfills the inequality in Theorem 2 (see [1] or [12]), that is, there is some constant C such that for all $s, t \in [t_0, \overline{t}]$

$$\mathrm{E}(\|x_t - x_s\|^{2n}) \le C|t - s|^n, \quad t, s \in [t_0, \overline{t}] \quad (7)$$

holds if the $2n$-th moment of the initial value is finite. The next proposition will give conditions under which the corresponding inequality with respect to t and a is fulfilled on a bounded subset of $[t_0, \overline{t}] \times \mathbb{A}$.

Proposition 2. Let $\{x_{t,a}\}_{(t,a)\in[t_0,\overline{t}]\times\mathbb{A}}$ denote the process defined by Equation (6), let $U \subseteq \mathbb{A}$ be an arbitrary bounded subset of $\mathbb{A}$ and let $n \in \mathbb{N}$. Assume that Conditions (IV), (Lip) and (RG) are fulfilled and in addition, the following conditions hold:

(C1) $L : \mathbb{A} \to \mathbb{R}_{\ge 0}$ from (Lip) and $K : \mathbb{A} \to \mathbb{R}_{\ge 0}$ from (RG) are bounded on U.

(C2) Local Lipschitz condition with respect to a: For all $x \in \mathbb{R}^d$ there exists a constant $\tilde{L} = \tilde{L}(U, x) > 0$ such that for all $t \in [t_0, \overline{t}]$ and for all $a, b \in U$ it holds that

$$\|f(t, a, x) - f(t, b, x)\| + \|G(t, a, x) - G(t, b, x)\| \le \tilde{L}(U, x)\|a - b\|$$

where the growth of $\tilde{L}$ is bounded by a polynomial in $\|x\|$, that is, there is an $M = M(U) > 0$ and a $k = k(U) \in \mathbb{N}$ such that for all $x \in \mathbb{R}^d$

$$\tilde{L}(U, x) \le M(U)(1 + \|x\|)^k.$$

(C3) The $2nk$-th moments of the initial values $x_{t_0,a}$ exist and are bounded on U, that is,

$$\sup_{a\in U} \mathrm{E}(\|x_{t_0,a}\|^{2nk}) < \infty.$$

In addition, there is a constant $c = c(U, n)$ such that for all $a, b \in U$ it holds that

$$\mathrm{E}(\|x_{t_0,a} - x_{t_0,b}\|^{2n}) \le c\|a - b\|^{2n}.$$

Then there is a constant $C = C(U, n) > 0$ such that for all $s, t \in [t_0, \overline{t}]$ and for all $a, b \in U$ the following inequality holds

$$\mathrm{E}(\|x_{s,a} - x_{t,b}\|^{2n}) \le C \left\| \begin{pmatrix} s - t \\ a - b \end{pmatrix} \right\|^n. \quad (8)$$

The rather technical proof is omitted since it is similar to the proof of (7) (see [1, 12]).

Now, we can conclude that a separable version of our process (6) is almost surely continuous with respect to (t, a) if the conditions of the above proposition are satisfied for $n \in \mathbb{N}$ big enough.

Proposition 3. The stochastic process $\{x_{t,a}\}_{(t,a)\in[t_0,\overline{t}]\times\mathbb{A}}$ defined by (6) is almost surely continuous with respect to (t, a) if there is an $n \ge p + 2$ such that the conditions of Proposition 2 are satisfied for each bounded subset $U \subseteq \mathbb{A}$.

Proof. Let $c \in \mathbb{A}$ and let $U(c) \subseteq \mathbb{A}$ denote a bounded neighborhood of c. Since the conditions of Proposition 2 are fulfilled for some $n \ge p+2$ we know that (8) holds for all $(s, a), (t, b) \in [t_0, \overline{t}] \times U(c)$ which means that, according to Proposition 1, x is an almost surely continuous process on $[t_0, \overline{t}] \times U(c)$, that is, there is a measure-zero set $N(c) \in \Sigma$ such that for all $\omega \in N(c)^{\mathsf{c}}$ the sample function $x_{\cdot,\cdot}(\omega)$ is continuous. Since $\mathbb{A}$ can be covered by bounded neighborhoods of countably many $c \in \mathbb{A}$ the set $N^{\mathsf{c}} = \bigcap_c N(c)^{\mathsf{c}}$ is measurable and has probability 1 which means that x is an almost surely continuous process on $[t_0, \overline{t}] \times \mathbb{A}$. □

If we replace, as described at the end of Section 2.1, $\{x_{t,a}\}_{(t,a)\in[t_0,\overline{t}]\times\mathbb{A}}$ by a continuous version, we can infer measurability from Theorem 3.

Corollary 1. Let $\mathbb{A} \in \mathcal{B}(\mathbb{R}^p)$ be a Borel set and let $\{x_{t,a}\}_{(t,a)\in[t_0,\overline{t}]\times\mathbb{A}}$ be a continuous process of the form (6). If we choose an $\omega \in \Omega$ such that $x_{\cdot,\cdot}(\omega)$ is continuous and replace all discontinuous sample functions by $x_{\cdot,\cdot}(\omega)$ we get a continuous version which is, according to Theorem 3, measurable with respect to $\mathcal{B}([t_0, \overline{t}]) \otimes \mathcal{B}(\mathbb{A}) \otimes \Sigma$.

3.2 Parameters modelled by random variables

From now on the probability space on which the Wiener process $\{w_t\}_{t\ge t_0}$ is defined shall be denoted $(\Omega_w, \Sigma_w, P_w)$. We assume that the stochastic process $\{x_{t,a}\}_{(t,a)\in[t_0,\overline{t}]\times\mathbb{A}}$ defined by Equation (6) is measurable with respect to the product-σ-algebra $\mathcal{B}([t_0, \overline{t}]) \otimes \mathcal{B}(\mathbb{A}) \otimes \Sigma_w$ and all sample functions are continuous on $[t_0, \overline{t}] \times \mathbb{A}$. The measurability of x allows us to model the parameter uncertainty of a by a random

variable, that is, a measurable function $\alpha : \Omega_{\mathbb{A}} \to \mathbb{A}$ on some probability space $(\Omega_{\mathbb{A}}, \Sigma_{\mathbb{A}}, P_{\mathbb{A}})$. Consequently, the map

$$\begin{array}{rccc} \hat{\alpha}: & [t_0,\overline{t}]\times\Omega_{\mathbb{A}}\times\Omega_w & \to & [t_0,\overline{t}]\times\mathbb{A}\times\Omega_w \\ & (t,\omega_{\mathbb{A}},\omega_w) & \mapsto & (t,\alpha(\omega_{\mathbb{A}}),\omega_w) \end{array}$$

is measurable with respect to the product σ-algebra $\mathcal{B}([t_0,\overline{t}])\otimes\Sigma_{\mathbb{A}}\otimes\Sigma_w$. Composing $\hat{\alpha}$ and x leads to the measurable map $\xi = x\circ\hat{\alpha}$

$$\begin{array}{rccc} \xi: & [t_0,\overline{t}]\times\Omega_{\mathbb{A}}\times\Omega_w & \to & \mathbb{R}^d \\ & (t,\omega_{\mathbb{A}},\omega_w) & \mapsto & x(t,\alpha(\omega_{\mathbb{A}}),\omega_w) \end{array} \tag{9}$$

which can be interpreted as a stochastic process $\{\xi_t\}_{t\in[t_0,\overline{t}]}$ on the time interval $[t_0,\overline{t}]$ and the product space $(\Omega,\Sigma,P) = (\Omega_{\mathbb{A}}\times\Omega_w, \Sigma_{\mathbb{A}}\otimes\Sigma_w, P_{\mathbb{A}}\otimes P_w)$.

Proposition 4. The map ξ defined by (9) can be interpreted as a stochastic process $\{\xi_t\}_{t\in[t_0,\overline{t}]}$ on the time interval $[t_0,\overline{t}]$ and the probability space (Ω,Σ,P). The process $\{\xi_t\}_{t\in[t_0,\overline{t}]}$ is measurable and all sample functions are continuous.

Proof. The map $\xi = x\circ\hat{\alpha}$ is measurable since it is the composition of the two measurable functions $\hat{\alpha}$ and x where the domain of x is the same measure space as the range of $\hat{\alpha}$. Consequently, for each $t\in[t_0,\overline{t}]$ the partial map

$$\xi_t : \Omega \to \mathbb{R}^d, \omega \mapsto x_{t,\alpha(\omega_{\mathbb{A}})}(\omega_w)$$

is a random variable which means that ξ is a measurable stochastic process. Note that for each $a\in\mathbb{A}$ and each $\omega_w\in\Omega_w$ the partial map $x_{\cdot,a}(\omega_w)$ is continuous because the sample function $x_{\cdot,\cdot}(\omega_w)$ is continuous. Since for all $\omega_{\mathbb{A}}$ we have $\alpha(\omega_{\mathbb{A}})\in\mathbb{A}$ we can infer that $\xi_\cdot(\omega) = x_{\cdot,\alpha(\omega_{\mathbb{A}})}(\omega_w)$ is continuous for all $\omega\in\Omega$. □

3.3 Parameters modelled by random sets

The uncertainty of the parameter a in Equation (5) shall now be modelled by a random compact set

$$A : \Omega_{\mathbb{A}} \to \mathcal{K}'(\mathbb{A})$$

where $\mathcal{K}'(\mathbb{A})$ denotes the set of all non-empty compact subsets of $\mathbb{R}^p$ being also a subset of $\mathbb{A}$. Then we can define a set-valued function X by

$$X : (t,\omega) \mapsto \{x_{t,a}(\omega_w) : a\in A(\omega_{\mathbb{A}})\} \tag{10}$$

where $(t,\omega)\in[t_0,\overline{t}]\times\Omega$ and x is the process defined by (6) which is still assumed to be measurable and continuous. The next proposition states that X is a set-valued process with compact values, that is, for each $t\in[t_0,\overline{t}]$ it holds that X_t is a random compact set which particularly means that the measurability condition (4) is fulfilled.

Proposition 5. Let $A : \Omega_{\mathbb{A}} \to \mathcal{K}'(\mathbb{A})$ be a random compact set and let X be the set-valued map defined by Equation (10). Then the following holds:

1. X can be interpreted as a set-valued process on the time interval $[t_0,\overline{t}]$ and the completed probability space $(\Omega,\overline{\Sigma}^P,\overline{P})$ with values in $\mathcal{K}'(\mathbb{R}^d)$,
2. All sample functions of X are continuous with respect to the Hausdorff-metric H on $\mathcal{K}'(\mathbb{R}^d)$.
3. X is measurable with respect to the product-σ-algebra $\mathcal{B}([t_0,\overline{t}])\otimes\overline{\Sigma_{\mathbb{A}}\otimes\Sigma_w}^{P_{\mathbb{A}}\otimes P_w}$.
4. For a Castaing representation $\{\alpha_n\}_{n\in\mathbb{N}}$ of A the processes $\{\xi^n\}_{n\in\mathbb{N}}$ defined by
$$\xi^n_t(\omega) = x_{t,\alpha_n(\omega_{\mathbb{A}})}(\omega_w),\ (t,\omega)\in[t_0,\overline{t}]\times\Omega$$
form a Castaing representation of X and for each $t\in[t_0,\overline{t}]$ the family $\{\xi^n_t\}_{n\in\mathbb{N}}$ forms a Castaing representation of X_t.

Proof. First note that $X_t(\omega)$ is a non-empty compact subset of $\mathbb{R}^d$ for all $t\in[t_0,\overline{t}]$ and all $\omega\in\Omega$ since $x_{t,\cdot}(\omega_w)$ is continuous in a and $A(\omega_{\mathbb{A}})$ is a non-empty compact subset of $\mathbb{R}^p$ for all $\omega_{\mathbb{A}}\in\Omega_{\mathbb{A}}$. Since for the proof of the first three statements the Castaing representation $\{\xi^n\}_{n\in\mathbb{N}}$ is used Assertion 4 is proved first. Hence, we show that for all $(t,\omega)\in[t_0,\overline{t}]\times\Omega$ it holds that

$$\{x_{t,a}(\omega_w) : a\in A(\omega_{\mathbb{A}})\} = \mathrm{cl}(\{\xi^n_t(\omega)\}_{n\in\mathbb{N}}).$$

In fact, since $\{\alpha_n\}_{n\in\mathbb{N}}$ is a Castaing representation of A we know that for all $a\in A(\omega_{\mathbb{A}})$ there is a subsequence $\{\alpha_{n_j}\}_{j\in\mathbb{N}}$ such that $\alpha_{n_j}(\omega_{\mathbb{A}})\to a$ for $j\to\infty$. Continuity of $x_{t,\cdot}(\omega_w)$ in a implies $\xi^{n_j}_t(\omega) = x_{t,\alpha_{n_j}(\omega_{\mathbb{A}})}(\omega_w) \to x_{t,a}(\omega_w)$ which means that $x_{t,a}(\omega_w)\in\mathrm{cl}(\{\xi^n_t(\omega)\})$. On the other hand, it is clear that $\alpha_n(\omega_{\mathbb{A}})\in A(\omega_{\mathbb{A}})$ for all $\omega_{\mathbb{A}}\in\Omega_{\mathbb{A}}$, $n\in\mathbb{N}$ and consequently $\xi^n_t(\omega)\in X_t(\omega)$ for all $\omega\in\Omega$ and $n\in\mathbb{N}$. Since $X_t(\omega)$ is closed, it holds that $\mathrm{cl}(\{\xi^n_t(\omega)\}_{n\in\mathbb{N}})\subseteq X_t(\omega)$. Hence, for each $t\in[t_0,\overline{t}]$ it follows that $X_t(\omega) = \mathrm{cl}(\{\xi^n_t(\omega)\}_{n\in\mathbb{N}})$ for all $\omega\in\Omega$. According to the Fundamental Measurability Theorem 4, this means that X_t is a random compact set on the completion of the probability space (Ω,Σ,P), that is,

$$(\Omega_{\mathbb{A}}\times\Omega_w, \overline{\Sigma_{\mathbb{A}}\otimes\Sigma_w}^{P_{\mathbb{A}}\otimes P_w}, \overline{P_{\mathbb{A}}\otimes P_w}).$$

The continuity of X is a consequence of the continuity of the processes ξ^n ($n\in\mathbb{N}$). Indeed, after recalling that for $A,B\in\mathcal{K}'(\mathbb{R}^d)$ the Hausdorff-metric H is defined by

$$H(A,B) = \max(\sup_{a\in A}\inf_{b\in B}\|a-b\|, \sup_{b\in B}\inf_{a\in A}\|a-b\|)$$

suppose that for arbitrary $\omega \in \Omega$ there is a $t \in [t_0, \overline{t}]$ and an $\varepsilon_0 > 0$ such that for all $\delta > 0$ there is an $s = s(\delta)$ such that $|s - t| < \delta$ and

$$H(X_s(\omega), X_t(\omega)) \geq \varepsilon_0.$$

Because of the closedness of $X_t(\omega)$ and $X_s(\omega)$ this corresponds to the assumption that at least one of the following two inequalities holds

$$\begin{aligned} \sup_{n\in\mathbb{N}} \inf_{m\in\mathbb{N}} \|\xi_s^n(\omega) - \xi_t^m(\omega)\| &\geq \varepsilon_0, \\ \sup_{m\in\mathbb{N}} \inf_{n\in\mathbb{N}} \|\xi_s^n(\omega) - \xi_t^m(\omega)\| &\geq \varepsilon_0. \end{aligned}$$

From the first inequality one can infer that there is an $n \in \mathbb{N}$ such that for all $m \in \mathbb{N}$ it holds that

$$\|\xi_s^n(\omega) - \xi_t^m(\omega)\| \geq \inf_{m\in\mathbb{N}} \|\xi_s^n(\omega) - \xi_t^m(\omega)\| \geq \frac{\varepsilon_0}{2}.$$

Of course, this inequality also holds for the choice $m = n$ which leads to

$$\|\xi_s^n(\omega) - \xi_t^n(\omega)\| \geq \frac{\varepsilon_0}{2},$$

but this would mean that ξ^n is not continuous at t. If we apply the same argument to the second inequality we can conclude that $H(X_s(\omega), X_t(\omega)) \geq \varepsilon_0$ cannot hold. Hence, X is a continuous process.

Since $\mathcal{K}'(\mathbb{R}^d)$ together with the Hausdorff metric H is a metric space the measurability of X is a direct consequence of the continuity of all sample functions $X_\cdot(\omega)$ and Theorem 3. □

The different maps that appeared in this section together with the underlying measure spaces are summarized in the following table. (Note that λ and λ^p denote the Lebesgue measures on $\mathcal{B}([t_0, \overline{t}])$ and $\mathcal{B}(\mathbb{A})$, respectively.)

map	underlying measure space
x	$([t_0, \overline{t}] \times \mathbb{A} \times \Omega_w,$ $\mathcal{B}([t_0, \overline{t}]) \otimes \mathcal{B}(\mathbb{A}) \otimes \Sigma_w, \lambda \otimes \lambda^p \otimes P_w)$
α, A	$(\Omega_\mathbb{A}, \Sigma_\mathbb{A}, P_\mathbb{A})$
$\hat{\alpha}$, ξ	$([t_0, \overline{t}] \times \Omega_\mathbb{A} \times \Omega_w,$ $\mathcal{B}([t_0, \overline{t}]) \otimes \Sigma_\mathbb{A} \otimes \Sigma_w, \lambda \otimes P_\mathbb{A} \otimes P_w)$
X	$([t_0, \overline{t}] \times \Omega_\mathbb{A} \times \Omega_w,$ $\mathcal{B}([t_0, \overline{t}]) \otimes \overline{\Sigma_\mathbb{A} \otimes \Sigma_w}^{P_\mathbb{A} \otimes P_w}, \lambda \otimes \overline{P_\mathbb{A} \otimes P_w})$

4 First Entrance and Inclusion Times for Set-valued Processes

In many applications, it is useful to observe the first time where a single-valued stochastic process enters some subset of the state space or the last time where it leaves this subset. For example, one could be interested in the first exceedance of a certain level by a real-valued process to assess the reliability of a system described by this process (see for example [30]). In his book [7], Dynkin discusses the theory of first entrance and exit times of right-continuous Markov processes. Other theoretical background can be found in [4, 17].

For a (single-valued) d-dimensional process $\{\xi_t\}_{t\in[t_0,\overline{t}]}$ on a probability space (Ω, Σ, P) and a subset $B \subseteq \mathbb{R}^d$ we shall call

$$\tau_\xi^B : \Omega \to [t_0, \overline{t}], \omega \mapsto \inf\{t : \xi_t(\omega) \in B\} \tag{11}$$

the first entrance time of ξ into B. Note that if the infimum does not exist we set $\tau_\xi^B(\omega) = \overline{t}$. One can show (see [7]) that (11) is measurable if $B \in \mathcal{B}(\mathbb{R}^d)$ and ξ is right-continuous. Furthermore, if ξ is continuous and B is a closed subset of $\mathbb{R}^d$ then τ_ξ^B is a stopping time w.r.t. the natural filtration $\{\mathcal{A}_t\}_{t\in[t_0,\overline{t}]}$ defined by

$$\mathcal{A}_t = \sigma(\xi_s^{-1}(B) : s \in [t_0, \overline{t}], B \in \mathcal{B}(\mathbb{R}^d)), \tag{12}$$

and if B is open then τ_ξ^B is a stopping time w.r.t. the right-continuous filtration $\{\mathcal{A}_{t+}\}_{t\in[t_0,\overline{t}]}$ where

$$\mathcal{A}_{t+} = \bigcap_{t<s\leq\overline{t}} \mathcal{A}_s, \ \mathcal{A}_{\overline{t}+} = \mathcal{A}_{\overline{t}}. \tag{13}$$

If we consider a continuous process $\{X_t\}_{t\in[t_0,\overline{t}]}$ with values in $\mathcal{K}'(\mathbb{R}^d)$ we can define the following two maps that correspond to (11):

$$\underline{\tau}^B : \Omega \to [t_0, \overline{t}], \quad \omega \mapsto \inf\{t : X_t(\omega) \cap B \neq \emptyset\} \tag{14}$$

$$\overline{\tau}^B : \Omega \to [t_0, \overline{t}], \quad \omega \mapsto \inf\{t : X_t(\omega) \subseteq B\} \tag{15}$$

If the infimum does not exist, we set $\underline{\tau}^B(\omega) = \overline{t}$ or $\overline{\tau}^B(\omega) = \overline{t}$, respectively. We call $\underline{\tau}^B$ the first entrance time of X into B, and we call $\overline{\tau}^B$ the first inclusion time of X in B.

Considering the natural filtration $\{\Sigma_t\}_{t\in[t_0,\overline{t}]}$ of X defined by

$$\Sigma_t = \sigma(X_s^-(B) : s \in [t_0, t], B \in \mathcal{B}(\mathbb{R}^d)) \subseteq \Sigma \tag{16}$$

the next proposition (which is the set-valued analogue of Dynkin's Lemma 4.1 in [7]) gives conditions under which $\underline{\tau}^B$ and $\overline{\tau}^B$ are measurable or even stopping times w.r.t. the augmented filtration $\{\hat{\Sigma}_t^P\}_{t\in[t_0,\overline{t}]}$, that is the ascending family of complete σ-algebras defined by

$$\hat{\Sigma}_t^P = \sigma(\Sigma_t \cup \mathcal{N}) \subseteq \overline{\Sigma}^P \tag{17}$$

where $\mathcal{N}$ is the set of all subsets of measure-zero sets in Σ.

Proposition 6. Suppose that $\{X_t\}_{t\in[t_0,\overline{t}]}$ is a continuous $\mathcal{K}'(\mathbb{R}^d)$-valued process on a probability space (Ω,Σ,P) and $\{\Sigma_t\}_{t\in[t_0,\overline{t}]}$ is its natural filtration defined by (16).

1. If $B\in\mathcal{G}(\mathbb{R}^d)$ is an open subset of $\mathbb{R}^d$ then
$$\{\omega:\underline{\tau}^B(\omega)\le t\},\ \{\omega:\overline{\tau}^B(\omega)\le t\}\in\hat{\Sigma}^P_{t+}.$$
2. If $B\in\mathcal{F}(\mathbb{R}^d)$ is a closed subset of $\mathbb{R}^d$ then
$$\{\omega:\underline{\tau}^B(\omega)\le t\},\ \{\omega:\overline{\tau}^B(\omega)\le t\}\in\hat{\Sigma}^P_{t}.$$

Proof. The proof is omitted here since it is very similar to the proof of Lemma 4.1 in [7]. □

An interesting question is if $\underline{\tau}^B$ and $\overline{\tau}^B$ can be attained by first entrance times of selections of X. The next proposition states that this is possible.

Proposition 7. Let $X:[t_0,\overline{t}]\times\Omega\to\mathcal{K}'(\mathbb{R}^d)$ be a continuous set-valued process with non-empty compact values and let $B\subseteq\mathbb{R}^d$ be an arbitrary subset of $\mathbb{R}^d$. Then for all $\omega\in\Omega$ it holds that
$$\begin{aligned}\inf_{\xi\in\mathcal{S}(X)}\tau^B_\xi(\omega) &= \underline{\tau}^B(\omega),\\ \sup_{\xi\in\mathcal{S}(X)}\tau^B_\xi(\omega) &\le \overline{\tau}^B(\omega).\end{aligned}$$
If (Ω,Σ,P) is complete and $B\in\mathcal{G}(\mathbb{R}^d)$ then for all $\omega\in\Omega$ the second inequality becomes an equality.

Proof. The equality for $\underline{\tau}^B$ and the inequality for $\overline{\tau}^B$ can be seen easily by using the equation
$$X_t(\omega)=\{\xi_t(\omega):\xi\in\mathcal{S}(X)\}$$
which holds for all $t\in[t_0,\overline{t}]$ and $\omega\in\Omega$. If (Ω,Σ,P) is complete and B is an open subset of $\mathbb{R}^d$ then $\overline{\tau}^B$ is Σ-measurable by Proposition 6. Consider the map
$$Y:(t,\omega)\mapsto\begin{cases}X_t(\omega) & \text{if } \overline{\tau}^B(\omega)\le t\\ X_t(\omega)\cap B^{\mathsf{c}} & \text{if } \overline{\tau}^B(\omega)>t\end{cases}$$
which has non-empty closed values. Note that
$$M=\{(t,\omega)\in[t_0,\overline{t}]\times\Omega:\overline{\tau}^B(\omega)\le t\}\in\mathcal{B}([t_0,\overline{t}])\otimes\Sigma$$
since $(t,\omega)\mapsto\overline{\tau}^B(\omega)-t$ is a measurable function. Furthermore, it can be checked easily that for any $C\in\mathcal{B}(\mathbb{R}^d)$ it holds that
$$Y^-(C)=(X^-(C)\cap M)\cup(X^-(B^{\mathsf{c}}\cap C)\cap M^{\mathsf{c}})$$
which means that Y is a random closed set. From Theorem 4 one can infer that there is a selection $\xi\in\mathcal{S}(Y)$ which implies that $\tau^B_\xi(\omega)=\overline{\tau}^B(\omega)$ for all $\omega\in\Omega$. Since $Y(\omega)\subseteq X(\omega)$ for all $\omega\in\Omega$ the map ξ is also a selection of X. □

For a set-valued process defined by (10) which fulfills the conditions of Proposition 5 we can consider for each $\alpha\in\mathcal{S}(A)$ and $a\in\mathbb{A}$ the special entrance times
$$\begin{aligned}\tau^B_\alpha:\omega &\mapsto \inf\{t\in[t_0,\overline{t}]:x_{t,\alpha(\omega_\mathbb{A})}(\omega_w)\in B\},\\ \tau^B_a:\omega_w &\mapsto \inf\{t\in[t_0,\overline{t}]:x_{t,a}(\omega_w)\in B\}.\end{aligned}$$

Proposition 8. Let $X:[t_0,\overline{t}]\times\Omega\to\mathcal{K}'(\mathbb{R}^d)$ be a set-valued process defined by (10) which fulfills the conditions of Proposition 5. Then the following relations hold for all $\omega\in\Omega$
$$\begin{aligned}\inf_{a\in A(\omega_\mathbb{A})}\tau^B_a(\omega_w)=\inf_{\alpha\in\mathcal{S}(A)}\tau^B_\alpha(\omega) &= \inf_{\xi\in\mathcal{S}(X)}\tau^B_\xi(\omega)\\ \sup_{a\in A(\omega_\mathbb{A})}\tau^B_a(\omega_w)=\sup_{\alpha\in\mathcal{S}(A)}\tau^B_\alpha(\omega) &\le \sup_{\xi\in\mathcal{S}(X)}\tau^B_\xi(\omega).\end{aligned}$$

Proof. Let $\omega\in\Omega$. Note that $\tau^B_\alpha(\omega)=\tau^B_{\alpha(\omega_\mathbb{A})}(\omega_w)$ for all $\alpha\in\mathcal{S}(A)$ and $A(\omega_\mathbb{A})=\{\alpha(\omega_\mathbb{A}):\alpha\in\mathcal{S}(A)\}$. Then in both lines the left equality is obvious. According to Proposition 7 the second equality in the first line is proved by showing
$$\inf_{a\in A(\omega_\mathbb{A})}\tau^B_a(\omega_w)=\underline{\tau}^B(\omega).$$
From the relations $\{x_{\cdot,\alpha}:\alpha\in\mathcal{S}(A)\}\subseteq\mathcal{S}(X)$ and $\tau^B_{x_{\cdot,\alpha}}(\omega)=\tau^B_\alpha(\omega)$ we get the inequality in the second line. □

This means that for processes of the form (10) the first entrance time $\underline{\tau}^B$ can be attained by observing the first entrance times of the special selections $x_{\cdot,a}$ or $x_{\cdot,\alpha}$. This can be useful for the practical calculation of $\underline{\tau}^B$. Unfortunately, there does not seem to be an obvious condition under which the attainability of $\overline{\tau}^B$ holds.

5 Example

In the following we shall give an illuminating example how the concept described in the foregoing sections can be applied to problems from structural mechanics where systems of ODEs of order one and two play an important role.

For the sake of simplicity we consider the so-called Langevin equation
$$dx_t=-a_1x_tdt+a_2dw_t$$
with initial value x_0 where w_t is a one dimensional Wiener process, $a_1>0$ and $a_2\in\mathbb{R}$ ($d=m=1$, $t_0=0$). Its unique solution is the so-called Ornstein-Uhlenbeck process
$$x_t=e^{-a_1t}x_0+a_2\int_0^t e^{-a_1(t-s)}dw_s \qquad (18)$$

which is a Gaussian stochastic process if and only if x_0 is normally distributed or constant. For modelling the uncertainty of the parameters a_1 and a_2 we shall use the following two finite random sets

$$\begin{array}{lll} A_1: & \omega_{\mathbb{A}11} \mapsto [1,3], & P_{\mathbb{A}1}(\omega_{\mathbb{A}11}) = 2/5 \\ & \omega_{\mathbb{A}12} \mapsto [2,4], & P_{\mathbb{A}1}(\omega_{\mathbb{A}12}) = 3/5 \\ A_2: & \omega_{\mathbb{A}21} \mapsto [0.5,1.5], & P_{\mathbb{A}2}(\omega_{\mathbb{A}21}) = 1/3 \\ & \omega_{\mathbb{A}22} \mapsto [1,2], & P_{\mathbb{A}2}(\omega_{\mathbb{A}22}) = 2/3 \end{array}$$

which can be written in the shorter form

$$\begin{aligned} A_1 &= \{([1,3], 2/5), ([2,4], 3/5)\}, \\ A_2 &= \{([0.5,1.5], 1/3), ([1,2], 2/3)\}. \end{aligned}$$

From these random sets we construct the following joint random set on a probability space $\Omega_{\mathbb{A}} = \{\omega_{\mathbb{A}i}\}_{1\le i\le 4}$ with values in $\mathcal{K}'(\mathbb{R}^2)$

$$\begin{aligned} A = \{&([1,3]\times[0.5,1.5], 2/15), ([1,3]\times[1,2], 4/15), \\ &([2,4]\times[0.5,1.5], 1/5), ([2,4]\times[1,2], 2/5)\} \end{aligned}$$

by taking as focal elements the Cartesian products of each focal element of the first with each focal element of the second random set and multiplying the respective weights. This is a kind of independence which is called random set independence (see [3, 8, 9]). According to Equation (10) we get a set-valued process X with values in $\mathcal{K}'(\mathbb{R})$ which can be bounded by the single-valued processes L and U defined by

$$L_t(\omega) = \inf_{x\in X_t(\omega)} x, \quad U_t(\omega) = \sup_{x\in X_t(\omega)} x.$$

Furthermore, we consider the selection

$$\begin{array}{ll} \alpha: & \omega_{\mathbb{A}1} \mapsto (1.7, 1.1), \quad P_{\mathbb{A}}(\omega_{\mathbb{A}1}) = 2/15 \\ & \omega_{\mathbb{A}2} \mapsto (2.3, 1.5), \quad P_{\mathbb{A}}(\omega_{\mathbb{A}2}) = 4/15 \\ & \omega_{\mathbb{A}3} \mapsto (3.0, 0.9), \quad P_{\mathbb{A}}(\omega_{\mathbb{A}3}) = 1/5 \\ & \omega_{\mathbb{A}4} \mapsto (3.2, 1.4), \quad P_{\mathbb{A}}(\omega_{\mathbb{A}4}) = 2/5 \end{array}$$

and the corresponding process ξ defined by (9).

Figure 1 shows details of sample functions of the boundary processes L and U (solid lines) with respect to the same sample function of the Wiener process and the choice $\omega_{\mathbb{A}} = \omega_{\mathbb{A}1}$. The dashed line shows the corresponding sample function of ξ. The graphs were simulated by using the Euler method (see for example [14]) with 1000 time steps from $t_0 = 0$ to $\overline{t} = 10$, $x_0 \equiv 0$. The interval $[1,3]\times[0.5,1.5]$ was discretized by a grid of 101×101 points applying to each of the grid points the Euler scheme and choosing in each time step the greatest value for U and the smallest value for L.

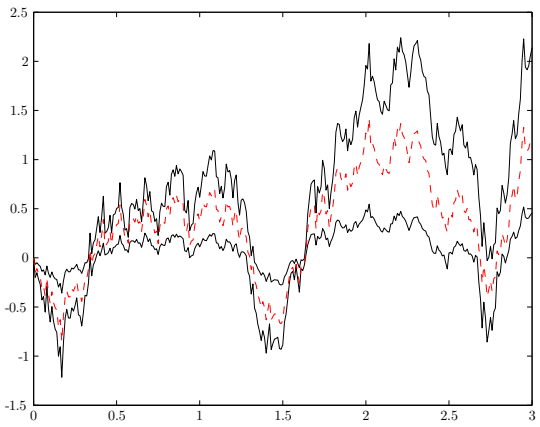


Figure 1: Sample functions of X (boundaries in solid lines) and ξ (dashed line).

Figure 2 shows upper and lower cumulative distribution functions

$$\begin{aligned} \underline{F}_t(x) &= P(X_t \subseteq (-\infty, x)) = P(U_t < x) \\ \overline{F}_t(x) &= P(X_t \cap (-\infty, x) \neq \emptyset) = P(L_t < x) \end{aligned}$$

of the random set X_t at time $t = 10$. They were calculated by simulating 1000 sample functions of the Wiener process and considering all four possible focal elements of A. The dashed line shows the cumulative distribution function $F_{t,\alpha}$ of the random variable ξ_t $(t = 10)$.

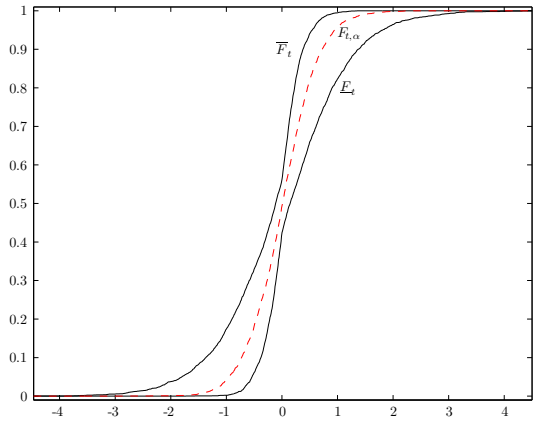


Figure 2: P-box of X_t (solid lines) and cumulative distribution function of ξ_t (dashed line) $(t = 10)$.

Finally, one can consider the first entrance times $\underline{\tau}^B$, τ_α^B and the first inclusion time $\overline{\tau}^B$ for $B = (0.5, \infty)$. The corresponding cumulative distribution functions are displayed in Figure 3.

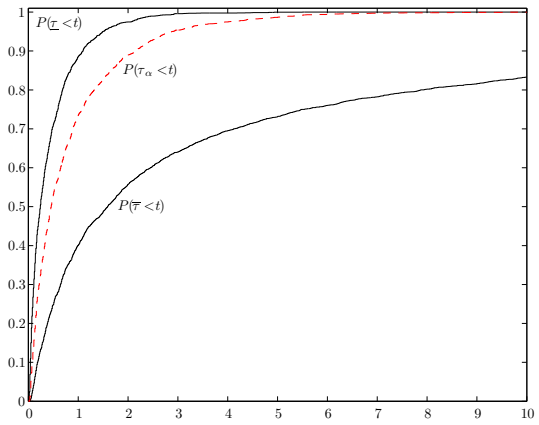


Figure 3: CDFs of first entrance time $\underline{\tau}^B$ and first inclusion time $\overline{\tau}^B$ (solid lines), CDF of first entrance time τ_α^B (dashed line).

6 Summary and Conclusions

In this paper, we consider ordinary stochastic differential equations whose coefficients depend on parameters. Conditions are given under which solution processes continuously depend on these parameters. If this is the case then modelling parameter uncertainty by using random compact sets leads to set-valued processes with compact values which are continuous with respect to the Hausdorff metric. We show that the single-valued solutions of the stochastic differential equation under scrutiny obtained by choosing single parameter values are selections which can be used to represent the set-valued process. Furthermore, analogues of first entrance times for set-valued processes are defined and their attainability by selections is discussed. Finally, an example is given to illustrate the theoretical concept.

As a topic for future research, we plan the investigation of further properties of the set-valued processes of the form (10). Furthermore, this theoretical concept will be applied to engineering problems (from structural mechanics) and it will be explored how first entrance and inclusion times (defined by (14), (15)) can be calculated or simulated.

Acknowledgements

I would like to thank Michael Oberguggenberger for helpful discussions and comments.

References

[1] L. Arnold. *Stochastic Differential Equations: Theory and Applications*. Wiley, 1974

[2] C. Castaing, M. Valadier. *Convex analysis and measurable multifunctions*. Lecture notes in mathematics 580, Springer, 1977.

[3] I. Couso, S. Moral, and P. Walley. Examples of independence for imprecise probabilities. In G. De Cooman, F. G. Cozman, S. Moral, and P. Walley, editors, *ISIPTA '99, Proceedings of the First International Symposium on Imprecise Probabilities and Their Applications, held at the Conference Center "Het Pand" of the Universiteit Gent, Ghent, Belgium, 29 June - 2 July 1999*, pages 121–130, 1999.

[4] H. Cramer, M. R. Leadbetter. *Stationary and Related Stochastic Processes, Sample Function Properties and Their Applications*. Wiley, 1967

[5] J. L. Doob. *Stochastic Processes*. Wiley, 1990.

[6] D. Dubois, M. A. Lubiano, H. Prade, M. A. Gil, P. Grzegorzewski, and O. Hryniewicz, editors. *Soft Methods for Handling Variability and Imprecision, Selected papers from the 4th International Conference on Soft Methods in Probability and Statistics, SMPS 2008, Toulouse, France, September 8-10, 2008*, volume 48 of *Advances in Soft Computing*. Springer, 2008.

[7] E. B. Dynkin. *Markov Processes Vol. 1*. Springer, 1965

[8] Thomas Fetz. Sets of joint probability measures generated by weighted marginal focal sets. In G. De Cooman, T. Fine, and T. Seidenfeld, editors, *ISIPTA '01, Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications, Ithaca, NY, USA*, pages 171–178. Shaker, 2001.

[9] Th. Fetz, M. Oberguggenberger. Propagation of uncertainty through multivariate functions in the framework of sets of probability measures. *Reliability Engineering and System Safety*, 85:73-87, 2004.

[10] A. Friedman. *Stochastic differential equations and applications, Volume 1*. Academic Press, 1975.

[11] I. I. Gikhman, A. V. Skorokhod. *Introduction to the theory of random processes*. Saunders Company, 1969.

[12] I. I. Gikhman, A. V. Skorokhod. *Stochastische Differentialgleichungen*. Akademie-Verlag Berlin, 1971

[13] C. J. Himmelberg. Measurable relations. *Fundamenta Mathematicae*, 87:53–72, 1975.

[14] S. M. Iacus. *Simulation and Inference for Stochastic Differential Equations*. Springer, 2008

[15] Eun Ju Jung, Jai Heui Kim. On Set-Valued Stochastic Integrals. *Stochastic Analysis and Applications*, 21:401-418, 2003

[16] O. Kallenberg. *Foundations of modern probability*. Springer, 1997

[17] M. R. Leadbetter, G. Lindgren, H. Rootzen. *Extremes and Related Properties of Random Sequences and Processes*. Springer, 1983

[18] Shoumei Li, Aihong Ren. Representation theorems, set-valued and fuzzy set-valued Ito integral. *Fuzzy Sets and Systems*, 158:949-962, 2007

[19] Jungang Li, Shoumei Li. Set-Valued Stochastic Lebesgue Integral and Representation Theorems. *International Journal of Computational Intelligence Systems*, 1:177-187, 2008

[20] Jungang Li and Shoumei Li. Strong solution of set-valued stochastic differential equation. In Dubois et al. [6], pages 271–277.

[21] G. Matheron. *Random Sets and Integral Geometry*. Wiley, 1975.

[22] I. Molchanov. *Theory of random sets*. Springer, 2005.

[23] M. Oberguggenberger, W. Fellin. Reliability bounds through random sets: nonparametric methods and geotechnical applications. *Computers & Structures*, 86:1093-1101, 2008

[24] M. Oberguggenberger. The mathematics of uncertainty: models, methods and interpretations. In *Analyzing Uncertainty in Civil Engineering*, W. Fellin, H. Lessman, M. Oberguggenberger, R. Vieider (eds.), Springer Verlag, Berlin, 2005.

[25] M. Oberguggenberger, J. King, B. Schmelzer. Imprecise probability methods for sensitivity analysis in engineering. *In: G. de Cooman, J. Vejnarova, M. Zaffalon (Eds.), ISIPTA'07, Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications. Action M Agency, SIPTA, Prague 2007*, 317-326

[26] M. Oberguggenberger, J. King, B. Schmelzer. Classical and imprecise probability methods for sensitivity analysis in engineering: a case study. *International Journal of Approximate Reasoning*, 50(4):680-693, 2009

[27] Y. Ogura. On stochastic differential equations with set coefficients and the Black-Scholes model. In *Proceedings of the Eighth International Conference on Intelligent Technologies, Sidney*, 300–304, 2007

[28] Y. Ogura. On stochastic differential equations with fuzzy set coefficients. In Dubois et al. [6], pages 263–270.

[29] H. T. Nguyen. *An Introduction to Random Sets*. Chapman and Hall/CRC Press, Boca Raton, 2006

[30] T. T. Soong, M. Grigoriu. *Random Vibration of Mechanical and Structural Systems*. Prentice Hall, 1993

Coefficients of ergodicity for imprecise Markov chains

Damjan Škulj
University of Ljubljana
Faculty of Social Sciences
damjan.skulj@fdv.uni-lj.si

Robert Hable
University of Bayreuth
Department of Mathematics
robert.hable@uni-bayreuth.de

Abstract

Coefficients of ergodicity are an important tool in measuring convergence of Markov chains. We explore possibilities to generalise the concept to imprecise Markov chains. We find that this can be done in at least two different ways, which both have interesting implications in the study of convergence of imprecise Markov chains. Thus we extend the existing definition of the uniform coefficient of ergodicity and define a new so-called weak coefficient of ergodicity. The definition is based on the endowment of a structure of a metric space to the class of imprecise probabilities. We show that this is possible to do in some different ways, which turn out to coincide.

Keywords. Markov chain, imprecise Markov chain, coefficient of ergodicity, lower expectation, upper expectation

1 Introduction

Markov chains are a very popular mathematical model used to describe various dynamical systems. Their properties have been studied in great detail. The modelling of a Markov chain requires estimating a relatively large number of parameters, which is in many practical situations very difficult to achieve precisely. Thus sometimes parameters are estimated with high imprecision, and the theory provides virtually no better answer than regarding the most likely estimates as precise, leading to seemingly precise results that do not reflect the lack of certainty in the input data.

The rapid development of the methods of imprecise probabilities has allowed the study of Markov chains where the imprecision in input data can be incorporated in the results. A detailed study in this topic has been presented by Hartfiel [8] who considered the model where precise initial and transition probability matrices are replaced by sets of possible initial probabilities and transition matrices. This model is known under the name Markov set-chains (see also Hartfiel and Seneta [9]). He pays special attention to the case where the sets can be described using probability intervals. This basically means that every probability of an elementary event is bounded by a lower and upper bound. A similar model was studied from the perspective of the theory of interval probabilities by Kozine and Utkin [11]. The more general interval probabilities based on the Weichselberger's model [20] were involved in the study of Markov chains by Škulj [16, 17]. A more recent approach by de Cooman et al. [2] further generalises the way imprecision is involved into Markov chains, taking an approach based on upper expectation operators. This approach is known from the study of the related field of Markov decision processes used by Satia and Lave [14], followed by [7, 10, 12, 21].

In this paper we follow the approach of de Cooman et al. The topic we study here is the convergence of imprecise Markov chains. The most common result in the classical theory is the Perron-Frobenius theorem that implies unique convergence for the case of regular Markov chains. In [17] the concept of regularity was generalised to imprecise Markov chains and a similar theorem was proved. However, it turns out that weaker conditions than regularity are sufficient to ensure convergence of Markov chains, both in precise and imprecise case. In both cases coefficients of ergodicity prove to be very useful tools. They have been widely used in the precise case (see e.g. Seneta [15]), while Hartfiel [8] generalises them to imprecise Markov chains.

Recently, de Cooman et al. give conditions for convergence of imprecise Markov chains that are substantially weaker than those used by Hartfiel [8], although in the precise case they seem to be very similar. The different generalisations of the conditions for convergence suggest that there may be different possibilities to define coefficients of ergodicity for the case of im-

precise Markov chains. In this paper we show that indeed a generalisation different from the one used by Hartfiel is possible. We also believe, although we have not yet explored this relation, that conditions implied by our new generalised coefficients are closely related to those found by de Cooman et al. The definition of the new coefficient of ergodicity is based on endowing the set of imprecise probabilities with a structure of a metric space.

The paper has the following structure. In the next section we review some theory on lower expectation operators that form a basis for the model of imprecise Markov chains. Further, in Section 3 we explore some possibilities to endow the family of imprecise probabilities with the structure of a metric space, and in Section 4 we describe the model of imprecise Markov chains that we use. Finally, in Section 5 we study the generalisations of coefficients of ergodicity and compare them to the existing generalisations.

2 Lower expectation operators

Let Ω be a finite set of states and let $\mathcal{F}$ be the set of real-valued maps on Ω. Further let $\mathcal{F}_1$ denote the subset of all non-negative real-valued maps with $f(\omega) \leq 1$ for every $\omega \in \Omega$. We denote by 1_Ω, or sometimes just 1, the constant map on Ω such that $f(\omega) = 1$ for all $\omega \in \Omega$. For a pair of maps f and g such that $f(\omega) \geq g(\omega)$ for every $\omega \in \Omega$ we write $f \geq g$, and if at least one of the inequalities is strict we write $f > g$.

The set $\mathcal{F}$ can be equipped with the *maximum norm* given by

$$||f||_\infty = \max_{\omega\in\Omega} |f(\omega)|,$$

which induces the *Chebyshev distance*:

$$d_c(f,g) = \max_{\omega\in\Omega} |f(\omega) - g(\omega)|.$$

We can write $\mathcal{F}_1 = \{f \in \mathcal{F} \mid f \geq 0, ||f||_\infty \leq 1\}$.

We characterise a *probability measure* or a *probability* p as a real valued map on Ω such that

$$\sum_{\omega\in\Omega} p(\omega) = 1$$

and

$$p(\omega) \geq 0 \quad \text{for every } \omega \in \Omega.$$

Therefore $p(A) = \sum_{\omega\in A} p(\omega)$ for every $A \subseteq \Omega$. Thus every probability can be considered to belong to the set $\mathcal{F}_1$. We also consider sets of probabilities, which we usually assume to be closed and convex. Sometimes we assume an enumeration of elements of Ω and for short denote, for instance, $f_i = f(\omega_i)$.

There is a one-to-one correspondence between closed convex sets of probabilities and the corresponding *lower* and *upper expectation operators*. We denote the lower expectation operator of a closed convex set of probabilities $\mathcal{M}$ by $\underline{P}$ and the upper expectation operator by $\overline{P}$. So for any $f \in \mathcal{F}$ we define:

$$\underline{P}(f) = \min_{p\in\mathcal{M}} E_p f \tag{1}$$

and

$$\overline{P}(f) = \max_{p\in\mathcal{M}} E_p f. \tag{2}$$

The min and max in the above equations can be written because of the finiteness of the probability space which assures that all closed sets of probabilities are compact and therefore all minima and maxima exist. In the case of the above correspondence between a set of probabilities and an expectation operator we say that $\mathcal{M}$ is a *credal set* of $\underline{P}$ and we may denote

$$\mathcal{M} = \mathcal{M}(\underline{P}).$$

Every lower expectation operator $\underline{P}$ has the following properties. Let f, f_1, f_2 be arbitrary elements from $\mathcal{A}$. Then:

superadditivity: $\underline{P}(f_1 + f_2) \geq \underline{P}(f_1) + \underline{P}(f_2)$;

non-negative homogeneity: $\underline{P}(\lambda f) = \lambda \underline{P}(f)$ for every $\lambda \geq 0$;

constant additivity: $\underline{P}(f + \mu 1_\Omega) = \underline{P}(f) + \mu$ for every real μ.

Further we note that any expectation operator is completely determined by its values on the space $\mathcal{F}_1$. To see this take any map $f \in \mathcal{F}$ and define the corresponding $\tilde{f} \in \mathcal{F}_1$ with

$$\tilde{f} = \frac{f}{2||f||_\infty} + \frac{1}{2}1_\Omega,$$

if $||f||_\infty > 0$, and $\tilde{f} = \frac{1}{2}1_\Omega$ otherwise. The value $\tilde{a} = \underline{P}(\tilde{f})$ then determines

$$\underline{P}(f) = \left(\tilde{a} - \frac{1}{2}\right) \cdot 2||f||_\infty,$$

as follows from non-negative homogeneity and constant additivity.

3 Distance measures between imprecise probabilities

The set of probability measures on a measurable space $(\Omega, \mathcal{A})$ can be metricised using the following metric:

$$d(p,p') = \max_{A\in\mathcal{A}} |p(A) - p'(A)| = \frac{1}{2}\sum_{\omega\in\Omega} |p(\omega) - p'(\omega)|, \tag{3}$$

for every pair of probability measures p and p'.

Given a metric space M and non-empty compact subsets $X, Y \subset M$ the *Hausdorff metric* (see e.g. [1]) is defined as

$$d_H(X,Y) = \max\left\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\right\}. \tag{4}$$

This metric makes the set of non-empty compact sets a metric space denoted by $F(M)$. Moreover, if M is a compact space, so is $F(M)$. Note also that every compact metric space is complete. The Hausdorff distance can be applied to the family of compact sets of probabilities using the distance function (3) in (4), making it, in the case of a finite space, a complete metric space.

Let $\underline{P}$ and $\underline{P}'$ be lower expectation operators. Then we define the following distance between them:

$$\tilde{d}(\underline{P}, \underline{P}') = \max_{f\in\mathcal{F}_1} |\underline{P}(f) - \underline{P}'(f)|. \tag{5}$$

Because of the finiteness of Ω the max in the above equation exists. If f is any non-negative real map on Ω then we have that $\tilde{f} = \dfrac{f}{||f||_\infty} \in \mathcal{F}_1$. Because of positive homogeneity of lower expectation operator we conclude that

$$|\underline{P}_1(f) - \underline{P}_2(f)| \leq \tilde{d}(\underline{P}_1, \underline{P}_2)||f||_\infty. \tag{6}$$

The next proposition shows that the metrics (5) and (3) coincide for probability measures. Therefore, from now on we denote both distances with d.

Proposition 1. *Let p and p' be probability measures on $(\Omega, \mathcal{A})$. Then we have that*

$$\max_{f\in\mathcal{F}_1} |E_p f - E_{p'} f| = d(p, p').$$

Proof. Define the function

$$F(\omega) = \begin{cases} 1, & p(\omega) \geq p'(\omega); \\ 0, & \text{otherwise.} \end{cases}$$

For any real function $f \in \mathcal{F}_1$ we have

$$\begin{aligned} |E_p f - E_{p'} f| &= \left|\sum_i (p_i - p'_i) f_i\right| \\ &\leq \left|\sum_i (p_i - p'_i) F_i\right| \\ &= \max_{A\subset\Omega} |p(A) - p'(A)| \\ &= d(p, p'). \end{aligned}$$

□

The following theorem shows that the metric (5) between lower expectation operators coincides with the Hausdorff metric between their credal sets. (A similar result can be found in [6], Lemma 6.7.)

Theorem 1. *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be closed convex sets of probabilities and let $\underline{P}_1$ and $\underline{P}_2$ be their lower expectation operators. Then we have that*

$$d(\underline{P}_1, \underline{P}_2) = d_H(\mathcal{M}_1, \mathcal{M}_2). \tag{7}$$

Proof. First we show that for any probabilities p_1 and p_2 we have that

$$\max_{f\in\mathcal{F}_1} |E_{p_1} f - E_{p_2} f| = \max_{f\in\mathcal{F}_1} E_{p_1} f - E_{p_2} f. \tag{8}$$

This follows from the fact that $f \in \mathcal{F}_1$ implies $1_\Omega - f \in \mathcal{F}_1$ and $E_{p_1} f - E_{p_2} f = -(E_{p_1}(1-f) - E_{p_2}(1-f))$ which implies

$$\begin{aligned} \max_{f\in\mathcal{F}_1} |E_{p_1} f - E_{p_2} f| &= \max_{f\in\mathcal{F}_1} \max\{E_{p_1} f - E_{p_2} f, \\ &\qquad E_{p_1}(1-f) - E_{p_2}(1-f)\} \\ &= \max_{f\in\mathcal{F}_1} E_{p_1} f - E_{p_2} f. \end{aligned}$$

The definition of the Hausdorff distance and the equation (8) implies that

$$d_H(\mathcal{M}_1, \mathcal{M}_2) = \max_{p_1\in\mathcal{M}_1} \min_{p_2\in\mathcal{M}_2} \max_{f\in\mathcal{F}_1} E_{p_1} f - E_{p_2} f \tag{9}$$

or in the last expression the roles of $\mathcal{M}_1$ and $\mathcal{M}_2$ can be exchanged, and that case would be treated equally because of symmetry. Now fix any $p_1 \in \mathcal{M}_1$ and consider the map:

$$\Gamma\colon \mathcal{M}_2 \times \mathcal{F}_1 \to \mathbb{R}$$

where

$$(p_2, f) \mapsto E_{p_1} f - E_{p_2} f.$$

Now the set $\mathcal{M}_2$ is compact by definition, and the mapping $p_2 \mapsto \Gamma(p_2, f)$ is continuous and linear, therefore also convex, for any fixed $f \in \mathcal{F}_1$. Furthermore for a fixed p_2 the mapping $f \mapsto \Gamma(p_2, f)$ is also linear, and therefore concave. Now we can use the minimax theorem (see [5]: Theorem 2) to obtain:

$$\min_{p_2\in\mathcal{M}_2} \max_{f\in\mathcal{F}_1} \Gamma(p_2, f) = \max_{f\in\mathcal{F}_1} \min_{p_2\in\mathcal{M}_2} \Gamma(p_2, f).$$

That is

$$\min_{p_2\in\mathcal{M}_2} \max_{f\in\mathcal{F}_1} E_{p_1} f - E_{p_2} f = \max_{f\in\mathcal{F}_1} \min_{p_2\in\mathcal{M}_2} E_{p_1} f - E_{p_2} f.$$

Using the above equality we obtain:

$$\begin{aligned}
&\max_{p_1\in\mathcal{M}_1}\min_{p_2\in\mathcal{M}_2} d(p_1,p_2)\\
&\quad=\max_{p_1\in\mathcal{M}_1}\min_{p_2\in\mathcal{M}_2}\max_{f\in\mathcal{F}_1} E_{p_1}f-E_{p_2}f\\
&\quad=\max_{p_1\in\mathcal{M}_1}\max_{f\in\mathcal{F}_1}\min_{p_2\in\mathcal{M}_2} E_{p_1}f-E_{p_2}f\\
&\quad=\max_{f\in\mathcal{F}_1}\max_{p_1\in\mathcal{M}_1}\min_{p_2\in\mathcal{M}_2} E_{p_1}f-E_{p_2}f\\
&\quad=\max_{f\in\mathcal{F}_1}\overline{P}_1(f)-\overline{P}_2(f)\\
&\quad=\max_{f\in\mathcal{F}_1}\overline{P}_1(1-f)-\overline{P}_2(1-f)\\
&\quad=\max_{f\in\mathcal{F}_1}\underline{P}_2(f)-\underline{P}_1(f).
\end{aligned}$$

Finally, using this and the symmetry between $\mathcal{M}_1$ and $\mathcal{M}_2$, we get

$$\begin{aligned}
d_H(\mathcal{M}_1,\mathcal{M}_2)&=\max\{\max_{p_1\in\mathcal{M}_1}\min_{p_2\in\mathcal{M}_2} d(p_1,p_2),\\
&\qquad\max_{p_2\in\mathcal{M}_2}\min_{p_1\in\mathcal{M}_1} d(p_1,p_2)\}\\
&=\max_{f\in\mathcal{F}_1}\{\underline{P}_2(f)-\underline{P}_1(f),\underline{P}_1(f)-\underline{P}_2(f)\}\\
&=\max_{f\in\mathcal{F}_1}|\underline{P}_1(f)-\underline{P}_2(f)|\\
&=d(\underline{P}_1,\underline{P}_2),
\end{aligned}$$

which completes the proof. □

We will also need the maximal distance between probability measures belonging to a pair of credal sets $\mathcal{M}_1$ and $\mathcal{M}_2$ with the corresponding lower and upper expectation operators $\underline{P}_1,\overline{P}_1$ and $\underline{P}_2,\overline{P}_2$ respectively. Using Proposition 1 we have that

$$\begin{aligned}
\max_{\substack{p_1\in\mathcal{M}_1\\p_2\in\mathcal{M}_2}} d(p_1,p_2)&=\max_{\substack{p_1\in\mathcal{M}_1\\p_2\in\mathcal{M}_2}}\max_{f\in\mathcal{F}_1}|E_{p_1}f-E_{p_2}f|\\
&=\max_{f\in\mathcal{F}_1}\max_{\substack{p_1\in\mathcal{M}_1\\p_2\in\mathcal{M}_2}}|E_{p_1}f-E_{p_2}f|\\
&=\max_{f\in\mathcal{F}_1}\max\{\overline{P}_1(f)-\underline{P}_2(f),\\
&\qquad\overline{P}_2(f)-\underline{P}_1(f)\}.
\end{aligned}$$

However, instead of taking the maxima over the whole $\mathcal{F}_1$ in the above equation it would be enough to only consider characteristic functions of subsets of Ω, as follows from Proposition 1. Therefore

$$\begin{aligned}
\max_{\substack{p_1\in\mathcal{M}_1\\p_2\in\mathcal{M}_2}} d(p_1,p_2)&=\max_{A\subset\Omega}\max\{\overline{P}_1(1_A)-\underline{P}_2(1_A),\\
&\qquad\overline{P}_2(1_A)-\underline{P}_1(1_A)\}.
\end{aligned}$$

It follows that for any pair of lower and upper expectation operators $\underline{P}_1$ and $\overline{P}_2$ we have that

$$\max_{f\in\mathcal{F}_1}\{\overline{P}_2(f)-\underline{P}_1(f)\}=\max_{A\subset\Omega}\{\overline{P}_2(1_A)-\underline{P}_1(1_A)\}. \tag{10}$$

We will also need some results on convergence of lower expectation operators. We study the convergence in the metric (5). In proving the convergence results we will use the result that any decreasing sequence of non-empty compact sets is non-empty (see [4]: Lemma I.5.6).

Proposition 2. *Let $\{\underline{P}_n\}_{n\in\mathbb{N}}$ be an increasing sequence of lower expectation operators and $\{\mathcal{M}_n\}_{n\in\mathbb{N}}$ the sequence of the corresponding credal sets. Then the sequence $\{\mathcal{M}_n\}_{n\in\mathbb{N}}$ is decreasing with respect to set inclusion and the limit*

$$\underline{P}_\infty=\lim_{n\to\infty}\underline{P}_n$$

exists and

$$\mathcal{M}(\underline{P}_\infty)=\bigcap_{n\in\mathbb{N}}\mathcal{M}_n.$$

Moreover, the above credal set is non-empty.

Proof. For every $f\in\mathcal{F}_1$ we have that the sequence $\{\underline{P}_n(f)\}$ is an increasing sequence bounded from above by 1 and is therefore convergent. Now take any $p\in\bigcap_{n\in\mathbb{N}}\mathcal{M}_n$. Then by definition, for every $f\in\mathcal{F}_1$ we have that $E_pf\geq\underline{P}_\infty(f)$, so $\bigcap_{n\in\mathbb{N}}\mathcal{M}_n\subseteq\mathcal{M}(\underline{P}_\infty)$. To see the converse inclusion take any probability p such that $E_pf\geq\underline{P}_\infty(f)\geq\underline{P}_n$ for every $n\in\mathbb{N}$. Therefore $p\in\mathcal{M}_n$ for every $n\in\mathbb{N}$ and every $f\in\mathcal{F}_1$, which implies that $p\in\bigcap_{n\in\mathbb{N}}\mathcal{M}_n$. Thus, $\mathcal{M}(\underline{P}_\infty)\subseteq\bigcap_{n\in\mathbb{N}}\mathcal{M}_n$. As follows from the above remark, the set $\bigcap_{n\in\mathbb{N}}\mathcal{M}_n$ is non-empty. □

Proposition 3. *Let $\{\underline{P}_n\}_{n\in\mathbb{N}}$ be any convergent sequence of lower expectation operators and $\{\mathcal{M}_n\}_{n\in\mathbb{N}}$ the sequence of the corresponding credal sets. Then the set*

$$\mathcal{M}_\infty=\bigcap_{n\in\mathbb{N}}\mathrm{co}\Big(\overline{\bigcup_{m\geq n}\mathcal{M}_m}\Big),$$

where co *denotes the convex hull, is the credal set of the limit lower expectation operator $\underline{P}_\infty=\lim_{n\to\infty}\underline{P}_n$. Moreover, the set $\mathcal{M}_\infty$ is non-empty and therefore the lower expectation operator $\underline{P}_\infty$ is well defined.*

Proof. First we define the following sequence of lower expectation operators:

$$\tilde{\underline{P}}_n=\inf_{m\geq n}\underline{P}_m.$$

Clearly, the convergence of the sequence $\{\underline{P}_n\}$ implies the convergence of $\{\tilde{\underline{P}}_n\}$ with the same limit. We only need to see that the credal set of $\tilde{\underline{P}}_n$ is $\mathrm{co}(\overline{\bigcup_{m\geq n}\mathcal{M}_m})$.

To see this take any convergent sequence $\{p_r\}$ in $\bigcup_{m\geq n}\mathcal{M}_m$. For every $f\in\mathcal{F}$ we have that $E_{p_r}f\geq\tilde{\underline{P}}_n(f)$ and therefore $\lim_{r\to\infty}E_{p_r}f=E_{\lim_{r\to\infty}p_r}f\geq$

$\underline{\tilde{P}}_n(f)$, and thus $\lim_{r\to\infty} p_r$ belongs to the credal set of $\underline{\tilde{P}}_n(f)$. Further, given any $f \in \mathcal{F}_1$ there is some $p_r \in \mathcal{M}_m$, for $m \geq n$ so that $E_{p_r} f \leq \underline{\tilde{P}}_n(f) + \frac{1}{r}$. Since the set of all probabilities on a finite set is compact, the sequence $\{p_r\}$ has a convergent subsequence converging to a probability p and $E_p f = \underline{\tilde{P}}_n(f)$. Thus, $\underline{\tilde{P}}_n$ is the lower expectation operator of the set $\overline{\bigcup_{m\geq n} \mathcal{M}_m}$ which implies that its closure is the credal set of $\underline{\tilde{P}}_n$.

To finish the proof we apply Proposition 2 to the increasing sequence $\{\underline{\tilde{P}}_n\}$ and the corresponding credal sets $\mathrm{co}(\bigcup_{m\geq n} \mathcal{M}_n)$. □

Corollary 1. *The set of all lower expectation operators is complete in the metric* (5).

4 Imprecise Markov chains

One of the most natural ways to involve imprecision in a probabilistic model is to allow a set of possible probability distributions instead of a single one. In the case of Markov chains such sets can be allowed in place of transition probabilities as well as initial probability distributions. Additionally, we usually assume such sets are closed and convex. This assumption is particularly useful because, as described in Section 2, the sets can be equivalently described using lower or upper expectation operators. There are of course many models that allow description of sets of probabilities, such as *interval probabilities* (see e.g. [20]) or *lower* and *upper previsions* (see e.g. [18, 19]).

The most basic form used in most of the approaches taken until now is to put constraints, usually in the form of intervals, on the probabilities belonging to the elementary sets (see [8], [11]). The imprecision concerning the initial distribution is thus presented through the intervals $[p_i, q_i]$ which are supposed to contain the unknown initial probability $P(X_0 = i)$. Similarly, the probabilities of transition from the state i to j are given in the form of intervals $[p_{ij}, q_{ij}]$ supposed to contain the unknown true transition probability $P(X_{n+1} = j | X_n = i)$. Even though the true probabilities are unknown, it is certain that the sum of all probabilities is 1. Thus the values within the intervals must be taken so that they sum to 1, or in the case of transition interval matrices, all rows must sum to 1. An additional assumption that is usually made about the intervals is that all values within the interval are reachable or, in particular, that the interval bounds are reachable. In the common terminology of imprecise probabilities this requirement is named *coherence*. To each set of intervals, the set of probabilities assuming their values within those intervals can be assigned.

One of the crucial differences between precise and imprecise probabilities is that a precise probability can be fully determined by far less information than an imprecise probability. Thus to determine any precise probability, only its values on elementary sets are needed to be found, while the sets of probabilities able to be represented via simple intervals described above is fairly limited. (Many examples can be found e.g. in [20], [19], [18].) Another difference compared to the classical model is that transition probabilities that govern transitions of a Markov chain in the imprecise case may change in time. Thus, we are dealing with possibly non-homogeneous chains, which consequently require considering non-homogeneous matrix products.

Now we introduce the terminology used to describe imprecise Markov chains. We will assume a nonempty set Ω whose elements are called *states*. For simplicity we will assume they are the consecutive integers $1, \ldots, m$, since in the basic model their values have no special consequences. We will follow the approach similar to the one taken by de Cooman et al. [2] to describe the sets of probabilities using the corresponding expectation operators, usually this will mean lower expectation operators.

We will thus assume a set $\mathcal{M}_0$ of *initial probability distributions* and let $\underline{P}_0$ be its lower expectation operator (see (1)). Further, we assume a set of transition matrices $\mathcal{P}$, whose rows are *separately specified*, i.e. for any two transition matrices p and p' with rows p_i and p'_i replacing the ith row of p with p'_i results in a matrix that still belongs to $\mathcal{P}$. By adopting this property we can associate row sets of distributions $\mathcal{P}_i$ to $\mathcal{P}$ so that any independent choice of rows from the row sets gives a transition matrix in $\mathcal{P}$. If additionally we assume that row sets are closed and convex, we have the following important property.

Lemma 1. *Let $\mathcal{P}$ be a convex set of transition matrices with separtely specified rows and let $\mathcal{M}$ be a convex set of probabilities. Then the set of probability distributions at the next step $\mathcal{M} \cdot \mathcal{P}$ is a convex set.*

We slightly modify the proof of [8]: Lemma 2.5.

Proof. We prove the lemma by showing that given the probabilities q and $q' \in \mathcal{M}$ and transition matrices p and $p' \in \mathcal{P}$ then, whenever $\alpha, \beta \geq 0$ and $\alpha + \beta = 1$,

$$(\alpha q \cdot p + \beta q' \cdot p') = (\alpha q + \beta q')r \tag{11}$$

with $r \in \mathcal{P}$.

Take $j \in \Omega$. We have

$$\begin{aligned}(\alpha q \cdot p + \beta q' \cdot p')_j &= \alpha \sum_{i=1}^{m} q_i p_{ij} + \beta \sum_{i=1}^{m} q'_i p'_{ij} \\ &= \sum_{i=1}^{m} (\alpha q_i p_{ij} + \beta q'_i p'_{ij}) \\ &= \sum_{i=1}^{m} (\alpha q_i + \beta q'_i) \left(\frac{\alpha q_i}{\alpha q_i + \beta q'_i} p_{ij} + \frac{\beta q'_i}{\alpha q_i + \beta q'_i} p'_{ij} \right).\end{aligned}$$

Thus taking r with $r_{ij} = \frac{\alpha q_i}{\alpha q_i + \beta q'_i} p_{ij} + \frac{\beta q'_i}{\alpha q_i + \beta q'_i} p'_{ij}$ satisfies (11). Notice that ith row of r is a convex combination of some elements of $\mathcal{P}_i$ and therefore itself a member of $\mathcal{P}_i$ too. Now, because rows are separately specified the resulting matrix is also a member of $\mathcal{P}$. □

To each row set of probabilities we associate the lower expectation operator $\underline{T}_i$. Let $\underline{T}$ then be the matrix lower expectation operator whose ith row is $\underline{T}_i$. We will say that the set $\mathcal{P}$ is the *credal set* of $\underline{T}$.

Let $X_0, X_1, \ldots, X_n, \ldots$ be a sequence of random variables assuming the values in Ω. According to the given assumptions we have

$$P(X_0 = i) = q_i^0,$$

where $q^0 \in \mathcal{M}_0$. The role of the transition matrices is given by

$$P(X_{n+1} = j | X_n = i) = p_{ij}^n,$$

where $p^n \in \mathcal{P}$.

A basic feature of the theory of Markov chains is the ability to calculate the probability of being in some state j at time n given an initial probability. Of course, since the initial and transition probabilities are imprecise, the answer will also be given in the form of an imprecise probability, that is, in the form of a set of probabilities. Previous works such as Hartfiel's [8] provide the general answer to this question based on the classical theory. The set of possible probability distributions at step n is equal to the set of all possible initial distributions multiplied by all possible sequences of transition matrices. Let $\mathcal{M}_n$ denote the set of possible probability distributions at step n given the initial distribution $\mathcal{M}_0$. Then we have

$$\begin{aligned}\mathcal{M}_n = \{q \cdot p_1 \cdot \ldots \cdot p_n \mid q \in \mathcal{M}_0, p_i \in \mathcal{P} \\ \text{for every } i = 1, \ldots, n\} = \mathcal{M}_{n-1} \cdot \mathcal{P}. \quad (12)\end{aligned}$$

It follows from Lemma 1 that in the case where the set of transition matrices $\mathcal{P}$ has closed convex separately specified row sets, every $\mathcal{M}_n$ is also a closed convex set of probabilities. Therefore, they can be equivalently represented using lower expectation operators. The lower expectation operator corresponding to the set $\mathcal{M}_n$ is denoted by $\underline{P}_n$.

To calculate the values of $\underline{P}_n$ on real functions on Ω we follow the approach proposed in [2]. They first calculate the nth power of the transition operator $\underline{T}$ using so-called backwards recursion. This method can be described in the following way. Let f be any real valued map on Ω. Every expectation operator assigns to it a real number corresponding to the lower expectation. In particular, every row lower expectation operator $\underline{T}_i$ assigns to it the value $\underline{T}_i(f)$. A transition operator $\underline{T}$ thus assign to every f a vector of values

$$\underline{T}(f) = \begin{pmatrix} \underline{T}_1(f) \\ \underline{T}_2(f) \\ \vdots \\ \underline{T}_m(f) \end{pmatrix}. \quad (13)$$

Now $\underline{T}(f)$ is another real valued function on Ω to which a new instance of T can be applied to obtain $\underline{T}^2(f)$ and so on. Finally, applying $\underline{P}_0$ to $T^n(f)$ gives exactly the lower expectation of the lower expectation operator $\underline{P}_n$ corresponding to the set $\mathcal{M}_n$. For the proof see [2].

Once probabilities of states on different steps are calculated, we are often interested in the limiting behaviour of these probabilities. Thus, the question is what can be said about the probability $P(X_n = i)$ for a large n and how does it depend on the initial distribution? In the classical theory, Perron-Frobenius theorem assures convergence for the class of regular Markov chains (a Markov chain with the transition matrix p is *regular* if for some positive integer r the power p^r has only strictly positive entries). The Perron-Frobenius theorem states that the probabilities $p_i^{(n)} = P(X_n = i)$ converge to some unique limit probabilities independently on the initial distribution.

Regularity is therefore a sufficient condition for unique convergence of a Markov chain, but not also a necessary one. This is true already in the case of precise Markov chains, where a more general criteria are derived using *coefficients of ergodicity* that besides telling whether a chain is convergent also measure the rate of convergence (see e.g. Seneta [15]). Hartfiel [8] then applies a generalised coefficient of ergodicity to study the convergence of Markov set-chains. Recently, de Cooman et al. [2] find that the conditions applied by Hartfiel are in general too strong to assure the convergence of imprecise Markov chains. They define a class of *regularly absorbing* imprecise Markov chains, based on the accessibility between states, for which they show unique convergence.

5 Coefficients of ergodicity

Coefficients of ergodicity or *contraction coefficients* measure the rate of convergence of Markov chains. Seneta in his paper [15] defines a general coefficient of ergodicity for a stochastic matrix p with no zero columns to be

$$\tau(p) = \sup_{x,y} \frac{d(xp, yp)}{d(x, y)}$$

where d is some metric on the set of vectors with positive coordinates and whose components sum to 1 and x, y are such vectors. The value of $\tau(p)$ is between 0 and 1 and further τ has the following properties:

(i) $\tau(p_1 p_2) \leq \tau(p_1)\tau(p_2)$ for every pair of stochastic matrices with no zero columns p_1 and p_2;

(ii) $\tau(p) = 0$ whenever rank of p is 1 i.e. $p = \mathbf{1}v$ for some vector v.

Depending on the metrics, different coefficients of ergodicity are used. In this paper we are concerned with the coefficient generated by the metric (3). This coefficient was introduced by Dobrushin [3] and its direct evaluation is derived by Paz [13]:

$$\tau(p) = \frac{1}{2} \max_{i,j} \sum_{s=1}^{m} |p_{is} - p_{js}|.$$

In view of (3), the above can be stated as

$$\tau(p) = \max_{i,j} d(p_i, p_j). \tag{14}$$

where p_i and p_j denote the ith and jth row of p respectively.

For the case of imprecise Markov chains, Hartfiel [8] extends the concept of a coefficient of ergodicity to Markov chains where sets of transition probabilities are considered. For a set of transition matrices $\mathcal{P}$ he defines the *uniform coefficient of ergodicity* as

$$\tau(\mathcal{P}) = \sup_{p \in \mathcal{P}} \tau(p).$$

If $\mathcal{P}$ is an interval $[P, Q]$, i.e. $\mathcal{P} = \{p \mid p \text{ is a stochastic matrix such that } P \leq p \leq Q\}$, then he finds that

$$\tau(\mathcal{P}) \leq \frac{1}{2} \max_{i,j} \sum_{k=1}^{m} \max\{|q_{ik} - p_{jk}|, |q_{jk} - p_{ik}|\}.$$

where p_{ik} and q_{ik} are the components of P and Q respectively.

In our setting of lower and upper expectation operators, the calculation of the uniform coefficient of ergodicity is given by the following proposition.

Proposition 4. *Let $\mathcal{P}$ be a set of transition matrices and let $\underline{T}$ and $\overline{T}$ be its lower and upper expectation matrices. Then we have that*

$$\begin{aligned}\tau(\mathcal{P}) &= \max_{i,j} \max_{f \in \mathcal{F}_1} \overline{T}_i(f) - \underline{T}_j(f) \\ &= \max_{i,j} \max_{A \subset \Omega} \overline{T}_i(1_A) - \underline{T}_j(1_A).\end{aligned}$$

Proof. The second equality follows from (10). Let $p \in \mathcal{P}$ be arbitrary transition matrix. Then its ith and jth row are arbitrary probability distributions belonging to the credal sets of ith and jth row of $\mathcal{P}$. We have that

$$\begin{aligned}\tau(\mathcal{P}) &= \max_{p \in \mathcal{P}} \tau(p) \\ &= \max_{i,j} \max_{\substack{p_i \in \mathcal{M}_i \\ p_j \in \mathcal{M}_j}} d(p_i, p_j) \\ &= \max_{i,j} \max_{A \subset \Omega} \max\{\overline{T}_i(1_A) - \underline{T}_j(1_A), \\ &\qquad \overline{T}_j(1_A) - \underline{T}_i(1_A)\} \\ &= \max_{i,j} \max_{A \subset \Omega} \overline{T}_i(1_A) - \underline{T}_j(1_A),\end{aligned}$$

as required. □

Thus, we may define $\tau(\underline{T}) = \tau(\mathcal{M}(\underline{T}))$.

The uniform coefficient of ergodicity can be used as a contraction measure for a set of transition matrices. The following theorem holds ([8]: Theorem 3.3):

Theorem 2. *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be non-empty compact sets of probabilities. Then*

$$d_H(\mathcal{M}_1 \cdot \mathcal{P}, \mathcal{M}_2 \cdot \mathcal{P}) \leq \tau(\mathcal{P}) d_H(\mathcal{M}_1, \mathcal{M}_2).$$

A stochastic matrix p whose coefficient of ergodicity $\tau(p)$ is strictly smaller than 1 is called *scrambling* (see [15]). Further if $\mathcal{P}$ is a set of transition matrices such that $\tau(p_1 \cdot p_2 \cdots p_r) < 1$ for any matrices $p_i \in \mathcal{P}$ then such a set is called *product scrambling* (see [8]), and r is then called its *scrambling integer*. Thus we have that $\tau(\mathcal{P}^r) < 1$. Something very similar can be said about lower expectation matrices. We will say that a lower expectation matrix $\underline{T}$ is scrambling whenever $\tau(\underline{T}) < 1$ and if instead only $\tau(\underline{T}^r) < 1$ we will say that it is product scrambling with scrambling integer r.

Theorem 2 implies the following more general corollary ([8]: Theorem 3.4):

Corollary 2. *Let $\mathcal{P}$ be be product scrambling with scrambling integer r and let $\mathcal{M}_0$ be a non-empty compact set of probabilities. Then, for any positive integer h,*

$$d_H(\mathcal{M}_0 \mathcal{P}^h, \mathcal{M}_\infty) \leq K \beta^h$$

where $K = \tau(\mathcal{P}^r)^{-1} d_H(\mathcal{M}_0, \mathcal{M}_\infty)$ *and* $\beta = \tau(\mathcal{P}^r)^{\frac{1}{r}} < 1$ *and* $\mathcal{M}_\infty$ *is the unique compact set of probabilities such that*

$$\mathcal{M}_\infty \mathcal{P} = \mathcal{M}_\infty.$$

Thus,

$$\lim_{h\to\infty} \mathcal{M}_0 \mathcal{P}^h = \mathcal{M}_\infty.$$

Theorem 2 implies the convergence of a Markov set-chain in the Hausdorff metric. Moreover, if $\tau(\mathcal{P}) < 1$ for a set of transition matrices then given any initial probability distribution q_0 and a sequence of transition matrices $\{p_i\}_{i\in\mathbb{N}}$ such that every $p_i \in \mathcal{P}$ we have that the sequence $q_n = q_0 p_1 \cdots p_n$ converges to some p_∞. This is a consequence of the fact that $\tau(p_1 \cdots p_n) \to 0$ as n tends to infinity. Moreover, since clearly $\tau(\mathcal{P}') \leq \tau(\mathcal{P})$ for every $\mathcal{P}' \subseteq \mathcal{P}$, it follows that given a convergent Markov chain with the set of transition probabilities $\mathcal{P}$ then a Markov chain with the set of transition probabilities $\mathcal{P}'$ is also convergent.

De Cooman et al. [2] show that it not necessary to require that every possible transition matrix is a contraction, but instead, what is needed is only that the corresponding upper (or lower) expectations are becoming more and more similar. As a simple demonstration consider the following example.

Example 1. Let a set of transition matrices on the set $\Omega = \{1, 2\}$ be given by the following lower and upper transition matrix

$$P = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \qquad \text{and} \quad Q = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Clearly this set contains the matrix

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

which is not contractive. However, given any initial set of distributions the Markov chain with the above set of transition matrices converges to the set of all probability distributions on Ω.

De Cooman et al. further find sufficient conditions for unique convergence by studying the accessibility relation between states. Our aim here is to find a coefficient of ergodicity that would describe this type of convergence for imprecise Markov chains. We implement the following idea. Given a lower transition matrix $\underline{T}$, the backwards recursion allows the calculation of its powers $\underline{T}^n$ for every positive integer n. In the case of a precise transition matrix, the rows of its consequent powers get more and more similar, which is measured by the coefficient of ergodicity (14). In the case of a lower expectation matrix, the same effect will be achieved by measuring the distances between the row lower expectation operators corresponding to the powers of $\underline{T}$.

Definition 1. Let $\underline{T}$ be a transition lower expectation matrix. Then we define the *weak coefficient of ergodicity* as

$$\rho(\underline{T}) = \max_{\substack{f\in\mathcal{F}_1 \\ i,j}} |\underline{T}_i(f) - \underline{T}_j(f)|,$$

where $\underline{T}_i$ and $\underline{T}_j$ are ith and jth row lower expectation operators respectively.

The following proposition is an immediate consequence of the definitions.

Proposition 5. *Let* $\underline{T}$ *be a transition lower expectation matrix with rows* $\underline{T}_i$*. Then:*

$$\rho(\underline{T}) = \max_{i,j} d(\underline{T}_i, \underline{T}_j).$$

Proposition 6. *Let* $\underline{P}_1$ *and* $\underline{P}_2$ *be lower expectation operators and* $\underline{T}$ *a transition lower expectation matrix. Then we have that*

$$d(\underline{P}_1\underline{T}, \underline{P}_2\underline{T}) \leq \rho(\underline{T}) d(\underline{P}_1, \underline{P}_2).$$

Proof. Denote $c_f = \underline{T}(f)$ (see (13)) and let $\underline{c}_f$ and $\overline{c}_f$ be its minimal and maximal element respectively. Further let $\underline{\tilde{P}_1} = \underline{P}_1\underline{T}$ and $\underline{\tilde{P}_2} = \underline{P}_2\underline{T}$. Then using constant additivity and (6) we obtain

$$\begin{aligned} |\underline{\tilde{P}_1}(f) - \underline{\tilde{P}_2}(f)| &= |\underline{P_1}(c_f) - \underline{P_2}(c_f)| \\ &= |\underline{P}_1((c_f - \underline{c}_f) + \underline{c}_f) \\ &\qquad - \underline{P}_2((c_f - \underline{c}_f) + \underline{c}_f)| \\ &\leq d(\underline{P}_1, \underline{P}_2)||c_f - \underline{c}_f||_\infty \\ &= d(\underline{P}_1, \underline{P}_2)(\overline{c}_f - \underline{c}_f) \\ &\leq d(\underline{P}_1, \underline{P}_2)\rho(\underline{T}) \end{aligned}$$

□

Corollary 3. *Let* $\underline{R}$ *and* $\underline{S}$ *be any transition lower expectation matrices. Then:*

$$\rho(\underline{RS}) \leq \rho(\underline{R})\rho(\underline{S}).$$

Proof. Denote $\underline{T} = \underline{RS}$ and let $\underline{T}_i$ and $\underline{T}_j$ be the ith and jth row lower expectation operators. We have that, for instance,

$$\underline{T}_i(f) = \underline{R}_i\underline{S}(f).$$

Proposition 6 then yields

$$\begin{aligned} |\underline{T}_i(f) - \underline{T}_j(f)| &= |\underline{R}_i\underline{S}(f), \underline{R}_j\underline{S}(f)| \\ &\leq d(\underline{R}_i, \underline{R}_j)\rho(\underline{S}) \\ &\leq \rho(\underline{R})\rho(\underline{S}), \end{aligned}$$

as required. □

The next corollary is now immediate.

Corollary 4. *For any lower expectation operator $\underline{T}$ we have that*

$$\rho(\underline{T}^n) \leq \rho(\underline{T})^n.$$

Thus, it may happen that even if $\rho(\underline{T}) = 1$ it may be that $\rho(\underline{T}^n) < 1$.

The following proposition shows that the credal set of a contractive lower expectation operator contains at least one contractive transition matrix. The converse does not hold, as demonstrated by the example following the proposition.

Proposition 7. *Let $\underline{T}$ be a transition lower expectation matrix such that $\rho(\underline{T}) < 1$. Then there exists a precise transition matrix $p \in \mathcal{M}(\underline{T})$ such that $\tau(p) < 1$.*

Proof. Denote $\rho := \rho(\underline{T})$. Then for any pair of indices i and j we have $d(\underline{T}_i, \underline{T}_j) \leq \rho$. Coherence of $\underline{T}$ implies that for every set $A \subset \Omega$ we have a probability measure p^A such that $p_i^A(A) = \underline{T}(1_A)$ for every $1 \leq i \leq m$. Then $|p_i^A(A) - p_j^A(A)| < 1$ and $|p_i^A(A') - p_j^A(A')| \leq 1$ for any $A' \subset \Omega$. Let $\lambda_A > 0$ for every $A \subset \Omega$ and let $\sum_{A\subset\Omega} \lambda_A = 1$. Let $p = \sum_{A\subset\Omega} \lambda_A p^A$. Clearly then $p_i(A) - p_j(A) < 1$ for every $A \subset \Omega$ and thus $\tau(p) < 1$. □

Example 2. Let the lower expectation operator $\underline{T} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ be given. Thus the credal set of $\underline{T}$ contains all possible stochastic matrices with the first row equal to $(1, 0)$. Clearly, the weak coefficient of ergodicity of $\underline{T} = \underline{T}^n$, for every $n \in \mathbb{N}$, is equal to 1; however, the credal set contains, for instance, the matrix $p = \begin{pmatrix} 1 & 0 \\ 0.5 & 0.5 \end{pmatrix}$, whose coefficient of ergodicity is equal to 0.5.

Proposition 8. *Let $\underline{T}$ be a transition lower expectation matrix such that $\rho(\underline{T}) < 1$. Then there exists a lower expectation operator $\underline{P}_\infty$ satisfying the property:*

$$\underline{P}_\infty \underline{T} = \underline{P}_\infty. \tag{15}$$

We will call a lower expectation operator satisfying the property (15) an *invariant lower expectation operator* for a transition lower expectation matrix $\underline{T}$.

Proof. Consider the sequence $\underline{P}_n = \underline{P}_0 \underline{T}^n$. We will show that it is a Cauchy sequence in the metric (5). To see this, take some positive integers m and n with $m > n$. Using the fact that $d(\underline{P}, \underline{P}') \leq 1$ for any pair of expectation operators, we have that

$$\begin{aligned} d(\underline{P}_n, \underline{P}_m) &= d(\underline{P}_0 \underline{T}^n, \underline{P}_0 \underline{T}^m) \\ &= d(\underline{P}_0 \underline{T}^n, \underline{P}_0 \underline{T}^{m-n} \underline{T}^n) \\ &\leq d(\underline{P}_0, \underline{P}_0 \underline{T}^{m-n}) \rho(\underline{T}^n) \\ &\leq \rho(\underline{T}^n) \\ &\leq \rho(\underline{T})^n, \end{aligned}$$

and since $\rho(\underline{T}) < 1$ it follows that, with n large enough, this distance can be arbitrarily small. Because of the completeness of the set of lower expectation operators (Corollary (1)), the sequence converges to some lower expectation operator $\underline{P}_\infty$. □

Clearly the invariant lower operators of $\underline{T}$ is the same as the one for $\underline{T}^n$, and thus the above result also holds for a transition lower expectation matrix $\underline{T}$ such that $\rho(\underline{T}^n) < 1$.

Theorem 3. *Let $\underline{T}$ be a transition lower expectation matrix with $\rho(\underline{T}) < 1$ and $\underline{P}_0$ an initial lower expectation operator and $\underline{P}_\infty$ the invariant lower expectation operator for $\underline{T}$. Then*

$$d(\underline{P}_0 \underline{T}^n, \underline{P}_\infty) \leq d(\underline{P}_0, \underline{P}_\infty) \rho(\underline{T})^n.$$

Therefore,

$$\lim_{n\to\infty} \underline{P}_0 \underline{T}^n = \underline{P}_\infty$$

independently of $\underline{P}_0$, and $\underline{P}_\infty$ is thus the unique invariant lower expectation operator for $\underline{T}$.

Proof. Using (15) and Proposition 6 and Corollary 4 we obtain

$$\begin{aligned} d(\underline{P}_0 \underline{T}^n, \underline{P}_\infty) &= d(\underline{P}_0 \underline{T}^n, \underline{P}_\infty \underline{T}^n) \\ &\leq d(\underline{P}_0, \underline{P}_\infty) \rho(\underline{T})^n. \end{aligned}$$

Now since $\rho(\underline{T}) < 1$ the right hand side converges to 0. □

A corollary analogous to Corollary 2 of the last theorem can also be stated. We extend the notion of scrambling lower expectation matrices to the case where the weak coefficient of ergodicity is used. We will say a lower expectation matrix $\underline{T}$ is weakly scrambling if $\rho(\underline{T}) < 1$ and if $\rho(\underline{T}) = 1$ but $\rho(\underline{T}^r) < 1$ for some positive integer r that it is *weakly product scrambling* with *scrambling integer* r.

Corollary 5. *Let $\underline{T}$ be weakly product scrambling with scrambling integer r and let $\underline{P}_0$ be a lower expectation operator. Then, for any positive integer h,*

$$d(\underline{P}_0 \underline{T}^h, \underline{P}_\infty) \leq K \beta^h$$

where $K = \rho(\underline{T}^r)^{-1} d(\underline{P}_0, \underline{P}_\infty)$ and $\beta = \rho(\underline{T}^r)^{\frac{1}{r}}$. Thus,

$$\lim_{k\to\infty} \underline{P}_0 \underline{T}^k = \underline{P}_\infty.$$

The type of convergence measured by the weak coefficient of ergodicity is clearly closely related to that described in [2]. This suggests that regularly absorbing and weakly scrambling lower expectation matrices are closely related, if not identical. One of the directions in our future research is therefore to clarify this relation.

Acknowledgement

We thank the referees for their helpful suggestions that helped us to improve our paper.

References

[1] G. Beer. *Topologies on closed and closed convex sets.* Kluwer Academic Publishers, Dordrecht, 1993.

[2] G. de Cooman, F. Hermans, and E. Quaeghebeur. Imprecise Markov chains and their limit behaviour. *Probability in the Engineering and Informational Sciences*, 2009.

[3] R.L. Dobrushin. Central limit theorem for nonstationary Markov chains, I, II. *Theory of Probability and its Applications*, 1(4):329–383, 1956.

[4] N. Dunford and J.T. Schwartz. *Linear Operators. Part I: General Theory.* Wiley, New York, 1988.

[5] Ky Fan. Minimax theorems. *Proceedings of the National Academy of Sciences of the United States of America*, 39:42–47, 1953.

[6] R. Hable. *Data-based decisions under complex uncertainty.* PhD thesis, Ludwig-Maximilians-Universität (LMU) Munich, 2009.

[7] D. Harmanec. Generalizing Markov decision processes to imprecise probabilities. *Journal of Statistical Planning and Inference*, 105:199–213, 2002.

[8] D.J. Hartfiel. *Markov Set-Chains.* Springer-Verlag, Berlin, 1998.

[9] D.J. Hartfiel and E. Seneta. On the theory of Markov set-chains. *Advances in Applied Probability*, 26(4):947–964, 1994.

[10] H. Itoh and K. Nakamura. Partially observable Markov decision processes with imprecise parameters. *Artificial Intelligence*, 171(8–9):453–490, 2007.

[11] I. Kozine and L.V. Utkin. Interval-valued finite Markov chains. *Reliable Computing*, 8(2):97–113, 2002.

[12] A. Nilim and L. El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53:780–798, 2005.

[13] A. Paz. Ergodic theorems for infinite probabilistic tables. *Annals of Mathematical Statistics*, 41(2):539–550, 1970.

[14] J.K. Satia and R.E. Lave. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, 1973.

[15] E. Seneta. Coefficients of ergodicity - structure and applications. *Advances in Applied Probability*, 11(2):270–271, 1979.

[16] D. Škulj. Finite discrete time Markov chains with interval probabilities. In J. Lawry, E. Miranda, A. Bugarín, S. Li, M. A. Gil, P. Grzegorzewski, and O. Hryniewicz, editors, *SMPS*, volume 37 of *Advances in Soft Computing*, pages 299–306. Springer, 2006.

[17] D. Škulj. Regular finite Markov chains with interval probabilities. In G. de Cooman, M. Zaffalon, and J. Vejnarová, editors, *ISIPTA'07 - Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications*, pages 405–413. SIPTA, 2007.

[18] P. Walley. *Statistical reasoning with imprecise probabilities.* Chapman and Hall, London, New York, 1991.

[19] P. Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24:125–148, 2000.

[20] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung. I: Intervallwahrscheinlichkeit als umfassendes Konzept.* Physica-Verlag, Heidelberg, 2001.

[21] C.C. White and H.K. Eldeib. Markov decision processes with imprecise transition probabilities. *Operations Research*, 42(4):739–749, 1994.

Buying and Selling Prices under Risk, Ambiguity and Conflict

Michael Smithson
The Australian National University
michael.smithson@anu.edu.au

Paul D. Campbell
Australian Bureau of Statistics
paul.campbell@abs.gov.au

Abstract

This paper reports an empirical study of buying and selling prices for three kinds of gambles: Risky (with known probabilities), ambiguous (with lower and upper probabilities), and conflictive (with disagreeing probability assessments). The latter two types of gambles were constructed so that the variances in their probabilities were approximately equal, thereby ensuring that uncertainty type was not confounded with variance. Participants devaluated both ambiguous and conflictive gambles relative to risky gambles with equivalent expected utilities, but the ambiguous and conflictive valuation means did not significantly differ. Moreover, the endowment effect (the gap between buying and selling prices) was exaggerated for these two kinds of gambles in comparison with risky gambles. Conflictive gambles also were found to be devalued less than ambiguous gambles, relative to their risky counterparts. Several self-report measures of attitudes towards uncertainty and risk were included as potential predictors of pricing. The most effective predictors were a measure of instrumental risk orientation and a functional impulsivity scale. Instrumental risk positively predicted valuation of ambiguous and conflictive gambles but not of risky gambles. Functional impulsivity positively predicted valuation of risky gambles but neither of the other two kinds. No individual differences measures predicted relative devaluation.

Keywords. Ambiguity, conflict, prices, risk aversion, buying, selling.

1 Introduction

1.1 Preferences for Risk, Ambiguity and Conflict

The subject of this paper is the valuation of uncertain prospects when the uncertainty is not limited to known probabilities. We investigate two kinds of imprecise probabilities. Numerous studies since Ellsberg's [1] classic paper have demonstrated a general tendency for people to prefer *risky* gambles, i.e., with precise probabilities to *ambiguous* gambles, i.e., whose probabilities are imprecise in the sense of having a lower and upper bound. There have been only a few studies examining the effect of conflicting information [2], [3], and these have indicated that people prefer agreeing but ambiguous sources of information to conflictive but precise sources. To our awareness only one study has investigated *conflictive* gambles, i.e., gambles in which there are conflicting assessments of outcome probabilities [3].

Smithson [2] has argued that people treat ambiguity and conflict as distinct kinds of uncertainty in the sense that attitudes towards one may not correlate with attitudes toward the other, and his experiments and their replication by Cabantous [3] suggest that people prefer ambiguity to conflict. Several researchers also have investigated whether attitudes towards risk and ambiguity are correlated. An early study by Curley et al. [4] found no significant correlation, but later more nuanced investigations by Lauiola and his colleagues did find a positive correlation [5],[6]. Only one study to our knowledge has investigated the correlation between ambiguity and conflict attitudes [7], and found no significant correlation.

Nearly all of the studies in this vein have been based on choice tasks. However, a few have examined pricing, mainly regarding insurance premiums. There is a well-known reluctance for insurers to offer insurance on risks whose probabilities are unknown. When subjective probabilities are used by insurers such as Lloyds of London to estimate such risks, they regard those probabilities as ambiguous and charge higher premiums than they would if the probabilities were based on relative frequency data. The earliest empirical studies to test this effect found that insurers demand higher premiums under ambiguity than un-

der risk [8], and clients are willing to pay more for insurance under ambiguity than under risk [9]. The only study to include conflict [3] found that insurers demand higher premiums under conflict than under ambiguity. These findings suggest that ambiguous and conflictive gambles are devalued relative to expected-utility equivalent risky gambles, and conflictive gambles may be viewed as having less value than ambiguous ones.

There are two ways preferences among gambles may be inferred from buying and selling prices. The first is simply through the prices themselves, i.e., valuation. The second is by comparing the price assigned to a gamble against an appropriate subjective benchmark, i.e., relative valuation. Such comparisons operationalize uncertainty aversion or seeking in terms of prices. The benchmark in this study was the individual's price for a risky gamble with an expected utility equal to that of the ambiguous or conflictive gamble under comparison. In turn, the comparison was operationalized by the log of the ratio of the two prices.

On the basis of the literature reviewed thus far, we propose the following hypotheses.
Hypothesis 1: For mid-range probabilities, both valuation and relative valuation will be lowest for conflictive gambles, second lowest for ambiguous gambles, and highest for risky gambles.
Hypothesis 2: Valuation and relative valuation of risky and ambiguous gambles will be positively correlated, but neither will be correlated with valuation of conflictive gambles.

To our knowledge, none of the aforementioned studies investigated the effect of ambiguity or conflict on the difference between buying and selling prices. In a well-known violation of subjective expected utility known as the *endowment effect* [10], people tend to offer higher selling than buying prices for risky gambles. The standard betting interpretation of lower and upper probabilities also stipulates a higher selling than buying price for ambiguous gambles, but there appears to be no similar standard interpretation for conflictive gambles. Moreover, although it is psychologically plausible that an endowment effect should be greater for ambiguous than for risky gambles, it is not clear how that effect for conflictive gambles would compare. Thus, we posit
Hypothesis 3: For mid-range probabilities, the difference between buying and selling prices will be higher for ambiguous and conflictive gambles than for risky gambles.

1.2 Individual Differences

Research on risk and ambiguity attitudes has paid only limited attention to individual differences, despite obvious variability among individual responses to risk or ambiguity. By far the most widely documented individual difference is due to gender: Men are more risk-seeking than women [11]. Nevertheless, several psychological traits have emerged in the literature as potential predictors of attitudes toward risk and ambiguity.

Research into dispositional components of risk attitudes and risky behaviour has revealed several key relationships. Dispositional traits such as Impulsivity, Locus of Control, and Sensation Seeking have been linked as predictors of risk preferences and risky behaviour in activities ranging from simple games of chance to financial risks, stimulatory hobbies such as rock climbing [12]. In the Big Five personality framework, openness has most commonly been linked with risk-seeking. Our study has included the ten-item-personality inventory (TIPI), a short version of the five-factor model [13]. Finally, Zaleskiewicz [14] developed a two-factor model of risk-taking disposition, with *stimulating risk* correlating with risk-taking in domains such as recreation, and *instrumental risk* correlating with risk-taking in the financial domain. We have included his scales in our study.

We propose the following hypothesis involving the measures described above.
Hypothesis 4: Openness and the stimulating risk scales will be positively correlated with valuation and relative valuation for risky gambles. We leave as exploratory matters the question of whether openness, stimulating risk, or instrumental risk will be correlated with valuation or relative valuation for the ambiguous and conflictive gambles.

Likewise, a few researchers have posited individual difference predictors of attitudes towards ambiguity. In Lauriola and Levin's first paper [5], interviews with participants showing marked ambiguity seeking suggested that they preferred the ambiguous to the risky gamble because they were curious. Huettel et al. [15] found that a measure of impulsivity predicted ambiguity seeking in their fMRI study. These findings suggest including measures of analogs to curiosity and impulsivity. For the first, we have incorporated two recently developed measures based on the theory of uncertainty orientation [16], namely need for discovery and need for certainty [17]. *Need for discovery* measures the extent to which people actively seek novel information, and *need for certainty* measures the disposition to bolster and maintain current beliefs. For the second, we have included Dickman's [18] measures

of functional and dysfunctional impulsivity.

Finally, we propose the following hypothesis.
Hypothesis 5: Instrumental risk, need for discovery and functional impulsivity will be positively correlated and need for certainty negatively correlated with valuation and relative valuation for ambiguous and conflictive gambles.

2 Method

2.1 Participants, Design and Procedures

There were 88 participants with valid responses (58 females and 30 males), ranging in age from 18 to 57 ($M = 26.9$, $SD = 7.5$). A majority (78) of participants were friends and colleagues of the second author and were recruited via email. All participants had little background in probability or mathematics generally. The remaining participants were first year Australian National University psychology students participating for partial course credit. Participants gave informed consent, and were notified prior to commencement that their participation was voluntary and were given online feedback on the study's aims upon completion of the survey.

The study was administered via an online survey with two components, the second of which contained experimental stimuli. In the experimental component described below, participants were presented with 11 Card Game gambles. They were randomly assigned to one of two conditions: Vendor, where they were asked for a minimum selling price for each gamble, or Purchaser, where they were asked for a maximum buying price for each gamble.

2.2 Materials and Tasks

The first section of the study consisted of the individual differences measures. These included the need for discovery and need for certainty scales, the stimulating and instrumental risk inventories, functional and dysfunctional impulsivity scales, and the TIPI.

The second major component of the study consisted of three different tasks, designed to elicit uncertainty preferences. These tasks were extensively pilot-tested before the experiment was launched online. We restrict attention in this paper to the first task, the Card Game. The Card Game is comparable to Ellsberg's (1961) original two-colour task. It required participants to consider a gambling game in which players select a single card from a deck of 100. The deck consists of Old Maid and Go Fish cards in varying proportions. A player wins 10 dollars if they select a Go Fish card, and nothing if they select an Old Maid card. Participants were asked to consider 11 such games, and rate their preferences for each by either specifying the most they would be willing to pay to play the game (Pay to Play endowment condition) or the lowest price for which they would sell a free ticket to play (Selling Price endowment condition).

In the first five scenarios, the full contents of the deck were specified, and risk was manipulated by varying the number of Go Fish (winning) cards in the deck. The proportions of winning cards in the deck for these scenarios were .25, .4, .5, .6, and .75. The proportions were varied to enable estimation of the effect of probability on each participant's valuations of the gambles.

The next three scenarios contained ambiguous information about the deck. The probability intervals were $[.3, .7]$, $[.15, .85]$, and $[0, 1]$. Because the midpoint for each interval was .5, the expected value of each gamble was 5 dollars.

In the final three scenarios, participants were presented with conflicting pieces of information about the contents of the deck from two previous players, and were told that in each case one of the players was approximately correct. The expected value was again maintained at 5 dollars (probability 1/2 of winning 10 dollars), and the conflicting proportions of winning cards claimed by the two sources were $\{.4, .6\}$, $\{.3, .7\}$, and $\{.2, .8\}$. Because the average of each conflictive pair of probabilities was .5, the expected value of each gamble was 5 dollars.

The conflictive probabilities in each scenario were set such that the variance of the probabilities associated with each gamble was approximately equal to the variance in a corresponding ambiguous gamble. Sensitivity to variance has been posited as an explanation for ambiguity aversion, and this eliminates variance as a potential differentiating factor between ambiguous and conflictive gambles. Assuming a uniform distribution, the variance of the probability for an ambiguous gamble with winning probability $[p, 1-p]$ is

$$\sigma_a^2 = (1-2p)^2 /12.$$

Likewise, the variance of the probability for a conflictive gamble with winning probability $\{p, 1-p\}$ is

$$\sigma_c^2 = ((1-2p)/2)^2 .$$

Thus, the variances for the three ambiguous gambles are 0.083, 0.041, and 0.013 respectively; and the variances for the conflictive gambles are 0.09, 0.04, and 0.01 respectively.

3 Results

3.1 Uncertainty and Endowment Effects

The raw dependent variable was valuation, the buying or selling price (in Australian dollars) elicited from respondents. As described earlier, a relative valuation measure also was analyzed. We begin by analyzing valuation.

A minority of participants' valuations were equivalent to the expected utilities (EU's) of the gambles (e.g., valuing at 5 dollars a gamble with probability of .5 of gaining 10 dollars). In the Purchaser condition there were 13 EU responses for risky gambles, 13 for ambiguous gambles and 14 for conflictive gambles. In the Vendor condition, however, these dropped to 5, 3, and 9 EU responses respectively. A two-level logistic regression model found that the difference between the Vendor and Purchaser conditions was significant ($p = .031$), but found no difference among the three types of gambles.

All of the valuations were analyzed with a 2-level GLMM to test Hypotheses 1 and 3 on the valuation data. The GLMM is a choice model without a weighting parameter for probabilities, to ensure model identifiability. The final version of the choice model has the form

$$y_{ij} \approx N\left(\mu_{ij}, \sigma^2\right).$$

The μ_{ij} are defined as subjective expected utilities:

$$\mu_{ij} = U_{ij}\,\pi_i,$$

where U_{ij} is the subjective utility and π_i is the expected probability for the i^{th} gamble and j^{th} subject. In turn, the U_{ij} comprise a 2-level model:

$$\begin{aligned} U_{ij} &= \beta_{0j} + \beta_{1j}\,x_{1i} + \left(\beta_{2j} + \beta_{22j}\,x_{1i}\right) z_{1i} \\ &+ \left(\beta_{3j} + \beta_{33j}\,x_{1i}\right) z_{2i} + \left(\beta_{4j} + \beta_{44j}\,z_{1i}\right) x_{2i}, \end{aligned}$$

where
$x_{1i} = 0$ for the purchaser condition and 1 for the vendor condition,
x_{2i} is the variance of the probability in the i^{th} gamble,
$z_{1i} = 0$ for a precise or conflictive probability and 1 an ambiguous probability, and
$z_{2i} = 0$ for a precise or ambiguous probability and 1 a conflictive probability.
The random-effects coefficients are defined as follows: $\beta_{kj} = \nu_k + u_{kj}$, with $u_{kj} \approx N\left(0, \sigma^2_{kj}\right)$.
The model was estimated via Bayesian MCMC using WinBUGs1.4, in a 2-chain model with a burn-in length of 5,000 iterations and estimations based on a subsequent 10,000 iterations. Convergence diagnostics were favorable for all parameters.

The fixed-effects parameter ν_1 establishes the classic effect of devaluation in the vendor condition if it is negative. The ν_2 and ν_3 parameters compare valuation of ambiguous and conflictive gambles with risky gambles under the purchaser condition, whereas ν_{22} and ν_{33} do so under the vendor condition. All four of these parameters are engaged for testing Hypothesis 1 and the latter two for testing Hypothesis 3. Finally, the ν_4 parameter tests the effect of variance in the probabilites for conflictive gambles and $\nu_4 + \nu_{44}$ does so for ambiguous gambles.

The parameter estimates are displayed in Table 1, along with their standard errors and 95% credible intervals. For risky gambles, the ν_0 estimate suggests a tendency to devalue the \$10 monetary amount slightly in the purchaser condition and the negative ν_1 estimate reproduces the classic further devaluation under the vendor condition.

parameter	estimate	se	lower credib.	upper credib.
ν_0	9.298	0.177	8.954	9.651
ν_1	-0.772	0.290	-1.341	-0.205
ν_2	-1.462	0.201	-1.856	-1.071
ν_{22}	-0.782	0.290	-1.347	-0.208
ν_3	-1.317	0.200	-1.709	-0.924
ν_{33}	-0.520	0.296	-1.100	0.063
ν_4	0.092	0.024	0.044	0.139
ν_{44}	-0.088	0.033	-0.153	-0.022

Table 1: Fixed-Effect Parameter Estimates

Although it is not immediately clear from Table 1, Hypothesis 1 receives only partial support from the findings. The risky gambles are valued more highly ($M = 4.320$) than the ambiguous ($M = 3.166$) and conflictive ($M = 3.568$) gambles, but the ambiguous and conflictive valuation means do not significantly differ. Hypothesis 3, on the other hand, is well-supported. Both ν_{22} and ν_{33} are negative and not significantly different from each other, reflecting greater differences between buying and selling prices for the ambiguous and conflictive gambles than for risky gambles.

Additionally, the effect of variance in the probabilities on valuation was positive for conflictive gambles ($\nu_4 = 0.092$). However, this effect did not emerge for ambiguous gambles because $\nu_4 + \nu_{44} = 0.004$ which did not differ significantly from 0.

We now turn to relative valuation. Recall that the relative valuation measure was the log-ratio of the valuation of the benchmark risky gamble v_r and an alternative gamble v_a:

$$c_r = \ln\left(v_r \,/\, v_a\right)$$

The measure is defined so that higher scores indicate greater relative devaluation of the alternative gamble, so it behaves much like an uncertainty aversion measure.

A mixed ANOVA yielded significant main effects for variance and endowment, and type of gamble. The variance and endowment effects were in the expected directions, so that greater variance resulted in greater relative devaluation ($F(2, 59) = 5.695, p = .005$) and purchasers gave greater relative devaluations than vendors ($F(1, 60) = 9.327, p = .003$). Likewise, there was a significant tendency for conflictive gambles to be relatively devalued less than ambiguous ones ($F(1, 60) = 4.557, p = .037$). There were no interaction effects.

Finally, Hypothesis 2 was tested initially by examining correlations among the valuation and relative valuation measures. These revealed that although valuations and relative valuations of risky and ambiguous gambles were indeed positively correlated, so were they with their counterparts in the conflictive gambles. There were no discernible differences in the strength of correlations between the different types of gambles. The correlations of valuations among gambles were relatively high, ranging from .625 to .950, with RMS $r = .786$. The corresponding findings were similar for both measures of relative valuation (difference and log-ratio), although the correlations were not as strong.

A major limitation of simply correlating valuations across gambles is its inability to address correlations between specific effects. This limitation can be overcome by examining correlations between random-effects parameter estimates in the choice model developed earlier. Table 2 displays these correlations.

β_{0j}							
0.67	β_{1j}						
-0.21	-0.10	β_{2j}					
0.18	0.23	0.41	β_{22j}				
-0.24	-0.12	0.63	-0.05	β_{3j}			
0.27	0.39	0.11	0.52	0.27	β_{33j}		
-0.01	0.01	-0.14	0.02	-0.31	-0.18	β_{4j}	
0.04	0.02	-0.38	-0.42	0.09	0.08	-0.50	β_{44j}

Table 2: Random-Effect Parameter Correlations

The parameters relevant to risky gambles alone (β_{0j} and β_{1j}) are more strongly correlated with each other than with any of the other parameters. Likewise, the parameters measuring effects relevant to the ambiguous and conflictive gambles are more strongly correlated among each other than they are with β_{0j}, β_{1j} or β_{4j}. These findings contradict Hypothesis 2 and suggest a moderately strong link between ambiguous and conflictive gambles in terms of the effects that endowment and variance have on them.

3.2 Individual Differences

Hypotheses 4 and 5 were assessed by excluding the responses that conformed to expected utility theory, because those cases would not be predicted by anything other than the value of the gamble and its probability. To enhance statistical power, the variance in the probabilities was ignored in these analyses, so that only endowment and gamble type were taken into account. Individual differences variables were entered one at a time on their own and a final model was built up by forward addition and likelihood-ratio tests.

Hypothesis 4 was not supported by the prediction of valuation, relative devaluation, or random-effects coefficients. Neither the Openness nor stimulating risk scales predicted any of these dependent variables. Only functional impulsivity predicted valuation of risky gambles, with a positive coefficient ($z = 0.540, p = .005$). However, functional impulsivity did not predict relative devaluation of risky gambles. The relevant random-effects coefficients, β_{0j} and β_{1j}, were weakly positively correlated with scores on the instrumental risk scale ($r = .22$ and .23 respectively).

Hypothesis 5 received some support only for the prediction of valuation and random-effects coefficients. No individual differences measures predicted relative devaluation. For valuation data, there were significant two-way interaction terms between gamble type and instrumental risk and functional impulsivity. The functional impulsivity interaction term was significantly negative for ambiguous gambles ($z = -0.452, p = .005$) and nearly so for conflictive gambles ($z = -0.358, p = .063$). The instrumental risk interaction term, on the other hand, was significantly positive for ambiguous gambles ($z = 0.426, p = .012$) and nearly so for conflictive gambles ($z = 0.337, p = .058$). As for random-effects coefficients, two of the relevant coefficients, β_{2j} and β_{3j}, were positively correlated with scores on the instrumental risk scale (r = .27 for both).

4 Discussion

Our data reproduced the classic endowment effect, the routine violation of expected utility theory whereby people nominate higher selling prices than buying prices for gambles with precise probabilities. The fact that this effect emerged clearly in this study suggests that the experimental manipulation of endowment condition was effective, despite the fact that the

gambles did not yield actual monetary rewards.

Hypotheses 1 and 2 received partial support, but there were some unexpected findings. Conflictive and ambiguous gambles were valued less than expected-utility-equivalent risky gambles. This finding is in line with the aforementioned insurance literature regarding ambiguous gambles, and establishes a similar result for conflictive gambles. However, valuations of ambiguous and conflictive gambles with equivalent variances in the probabilities did not differ. The finding that the random-effects coefficients for ambiguous and conflictive gambles were correlated with each other but not with risky gambles adds weight to the impression that people may evaluate these two kinds of nonprobabilistic uncertainty in similar ways.

However, relative devaluation behaved differently: A significant tendency for conflictive gambles to be relatively devalued less than ambiguous ones and no interaction with endowment or variance. The main effect is unexpected and directly counterindicative of hypothesis 1. It is possible that respondents are more willing to bet on a gamble where the probability of winning is either very high or very low, and this suggests investigating this effect for much higher stakes and also for loss frames.

These findings appear contrary to the preference for ambiguity over conflict established in [2] and replicated in [3]. Moreover, in a recent study of choices among gambles quite similar to those used in this study [7], conflictive gambles were selected less often than ambiguous ones. However, it certainly is possible for people to show preferences in their choices that do not emerge in their valuations (and vice-versa). Preference reversals, after all, are one of the most thoroughly studied violations of expected utility theory. More specifically, response mode (direct comparison versus rating or pricing) has been shown to affect the strength of ambiguity aversion ([19], [20]), with stronger effects found in forced-choice tasks.

A worthwhile extension of the current study would include appropriate choice tasks along with valuation. However, Bowen et al. [21] have observed that when forced to choose, individuals would choose the less ambiguous option and their choice in turn motivates them to overly value the unambiguous option precisely because they need to justify having chosen it. An obvious way around this problem would be to randomize the order of response mode (i.e., half choosing first and half valuing first).

Hypothesis 3 received fairly strong support. The endowment effect was decidedly stronger for conflictive and ambiguous gambles than for risky ones. The random-effects coefficients for these endowment effects were moderately correlated (r = .52) but they also were weakly but positively correlated with the endowment effect for risky gambles (r = .23 and .39).

Could the extra endowment effect for ambiguous and conflictive gambles be explained by the standard betting interpretation of lower and upper probabilities, and therefore by a function of the variance in probabilities? Our findings indicate otherwise, and in fact when variance is taken into account by introducing the appropriate variance*endowment and variance*endowment*gamble-type interaction terms into the choice model, these terms do not significantly improve model fit. Therefore, the betting interpretation of lower and upper probabilities does not explain the extra devaluation of ambiguous and conflictive probabilities, so the cause probably is an alternative psychological response to those types of gambles.

Almost all evidence for candidate explanations comes from studies of ambiguous gambles [22]. However, there is also direct evidence that people simply regard options with missing information as inferior to those with complete information [23], and that this view holds even when the outcomes are losses instead of gains [24]. There appears to be no difference between ambiguous and conflictive gambles; the endowment effect is enhanced equally for both. Respondents appear to devalue both types of gamble as if they perceive a solitary feature that makes both of them inferior to gambles with known probabilities. These findings are compatible with the missing-information explanation.

The absence of correlations between the stimulating risk scale, openness, need for discovery or need for certainty and the valuation of risky gambles (Hypothesis 4) is somewhat surprising, although not very unusual for research in this area. Self-report measures of risk-taking dispositions, tolerance of uncertainty, and the like often do not correlate strongly and can vary considerably across different domains [25]. The study of attitudes towards and responses to nonprobabilistic uncertainty is beset with difficult issues in terminology and measurement [26].

Functional impulsivity and the instrumental risk scale, on the other hand, predicted valuation and random-effects coefficients, albeit in some ways not anticipated in Hypothesis 5. Instrumental risk positively predicted valuation in the ambiguous and conflictive gambles but not in the risky gambles, in line with Hypothesis 5. Likewise, instrumental risk was positively associated with the random-effects coefficients that differentiate the valuation of the ambiguous and conflictive gambles from risky gambles. In other words, higher instrumental risk scores predicted greater valuation of conflictive and ambiguous gam-

bles relative to risky ones. Functional impulsivity, on the other hand, positively predicted valuation only in risky gambles. That effect was reduced to insignificance in the ambiguous and conflictive gambles, in contrast to the Huettel et al. [15] finding that related functional impulsivity to ambiguity seeking.

The instrumental risk scale measures the extent to which people are willing to bear risks in the pursuit of goals or achievements, in contrast to enjoying risks for thrill or excitement. One consequence of this effect is that people scoring high on functional impulsivity value ambiguous and conflictive gambles more like a subjective expected utility agent. A goal-oriented attitude towards risk-taking may lessen the deleterious impact of missing information on the valuation of uncertain prospects, perhaps by motivating people to seek additional information about such prospects. This explanation is compatible with Lauriola and Levin's [5] surmise about the role of curiosity in ambiguity-seeking.

We have already suggested extending this study by comparing preferences as revealed in choice and pricing tasks. We conclude with three additional suggestions for future experimental research on this topic. The most severe limitation on our study is the restriction of the expected probability in the ambiguous and conflictive gambles to a single value (.5) and the prize to \$10. Those restrictions make it impossible to ascertain whether devaluation of ambiguous and conflictive gambles is due to decreasing subjective utility, pessimistic down-weighting of probabilities, or both. Systematically varying the monetary amounts and expected values of the imprecise probabilities would enable separate estimation of probability weighting and subjective utility functions. Second, loss frames need to be studied as well as gain frames. Although Einhorn and Hogarth [24] found ambiguity aversion for loss frames, Smithson [2] found a reflection effect for conflictive scenarios in line with prospect theory's claim that people become risk-seeking under the prospect of loss. Third, the effects of ambiguous versus conflicting utility assessments have yet to be investigated. Taken together, these four suggestions offer a research program that should enrich our understanding of judgment and choice under imprecise probabilities.

Acknowledgements

The design of the experiment reported here and data collection were carried out by the second author as part of his Honours Thesis in Psychology at The Australian National University, under the supervision of the first author.

References

[1] D. Ellsberg. Risk, Ambiguity and the Savage Axioms. *Quarterly Journal of Economics*, 75:643–669, 1961.

[2] M. Smithson. Conflict Aversion: Preference for Ambiguity vs. Conflict in Sources and Evidence. *Organizational Behavior and Human Decision Processes*, 79:179–198, 1999.

[3] L. Cabantous. Ambiguity Aversion in the Field of Insurance: Insurers' Attitude to Imprecise and Conflicting Probability Estimates. *Theory and Decision*, 62:2219–240, 2007.

[4] S.P. Curley, F. Yates, and R.A. Abrams. Psychological Sources of Ambiguity Avoidance. *Organizational Behavior and Human Decision Processes*, 38:230–256, 1986.

[5] M. Lauriola and I.P. Levin. Relating Individual Differences in Attitude toward Ambiguity to Risky Choices. *Journal of Behavioral Decision Making*, 14:107–122, 2001.

[6] M. Lauriola, I.P. Levin, and S.S. Hart. Common and Distinct Factors in Decision Making under Ambiguity and Risk: A Psychometric Study of Individual Differences. *Organizational Behavior and Human Decision Processes*, 104:130–149, 2007.

[7] H. Pushkarskaya, X. Liu, M. Smithson, and J. Joseph. Neurobiological Responses in Individuals Making Choices in Uncertain Environments: Ambiguity and Sample Space Ignorance. *Unpublished Manuscript*, 2009.

[8] R. Hogarth and H. Kunreuther. Risk, Ambiguity and Insurance. *Journal of Risk and Uncertainty*, 2:5–35, 1989.

[9] H. Kunreuther, J. Mezaros, R. Hogarth, and M. Spranca. Ambiguity and Underwriter Decision Processes. *Journal of Economic Behavior and Organization*, 26:337–352, 1995.

[10] R. Thaler. Toward a Positive Theory of Consumer Choice. *Journal of Economic Behavior and Organization*, 1:39–60, 1980.

[11] J.P. Byrnes, D.C., Miller, D.C., and W.D. Schafer. Gender Differences in Risk Taking: A Meta-Analysis. *Psychological Bulletin*, 125:367–383, 1999.

[12] P. Horvath and M. Zuckerman. Sensation Seeking, Risk Appraisal, and Risky Behavior. *Personality and Individual Differences*, 14:41–52, 1993.

[13] S.D. Gosling, P.J. Rentfrow, and W.B. Swann. A Very Brief Measure of the Big-Five Personality Domains. *Journal of Research in Personality*, 37:504–528, 2003.

[14] T. Zaleskiewicz. Beyond Risk Seeking and Risk Aversion: Personality and the Dual Nature of Economic Risk Taking. *European Journal of Personality*, 15:S105–S122, 2001.

[15] S.A. Huettel, C.J. Stowe, E.M. Gordon, B.T. Warner, and M.L. Platt. Neural Signatures of Economic Preferences for Risk and Ambiguity. *Neuron*, 49:765–775, 2006.

[16] R.M. Sorrentino and C.J.R. Roney. *The Uncertain Mind: Individual Differences in Facing the Unknown*. London: Taylor and Francis, 2000.

[17] J. Schuurmans-Stekhoven and M. Smithson. Orientation to Uncertainty as a Possible Dual Process. *Unpublished Manuscript*, 2009.

[18] S.J. Dickman. Functional and Dysfunctional Impulsivity: Personality and Cognitive Correlates. *Journal of Personality and Social Psychology*, 58: 95–102, 1990.

[19] C.R. Fox and A. Tversky. Ambiguity Aversion and Comparative Ignorance. *Quarterly Journal of Economics*, 110:879–895, 1995.

[20] C.R. Fox and M. Weber. Ambiguity Aversion, Comparative Ignorance, and Decision Context. *Organizational Behavior and Human Decision Processes*, 88:476–498, 2002.

[21] J. Bowen, Z.L. Qiu, and Y. Li. Robust Tolerance for Ambiguity. *Organizational Behavior and Human Decision Processes*, 57:155–165, 1994.

[22] D. Frisch and J. Baron. Ambiguity and Rationality. *Journal of Behavioral Decision Making*, 1:149–157, 1988.

[23] I. Ritov and J. Baron. Reluctance to Vaccinate: Commission Bias and Ambiguity. *Journal of Behavioral Decision Making*, 3:263-277, 1990.

[24] H.J. Einhorn and R.M. Hogarth. Decision Making under Ambiguity. *Journal of Business*, 59:S225–S250, 1986.

[25] E.U. Weber, A.R. Blais, and N. Betz. A Domain-Specific Risk-Attitude Scale: Measuring Risk Perceptions and Risk Behaviors. *Journal of Behavioral Decision Making*, 15:1–28, 2002.

[26] M. Smithson. The Many Faces and Masks of Uncertainty. In Bammer, G. and Smithson, M. (Eds.), *Uncertainty and Risk: Multidisciplinary Perspectives*. London: Earthscan, 13–26, 2008.

6th International Symposium on Imprecise Probability: Theories and Applications, Durham, United Kingdom, 2009

Statistical Inference for Interval Identified Parameters

Jörg Stoye
New York University
j.stoye@nyu.edu

Abstract

This paper analyzes the construction of confidence intervals for a parameter θ_0 that is "interval identified," that is, the sampling process only reveals upper and lower bounds on θ_0 even in the limit. Analysis of inference for such parameters requires one to reconsider some fundamental issues. To begin, it is not clear which object – the parameter or the set of parameter values characterized by the bounds – should be asymptotically covered by a confidence region. Next, some straightforwardly constructed confidence intervals encounter problems because sampling distributions of relevant quantities can change discontinuously as parameter values change, leading to problems that are familiar from the pre-testing and model selection literatures. I carry out the relevant analyses for the simple model under consideration, but also emphasize the generality of problems encountered and connect developments to general themes in the rapidly developing literature on inference under partial identification. Results are illustrated with an application to the Survey of Economic Expectations.

Keywords. Partial identification, bounds, confidence regions, hypothesis testing, uniform inference, moment inequalities, subjective expectations.

1 Introduction

Analysis of partial identification is an area of recent growth in statistics and econometrics. To understand its premise, recall the classic definition of identification [16]: A parameter is *identified* if the mapping from its true value to population distributions of observables is invertible; thus, if we knew the latter distribution, we could back out the parameter value. In benevolent settings like those of this paper, identification implies that the parameter's true value can be learned as data accumulate.[1] In contrast, partial identification means that even in the limit, one will only learn some restrictions on this value. Somewhat more formally, if the parameter of interest is θ_0 and is contained in some parameter set Θ, then partial identification means that the population distribution of observables is consistent with any parameter value $\theta \in \Theta_0$, where Θ_0 is an *identified set* containing θ_0. Conventional identification ("point identification") obtains when $\Theta_0 = \{\theta_0\}$; the data generating process reveals nothing of interest if $\Theta_0 = \Theta$. Partial identification ("set identification") obtains in between.

Standard theories of (frequentist) estimation and inference presuppose point identification and require significant adaptation to be applicable to partially identified models. Estimation is the somewhat easier case because it is immediately clear that consistent estimators of θ_0 are unavailable, whereas the object Θ_0 itself is identified in the usual sense (if one thinks of the power set of Θ as a set of feasible parameter values). Questions that arise in estimating this set are typically more of a technical than a conceptual nature. Indeed, in many applications including this paper's, Θ_0 is a well-behaved set whose boundary can be parametrically characterized, so that consistent estimators of Θ_0 obtain straightforwardly. Theories of estimation for more general cases were provided in [5] and [9], among others.

The construction of confidence regions, on the other hand, raises a fundamental question. Should a confidence interval be constructed to cover (with some pre-specified probability) Θ_0 or rather θ_0? Beyond that, a specific technical problem emerges. Construction of confidence intervals typically requires estimation of the limiting sampling distribution of some criterion function or test statistic. These limiting distributions may change discontinuously as the shape of Θ_0 changes qualitatively, e.g. as Θ_0 loses measure.

[1] In general, identifiability is a necessary but not sufficient condition for learnability; e.g., consider incidental parameters or parameters that are discontinuous functions of population distributions.

To be uniformly valid in such critical regions, confidence regions have to implicitly or explicitly deal with a "model selection" or "pre-testing" problem.

This paper discusses these issues and illustrates their impact in a simple but, as it turns out, already quite subtle problem of inference under partial identification. I will discuss the methodological differences between confidence intervals for Θ_0 and for θ_0 and, for either case, provide confidence regions that deal with the aforementioned model selection problem as well as simple ones that do not. I also illustrate all of these in a simple application to real-world data. Parts of the paper have survey character; in particular, section 5.2 reprises results that were recently derived by this author elsewhere [28]. What's new is some technical arguments in section 5.1, the methodological discussion, the intuitions in sections 5.2 and 5.3, and the numerical examples. But to some degree, the purpose of the paper is to provide an entry point to a rapidly developing literature that might be of interest to members of the interval probabilities community.

2 The Setting

Consider the real-valued parameter $\theta_0 \equiv \theta(P_0)$ of a probability distribution $P_0(X)$; here P_0 is known a priori to lie in a set $\mathcal{P}$ that is characterized by ex ante constraints (maintained assumptions), and θ_0 is known to lie in $\Theta \equiv \theta(\mathcal{P})$. The nonstandard feature is that the random variable X is not completely observable, thus θ_0 may not be identifiable: even perfect knowledge of the observable aspects of P_0 might not reveal it. Assume, however, that those observable aspects identify bounds $\theta_l(P_0)$ and $\theta_u(P_0)$ s.t. $\theta_u > \theta_l$ and $\theta_0 \in [\theta_l, \theta_u]$ almost surely. The interval $\Theta_0 \equiv [\theta_l, \theta_u]$ will also be called *identified set*. Let $\Delta \equiv \theta_u - \theta_l$ denote its length.

Here is a motivating example that will later be analyzed numerically. Between 1994 and 1998, the Survey of Economic Expectations elicited worker expectations of job loss by asking the following question:

I would like you to think about your employment prospects over the next 12 months. What do you think is the percent chance that you will lose your job during the next 12 months?

Responses could be any number in $[0, 100]$; with extremely few exceptions near the extremal values, integers were chosen. The survey also elicited covariates, which will be ignored here. The quantity of interest is the population average of subjectively expected probability of job loss, a number that can alternatively be read as the aggregate expected fraction of jobs lost. 3688 of $n = 3860$ sample subjects answered the question, and the average subjective probability expressed by them was 14.8%. However, there was significant item nonresponse: 172 respondents refused to provide an answer. Their subjective expectations of job loss are naturally unknown, although they must lie between 0 and 100 percent. One could pin down an aggregate job loss expectation by making sufficiently strong assumptions about the missing data. For example, if it is assumed that data are missing completely at random, i.e. nonresponders entirely resemble responders other than by not responding, then the aggregate expectation is estimated as 14.8%. As the original data set contains covariates, one could – somewhat more sophisticatedly – assume that data are missing at random conditional on observables. Propensity score or other estimation methods would then lead to a somewhat different estimate that takes into account the distribution of covariates among nonresponders.[2] While they lead to sharp conclusions, these assumptions are very strong and may be accordingly controversial. Partial identification analysis seeks to avoid them, accepting that conclusions may become weaker as a result. An extreme example of this are *worst-case bounds*. In the present example, one could estimate such bounds on aggregate expectations by imputing answers of 0 respectively 100 for all missing data. Numerically, this leads to a lower bound of 14.1% and an upper one of 18.6%. In a next step, these bounds can be refined by re-introducing additional (but not fully identifying) information, and analyses of this kind now constitute a lively literature (see [18] or [19] for surveys). Worst-case bounds suffice to exhibit the inference problem, though, and I will be content with doing that here.

The example is an instance of the "mean with missing data" problem, about the simplest scenario of partial identification that one can think up.[3] In general, assume that X is supported on $[0, 1]$ and that the quantity of interest is $\mathbb{E}X$, but X is observable only if a second, binary random variable $D \in \{0, 1\}$ equals 1. Technically, the sampling process generates a random sample not of realizations x_i, but of realizations $(d_i, x_i d_i)$ which are informative about x_i only if $d_i = 1$. This sampling process identifies the following worst-case bounds:

$$\mathbb{E}(X|D=1)\Pr(D=1) \leq \mathbb{E}X \leq \mathbb{E}(X|D=1)\Pr(D=1) + 1 - \Pr(D=1).$$

These bounds are best possible without further as-

[2] The classic reference on these assumptions is [26]; for a textbook treatment, see [25].

[3] There are many natural examples in which pure identification analysis, i.e. characterization of bounds that are implied by identifiable quantities, amounts to a nontrivial optimization problem ([6], [12], [14], [27]).

sumptions; they are attained if all missing data equal 0 respectively 1.[4]

It is obvious that θ_0 cannot be estimated consistently. At the same time, I will impose assumptions that render trivial the problem of estimating Θ_0. Specifically, assume that estimators $\widehat{\theta}_l$ and $\widehat{\theta}_u$ exist and are uniformly jointly asymptotically normal:

$$\sqrt{n}\left[\begin{array}{c}\widehat{\theta}_l-\theta_l\\ \widehat{\theta}_u-\theta_u\end{array}\right]\xrightarrow{d}N\left(\left[\begin{array}{c}0\\0\end{array}\right],\left[\begin{array}{cc}\sigma_l^2 & \rho\sigma_l\sigma_u\\ \rho\sigma_l\sigma_u & \sigma_u^2\end{array}\right]\right)$$

uniformly in $P\in\mathcal{P}$, where $\left(\sigma_l^2,\sigma_u^2,\rho\right)$ is known. Also, let $\widehat{\Delta}\equiv\widehat{\theta}_u-\widehat{\theta}_l$,

The full strength of $\sqrt{n}$-consistency and asymptotic joint normality of $\left(\widehat{\theta}_l,\widehat{\theta}_u\right)$ is required only to simplify the presentation. For example, $\left(\widehat{\theta}_l,\widehat{\theta}_u\right)$ could also converge at a nonparametric rate, and it would suffice for its distribution to be consistently estimated by the bootstrap. Similarly, assuming that $\left(\sigma_l^2,\sigma_u^2,\rho\right)$ is unknown but can be uniformly consistently estimated (as is the case in the numerical example) would only add notation and require some additional regularity conditions exhibited in [28]. The important substantive assumption that I do make is that the problem of estimating the asymptotic distribution of $\sqrt{n}\left[\widehat{\theta}_l-\theta_l,\widehat{\theta}_u-\theta_u\right]$ has been solved. This assumes away many issues which are not particular to partial identification problems. Note right away that in the motivating example, if one assumes that $\mathbb{E}\left(X|D=1\right)$ and $\Pr(D=1)$ are boundedly away from $\{0,1\}$, then the Berry-Esseen theorem implies uniform joint normality of the obvious estimators

$$\begin{aligned}\widehat{\theta}_l &= \frac{1}{n}\sum_{i=1}^{n}y_id_i\\ \widehat{\theta}_u &= \frac{1}{n}\sum_{i=1}^{n}\left(y_id_i+1-d_i\right)\\ \widehat{\Delta} &= 1-\frac{1}{n}\sum_{i=1}^{n}d_i.\end{aligned}$$

In this application, Θ_0 would naturally be estimated by the plug-in estimator $\widehat{\Theta}\equiv\left[\widehat{\theta}_l,\widehat{\theta}_u\right]$, which was already discovered to numerically equal $[14.1\%,18.6\%]$. I now turn to the difficult problem, namely how to compute confidence regions.

[4] In the specific example, the identified bounds can be seen as characterizing an interval probability for X. This generally occurs with missing data problems because these identify probability distributions up to contamination neighborhoods, and also in many but not all other settings of partial identification.

3 What Should a Confidence Region Cover?

If a parameter θ_0 is conventionally identified, one would like a confidence region CI to fulfil

$$\Pr(\theta_0\in CI)\geq 1-\alpha$$

for some pre-specified α, at least asymptotically as $n\to\infty$. Subject to this constraint, confidence regions should be short or fulfil some other desiderata. However, it is not obvious how to generalize this condition to situations of partial identification. The earlier strand of this literature aimed at the coverage condition

$$\Pr(\Theta_0\subseteq CI)\geq 1-\alpha,$$

thus the idea was to cover the identified set. The methodological contribution of [15] was to rather define coverage by

$$\inf_{\theta_0\in\Theta_0}\Pr(\theta_0\in CI)\geq 1-\alpha,$$

i.e. to attempt coverage of the parameter. This has to be expressed in terms of an infimum over Θ_0 because it is not generally feasible to make coverage probabilities constant over Θ_0. For example, if Θ_0 has an interior, then under regularity conditions any reasonable (i.e. consistent in the Hausdorff metric) estimator $\widehat{\Theta}$ of Θ_0 covers any point in this interior with a limiting probability of 1. The probability limit of $(1-\alpha)$ must, therefore, apply only in some least favorable case that is typically attained on the boundary of Θ_0. Note the following, one-sided implication:

$$\left[\Theta_0\subseteq CI\Longrightarrow\theta_0\in CI\right],\forall\theta_0\in\Theta_0.$$

Thus, if one is content with coverage of the parameter, then a confidence region for the identified set will be valid but generally conservative and therefore needlessly large. On the other hand, if one strives for coverage of the set, coverage of the parameter is simply not sufficient.

Before even attempting to define a confidence region, a researcher must decide which type of coverage is desired. The answer seems to be that it depends on whether Θ_0 or θ_0 is the ultimate object of interest. A reasonable case can be made for either, and I will now attempt to do so.[5]

[5] A superficial answer to this question would be that "it depends on the loss function." In general, one will want to cover the parameter if in the corresponding hypothesis testing problem, loss is incurred from falsely rejecting a null hypothesis about θ_0 as opposed to Θ_0. However, the analogy is not quite precise because coverage of Θ_0 can be justified from testing of compound nulls about θ_0, especially if one is interested in familywise control of the error rate. Also, this would only push back the methodological question by one level. Why, after all, is θ_0 and not Θ_0 in the loss function?

An interest in covering θ_0 seems to hinge on the premise that θ_0 is indeed a true parameter value in the sense of being descriptive of some feature of the real world in a way that other, observationally equivalent values $\theta \in \Theta_0$ are not. This presupposes what one might call a realist interpretation of one's statistical model, meaning that (i) different parameter values correspond to substantially different facts about the real world, (ii) we can on principle learn, at least in some approximate way, the truth about these facts, even though the data set at hand allow this only to a degree that is limited even beyond the usual issues of sampling variation. An analogy from physics for this setting might be that observations generated by a particular experiment generate very imprecise information about some object of interest, but this is because of limitations of measurement, e.g. the resolution of telescopes, and it is accepted that better experimental methods could on principle lead to more precise learning. Among the schools of thought that can be found within the interval probabilities community, this attitude might particularly appeal to researchers who think of interval probabilities mainly as a robustness or sensitivity tool.

In contrast, a statistician who accepts that Θ_0 is all that could ever be learned might find specious the aim of covering θ_0. This attitude would seem especially apt if the underspecified (e.g., interval) probabilities that partially identified models reveal in the limit correspond to fundamental limits to our ability to model underlying phenomena. An analogy from physics might be that observations are imprecise due to fundamental limitations as famously encountered in quantum physics. I conjecture that this attitude might particularly appeal to researchers who think of interval probabilities as a philosophical alternative to conventional probabilities, which they may think of as hopelessly optimistic.

I generally believe that both approaches have merit, and I will discuss both types of confidence regions below. In this paper's specific example, it is this author's feeling that coverage of θ_0 might have special merit. With item nonresponse in surveys, there is often a clear sense in which some precise answer to the item is a matter of fact; sometimes, this answer could even be gleaned from alternative data sources except for legal or practical reasons. (Income and age are salient examples.) In these cases, underidentification of θ_0 seems to stem from practical as opposed to epistemological problems; losses incurred by future policy decisions might well depend on θ_0 rather than Θ_0; and it might be reasonable to think of θ_0 as the quantity of ultimate interest.

4 A (Too) Straightforward Approach

The simplest extension of Wald-type confidence regions to inference on Θ_0 is the following construction which has been used frequently in the literature:

$$CI_{1-\alpha}(\Theta) = \left[\widehat{\theta}_l - \frac{c_\alpha \sigma_l}{\sqrt{n}}, \widehat{\theta}_u + \frac{c_\alpha \sigma_u}{\sqrt{n}}\right],$$

where $c_\alpha = \Phi^{-1}(1-\alpha/2)$ and Φ is the standard normal c.d.f.; e.g. $c_\alpha \approx 1.96$ for a 95%-confidence interval. In words, just enlarge the plug-in- estimator of Θ_0 by the usual number of standard errors. A Bonferroni argument establishes that

$$\begin{aligned}
&\lim_{n\to\infty} \Pr(\Theta_0 \subsetneq CI_{1-\alpha}(\Theta)) \\
= &\lim_{n\to\infty} \Pr\left(\widehat{\theta}_l - \frac{c_\alpha \sigma_l}{\sqrt{n}} > \theta_l \vee \widehat{\theta}_u + \frac{c_\alpha \sigma_u}{\sqrt{n}} < \theta_u\right) \\
\leq &\lim_{n\to\infty} \left(\Pr\left(\widehat{\theta}_l - \frac{c_\alpha \sigma_l}{\sqrt{n}} > \theta_l\right)\right. \\
&\left. + \Pr\left(\widehat{\theta}_u + \frac{c_\alpha \sigma_u}{\sqrt{n}} < \theta_u\right)\right) \\
= &\lim_{n\to\infty} \Pr\left(\frac{\sqrt{n}}{\sigma_l}\left(\widehat{\theta}_l - \theta_l\right) < c_\alpha\right) \\
&+ \lim_{n\to\infty} \Pr\left(\frac{\sqrt{n}}{\sigma_l}\left(\widehat{\theta}_u - \theta_u\right) < -c_\alpha\right) \\
\to &\ 1 - \Phi(c_\alpha) + \Phi(-c_\alpha) = \alpha,
\end{aligned}$$

thus this interval appears valid (if potentially conservative). By the preceding section's reasoning, it must then be conservative for θ_0. Indeed, one can define a confidence region for θ_0 by using the above construction but lowering its confidence level. To see this, observe that

$$\begin{aligned}
&\lim_{n\to\infty} \Pr(\theta_0 \notin CI_{1-\alpha}(\Theta)) \\
= &\lim_{n\to\infty} \Pr\left(\widehat{\theta}_l - \frac{c_\alpha \sigma_l}{\sqrt{n}} > \theta_0 \vee \widehat{\theta}_u + \frac{c_\alpha \sigma_u}{\sqrt{n}} < \theta_0\right).
\end{aligned}$$

If $\theta_l < \theta_0 < \theta_u$, then both $\Pr\left(\widehat{\theta}_l - c_\alpha \sigma_l/\sqrt{n} > \theta_0\right)$ and $\Pr\left(\widehat{\theta}_u + c_\alpha \sigma_u/\sqrt{n} < \theta_0\right)$ vanish at exponential rate as $n \to \infty$, thus

$$\lim_{n\to\infty} \Pr(\theta_0 \notin CI_{1-\alpha}(\Theta)) = 0.$$

If $\theta_0 = \theta_l$, then this reasoning still holds for $\Pr\left(\widehat{\theta}_u + c_\alpha \sigma_u/\sqrt{n} < \theta_0\right)$, but one has

$$\begin{aligned}
&\lim_{n\to\infty} \Pr(\theta_l \notin CI_{1-\alpha}(\Theta)) \\
= &\lim_{n\to\infty} \Pr\left(\widehat{\theta}_l - \frac{c_\alpha \widehat{\sigma}_l}{\sqrt{n}} > \theta_0\right) = \alpha/2.
\end{aligned}$$

A similar reasoning applies if $\theta_0 = \theta_u$, thus

$$\lim_{n\to\infty} \inf_{\theta_0 \in \Theta_0} \Pr(\theta_0 \notin CI_{1-\alpha}(\Theta)) = \alpha/2,$$

and $CI_{1-\alpha}(\Theta)$ is a (non-conservative) $(1-\alpha/2)$ confidence interval for θ_0. Thus one can simply generate a $(1-\alpha)$ confidence interval for θ_0 by writing $CI_{1-\alpha}(\theta) = CI_{1-2\alpha}(\Theta)$. The intuition for this trick is that in the limit as $n \to \infty$, at least one end of the true identified set is far away from the true parameter value, so the hypothesis testing problem that corresponds to the confidence region is really one-sided.

5 Uniform Confidence Regions

The preceding, simple constructions may be compelling at first look, but they suffer from a severe problem: Coverage fails to be uniform over interesting regions of parameter space. This is especially easy to see with respect to coverage of θ_0. While it is true for any fixed (P_0, Θ_0) that $\lim_{n\to\infty} \inf_{\theta_0 \in \Theta_0} \Pr(\theta_0 \in CI_{1-\alpha}(\Theta)) = 1-\alpha/2$, one also finds that $\Pr(\theta_0 \in CI_{1-\alpha}(\Theta)) \to 1-\alpha$ along any local sequence of parameters where $\Delta = o(n^{-1/2})$, i.e. when Δ is asymptotically small relative to sampling error. The algebraic reason is a failure, under this condition, of the above observation that $\Pr\left(\widehat{\theta}_u + c_\alpha \sigma_u/\sqrt{n} < \theta_l\right) \to 0$. The intuitive reason is that the testing problem remains two-sided in the limit. In any case, the confidence region fails to be valid precisely when conventional identifiability of θ_0 is approached, i.e. when the underlying problem actually becomes easier.

Uniformity failures are standard in statistics. Indeed, they are unavoidable if the set of distributions $\mathcal{P}$ is large enough so that the information contained in a sample cannot be bounded away from zero, as famously demonstrated in [4]. The assumption of uniform joint normality is more than sufficient to exclude such situations, however. Accordingly, the present uniformly failure has a much more avoidable cause, namely that Δ is assumed to be large relative to standard errors. If cases of near point identification are of substantive interest, as they often will be, this assumption plainly reveals an inappropriate asymptotic framework. Indeed, were one to neglect this uniformity failure, one would be led to construct confidence intervals that *shrink* as a parameter moves from point identification to slight underidentification. I therefore now turn to constructions that are valid uniformly over possible values of Δ.

The uniformity failure in the coverage argument for θ_0, and different ways to fix the construction, have received significant attention in the literature, and relevant results will be reported. Somewhat surprisingly, $CI_{1-\alpha}(\Theta)$ has seen application even though it is not uniformly valid either. The problem can be intuitively seen as follows. Suppose that $\sigma_l = 1$ but $\sigma_u = 10$. An oracle version of $CI_{95\%}(\Theta)$ that uses infeasible knowledge of these values would be

$$CI_{95\%}(\Theta) = \left[\widehat{\theta}_l - \frac{1.96}{\sqrt{n}}, \widehat{\theta}_u + \frac{19.6}{\sqrt{n}}\right],$$

but for Δ small enough, this interval is strictly contained in the standard Wald confidence region for θ_u,

$$\left[\widehat{\theta}_u - \frac{19.6}{\sqrt{n}}, \widehat{\theta}_u + \frac{19.6}{\sqrt{n}}\right],$$

thus it cannot possibly be valid for Θ_0 in such cases. The upshot is that $CI_{1-\alpha}(\Theta_0)$ is simultaneously conservative, and hence potentially too large, under pointwise asymptotics and invalid under uniform ones, a rather unsatisfactory state of affairs.

5.1 A Confidence Region for Θ_0

If CI_α is interpreted as confidence region for Θ_0, the root cause of its uniformity failure is the same one that underlies its potential conservativeness: Its construction fails to properly account for the fact that the underlying estimation problem is bivariate. This can be fixed by an alternative construction that takes just that bivariate problem – i.e., estimation of (θ_l, θ_u) – as its starting point. Thus, define an arbitrary *joint* confidence region $CI_{1-\alpha}(\theta_l, \theta_u)$ for $\{\theta_l, \theta_u\}$. Denote by $\Theta_l \subset \mathbb{R}$ the projection of this confidence region onto the θ_l-axis and by $\Theta_u \subset \mathbb{R}$ its projection onto the θ_u-axis. Then $\lim_{n\to\infty} \Pr(\theta_l \in \Theta_l, \theta_u \in \Theta_u) \geq 1-\alpha$. Let $CI'_{1-\alpha}$ be the convex hull of $\Theta_l \cup \Theta_u$, then it follows that

$$\begin{aligned} & \lim_{n\to\infty} \Pr(\theta_l \in CI'_{1-\alpha} \wedge \theta_u \in CI'_{1-\alpha}) \geq 1-\alpha \\ \implies & \lim_{n\to\infty} \Pr([\theta_l, \theta_u] \in CI'_{1-\alpha}) \geq 1-\alpha, \end{aligned}$$

where the conclusion uses convexity of $CI'_{1-\alpha}$.

This construction will be uniformly valid as long as normal approximations apply uniformly. Of course, due to the two steps of first forming projections and then computing convex hulls, it is in general conservative, and potentially very much so. This conservatism can be avoided by appropriately choosing the initial confidence region $CI_{1-\alpha}(\theta_l, \theta_u)$. In particular, one should not pick the confidence region of smallest area, i.e. the usual confidence ellipse for bivariate normal means. A better choice is the confidence region that minimizes the length of the convex hull of its projections onto the axes. This confidence region is easily identified as the smallest one to be expressed as $[a,b]^2$ for $a, b \in \mathbb{R}$, i.e. the optimal choice for $CI_{1-\alpha}(\theta_l, \theta_u)$ is

$$\begin{aligned} & CI^*_{1-\alpha}(\theta_l, \theta_u) = \arg\min\{b-a\} \\ \text{s.t.} & \int_{[a,b]^2} dF_{\mathcal{N}}\left(\widehat{\theta}_l, \widehat{\theta}_u, \sigma_l n^{-1/2}, \sigma_u n^{-1/2}, \rho\right) = 1-\alpha, \end{aligned}$$

where $F_{\mathcal{N}}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ denotes a bivariate normal distribution with the specified parameters. Write $CI^*_{1-\alpha}(\theta_l, \theta_u) = [a^*, b^*]$[2], then the convex hull of the projection of this region onto the axes is $CI^*_{1-\alpha}(\Theta) = [a^*, b^*]$, and one obtains

$$\lim_{n\to\infty} \Pr([\theta_l, \theta_u] \subseteq CI^*_{1-\alpha}(\Theta)) = 1 - \alpha$$

uniformly. This construction does not seem to appear in the relevant literature, although projection techniques were used before. In particular, [8] propose to make the initial confidence region $CI_{1-\alpha}(\theta_l, \theta_u)$ balanced, that is, to equalize each parameter's contribution to noncoverage risk. A new justification for this idea in the present context will be encountered below.[6]

5.2 A Confidence Region for θ_0

Uniform confidence regions for θ_0 were recently developed in the literature, with an initial proposal by [15], some issues with which were diagnosed and alleviated in [28]. I will here provide an intuitive development that differs from the original one but connects this section to the preceding one.

The basic idea is the same as before, namely to start from the bivariate problem of estimating (θ_l, θ_u). The difference is that as interest is in covering θ_0 and not Θ_0, the intuitive starting point would be an interval that exhibits pre-specified coverage probability for both θ_l and θ_u, but not necessarily jointly. Some tedious algebra reveals that the shortest such construction is

$$CI^*_{1-\alpha}(\theta) \equiv \left[\widehat{\theta}_l - \frac{\sigma_l c_l}{\sqrt{n}}, \widehat{\theta}_u + \frac{\sigma_u c_u}{\sqrt{n}}\right],$$

where (c_l, c_u) minimize the length of $CI^*_{1-\alpha}(\theta)$ s.t.

$$\int_{-\infty}^{c_l} \Phi\left(\frac{\rho z + c_u + \frac{\sqrt{n}\Delta}{\sigma_u}}{\sqrt{1-\rho^2}}\right) d\Phi(z) \quad \geq \quad 1-\alpha \quad (1)$$

$$\int_{-\infty}^{c_u} \Phi\left(\frac{\rho z + c_l + \frac{\sqrt{n}\Delta}{\sigma_l}}{\sqrt{1-\rho^2}}\right) d\Phi(z) \quad \geq \quad 1-\alpha. \quad (2)$$

(These expressions simplify if $\rho = \pm 1$.) The constraints separately calibrate coverage probabilities at θ_l and θ_u and can be generated by writing out bivariate normal approximations to sampling distributions.

There is a catch however: Expression (1-2) includes Δ, which is not known, thus I just defined an infeasible or "oracle" confidence region. In more elementary inference problems, it is routine to initially do just that and then show that estimators can be substituted for unknown population quantities. But this does not work out here. Under the joint normality assumption, one generally has $\left(\widehat{\Delta} - \Delta\right) = O(n^{-1/2})$, thus $\sqrt{n}\widehat{\Delta}$ does not converge to $\sqrt{n}\Delta$. This will not matter if $\sqrt{n}\Delta$ diverges, in which case $\sqrt{n}\widehat{\Delta}$ diverges as well, but it renders $CI^*_{1-\alpha}(\theta)$ invalid along local parameter sequences where $\sqrt{n}\Delta$ converges.

To resolve this issue, one must ensure that the estimator Δ^* of Δ substituted into (1-2) is *superefficient at zero*. More precisely, Δ^* must have the property that there exists some sequence $\{a_n\}$ that vanishes slowly (i.e., $a_n \to 0$ but $\sqrt{n}a_n \to \infty$) s.t. if the sequence $\{\Delta_n\}$ is dominated by $\{a_n\}$, then $\sqrt{n}(\Delta^* - \Delta_n) \to 0$. Verbally, Δ^* converges at a faster rate than $n^{-1/2}$ for parameter sequences Δ_n that vanish sufficiently fast, including all sequences s.t. $\Delta_n \leq O(n^{-1/2})$.

A striking finding in [28] is that $\widehat{\Delta} \equiv \widehat{\theta}_u - \widehat{\theta}_l$ itself fulfils just this condition in a rather wide set of applications, namely whenever (i) $\left(\widehat{\theta}_l, \widehat{\theta}_u\right)$ are uniformly jointly asymptotically normal, as assumed here, and (ii) $\widehat{\Delta} \geq 0$ almost surely, e.g. $\widehat{\theta}_l \leq \widehat{\theta}_u$ by construction. Thus, if estimators of upper and lower bounds are jointly asymptotically normal and are necessarily ordered in the right way, then the implied estimator of the difference between the bounds is superefficient at zero. This condition turns out to have reasonably wide applicability. Among other things, it means that the estimator $\widehat{\Delta}$ in this paper's example – the mean with missing data – is superefficient.[7]

However, there are also many cases (e.g. in [22] and [24]) where superefficiency of $\widehat{\Delta}$ will not obtain naturally. It must then be induced artificially. A simple way to do this is to shrink $\widehat{\Delta}$ toward zero, writing

$$\Delta^* = \widehat{\Delta} \cdot \mathbb{I}\{\widehat{\Delta} \geq a_n\}, \quad (3)$$

where $\mathbb{I}\{\cdot\}$ is the indicator function and a_n is a user-specified sequence of numbers s.t. $a_n \to 0$ but $\sqrt{n}a_n \to \infty$. One of the main results in [28] is that $CI^*_{1-\alpha}(\theta)$ is uniformly valid for θ_0 upon substitution of Δ^* for Δ in (1-2).

A second, less troublesome issue with $CI^*_{1-\alpha}(\theta)$ is that it may not be well defined as written, namely if $\widehat{\theta}_l - \sigma_l c_l/\sqrt{n} > \widehat{\theta}_u + \sigma_u c_u/\sqrt{n}$, which absent superefficiency of $\widehat{\Delta}$ is an event with nonvanishing (more precisely: not uniformly vanishing) finite sample probability. This author's proposal is to leave the interval empty in such cases. This does not affect its

[6] [13] also propose a similar construction but make it symmetric about $\left\{\widehat{\theta}_l, \widehat{\theta}_u\right\}$. [15] and [28] mention $CI_{1-\alpha}(\Theta)$ as confidence region for Θ_0; in fairness, their focus is squarely elsewhere.

[7] In the specific example, superefficiency of $\widehat{\Delta}$ can also be seen heuristically. The estimator $\widehat{\Delta}$ is the sample analog of a population probability $\Delta = \Pr(D = 0)$, thus it has variance $\Delta(1-\Delta)/N$, the numerator of which vanishes as $\Delta \to 0$.

validity; hence, any other fix will lead to a needlessly long interval. It can also be interpreted as an embedded specification test: Samples which induce $\widehat{\theta}_l - \sigma_l c_l/\sqrt{n} > \widehat{\theta}_u + \sigma_u c_u/\sqrt{n}$ really cast doubt on the maintained hypothesis that $\theta_u \geq \theta_l$. Having said that, some users might not like confidence sets that can be empty. They could define $CI^*_{1-\alpha}(\theta)$ in an arbitrary manner whenever $\widehat{\theta}_l - \sigma_l c_l/\sqrt{n} > \widehat{\theta}_u + \sigma_u c_u/\sqrt{n}$. A natural solution might be to proceed as if one had learned that $\theta_u = \theta_l$, thus one could write

$$CI^*_{1-\alpha}(\theta) = \left[\widehat{\theta} - \frac{c_\alpha \sigma}{\sqrt{n}}, \widehat{\theta} + \frac{\sigma c_\alpha}{\sqrt{n}}\right],$$

where $\widehat{\theta} \equiv \left(\widehat{\theta}_l/\sigma_l^2 + \widehat{\theta}_u/\sigma_u^2\right) / \left(1/\sigma_l^2 + 1/\sigma_u^2\right)$ is a variance weighted average of $\widehat{\theta}_l$ and $\widehat{\theta}_u$ and $\sigma^2 \equiv 1/\left(1/\sigma_l^2 + 1/\sigma_u^2\right)$ is its sampling variance.

5.3 Relation to Model Selection and to Moment Inequalities

To understand the workings of $CI^*_{1-\alpha}(\theta)$, it is instructive to emphasize the model selection, or "pre-testing," issue that is lurking below the surface here. Recall that confidence regions typically correspond to hypothesis tests, that is, they can be thought of as lower contour set of some test statistic, thus collecting parameter values ξ for which the data do not reject the null hypothesis $H_0 : \theta_0 = \xi$. When constructing a confidence region for θ_0, the corresponding hypothesis test appears one-sided in the pointwise limit as $n \to \infty$ for any $\Delta > 0$, thus one seemingly gets away with lower cutoff values c_α than would be required for two-sided tests. Yet the test remains two-sided if $\Delta = 0$, in which case the confidence region would surely have to be a standard Wald confidence region. The pointwise limit distributions of relevant test statistics thus change discontinuously as $\Delta \to 0$. Of course, their true finite sampling distribution are continuous in Δ for any n. It follows that for any n, the pointwise approximations must be misleading for some Δ. This is why $C_{1-\alpha}(\theta)$ fails to be uniformly valid.

This type of problem is familiar to researchers investigating model selection or pre-testing. Essentially the same issues occur at the boundary between models that a pre-test or model selection procedure aims to separate. Indeed, one can think of the present problem as one of model selection, namely as deciding whether a point identified ($\Delta = 0$) or partially identified ($\Delta > 0$) model better describes the data. The shrinkage step (3) can then be interpreted as a pre-test that decides among these models, with $\Delta^* = 0$ indicating that point identification should be presumed.[8]

A general problem with pre-tests is that their sampling error must be taken into account in subsequent inference and will frequently invalidate it. To avoid this, the test underlying $CI^*_{1-\alpha}(\theta)$ has a conservative slant. Point identification requires more conservative inference in the sense of larger cutoff values, therefore one can achieve validity (at cost of having longer confidence intervals) by erring in favor of presuming point identification. This is here implemented because the sequence a_n vanishes at a rate slower than $O(n^{-1/2})$, thus along any local sequence where $\Delta \leq O(n^{-1/2})$, point identification will eventually be presumed with probability 1. The price is that $CI^*_{1-\alpha}(\theta)$ will be uniformly valid (i.e. valid along all moving parameter sequences) and pointwise exact (i.e., not conservative under asymptotics that hold true parameter values fixed), but conservative along certain local sequences. Some features of this sort are essentially unavoidable when working with pre-tests; the question is mainly whether researchers acknowledge them or not, an issue on which [17] offer some cautionary tales.

It is also noted that upper and lower bounds on a real-valued parameter θ_0 are a special case of moment inequalities, a rather general framework that recently attracted much interest ([1], [2], [3], [7], [10], [21]). Moment inequalities occur when a true parameter value θ_0 is incompletely characterized by a set of inequalities

$$\mathbb{E}(m_j(x_i, \theta_0)) \geq 0, j = 1, \ldots, J,$$

where the expectations are population expectations and the m_j are known functions. Clearly such a set of conditions generally identifies a set, e.g. a polyhedron if the m_j are linear. This paper's scenario fits this framework as the special case of two moment inequalities

$$\begin{aligned} \mathbb{E}(\theta_0 - d_i x_i) &\geq 0 \\ \mathbb{E}(d_i x_i + 1 - d_i - \theta_0) &\geq 0. \end{aligned}$$

Many of the problems encountered for moment inequalities are just more intricate versions of the ones analyzed here. In particular, the adequate definition of confidence regions will depend on which moment inequalities bind, which can potentially be determined via a pre-test; but this will encounter the problem just described. Sure enough, numerous papers on moment inequalities ([2], [3], [7], [10], [21]; see also [11] for related ideas about compound hypothesis testing more generally) contain a step in which sample analogs of moment inequalities are shrunk toward zero, i.e. they

[8] In the specific example, the discontinuity issue could also be avoided by calibrating cutoff values through subsampling [23] although not through the bootstrap [7]. See [1] for a more general analysis of subsampling and its limits in cases of partial identification.

perform the exact trick introduced in the previous subsection.[9]

5.4 Unbiasedness of Confidence Regions

I conclude the theoretical analysis with some remarks about unbiasedness of confidence intervals under partial identification.[10] Recall that a confidence region CI for θ_0 is unbiased if $\Pr(\theta \in CI)$, seen as a function of θ, is maximized at θ_0. The corresponding concept for hypothesis tests is that the probability of rejection should be minimized on the null.[11]

Unbiasedness in this sense will not apply here. Consider first coverage of θ_0 when the identified set is $[\theta_l, \theta_u]$. Any reasonable confidence region will cover points in the interior of this set with probability approaching one and thus cannot be unbiased when the truth is $\theta_0 = \theta_u$. The situation is not better regarding coverage of Θ_0. Clearly any subset of Θ_0 will be covered more frequently than Θ_0 itself. Even excluding subsets from the comparison, problems with small sets remain. For example, as long as some noncoverage risk stems from the lower end of $[\theta_l, \theta_u]$, some set of the form $[\theta_u - \sqrt{n}\gamma, \theta_u + \sqrt{n}\gamma]$ will be covered more frequently than $[\theta_l, \theta_u]$.

It seems more promising to take a cue from compound hypothesis testing and be content with the requirement that Θ_0 is an upper contour set of $\Pr(\theta \in CI)$. Yet even this aim seems unrealistic when Δ is allowed to be small. For example, if $\Delta = n^{-1/2}$ and σ_u sufficiently exceeds σ_l, then any convex 95% confidence region for θ_u is conservative for θ_l and hence for a parameter value locally below θ_l. Unbiasedness could then only be achieved at the price of substantial conservatism, if at all. Thus, one might further weaken the unbiasedness criterion by requiring it only to hold along parameter sequences that hold $(\Delta, \sigma_l, \sigma_u)$ fixed.

With these adjustments in place, $CI^*_{1-\alpha}(\theta)$ is (asymptotically) unbiased. In particular, (1-2) bind with probability approaching 1, and in the limit, $\Pr(\theta \in CI^*_{1-\alpha}(\theta)) \geq 1-\alpha$ on Θ_0 but $\Pr(\theta \in CI^*_{1-\alpha}(\theta)) < 1-\alpha$ otherwise. $CI^*_{1-\alpha}(\Theta)$, on the other hand, does not fulfil the requirement because it is based on an unbalanced simultaneous confidence region for (θ_l, θ_u). If these parameters are measured with different precision, then $CI^*_{1-\alpha}(\Theta)$ will be more likely to cover the more precisely measured one because some such allocation of noncoverage risk minimizes length. As a result, if $\sigma_u > \sigma_l$, say, then some local value of the form $\theta_l - \sqrt{n}\gamma$ is covered more frequently than θ_u. This may be acceptable because it is not obvious that a confidence region designed for Θ_0 as object of interest need be unbiased for θ_0. Having said that, such unbiasedness is achieved by the balanced construction in [8], so one arguably encounters a trade-off between unbiasedness and length of confidence regions.

6 Numerical Illustrations

This section illustrates the above findings with some numerical examples. The first one is the empirical application described in section 2; the other two use artificial data. Recall that interest was in an average subjective probability of one-year-ahead job loss. Sample size is $n = 3860$; using the notation from section 2, the sample analog of $\mathbb{E}(X|D=1)$ is 14.8% and the sample analog of $\Pr(D=1)$, i.e. the probability of response, is 95.5%. These numbers imply that apart from their asymptotic validity, normal approximations should be expected to work well for the given sample. Simple computations establish that furthermore

$$\left(\widehat{\theta}_l, \widehat{\theta}_u, \widehat{\Delta}, \widehat{\sigma}_l, \widehat{\sigma}_u, \widehat{\rho}\right) = (14.10, 18.55, 4.45, 23.53, 29.22, 0.714).$$

The estimator of the identified set and the different confidence regions then compute as follows:

$$\begin{aligned} \widehat{\Theta} &= [14.10, 18.55] \\ CI_{95\%}(\Theta_0) &= [13.36, 19.48] \\ CI_{95\%}(\theta_0) &= [13.48, 19.33] \\ CI^*_{95\%}(\Theta_0) &= [13.33, 19.45] \\ CI^*_{95\%}(\theta_0) &= [13.48, 19.33]. \end{aligned}$$

The results show the expected features: $CI_{5\%}(\theta_0) \subset CI_{5\%}(\Theta_0)$ (as is the case by construction), and $CI^*_{5\%}(\Theta_0)$ differs from $CI_{5\%}(\Theta_0)$ without nesting it. Having said that, the quantitative differences are small. This comes from two facts: First, in the example, $\widehat{\Delta}$ is large relative to $\widehat{\sigma}_l/\sqrt{n-1}$, so that the uniformity issues are not salient and the fixes hence marginal; indeed $CI_{5\%}(\theta_0)$ and $CI^*_{5\%}(\theta_0)$ cannot be distinguished numerically. Second, the estimators of the bounds have strong positive correlation ($\widehat{\rho} = 0.714$), so that the construction of $CI_{5\%}(\Theta_0)$ is not all that conservative.

To bring these issues a bit more to the forefront, I also generate intervals for a hypothetical dataset in which $n = 100$, I continue to assume that $\widehat{\Delta}$ is superefficient, and

$$\left(\widehat{\theta}_l, \widehat{\theta}_u, \widehat{\Delta}, \sigma_l, \sigma_u, \rho\right) = (15, 17, 2, 20, 30, -.3).$$

[9] Note that $\Delta = \mathbb{E}(1-d_i)$, thus shrinking $\widehat{\Delta}$ amounts to artificially tightening the second of the above moment inequalities.

[10] I thank a referee for raising this question.

[11] None of this can here be shown for finite samples because this paper's assumptions do not restrict finite sample distributions. I therefore mean unbiasedness to apply asymptotically as $n \to \infty$; this is a nontrivial requirement because it is understood to apply to ($\sqrt{n}$-)local alternatives.

Results then are:

$$\begin{aligned} \widehat{\Theta} &= [15, 17] \\ CI_{95\%}(\Theta_0) &= [11.08, 22.88] \\ CI_{95\%}(\theta_0) &= [11.71, 21.93] \\ CI^*_{95\%}(\Theta_0) &= [10.28, 22.63] \\ CI^*_{95\%}(\theta_0) &= [11.54, 22.01]. \end{aligned}$$

This example is somewhat rigged to showcase the effect of Δ being small. The difference between $CI_{5\%}(\Theta_0)$ and $CI^*_{5\%}(\Theta_0)$ is much larger. The former is substantially too small at its left end and must be extended to account for the large sampling variation in $\widehat{\theta}_u$. At the same time, the negative correlation means that noncoverage at the upper and lower end of the interval are likely to occur in the same samples, thus the overall probability of noncoverage is noticeably less than the sum of those two individual probabilities. This can be exploited to make the interval shorter, and it is this effect that dominates at its right end. Finally, the higher precision of $\widehat{\theta}_l$ is exploited by $CI^*_{95\%}(\Theta_0)$ to minimize interval length at the price of unbalancedness as discussed above; a balanced version of the interval would have a higher minimum as well as maximum but be longer.

The second hypothetical example features a large Δ but a very negative correlation between estimators, implying that the Bonferroni construction $CI_{95\%}(\Theta_0)$ is quite conservative. With $n = 100$ and

$$\left(\widehat{\theta}_l, \widehat{\theta}_u, \widehat{\Delta}, \sigma_l, \sigma_u, \rho\right) = (10, 20, 10, 20, 20, -.9),$$

one accordingly gets

$$\begin{aligned} \widehat{\Theta} &= [10, 20] \\ CI_{95\%}(\Theta_0) &= [6.08, 23.92] \\ CI_{95\%}(\theta_0) &= [6.71, 23.29] \\ CI^*_{95\%}(\Theta_0) &= [6.40, 23.59] \\ CI^*_{95\%}(\theta_0) &= [6.71, 23.29] \end{aligned}$$

and $CI^*_{95\%}(\Theta_0)$ is noticeably smaller than $CI_{95\%}(\Theta_0)$.

7 Summary and Outlook

Analysis of partial identification aims to provide conclusions which are robust, even at the price of not always being very strong. It is close in spirit and in methods to much work on interval probabilities (and also to robust Bayesian approaches). The systematic analysis of estimation and inference under partial identification is the object of a currently active literature. One general finding is that compared to well known methods that apply to conventionally identified methods, basic questions about inference have to be asked anew, and findings become substantially more nuanced.

This paper illustrated some of these issues in the very simple setting of an interval identified real-valued parameter. Inference toward an expected value when some data are missing served as motivating example that was carried out with real-world data. The issues encountered along the way range from the methodological or even philosophical to the pragmatic and quite technical. In particular, it was seen that simple asymptotic frameworks can inform misleading results, and that there are some nontrivial complications which link the inference problem to the large and growing literature on post model selection estimation and inference. Work on much more general settings than the one investigated here is under way; it encounters essentially the same problems, and then some. It is hoped that once these general theories are in place, thinking in terms of partial identification, rather than assuming away all identification problems, becomes part of many statisticians' and applies researchers' toolkit.

Acknowledgements

I thank Nick Kiefer for a question that ultimately led to the construction of $CI^*_{1-\alpha}(\Theta)$ and two referees as well as a seminar audience at Yale's statistics department for helpful comments. This paper was written while the author visited the Cowles Foundation at Yale University, whose hospitality is gratefully acknowledged. Financial support from a University Research Challenge Fund, New York University, is gratefully acknowledged.

References

[1] D.W.K. Andrews and P. Guggenberger. Validity of Subsampling and 'Plug-in Asymptotic' Inference for Parameters Defined by Moment Inequalities, *Econometric Theory*, forthcoming.

[2] D.W.K. Andrews and P. Jia. Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure. Cowles Foundation Discussion Paper 1676, 2008.

[3] D.W.K. Andrews and G. Soares. Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection, Cowles Foundation Discussion Paper 1631, 2007.

[4] R.R. Bahadur and L.J. Savage. The Nonexistence of Certain Statistical Procedures in Nonparamet-

ric Problems. *Annals of Mathematical Statistics* 25:1115–1122, 1955.

[5] A. Beresteanu and F. Molinari. Asymptotic Properties for a Class of Partially Identified Models. *Econometrica*, 76:763-814, 2008.

[6] A. Beresteanu, I. Molchanov, and F. Molinari. Sharp Identification Regions in Games. CEMMAP working paper 15/08, 2008.

[7] F.A. Bugni. Bootstrap Inference in Partially Identified Models. Preprint, Northwestern University, 2007.

[8] J. Cheng and D.S. Small. Bounds on Causal Effects in Three-Arm Trials with Non-Compliance. *Journal of the Royal Statistical Society, Series B*, 68:815-836, 2006.

[9] V. Chernozhukov, H. Hong, and E.T. Tamer. Parameter Set Inference in a Class of Econometric Models. *Econometrica*, 75:1243-1284, 2007.

[10] Y. Fan and S.S. Park. Confidence Intervals for Some Partially Identified Parameters. Preprint, Vanderbilt University, 2007.

[11] P.R. Hansen. Asymptotic Tests of Composite Hypotheses. Preprint, Brown University, 2003.

[12] B.E. Honoré and E.T. Tamer. Bounds on Parameters in Panel Dynamic Discrete Choice Models. *Econometrica* 74:611–629, 2006.

[13] J.L. Horowitz and C.F. Manski. Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data. *Journal of the American Statistical Association* 95:77–84, 2000.

[14] J.L. Horowitz, C.F. Manski, M. Ponomareva, and J. Stoye: Computation of Bounds on Population Parameters When the Data are Incomplete. *Reliable Computing* 9:419-440, 2003.

[15] G. Imbens and C.F. Manski. Confidence Intervals for Partially Identified Parameters. *Econometrica*, 72:1845-1857, 2004.

[16] T. Koopmans. Identification Problems in Economic Model Construction. *Econometrica* 17:125-144, 1949.

[17] H. Leeb and B. Pötscher. Model Selection and Inference: Facts and Fiction. *Econometric Theory* 21:21-59, 2005.

[18] C.F. Manski. *Partial Identification of Probability Distributions*. Springer Verlag, 2003.

[19] C.F. Manski. *Identification for Prediction and Decision*. Harvard University Press, 2007.

[20] C.F. Manski and J. Straub. Worker Perceptions of Job Insecurity in the Mid-1990s: Evidence from the Survey of Economic Expectations. *Journal of Human Resources* 35:447-479, 2000.

[21] K. Menzel. Estimation and Inference with Many Moment Inequalities. Preprint, Massachussetts Institute of Technology, 2008.

[22] A. Pakes, J. Porter, K. Ho, and J. Ishii. Moment Inequalities and their Application. Preprint, Harvard University, 2006.

[23] J.P. Romano and A.M. Shaikh. Inference for Identifiable Parameters in Partially Identified Econometric Models. *Journal of Statistical Planning and Inference*, 138:2786-2807, 2008.

[24] A.M. Rosen. Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities. *Journal of Econometrics*, 146:107-117, 2008.

[25] P.R. Rosenbaum. *Observational Studies (2nd Edition)*. Springer Verlag, 2002.

[26] D.B. Rubin. Inference and Missing Data. *Biometrika* 63:581–592, 1976.

[27] J. Stoye. Partial Identification of Spread Parameters. Preprint, New York University, 2005.

[28] J. Stoye. More on Confidence Intervals for Partially Identified Parameters. *Econometrica*, forthcoming.

Shifted Dirichlet Distributions as Second-Order Probability Distributions that Factors into Marginals

David Sundgren
University of Gävle
dsn@hig.se

Love Ekenberg
Stockholm University
and KTH
lovek@dsv.su.se

Mats Danielson
Stockhom University
and KTH
mad@dv.su.se

Abstract

In classic decision theory it is assumed that a decision-maker can assign precise numerical values corresponding to the true value of each consequence, as well as precise numerical probabilities for their occurrences. In attempting to address real-life problems, where uncertainty in the input data prevails, some kind of representation of imprecise information is important. Second-order distributions, probability distributions over probabilities, is one way to achieve such a representation. However, it is hard to intuitively understand statements in a multi-dimensional space and user statements must be provided more locally. But the information-theoretic interplay between joint and marginal distributions may give rise to unwanted effects on the global level. We consider this problem in a setting of second-order probability distributions and find a family of distributions that normalised over the probability simplex equals its own product of marginals. For such distributions, there is no flow of information between the joint distributions and the marginal distributions other than the trivial fact that the variables belong to the probability simplex.

Keywords. Second-order probability distribution, Dirichlet distribution, Beta distribution, Kullback-Leibler divergence, relative entropy, product of marginal distributions.

1 Introduction

In attempting to address real-life decision problems, where uncertainty about data prevails, some kind of representation of imprecise information is important and several have been proposed. In particular, first-order representations, such as sets of probability measures [9], upper and lower probabilities [2], and interval probabilities and utilities of various kinds, see e.g. [15, 16], have been suggested for enabling a better representation of the input sentences for a subsequent decision analysis. To facilitate a better qualification of the various possible functions, higher-order estimates, such as distributions expressing various beliefs, can be introduced over n-dimensional spaces, where each dimension corresponds to possible probabilities of events or utilities of consequences. Such hierarchical model approaches are sometimes better suited for modelling incomplete knowledge and can add important information when handling aggregations of imprecise representations, as is the case in decision trees or probabilistic networks [3]. There are, however, at least two problems herein. Firstly, a normal decision maker cannot have any meaningful intuition regarding a multi-dimensional space and the information must be provided more locally, and secondly, it is hard to obtain global information from such local information. Of particular interest in this context is therefore to investigate the relation between global and local distributions.

We will use second-order probabilities, formally defined below in Definition 4, in short, these are probability distributions on random variables that take values on $[0, 1]$ and sum to 1. The intuition of a second-order probability distribution is that it is a distribution that assigns probabilities to the probabilities of the possible outcomes of an event. So such a distribution will have to be defined on the hyper-surface defined by $\sum_{i=1}^{n} x_i = 1, x_i \geq 0, i = 1, \ldots, n$ or, equivalently on the $n-1$-dimensional simplex where $\sum_{i=1}^{n-1} x_i \leq 1, x_i \geq 0, i = 1, \ldots, n-1$ and x_n is an abbreviation of $1 - \sum_{i=1}^{n-1} x_i$. In this paper we will only consider continuous distributions.

The uniform distribution with support on the simplex where $\sum_{i=1}^{n-1} x_i \leq 1, x_i \geq 0, i = 1, \ldots, n-1$ with constant value the inverse of the volume of the simplex and the Dirichlet distribution are examples of second-order probability distributions.

Such second-order probability distributions is one way of handling uncertainty of probabilities in a decision situation, see e.g. [11], [14] and [5]. Instantly, new

difficulties appear; on the one hand it may seem that there are too many distributions to choose from given the available knowledge, on the other hand it is not certain that any set of univariate second-order distributions is consistent with the fact that the variables are themselves probabilities. Even if they are consistent, the marginal distributions and the joint distribution may represent different information, e.g. it is shown in [13] that the uniform joint distributions have marginal distributions that are far from uniform.

1.1 Definitions

Definition 1 *[1] For a k-dimensional random vector* $(X_1, \ldots, X_k)$ *the (joint)* distribution μ *is defined by*

$$\mu(A) = \Pr[(X_1, \ldots, X_k) \in A], A \in \mathcal{R}^k$$

where $\mathcal{R}^k$ *is the* σ*-field generated by the bounded rectangles* $[x = (x_1, \ldots, x_k) : a_i < x_i \leq b_i, i = 1, \ldots, k]$.

Definition 2 *[1] A k-dimensional random vector* $(X_1, \ldots, X_k)$ *and its distribution have* density f *with respect to Lebesgue measure if* f *is a nonnegative Borel function on* R^k *and*

$$\mu(A) = \int_A f(\mathbf{x})\, d\mathbf{x}, A \in \mathcal{R}^k\,.$$

Definition 3 *[1] If the k-dimensional vector* $X = (x_1, \ldots x_k)$ *has distribution* μ *and if* $\pi_j : R^k \to R$ *is defined by* $\pi_j(x_1, \ldots, x_n) = x_j$*, the (univariate)* marginal distributions *of* μ *are* $\mu_j = \mu \circ \pi_j^{-1}$ *given by* $\mu_j(A) = \mu[(x_1, \ldots, x_k) : x_j \in A] = \Pr[X_j \in A]$ *for all* $A \in \mathcal{R}$.

Definition 4 *A* second-order probability distribution *is a distribution* μ *with support on a set* $\mathcal{P} = \{(x_1, \ldots, x_k) : 0 \leq a_i \leq x_i \leq b_i, i = 1, \ldots, k, \sum_{i=1}^{k} \leq 1\}$.

1.2 The Problem

Below we will only consider densities and for simplicity abuse terminology as to identify distributions with their probability density functions.

For most decision makers it would be easiest to consider univariate distributions since it is harder to think in several dimensions [4]. In general, though, the marginal distributions together contain more information than the corresponding multivariate distribution. The random variables are the probabilities of the possible outcomes of an event. If the variables are dependent in other than relating to the same event this information discrepancy between local and global is natural since information would be shared between the local variables. But settings where the opposite holds comes easier to mind, and such cases would be better modelled with random variables that are as independent as possible modulo that they sum to one.

The above reasoning motivates us to consider whether there are joint second-order probability distributions that have the same information content as its univariate marginal distributions. This condition will be seen to be equivalent to the joint probability distribution function being equal to the product of its own univariate marginal distributions multiplied with a normalising constant that comes from us working in the probability simplex rather than in the unit cube. That is, the information-theoretic constraint of not losing information when taking marginals coincides with the practical concern of being able to construct a joint probability density from given marginals in the simplest possible way. In terms of copulas (see e.g. [10] or [12]), the condition is that the copula is the product copula multiplied by some constant.

We show that the condition of a joint probability distribution function being equal to the product of its own univariate marginal distributions multiplied with a normalising constant is met by a family of distributions that have the same shape as the Dirichlet distribution. The first-order probability variables can be given arbitrary bounds only from below. When the lower bounds are zero, we have a special case of the Dirichlet distribution where all parameters are equal. With general lower bounds $x_i \geq a_i$, the support of x_i is the interval $[a_i, 1 + a_i - \sum_{i=1}^{n} a_i]$, the joint Dirichlet distribution and the corresponding marginal Beta distribution are shifted and re-scaled accordingly.

2 Minimal Kullback-Leibler Divergence

To capture the notion that no information other than that of being on the probability simplex is either lost or gained when going between a joint probability distribution and its marginals, we use the *Kullback-Leibler divergence* or *relative entropy* [8], see also [6, 7].

Definition 5 *If* P *and* Q *are probability measures over a set* X *and if* μ *is a measure such that* $p = \frac{dP}{d\mu}$ *and* $q = \frac{dQ}{d\mu}$ *exist the* Kullback-Leibler divergence *from* P *to* Q *is*

$$D_{\mathrm{KL}}(P\|Q) = \int_X p \log \frac{p}{q}\, d\mu\,.$$

Since we want, as far as possible given that we are on the probability simplex $\mathcal{P}$, that the joint distribution $f(x_1, \ldots, x_n)$ contains the same information as

the product of marginals $\prod_{i=1}^n f_i(x_i)$, we want the Kullback-Leibler divergence $D_{\mathrm{KL}}(f\|\prod_{i=1}^n f_i)$, also known as the *total correlation* [17] of $X_1,\ldots,X_n$ to be minimal. *Gibbs' inequality* states that $D_{\mathrm{KL}}(P\|Q) \geq 0$ with equality only if $P = Q$. Since the probability simplex $\mathcal{P}$ is measurable we can calculate $D_{\mathrm{KL}}(f\|\prod_{i=1}^n f_i)$ as a Lebesgue integral.

But restricting the support of $\prod_{i=1}^n f_i$ to the probability simplex $\mathcal{P}$ means that $\prod_{i=1}^n f_i$ must be normalised in order to be a distribution, i.e. we must find a real number K such that $\int_{\mathcal{P}} \prod_{i=1}^n f_i(x_i)/K\, d\mathbf{x} = 1$. So minimising the Kullback-Leibler divergence of $\prod_{i=1}^n f_i$ from the joint probability distribution entails finding f such that

$$f(x_1,\ldots x_n) = \frac{1}{K}\prod_{i=1}^n f_i(x_i)\,,$$

where $f_i(x_i)$ is the marginal distribution of $f(\mathbf{x})$ with respect to x_i and $K = \int_{\mathcal{P}} \prod_{i=1}^n f_i(x_i)\, d\mathbf{x}$. Let us say that such distributions *factors into marginals*.

3 Characterisation of Distributions that Factors into Marginals

Theorem 1 *A probability distribution $f(\mathbf{x})$ factors into marginals if and only if its marginal distributions are*

$$f_i(x_i) = \frac{1}{(n-1)\left(1-\sum_{j=1}^n a_j\right)^{\frac{1}{n-1}}(x_i-a_i)^{\frac{n-2}{n-1}}}$$

with support $[a_i, 1-\sum_{j\neq i} a_j]$, where $\sum_{j=1}^n a_j < 1$.

Corollary 1 *A joint probability distribution function $f(\mathbf{x})$ on the probability simplex $\mathcal{P}$ factors into marginals if and only if*

$$f(x_1,\ldots,x_{n-1}) = \frac{(1-\sum_{i=1}^n a_i)\prod_{i=1}^n f_i(x_i)}{\Gamma^{1-n}\left(\frac{n}{n-1}\right)}\,,$$

where $x_n = 1-\sum_{i=1}^{n-1} x_i$, $f_i(x_i), i = 1,\ldots,n$ are the marginal distributions of f and $f_i(x_i) = 0$ for $x_i \geq a_i, x_i \leq 1-\sum_{j\neq i} a_j$.

When the first-order probability variables x_i are minimally restricted, i.e. $a_i = 0$, $\mathbf{x} = (x_1,\ldots,x_n)$ are Dirichlet distributed with parameters $\alpha_i = \frac{1}{n-1}$ and the marginal distributions f_i are Beta distributions $f(x;\alpha,\beta)$ with parameters $\alpha = \frac{1}{n-1}$ and $\beta = 1$. The marginal distributions f_i also have the same shape as Pareto distributions, but cut off so that the support has upper bound $1-\sum_{j\neq i} a_j$ rather than infinity.

We make a quick note on the degenerate case where $\sum_{j=1}^n a_j = 1$; then the marginal distributions are Dirac pulses $f_i(x_i) = \delta(x_i - a_i)$, i.e. all belief is concentrated in the points $x_i = a_i$ and the joint probability distribution is $\prod_{i=1}^n \delta(x_i - a_i)$.

We proceed with the proof of Theorem 1. The proof is based on the fact that an integral $\int_{\mathcal{P}} \prod_{i=1}^n g_i(x_i)\, d\mathbf{x}$ of a product of univariate functions over the probability simplex $\mathcal{P}$ is the repeated convolution $g_1 * g_2 * \cdots * g_n(1)$. E.g. when $n = 3$ we have

$$\int_0^1 \int_0^{1-x_1} g *_1 (x_1) g_2(x_2) g_3(1-x_1-x_2)\, dx_2\, dx_1 = \int_0^1 g_1(x_1)[g_2 * g_3(1-x_1)]\, dx_1 = g_1 * g_2 * g_3(1)\,.$$

If $f(\mathbf{x})$ factors into marginals the marginal distribution with respect to x_i is

$$\frac{1}{K} f_i(x_i) \underset{j\neq i}{\ast} f_j\,(1-x_i)\ ,$$

where $*_{i\neq j} f_j$ is the $n-1$-fold repeated convolution $f_1 * f_2 * \cdots * f_{i-1} * f_{i+1} * \cdots * f_n$ and K is the n-fold convolution $*_{i=1}^n f_i(1)$. Assume that $\{f_i\}_{i=1}^n$ are the marginal distributions of a joint distribution that factors into marginals. Then for all $i, i = 1,\ldots,n$,

$$\underset{j\neq i}{\ast} f_j(1-x_i) = KH(c_i - x_i) = KH((1-x_i)-(1-c_i))\,,$$

where c_i is such that $f_i(x_i) = 0$ when $x_i > c_i$.

Then the distributions f_k must have Laplace transforms F_k such that

$$\prod_{k\neq i} F_k = \frac{Ke^{-(1-c_i)s}}{s}$$

and if f_k is on the form $g_k(x_k - a_k)H(x_k - a_k)$ where $f_k(x_k) = 0$ when $x_k < a_k$, g_k must have Laplace transform $\left(\frac{K}{s}\right)^{\frac{1}{n-1}}$, that is $g_k(x_k) = \frac{K^{\frac{1}{n-1}}}{\Gamma\left(\frac{1}{n-1}\right)x_k^{\frac{n-2}{n-1}}}$ and

$$f_k(x_k) = \frac{K^{\frac{1}{n-1}} H(x_k - a_k)}{\Gamma\left(\frac{1}{n-1}\right)(x_k - a_k)^{\frac{n-2}{n-1}}}$$

since the Laplace transform of t^α is $\frac{\Gamma(1+\alpha)}{s^{1+\alpha}}$, where $\Gamma(1+\alpha) = \int_0^\infty e^{-x}x^\alpha\, dx$.

Further, since the Laplace transform of $f_k(x_k)$ is $\frac{K^{\frac{1}{n-1}}e^{-sa_k}}{s^{\frac{1}{n-1}}}$,

$$\underset{j\neq i}{\ast} f_j(1-x_i) = \frac{Ke^{-s(\sum_{j\neq i} a_j)}}{s}\,,$$

the upper limit of the support of x_i is $c_i = 1-\sum_{j\neq i} a_j$ and the n-fold convolution $\overset{n}{\underset{i=1}{\ast}} f_i(t)$ is the inverse

Laplace transform of $\frac{K^{\frac{n}{n-1}} e^{-s\sum_{i=1}^n a_i}}{s^{\frac{n}{n-1}}}$, i.e. $\overset{n}{\underset{i=1}{\circledast}} f_i(t) = \frac{K^{\frac{n}{n-1}} H(t-\sum_{i=1}^n a_i)(t-\sum_{i=1}^n a_i)^{\frac{1}{n-1}}}{\Gamma\left(\frac{n}{n-1}\right)}$, so

$$K = \overset{n}{\underset{i=1}{\circledast}} f_i(1) = \frac{K^{\frac{n}{n-1}} (1-\sum_{i=1}^n a_i)^{\frac{1}{n-1}}}{\Gamma\left(\frac{n}{n-1}\right)}$$

and $K = \frac{\Gamma^{n-1}\left(\frac{n}{n-1}\right)}{1-\sum_{i=1}^n a_i}$.

But since $\Gamma(z+1) = z\Gamma(z)$, $\Gamma\left(\frac{n}{n-1}\right) = \frac{1}{n-1}\Gamma\left(\frac{1}{n-1}\right)$ and

$$K = \frac{\Gamma^{n-1}\left(\frac{1}{n-1}\right)}{(n-1)^{n-1}\left(1-\sum_{i=1}^n a_i\right)}.$$

4 Some Properties of Second-Order Distributions that Factors into Marginals

The second-order probability distributions that factors into marginals are, as we have seen above, determined by the n-dimensional vector $(a_1, \ldots, a_n)$, where a_i is the lower bound of the support of the marginal distribution f_i. Thus we can by Corollary 1 define its probability density function with respect to the Lebesgue measure as

$$f(x_1, \ldots, x_{n-1}; a_1, \ldots, a_n) = \frac{(1-\sum_{i=1}^n a_i)\prod_{i=1}^n f_i(x_i)}{\Gamma^{1-n}\left(\frac{n}{n-1}\right)},$$

for $x_1, \ldots, x_{n-1}$ such that $\sum_{i=1}^n x_i \leq 1$ and where $\sum_{i=1}^n a_i < 1$.

Likewise the marginal distributions are

$$f_i(x_i, a_1, \ldots, a_n) = \frac{1}{(n-1)\left(1-\sum_{i=1}^n a_i\right)^{\frac{1}{n-1}} (x_i - a_i)^{\frac{n-2}{n-1}}}$$

with support $[a_i, 1-\sum_{j\neq i} a_j]$. When $a_i = 0$ for all $i = 1, \ldots, n$, $(x_1, x_2, \ldots, x_{n-1})$ have the Dirichlet distribution $f(x_1, \ldots, x_n; 1/(n-1), \ldots, 1/(n-1))$ and the individual variables x_i have Beta distributions $f(x; 1/(n-1), 1)$.

Regarding the intervals of support, one may choose the lower bounds a_i freely as long as the sum $a_1 + \cdots + a_n$ is less than one (and lower bounds summing to a number greater than one is unreasonable since the variables x_i have a sum less than one). But if we want a joint second-order distribution that factors into marginals, the upper bounds are determined by the lower bounds a_i. A consequence of this is that arbitrary support intervals are not in general possible to reconcile with this type of distributions. If the support intervals of the marginal distributions are $[a_i, b_i]$, we cannot form the joint second-order probability distribution as the normalised product of the marginal distributions unless $b_i = 1 - \sum_{j\neq i} a_i$.

Let us list some properties of the marginal distributions; if a second-order probability distribution f with parameters $a_1, \ldots a_n$ factors into marginals the univariate marginal distributions have

- mean $a_i + \frac{1-\sum_{j=1}^n a_j}{n}$,
- median $a_i + \left(\frac{1-\sum_{i=1}^n a_i}{2}\right)^{n-1}$ and
- variance $\frac{(n-1)^2}{n^2(2n-1)}\left(1-\sum_{j=1}^n a_j\right)^2$.

4.1 Multivariate Marginal Distributions

We may generalise the argument in the proof of Theorem 1 to achieve the multivariate marginal distribution of $x_1, \ldots x_k, k < n$ as

$$\frac{1}{K}\prod_{i=1}^k f_i(x_i) \overset{n}{\underset{i=k+1}{\circledast}} f_i\left(1-\sum_{i=1}^k x_i\right).$$

Since the Laplace transform of $f_i(x_i)H(x_i - a_i)$ is $\frac{K^{\frac{1}{n-1}} e^{-sa_i}}{s^{\frac{1}{n-1}}}$ with $K = \frac{\Gamma^{n-1}\left(\frac{n}{n-1}\right)}{1-\sum_{i=1}^n a_i}$ we have the following Corollary.

Corollary 2 *If $f(x_1, \ldots, x_{n-1})$ is a second-order probability distribution that factors into marginals, the multivariate marginal distribution $f(x_1, x_2, \ldots, x_k)$ is*

$$\frac{\prod_{i=1}^k f_i(x_i)\left(1-\sum_{i=1}^n a_i\right)^{\frac{k-1}{n-1}}}{\Gamma\left(\frac{n-k}{n-1}\right)\Gamma^{k-1}\left(\frac{n}{n-1}\right)\left(1-\sum_{i=1}^k x_i - \sum_{i=k+1}^n a_i\right)^{\frac{k-1}{n-1}}}.$$

Corollary 2 in turn gives us a result on conditional distributions.

Corollary 3 *The conditional distribution of x_k given $x_1, x_2, \ldots, x_{k-1}$ is*

$$C\frac{f_k(x_k)\left(1-\sum_{i=1}^{k-1} x_i - \sum_{i=k}^n a_i\right)^{\frac{k-2}{n-1}}}{\left(1-\sum_{i=1}^k x_i - \sum_{i=k+1}^n a_i\right)^{\frac{k-1}{n-1}}},$$

where

$$C = \frac{\Gamma\left(\frac{n-k+1}{n-1}\right)\left(1-\sum_{i=1}^n a_i\right)^{\frac{1}{n-1}}}{\Gamma\left(\frac{n}{n-1}\right)\Gamma\left(\frac{n-k}{n-1}\right)}$$

if $x_i, i = 1, \ldots n-1$, are distributed by a second-order probability distribution that factors into marginals.

5 Examples

Example 1 *With* $n = 3$, *let us take* $a_1 = 1/3, a_2 = 1/5$ *and* $a_3 = 1/8$. *Then* $1 - \sum_{i=1}^{3} a_i = \frac{120-40-24-15}{120} = \frac{41}{120}$.

$$f_1(x_1) = \frac{1}{2\sqrt{41/120(x_1 - 1/3)}},$$

$$f_2(x_2) = \frac{1}{2\sqrt{41/120(x_2 - 1/5)}}$$

and

$$f_3(x_3) = \frac{1}{2\sqrt{41/120(x_3 - 1/8)}},$$

with support $[1/3, 27/40], [1/5, 13/24]$ *and* $[1/8, 7/15]$ *and mean* $\frac{161}{360}, \frac{113}{360}$ *and* $\frac{43}{180}$, *respectively.*

The joint distribution $f(x_1, x_2)$ *is*

$$\frac{41 f_1(x_1) f_2(x_2) f_3(1 - x_1 - x_2)}{120\Gamma^2(3/2)} = \frac{\sqrt{120/41}}{\Gamma^2(3/2)\sqrt{(x_1 - 1/3)(x_2 - 1/5)(7/8 - x_1 - x_2)}},$$

see Figure 1 for a plot.

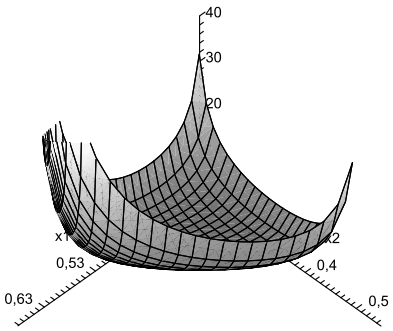


Figure 1: The joint probability density function of Example 1

Given that we in Example 1 wanted to represent knowledge about lower bounds on probabilities, the joint and marginal distribution seem rather rich in information and far from uniform. But we do not wish to minimise entropy in either the joint or the marginal distributions, instead the goal is to balance the entropy of the joint distributions and the marginals. The local effects of maximising entropy globally have partly been studied in [13]. To further study this effect and the converse global effect of local entropy maximisation is a topic for future research.

Example 2 *Let* $n = 5$ *and* $a_i = 1/10$. *Then* f_1, f_2, f_3, f_4, f_5 *where* $f_i(x) = f(x) = \frac{1}{4(1/2)^{1/4}(x-1/10)^{1/4}} = \frac{1}{2^{7/4}(x-1/10)^{3/4}}$ *are the marginal distributions of a second-order distribution that factors into its own marginals, in Figure 2 we see a plot of the marginal distributions.*

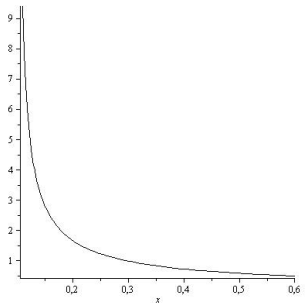


Figure 2: The marginal probability density functions $f_i(x) = 2^{-7/4}(x - 1/10)^{-3/4}$ of Example 2 for $1/10 \leq x \leq 3/5$.

The support of $f_i(x_i)$ *is* $[1/10, 3/5]$. *The means are* $\mu_i = 1/5$ *and*

$$f(x_1, x_2, x_3, x_4) = \frac{(1 - 1/2)\prod_{i=1}^{5} f_i(x_i)}{\Gamma^4(5/4)}.$$

The three variable marginal distribution $f(x_1, x_2, x_3)$ *is*

$$\frac{f(x_1) f(x_2) f(x_3)}{\Gamma(1/2)\Gamma^2(5/4)\sqrt{2}\sqrt{4/5 - x_1 - x_2 - x_3}},$$

and the conditional distribution $f(x_4|x_1, x_2, x_3)$ *is*

$$\frac{f(x_4)\Gamma(1/2)\sqrt{4/5 - x_1 - x_2 - x_3}}{2^4\Gamma(5/4)\Gamma(1/4)(9/10 - x_1 - x_2 - x_3 - x_4)^{3/4}}$$

E.g. The conditional distribution $f(x_4|x_1 = 1/10, x_2 = 1/5, x_3 = 2/5)$ *is shown in Figure 3.*

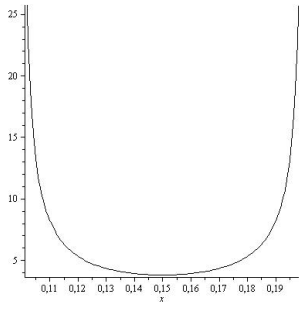


Figure 3: The conditional density of x_4 given $x_1 = 1/10, x_2 = 1/5, x_3 = 2/5$ in Example 2

The two variable marginal distribution $f(x_2, x_4)$ *is*

$$\frac{f(x_2)f(x_4)}{2^{1/4}\Gamma(3/4)\Gamma(5/4)(7/10 - x_2 - x_4)^{1/4}},$$

and the conditional distribution $f(x_1|x_2, x_4)$ *is*

$$\frac{f_1(x_1)\Gamma(3/4)(7/10 - x_2 - x_4)^{1/4}}{2^{1/4}\Gamma(5/4)\Gamma(1/2)(3/5 - x_1 - x_2 - x_4)^{1/2}}$$

The variance of f_i *is*

$$\frac{4^2}{5^2 9}(1/2)^2 = \frac{4}{225}.$$

6 Conclusion

We have found a characterisation of the second-order probability distributions that can be expressed as a normalised product of its own marginal distributions. For such distributions there is a direct path from local to global information. From an information-theoretical standpoint, such probability distributions are unique in that given that the variables are probabilities, no information is either lost or gained when going between the joint distribution and the univariate marginal distributions.

The family of distributions with the properties mentioned above can be said to be a generalisation of a special case of the Dirichlet distribution. When all lower bounds on the first-order probabilities are zero, we get the Dirichlet distribution will all parameters equal to $1/(n-1)$, where n is the number of possible outcomes whose probabilities are the variables. In this case, of course, the marginal distributions are Beta distributions. But in general, with first-order probability variables bounded from below by positive numbers, we have shifted and re-scaled versions of the Dirichlet and Beta distributions, respectively. It is a matter for future research to investigate to which degree properties of Dirichlet and Beta distributions carry over to their shifted counterparts.

References

[1] P. Billingsley. *Probability and Measure*. John Wiley & Sons, Inc., 1995.

[2] M. Danielson and L. Ekenberg. A Framework for Analysing Decisions under Risk. *European Journal of Operational Research, Vol. 104, Issue 3*, pages 474–484, 1998.

[3] L. Ekenberg, M. Andersson, M. Danielson, and A. Larsson. Distributions over Expected Utilities in Decision Analysis. *Proceedings of ISIPTA '07*, 2007.

[4] L. Ekenberg and J. Thorbiörnson. Second-Order Decision Analysis. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 9, No 1*, 9(1):13–38, 2001.

[5] L. Ekenberg, J. Thorbiörnson, and T. Baidya. Value Differences Using Second-order Distributions. *International Journal of Approximate Reasoning*, 38(1):81–97, 2005.

[6] E.T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review*, 106(4):620–630, 1957.

[7] E.T. Jaynes. Information Theory and Statistical Mechanics II. *Physical Review*, 108(2):171–190, 1957.

[8] S. Kullback and R.A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[9] I. Levi. *The Enterprise of Knowledge*. MIT Press, 1980.

[10] A. W. Marshall and I. Olkin. Families of Multivariate Distributions. *Journal of the American Statistical Assoiciation*, 83(403):834–841, September 1988.

[11] R. F. Nau. Uncertainty Aversion with Second-Order Utilities and Probabilities. *Management Science*, 52(1):136–145, 2006.

[12] R. B. Nelsen. *An introduction to copulas, Volume 139 of Lecture Notes in Statistics*. Springer Verlag, 1999.

[13] D. Sundgren, M. Danielson, and L. Ekenberg. Some Second Order Effects on Interval Based Probabilities. In *Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference*, pages 848–853. AAAI Press, 2006.

[14] L. Utkin and T. Augustin. Decision Making with Imprecise Second-Order Probabilities. In *ISIPTA '03 - Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications*, pages 547–561, 2003.

[15] P. Walley. *Statistical reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

[16] P. Walley. Towards a Unified Theory of Imprecise Probability. *International Journal of Approximate Reasoning*, 24(2-3):125–148, 2000.

[17] S. Watanabe. Infomation Theoretical Analysis of Multivariate Correlation. *IBM Journal of Research and Development*, 4(1):66–82, 1960.

Multi-criteria decision making with a special type of information about importance of groups of criteria

Lev V. Utkin
Department of Computer Science
St.Petersburg Forest Technical Academy
lev.utkin@mail.ru

Abstract

An axiomatic approach for solving a multi-criteria decision making problem is studied in the paper, which generally allows reducing a set of Pareto optimal solutions. The information about criteria in the problem is represented as the decision maker judgments of a special type. The judgments have a clear behavior interpretation and can be used in various decision problems. It is shown in the paper how to combine the judgments and to use them for reducing the Pareto set when they are provided by several decision makers. Two global criteria of decision making are introduced for comparing of decision alternatives. The first criterion based on the lower expectation, the second one is based on determining the belief and plausibility functions in the framework of Dempster-Shafer theory and uses the "threshold" probability for the final decision making. The numerical examples illustrate the proposed approach.

Keywords. Multi-criteria decision making, desirable gambles, Dempster-Shafer theory, judgments, preferences, Pareto set

1 Introduction

Most methods of *multi-criteria decision making* (MCDM) problems are somehow or other based on combining or aggregating *criteria*. According to these methods, *decision alternatives* (DA's) are compared by using the aggregated criterion. There are different ways for criteria combining. The widely-spread ways are linear, multiplicative and maximin combinations [3, 8]. For instance, the well-known analytic hierarchy process method proposed by Saaty [8] is based on the linear combination of criteria. However, in spite of the popularity of the aggregation methods for solving MCDM problems, they are ad hoc and have some justification related to certain applied areas. The main shortcoming of ad hoc methods is that it is often difficult to validate or to justify the optimal solutions. Another shortcoming is the necessity to have criteria with identical numerical scales.

Another part of methods is axiomatic, i.e., they are based on some axioms or properties and can be called strong methods. One of such the methods is reducing the so-called Pareto set of non-dominated solutions by utilizing some additional information about importance of criteria provided by experts, decision makers (DM's), etc. The amount of the additional information and its consistency determines the number of solutions in a reduced Pareto set. Ideally, the reduced Pareto set should consist of one solution.

Procedures for processing the additional information and for reducing the Pareto set totally depend on the type of available data or judgments. Many authors use the "weights" of criteria $\mathbf{v} = (v_1, ..., v_r)$ and different kinds of their ranking. For instance, Park and Kim [7], Kim and Ahn [4] distinguish between the following approaches to the elicitation of attribute weights: weak ranking ($v_i \geq v_j$); strict ranking ($v_i - v_j \geq \lambda_i$); ranking with multiples ($v_i \geq \lambda_i v_j$); interval form: ($\lambda_i \leq v_i \leq \lambda_i + \epsilon_i$); ranking of differences ($v_i - v_j \geq v_k - v_l$). Here $\lambda_i \geq 0$, $\epsilon_i \geq 0$.

Another very interesting type of judgments elicited from DM's or experts has been proposed by Noghin [5, 6] for reducing the Pareto optimal set in the framework of his theory of relative importance of criteria. Some details of the theory will be considered below. This type of judgments does not require to have identical numerical scales for criteria. It has a simple and clear behavior interpretation. Moreover, it turns out that many statements of the theory have analogues in the framework of desirable gambles [13, 14]. Therefore, MCDM problems in the framework of desirable gambles by relying on the Noghin's theory of relative importance of criteria are studied in the paper.

An interesting approach for eliciting the additional judgments from DM's or experts in MCDM problems (called the *DS/AHP method*) has been proposed by

Beynon *et al* [1, 2]. This method uses *Dempster-Shafer theory* in a framework of the analytic hierarchy process and allows to compare not only single DA's, but also groups of DA's. Beynon *et al* proposed to compare DA's, not criteria. However, a similar elicitation procedure can be applied to criteria [12]. Nevertheless, this method is also ad hoc and uses in the long run the linear aggregation (the analytic hierarchy process). Therefore, an attempt to modify it for reducing the Pareto optimal set in the framework of desirable gambles and Noghin's theory is made in the paper.

In the paper, "interval-valued" judgments as the extension of judgments proposed by Beynon *et al* are analyzed. These judgments have the following form: "I do not know which criterion is the most important, but this criterion belongs to the subset B of the set of criteria". Then these simple judgments are generalized to a more complex form, for instance, "I'm willing to pay w_i for the i-th criterion in order to get w_j for the j-th criterion. I'm also willing to pay w_k for the k-th criterion in order to get w_l for the l-th criterion. However, I do not know what is better." Their analysis is the next task for solving in the paper.

At the same time, we have to note that these judgments can be provided by several DM's. Therefore, the following problem to be solved is to combine them by taking into account the quality or weights of DM's. This will be done by introducing two global criteria of decision making, which are based on some statements of the Dempster-Shafer theory. In fact, these criteria can be regarded in the framework of second-order models [11]. It will be shown in the paper how to reduce the Pareto set of optimal solutions by applying the given information in the above forms and by using two proposed criteria.

The paper is organized as follows. The main definitions of MCDM are provided in Section 2. Some statements of Noghin's theory of relative importance of criteria are considered in Section 3. Noghin's theory is formulated in the framework of desirable gambles in Section 4. Different types of "interval-valued" judgments about criteria and their use for reducing a Pareto set are studied in Section 5. An illustrative numerical example is considered in the same section.

2 The MCDM problem statement

A general MCDM problem can be formulated in the following way. Suppose that there is a set of DA's $\mathbb{X} = \{X_1, ..., X_n\}$ consisting of n elements. Moreover, there is a set of criteria $\mathbb{C} = \{C_1, ..., C_r\}$ consisting of r elements, $r \geq 2$. For every DA, say the k-th DA, we can write the value of the i-th criterion $C_i(X_k)$ briefly denoted x_{ki}, $k = 1, ..., n$, $i = 1, ..., r$. Below, we will say that the i-th DA is characterized by the vector $X_i = (x_{i1}, ..., x_{ir})$. We will assume that the number of criteria and the number of DA's are finite.

To solve a MCDM problem is to find a set of all optimal solutions denoted by $\text{Opt}\mathbb{X} \subseteq \mathbb{X}$, which can be regarded as the best solutions under certain conditions.

By making decisions, we usually have to take many objectives or criteria into account. The main feature here is that the different objectives are most likely conflicting and the final decision is commonly called a trade-off. When dealing with multiple objectives, solutions can be incomparable since they can dominate each other in different objectives. This lead to the notion of *Pareto optimality*, which is based on a partial order among the solutions. A solution is called Pareto optimal, if it is not dominated by any other solution, that is, if there is no other solution that is better in at least one objective and not worse in any of the other objectives. Naturally, Pareto optimal solutions are the candidates for a trade-off.

Let us give some standard definitions related to Pareto optimal solutions under assumption that there is no information about importance of criteria.

Definition 1 *$X \in \mathbb{X}$ dominates $Y \in \mathbb{X}$, denoted $X \succ Y$ iff $\forall i = 1, ..., r$, $x_i \geq y_i$ with at least one strict inequality.*

Definition 2 *$Y \in \mathbb{X}$ is a Pareto optimal alternative, also called an efficient alternative, iff $\nexists X \in \mathbb{X}$ such that $X \succ Y$. The set of all Pareto optimal alternatives in $\mathbb{X}$ or Pareto set is denoted $\mathcal{P}(\mathbb{X})$.*

It follows from the above definitions that the following inclusions are valid $\text{Opt}\mathbb{X} \subseteq \mathcal{P}(\mathbb{X}) \subseteq \mathbb{X}$.

For many optimization problems, the number of Pareto optimal solutions can be rather large. Therefore, the problem of reducing Pareto optimal sets by obtaining the additional information is very important.

3 Noghin's relative importance of criteria

For reducing the Pareto optimal set, Noghin in [5] proposed the so-called theory of relative importance of criteria. This theory is based on the standard axioms and definitions of Pareto optimal solutions and the following additional axiom.

Axiom 1 *The preference relation $\succ$ is invariant with*

respect to positive linear transformation[1].

The main idea of Noghin's theory is to compare criteria by means of parameters.

Definition 3 *Let $i, j \in N = \{1, 2, ..., r\}$, $i \neq j$. We say that the i-th criterion is more important than the j-th criterion with two positive parameters w_i and w_j if for any two vectors $X, Y \in \mathbb{X}$ such that*

$$x_i > y_i,\ x_j < y_j,\ x_k = y_k,\ \forall k \in N \backslash \{i, j\},$$

$$x_i - y_i = w_i,\ x_j - y_j = -w_j,$$

the relationship $X \succ Y$ is valid.

A behavior interpretation of the parameters w_i and w_j is the following. The DM is willing to pay w_j units for the j-th criterion in order to get w_i units for the i-th criterion. The relative importance coefficient is defined as

$$\theta_{ij} = \frac{w_j}{w_i + w_j}.$$

It can be seen that $0 < \theta_{ij} < 1$. At that, θ_{ij} is close to 1 if $w_j \gg w_i$. Moreover, θ_{ij} is close to 0 if $w_j \ll w_i$.

Introduce the following vector

$$W_{ij} = (0, ..., 0, w_i, 0, ..., -w_j, 0, ..., 0),$$

whose $r - 2$ elements are zero, the i-th element is w_i, the j-th element is $-w_j$. If the relation $X \succ Y$ is valid with the given parameters w_i and w_j, then we can write that the relation $W_{ij} \succ 0_r$ is valid. Here 0_r is the vector of r zero elements. The relation $W_{ij} \succ 0_r$ is equivalent to the relation $\Theta_{ij} \succ 0_r$, where

$$\Theta_{ij} = (0, ..., 0, 1 - \theta_{ij}, 0, ..., -\theta_{ij}, 0, ..., 0).$$

One of the main results of Noghin's theory of the relative importance of criteria is the following theorem.

Theorem 1 (Noghin [5]) *Let the i-th criterion be more important than the j-th criterion with the pair of positive parameters w_i and w_j. Then for any nonempty set of optimal vectors* Opt$\mathbb{X}$*, it follows that*

$$\text{Opt}\mathbb{X} \subseteq \mathcal{P}^*(\mathbb{X}) \subseteq \mathcal{P}(\mathbb{X}),$$

where $\mathcal{P}(\mathbb{X})$ is a set of Pareto-optimal vectors with respect to criteria $\mathbb{C} = \{C_1, ..., C_r\}$; $\mathcal{P}^*(\mathbb{X})$ is a set of Pareto-optimal vectors with respect to criteria $\mathbb{C}^* = \{C_1^*, ..., C_r^*\}$ such that

$$C_j^* = w_j C_i + w_i C_j,\ C_k^* = C_k,\ k \neq j.$$

[1] A binary relation $\mathcal{R}$ defined on $\mathbb{R}^r$ is said to be invariant with respect to positive linear transformation if for any vectors $X, Y, c \in \mathbb{R}^r$ and each positive number α the relationship $X \mathcal{R} Y$ implies $(\alpha X + c)\, \mathcal{R}\, (\alpha Y + c)$.

In other words, Theorem 1 provides a simple computation way for reducing the Pareto optimal set $\mathcal{P}(\mathbb{X})$. Its proof is based on properties of convex cones [6] produced by preferences of the form $W_{ij} \succ 0$. Theorem 1 is very important because it is a tool for dealing with the information about the relative importance of criteria. It can be easy written in terms of the relative importance coefficients θ_{ij}.

4 Sets of desirable gambles

A goal of this section is to consider Noghin's theory of the relative importance of criteria in the framework of desirable gambles [13, 14] and to show that its results and statements can be rather simply obtained on the basis of the framework. Preliminaries of the framework of desirable gambles given below can be found in [14].

Let Ω denote the set of possible outcomes under consideration. A bounded mapping from Ω to $\mathbb{R}$ (the real numbers) is called a *gamble*. Let $\mathcal{L}$ be a nonempty set of gambles. A mapping $\underline{P} : \mathcal{L} \to \mathbb{R}$ is called a lower prevision or lower expectation. The lower prevision of a gamble X is interpreted as a supremum buying price for X, meaning that it is acceptable to pay any price smaller than $\underline{P}(X)$ for the uncertain reward X. A lower prevision is said to be coherent when it is the lower envelope of some set of linear expectations, i.e., when there is a nonempty set of probability measures, $\mathcal{M}$, such that $\underline{P}(X) = \inf\{E_P(X) : P \in \mathcal{M}\}$ for all $X \in \mathcal{L}$, where $E_P(X)$ denotes the expectation of X with respect to P. The conjugate upper prevision is determined by $\overline{P}(X) = -\underline{P}(-X)$. It is interpreted as an infimum selling price for X.

For $X, Y \in \mathcal{L}$, write $X \geq Y$ to mean that $X(\omega) \geq Y(\omega)$ for all $\omega \in \Omega$, and write $X > Y$ to mean $X \geq Y$ and $X(\omega) > Y(\omega)$ for some $\omega \in \Omega$. According to Walley [13], a gamble X is *inadmissible* in $\mathcal{L}$ when there is $Y \in \mathcal{L}$ such that $Y \geq X$ and $Y \neq X$. Otherwise X is *admissible* in $\mathcal{L}$. The subset $\mathcal{P}$ of admissible gambles in $\mathcal{L}$ is an analogue of the Pareto set in MCDM. A set of desirable gambles, denoted by $\mathcal{D}$, is a subset of $\mathcal{L}$. A set of desirable gambles is said to be *coherent* when it satisfies the four axioms:

D1. $0 \notin \mathcal{D}$.

D2. if $X \in \mathcal{L}$ and $X > 0$, then $X \in \mathcal{D}$.

D3. if $X \in \mathcal{D}$ and $c \in \mathbb{R}_+$, then $cX \in \mathcal{D}$.

D4. if $X \in \mathcal{D}$ and $Y \in \mathcal{D}$, then $X + Y \in \mathcal{D}$.

Thus a coherent set of desirable gambles is a convex cone of gambles that contains all positive gambles ($X > 0$) but not the zero gamble. Consequence of the axioms: If $X \in \mathcal{D}$ and $Y \geq X$, then $Y \in \mathcal{D}$.

It can be seen from the axioms of coherence that D3 and D4 coincide with Axiom 1 about positive linear transformation used by Noghin in his theory. Moreover, it can be seen from Definition 3 that assessments of the parameters w_i and w_j can be regarded as some extension of the probability ratios studied by Walley [13]. The probability ratios generalize the comparative probability judgments and have the form "A is at least l times as probable as B", where l is a positive number. The gamble $A - lB$ is almost desirable. This implies that $A \succ lB$.

Walley states that there is a one-to-one correspondence between coherent sets of desirable gambles and coherent partial preference orderings, defined by $X \succ Y$ if and only if $X - Y \in \mathcal{D}$. This is very important statement which allows to find the same correspondence between the framework of desirable gambles and Noghin's theory.

If a closed convex set of probability measures $\mathcal{M}$ is given, then we can define a set of desirable gambles as follows:

$$\mathcal{D} = \{X \in \mathcal{L} : X > 0 \text{ or } \mathbb{E}_P(X) > 0, \forall P \in \mathcal{M}\}. \quad (1)$$

Then $\mathcal{D}$ is coherent and $\mathcal{M}$ can be recovered from it by

$$\mathcal{M} = \{P : \mathbb{E}_P(X) \geq 0, \ \forall X \in \mathcal{D}\}. \quad (2)$$

Note that (1) can be rewritten as

$$\mathcal{D} = \{X \in \mathcal{L} : X > 0 \text{ or } \ \underline{\mathbb{E}}_{\mathcal{M}}(X) > 0\}. \quad (3)$$

Suppose that we have information about the relative importance of the i-th and the j-th criteria, i.e. the i-th criterion is more important than the j-th criterion with two positive parameters w_i and w_j. Let us return to the vector W_{ij} produced by the parameters w_i, w_j and consider again the relation $W_{ij} \succ 0_r$ (see Section 3). This relation can be written in the framework of desirable gambles as the condition $W_{ij} - 0_r \in \mathcal{D}$ or just $W_{ij} \in \mathcal{D}$. In other words, the information about the relative importance of the i-th and the j-th criteria can be represented as the condition that the vector W_{ij} belongs to the set of desirable gambles.

Now we reformulate Noghin's theorem and prove it in terms of desirable gambles.

Let X and Y be two DA's. We will denote below the vector $Z = X - Y$ and its components $z_k = x_k - y_k$ for short.

Theorem 2 *The preference $X \succ Y$ is valid if $X^* > Y^*$ and $W_{ij} \in \mathcal{D}$. Here $X^* = (x_1^*, ..., x_r^*)$ and $Y^* = (y_1^*, ..., y_r^*)$ such that*

$$x_j^* = w_j x_i + w_i x_j, \ x_k^* = x_k, \ k \neq j,$$

$$y_j^* = w_j y_i + w_i y_j, \ y_k^* = y_k, \ k \neq j.$$

Proof. Note that $X \succ Y$ if $X - Y = Z \in \mathcal{D}$ or $\mathbb{E}_P(Z) > 0$ for all $P \in \mathcal{M}$. The condition $W_{ij} \in \mathcal{D}$ restricts the set $\mathcal{M}$ of possible probability measures by the constraint $\mathbb{E}_P(W_{ij}) \geq 0$. If we denote $P = (\pi_1, ..., \pi_r)$, then the above constraint can be rewritten as $w_i\pi_i - w_j\pi_j \geq 0$. This implies that the set of all probability measures $\mathcal{M}$ is reduced to the subset $\mathcal{M}(ij) \subseteq \mathcal{M}$. The subset $\mathcal{M}(ij)$ is defined by the constraints

$$\sum_{k=1}^{r} \pi_k = 1, \ \pi_k \geq 0, \ \forall k \in N,$$

$$w_i\pi_i - w_j\pi_j \geq 0.$$

Here $N = \{1, 2, ..., r\}$.

Let us find extreme points of $\mathcal{M}(ij)$. They are

$$(0, ..., 0, 1_k, 0, ..., 0), \ \forall k \in N \backslash \{j\},$$

and

$$\pi_i = \frac{w_j}{w_i + w_j}, \ \pi_j = \frac{w_i}{w_i + w_j},$$
$$\pi_k = 0, \ \forall k \in N \backslash \{j\}.$$

The last extreme point is produced by the equality $w_i\pi_i - w_j\pi_j = 0$.

The extreme points define the set of probability distributions $\mathcal{M}(ij)$. Therefore, if we prove that the inequality $\mathbb{E}_P(Z) > 0$ is valid for extreme points, then this inequality will be valid for all $P \in \mathcal{M}(ij)$. The first $k - 2$ extreme points give

$$\mathbb{E}_P(Z) = z_k, \ \forall k \in N \backslash \{i, j\}.$$

The last extreme point gives

$$\mathbb{E}_P(Z) = \pi_i z_i + \pi_j z_j = \frac{w_j z_i}{w_i + w_j} + \frac{w_i z_j}{w_i + w_j}.$$

At the same time, the condition $X^* > Y^*$ implies that $z_k > 0$ or $z_k = 0$ for all $k \neq j$, and $w_j z_i + w_i z_j > 0$. Hence $\mathbb{E}_P(Z) > 0$ for all $P \in \mathcal{M}(ij)$ and $X \succ Y$, as was to be proved. ■

Example 1 *Consider the simplified and modified example of the optimal choice of a place for the airport construction given by Keeney and Raiffa in their book [3] and solve it by using Noghin's theory. The problem is to decide where a new airport should be constructed in accordance with the following criteria[2]: minimize investment of capital in million dollars (C_1), maximize carrying capacity in the daily number of air travellers (C_2), maximize safety expressed in the 9-point*

[2]The example is given with some changes.

scale from 1 *till* 9 *(C_3), minimize remoteness in kilometers (C_4). There are four places for the construction (DA's) denoted X_1, X_2, X_3, X_4. The MCDM problem can also be represented by means of the matrix*

	C_1	C_2	C_3	C_4
X_1	-20	15000	6	-3
X_2	-30	25000	4	-1
X_3	-40	18000	7	-5
X_4	-25	20000	5	-2

.

Here negative values are taken in order to replace the "minimization" goals by the "maximization" ones.

It can be seen from the matrix that all the DA's belong to the Pareto set.

The DM is willing to pay $w_3 = 3$ units for the third criteria in order to get $w_1 = 2$ units for the first criterion. The provided information can be represented by the gamble $W_{13} = (2, 0, -3, 0) \in \mathcal{D}$ or equivalently by the gamble $\Theta_{13} = (2/5, 0, -3/5, 0) \in \mathcal{D}$.

Then we write the modified matrix by using Noghin's theorem

	C_1	C_2	$3 \cdot C_1 + 2 \cdot C_3$	C_4
X_1	-20	15000	-48	-3
X_2	-30	25000	-82	-1
X_3	-40	18000	-106	-5
X_4	-25	20000	-65	-2

.

Hence, we reduce the Pareto set which now consists of three DA's X_1, X_2 and X_4. It can be seen that it is not enough to have the supplied judgment for getting one optimal solution.

5 Groups of the most important or preferable criteria

Let us quickly return to the analytic hierarchy process method. In addition to the fact that it must perform very complicated and numerous pairwise comparisons amongst alternatives the method uses precise estimates of experts or DM's. This condition can not be satisfied in many applications because judgments elicited from experts are usually imprecise and unreliable due to the limited precision of human assessments. In order to overcome these difficulties and to extend the analytic hierarchy process on a more real elicitation procedures, Beynon *et al* [1, 2] proposed a method using Dempster-Shafer theory and called the DS/AHP method. The method was developed for decision making problems with a single DM, and it applies the analytic hierarchy process method for collecting the preferences from the DM and for modelling the problem as a hierarchical decision tree. It should be noted that the main idea underlying the DS/AHP method is not applying Dempster-Shafer theory to the analytic hierarchy process method. It is comparison of groups of alternatives with a whole set of alternatives. Such the type of comparison is equivalent to the preferences stated by the DM. In other words, Beynon *et al* [1, 2] proposed to consider preferences of the form $B \succ \mathbb{X}$ with some degree v of it, where B is a subset or a group of DA's, $\mathbb{X}$ is the set of all alternatives, v is a positive number in accordance with some scale [8]. The same can be carried out for the criteria, i.e., we can consider preferences of the form $D \succ \mathbb{C}$, where D is a subset of criteria. It is obvious that this preference can be rewritten in the form $D \succ \mathbb{C} \backslash D$. The authors of the papers [1, 2] assign to every subset B some basic probability assignment (BPA) [9] denoted $m(B)$. The same can be done for criteria.

Such the elicitation procedure has some virtues. First, a DM does not need to choose the most important criterion from the set of criteria. The DM chooses a subset of criteria by assuming that one of these criteria is the most important or important with some degree of importance. However, these judgments are used in the aforementioned aggregating criteria methods which are ad hoc. As a result, it is difficult to validate the approach in specific applied problems.

Now we will formalize the above elicitation procedure in the framework of Noghin's theory and desirable gambles. Then we will study how this procedure can be applied to reducing the set of Pareto optimal solutions.

Suppose that there is a set of t judgments of the form $D_l \succ \mathbb{C}$ with the corresponding BPA's $m(D_l)$, $l = 1, ..., t$. The first question is to construct a criterion (criteria) for the validity of the preference $X \succ Y$. These criteria will be called global in order to distinguish them from the criteria $C_1, ..., C_r$ of the considered MCDM problem.

The second question is the computation rules for the validity of $X \succ Y$.

5.1 Simple comparison judgments

First we consider simple comparison judgments of the form: "I do not know which criterion is the most important, but this criterion belongs to the subset $B \subseteq \mathbb{C}$". Here the degree v is assumed to be unknown. Suppose that the unknown important criterion has the number k and the subset B contains t elements with numbers from the index set[3] B^0. Denote $B^1 = N \backslash B^0$, $N = \{1, ..., r\}$. Then we can provide

[3]The set of indices of elements of B will be denoted B^0. The set of indices of elements of $\mathbb{C} \backslash B$ will be denoted B^1.

$r-t$ judgments:

"The k-th criterion is more important than the j-th criterion from $\mathbb{C}\backslash B$ with the pair of positive parameters $w_k = 1$ and $w_j = 1$".

Here $k \in B^0$ and $j \in B^1$. So, every judgment produces the gamble[4]

$$W_{kj} = (0, ..., 0, 1_k, 0, ..., -1_j, 0, ..., 0), \qquad (4)$$

such that $W_{kj} \succ 0_r$, $k \in B^0$, $j \in B^1$.

It should be noted that the simple comparison judgment with the above desirable gamble W_{kj} can be applied to decision problems with uniform criteria, i.e., criteria have identical numerical scales.

Now we can find the subset $\mathcal{M}(k, B^1) \subseteq \mathcal{M}$ of probability distributions $P = (\pi_1, ..., \pi_r)$ restricted by the desirable gambles W_{kj}, $j \in B^1$, or equivalently its extreme points. The subset $\mathcal{M}(k, B^1)$ is produced by the judgment about comparison of the k-th criterion and the j-th criterion.

Proposition 1 *Given the additional information in the form (4), the preference $X \succ Y$ is valid if the condition*

$$z_k + \sum_{j \in L} z_j \geq 0$$

is valid for all $L \subseteq B^1$ and $z_i \geq 0$ for all $i \in B^0$.

Proof. Let us find the subset $\mathcal{M}(k, B^1)$. It follows from (2) and from (4) that this set is produced by the constraints[5]

$$\pi_k - \pi_j \geq 0,\ j \in B^1,\ \pi_i \geq 0,\ i \in N,$$

$$\pi_1 + \pi_2 + ... + \pi_r = 1.$$

Consider r equalities instead of inequalities in the above constraints. Hence, we get extreme points of the form:

$$\begin{aligned}
&\pi_k = 1,\ \pi_i = 0,\ \forall i \neq k,\\
&\pi_k = 1/2,\ \pi_{j_1} = 1/2,\ j_1 \in B^0,\\
&\pi_k = 1/3,\ \pi_{j_1} = \pi_{j_2} = 1/3,\ j_1, j_2 \in B^0,\\
&\quad ...\\
&\pi_k = 1/(r-t+1),\ \pi_{j_i} = 1/(r-t+1),\\
&j_i, \in B^0,\ i = 1, ..., r-t.
\end{aligned}$$

Only non-zero elements of extreme points are written here. The proof directly follows from the condition of desirability of gambles $X - Y$, which is of the form: $\mathbb{E}_P(X - Y) \geq 0$, $\forall P \in \text{extr}(\mathcal{M}(k, B^1))$. ■

Several conditions in Proposition 1 can be replaced by one equivalent condition

$$z_k + \min_{L \subseteq B^1} \sum_{j \in L} z_j \geq 0. \qquad (5)$$

So, the Pareto set can be reduced by using condition (5) for every pair of DA's.

It also follows from the proof of Proposition 1 and from (5) that the lower expectation of the gamble $Z = X - Y$ under conditions $W_{kj} \succ 0_r$, $j \in D_l^1$, denoted $\underline{\mathbb{E}}_{\mathcal{M}(k, D_l^1)}(Z)$ is of the form

$$\begin{aligned}
\underline{\mathbb{E}}_{\mathcal{M}(k, D_l^1)}(Z) &= \min_{L \subseteq D_l^1} \mathbb{E}_P(Z)\\
&= \min_{L \subseteq D_l^1} \frac{1}{q_L + 1} \left(z_k + \sum_{j \in L} z_j \right). \qquad (6)
\end{aligned}$$

Here L is a subset of D_l^1; q_L is the number of elements in L ($q_L = \text{card}\,(L)$).

We have considered how to formalize "one-side interval" preference[6]. However, the additional information about BPA's of the corresponding "intervals" has not been applied to the studied MCDM problem. In order to take this additional information into account, we have to introduce the so-called *global criteria* which establish how to compare two DA's from the Pareto set in accordance with all the available information. It should be noted that the global criteria differ from the criteria (goals) $C_1, ..., C_r$.

Below two global criteria for comparison DA's X and Y are proposed.

5.1.1 The first global criterion

The first global criterion is based on the definition of the desirability (3) and can be written as follows. The preference $X \succ Y$ is valid if $\underline{\mathbb{E}}_{\mathcal{P}} \mathbb{E}_P(X - Y) > 0$. Here $\mathcal{P}$ is a set of probability distributions defined on the partition of $\mathcal{M}$ produced by the given information in the form of preferences $D_l \succ \mathbb{C}$ with BPA's $m(D_l)$, $l = 1, ..., t$. For computing the lower expectation, we can use the approach introduced by Strat [10], which directly relies on belief functions based on some basic probability assignment $m(\cdot)$. According to this approach, the lower expectation of $\underline{\mathbb{E}}h$ of a function h is

[4] The reason why the parameters $w_k = 1$ and $w_j = 1$ are taken for formalizing the simple comparison judgments is clearly seen from the proof of Proposition 1.

[5] One can see from the first $r - t$ constraints that they correspond to the comparison of probabilities π_k and π_j, i.e., they formalize the judgment "the k-criterion is as probable as j-th criterion". This implies that the parameters $w_k = 1$ and $w_j = 1$ form the simple comparison.

[6] We have still studied judgments with a fixed k and "interval" $\mathbb{C}\backslash D$ without analyzing the interval D.

determined as follows:

$$\underline{\mathbb{E}}h = \sum_{l=1}^{t} m(D_l) \min_{x \in D_l} h(x).$$

Let $\mathcal{M}(k, D_l^1)$ be a subset of probability distributions produced by conditions $W_{kj} \succ 0_r$, $j \in D_l^1$. Then $m(D_l)$ corresponds to the union of subsets

$$\mathcal{M}(D_l) = \cup_{k \in D_l^0} \mathcal{M}(k, D_l^1).$$

Then the function $h(x)$ in the considered case is the expectation $\sum_{i=1}^{r} \pi_i z_i$. Hence, we get

$$\underline{\mathbb{E}}_{\mathcal{P}} \mathbb{E}_P(Z) = \sum_{l=1}^{t} m(D_l) \left(\min_{k \in D_l^0} \inf_{P \in \mathcal{M}(D_l)} \sum_{i=1}^{r} \pi_i z_i \right).$$

However, there holds

$$\inf_{P \in \mathcal{M}(D_l)} \sum_{i=1}^{r} \pi_i z_i = \underline{\mathbb{E}}_{\mathcal{M}(k, D_l^1)}(X - Y).$$

Hence, Proposition 2 can be stated from the above reasoning.

Proposition 2 *Suppose that there is a set of t judgments of the form $D_l \succ \mathbb{C}$ with the corresponding BPA's $m(D_l)$, $l = 1, ..., t$. The preference $X \succ Y$ is valid in accordance with the first global criterion if the condition*

$$\underline{\mathbb{E}}_{\mathcal{P}} \mathbb{E}_P(X - Y) = \sum_{l=1}^{t} m(D_l) \min_{k \in D_l^0} \underline{\mathbb{E}}_{\mathcal{M}(k, D_l^1)}(X - Y) \geq 0$$

is valid. Here $\underline{\mathbb{E}}_{\mathcal{M}(k, D_l^1)}(X - Y)$ is defined from (6).

5.1.2 The second global criterion

The second criterion is based on the definition of belief and plausibility functions. According to this criterion, we can say about the preference $X \succ Y$ with some "threshold" or confident probability which lies between the belief and plausibility functions. Note that the set $\mathcal{M}$ of all probability distributions can be divided into two subsets $\mathcal{M}_1$ and $\mathcal{M}_2$. The subset $\mathcal{M}_1$ satisfies the condition $X - Y \in \mathcal{D}$. The subset $\mathcal{M}_2$ satisfies the condition $X - Y \notin \mathcal{D}$. Then all subsets $\mathcal{M}(D_l)$ belonging to $\mathcal{M}_1$ form the belief function $\mathrm{Bel}\,(X - Y \in \mathcal{D})$. Note that the subset $\mathcal{M}(D_l)$ intersects $\mathcal{M}_1$ if at least for one of the values k from D_l^0 the subset $\mathcal{M}(k, D_l^1)$ belongs to $\mathcal{M}_1$. The proposition follows from the above.

Proposition 3 *Suppose that there is a set of t judgments of the form $D_l \succ \mathbb{C}$ with the corresponding BPA's $m(D_l)$, $l = 1, ..., t$. The preference $X \succ Y$ is valid in accordance with the second global criterion with a probability belonging to the interval with the following bounds*

$$\mathrm{Bel}\,(X - Y \in \mathcal{D}) = \sum_{l \in R} m(D_l),$$

$$\mathrm{Pl}\,(X - Y \in \mathcal{D}) = \sum_{l \in G} m(D_l),$$

where R is a set of indices such that for every $l \in R$, there holds

$$\min_{k \in D_l^0} \underline{\mathbb{E}}_{\mathcal{M}(k, D_l^1)}(X - Y) > 0,$$

G is a set of indices such that for every $l \in G$, there holds

$$\max_{k \in D_l^0} \underline{\mathbb{E}}_{\mathcal{M}(k, D_l^1)}(X - Y) > 0.$$

Here $\underline{\mathbb{E}}_{\mathcal{M}(k, D_l^1)}(X - Y)$ is defined from (6).

The belief function is the lower (pessimistic or conservative) bound for the probability of the preference $X \succ Y$.

Note that Propositions 2 and 3 are rather general and their main results do not depend on the way of obtaining the lower expectation $\underline{\mathbb{E}}_{\mathcal{M}(k, D_l^1)}(X - Y)$. This implies that the propositions can be generalized by studying a more practical case when we have parameters of the criteria importance w_k and w_j (see Section 3).

5.2 General case

In this section, we generalize the simple comparison judgments by introducing parameters for every pair of criteria, i.e. for every k, DM's supply different parameters $w_j^{(k)}$ for all $j \in B^1$. This is a possible formalization of judgments: "The k-th criterion from B is more important than the j-th criterion from $\mathbb{C} \backslash B$ with the pair of positive parameters w_k and w_j". A special case of the above judgment is the preferences provided by DM's with some degree v under condition that the criteria have identical scales. In this case, we have $v = w_j / w_k$ or $v = \theta_{kj} / (1 - \theta_{kj})$. However, we consider the general case.

Assume for example that $\mathbb{C} = \{C_1, C_2, C_3\}$, $B = \{C_1, C_2\}$, and $\mathbb{C} \backslash B = \{C_3\}$. Then the corresponding judgment of a DM might also have the form: "I'm willing to pay w_3 for C_3 in order to get w_1 for C_1. I'm also willing to pay w_3 for C_3 in order to get w_2 for C_2. However, I do not know what is better. "

Suppose that we have a set of judgments such that every judgment produces the gamble

$$W_{kj} = (0, ..., 0, w_k, 0, ..., -w_j, 0, ..., 0), \qquad (7)$$

such that $W_{kj} \succ 0_r$, $k \in B^0$, $j \in B^1$.

Now we can find the set $\mathcal{M}$ restricted by the desirability of gambles W_{kj} or equivalently its extreme points.

Proposition 4 *Given the additional information in the form (7), the preference $X \succ Y$ is valid if the condition*

$$z_k + \sum_{j \in L} \frac{w_k}{w_j} z_j \geq 0$$

is valid for all $L \subseteq B^1$ and $z_i \geq 0$ for all $i \in B^0$.

Proof. Denote $v_{kj} = w_k/w_j$. It follows from (2) and from (7) that the set $\mathcal{M}$ is produced by the constraints

$$v_{kj}\pi_k - \pi_j \geq 0, \ j \in B^1,$$

$$\pi_i \geq 0, \ i \in N,$$

$$\pi_1 + \pi_2 + ... + \pi_r = 1.$$

Case 1. $v_{kj}\pi_k = \pi_j$, $\pi_i = 0$, $\forall i \in N \backslash \{k, j\}$. Then for every $j \in B^1$, we get the extreme points

$$\pi_k = \frac{1}{1 + v_{kj}}, \ \pi_j = \frac{v_{kj}}{1 + v_{kj}},$$
$$\pi_i = 0, \forall i = N \backslash \{k, j\}.$$

Case 2. $v_{kj_1}\pi_k = \pi_{j_1}$, $v_{kj_2}\pi_k = \pi_{j_2}$, $\pi_i = 0$, $\forall i \in N \backslash \{k, j_1, j_1\}$. Then for every pair $j_1, j_2 \in B^1$, we get the extreme points

$$\pi_k = \frac{1}{1 + v_{kj_1} + v_{kj_2}}, \ \pi_{j_1} = \frac{v_{kj_1}}{1 + v_{kj_1} + v_{kj_2}},$$
$$\pi_{j_2} = \frac{v_{kj_2}}{1 + v_{kj_1} + v_{kj_2}}, \ \pi_i = 0, \forall i = N \backslash \{k, j_1, j_2\}.$$

By continuing the analysis of the cases, we write the following last case.

Case $r - t + 1$. $v_{kj_i}\pi_k = \pi_{j_i}$, $i = 1, ..., r - t$, $\pi_l = 0$, $\forall l \in B^0 \backslash \{k\}$. Then we get the extreme points

$$\pi_k = \frac{1}{1 + \sum_{i=1}^{r-t} v_{kj_i}},$$
$$\pi_{j_i} = \frac{v_{kj_i}}{1 + \sum_{i=1}^{r-t} v_{kj_i}}, \ i = 1, ..., r - t,$$
$$\pi_l = 0, \ \forall l \in B^0 \backslash \{k\}.$$

The proof directly follows from the condition of desirability of the gamble $Z = X - Y$, which is of the form: $\mathbb{E}_P(Z) \geq 0$, $\forall P \in \text{extr}(\mathcal{M})$. Hence, for every subset $L \subseteq B^1$, we can write the expectations as follows:

$$\mathbb{E}_P(Z) = \frac{z_k}{1 + \sum_{i \in L} v_{ki}} + \sum_{j \in L} \frac{v_{kj} z_j}{1 + \sum_{i \in L} v_{ki}}.$$

Since $v_{kj_i} \geq 0$ for all k, j, i, then $\mathbb{E}_P(Z) \geq 0$ for every extreme point if

$$z_k + \sum_{j \in L} v_{kj} z_j \geq 0, \ L \subseteq B^1,$$

as was to be proved. ■

We get the rather simple expressions for reducing the Pareto set.

Generally speaking, the values w_k in Proposition 4 may be different for different values of $j \in L$ and the index k_j should be used. However, we assume for simplicity that the parameters w_k are identical for every W_{kj}. Moreover, it can be seen from the proof of Proposition 4 that the condition of the preference $X \succ Y$ depends only on the ratio w_k/w_j and we can always change w_k and w_j without changing the above ratio.

Let us considers a special case when each of the subsets B^1 and B^0 consists of one element.

Corollary 1 *Suppose that $B^1 = \{k\}$ and $B^0 = \{j\}$. Then the preference $X \succ Y$ is valid if the conditions*

$$w_j z_k + w_k z_j \geq 0, \ z_i \geq 0, \ \forall i \neq j,$$

are valid.

One can see that the conditions in Corollary 1 coincide with the conditions in Theorems 1 and 2.

Several conditions in Proposition 4 can be replaced by one equivalent condition

$$z_k + w_k \min_{L \subseteq B^1} \sum_{j \in L} z_j w_j^{-1} \geq 0. \qquad (8)$$

By using (8) and Propositions 2, 3 we can write the following corollary.

Corollary 2 *If there are judgments of one DM ($l = 1$, $D_1 = D$) with the BPA $m(D_l) = 1$, then the preference $X \succ Y$ is valid in accordance with the first global criterion if the conditions*

$$\min_{k \in D^0} \{z_k + T w_k\} \geq 0$$

are valid. Here

$$T = \min_{L \subseteq D^1} \sum_{j \in L} z_j w_j^{-1}.$$

Moreover, the belief function $\mathrm{Bel}\,(X - Y \in \mathcal{D})$ *is* 1 *in accordance with the second global criterion if the above conditions are valid.*

It also follows from the proof of Proposition 4 and from (8) that the lower expectation of the gamble $Z = X - Y$ under conditions $W_{kj} \succ 0_r$, $j \in D_l^1$, denoted $\underline{\mathbb{E}}_{\mathcal{M}(k,D_l^1)}(Z)$ is of the form

$$\underline{\mathbb{E}}_{\mathcal{M}(k,D_l^1)}(Z) = \min_{L \subseteq D_l^1} \mathbb{E}_P(Z) = \min_{L \subseteq D_l^1} \frac{z_k + w_k \sum_{j \in L} z_j w_j^{-1}}{1 + w_k \sum_{i \in L} w_i^{-1}}. \quad (9)$$

Then Propositions 2 and 3 can be used in the considered case of the elicited information if we replace (6) by (9).

Example 2 *Let us return to Example 1. The judgment of the first DM is the following:*

"I'm willing to pay $w_2 = 15000$ *for* C_2 *in order to get* $w_1 = 15$ *for* C_1 *and I'm willing to pay* $w_4 = 7$ *for* C_4 *in order to get* $w_1 = 15$ *for* C_1*. I'm also willing to pay* $w_2 = 24000$ *for* C_2 *in order to get* $w_3 = 1$ *for* C_3 *and I'm willing to pay* $w_4 = 10$ *for* C_4 *in order to get* $w_3 = 1$ *for* C_3*. However, I do not know what is better. "*

The above judgment can be formalized as $D_1 = \{C_1, C_3\} \succ \{C_2, C_4\}$*. The judgment of the second DM is the following:*

"I'm willing to pay $w_3 = 6$ *for* C_3 *in order to get* $w_1 = 30$ *for* C_1*. I'm also willing to pay* $w_3 = 8$ *for* C_3 *in order to get* $w_2 = 10000$ *for* C_4*. I'm also willing to pay* $w_3 = 20$ *for* C_3 *in order to get* $w_4 = 1$ *for* C_4*. However, I do not know what is better. "*

This judgment can be formalized as $D_2 = \{C_1, C_2, C_4\} \succ \{C_3\}$.

The BPA of the first DM is $m(D_1) = 0.6$*. The BPA of the second DM is* $m(D_2) = 0.4$.

Let us find $\underline{\mathbb{E}}_{\mathcal{M}(k,D_l^1)}(X - Y)$*. If* $D_1^1 = \{2, 4\}$ *and* $k = 1, 3$*, then it follows from (9) that*

$$\underline{\mathbb{E}}_{\mathcal{M}(1,D_1^1)} = \min\left(z_1, \frac{z_1 + 15z_2/15000}{1 + 15/15000}, \frac{z_1 + 15z_4/7}{1 + 15/7}, \frac{z_1 + 15z_2/15000 + 15z_4/7}{1 + 15/15000 + 15/7}\right),$$

Table 1: Comparison of DA's by using two criteria

$X \succ Y$	$\underline{\mathbb{E}}_{\mathcal{P}}\mathbb{E}_P(X-Y)$	Bel	Pl
$X_1 \succ X_2$	-1199	0.6	1
$X_1 \succ X_3$	0.14	0.4	1
$X_1 \succ X_4$	0.09	0.6	1
$X_2 \succ X_1$	-10	0	0.4
$X_3 \succ X_1$	-1614	0	0.6
$X_4 \succ X_1$	-1614	0	0.6
$X_2 \succ X_3$	-2.13	0	1
$X_2 \succ X_4$	-5	0	0.4
$X_3 \succ X_2$	-2810	0	0.6
$X_3 \succ X_4$	-810.2	0	0.6

$$\underline{\mathbb{E}}_{\mathcal{M}(3,D_1^1)} = \min\left(z_3, \frac{z_3 + 1z_2/24000}{1 + 1/24000}, \frac{z_3 + 1z_4/10}{1 + 1/10}, \frac{z_3 + 1z_2/24000 + 1z_4/10}{1 + 1/24000 + 1/10}\right).$$

If $D_2^1 = \{3\}$ *and* $k = 1, 2, 4$*, then it follows from (9) that*

$$\underline{\mathbb{E}}_{\mathcal{M}(1,D_2^1)} = \min\left(z_1, \frac{z_1 + 30z_3/6}{1 + 30/6}\right),$$

$$\underline{\mathbb{E}}_{\mathcal{M}(2,D_2^1)} = \min\left(z_2, \frac{z_2 + 10000z_3/8}{1 + 10000/8}\right),$$

$$\underline{\mathbb{E}}_{\mathcal{M}(4,D_2^1)} = \min\left(z_4, \frac{z_4 + 1z_3/20}{1 + 1/20}\right).$$

The computation results with using Propositions 2 and 3 are shown in Table 1. It can be seen from Table 1 that the reduced Pareto set ***in accordance with the first criterion*** $\underline{\mathbb{E}}_{\mathcal{P}}\mathbb{E}_P(X-Y) > 0$ *of decision making consists of two DA's* X_1 *and* X_2 *because* $\underline{\mathbb{E}}_{\mathcal{P}}\mathbb{E}_P(X_1 - X_3) = 0.14 > 0$ *and* $\underline{\mathbb{E}}_{\mathcal{P}}\mathbb{E}_P(X_1 - X_4) = 0.09 > 0$*. However, by using* ***the second criterion*** *of decision making* ***with the "threshold" probability*** 0.6 *for the belief function, we can construct the reduced Pareto set consisting of two DA's* X_1 *and* X_3.

6 Conclusion

A method for solving a MCDM problem with the elicited information about criteria of a special form has been proposed in the paper. The main feature of the method is that it is based on reducing a set of Pareto optimal solutions and does not use aggregation of criteria for solving the problem. The additional information applied in the proposed method is rather natural because DM's or experts are usually able to provide parameters w_i and w_j whose simple behavior

interpretation is considered in Section 3 and in Example 2.

It has been shown in the paper that Noghin's theory of relative importance of criteria can be easy represented in terms of sets of desirable gambles and many statements of the theory can be proved by means of desirable gambles and the imprecise probability theory.

Two global criteria of decision making are introduced. The first criterion based on the lower expectation uses the second-order models as a main tool for determining whether a preference $X \succ Y$ is valid or not. The second criterion is based on determining the belief and plausibility function in the framework of Dempster-Shafer theory. It uses the so-called "threshold" probability for the final decision making.

One can see from the proposed expressions (6), (9) and Propositions 2 and 3 that all the mathematical expressions are rather simple from the computation point of view and they do not require special procedures for computing the lower expectation $\underline{\mathbb{E}}_{\mathcal{P}}\mathbb{E}_P(X-Y)$ and the belief and plausibility functions.

Some specialists in Dempster-Shafer theory might object that the condition of independence of DM's in combining their judgments is not taken into account. Of course, we could assume that the DM's are independent and use, for instance, Dempster rule of combination. However, the main aim of the paper is to propose an approach for reducing the Pareto set on the basis of the special information, in particular, on judgments producing sets of gambles (7). Various modifications and features of the approach can be studied in further research.

It should be noted that the simple case has been studied in the paper when only judgments of the special type are provided by DM's. However, the proposed approach can be extended on a more complicated case. Therefore, a direction for further work is to investigate the general cases.

Acknowledgement

I thank referees for useful and detailed suggestions that improved the paper.

References

[1] M. Beynon. DS/AHP method: A mathematical analysis, including an understanding of uncertainty. *European Journal of Operational Research*, 140:148–164, 2002.

[2] M. Beynon, B. Curry, and P. Morgan. The Dempster-Shafer theory of evidence: An alternative approach to multicriteria decision modelling. *Omega*, 28:37–50, 2000.

[3] R.L. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. John Wiley and Sons, NY, 1976.

[4] S.H. Kim and B.S. Ahn. Interactive group decision making procedure under incomplete information. *European Journal of Operational Research*, 116:498–507, 1999.

[5] V.D. Noghin. Relative importance of criteria: a quantitative approach. *Journal of Multi-Criteria Decision Analysis*, 6:355–363, 1997.

[6] V.D. Noghin. *Decision Making in Multicriteria Environment: A Quantitative Approach*. Fizmatlit, Moscow, 2002. http://www.apmath.spbu.ru/en/staff/nogin.

[7] K.S. Park and S.H. Kim. Tools for interactive multi-attribute decision making with incompletely identified information. *European Journal of Operational Research*, 98:111–123, 1997.

[8] T.L. Saaty. *Multicriteria Decision Making: The Analytic Hierarchy Process*. McGraw Hill, New York, 1980.

[9] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[10] T.M. Strat. Decision analysis using belief functions. *International Journal of Approximate Reasoning*, 4(5):391–418, 1990.

[11] L.V. Utkin. Imprecise second-order hierarchical uncertainty model. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 11(3):301–317, 2003.

[12] L.V. Utkin and N.V. Simanova. Multi-criteria decision making by incomplete preferences. *Journal of Uncertain Systems*, 2(4):255–266, 2008.

[13] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[14] P. Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24:125–148, 2000.

Combining imprecise Bayesian and maximum likelihood estimation for reliability growth models

Lev V. Utkin
Department of Computer Science
Forest Technical Academy
St.Petersburg
lev.utkin@mail.ru

Svetlana I. Zatenko
Department of Mathematics
Forest Technical Academy
St.Petersburg
s_lana2004@mail.ru

Frank P.A. Coolen
Department of Mathematical Sciences
Durham University
Durham
frank.coolen@durham.ac.uk

Abstract

A new framework is explored for combining imprecise Bayesian methods with likelihood inference, and it is presented in the context of reliability growth models. The main idea of the framework is to divide a set of the model parameters of interest into two subsets related to fundamentally different aspects of the overall model, and to combine Walley's idea of imprecise Bayesian models related to one of the subsets of the model parameters with maximum likelihood estimation for the other subset. In accordance with the first subset and statistical data, the imprecise Bayesian model is constructed, which provides lower and upper predictive probability distributions depending on the second subset of parameters. These further parameters are then estimated by a maximum likelihood method, based on a novel proposition for maximum likelihood estimation over sets of distributions following from imprecise Bayesian models for the other subset of parameters. Use of this hybrid method is illustrated for reliability growth models and regression models, and some essential topics that need to be addressed in order to fully justify and further develop this framework are discussed.

Keywords. Bayesian inference, imprecise probabilities, linear regression, lower and upper probability distributions, maximum likelihood estimation, reliability growth models

1 Introduction

One of the main goals of system analysis is to predict its future behaviour on the basis of past experience, for which one typically constructs a statistical model to quantify uncertainties and to enable learning from data. There is a variety of statistical theories and methods for such inference, and researchers often strongly advocate one specific general theory, e.g. the Bayesian approach, whilst rejecting other approaches that also have their merits. In this paper we explore combined use of imprecise Bayesian methods, where sets of prior distributions are used, with maximum likelihood estimation, both on different subsets of all parameters appearing in a statistical model. At first look, these methods may appear to have little in common and one may favour either a complete (imprecise) Bayesian approach or maximum likelihood estimation of all parameters. However, if one considers a Bayesian approach as using a weighted likelihood function, with weights reflecting prior knowledge, the two are less contradictory and exploration of the opportunity to combine both into a hybrid method can be of interest. In this paper we set the first steps in this direction, which include a crucial proposition on maximum likelihood estimation for a subset of parameters following imprecise Bayesian inference on a different subset of parameters. Detailed fundamental analysis and further exploration of this hybrid approach will be important for its full justification, in particular with regard to possible interpretations of the resulting inferences. We present our ideas in the context of reliability growth models.

An important feature of many systems is growth or change of some of their characteristics over time, which has to be taken into account when constructing a statistical model for the system. For example, a common approach for measuring software reliability [18] is by using a statistical model whose parameters are generally estimated from available data on software failures, and the model may be obtained by observing the overall trend of reliability growth during the debugging process. In other words, a software reliability growth model describes how observation of failures, and correcting the underlying faults – such as occurs in software development when the software is being tested and debugged – affect the reliability of software. The word "growth" is rather conventional to describe reliability models with important characteristics changing over time, it does not restrict use of such models to systems whose reliability actually improves. In other words, a growth model can be

regarded to be a mathematical expression which fits experimental data from systems with some important changes over time.

Suppose that $X_1, ..., X_n$ is a series of random variables, for instance, numbers of successful software runs between the $(i-1)$-th and i-th software failures. We suppose that variable X_i is governed by a probability distribution function $p_i(x \mid \mathbf{b}, \mathbf{d})$ depending on two vectors of parameters $\mathbf{b}$ and $\mathbf{d}$. The vector $\mathbf{b}$ contains parameters of the probability distribution under consideration. The vector $\mathbf{d}$ of parameters characterizes the growth, i.e., the growth is modelled by a function $f(i, \mathbf{d})$ which characterizes the change of the system behavior ('growth'). For example in software reliability analysis, the function f mainly shows how parameters $\mathbf{b}$ of the probability distribution p_i change with the number of corrected errors or faults i. Generally, the vector $\mathbf{b}$ depends on $\mathbf{d}$ and the number i of the random variable X_i under consideration.

It should be noted that the growth function in some models is explicitly stated. For instance, Littlewood and Verrall [8] suggest software reliability models with linear and quadratic forms for the function f with two parameters $\mathbf{d} = (d_0, d_1)$: $f(i, \mathbf{d}) = d_0 + d_1 i$ or $f(i, \mathbf{d}) = d_0 + d_1 i^2$. In these models, the growth function is included as parameter of a gamma distribution, which changes with the number of corrected errors in the software.

A similar feature occurs in regression models [9], which in their simplest form provide a relation between predictor variables X_i, $i = 1, ..., n$, and a response variable Y. A typical regression model can be written as

$$Y = f(\mathbf{X}, \mathbf{d}) + \epsilon.$$

Here $\mathbf{X} = (1, X_1, ..., X_n)$; $\mathbf{d}$ is the vector of parameters; ϵ are uncorrelated random errors or noise, usually assumed to have expected value 0 and unknown variance σ^2. In such a model, $\mathbf{d}$ can be a set of growth parameters, for instance, coefficients in a linear regression model, while setting $\mathbf{b} = (\sigma^2)$ fits with the generic notation suggested above.

Clearly, the growth function f may model different characteristics. In software reliability models, it typically enables possible changes of the parameters $\mathbf{b}$ of the probability distribution of random variables X_i to reflect actual changes to software systems, mostly due to error corrections. In regression models, the parameter $\mathbf{b} = (\sigma^2)$ is assumed to be constant, but the growth function characterizes the system behaviour. Nevertheless, both types of models are equivalent from mathematical point of view[1]. In both the cases, we assume a form of f and wish to learn about the parameters $\mathbf{d}$ of f from data.

There are several approaches for inference about growth models on the basis of statistical data. Nowadays, the most popular inferential methods tend to use the likelihood function as main mechanism to link model parameters and statistical data. For models such as reliability growth models, estimation is required both for parameters of the basic probability model and parameters explicitly modelling the growth behaviour. This may involve a substantial number of parameters, with possibly relatively few data available. In this paper, we explore a possible way for dealing with this, by considering imprecise Bayesian inference for one subset of parameters, and a maximum likelihood approach to estimate the other subset of parameters. Such imprecise Bayesian inference has been presented, without a link to maximum likelihood for further parameters, by Walter, Augustin and Peters [17] with application to linear regression models. Typically, a precise parametric model is assumed, with imprecision following through the use of sets of conjugated priors [1, 11, 16]. It is theoretically feasible to use sets of priors for all parameters combined, but this may well lead to very wide posterior intervals for inferences of interest, and if one can estimate some of the parameters by means of maximum likelihood methods, it could be also be attractive with regard to not needing to attempt to assign informative (sets of) prior distributions, in particular if they are on a feature about which no clear expert judgement is available or which one strongly wishes to infer from the data.

The approach we propose in this paper is as follows. By using imprecise Bayesian inference, we can exclude all the parameters of the vector $\mathbf{b}$ from the model, and derive a set of predictive *cumulative distribution functions* (CDFs) such that their lower and upper bounds are conditional on all the parameters of the vector $\mathbf{d}$. This is followed by estimation of the parameters of the vector $\mathbf{d}$, for which we use a modified maximum likelihood estimation method described and justified in Section 3. This approach allows us to reduce the number of parameters in the model and to maximize the likelihood function only over parameters of the vector $\mathbf{d}$ without considering the parameters of vector $\mathbf{b}$. Even further, it can be applied if one explicitly wishes to take expert judgement into account on the part of the model corresponding to parameters $\mathbf{b}$, and this expert judgement is best reflected by imprecise probabilities, while no such prior information is available for the model aspects related to parameters $\mathbf{d}$, for which, however, one can use process data.

[1] The software reliability growth models in the literature are often called regression models due to some common features of the models.

To simplify the presentation of the proposed approach, we study discrete random variables X_i corresponding to the number of successful software runs between the $(i-1)$-th and i-th software failures (for the first software reliability growth model, Section 6) or to the random number of failures between t_{i-1} and t_i (for the second software reliability growth model, Section 7), $i = 1, ..., n$. A general scheme for such combined inference for regression models will be briefly considered in Section 8, to demonstrate that the proposed framework can be applied to various problems.

2 The likelihood principle for constructing standard models

Let $\mathbf{K} = (k_1, ..., k_n)$ be a realization of $X_1, ..., X_n$, with k_i non-negative integers. If probability distributions $p_i(k_i \mid \mathbf{b}, \mathbf{d})$ of the random variables X_i, $i = 1, ..., n$, are known or assumed, then the standard way for obtaining the parameters $\mathbf{b}$ and $\mathbf{d}$ of a growth model is to maximize the likelihood function

$$L(\mathbf{K} \mid \mathbf{b}, \mathbf{d}) = \prod_{i=1}^{n} p_i(k_i \mid \mathbf{b}, \mathbf{d})$$

over a set of parameters $\mathbf{b}$ and $\mathbf{d}$. Values of the parameters $\mathbf{b}$ and $\mathbf{d}$ should be chosen in such a way that makes $L(\mathbf{K} \mid \mathbf{b}, \mathbf{d})$ achieve its maximum.

Many well-known software reliability growth models presented in the literature have been implemented with such standard maximum likelihood estimation. Such models differ only by assumptions about the probability distributions p_i and the growth function f. For example, p_i in the Jelinski-Moranda model [6] is exponential, the Rayleigh distribution is used in the Schick-Wolverton model [13], and the Littlewood-Verrall model [8] uses a Beta distribution.

3 Maximization of the likelihood function over a set of distributions

Suppose that the random variable X_i is governed by an unknown CDF $F_i(k)$ which is only known to belong to the set $\mathcal{M}_i(\mathbf{d})$ defined by the lower and upper CDFs

$$\underline{F}_i(k \mid \mathbf{d}) = \inf_{\mathcal{M}_i(\mathbf{d})} F(k), \tag{1}$$

$$\overline{F}_i(k \mid \mathbf{d}) = \sup_{\mathcal{M}_i(\mathbf{d})} F(k). \tag{2}$$

It should be noted that the set $\mathcal{M}_i(\mathbf{d})$ is the set of *all* CDFs bounded by $\underline{F}_i(k \mid \mathbf{d})$ and $\overline{F}_i(k \mid \mathbf{d})$, so it is *not* the set of parametric distributions having the same parametric form as the bounding distributions. This is an important feature of the proposed approach for combined imprecise Bayesian and likelihood inference in this paper. Moreover, the bounds $\underline{F}_i(k \mid \mathbf{d})$ and $\overline{F}_i(k \mid \mathbf{d})$ are assumed not to depend on the parameters $\mathbf{b}$, which is achieved by taking the predictive CDFs resulting from the imprecise Bayesian approach applied with regard to the parameters $\mathbf{b}$.

The likelihood function can be written in the following form:

$$L(\mathbf{K} \mid \mathbf{d}) = \Pr\{X_1 = k_1, ..., X_n = k_n\}.$$

Proposition 1 explains how the above likelihood function is maximized over all distributions belonging to $\mathcal{M}_1(\mathbf{d}), ..., \mathcal{M}_n(\mathbf{d})$.

Proposition 1 *Suppose that discrete random variables $X_1, ..., X_n$ are governed by a probability distribution $F(k)$ from sets $\mathcal{M}_i$ defined by bounds (1)-(2), respectively. If $X_1, ..., X_n$ are independent, then there holds*

$$\max_{\mathcal{M}_1, ..., \mathcal{M}_n} \Pr\{X_1 = k_1, ..., X_n = k_n\} = \prod_{i=1}^{n} \left\{\overline{F}_i(k_i) - \underline{F}_i(k_i - 1)\right\}. \tag{3}$$

Proof. Denote $N = \{1, 2, ..., n\}$, $\mathbf{M} = (m_1, ..., m_n)$. Let $I_{\{1,...,k_i\}}(m)$ be the indicator function taking the value 1 if $m \le k_i$. The indicator functions are used in the proof to represent all probabilities as expectations of indicator functions, and to write the natural extension in its standard form. The upper bound for the joint probability $\Pr\{X_1 = k_1, ..., X_n = k_n\}$ can be found by solving the following optimization problem:

$$\max \sum_{m_1=1}^{\infty} \cdots \sum_{m_n=1}^{\infty} I_{\{k_1,...,k_n\}}(\mathbf{M}) \prod_{i=1}^{n} p_i(m_i),$$

subject to

$$\sum_{m=1}^{\infty} p_i(m) = 1,$$

$$\underline{F}_i(j) \le \sum_{m=1}^{\infty} I_{\{1,...,j\}}(m) p_i(m) \le \overline{F}_i(j),$$

$$i = 1, ..., n,\ j = 1, 2...$$

The objective function can be rewritten as follows:

$$\prod_{i=1}^{n} \sum_{m_i=1}^{\infty} \left(I_{\{1,..,k_i\}}(m_i) - I_{\{1,..,k_i-1\}}(m_i) p_i(m_i)\right).$$

Introduce new variables

$$F_i(j) = \sum_{m_i=1}^{\infty} I_{\{1,...,j\}}(m_i) p_i(m_i).$$

Then we can rewrite the optimization problem as

$$\max \prod_{i=1}^{n} \{F_i(j) - F_i(j-1)\},$$

subject to

$$\underline{F}_i(j) \le F_i(j) \le \overline{F}_i(j),$$
$$\underline{F}_i(j-1) \le F_i(j-1) \le \overline{F}_i(j-1),\ i = 1, ..., n.$$

By using the known rules of interval analysis, we obtain (3), which completes the proof. ■

Proposition 1 generalizes the standard likelihood estimation for precise probability models.

4 Imprecise Bayesian models as a way for obtaining the set $\mathcal{M}$

We now consider how to derive the set $\mathcal{M}(\mathbf{d})$. A straightforward way is to use ideas similar to Walley's imprecise Bayesian approach [16].

4.1 Standard Bayesian analysis

One of the efficient approaches to estimation of the model parameters is Bayesian analysis [2, 4, 12]. It treats parameters of concern as random variables which are assigned a prior probability distribution before observations become available. If we assume that the random variable has a probability distribution with vector of unknown parameters $\mathbf{b}$, then these parameters would be regarded as random variables with a prior probability density $\pi(\mathbf{b} \mid \mathbf{c})$, characterized by (hyper-)parameters $\mathbf{c}$. In this case, the Bayesian approach can be applied for computing the CDF for the random variable of interest, with the parameter $\mathbf{b}$ integrated out:

$$F(k \mid \mathbf{c}) = \int_{\Omega} F(k \mid \mathbf{b}) \cdot \pi(\mathbf{b} \mid \mathbf{c}) \mathrm{d}\mathbf{b}.$$

Here Ω is the set of values of $\mathbf{b}$.

Central to the Bayesian approach is the derivation of the posterior distribution of the unknown parameters, given both the data and the assumed prior density for these parameters, and achieved by application of Bayes' theorem. Suppose that the prior distribution $\pi(\mathbf{b} \mid \mathbf{c})$ represents our uncertainty with regard to $\mathbf{b}$ prior to collecting information in the form of a set $\mathbf{K} = (k_1, ..., k_n)$ of observed values of independent random variables $X_1, ..., X_n$. Let $p(k)$ be the probability mass function for the observed data k given $\mathbf{b}$. Then the posterior distribution $\pi(\mathbf{b} \mid \mathbf{K}, \mathbf{c})$ as the conditional distribution of $\mathbf{b}$ given the observed data $\mathbf{K}$ and prior parameters $\mathbf{c}$ is computed as

$$\pi(\mathbf{b} \mid \mathbf{K}, \mathbf{c}) \propto p(k_1) \cdot \cdots p(k_n) \cdot \pi(\mathbf{b} \mid \mathbf{c}).$$

Here $\pi(\mathbf{b} \mid \mathbf{K}, \mathbf{c})$ represents updated beliefs about $\mathbf{b}$, with information $\mathbf{K}$ taken into account.

The prior distribution is often chosen to facilitate calculation of the prior, especially through the use of *conjugate priors* [2]. If the posterior distribution $\pi(\mathbf{b} \mid \mathbf{K}, \mathbf{c})$ and the prior distribution $\pi(\mathbf{b} \mid \mathbf{c})$ both belong to the same family of distributions, the π and p are called conjugate distributions and π is called a conjugate prior for p.

4.2 Imprecise prior models

A critical feature of any Bayesian analysis is the choice of a prior distribution, which is often done by considering the choice of (hyper-)parameters of an assumed parametric prior probability distribution. This is both important if one aims at modelling prior information and if one aims to choose a prior distribution in order to reflect the absence of prior information about the parameters. In this paper we focus on the latter case, where a so-called non-informative prior has to be constructed. Many criteria for non-informativeness, and methods to determine non-informative priors, have been proposed in the literature [2, 12], with many methods applying the Bayes-Laplace postulate or the principle of insufficient reason. However, this choice meets some difficulties or problems. In particular, Walley [16] provides examples illustrating possible problems and shortcomings of the principle of insufficient reason. Syversveen [14] presents a detailed review of methods for constructing non-informative priors.

An alternative way for using the Bayesian approach if one wishes not to take prior knowledge into account is through the use of a class $\mathcal{P}$ of (non-informative) prior distributions π [15], which can overcome most problems that can occur when single non-informative priors are used. Such a class of priors can be considered through the lower $\underline{P}$ and upper $\overline{P}$ probabilities of an event A as

$$\underline{P}(A) = \sup\{P_\pi(A) : \pi \in \mathcal{P}\},$$
$$\overline{P}(A) = \inf\{P_\pi(A) : \pi \in \mathcal{P}\}.$$

As pointed out by Syversveen [14] and Walley [16], the class $\mathcal{P}$ under some conditions is "not a class of reasonable priors, but a reasonable class of priors". This means that each single member of the class is not a reasonable model for prior ignorance, because no single distribution can model ignorance satisfactorily, but the whole class is a reasonable model for prior ignorance. When we have little prior information, the upper probability of a non-trivial event should be close to one and the lower probability should be

close to zero. This means that the prior probability of the event may be arbitrary from 0 to 1.

Quaeghebeur and de Cooman [11] proposed a class of imprecise probability models in the framework of the so-called exponential families of probability distributions [2]. These models significantly extend a set of Bayesian imprecise models and give a possibility to develop a framework for imprecise growth models. In our approach, the set $\mathcal{P}$ is used in the imprecise Bayesian framework to take data into account with regard to parameters $\mathbf{b}$, and thus to generate the set $\mathcal{M}$ of predictive distributions with lower and upper bounds which allow us to apply Proposition 1 for maximum likelihood estimation of the parameters $\mathbf{d}$.

5 A general scheme of the model construction

We now present a general scheme for our proposed method that combines imprecise Bayesian inference and maximum likelihood estimation. We present it using the setting of reliability growth models discussed earlier in this paper, but the general idea is more widely applicable. The first task is to define the sets $\mathcal{M}_1(\mathbf{d}), ..., \mathcal{M}_n(\mathbf{d})$ or their bounds by using an appropriate imprecise Bayesian model. It consists of four steps.

1. We divide the set of parameters into two subsets. The first subset contains the parameters $\mathbf{b}$ of the assumed probability distribution p of the random variables $X_1, ..., X_n$. The second subset consists of the growth parameters $\mathbf{d}$.

2. For the assumed probability distribution p of the random variables, we choose an appropriate type of the conjugate prior $\pi(\mathbf{b} \mid \mathbf{c})$ with parameters $\mathbf{c}$.

3. We construct the corresponding Bayesian imprecise model on the basis of results of Walley [16] or Quaeghebeur and de Cooman [11]. At that point we replace the parameters $\mathbf{c}$ by new parameters including the hyperparameter s (see [11, 16] and examples below). The produced set $\mathcal{P}$ depends on the hyperparameter s.

4. By using n observations $k_1, ..., k_n$, we write the lower $\underline{F}_i(k \mid \mathbf{d}, s)$ and upper $\overline{F}_i(k \mid \mathbf{d}, s)$ predictive CDFs as functions of the parameters $\mathbf{d}$ and the hyperparameter s for every debugging period. These functions form the sets[2] $\mathcal{M}_1(\mathbf{d}), ..., \mathcal{M}_n(\mathbf{d})$.

[2]It should be noted that the set $\mathcal{M}_i(\mathbf{d})$ also depends on the hyperparameter s. However, we omit this parameter for shorter notation.

After completing the four steps of the first task, the sets $\mathcal{M}_1(\mathbf{d}), ..., \mathcal{M}_n(\mathbf{d})$ have been derived and these sets do not depend on the parameters $\mathbf{b}$ or $\mathbf{c}$. They depend only on the growth parameters $\mathbf{d}$, the hyperparameter s for the imprecise prior class, and the number of debugging periods i. The second task is to estimate the parameters $\mathbf{d}$, it consists of two steps.

1. The likelihood function $L(\mathbf{K} \mid \mathbf{d}, s)$ is derived by applying Proposition 1.

2. Values of the parameters $\mathbf{d}$ for a fixed s should be chosen in such a way that makes $L(\mathbf{K} \mid \mathbf{d}, s)$ achieve its maximum.

Note that the parameters $\mathbf{b}$ do not appear in the process, as they have been integrated out with the use of a class of priors to derive predictive distributions, and this process also implicitly replaced the parameters $\mathbf{c}$ by s. Clearly, the step to get $\mathbf{b}$ out of the model, without explicitly estimating their values, is imprecise and leads to predictive imprecise probabilities for the random variables of interest. For example, if we construct a software reliability model, then we are looking for the predictive behavior of the analyzed software after n corrections of errors. In other words, we have to compute the probability measures of time to the $(n+1)$-th failure, in particular, the lower and upper probability distributions of time to the $(n+1)$-th failure. These bounds are totally determined by the parameters $\mathbf{d}$ and s in our approach, with s chosen to specify the class of priors, and $\mathbf{d}$ to be estimated by our proposed maximum likelihood approach in the second stage of our method.

In the following sections, we illustrate our method by considering some special cases which apply known imprecise Bayesian models and consider well-known software reliability growth models.

6 A software run reliability growth model

The detailed description of software run reliability models is given in [3]. A run is a minimum execution unit of software. Any software execution process can be divided into a series of runs. When a run is executed, the software either passes or fails. Usually it is assumed that after observing a software failure, the software is corrected and it is usually assumed that this action actually removes the software error that caused the failure, hence the software improves due to this action and therefore the term reliability growth tends to be used. There are many variations to this basic scenario in the software reliability literature, we do not address these here.

Let X be a run lifetime of software, that is, X is a discrete random variable taking the value k if the software fails during the k-th run after $k-1$ successful runs. The run lifetime distribution (probability mass function) is defined as $p(k) = \Pr\{X = k\}$.

6.1 The imprecise beta-geometric model

If we assume that the random variable X is governed by the geometric distribution with parameter r and the probability mass function

$$p(k \mid r) = (1-r)^{k-1}r, \ k = 1, 2, ...,$$

then the set $\mathcal{M}$ can be constructed by using an imprecise model that is very similar to the beta-binomial model proposed by Walley [16]. The prior Beta distribution of the random variable r, denoted $\text{Beta}(\alpha, \beta)$ with parameters $\alpha > 0$ and $\beta > 0$, has probability density function

$$\pi(r) = \frac{1}{\text{B}(\alpha, \beta)} r^{\alpha-1}(1-r)^{\beta-1}, \ 0 \le r \le 1.$$

Here $\text{B}(\alpha, \beta)$ is the standard beta function.

Using the general notation introduced before in this paper for our new method, we write $\mathbf{b} = (r)$, $\mathbf{c} = (\alpha, \beta)$. If we observe k runs of software between the $(i-1)$-th and i-th software failures, and we assume that the number of such runs is geometrically distributed with parameter r, then the posterior distribution $\pi(r \mid k, \mathbf{c})$ is again a beta distribution, namely

$$\pi(r \mid k, \mathbf{c}) = \text{Beta}(\alpha + 1, \beta + k).$$

Here Bayesian analysis leads to the probability distribution of the number of events with parameters α and β. We can call this a beta-geometric model. In the beta-binomial model, Walley proposed to replace these parameters by introducing s and γ, with $\alpha = s\gamma$ and $\beta = s - s\gamma$, and then the parameter γ is allowed to take on any value in the interval from 0 to 1, hence a set of prior distributions is created which only depends on the choice of $s > 0$, and which trivially leads to a corresponding set of posterior distributions. The hyperparameter s determines the influence of the prior distribution on posterior probabilities [16]. The beta-geometric model proposed here can be given exactly the same imprecise Bayesian treatment, resulting in what we call the imprecise beta-geometric model. The lower and upper bounds can be obtained by minimizing and maximizing the probabilities of events over all values γ in $[0, 1]$.

6.2 The imprecise beta-geometric growth model

Suppose that the probability $r = r_i$ is a random variable having a beta distribution with prior parameters α and $\beta + f(i, \varphi)$. Here $f(i, \varphi)$ is a function characterizing the software reliability growth, in particular, assume for simplicity that $f(i, \varphi) = (i-1) \cdot \varphi$. In this case, we get a model with three parameters, including two prior parameters α and β of the probability distribution and one parameter φ which characterises the reliability growth. The notation introduced above can be used by defining $\mathbf{c} = (\alpha, \beta)$ and $\mathbf{d} = (\varphi)$.

The construction of the model is based on the idea of dividing the set of parameters α, β, φ into two subsets and to consider the imprecise Bayesian model on the set $\mathcal{M}_i(\varphi)$ of CDFs bounded by some lower $\underline{F}_i(k \mid \varphi, \alpha, \beta)$ and upper $\overline{F}_i(k \mid \varphi, \alpha, \beta)$ CDFs which are defined by the set of parameters $\mathbf{c} = (\alpha, \beta)$ for a fixed parameter φ, for $i = 1, ..., n$. In other words, we fix φ and construct the sets of CDFs $F_i(k)$ with bounds depending on $f(i, \varphi)$ by using the imprecise beta-geometric model.

After constructing the set $\mathcal{M}_i(\varphi)$ of CDFs $F_i(k \mid \varphi)$ having the lower $\underline{F}_i(k \mid \varphi, \alpha, \beta)$ and upper $\overline{F}_i(k \mid \varphi, \alpha, \beta)$ CDFs for every $i = 1, ..., n$, and by assuming that the random variables $X_1, ..., X_n$ are independent, the likelihood function can be written and maximized by application of Proposition 1, leading to the value φ_0 that maximises this likelihood, so which we consider an appropriate estimate of φ.

Denote the parameters of the i-th posterior beta distribution after n observations

$$\alpha^* = \alpha + n - 1, \quad \beta_i^* = \beta + D_i(\varphi),$$

where

$$D_i(\varphi) = K_n + f(i, \varphi), \ K_n = \sum_{j=1}^{n-1} (k_j - 1).$$

We have to draw attention that the prior parameter β for the i-th posterior beta distribution is $\beta_i^* = \beta + f(i, \varphi)$. In addition, we get K_n runs of the software during n periods of observations. This implies that the posterior parameter β_i^* for i-th period of debugging is defined by n periods of observations. This is a very important feature and that is why we use index i for the posterior parameter β^*.

It can be also seen from the above that the posterior parameters depend on $\mathbf{d}$. In the considered special case, β^* depends on $f(i, \varphi)$.

Now we can write the predictive CDF for the i-th step of the software debugging after n observations as

follows:

$$F_i(k \mid \varphi, \alpha, \beta) = \int_0^1 (1-(1-p)^k) \cdot \mathrm{Beta}(\alpha^*, \beta_i^*)\mathrm{d}p = 1 - \frac{\mathrm{B}(\alpha^* + \beta_i^*, k)}{\mathrm{B}(\beta_i^*, k)}.$$

By using the introduced notation $\alpha = s\gamma$, $\beta = s - s\gamma$, we write

$$F_i(k \mid \varphi, \gamma, s) = 1 - \frac{\mathrm{B}(s+n-1+D_i(\varphi), k)}{\mathrm{B}(s - s\gamma + D_i(\varphi), k)}.$$

The function $F_i(k \mid \varphi, \gamma, s)$ increases as γ increases in the interval $[0,1]$, because the beta function $\mathrm{B}(x,y)$ is decreasing in x for $x > 0$. This implies that the lower bound for $\mathcal{M}_i(\varphi)$ is determined as

$$\underline{F}_i(k \mid \varphi, s) = \sup_{\gamma \in (0,1)} F_i(k \mid \varphi, \gamma, s) = 1 - \frac{\mathrm{B}(s+n-1+D_i(\varphi), k)}{\mathrm{B}(s + D_i(\varphi), k)}.$$

The upper bound is determined as

$$\overline{F}_i(k \mid \varphi, s) = \inf_{\gamma \in (0,1)} F_i(k \mid \varphi, \gamma, s) = 1 - \frac{\mathrm{B}(s+n-1+D_i(\varphi), k)}{\mathrm{B}(D_i(\varphi), k)}.$$

By having the lower and upper CDFs, it follows from Proposition 1 that the likelihood function maximized over $\mathcal{M}_i(\varphi)$ by given s and φ is of the form:

$$\begin{aligned}&\max_{\mathcal{M}(\varphi)} L(\mathbf{K} \mid \varphi, s) \\ &= \prod_{i=1}^{n} \left(\overline{F}_i(k_i \mid \varphi, s) - \underline{F}_i(k_i - 1 \mid \varphi, s)\right) \\ &= \prod_{i=1}^{n} \left(\frac{\mathrm{B}(C_i,\ k_i - 1)}{\mathrm{B}(s + D_i(\varphi),\ k_i - 1)} - \frac{\mathrm{B}(C_i,\ k_i)}{\mathrm{B}(D_i(\varphi),\ k_i)} \right).\end{aligned}$$

Here $C_i = s + n - 1 + D_i(\varphi)$.

The parameter φ should be chosen in such a way that makes $\ln L(\mathbf{K} \mid \varphi, s)$ achieve its maximum. The optimal value φ_0 of φ can be found by numerically solving the equation $\partial \ln L(\mathbf{K} \mid \varphi, s)/\partial\varphi = 0$. Once we have calculated the estimate of the parameter φ, we can derive the lower and upper software run failure functions after the n-th software failure, i.e., we can compute the lower and upper CDFs of the $(n+1)$-th failure

$$\underline{F}_{n+1}(k, s) = 1 - \frac{\mathrm{B}(s+n+D_{n+1}(\varphi_0),\ k)}{\mathrm{B}(s + D_{n+1}(\varphi_0),\ k)},$$

$$\overline{F}_{n+1}(k, s) = 1 - \frac{\mathrm{B}(s+n+D_{n+1}(\varphi_0),\ k)}{\mathrm{B}(D_{n+1}(\varphi_0),\ k)}.$$

7 NHPP software reliability models

One of the important frameworks for developing software reliability models dealing with numbers $N(t)$ of software failures occurring up to a certain time period t is the non-homogeneous Poisson process (NHPP). Let $X_i = N(t_i) - N(t_{i-1})$ be the random number of failures between t_{i-1} and t_i. For any time points $0 < t_1 < t_2 < ...$ (for ease of notation, let $t_0 = 0$), the probability that the number of failures between t_{i-1} and t_i is k, $k = 0, 1, 2, ...$, can be written as

$$\begin{aligned}&\Pr\{N(t_i) - N(t_{i-1}) = k\} \\ &= \frac{\{m(t_i) - m(t_{i-1})\}^k}{k!} e^{\{-(m(t_i) - m(t_{i-1}))\}}. \qquad (4)\end{aligned}$$

Here $m(t)$ is the mean number of failures occurring up to time t. The NHPP models differ through the function $m(t)$, popular examples of which for software reliability models are $m(t) = a(1 - \exp(-bt))$ (Goel-Okumoto model [5]) and $m(t) = a\ln(1+bt)$ (Musa-Okumoto model [10]). Our goal is to estimate the parameters a and b for such a model, based on statistical data consisting of numbers of failures k_i per subintervals $(t_{i-1}, t_i]$, $i = 1, ..., n$. As before, we denote these data by the vector $\mathbf{K} = (k_1, ..., k_n)$.

7.1 The imprecise negative binomial model

When the number of failures has a Poisson distribution with the parameter λ, gamma distributions are conjugate priors, denoted by $\mathrm{Gamma}(\alpha, \beta)$. If we observed K failures during a period of time T, then the posterior distribution is $\mathrm{Gamma}(\alpha^*, \beta^*)$, where $\alpha^* = \alpha + K$ and $\beta^* = \beta + T$. Hence, the predictive probability of k failures during time t under condition that K failures were observed during time T is [2]

$$P(k) = \int_0^\infty \frac{(\lambda t)^k e^{-\lambda t}}{k!} \mathrm{Gamma}(\alpha^*, \beta^*)\mathrm{d}\lambda = \frac{\Gamma(\alpha^* + k)}{\Gamma(\alpha^*)k!} \left(\frac{\beta^*}{\beta^* + t}\right)^{\alpha^*} \left(\frac{t}{\beta^* + t}\right)^k. \qquad (5)$$

Here $\Gamma(\alpha)$ is the standard gamma function.

7.2 The imprecise negative binomial growth model

A wide range of suitable mean value functions can be represented in the form $m(t; a, b) = a \cdot \tau(t, b)$. The parameter λ of the Poisson distribution in (5) and the argument t can be replaced by the parameter a and the discrete time $\tau(t_i, b) - \tau(t_{i-1}, b)$, respectively. In fact, by replacing λ by a, we get the Poisson process with a scaled time of the software testing, i.e., every time interval $[t_{i-1}, t_i]$ is replaced by the interval

$[\tau(t_{i-1},b),\tau(t_i,b)]$. Then we can write the predictive CDF of the number of failures in the time interval between t_i and t ($t\in[t_i,t_{i+1}]$) after n observation periods through the regularized incomplete Beta-function [7] as follows:

$$F_i(k,t|\mathbf{c},b)=1-\frac{\mathrm{B}_{q(i,t)}(k+1,\alpha+K_n)}{\mathrm{B}(k+1,\alpha+K_n)}$$
$$=1-I\left(q(i,t),k+1,\alpha+K_n\right).$$

Here $t_0=0$, $k_0=0$,

$$q(i,t)=\frac{T_i(t,b)}{\beta+\tau(t_n,b)+T_i(t,b)},$$
$$T_i(t,b)=\tau(t,b)-\tau(t_i,b),\ K_n=\sum_{j=1}^{n}k_j,$$

$\mathrm{B}_q(k+1,r)$ is the incomplete Beta-function with $I\left(q,k,r\right)$ the regularized incomplete Beta-function.

We must select a bounded set for the vector (α,β), in order to avoid ending up with vacuous posterior predictive distributions. In analogy with imprecise prior classes described above, we want this set to be described by a single hyper-parameter s, and we choose all vectors (α,β) within the triangle $(0,0)$, $(s,0)$, $(0,s)$. This implies that all possible prior 'rates of occurrence of failures' are represented, as the prior allows interpretation of $\alpha/\beta=\gamma$ as this rate, hence this would include all such rates in $(0,\infty)$. This prior set, and related inferences, is of course similar in nature to the work by Quaeghebeur and de Cooman [11], yet it is slightly different. This prior set leads to the lower and upper bounds for $\mathcal{M}_i(b)$ by $t\in[t_i,t_{i+1}]$

$$\underline{F}_i(k,t\mid s,b)=1$$
$$-I\left(\frac{T_i(t,b)}{\tau(t_n,b)+T_i(t,b)},k+1,s+K_n\right),$$

$$\overline{F}_i(k,t\mid s,b)=1$$
$$-I\left(\frac{T_i(t,b)}{s+\tau(t_n,b)+T_i(t,b)},k+1,K_n\right).$$

The next step is to use Proposition 1 and to maximize the likelihood function over the set of b

$$L(\mathbf{K}|b,s)=\prod_{i=1}^{n}\left(\overline{F}_i(k_i,t_i\mid s,b)-\underline{F}_i(k_i-1,t_i\mid s,b)\right).$$

Once we have the maximum likelihood estimator, following Proposition 1, of the parameter b, we can construct the lower and upper bounds for the CDF of the number of failures in time interval $[t_n,t]$ after n periods of debugging.

8 Regression model (general scheme)

We briefly explain how the combined imprecise Bayes and likelihood approach, proposed in this paper, can be applied to basic regression problems. Suppose that we have $n+1$ variables Y and X_j, $j=1,...,n$, with Y being a dependent variable and $\{X_1,...,X_n\}$ being n independent predictor variables, related to Y according to the relation $Y=f(X_1,...,X_n)$. The standard linear regression model[3] is a special case and can be written as

$$Y=\mathbf{X}\mathbf{d}+\epsilon.$$

Here $\mathbf{X}=(1,X_1,...,X_n)$; $\mathbf{d}=(d_0,...,d_n)^{\mathrm{T}}$ is the vector of parameters; ϵ are random errors or noise having zero mean and the unknown variance σ^2.

To fit with the presentation in this paper, we assume that ϵ is a discrete variable[4]. Let us construct the imprecise Bayesian model for ϵ. If ϵ is governed by some probability distribution $p(z\mid\sigma)$ and there is the corresponding conjugate distribution $\pi(\sigma\mid\mathbf{c})$, then we can find the predictive CDF $F_n(z\mid s,\gamma)$ after having n observations $(y_1,\mathbf{x}_1),...,(y_n,\mathbf{x}_n)$ depending on new parameters s,γ [11] and its bounds $\underline{F}(z\mid s)$, $\overline{F}(z\mid s)$.

Denote $z_i=y_i-\mathbf{x}_i\mathbf{d}$ and $\mathbf{Z}=(z_1,...,z_n)$. Having derived the lower and upper CDFs, it follows from Proposition 1 that the likelihood function that is to be maximized over $\mathcal{M}$, with given s, is of the form:

$$\max_{\mathcal{M}}L(\mathbf{Z}\mid s)=\prod_{i=1}^{n}\left(\overline{F}(z_i\mid s)-\underline{F}(z_i-1\mid s)\right).$$

Denote $z_i=y_i-\mathbf{x}_i\mathbf{d}$. Hence

$$\max_{\mathcal{M}}L(\mathbf{Z}\mid s)$$
$$=\prod_{i=1}^{n}\left(\overline{F}(y_i-\mathbf{x}_i\mathbf{d}\mid s)-\underline{F}(y_i-\mathbf{x}_i\mathbf{d}-1\mid s)\right).$$

Now we can find parameters $\mathbf{d}$ by maximizing the obtained likelihood function.

In the regression model, we again separate the parameters of the probability distribution of ϵ and the parameters $\mathbf{d}$. However, in contrast to the software reliability models, the parameters $\mathbf{c}$ directly do not change with the growth parameters $\mathbf{d}$ (see the parameter β^* and the function $f(i)$ in Subsection 6.2 for comparison). Moreover, the set $\mathcal{M}$ and its bounds do not depend on the parameters $\mathbf{d}$. This allows us to avoid the index i and to consider identical sets $\mathcal{M}$. Nevertheless, the general approach for modelling and inference is the same as described in this paper.

[3] The more general model $Y=f(\mathbf{X},\mathbf{d})+\epsilon$ which can be analyzed in the same way.

[4] See Section 9 for comments relevant to the more usual case with continuous ϵ

9 Concluding remarks

In this paper we have proposed a way towards development of statistical methods that combine imprecise Bayesian inference for one subset of all parameters with maximum likelihood estimation for the other parameters. The key to this approach is Proposition 1, which provides a generalization of maximum likelihood estimation for discrete variables with sets of distributions. There are many important research questions that need answering, in particular with regard to the interpretations of these inferences and their application to large scale problems. We particularly see a benefit in models with differing features related to different parameters, for example the reliability growth models discussed in some detail and used to present and illustrate the novel approach in this paper, where some parameters are specifically used to model the growth aspect. It should also be studied in which situations this approach is most valuable. For example, it may well be most suitable in situations where one has significant prior knowledge on some parameters, yet does not feel confident enough to assign precise prior distributions to them, whereas on another aspect of the model one has no prior knowledge and explicitly wishes only to estimate those parameters using the data. Some statisticians might object if the same data set is used for related inference in two different stages, feeling that the same data might be used twice. This would be wrong, as the parameters estimated at the different stages play different roles, and hence estimates are based on different aspects of the information within the total data set available.

We presented the main idea of the new framework in this paper as an extension of the known imprecise Bayesian models [11, 16] to situations where the process considered has some changeable behaviour, which we also wish to estimate using the data. In line with most reported developments in such imprecise Bayesian models, we presented it from the perspective of a non-informative prior set of distributions, but it may indeed well be more useful to apply this combined method with an informative prior set of distributions. When such sets are also defined using conjugate priors in the same way as for these non-informative prior sets, that is done in a straightforward manner which we will discuss and explore further elsewhere. We chose to focus our presentation on software reliability growth models, as these typically have clear divisions of the parameters according to the different roles, which we consider very suitable for the method proposed. As indicated, the general approach might also provide a promising method for imprecise regression models.

We have stated in Section 3 that the set $\mathcal{M}_i(\mathbf{d})$ is the set of *all* CDFs bounded by $\underline{F}_i$ and $\overline{F}_i$. One could also consider the use of only a set of parametric distributions, all with the same parametric form as the bounding distributions. However, following this approach, maximization of the likelihood function over a set of distributions with parameters $\mathbf{c}$ derived in Section 3 is reduced to its maximization over a set of parameters $\mathbf{c}$. In this case, we get the standard statistical model completely based on the maximum likelihood estimation, which does not differ from many well-known models of software reliability and regression models.

Due to limited size of this paper, we did not illustrate the proposed models by data examples, such examples will be included in specific topic oriented presentations elsewhere, where we also compare these inferences to other inferences including full Bayesian and full likelihood approaches. Nevertheless, we wish to point out that initial indications from computational examples suggest that this new combined method performs well, also so if there are relatively few data, but further study is required in order to draw general conclusions.

We did not consider continuous random variables, but of course this case is very important. However, Proposition 1 can be extended on the continuous case, so it looks like the method can also be applied for continuous random variables $X_1, ..., X_n$. In this case, the likelihood function can be written as

$$L(\mathbf{X}) = \lim_{\triangle_1 \to 0, \ldots, \triangle_n \to 0} \frac{\Pr\{x_1 \le X_1 \le x_1 + \triangle_1, .., x_n \le X_n \le x_n + \triangle_n\}}{\triangle_1 \cdots \triangle_n},$$

and this suggests that maximum likelihood estimates for the parameters can be derived by maximizing

$$\max_{\mathcal{M}_1, \ldots, \mathcal{M}_n} L(\mathbf{X}) = \prod_{i=1}^{n} \left(\overline{F}_i(x_i) - \underline{F}_i(x_i)\right) \delta(x_i). \quad (6)$$

Here $\delta(x_i)$ is Dirac function which has unit area concentrated in the immediate vicinity of points x_i. The likelihood function achieves its maximum by taking the probability density functions such that $\rho_i(x_i) = \left(\overline{F}_i(x_i) - \underline{F}_i(x_i)\right) \delta(x_i)$. However, whether or not condition (6) is fully correct is yet to be established, which is an important topic for further research.

The continuous case would enable many application models. For example, it would enable our combined method to be applied to regression models with the common assumption that the random variable ϵ is normally distributed, $\mathcal{N}(0, \sigma^2)$, where a gamma distribution Gamma(α, β) can be used as conjugate prior

for $1/\sigma^2$. Hence, the predictive probability density function after having n observations is of the form:

$$p(z|s,\gamma) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{s+k+3}{2}\right)}{\Gamma\left(\frac{s+k+2}{2}\right)} \frac{(s\gamma+\tau_k)^{\frac{s+k+2}{2}}}{(s\gamma+\tau_k+z^2)^{\frac{s+k+3}{2}}},$$

where

$$\tau_k = \sum_{j=1}^{k} z_j^2 = \sum_{j=1}^{k} (y_i - f(\mathbf{X}_i, \mathbf{d}))^2 .$$

By using the imprecise Bayesian normal model [11], we can then construct the imprecise regression model combining imprecise Bayesian inference with maximum likelihood estimation as briefly discussed in Section 8 where only discrete random variables were used in line with the general presentation in this paper. So, establishing the detailed and fully justified generalization of the approach in this paper to continuous random variables is very important, and we are hopeful to report on this in the near future.

Acknowledgement

We thank referees for useful and detailed suggestions that improved the paper.

References

[1] J.-M. Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39(2-5):123–150, 2005.

[2] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley, Chichester, 1994.

[3] K.Y. Cai. Towards a conceptual framework of software run reliability modeling. *Information Sciences*, 126:137–163, 2000.

[4] S. Ferson. Bayesian methods in risk assessment. Technical report, RAMAS, 2006.

[5] A.L. Goel and K. Okomoto. Time dependent error detection rate model for software reliability and other performance measures. *IEEE Trans. Reliab.*, R-28:206–211, 1979.

[6] Z. Jelinski and P.B. Moranda. Software reliability research. In W. Greiberger, editor, *Statistical Computer Performance Evaluation*, pages 464–484. Academic Press, New York, 1972.

[7] N.L. Jonson, S. Kotz, and A.W. Kemp. *Univariate discrete distributions*. Wiley, New York, 1992.

[8] B. Littlewood and J. Verall. A Bayesian reliability growth model for computer software. *Applied Statistics*, 22:332–346, 1973.

[9] C. Loader. *Local regression and likelihood*. Springer-Verlag, New York, 1999.

[10] J.D. Musa, A. Iannino, and K. Okumoto. *Software Reliability: Mesurement, Prediction, Application*. McGraw-Hill, 1987.

[11] E. Quaeghebeur and G. de Cooman. Imprecise probability models for inference in exponential families. In J.-M. Bernard, T. Seidenfeld, and M. Zaffalon, editors, *Proc. of the 4rd Int. Symposium on Imprecise Probabilities and Their Applications, ISIPTA'05*, Pittsburgh, Pennsylvania, July 2005. Carnegie Mellon University.

[12] C.P. Robert. *The Bayesian Choice*. Springer, New York, 1994.

[13] G.J. Schick and R.W. Wolverton. An analysis of competing software reliability models. *IEEE Trans. on Software Engineering*, SE-4:104–120, 1978.

[14] A.R. Syversveen. Noninformative Bayesian priors. Interpretation and problems with construction and applications. Preprint Statistics 3, Department of Mathematical Sciences, NTNU, Trondheim, 1998.

[15] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[16] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996. with discussion.

[17] G. Walter, Th. Augustin, and A. Peters. Linear regression analysis under sets of conjugate priors. In G. de Cooman, J. Vejnarova, and M. Zaffalon, editors, *Proceedings of the Fifth International Symposium on Imprecise Probabilities and Their Applications*, pages 445–455, Prague, Czech Republic, 2007.

[18] M. Xie. *Software Reliability Modeling*. World Scientific, 1991.

On Conditional Independence in Evidence Theory

Jiřina Vejnarová
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
&
University of Economics, Prague
vejnar@utia.cas.cz

Abstract

The goal of this paper is to introduce a new concept of conditional independence in evidence theory, to prove its formal properties, and to show in what sense it is superior to the concept introduced previously by other authors.

Keywords. Evidence theory, random sets independence, conditional independence, conditional noninteractivity.

1 Introduction

Any application of artificial intelligence models to practical problems must manage two basic issues: uncertainty and multidimensionality. The models currently most widely used to manage these issues are so-called *probabilistic graphical Markov models*.

In these models, the problem of multidimensionality is solved using the notion of conditional independence, which enables factorisation of a multidimensional probability distribution into small parts, usually marginal or conditional low-dimensional distributions (or generally into low-dimensional factors). Such a factorisation not only decreases the storage requirements for representation of a multidimensional distribution, but it usually induces efficient computational procedures allowing inference from these models as well. Many results analogous to those concerning conditional independence, Markov properties and factorisation from probabilistic framework were also achieved in possibility theory [12, 13].

It is easy to realise that our need of efficient methods for representation of probability and possibility distributions (requiring an exponential number of parameters) logically leads us to greater need of an efficient tool for representation of belief functions, which cannot be represented by a distribution (but only by a set function), and therefore the space requirements for their representation are superexponential.

After a thorough study of relationships among stochastic independence, possibilistic T-independence, random set independence and strong independence [14, 15], we came to the conclusion that the most proper independence concept in evidence theory is random set independence. Therefore, this contribution is fully devoted to two different generalisations of random set independence to conditional independence.

The contribution is organised as follows. After a short overview of necessary terminology and notation (Section 2), in Section 3 we introduce a new concept of conditional independence and show in which sense it is superior to the previously suggested independence notions [10, 1]. In Section 4 we prove its formal properties.

2 Basic Notions

The aim of this section is to introduce as briefly as possible basic notions and notations necessary for understanding the following text.

2.1 Set Projections and Extensions

For an index set $N = \{1, 2, \ldots, n\}$ let $\{X_i\}_{i\in N}$ be a system of variables, each X_i having its values in a finite set $\mathbf{X}_i$. In this paper we will deal with *multidimensional frame of discernment*

$$\mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_n,$$

and its *subframes* (for $K \subseteq N$)

$$\mathbf{X}_K = \bigtimes_{i\in K} \mathbf{X}_i.$$

When dealing with groups of variables on these subframes, X_K will denote a group of variables $\{X_i\}_{i\in K}$ throughout the paper.

A *projection* of $x = (x_1, x_2, \ldots, x_n) \in \mathbf{X}_N$ into $\mathbf{X}_K$ will be denoted $x^{\downarrow K}$, i.e., for $K = \{i_1, i_2, \ldots, i_k\}$

$$x^{\downarrow K} = (x_{i_1}, x_{i_2}, \ldots, x_{i_k}) \in \mathbf{X}_K.$$

Analogously, for $M \subset K \subseteq N$ and $A \subset \mathbf{X}_K$, $A^{\downarrow M}$ will denote a *projection* of A into $\mathbf{X}_M$:[1]

$$A^{\downarrow M} = \{y \in \mathbf{X}_M \mid \exists x \in A : y = x^{\downarrow M}\}.$$

In addition to the projection, in this text we will also need an opposite operation, which will be called an extension. By an *extension* of two sets $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_L$ $(K, L \subseteq N)$ we will understand a set

$$A \otimes B = \{x \in \mathbf{X}_{K\cup L} : x^{\downarrow K} \in A \ \ \& \ \ x^{\downarrow L} \in B\}.$$

Let us note that if K and L are disjoint, then

$$A \otimes B = A \times B.$$

2.2 Set Functions

In evidence theory (or Dempster-Shafer theory) two measures are used to model the uncertainty: belief and plausibility measures. Both of them can be defined with the help of another set function called a *basic (probability* or *belief) assignment* m on $\mathbf{X}_N$, i.e.,

$$m : \mathcal{P}(\mathbf{X}_N) \longrightarrow [0, 1] \tag{1}$$

for which

$$\sum_{A\subseteq\mathbf{X}_N} m(A) = 1. \tag{2}$$

Furthermore, we assume that $m(\emptyset) = 0$.

Belief and *plausibility* measures are defined for any $A \subseteq \mathbf{X}_N$ by the equalities

$$Bel(A) = \sum_{B\subseteq A} m(B),$$
$$Pl(A) = \sum_{B\cap A\neq\emptyset} m(B),$$

respectively.

It is well-known (and evident from these formulae) that for any $A \in \mathcal{P}(\mathbf{X}_N)$

$$Pl(A) = 1 - Bel(A^C) \tag{3}$$

holds, where A^C is a set complement of $A \in \mathcal{P}(\mathbf{X}_N)$. Furthermore, basic assignment can be computed from belief function via Möbius inversion:

$$m(A) = \sum_{B\subseteq A} (-1)^{|A\setminus B|} Bel(B), \tag{4}$$

i.e., any of these three functions is sufficient to define values of the remaining two.

[1] Let us remark that we do not exclude situations when $M = \emptyset$. In this case $A^{\downarrow\emptyset} = \emptyset$.

In addition to belief and plausibility measures, *commonality function* can also be obtained from basic assignment m:

$$Q(A) \quad = \quad \sum_{B\supseteq A} m(B).$$

The last notion plays an important role in the definition of so-called (conditional) noninteractivity of variables (cf. Section 3.2) and in Shenoy's valuation-based systems [10]. Similarly to (4), one can obtain basic assignment from commonality function via an analogous formula

$$m(A) = \sum_{B\supseteq A} (-1)^{|B\setminus A|} Q(B). \tag{5}$$

A set $A \in \mathcal{P}(\mathbf{X}_N)$ is a *focal element* if $m(A) > 0$. A pair $(\mathcal{F}, m)$, where $\mathcal{F}$ is the set of all focal elements, is called a *body of evidence*. A basic assignment is called *Bayesian* if all its focal elements are singletons. A body of evidence is called *consonant* if its focal elements are nested.

For a basic assignment m on $\mathbf{X}_K$ and $M \subset K$, a *marginal basic assignment* of m is defined (for each $A \subseteq \mathbf{X}_M$):

$$m^{\downarrow M}(A) = \sum_{B\subseteq\mathbf{X}_K : B^{\downarrow M}=A} m(B).$$

Analogously, $Bel^{\downarrow M}$, $Pl^{\downarrow M}$ and $Q^{\downarrow M}$ will denote the corresponding marginal belief measure, plausibility measure and commonality function, respectively.

Having two basic assignments m_1 and m_2 on $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively $(K, L \subseteq N)$, we say that these assignments are *projective* if

$$m_1^{\downarrow K\cap L} = m_2^{\downarrow K\cap L},$$

which occurs if and only if there exists a basic assignment m on $\mathbf{X}_{K\cup L}$ such that both m_1 and m_2 are marginal assignments of m.

3 Random Set Independence and Its Generalisations

3.1 Marginal Case

Let us start this section by recalling the notion of random sets independence [2].[2]

Definition 1 Let m be a basic assignment on $\mathbf{X}_N$ and $K, L \subset N$ be disjoint. We say that groups of

[2] Klir [6] uses the notion *noninteractivity*.

variables X_K and X_L are *independent with respect to basic assignment* m (and denote it by $K \perp\!\!\!\perp L\ [m]$) if

$$m^{\downarrow K\cup L}(A) = m^{\downarrow K}(A^{\downarrow K}) \cdot m^{\downarrow L}(A^{\downarrow L}) \qquad (6)$$

for all $A \subseteq \mathbf{X}_{K\cup L}$ for which $A = A^{\downarrow K} \times A^{\downarrow L}$, and $m(A) = 0$ otherwise.

It has been shown in [14] that application of Definition 1 to two consonant bodies of evidence leads to a body of evidence which is no longer consonant.

It seemed that this problem could be avoided if we took into account the fact that both evidence and possibility theories could be considered as special kinds of imprecise probabilities. Nevertheless, in [15] we showed that the application of strong independence to two general bodies of evidence (neither Bayesian nor consonant) leads to models beyond the framework of evidence theory.

From these examples one can see that although models based on possibility measures, belief measures and credal sets can be linearly ordered with respect to their generality, nothing similar holds for the corresponding independence concepts.

Therefore, random sets independence presently seems to be the most appropriate independence concept within the framework of evidence theory from the viewpoint of multidimensional models.[3] For this reason in this section we will deal with two generalisations of this concept.

Before doing that, let us present an assertion showing that conditional noninteractivity and conditional independence (presented in the following two subsections) are identical if the condition is empty.

Lemma 1 *Let K, L be disjoint, then $K \perp\!\!\!\perp L\ [m]$ if and only if*

$$Q^{\downarrow K\cup L}(A) = Q^{\downarrow K}(A^{\downarrow K}) \cdot Q^{\downarrow L}(A^{\downarrow L}) \qquad (7)$$

for all $A \subseteq \mathbf{X}_{K\cup L}$.

Proof can be found in [5].

From this lemma one can conjecture why the generalisation of (7) to the conditional case became widely used, while, as far as we know, no direct generalisation of (6) has been suggested up to now.

In the following example we will show that nothing similar holds for beliefs and plausibilities; more exactly, application of formulae analogous to (7) leads to models beyond the theory of evidence.

[3] Let us note that there exist different independence concepts suitable in other situations, for details the reader is referred to [3]

Table 1: Basic assignments m_X and m_Y.

$A \subseteq \mathbf{X}$	$m_X(A)$	$Bel_X(A)$	$Pl_X(A)$
$\{x\}$	0.3	0.3	0.8
$\{\bar{x}\}$	0.2	0.2	0.7
$\mathbf{X}$	0.5	1	1
$A \subseteq \mathbf{Y}$	$m_Y(A)$	$Bel_Y(A)$	$Pl_Y(A)$
$\{y\}$	0.6	0.6	0.9
$\{\bar{y}\}$	0.1	0.1	0.4
$\mathbf{Y}$	0.3	1	1

Table 2: Results of application of formula (8).

$C \subseteq \mathbf{X} \times \mathbf{Y}$	$Bel_{XY}(C)$	$m_{XY}(C)$
$\{xy\}$	0.18	0.18
$\{x\bar{y}\}$	0.03	0.03
$\{\bar{x}y\}$	0.12	0.12
$\{\bar{x}\bar{y}\}$	0.02	0.02
$\{x\} \times \mathbf{Y}$	0.3	0.09
$\{\bar{x}\} \times \mathbf{Y}$	0.2	0.06
$\mathbf{X} \times \{y\}$	0.6	0.3
$\mathbf{X} \times \{\bar{y}\}$	0.1	0.05
$\{xy, \bar{x}\bar{y}\}$	1	0.8
$\{x\bar{y}, \bar{x}y\}$	1	0.85
$\mathbf{X} \times \mathbf{Y} \setminus \{\bar{x}\bar{y}\}$	1	-1.08
$\mathbf{X} \times \mathbf{Y} \setminus \{\bar{x}y\}$	1	-0.43
$\mathbf{X} \times \mathbf{Y} \setminus \{x\bar{y}\}$	1	-0.92
$\mathbf{X} \times \mathbf{Y} \setminus \{xy\}$	1	-0.52
$\mathbf{X} \times \mathbf{Y}$	1	0.45

Example 1 Consider two basic assignments m_X and m_Y on $\mathbf{X} = \{x, \bar{x}\}$ $\mathbf{Y} = \{y, \bar{y}\}$ specified in Table 1 together with their beliefs and plausibilities.

Let us compute joint beliefs and plausibilities via formulae

$$Bel^{\downarrow K\cup L}(A) = Bel^{\downarrow K}(A^{\downarrow K}) \cdot Bel^{\downarrow L}(A^{\downarrow L}), \quad (8)$$
$$Pl^{\downarrow K\cup L}(A) = Pl^{\downarrow K}(A^{\downarrow K}) \cdot Pl^{\downarrow L}(A^{\downarrow L}). \quad (9)$$

Their values are contained in Tables 2 and 3, respectively, together with the corresponding values of basic assignments computed via (4) (and also (3), in the latter case). As some values of the "joint basic assignments" are negative, which contradicts to (1) it is evident that these models are beyond the framework of evidence theory. ♢

Therefore, it seems that a definition of independence in terms of beliefs or plausibilities would be much

Table 3: Results of application of formula (9).

$C \subseteq \mathbf{X} \times \mathbf{Y}$	$Pl_{XY}(C)$	$Bel_{XY}(C)$	$m_{XY}(C)$
$\{xy\}$	0.72	0	0
$\{x\bar{y}\}$	0.32	0	0
$\{\bar{x}y\}$	0.63	0	0
$\{\bar{x}\bar{y}\}$	0.28	0	0
$\{x\} \times \mathbf{Y}$	0.8	0.3	0.3
$\{\bar{x}\} \times \mathbf{Y}$	0.7	0.2	0.2
$\mathbf{X} \times \{y\}$	0.9	0.6	0.6
$\mathbf{X} \times \{\bar{y}\}$	0.4	0.1	0.1
$\{xy, \bar{x}\bar{y}\}$	1	0	0
$\{x\bar{y}, \bar{x}y\}$	1	0	0
$\mathbf{X} \times \mathbf{Y} \setminus \{\bar{x}\bar{y}\}$	1	0.72	−0.18
$\mathbf{X} \times \mathbf{Y} \setminus \{\bar{x}y\}$	1	0.37	0.07
$\mathbf{X} \times \mathbf{Y} \setminus \{x\bar{y}\}$	1	0.68	−0.12
$\mathbf{X} \times \mathbf{Y} \setminus \{xy\}$	1	0.28	−0.02
$\mathbf{X} \times \mathbf{Y}$	1	1	0.05

more complicated than Definition 1.

3.2 Conditional Noninteractivity

Ben Yaghlane et al. [1] generalised the notion of noninteractivity in the following way: Let m be a basic assignment on $\mathbf{X}_N$ and $K, L, M \subset N$ be disjoint, $K \neq \emptyset \neq L$. Groups of variables X_K and X_L are *conditionally noninteractive given X_M with respect to m* if and only if the equality

$$Q^{\downarrow K\cup L\cup M}(A) \cdot Q^{\downarrow M}(A^{\downarrow M}) = Q^{\downarrow K\cup M}(A^{\downarrow K\cup M}) \cdot Q^{\downarrow L\cup M}(A^{\downarrow L\cup M}) \quad (10)$$

holds for any $A \subseteq \mathbf{X}_{K\cup L\cup M}$.

Let us note that the definition presented in [1] is based on conjunctive Dempster's rule, but the authors proved its equivalence with (10). Let us also note that (10) is a special case of the definition of conditional independence in valuation-based systems[4] introduced by Shenoy [10].

The cited authors proved in [1] that conditional noninteractivity satisfies the so-called graphoid properties.[5]

Nevertheless, this notion of independence does not seem to be appropriate for construction of multidimensional models. As already mentioned by Studený [11], it is not consistent with marginalisation. What that means can be seen from the following definition and illustrated by a simple example from [1] (originally suggested by Studený).

An independence concept is *consistent with marginalisation* iff for arbitrary projective basic assignments (probability distributions, possibility distributions, etc.) m_1 on $\mathbf{X}_K$ and m_2 on $\mathbf{X}_L$ there exists a basic assignment (probability distribution, possibility distribution, etc.) on $\mathbf{X}_{K\cup L}$ satisfying this independence concept and having m_1 and m_2 as its marginals.

Example 2 Let X_1, X_2 and X_3 be three binary variables with values in $\mathbf{X}_1 = \{a_1, \bar{a}_1\}, \mathbf{X}_2 = \{a_2, \bar{a}_2\}$, $\mathbf{X}_3 = \{a_3, \bar{a}_3\}$ and m_1 and m_2 be two basic assignments on $\mathbf{X}_1 \times \mathbf{X}_3$ and $\mathbf{X}_2 \times \mathbf{X}_3$ respectively, both of them having only two focal elements:

$$\begin{aligned} &m_1(\{(a_1, \bar{a}_3), (\bar{a}_1, \bar{a}_3)\}) = .5, \\ &m_1(\{(a_1, \bar{a}_3), (\bar{a}_1, a_3)\}) = .5, \\ &m_2(\{(a_2, \bar{a}_3), (\bar{a}_2, \bar{a}_3)\}) = .5, \\ &m_2(\{(a_2, \bar{a}_3), (\bar{a}_2, a_3)\}) = .5. \end{aligned} \quad (11)$$

Since their marginals are projective

$$\begin{aligned} m_1^{\downarrow 3}(\{\bar{a}_3\}) = m_2^{\downarrow 3}(\{\bar{a}_3\}) = .5, \\ m_1^{\downarrow 3}(\{a_3, \bar{a}_3\}) = m_2^{\downarrow 3}(\{a_3, \bar{a}_3\}) = .5, \end{aligned}$$

there exists (at least one) common extension of both of them, but none of them is such that it would imply conditional noninteractivity of X_1 and X_2 given X_3. Namely, the application of equality (10) to basic assignments m_1 and m_2 leads to the following values of the joint "basic assignment":

$$\begin{aligned} &\bar{m}(\mathbf{X}_1 \times \mathbf{X}_2 \times \{\bar{a}_3\}) = .25, \\ &\bar{m}(\mathbf{X}_1 \times \{a_2\} \times \{\bar{a}_3\}) = .25, \\ &\bar{m}(\{a_1\} \times \mathbf{X}_2 \times \{\bar{a}_3\}) = .25, \\ &\bar{m}(\{(a_1, a_2, \bar{a}_3), (\bar{a}_1, \bar{a}_2, a_3)\}) = .5, \\ &\bar{m}(\{(a_1, a_2, \bar{a}_3)\}) = -.25, \end{aligned}$$

which is outside of evidence theory. ♢

Therefore, instead of the conditional noninteractivity, in [5] we proposed to use another notion of conditional independence which will be introduced in the following subsection.

3.3 Conditional Independence

Definition 2 Let m be a basic assignment on $\mathbf{X}_N$ and $K, L, M \subset N$ be disjoint, $K \neq \emptyset \neq L$. We say that groups of variables X_K and X_L are *conditionally independent given X_M with respect to m* (and denote it by $K \perp\!\!\!\perp L | M$ $[m]$), if the equality

$$\begin{aligned} m^{\downarrow K\cup L\cup M}(A) \cdot m^{\downarrow M}(A^{\downarrow M}) \\ = \; m^{\downarrow K\cup M}(A^{\downarrow K\cup M}) \cdot m^{\downarrow L\cup M}(A^{\downarrow L\cup M}) \end{aligned} \quad (12)$$

[4] Nevertheless, in valuation-based systems commonality function is a primitive concept (and basic assignment is derived by formula (5)).

[5] The reader not familiar with graphoid axioms is referred to the beginning of Section 4.

holds for any $A \subseteq \mathbf{X}_{K\cup L\cup M}$ such that $A = A^{\downarrow K\cup M} \otimes A^{\downarrow L\cup M}$, and $m(A) = 0$ otherwise.

Let us note that for $M = \emptyset$ the concept coincides with Definition 1, which enables us to use the term conditional independence. Let us also note that (12) resembles, from the formal point of view, the definition of stochastic conditional independence [7].

The following assertion expresses the fact (already mentioned above) that this concept of conditional independence is consistent with marginalisation. Moreover, it presents a form expressing the joint basic assignment by means of its marginals.

Theorem 1 *Let m_1 and m_2 be projective basic assignments on $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively. Let us define a basic assignment m on $\mathbf{X}_{K\cup L}$ by the formula*

$$m(A) = \frac{m_1(A^{\downarrow K}) \cdot m_2(A^{\downarrow L})}{m_2^{\downarrow K\cap L}(A^{\downarrow K\cap L})} \tag{13}$$

for $A = A^{\downarrow K} \otimes A^{\downarrow L}$ such that $m_1^{\downarrow K\cap L}(A^{\downarrow K\cap L}) > 0$ and $m(A) = 0$ otherwise. Then

$$m^{\downarrow K}(B) = m_1(B), \tag{14}$$
$$m^{\downarrow L}(C) = m_2(C) \tag{15}$$

for any $B \in \mathbf{X}_K$ and $C \in \mathbf{X}_L$, respectively, and $(K \setminus L) \perp\!\!\!\perp (L \setminus K) | (K \cap L)$ $[m]$. Furthermore, m is the only basic assignment possessing these properties.

Proof. To prove equality (14) we have to show that for any $B \subseteq \mathbf{X}_K$

$$\sum_{A\subseteq\mathbf{X}_{K\cup L}: A^{\downarrow K}=B} m(A) = m_1(B). \tag{16}$$

Since, due to the definition of m, $m(A) = 0$ for any $A \subseteq \mathbf{X}_{K\cup L}$ for which $A \neq A^{\downarrow K} \otimes A^{\downarrow L}$, we see that

$$\begin{aligned}
&\sum_{A\subseteq\mathbf{X}_{K\cup L}: A^{\downarrow K}=B} m(A) \\
&\quad = \sum_{\substack{A\subseteq\mathbf{X}_{K\cup L}: A^{\downarrow K}=B \\ A=A^{\downarrow K}\otimes A^{\downarrow L}}} m(A) \\
&\quad = \sum_{\substack{C\subseteq\mathbf{X}_L \\ C^{\downarrow K\cap L}=B^{\downarrow K\cap L}}} m(B\otimes C).
\end{aligned}$$

To prove formula (16), we have to distinguish between two situations depending on the value of $m_2^{\downarrow K\cap L}(B^{\downarrow K\cap L})$. If this value is positive then

$$\begin{aligned}
&\sum_{A\subseteq\mathbf{X}_{K\cup L}: A^{\downarrow K}=B} m(A) \\
&\quad = \sum_{\substack{C\subseteq\mathbf{X}_L \\ C^{\downarrow K\cap L}=B^{\downarrow K\cap L}}} \frac{m_1(B)\cdot m_2(C)}{m_2^{\downarrow K\cap L}(B^{\downarrow K\cap L})} \\
&\quad = \frac{m_1(B)}{m_2^{\downarrow K\cap L}(B^{\downarrow K\cap L})} \sum_{\substack{C\subseteq\mathbf{X}_L \\ C^{\downarrow K\cap L}=B^{\downarrow K\cap L}}} m_2(C) \\
&\quad = \frac{m_1(B)}{m_2^{\downarrow K\cap L}(B^{\downarrow K\cap L})} m_2^{\downarrow K\cap L}(B^{\downarrow K\cap L}) = m_1(B).
\end{aligned}$$

If $m_2^{\downarrow K\cap L}(B^{\downarrow K\cap L}) = 0$ then, according to the definition of m, $m(A) = 0$. But $m_1^{\downarrow K\cap L}(B^{\downarrow K\cap L}) = 0$ also, due to the projectivity of m_1 and m_2, and therefore also $m_1(B) = 0$.

The proof of equality (15) is completely analogous due to the projectivity of m_1 and m_2.

Now, let us prove that $X_{K\setminus L}$ and $X_{L\setminus K}$ are conditionally independent given $X_{K\cap L}$ with respect to a basic assignment m defined via (13) for any $A \subseteq \mathbf{X}_{K\cup L}$, such that $A = A^{\downarrow K} \otimes A^{\downarrow L}$ and $m^{\downarrow K\cap L}(A^{\downarrow K\cap L}) > 0$ and $m(A) = 0$ otherwise. First let us show, that

$$\begin{aligned}
&m^{\downarrow K\cup L}(A) \cdot m^{\downarrow K\cap L}(A^{\downarrow K\cap L}) \\
&\qquad = m^{\downarrow K}(A^{\downarrow K}) \cdot m^{\downarrow L}(A^{\downarrow L}),
\end{aligned} \tag{17}$$

holds for all $A = A^{\downarrow K} \otimes A^{\downarrow L}$. If $m^{\downarrow K\cap L}(A^{\downarrow K\cap L}) > 0$, then multiplying both sides of the formula (13) by $m^{\downarrow K\cap L}(A^{\downarrow K\cap L})$ we obtain the equality (17), as (14) and (15) are satisfied and $m^{\downarrow K\cup L}(A) = m(A)$ for any $A \subseteq \mathbf{X}_{K\cup L}$. If $m^{\downarrow K\cap L}(A^{\downarrow K\cap L}) = 0$ then $m^{\downarrow L}(A^{\downarrow L}) = 0$ also, and therefore both sides of (17) equal 0. If $A \neq A^{\downarrow K} \otimes A^{\downarrow L}$, then $m(A) = 0$ by assumption.

Let $X_{K\setminus L}$ and $X_{L\setminus K}$ be conditionally independent given $X_{K\cap L}$ with respect to a basic assignment m, and $A \subseteq \mathbf{X}_{K\cup L}$ be such that $A = A^{\downarrow K} \otimes A^{\downarrow L}$ and $m^{\downarrow K\cap L}(A^{\downarrow K\cap L}) > 0$. Then (17) holds and therefore

$$m^{\downarrow K\cup L}(A) = \frac{m^{\downarrow K}(A^{\downarrow K}) \cdot m^{\downarrow L}(A^{\downarrow L})}{m^{\downarrow K\cap L}(A^{\downarrow K\cap L})},$$

i.e., (13) holds due to (14) and (15) and the fact that $m^{\downarrow K\cup L}(A) = m(A)$ for any $A \subseteq \mathbf{X}_{K\cup L}$. If $m^{\downarrow K\cap L}(A^{\downarrow K\cap L}) = 0$ then also $m^{\downarrow K}(A^{\downarrow K}) = 0, m^{\downarrow K\cap L}(A^{\downarrow L}) = 0$ and $m(A) = 0$. If $A \neq A^{\downarrow K} \otimes A^{\downarrow L}$ then $m(A) = 0$, which directly follows from Definition 2. □

Let us close this section by demonstrating application of the conditional independence notion (and Theorem 1) to Example 2.

Example 2 *(Continued)* Let us go back to the problem of finding a common extension of basic assignments m_1 and m_2 defined by (11). Theorem 1 says that for basic assignment m defined as follows

$$m(\mathbf{X}_1 \times \mathbf{X}_2 \times \{\bar{a}_3\}) = .5,$$
$$m(\{(a_1, a_2, \bar{a}_3), (\bar{a}_1, \bar{a}_2, a_3)\}) = .5,$$

variables X_1 and X_3 are conditionally independent given X_2. ♢

4 Formal Properties of Conditional Independence

Among the properties satisfied by the ternary relation $K \perp\!\!\!\perp L|M\ [m]$, the following are of principal importance:

(A1) $K \perp\!\!\!\perp L|M\ [m] \ \Rightarrow\ L \perp\!\!\!\perp K|M\ [m]$,

(A2) $K \perp\!\!\!\perp L \cup M|I\ [m] \ \Rightarrow\ K \perp\!\!\!\perp M|I\ [m]$,

(A3) $K \perp\!\!\!\perp L \cup M|I\ [m] \ \Rightarrow\ K \perp\!\!\!\perp L|M \cup I\ [m]$,

(A4) $K \perp\!\!\!\perp L|M \cup I\ [m] \ \wedge\ K \perp\!\!\!\perp M|I\ [m] \implies K \perp\!\!\!\perp L \cup M|I\ [m]$,

(A5) $K \perp\!\!\!\perp L|M \cup I\ [m] \ \wedge\ K \perp\!\!\!\perp M|L \cup I\ [m] \implies K \perp\!\!\!\perp L \cup M|I\ [m]$.

Let us recall that stochastic conditional independence satisfies the so-called *semigraphoid* properties (A1)–(A4) for any probability distribution, while axiom (A5) is satisfied only for strictly positive probability distributions. Conditional noninteractivity referred to in Section 3.2, on the other hand, satisfies axioms (A1)–(A5) for general basic assignment m, as proven in [1].

Before formulating an important theorem justifying the definition of conditional independence, let us formulate and prove an assertion concerning set extensions.

Lemma 2 *Let $K \cap L \subseteq M \subseteq L \subseteq N$. Then, for any $C \subseteq \mathbf{X}_{K\cup L}$, condition* (a) *holds if and only if both conditions* (b) *and* (c) *hold true.*

(a) $C = C^{\downarrow K} \otimes C^{\downarrow L}$;

(b) $C^{\downarrow K\cup M} = C^{\downarrow K} \otimes C^{\downarrow M}$;

(c) $C = C^{\downarrow K\cup M} \otimes C^{\downarrow L}$.

Proof. Before proving the required implications let us note that $C \subseteq C^{\downarrow K} \otimes C^{\downarrow L}$, therefore $C = C^{\downarrow K} \otimes C^{\downarrow L}$ is equivalent to

$$\forall x \in \mathbf{X}_{K\cup L}\left(x^{\downarrow K} \in C^{\downarrow K}\ \&\ x^{\downarrow L} \in C^{\downarrow L} \implies x \in C\right).$$

(a) $\implies$ (b). Consider $x \in \mathbf{X}_{K\cup M}$, such that $x^{\downarrow K} \in C^{\downarrow K}$ and $x^{\downarrow M} \in C^{\downarrow M}$. Since $x^{\downarrow M} \in C^{\downarrow M}$ there must exist (at least one) $y \in C^{\downarrow L}$, for which $y^{\downarrow M} = x^{\downarrow M}$. Now construct $z \in \mathbf{X}_{K\cup L}$ for which $z^{\downarrow K} = x^{\downarrow K}$ and $z^{\downarrow L} = y$ (it is possible because $y^{\downarrow M} = x^{\downarrow M}$). From this construction we see that $z^{\downarrow K\cup M} = x$. Therefore $z^{\downarrow K} = x^{\downarrow K} \in C^{\downarrow K}$ and $z^{\downarrow L} = y \in C^{\downarrow L}$ from which, because we assume that (a) holds, we get that $z \in C$, and therefore also $x = z^{\downarrow K\cup M} \in C^{\downarrow K\cup M}$.

(a) $\implies$ (c). Consider now $x \in \mathbf{X}_{K\cup L}$, with projections $x^{\downarrow K\cup M} \in C^{\downarrow K\cup M}$ and $x^{\downarrow L} \in C^{\downarrow L}$. From $x^{\downarrow K\cup M} \in C^{\downarrow K\cup M}$ we immediately get that $x^{\downarrow K} \in C^{\downarrow K}$, which in combination with $x^{\downarrow L} \in C^{\downarrow L}$ (due to the assumption (a)) yields that $x \in C$.

(b) & (c) $\implies$ (a). Consider $x \in \mathbf{X}_{K\cup L}$ such that $x^{\downarrow K} \in C^{\downarrow K}$ and $x^{\downarrow L} \in C^{\downarrow L}$. From the latter property one also gets $x^{\downarrow M} \in C^{\downarrow M}$, which, in combination with $x^{\downarrow K} \in C^{\downarrow K}$ gives, because (b) holds true, that $x^{\downarrow K\cup M} \in C^{\downarrow K\cup M}$. And the last property in combination with $x^{\downarrow L} \in C^{\downarrow L}$ yields the required $x \in C$. □

Since all I, K, L, M are disjoint, we will omit symbol $\cup$ and use, for example, KLM instead of $K \cup L \cup M$ in the rest of the paper.

Theorem 2 *Conditional independence satisfies (A1)–(A4).*

Proof. **ad (A1)** The validity of the implication immediately follows from the commutativity of multiplication.

ad (A2) The assumption $K \perp\!\!\!\perp LM|I\ [m]$ means that for any $A \subseteq \mathbf{X}_{KLMI}$ such that $A = A^{\downarrow KI} \otimes A^{\downarrow LMI}$ the equality

$$m^{\downarrow KLMI}(A) \cdot m^{\downarrow I}(A^{\downarrow I}) = m^{\downarrow KI}(A^{\downarrow KI}) \cdot m^{\downarrow LMI}(A^{\downarrow LMI}) \tag{18}$$

holds, and if $A \neq A^{\downarrow KI} \otimes A^{\downarrow LMI}$, then $m(A) = 0$. Let us prove first that also for any $B \subseteq \mathbf{X}_{KMI}$ such that $B = B^{\downarrow KI} \otimes B^{\downarrow MI}$, the equality

$$m^{\downarrow KMI}(B) \cdot m^{\downarrow I}(B^{\downarrow I}) = m^{\downarrow KI}(B^{\downarrow KI}) \cdot m^{\downarrow MI}(B^{\downarrow MI}) \tag{19}$$

is valid. To do so, let us compute

$$\begin{aligned} m^{\downarrow KMI}(B) \cdot m^{\downarrow I}(B^{\downarrow I}) &= \sum_{\substack{A\subseteq \mathbf{X}_{KLMI} \\ A^{\downarrow KMI} = B^{\downarrow KI} \otimes B^{\downarrow MI}}} m^{\downarrow KLMI}(A) \cdot m^{\downarrow I}(A^{\downarrow I}) \\ &= \sum_{\substack{A\subseteq \mathbf{X}_{KLMI} \\ A = A^{\downarrow KI} \otimes A^{\downarrow LMI} \\ A^{\downarrow KMI} = B^{\downarrow KI} \otimes B^{\downarrow MI}}} m^{\downarrow KLMI}(A) \cdot m^{\downarrow I}(A^{\downarrow I}) \end{aligned}$$

$$= \sum_{\substack{A\subseteq \mathbf{X}_{KLMI}\\ A=A^{\downarrow KI}\otimes A^{\downarrow LMI}\\ A^{\downarrow KMI}=B^{\downarrow KI}\otimes B^{\downarrow MI}}} m^{\downarrow KI}(A^{\downarrow KI})\cdot m^{\downarrow LMI}(A^{\downarrow LMI})$$
$$= m^{\downarrow KI}(A^{\downarrow KI})\cdot \sum_{\substack{C\subseteq \mathbf{X}_{LMI}\\ C^{\downarrow MI}=B^{\downarrow MI}}} m^{\downarrow LMI}(C)$$
$$= m^{\downarrow KI}(B^{\downarrow KI})\cdot m^{\downarrow MI}(B^{\downarrow MI}),$$

as

$$\begin{aligned} m^{\downarrow I}(B^{\downarrow I}) &= m^{\downarrow I}(A^{\downarrow I}),\\ m^{\downarrow KI}(B^{\downarrow KI}) &= m^{\downarrow KI}(A^{\downarrow KI}). \end{aligned}$$

So, to finish this step we still must prove that if $B \neq B^{\downarrow KI}\otimes B^{\downarrow MI}$ then $m^{\downarrow KMI}(B)=0$. Also, in this case

$$m^{\downarrow KMI}(B) = \sum_{\substack{A\subseteq \mathbf{X}_{KLMI}\\ A^{\downarrow KMI}=B}} m^{\downarrow KLMI}(A),$$

but since $B = A^{\downarrow KMI} \neq A^{\downarrow KI}\otimes A^{\downarrow MI}$ then, because of Lemma 2, also $A \neq A^{\downarrow KI}\otimes A^{\downarrow LMI}$ for any A such that $A^{\downarrow KMI}=B$. But for these A's, $m^{\downarrow KLMI}(A)=0$ and therefore also $m^{\downarrow KMI}(B)=0$.

ad (A3) Again, let us suppose validity of $K \perp\!\!\!\perp LM|I$ $[m]$, i.e., for any $A\subseteq \mathbf{X}_{KLMI}$ such that $A = A^{\downarrow KI}\otimes A^{\downarrow LMI}$ equality (18) holds, and $m^{\downarrow KLMI}(A) = 0$ otherwise. Our aim is to prove that for any $C\subseteq \mathbf{X}_{KLMI}$ such that $C = C^{\downarrow KMI}\otimes C^{\downarrow LMI}$, the equality

$$\begin{aligned} &m^{\downarrow KLMI}(C)\cdot m^{\downarrow MI}(C^{\downarrow MI}) \\ &\quad = m^{\downarrow KMI}(C^{\downarrow KMI})\cdot m^{\downarrow LMI}(C^{\downarrow LMI}) \end{aligned} \tag{20}$$

is satisfied as well, and $m^{\downarrow KLMI}(C)=0$ otherwise. Let C be such that $m^{\downarrow I}(C^{\downarrow I})>0$. Since we assume that $K \perp\!\!\!\perp LM|I$ $[m]$ holds, we have for such a C

$$\begin{aligned} &m^{\downarrow KLMI}(C)\cdot m^{\downarrow I}(C^{\downarrow I}) \\ &\quad = m^{\downarrow KI}(C^{\downarrow KI})\cdot m^{\downarrow LMI}(C^{\downarrow LMI}), \end{aligned}$$

and therefore we can compute

$$\begin{aligned} &m^{\downarrow KLMI}(C)\cdot m^{\downarrow MI}(C^{\downarrow MI})\\ &\quad = m^{\downarrow KLMI}(C)\cdot m^{\downarrow I}(C^{\downarrow I})\cdot \frac{m^{\downarrow MI}(C^{\downarrow MI})}{m^{\downarrow I}(C^{\downarrow I})}\\ &\quad = m^{\downarrow KI}(C^{\downarrow KI})\cdot m^{\downarrow LMI}(C^{\downarrow LMI})\cdot \frac{m^{\downarrow MI}(C^{\downarrow MI})}{m^{\downarrow I}(C^{\downarrow I})}\\ &\quad = \frac{m^{\downarrow KI}(C^{\downarrow KI})\cdot m^{\downarrow MI}(C^{\downarrow MI})}{m^{\downarrow I}(C^{\downarrow I})}\cdot m^{\downarrow LMI}(C^{\downarrow LMI})\\ &\quad = m^{\downarrow KMI}(C^{\downarrow KMI})\cdot m^{\downarrow LMI}(C^{\downarrow LMI}), \end{aligned}$$

where the last equality is satisfied due to (A2) and the fact that $m^{\downarrow I}(C^{\downarrow I})>0$. If $m^{\downarrow I}(C^{\downarrow I})=0$ then also $m^{\downarrow KMI}(C^{\downarrow KMI})=0$, $m^{\downarrow LMI}(C^{\downarrow LMI})=0$ and $m^{\downarrow KLMI}(C)=0$ and therefore (20) also holds true.

It remains to be proven that $m(C)=0$ for all $C \neq C^{\downarrow KMI}\otimes C^{\downarrow LMI}$. But in this case, as a consequence of Lemma 2, also $C \neq C^{\downarrow KI}\otimes C^{\downarrow LMI}$, and therefore $m(C)=0$ due to the assumption.

ad (A4) First, supposing $K \perp\!\!\!\perp L|MI$ $[m]$ and $K \perp\!\!\!\perp M|I$ $[m]$, let us prove that for any $A\subseteq \mathbf{X}_{KLMI}$ such that $A = A^{\downarrow KI}\otimes A^{\downarrow LMI}$, the equality (18) holds. Since from $A = A^{\downarrow KI}\otimes A^{\downarrow LMI}$ it also follows due to Lemma 2 that $A = A^{\downarrow KMI}\otimes A^{\downarrow LMI}$, and therefore (since we assume $K \perp\!\!\!\perp L|MI$ $[m]$)

$$\begin{aligned} &m^{\downarrow KLMI}(A)\cdot m^{\downarrow MI}(A^{\downarrow MI}) \\ &\quad = m^{\downarrow KMI}(A^{\downarrow KMI})\cdot m^{\downarrow LMI}(A^{\downarrow LMI}). \end{aligned} \tag{21}$$

Now, let us further assume that $m^{\downarrow MI}(A^{\downarrow MI})>0$ (and thus also $m^{\downarrow I}(A^{\downarrow I})>0$). Since from $A = A^{\downarrow KI}\otimes A^{\downarrow LMI}$ Lemma 2 implies $A^{\downarrow KMI} = A^{\downarrow KI}\otimes A^{\downarrow MI}$, one gets from $K \perp\!\!\!\perp M|I$ $[m]$ that

$$\begin{aligned} &m^{\downarrow KMI}(A^{\downarrow KMI})\cdot m^{\downarrow I}(A^{\downarrow I})\\ &\quad = m^{\downarrow KI}(A^{\downarrow KI})\cdot m^{\downarrow MI}(A^{\downarrow MI}), \end{aligned}$$

which, in combination with equality (22), yields

$$\begin{aligned} &m^{\downarrow KLMI}(A)\cdot m^{\downarrow MI}(A^{\downarrow MI})\\ &\quad = \frac{m^{\downarrow KI}(A^{\downarrow KI})\cdot m^{\downarrow MI}(A^{\downarrow MI})}{m^{\downarrow I}(A^{\downarrow I})}\cdot m^{\downarrow LMI}(A^{\downarrow LMI}), \end{aligned}$$

which is (for positive $m^{\downarrow MI}(A^{\downarrow MI})$) evidently equivalent to (18). If, on the other hand, $m^{\downarrow MI}(A^{\downarrow MI})=0$, then also $m^{\downarrow LMI}(A^{\downarrow LMI})=0$ and $m^{\downarrow KLMI}(A)=0$ and both sides of (18) equal 0.

It remains to prove that $m^{\downarrow KLMI}(A)=0$ for all $A \neq A^{\downarrow KI}\otimes A^{\downarrow LMI}$. But $m^{\downarrow KLMI}(A)=0$ because Lemma 2 says that either $A \neq A^{\downarrow KMI}\otimes A^{\downarrow LMI}$ (and therefore $m^{\downarrow KLMI}(A)=0$ from the assumption that $K \perp\!\!\!\perp L|MI$ $[m]$) or $A^{\downarrow KMI} \neq A^{\downarrow KI}\otimes A^{\downarrow MI}$ (and then $m^{\downarrow KMI}(A^{\downarrow KMI})=0$ due to the assumption $K \perp\!\!\!\perp M|I$ $[m]$), and therefore also $m^{\downarrow KLMI}(A)=0$). □

Analogous to a probabilistic case, conditional independence $K \perp\!\!\!\perp L|MI$ $[m]$ does not generally satisfy (A5), as can be seen from the following simple example.

Example 3 Let X_1, X_2 and X_3 be three variables with values in $\mathbf{X}_1, \mathbf{X}_2$ and $\mathbf{X}_3$ respectively, $\mathbf{X}_i = \{a_i, \bar{a}_i\}, i = 1,2,3$, and their joint basic assignment is defined as follows:

$$\begin{aligned} m(\{(x_1,x_2,x_3\}) &= \tfrac{1}{16},\\ m(\mathbf{X}_1\times\mathbf{X}_2\times\mathbf{X}_3) &= \tfrac{1}{2}, \end{aligned}$$

for $x_i = a_i, \bar{a}_i$, values of m on the remaining sets being 0. Its marginal basic assignments on $\mathbf{X}_1\times\mathbf{X}_2, \mathbf{X}_1\times$

$\mathbf{X}_3, \mathbf{X}_2 \times \mathbf{X}_3$ and $\mathbf{X}_i, i = 1, 2, 3$ are

$$m^{\downarrow 12}(\{x_1, x_2\}) = \tfrac{1}{8},$$
$$m^{\downarrow 12}(\mathbf{X}_1 \times \mathbf{X}_2) = \tfrac{1}{2},$$
$$m^{\downarrow 13}(\{x_1, x_3\}) = \tfrac{1}{8},$$
$$m^{\downarrow 13}(\mathbf{X}_1 \times \mathbf{X}_3) = \tfrac{1}{2},$$
$$m^{\downarrow 23}(\{x_2, x_3\}) = \tfrac{1}{8},$$
$$m^{\downarrow 23}(\mathbf{X}_2 \times \mathbf{X}_3) = \tfrac{1}{2},$$

and

$$m^{\downarrow i}(x_i) = \tfrac{1}{4},$$
$$m^{\downarrow i}(\mathbf{X}_i) = \tfrac{1}{2},$$

respectively. It is easy (but somewhat time-consuming) to check that

$$m(A^{\downarrow 13} \otimes A^{\downarrow 23}) \cdot m^{\downarrow 3}(A^{\downarrow 3}) = m^{\downarrow 13}(A^{\downarrow 13}) \cdot m^{\downarrow 23}(A^{\downarrow 23})$$

and

$$m(A^{\downarrow 12} \otimes A^{\downarrow 23}) \cdot m^{\downarrow 2}(A^{\downarrow 2}) = m^{\downarrow 12}(A^{\downarrow 12}) \cdot m^{\downarrow 23}(A^{\downarrow 23}),$$

the values of remaining sets being zero, while e.g.

$$m(\{(a_1, \bar{a}_2, \bar{a}_3)\}) = \frac{1}{16} \neq \frac{1}{4} \cdot \frac{1}{8} = m^{\downarrow 1}(\{a_1\}) \cdot m^{\downarrow 23}(\{(\bar{a}_2, \bar{a}_3)\}),$$

i.e., $\{1\} \perp\!\!\!\perp \{2\}|\{3\}$ $[m]$ and $\{1\} \perp\!\!\!\perp \{3\}|\{2\}$ $[m]$ hold, but $\{1\} \perp\!\!\!\perp \{2,3\}|\emptyset$ $[m]$ does not. ♢

This fact perfectly corresponds to the properties of stochastic conditional independence. In probability theory (A5) need not be satisfied if the joint probability distribution is not strictly positive. But the counterpart of strict positivity of probability distribution for basic assignments is not straightforward. It is evident that it does not mean strict positivity on all subsets of the frame of discernment in question — in this case variables are not (conditionally) independent (cf. Definitions 1 and 2). On the other hand, it can be seen from Example 3 that strict positivity on singletons is not sufficient (and, surprisingly, as we shall see later, also not necessary). At present we are able to formulate Theorem 3. To prove it, we need the following lemma.

Lemma 3 *Let K, L, M be disjoint subsets of N, $K, L \neq \emptyset$ and m be a joint basic assignment on $\mathbf{X}_N$. Then the following statements are equivalent:*

(i) $K \perp\!\!\!\perp L|M$ $[m]$.

(ii) *The basic assignment $m^{\downarrow KLM}$ on $\mathbf{X}_{KLM}$ has for $A = A^{\downarrow KM} \otimes A^{\downarrow LM}$ the form*

$$m^{\downarrow KLM}(A) = f_1(A^{\downarrow KM}) \cdot f_2(A^{\downarrow LM}), \qquad (22)$$

where f_1 and f_2 are set functions on $\mathbf{X}_{KM}$ and $\mathbf{X}_{LM}$, respectively, and $m(A) = 0$ otherwise.

Proof. Let (i) be satisfied. Then for any $A = A^{\downarrow KM} \otimes A^{\downarrow LM}$ we have

$$m^{\downarrow KLM}(A) \cdot m^{\downarrow M}(A^{\downarrow M}) = m^{\downarrow KM}(A^{\downarrow KM}) \cdot m^{\downarrow LM}(A^{\downarrow LM}).$$

If $m^{\downarrow M}(A^{\downarrow M}) > 0$, we may divide both sides of the above equality by it and we obtain

$$m^{\downarrow KLM}(A) = \frac{m^{\downarrow KM}(A^{\downarrow KM}) \cdot m^{\downarrow LM}(A^{\downarrow LM})}{m^{\downarrow M}(A^{\downarrow M})}.$$

Therefore (ii) is obviously fulfilled, e.g. for

$$f_1(A^{\downarrow KM}) = m^{\downarrow K \cup M}(A^{\downarrow K \cup M})$$

and

$$f_2(A^{\downarrow LM}) = \frac{m^{\downarrow L \cup M}(A^{\downarrow L \cup M})}{m^{\downarrow M}(A^{\downarrow M})}.$$

If, on the other hand, $m^{\downarrow M}(A^{\downarrow M}) = 0$, then also $m^{\downarrow KM}(A^{\downarrow KM}) = 0$, $m^{\downarrow LM}(A^{\downarrow LM}) = 0$ and $m^{\downarrow KLM}(A^{\downarrow KLM}) = 0$, and therefore (22) trivially holds. To finish the proof of this implication we must prove that $m(A) = 0$ if $A \neq A^{\downarrow KM} \otimes A^{\downarrow LM}$, but it follows directly from the definition.

Let (ii) be satisfied. Then denoting

$$f_1^{\downarrow M}(A^{\downarrow M}) = \sum_{\substack{C \subseteq X_{KM} \\ C^{\downarrow M} = A^{\downarrow M}}} f_1(C)$$

and

$$f_2^{\downarrow M}(A^{\downarrow M}) = \sum_{\substack{C \subseteq X_{LM} \\ C^{\downarrow M} = A^{\downarrow M}}} f_2(C),$$

we have

$$\begin{aligned} m^{\downarrow KM}(A^{\downarrow KM}) &= \sum_{\substack{C \subseteq X_{KLM} \\ C^{\downarrow KM} = A^{\downarrow KM}}} m^{\downarrow KLM}(C) \\ &= \sum_{\substack{C \subseteq X_{KLM} \\ C^{\downarrow KM} = A^{\downarrow KM}}} f_1(C^{\downarrow KM}) \cdot f_2(C^{\downarrow LM}) \\ &= f_1(A^{\downarrow KM}) \cdot \sum_{\substack{D \subseteq X_{LM} \\ D^{\downarrow M} = A^{\downarrow M}}} f_2(D) \\ &= f_1(A^{\downarrow KM}) \cdot f_2^{\downarrow M}(A^{\downarrow M}) \end{aligned}$$

and similarly

$$m^{\downarrow LM}(A^{\downarrow LM}) \quad = f_2(A^{\downarrow LM}) \cdot f_1^{\downarrow M}(A^{\downarrow M}).$$

Therefore

$$\begin{aligned} & m^{\downarrow M}(A^{\downarrow M}) \\ & \quad = \sum_{\substack{C \subseteq X_{KLM} \\ C^{\downarrow M} = A^{\downarrow M}}} m^{\downarrow KLM}(C) = \sum_{\substack{D \subseteq X_{KM} \\ D^{\downarrow M} = A^{\downarrow M}}} m^{\downarrow KM}(D) \\ & \quad = \sum_{\substack{D \subseteq X_{KM} \\ D^{\downarrow M} = A^{\downarrow M}}} f_1(D^{\downarrow KM}) \cdot f_2^{\downarrow M}(D^{\downarrow M}) \\ & \quad = f_2^{\downarrow M}(A^{\downarrow M}) \cdot \sum_{\substack{D \subseteq X_{KM} \\ D^{\downarrow M} = A^{\downarrow M}}} f_1(D^{\downarrow KM}) \\ & \quad = f_2^{\downarrow M}(A^{\downarrow M}) \cdot f_1^{\downarrow M}(A^{\downarrow M}). \end{aligned}$$

Hence, multiplying both sides of (22) by $m^{\downarrow M}(A^{\downarrow M})$ one has

$$\begin{aligned} & m(A) \cdot m^{\downarrow M}(A^{\downarrow M}) \\ & \quad = f_1(A^{\downarrow KM}) \cdot f_2(A^{\downarrow LM}) \cdot f_1^{\downarrow M}(A^{\downarrow M}) \cdot f_2^{\downarrow M}(A^{\downarrow M}) \\ & \quad = f_1(A^{\downarrow KM}) \cdot f_2^{\downarrow M}(A^{\downarrow M}) \cdot f_2(A^{\downarrow LM}) \cdot f_1^{\downarrow M}(A^{\downarrow M}) \\ & \quad = m^{\downarrow KM}(A^{\downarrow KM}) \cdot m^{\downarrow LM}(A^{\downarrow LM}), \end{aligned}$$

i.e., (i) holds (as $m(A) = 0$ if $A \neq A^{\downarrow KM} \otimes A^{\downarrow LM}$ by assumption). □

Theorem 3 *Let m be a basic assignment on $\mathbf{X}_N$ such that $m(A) > 0$ if and only if $A = \bigtimes_{i \in N} A_i$, where A_i is a focal element on $\mathbf{X}_i$. Then (A5) is satisfied.*

Proof. Let $K \perp\!\!\!\perp L | MI$ $[m]$ and $K \perp\!\!\!\perp M | LI$ $[m]$. Then by Lemma 3 there exist functions f_1, f_2, g_1 and g_2 such that

$$\begin{aligned} m^{\downarrow KLMI}(A) &= f_1(A^{\downarrow KMI}) \cdot f_2(A^{\downarrow LMI}) \\ m^{\downarrow KLMI}(A) &= g_1(A^{\downarrow KLI}) \cdot g_2(A^{\downarrow LMI}) \end{aligned}$$

for any $A = A^{\downarrow KMI} \otimes A^{\downarrow LMI}$ and any $A = A^{\downarrow KLI} \otimes A^{\downarrow LMI}$, respectively.

If $m(A) > 0$ we can write

$$f_1(A^{\downarrow KMI}) = \frac{g_1(A^{\downarrow KLI}) \cdot g_2(A^{\downarrow LMI})}{f_2(A^{\downarrow LMI})}. \tag{23}$$

Let us note, that if $m(A) > 0$, then by assumption $A = \bigtimes_{i \in N} A_i$ and therefore it can be written as $A = A^{\downarrow K} \times A^{\downarrow L} \times A^{\downarrow M} \times A^{\downarrow I}$. Hence (23) may be rewritten into the form

$$\begin{aligned} & f_1(A^{\downarrow K} \times A^{\downarrow M} \times A^{\downarrow I}) \qquad (24) \\ & \quad = \frac{g_1(A^{\downarrow K} \times A^{\downarrow L} \times A^{\downarrow I}) \cdot g_2(A^{\downarrow L} \times A^{\downarrow M} \times A^{\downarrow I})}{f_2(A^{\downarrow L} \times A^{\downarrow M} \times A^{\downarrow I})}. \end{aligned}$$

Let us choose $B \subseteq \mathbf{X}_L$ such that $B = A^{\downarrow L}$. Then (24) can be written in the form

$$f_1(A^{\downarrow K} \times A^{\downarrow M} \times A^{\downarrow I}) = h_1(A^{\downarrow KI}) \cdot h_2(A^{\downarrow MI}),$$

where

$$\begin{aligned} h_1(A^{\downarrow KI}) &= g_1(A^{\downarrow K} \times B \times A^{\downarrow I}), \\ h_2(A^{\downarrow MI}) &= \frac{g_2(B \times A^{\downarrow M} \times A^{\downarrow I})}{f_2(B \times A^{\downarrow M} \times A^{\downarrow I})}. \end{aligned}$$

Therefore

$$\begin{aligned} m^{\downarrow KLMI}(A) &= h_1(A^{\downarrow KI}) \cdot h_2(A^{\downarrow MI}) \cdot f_2(A^{\downarrow LMI}) \\ &= h_1(A^{\downarrow KI}) \cdot h_2'(A^{\downarrow LMI}). \qquad (25) \end{aligned}$$

Now, we shall prove that (25) is valid also for $A = A^{\downarrow KI} \otimes A^{\downarrow LMI}$ such that $m(A) = 0$. The validity of

$$\begin{aligned} & m^{\downarrow KLMI}(A) \cdot m^{\downarrow M}(A^{\downarrow MI}) \\ & \quad = m^{\downarrow KMI}(A^{\downarrow KMI}) \cdot m^{\downarrow LMI}(A^{\downarrow LMI}) \end{aligned}$$

for $A = A^{\downarrow KMI} \otimes A^{\downarrow LMI}$ implies that at least one of $m^{\downarrow LMI}(A^{\downarrow LMI})$ and $m^{\downarrow KMI}(A^{\downarrow KMI})$ must also equal zero. In the first case, (25) holds for $h_2'(A^{\downarrow LMI}) = m^{\downarrow LMI}(A^{\downarrow LMI})$ and h_1 arbitrary.

If, on the other hand, $m^{\downarrow LMI}(A^{\downarrow LMI}) > 0$, then $m^{\downarrow KMI}(A^{\downarrow KMI})$ must equal zero. We also must prove that in this case $m^{\downarrow KI}(A^{\downarrow KI}) = 0$, from which (25) immediately follows. To prove it, let us suppose the contrary. Since $A^{\downarrow KMI} = \bigtimes_{i \in KMI} A_i$, there must exist at least one $j \in M$ such that A_j is not a focal element on $\mathbf{X}_j$. From this fact it follows that also $m^{\downarrow LMI}(A^{\downarrow LMI}) = 0$, as $m^{\downarrow j}(A_j)$ is marginal to $m^{\downarrow LMI}(A^{\downarrow LMI})$, and it contradicts the assumption that $m^{\downarrow LMI}(A^{\downarrow LMI}) > 0$.

It remains to be proven that $m(A) = 0$ if $A \neq A^{\downarrow KI} \otimes A^{\downarrow LMI}$. But it follows directly from the assumption, as $m(A) > 0$ only for $A = \bigtimes_{i \in N} A_i$. □

Example 3 suggests that the assumption of positivity of $m(A)$ on any $A = \bigtimes_{i \in N} A_i$, where A_i is a focal element on $\mathbf{X}_i$, is substantial. On the other hand, the assumption that $m(A) = 0$ otherwise may not be so substantial and (A5) may hold for more general bodies of evidence than those characterised by the assumption of Theorem 3 (at present we are not able to find a counterexample).

Let us note that, for Bayesian basic assignments, assumption of Theorem 3 seems to be more general than that of strict positivity of the probability distribution. But the generalisation is of no practical consequence — if probability of a marginal value is equal to zero, than this value may be omitted.

5 Summary and Conclusions

This paper started with a brief discussion, based on recently published results, why random sets independence is the most appropriate independence concept (from the viewpoint of multidimensional models) in evidence theory. Then we compared two generalisations of random sets independence — conditional noninteractivity and the new concept of conditional independence. We showed that, although from the viewpoint of formal properties satisfied by these concepts, conditional noninteractivity seems to be slightly better than conditional independence, from the viewpoint of multidimensional models the latter is superior to the former, as it is consistent with marginalisation.

There is still a problem to be solved, namely: can the sufficient condition be weakened while keeping the validity of (A5)?

Acknowledgements

Research presented in this paper is supported by GA ČR under grant 201/09/1891, GA AV ČR under grant A100750603 and MŠMT under grant 2C06019.

References

[1] B. Ben Yaghlane, Ph. Smets and K. Mellouli, Belief functions independence: II. the conditional case. *Int. J. Approx. Reasoning*, **31** (2002), 31–75.

[2] I. Couso, S. Moral and P. Walley, Examples of independence for imprecise probabilities, *Proceedings of ISIPTA'99,* eds. G. de Cooman, F. G. Cozman, S. Moral, P. Walley, 121–130.

[3] I. Couso, Independence concepts in evidence theory, *Proceedings of ISIPTA'07,* eds. G. de Cooman, J. Vejnarová, M. Zaffalon, 125–134.

[4] A. Dempster, Upper and lower probabilities induced by multivalued mappings. *Ann. Math. Statist.* **38** (1967), pp. 325–339.

[5] R. Jiroušek, J. Vejnarová, Compositional models and conditional independence in Evidence Theory, submitted to *International Journal of Approximate Reasoning.*

[6] G. J. Klir, *Uncertainty and Information. Foundations of Generalized Information Theory.* Wiley, Hoboken, 2006.

[7] S. L. Lauritzen, *Graphical Models.* Oxford University Press, 1996.

[8] S. Moral, A.Cano, Strong conditional independence for credal sets, *Ann. of Math. and Artif. Intell.*, **35** (2002), 295–321.

[9] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, New Jersey, 1976.

[10] P. P. Shenoy, Conditional independence in valuation-based systems. *Int. J. Approx. reasoning,* **10** (1994), 203-234.

[11] M. Studený, Formal properties of conditional independence in different calculi of artificial intelligence. *Proceedings of European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty ECSQARU93*, eds. K. Clarke, R. Kruse, S. Moral, Springer-Verlag, 1993, pp. 341-348.

[12] J. Vejnarová, Conditional independence relations in possibility theory. *Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems* **8** (2000), pp. 253–269.

[13] J. Vejnarová, Markov properties and factorization of possibility distributions. *Annals of Mathematics and Artificial Intelligence,* **35** (2002), pp. 357–377.

[14] J. Vejnarová, Conditional independence in evidence theory, *Proceedings of 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems IPMU'08,* eds. L. Magdalena, M. Ojeda-Aciego, J. L. Verdegay, pp. 938–945.

[15] J. Vejnarová, On two notions of independence in evidence theory, *Proceedings of 11th Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty*, eds. T. Itoh, A. Shirouha, Sendai University, 2008, pp. 69–74.

Bayes Linear Analysis of Imprecision in Computer Models, with Application to Understanding Galaxy Formation

Ian Vernon
Department of Mathematical Sciences
Durham University
I.R.Vernon@durham.ac.uk

Michael Goldstein
Department of Mathematical Sciences
Durham University
Michael.Goldstein@durham.ac.uk

Abstract

Imprecision arises naturally in the context of computer models and their relation to reality. An imprecise treatment of general computer models is presented, illustrated with an analysis of a complex galaxy formation simulation known as Galform. The analysis involves several different types of uncertainty, one of which (the Model Discrepancy) comes directly from expert elicitation regarding the deficiencies of the model. The Model Discrepancy is therefore treated within an Imprecise framework to reflect more accurately the beliefs of the expert concerning the discrepancy between the model and reality. Due to the conceptual complexity and computationally intensive nature of such a Bayesian imprecise uncertainty analysis, Bayes Linear Methodology is employed which requires consideration of only expectations and variances of all uncertain quantities. Therefore incorporating an Imprecise treatment within a Bayes Linear analysis is shown to be relatively straightforward. The impact of an imprecise assessment on the input space of the model is determined through the use of an Implausibility measure.

Keywords. Bayesian Inference, Computer models, Calibration, Imprecise model discrepancy, Implausibility, Galaxy Formation, Graphical Representation of Model Imprecision.

1 Introduction

Computer models make imprecise statements about physical systems. This arises because of compromises made in the physical theory and in approximations to solutions of very complex systems of equations. Therefore any statement about a physical system, for example climate change, which is derived from the analysis of computer models will be necessarily imperfect, as it will usually be very difficult to put a precise quantification on the discrepancy between the model analysis and the physical system [1]. A full probabilistic representation of the imprecision arising from such model discrepancy will typically be very complex and difficult to analyse. However, there is an alternative way to express such imprecision, based on viewing expectations rather than probability as the natural primitive for expressing uncertainty statements. This formulation allows us to focus directly on 'high level' summary expressions of imprecision. This approach is termed Bayes Linear Analysis; for a detailed treatment see [2].

In this paper we show how the Bayes Linear approach may be used to capture the most important features of the imprecision arising from the use of complex physical models. We illustrate our approach with the galaxy formation model known as Galform. Galform simulates the formation and evolution of approximately 1 million galaxies from the beginning of the Universe until the current day (a period of approximately 13 billion years). It gives outputs representing various physical features of each of the galaxies which can be compared with observational data [3].

This paper is structured as follows: in section 2 we discuss the Galform model in more detail, in section 3 the theory of computer models and the incorporation of the imprecise model discrepancy is described, and in section 4 we develop appropriate graphical displays for such imprecise analyses and demonstrate the application of these methods to the Galform model.

2 Cosmology and Galaxy Formation

2.1 Understanding the Universe

Over the last 100 years, major advances have been made in understanding the large scale structure of the Universe. Current theories of cosmology suggest that the Universe began in a hot, dense state approximately 13 billion years ago, and that it has been expanding rapidly ever since. However, there exists a major problem: observations of galaxies imply that

there must exist far more matter in the Universe than the visible matter that makes up stars, planets and us. This is referred to as 'Dark Matter' and understanding its nature and how it has affected the evolution of galaxies within our Universe is one of the most important problems in modern cosmology.

In order to study many of the effects of Dark Matter, cosmologists try to model Galaxy formation using complex computer models. In this paper, we develop the Bayesian treatment of imprecision for computer models, and illustrate our analysis using one such model, known as Galform (developed by the Galform group at the Institute for Computational Cosmology, Durham University).

2.2 Galform: a Galaxy Formation Simulation

Simulating the formation of large numbers of galaxies from the beginning of the Universe until the current day is a difficult task and so the process is split into two parts. First a Dark Matter simulation is performed to determine the behaviour of fluctuations of mass in the early Universe, and their subsequent growth into millions of galaxy sized lumps in the following 13 billion years. Second, the results of the Dark Matter simulation are used by a more detailed model called Galform which models the far more complicated interactions of normal matter including: gas cloud formation, radiative cooling, star formation and the effects of central black holes.

The first simulation is run on a volume of space of size $(1.63 \text{ billion light-years})^3$. This volume is split into 512 sub-volumes which are independently simulated using the second model Galform, which is the subject of the Imprecise Uncertainty Analysis in this paper (see figure 1). Each run of Galform takes 20-30 minutes per subvolume per processor.

2.3 Galform Inputs and Outputs

The Galform simulation provides many outputs related to approximately 1 million simulated galaxies. We consider the two most important types of output: the bj and K band luminosity functions. The bj band luminosity function gives the number of blue (i.e. young) galaxies of a certain luminosity per unit volume, while the K band luminosity function describes the number of red (i.e. old) galaxies (see Figure 1). The colour of a galaxy comes from the stars it contains, stars which on average burn bluer early in their lifecycle and redder as they age. These outputs can be compared to observational data gathered by the 2dFGRS galaxy survey (see [3] and references therein).

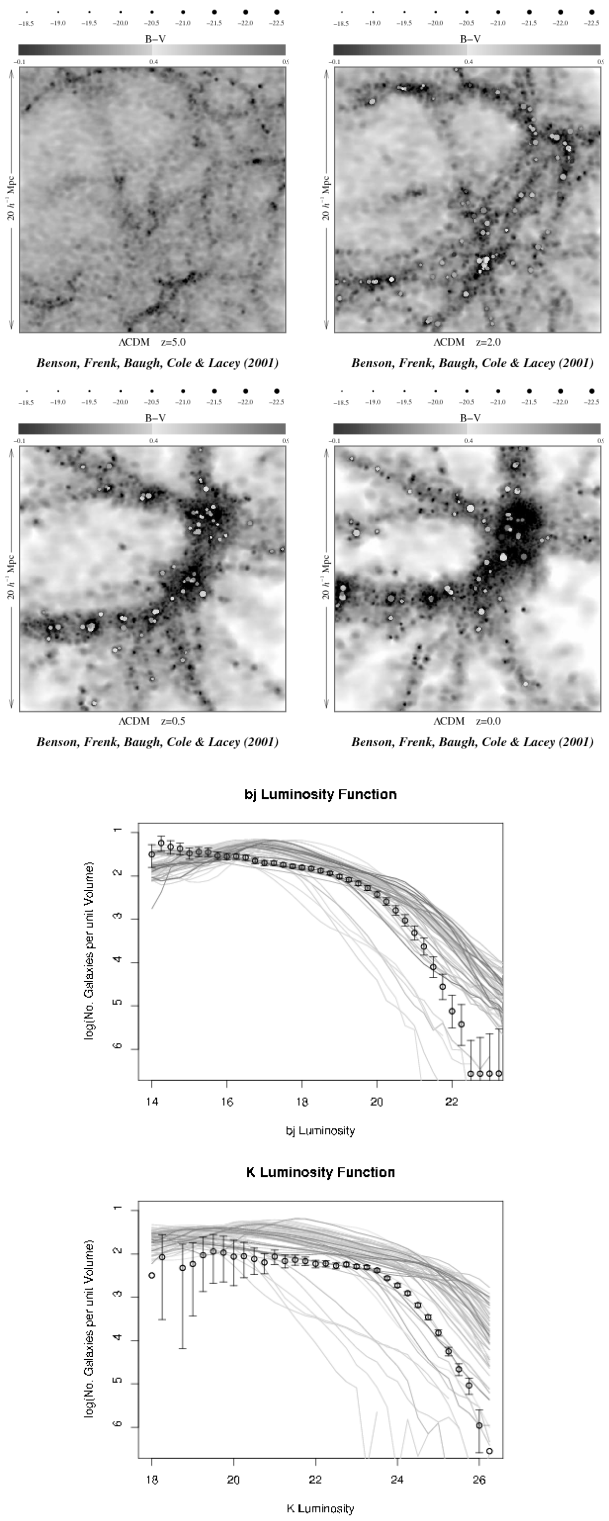


Figure 1: Top 4 panels: the evolution of both the Dark Matter Simulation and Galform over a 13 billion year period. Darker areas show higher concentrations of Dark Matter, leading to the formation of bright galaxies (the white dots). Bottom 2 panels: the bj and K luminosity functions. The grey lines are from 60 runs of the Galform simulation. The black points are observed data from the 2dFGRS survey with associated measurement errors.

Galform has 17 input parameters that the cosmologists were interested in varying. Due to expert judgements regarding the impact of these inputs on the luminosity functions we attempted to calibrate Galform over only 8 of the input parameters (while taking into account the possible effects of the remaining 9). These input parameters and their initial ranges are:

vhotdisk:	100 - 550
aReheat:	0.2 - 1.2
alphacool:	0.2 - 1.2
vhotburst:	100 - 550
epsilonStar:	0.001 - 0.1
stabledisk:	0.65 - 0.95
alphahot:	2 - 3.7
yield:	0.02 - 0.05

The other 9 parameters are: VCUT, ZCUT, alphastar, tau0mrg, fellip, fburst, FSMBH, epsilonSMBHEddington and tdisk.

2.4 Galaxy Formation: Main Issues

The main physical questions that the cosmologists are interested in are: do we understand how galaxies form, and could the galaxies we observe have been formed in the presence of large amounts of dark matter? In order to answer these questions it is vital to correctly analyse all relevant sources of uncertainty within this situation. Many of the sources of uncertainty derive from aspects of the problem for which we have a good physical understanding, for example, the various types of measurement error associated with the observational data (which mainly come from optical deficiencies of telescopes).

However, by far the most important uncertainties arise from the fact that we are uncertain about the discrepancy between the Galform model and the real system, and we are also uncertain about which choice of input should be made when running the model.

3 Bayes Linear Analysis for Computer Simulators

To understand and describe all the sources of uncertainty in the Galform simulator we apply computer model emulation techniques. Although here we will only discuss the Galform simulator, these techniques are very general and can be applied to any complex model of a physical system. Indeed they have been successfully applied to a wide variety of physical models (see [5] for a Bayes Linear approach, [4] for a fully Bayesian approach, and for an overview of computer experiments in general see [6] or the Managing Uncertainty in Complex Models website http://mucm.group.shef.ac.uk/index.html).

3.1 Main Objectives

A common aim of computer experiment analysis is to use observed data to reduce uncertainty about possible choices of the input parameters x (see [5] and [4]). In many problems the major interest lies in whether there is any choice of x that would lead to an acceptable match between model outputs and observed data. The larger the assessed discrepancy between model and system, the weaker the constraints the observations will impose on this choice. In this work we treat this discrepancy as imprecise. Therefore one of the most important aspects of the analysis of the model lies in identifying and quantifying the impact of such imprecision on the choice of possible input values.

3.2 Computer simulators

The simulator (Galform) is represented as a function, which maps the input parameters x to the outputs $f(x)$. We use the "Best Input Approach", where we assume there exists a value x^* independent of the function f such that the value $f^* = f(x^*)$ summarises all the information the simulator conveys about the system. In order to make meaningful statements about the system, denoted y, in relation to the model, we link the simulator to the system using the *model discrepancy* denoted ϵ_{md} via the equation:

$$y = f^* + \epsilon_{md}, \tag{1}$$

and assume that ϵ_{md} is independent of f and x^*, that is, independent in terms of our own beliefs.

The Model Discrepancy term ϵ_{md} links the real system y to the best evaluation of the model represented by f^*. This is distinct from other sources of uncertainty in our analysis and comes directly from expert opinion regarding the 'accuracy' of the model. Understanding the nature of ϵ_{md} is a non-trivial task as there are various other sources of uncertainty that are present that interfere with any assessment of ϵ_{md}. For example, we can never measure the real system y directly. Instead we have measurements z observed with experimental error ϵ_{obs} which are linked to the system by:

$$z = y + \epsilon_{obs}. \tag{2}$$

Another important source of uncertainty is due to lack of knowledge about the form of the function $f(x)$. As the model takes a significant time to run and has a high dimensional input space we only have limited knowledge about its behavior. Further, there is uncertainty regarding the best input value of x^* that features in the definition of ϵ_{md} (equation (1)).

These other types of uncertainty make understanding ϵ_{md} difficult, which is a significant problem as often ϵ_{md} is the most important source of uncertainty due to its size and nature. Due to these difficulties, the expert will often be imprecise over the assessment of the model discrepancy, and even more imprecision could occur when we consider the opinions of a group of experts. It is therefore reasonable to analyse ϵ_{md} within an imprecise framework, while treating other less significant (and more understood) sources of uncertainty as precise.

We need to understand the behavior of the Galform simulation $f(x)$: this is done by representing our beliefs about $f(x)$ as a statistical function known as an Emulator, described in the next section. We address the calibration problem (that of finding inputs x that give rise to good matches between the outputs of $f(x)$ and the observed data z) by use of a technique known as History Matching [5]. This involves discarding regions of the input parameter space that we are reasonably sure will give bad fits to the observed data, and we do this using an Implausibility measure. Analysing the effect on this measure of having an imprecise Model Discrepancy ϵ_{md} (and the corresponding effect on the History Match) is the main goal of this work.

3.3 Representing beliefs about f using emulators

An *emulator* is a stochastic belief specification for a deterministic function. This would be constructed after performing a large, space filling set of runs of the model [6]. Our emulator for component i of f is given by:

$$f_i(x) = \sum_j \beta_{ij}\, g_{ij}(x) + u_i(x)$$

where $B = \{\beta_{ij}\}$ are unknown scalars, g_{ij} are known deterministic functions of x, and $u(x)$ is a weakly stationary stochastic process. A simple specification is to suppose, for each x, that $u_i(x)$ has zero mean with constant variance and $\mathrm{Corr}(u_i(x), u_i(x'))$ is a function of $\|x - x'\|$. From the emulator, we may extract the mean, variance and covariance for the function, at each input value x.

$$\mu_i(x) = \mathsf{E}[f_i(x)], \qquad \kappa_i(x, x') = \mathrm{Cov}(f_i(x), f_i(x'))$$

Often, because of the mode of construction, the expectation of the emulator interpolates known runs of the model, while the variance represents uncertainty of the function at x inputs that have not been run. A key feature of an emulator is that it is (in most cases) several orders of magnitude faster to evaluate than the model itself. This is important as we will be exploring high dimensional input spaces that necessitate large numbers of evaluations. Emulator techniques are vital in the analysis of any model that has a moderate/long run time and a high dimensional input space.

3.4 Bayes Linear approach

For large scale problems involving computer models, a full Bayes analysis is hard for the following reasons. Firstly, it is very difficult to give a meaningful full prior probability specification over high dimensional input spaces. Secondly, the computations for learning from both observed data and runs of the model, and choosing informative runs, may be technically very challenging. Thirdly, in such computer model problems, often the likelihood surface is extremely complicated, and therefore any full Bayes calculation may be extremely non-robust. However, the idea of the Bayesian approach, namely capturing our expert prior judgements in stochastic form and modifying them by appropriate rules given observations, is conceptually appropriate.

The Bayes Linear approach is (relatively) simple in terms of belief specification and analysis, as it is based only on the mean, variance and covariance specification which, following de Finetti, we take as primitive. It also allows a relatively straightforward description of imprecision which is vital for this work.

We replace Bayes Theorem (which deals with probability distributions) by the Bayes Linear adjustment which is the appropriate updating rule for expectations and variances. The Bayes Linear adjustment of the mean and the variance of y given z is:

$$\mathsf{E}_z[y] = \mathsf{E}[y] + \mathrm{Cov}(y, z)\mathrm{Var}(z)^{-1}(z - \mathsf{E}[z]),$$
$$\mathsf{Var}_z[y] = \mathrm{Var}(y) - \mathrm{Cov}(y, z)\mathrm{Var}(z)^{-1}\mathrm{Cov}(z, y)$$

$\mathsf{E}_z[y]$, $\mathsf{Var}_z[y]$ are the expectation and variance for y adjusted by z.

The Bayes linear adjustment may be viewed as an approximation to a full Bayes analysis, or more fundamentally as the "appropriate" analysis given a partial specification based on expectation (with methodology for modelling, interpretation and diagnostic analysis). For more details see [2].

3.5 History Matching using Implausibility Measures.

We can now use the emulator, the model discrepancy and the measurement errors to calculate a Univariate Implausibility Measure, at any input parameter point x, for each component i of the computer model $f(x)$.

This is given by:

$$I^2_{(i)}(x) = |\mathsf{E}[f_i(x)] - z_i|^2 / \mathrm{Var}(f_i(x) - z_i) \quad (3)$$

which now becomes:

$$I^2_{(i)}(x) = |\mathsf{E}[f_i(x)] - z_i|^2 / (\mathrm{Var}(f_i(x)) + IMD + OE) \quad (4)$$

where $\mathsf{E}[f_i(x)]$ and $\mathrm{Var}(f_i(x))$ are the emulator expectation and variance, z_i are the observed data and $IMD = \mathrm{Var}(\epsilon_{md})$ and OE are the (univariate) Imprecise Model Discrepancy variance and Observational Error variance.

When $I_{(i)}(x)$ is large this implies that, even given all the uncertainties present in the problem, we would be unlikely to obtain a good match between model output and observed data were we to run the model at input x. This means that we can cut down the input space by imposing suitable cutoffs on the implausibility function (a process referred to as History Matching). Regarding the size of $I_{(i)}(x)$, if we assume that for fixed x the appropriate distribution of $(f_i(x^*) - z)$ is unimodal, then we can use the 3σ rule which implies that if $x = x^*$, then $I_{(i)}(x) < 3$ with a probability of approximately 0.95 (even if the distribution is asymmetric). Values higher than 3 would suggest that the point x should be discarded.

It should be noted that since the implausibility relies purely on means and variances (and therefore can be evaluated using Bayes Linear methodology), it is both tractable to calculate and simple to specify and hence to use as a basis of imprecise analysis.

One way to combine these univariate implausibilities is by maximizing over outputs:

$$I_M(x) = \max_i I_{(i)}(x) \quad (5)$$

Using the above unimodal assumptions, values of $I_M(x)$ of around 3.5 might suggest that x can be discarded, as is discussed in section 4.2.

If we construct a multivariate model discrepancy, then we can define a multivariate Implausibility measure:

$$I^2(x) = (\mathsf{E}[f(x)] - z)^T \mathrm{Var}(f(x) - z)^{-1} (\mathsf{E}[f(x)] - z),$$

which becomes:

$$(\mathsf{E}[f(x)] - z)^T (\mathrm{Var}(f(x)) + IMD + OE)^{-1} (\mathsf{E}[f(x)] - z).$$

Again, large values of $I(x)$ imply that we would be unlikely to obtain a good match between model output and observed data were we to run the model at input x. Choosing a cutoff for $I(x)$ is more complicated. As a simple heuristic, we might choose to compare $I(x)$ with the upper critical value of a χ^2 distribution with degrees of freedom equal to the number of outputs.

4 Application to a Galaxy Formation Simulation

One of the long-term goals of the Galform project is to identify the set of input parameters that give rise to acceptable matches between outputs of the Galform model and observed data. We do this using the History Matching ideas outlined above, the full details of which will be reported elsewhere. Before one can embark on such a process, the imprecise model discrepancy must be constructed, and its impact understood, as we now describe.

We proceed to analyse the Galaxy Formation model Galform using the computer model techniques described above. We choose to examine the mean of the first 40 sub-volumes (following the cosmologists' own attempts to calibrate) and select 11 output points from the bj and K luminosity graphs for use in our analysis, as shown in figure 2.

First, 1000 evaluations of the model were made (also shown in figure 2) using a space filling latin hypercube design across the 8-dimensional input space. These runs were used to construct an emulator for Galform as discussed in section 3.3.

We now describe the imprecise model discrepancy used to capture the cosmologist's assessment of the discrepancy between model and reality, and then go on to examine the imprecise implausibility measures this generates, and their impact on the judgement as to which inputs x are deemed acceptable.

4.1 Imprecise Model Discrepancy

At this stage we need to assess the Model Discrepancy ϵ_{md} related to all 11 outputs of interest. This is obtained from an expert opinion regarding the discrepancy between the model and reality, derived from opinions about potential deficiencies of the model. As this is a difficult assessment to make, an imprecise quantification of the model discrepancy will often be the most realistic representation of such uncertainty.

As we are doing a Bayes Linear analysis we only need to consider the assessment of $\mathsf{E}[\epsilon_{md}]$ and $\mathrm{Var}(\epsilon_{md})$. This is a major benefit of the Bayes Linear approach as we can represent any imprecision by letting some of these quantities vary over specified ranges and can then explore the consequences in the rest of our analysis. This is straightforward in comparison to a fully probabilistic analysis where such an imprecise specification would be extremely difficult, and a subsequent examination of the impact of such imprecision would often be intractable.

A leading expert stated that his beliefs regard-

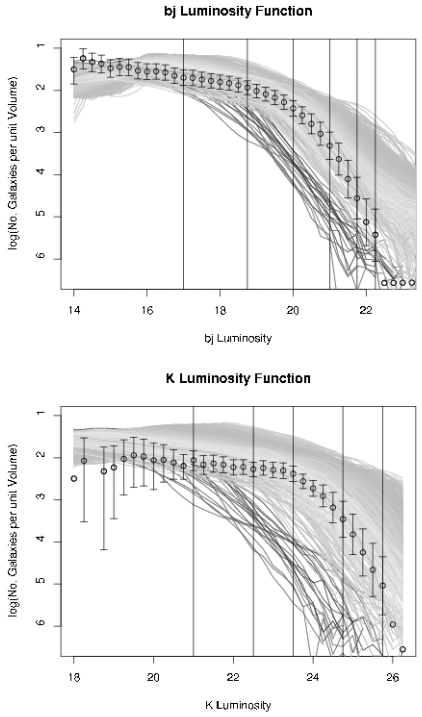


Figure 2: The bj and K luminosity outputs from 1000 runs of the model. The vertical black lines show the 11 outputs chosen for emulation. The error bars now incorporate the (univariate) model discrepancy with $a = \overline{a}$.

ing the model discrepancy were symmetric in that $\mathsf{E}[\epsilon_{md}] = 0$. Define $IMD = \mathrm{Var}(\epsilon_{md})$. Even for the univariate case (i.e. considering only one of the 11 outputs) the individual expert was unwilling to assess the size of *IMD* precisely. However, the expert was willing to make an imprecise assessment by specifying lower and upper bounds $\underline{IMD}$ and $\overline{IMD}$.

For the multivariate case, we needed to assess $IMD = \mathrm{Var}(\epsilon_{md})$ which is now an 11x11 matrix. The structure of this matrix will come from the expert's opinion as to the deficiencies of the model. In the case of Galform there are two major physical defects that can be identified. The first is the possibility that the model has too much (too little) mass in the simulated universe. This would lead to the 11 luminosity outputs all being too high (or too low), and would lead to positive correlation between all outputs in the MD matrix. The second possible defect is that the galaxies might age at the wrong rate leading to more/less blue galaxies and therefore less/more red galaxies. This would be represented as contributing a smaller negative correlation between the bj and K luminosity outputs. To respect the symmetries of these possible defects, the multivariate Imprecise Model Discrepancy (*IMD*) was parameterised in the following form:

$$IMD \;=\; a^2 \begin{pmatrix} 1 & b & .. & c & .. & c \\ b & 1 & .. & c & . & c \\ : & : & : & : & : & : \\ c & .. & c & 1 & b & .. \\ c & .. & c & b & 1 & .. \\ : & : & : & : & : & : \end{pmatrix} \qquad (6)$$

where now a, b and c are imprecise quantities, and we obtain the following expert assessments: $\underline{a} = 3.76 \times 10^{-2}, \overline{a} = 7.52 \times 10^{-2}, \underline{b} = 0.4, \overline{b} = 0.8,$ and $\underline{c} = 0.2, \overline{c} = b$.

It is possible to build in far more structure into *IMD* if required. The more detailed the structure, the more difficult eliciting expert information becomes. However, note the relative ease of specifying useful high-level imprecise statements using expectation as primitive, as compared to the corresponding effort for a fully probabilistic analysis. Exploring the effects of these specifications is also an easier task, as we now show by examining the effects of varying choices of a, b and c on the appropriate implausibility measures.

4.2 Implausibility Measures

In section 3.5 we showed how to construct the maximised and multivariate Implausibility measures $I_M(x)$ and $I(x)$. As these are derived using the imprecise model discrepancy we can write the dependence of these two implausibility measures on a, b and c explicitly. We can now explore the effects on $I_M(x, a)$ and $I(x, a, b, c)$ of varying a, b and c within the credal set C defined by:

$$\underline{a} < a < \overline{a}, \;\; \underline{b} < b < \overline{b}, \;\; \underline{c} < c < b,$$

as is described in the next section. As the implausibility measures are now imprecise, in order for regions of the input space x to be discarded as Implausible, they must violate the implausibility cutoff for all values of a, b and c, that is:

$$I(x, a, b, c) \;>\; I_{cut} \quad \forall\, a, b, c \in C, \qquad (7)$$

with a similar relation for $I_M(x, a)$:

$$I_M(x, a) \;>\; I_{Mcut} \quad \forall\, a \in C. \qquad (8)$$

In section 4.3 we set $I_{cut} = 26.75$ corresponding to a critical value of 0.995 from a χ^2 distribution with 11 degrees of freedom (and $I_{Mcut} = 3.5$) which were felt to be appropriate, conservative choices for the cutoffs. Note that if an input x satisfies either constraint (7) or constraint (8) then it is deemed implausible and will be discarded. As can be seen from equations (6),(3)

and (5), $I_M(x, a)$ is a monotonically decreasing function of a and hence constraint (8) will be equivalent to:

$$\min_{a \in C} I_M(x, a) \;=\; I_M(x, \overline{a}) > I_{Mcut} \qquad (9)$$

The constraint for $I(x, a, b, c)$ is more complex and in general no such monotonicity arguments can be used. In a full calibration analysis we would, for fixed x, evaluate $I(x, a, b, c)$ over a large number of points in the credal set C, and only discard the input x if it does not satisfy the implausibility cutoffs for every one of these points. However, here we are more interested in understanding the impact of different choices of a, b and c on the input space, which we do in the next section.

4.3 Effect of the Imprecise Model Discrepancy on the Assessment of the Best Input x^*

The most important effect of an imprecisely specified model discrepancy is its impact upon the choice of acceptable input parameters x^*. Above we showed how to construct the implausibility measures and described their use in deciding which inputs would be deemed acceptable. Here we will explore the impact of the imprecision on the multivariate measure itself, then on the percentage of input space remaining, by analysing the effects of varying a, b and c. Note that while we present all the pictures in greyscale, these displays are designed for presentation in colour.

Figure 3 shows the multivariate implausibility $I(x, a, b, c)$ as a function of a, b and c for two different fixed values of x. In the top (bottom) panel x_7 i.e. alphahot is set to its minimum (maximum) value of 2 (3.7). In both panels x_1 i.e. vhotdisk is at its maximum value of 550, and all other inputs are at their midrange values. In these and subsequent figures we examine slightly larger ranges for a, b and c than are defined by the Credal Set: here they satisfy $0.5\underline{a} < a < 2\overline{a}$, $0 < b < 0.95$ and $0 < c < b$. The top panel shows that $I(x, a, b, c)$ is minimised for large values of a, b and c attaining a minimum of approximately $I(x, a, b, c) = 14.2$. In the bottom panel however, the implausibility is minimised for low values of b and c and only attains a minimum of $I(x, a, b, c) = 38.3$. This shows the dramatically different behaviour of the implausibility measure as a function of a, b and c for two different parts of the input space, and specifically that general monotonicity arguments (such as used in equation (4.2)) cannot be applied to the imprecise parameters b and c. Plots such as those shown in figure 3 are very useful in helping to understand the impact of an imprecise assessment. However, one cannot examine such plots

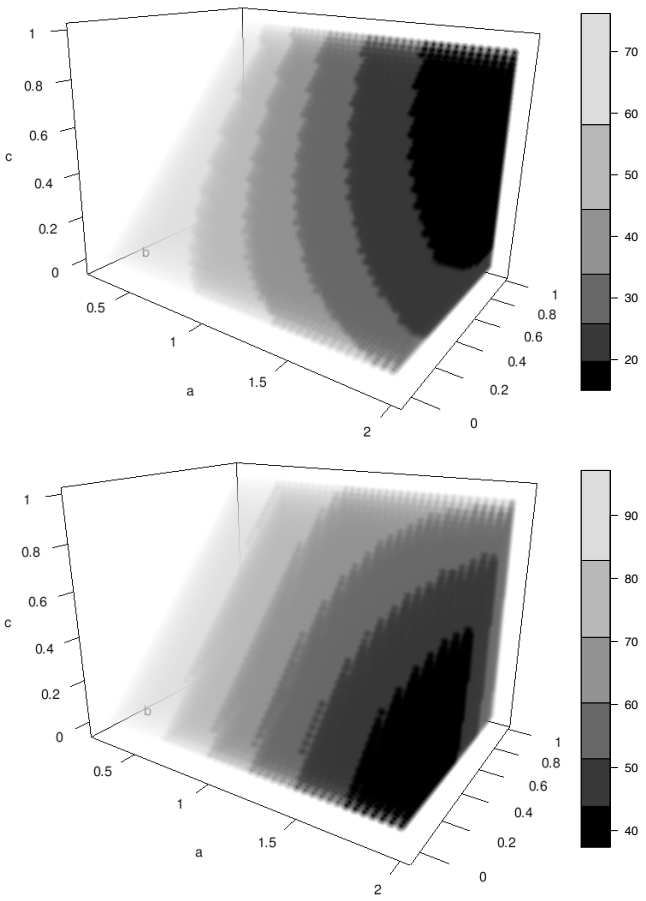


Figure 3: Both panels shows the multivariate implausibility $I(x, a, b, c)$ as a function of a, b and c for two different fixed values of x, with darker colours representing lower implausibility. Here a, b and c vary over the ranges $0.5\underline{a} < a < 2\overline{a}$, $0 < b < 0.95$ and $0 < c < b$. Note that the scale on the a-axis is in terms of multiples of $\overline{a}$. Top panel: vhotdisk = 550, alphahot = 2, Bottom panel: vhotdisk = 550, alphahot = 3.7, all other inputs set to their midrange values.

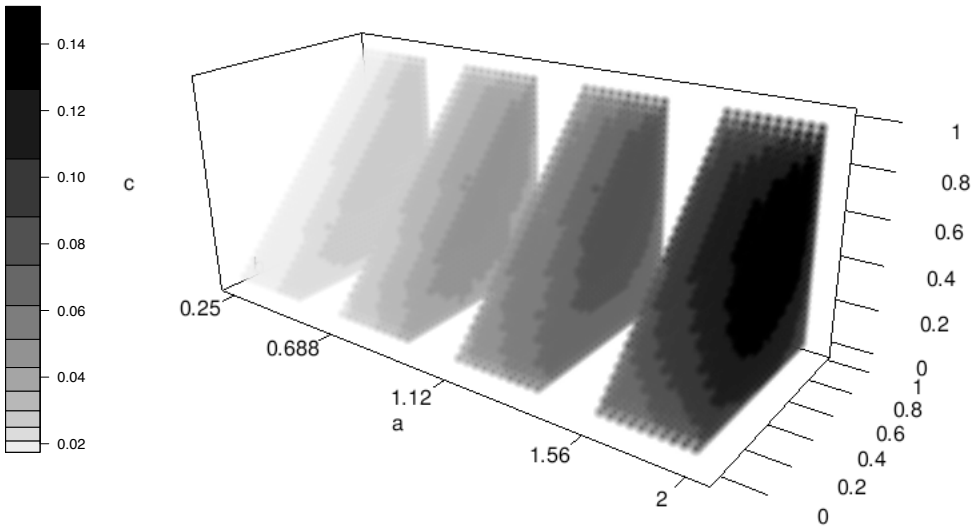


Figure 4: Fraction of input space that survives the multivariate implausibility cutoff given by equation (7) with $I_{cut} = 26.75$ as a function of a, b and c in the ranges $0.5\underline{a} < a < 2\overline{a}$, $0 < b < 0.95$ and $0 < c < b$. Note that the scale on the a-axis is in terms of multiples of $\overline{a}$.

for all points in the 8-dimensional input space. We therefore look at other ways to summarise and visualise the analysis.

We can summarise the effect of the imprecise multivariate implausibility cutoff given by equation (7) on the whole of the input space by looking at the fraction of space remaining once the cutoff has been imposed. Here we display the results corresponding to $I_{cut} = 26.75$, a value which was thought to be a reasonably conservative choice. Figure 4 shows this fraction of space remaining as a function of a, b and c in the ranges $0.5\underline{a} < a < 2\overline{a}$, $0 < b < 0.95$ and $0 < c < b$, with darker colours representing higher fractions. Figure 5 shows the same 3D plot from a different perspective. The 3D object has been cut in 3 places to allow one to see slices of the function at fixed values of a. This shows that for large values of a, the maximum space remaining would occur for intermediate values of b and c (approximately $b = 0.7$ and $c = 0.6$ for $a = 2$), however for smaller values of a the space remaining would be maximised by large b and c (e.g. for $a = 0.5\overline{a} = \underline{a}$, $b = 0.95$ and $c = 0.95$: see figure 5). These plots also suggest that the space remaining is far less sensitive to variation in b and c than in a: it is useful for the expert to know therefore that their assessment for a is more significant than for b and c.

Figure 6 shows the fraction of space remaining as a function of a for fixed choices of b and c. The boundaries of the Credal Set are shown by dotted vertical lines. Again one can see that to maximise the space remaining requires intermediate values of b and c for large a, and large values of b and c for small a. Also note that as a tends to small values, the fraction of space remaining varies only slowly: in fact setting $a = 0$ (which is not shown in this figure) leads to 0.017 of the input space remaining: this is important for the expert to know as it shows that some of the input space would survive the cutoff even for zero model discrepancy.

Examining the space remaining is useful in understand the effects of the imprecise specification of model discrepancy. However, it is also vital to assess the effect on the input space directly i.e. to determine which inputs x would not be discarded due to the imprecise specification. One way to analyze this is to ask what is the minimum value of a that is required to ensure that a particular input point x satisfies the implausibility cutoff. Figure 7 shows 3D plots of the required value of a as a function of the input parameters x_1 and x_7, and of b (with the other inputs at their midrange values , with $c = 0$, and the key in terms of multiples of $\overline{a}$). The darkest areas are those that have a required a of less than $\overline{a}$ and hence would survive the cutoff for the current specification. These plots show that while the value of b has effects in some parts of the input space, the region defined by required $a < \overline{a}$ is relatively independent of the value of b (a similar result is seen for plots with varying c and fixed b). This demonstrates that the required value of a is far more sensitive to the value of x_1 and x_7 as opposed to the specified range of the imprecise quantity b, and gives more evidence to suggest that the experts assessment

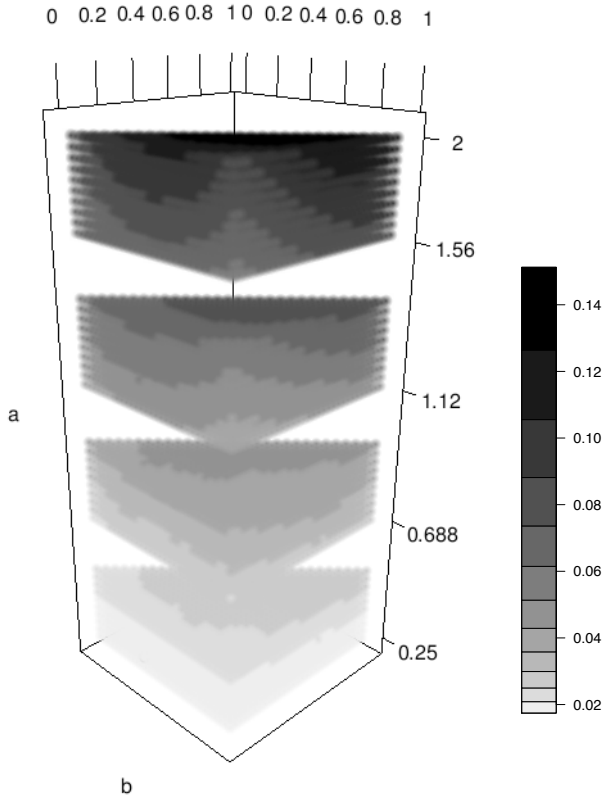


Figure 5: Alternative view of fraction of input space that survives the multivariate implausibility cutoff given by equation (7) with $I_{cut} = 26.75$ as a function of a, b and c in the ranges $0.5\underline{a} < a < 2\overline{a}$, $0 < b < 0.95$ and $0 < c < b$. Note that the scale on the a-axis is in terms of multiples of $\overline{a}$.

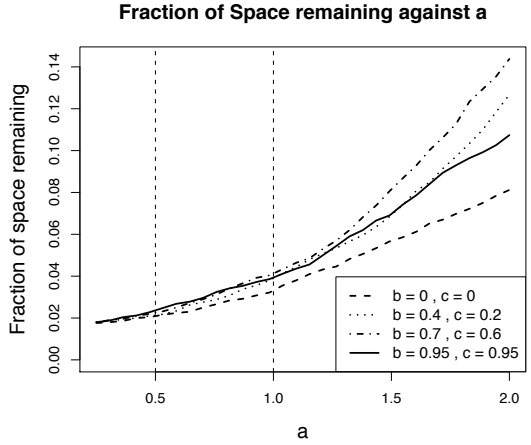


Figure 6: Fraction of input space that survives the multivariate implausibility cutoff given by equation (7) with $I_{cut} = 26.75$ as a function of a for the range $0.25\overline{a} < a < 2\overline{a}$, for various choices of b and c. The scale on the a axis is in terms of multiples of $\overline{a}$. Note that $\underline{a} = 0.5\overline{a}$. It can be seen that more space survives when $b = 0.7$ and $c = 0.6$ for large a, however, for smaller a the more extreme values $b = 0.95$ and $c = 0.95$ are preferred (which are not in the Credal Set).

for a is far more significant than that for b and c.

We have seen the effects of the imprecise assessment on the multivariate implausibility measure $I(x, a, b, c)$, on the fraction of space remaining after the cutoff is imposed, and on the set of allowed values of x_1 and x_7. We showed that these effects are non-trivial as the multivariate implausibility measure is a complicated function of x, a, b and c.

5 Conclusions

We have discussed how computer models make imprecise statements about physical systems. This imprecision arises due to the immense difficulty in giving a precise quantification on the discrepancy between the model analysis and the system. We have shown how use of Bayes Linear methods can provide a relatively straightforward description of this imprecision, allowing a meaningful elicitation of imprecise model discrepancy while leading to a tractable analysis of the issues involved in computer model calibration, which we demonstrated in the context of the galaxy formation simulation Galform.

The mathematical tractability of treating expectation as primitive also allows a detailed study of the effects of such imprecise assessments. In this case this in-

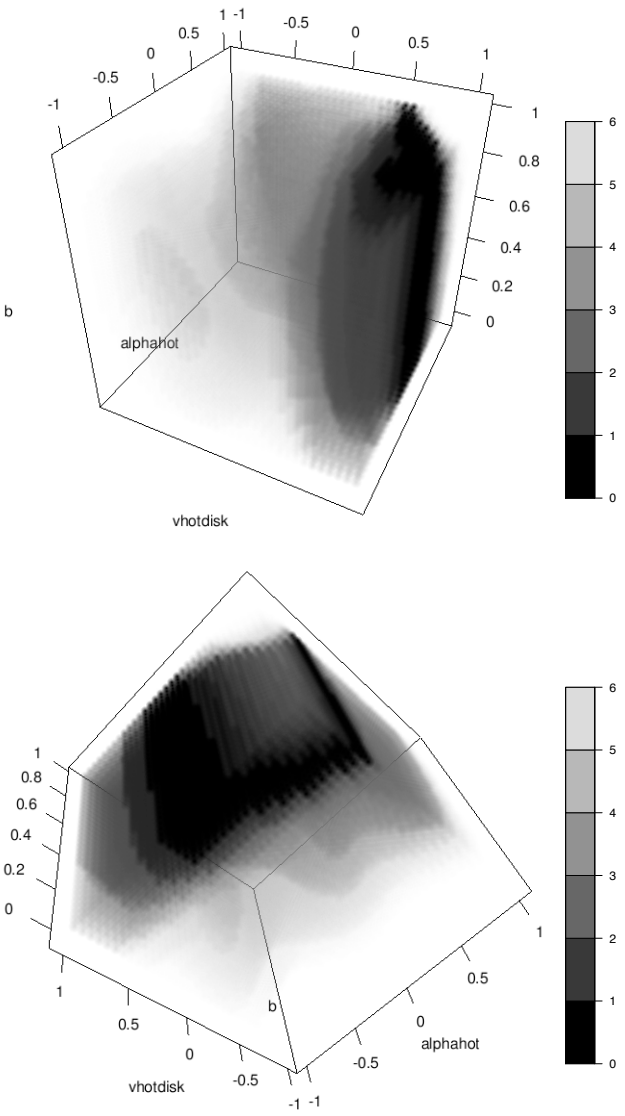


Figure 7: Plots showing the value of a that is required to ensure a point in input space satisfies the multivariate cutoff, as a function of the input parameters vhotdisk and alphahot, and of the imprecise quantity b (with c set to 0). The key is in terms of multiples of $\overline{a}$, and the darker areas represent low required a. All other input parameters have been set to their midrange values.

volved understanding the impact of the imprecision on the implausibility measures; measures that were used to discard regions on input parameter space thought to be very unlikely to give rise to acceptable matches between model output and observed data. In this way we were able to show the direct impact on parts of the input space of the expert's imprecise judgements regarding model deficiency. The effects of the imprecise assessments were found to be non-trivial and a variety of methods were used to summarise the data in order to produce meaningful visual representations of such effects.

Acknowledgements

This paper was produced with the support of the Basic Technology initiative as part of the Managing Uncertainty for Complex Models project, and with EPSRC funding through a mobility fellowship. We would like to thank Prof Richard Bower (Institute for Computational Cosmology, Physics Department, Durham University) for providing the expert assessments that feature in this work. We would also like to thank Prof Richard Bower and the Galfrom group (also based at the Institute for Computational Cosmology, Physics Department, Durham University) for access to the Galform model and to their computer resources.

References

[1] M. Goldstein and J.C.Rougier (2008). Reified Bayesian modelling and inference for physical systems (with discussion), *JSPI*, to appear, .

[2] Goldstein, M.,Wooff, D. (2007). Bayes Linear Statistics: Theory and Methods. *Wiley*

[3] Bower, R.G., Benson, A. J. et.al.(2006). The Broken hierarchy of galaxy formation, *Mon.Not.Roy.Astron.Soc.* 370, 645-655

[4] Kennedy, M.C. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society*, B,63, 425-464

[5] P.S. Craig, M. Goldstein, A.H. Seheult, J.A. Smith (1997). Pressure matching for hydocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments (with discussion), in *Case Studies in Bayesian Statistics*, vol. III, eds. C. Gastonis et al. 37-93. Springer-Verlag.

[6] Santner, T., Williams, B. and Notz, W. (2003). *The Design and Analysis of Computer Experiments*. Springer Verlag: New York.

Threat and Control in Military Decision Making

Christofer Waldenström
Department of computer and system sciences,
Stockholm University,
Sweden
christofer.waldenstrom@fhs.se

Love Ekenberg
Department of computer and system sciences,
Stockholm University,
Sweden
lovek@dsv.su.se

Mats Danielsson
Department of computer and system sciences,
Stockholm University
Sweden
mad@dsv.su.se

Abstract

This paper presents a model of how military commanders estimate the threat posed by the enemy in a tactical situation and how they use own forces to control that threat. The model is based on interviews with nine commanders from the Swedish navy and the purpose is to find automatic and adequate methods for reasoning about strategic issues based on the long-time experience of highly qualified military officers. The results show that the number of enemy units, the types of enemy units, the behavior of the enemy units, and the uncertainties regarding the number, types, and behavior determines the threat in a tactical situation. The own course of action works as a threat altering function to control that threat. When the commander should decide on a course of action, we suggest that it should be selected so it minimizes the expected threat.

Keywords. Military decision making, threat, worst case, expected value, imprecise probabilities

1 Introduction

Military decision-making means putting peoples life at stake in order to reach military objectives. The military decision makers are not only faced with risk of their own lives, their decisions also means subjecting own personnel and maybe even civilians to grave danger. Furthermore, the decisions often have to be made in highly stressful situations and in almost all cases under conditions of uncertainty and time pressure. When deciding what to do the military commander has to weigh possible gains against possible losses to determine the worth of each alternative. If an alternative where the possible gain outweighs the possible losses can be found, the risk of that alternative is considered worth taking, and it is chosen and implemented.

How military decision makers make such tradeoffs have not been studied to any great extent and empiric data in this field is almost nonexistent [1, 2]. Consequently, research is needed to investigate how military decision makers judge the risk of a certain course of action, and how they decide if that risk is worth taking. The rationale for this is that if we want to devise proper decision support we must first understand how such decisions are made in order to identify possible difficulties and pitfalls. This study is based on the assumption that determining acceptable risk means making a decision that strikes a balance between the factors that increase risk, the factors that decrease risk and the factors that justify risk. If such balance can be found, the risks following from the decision are acceptable and are worth taking. This paper focuses on how a commander estimates the threat posed by an enemy in a tactical situation and what he or she does to controls that threat. The results will be used as the groundwork aiming at devising a military decision support system.

2 Background

How a rational human being should make choices under conditions of uncertainty have been extensively studied in the field of normative decision-making, and a widespread opinion is that utility theory captures the concept of rationality [3-6]. Nevertheless, people seem to make decision in other ways but those stated by expected utility theory as has been demonstrated in a vast of psychological experiments [7, 8]. To accommodate deviations between the normative theories and the experimental results descriptive theories have been proposed [9, 10].

Luce and Raiffa's [5] distinction between certainty, risk or uncertainty has been further developed by Einhorn & Hogarth [11]. They distinguish between ignorance, ambiguity and risk according to the degree to which one can rule out alternative distributions. In a state of ignorance

no distributions are ruled out, while in a state of risk all but one distribution are ruled out. Ambiguity is an intermediate state between ignorance and risk and results from the uncertainty of specifying which of a set of distributions is appropriate in a given situation. Thus, ambiguity refers to not knowing the structure of the system that produces the outcomes. As showed by Ellsberg [12] people prefer risk to ambiguity).

This observation is of special interest in the military domain. The problem facing a military decision maker is to decide how to solve a mission in a hostile environment, and the decision is made difficult by the uncertainties regarding the enemy [13]. These uncertainties regard both what the enemy looks like (the structure of the system) as well as what the enemy will do (the outcome of the system).

Further, military decision making comprises of more than just selecting the best course of action from a given set. Courses of actions do not present themselves in a ready-made fashion, they must be developed, and this is done according the methods prescribed in military planning manuals [14]. These manuals prescribe the military decision-making process as a process aiming at procedural rationality [15] where course of action are first developed, and then the best is selected according to some criteria. Nevertheless, empirical research show that in many cases the decision maker only develops one 'good enough' course of action that is put to action [16, 17]. Thunholm [16] further showed that under conditions of time pressure rational methods do not produce better courses of action than intuitive methods. Even if this seems to spell bad news for the rational methods, it probably only means that better normative methods are yet to be found[1].

Another distinct feature of tactical decision-making in the navy is the decision-making tempo. The commander may only have a few hours to plan a mission before execution must begin. Once execution begins the focus changes; instead of devoting resources to decide what has to be achieved in the future, resources are redirected to figure out how the current operations are proceeding. Any difference between perceived state and the state predicted by the plan might be a potential problem. The commander must identify the situations that pose threats to the successful accomplishment of the mission. If a potential problem is detected, appropriate action must be devised and implemented in order to prevent derailing of operations. This makes military decision making an ongoing process. New courses of actions have to be developed and implemented as a reaction to the changing events [18].

[1] It should be noted that in some situations 'good enough' solutions, i.e., statisficing solutions, can be considered normative or even the only possible solution [26].

How people make decisions in such an environment been studied in the fields of dynamic and naturalistic decision-making. In the field of dynamic decision making the focus has been on how people in general control a dynamic system, and the difficulties they face in that task [19]. The results, however, are only on a general level and not immediately applicable to how military decision makers make judgments of threat and control.

Naturalistic decision making (NDM), on the other hand, are interested in how experts make decisions within their own fields and some studies have focused on military personnel [20, 21]. Results from this field indicate that decision makers employ quite stable strategies that, despite the presence of uncertainty, make it possible to make decisions.

In one NDM study Lipshitz and Strauss [22] studied how Israeli Army officers coped with uncertainty and concluded that the participants distinguished between three types of uncertainty: uncertainty caused by inadequate understanding, uncertainty caused by incomplete information and uncertainty caused by undifferentiated alternatives. They coped with these by applying five different strategies: i) reducing uncertainty (by collecting more information), ii) assumption-based reasoning, iii) weighing pros and cons of competing alternatives, iv) suppressing uncertainty, and v) forestalling. Similar strategies have been obtained by others, although the context in their studies was not military [10, 23]. Hutton [24] has made an extensive review of strategies with focus on the military context. But as in the case of dynamic decision-making no studies have explicitly focused on threat or control judgments.

Even if some effort has been made to describe how military decision makers cope with uncertainty, very few attempts have been made at investigating how they judge risk. What increases or decreases military risk, how uncertainty affects military risk and what makes military risks worth taking have neither been investigated to any great extent. This paper presents a model of how military decision maker judge the threat posed by the enemy and what he or she does to control that threat, and will be used to establish the requirements for a military decision support system. It should be noted that what people do is not necessary a good guide to what they should do. Nevertheless, a practical approach when designing support systems is to start with the problems people have in a task, helping people with things they find easy will probably leave that support unused. Thus, to identify these potential problems you need a descriptive account the task.

3 Method

The participants were nine officers who either were or had been in active duty in the Swedish navy. One of the participants had served as Chief of Navy, the highest commander of the Navy and a direct subordinate com-

mander to the Supreme Commander. One had served as Chief of Fleet, the highest commander of the Fleet. Two participants had served as Commander of a Surface Warfare Flotilla (the highest tactical commander of a naval mission consisting of 15-20 navy ships often coupled with support units such as helicopters, attack, fighter, or surveillance aircrafts). Three participants had served as Commanders of Surface Warfare Divisions (subordinate to a Flotilla Commander and in charge of approximately four to six navy ships). Two participants had served as Commanding Officers of a ship. Eight of the participants were specialized in anti surface warfare and/or anti submarine warfare and one officer in mine warfare. The participants had led between 10 to 100+ military planning processes on the tactical level or above, and they had led between 10 and 100 naval missions (exercise and/or live)[2]. All respondents were men.

The study was conducted using semi-structured interviews, duration ranging between 0.5-1.5 hours. The questions were based on the steps and tasks prescribed by the Swedish Navy's decision-making process (SNDMP), which like other military decision making processes is highly proceduralized process where of a number of distinct steps should be completed in sequence [25]. However, none of the steps or tasks in SNDMP explicitly states that the decision maker should carry out risk estimates, so asking how the respondents made such estimates would probably yield little or no data. Instead it was assumed that risk estimates would be embedded in the decision-making process and consequently all respondents had to describe how they carried out each of the steps.

The interviews were transcribed verbatim, leaving out pauses, humming etcetera and analyzed using content analysis. As no stage of the SNDMP explicitly calls for risk estimates it was suspected that the participants would use other phrases together with 'risk' when they accounted for how they made such considerations. Consequently, all statements containing the words "risk", "threat" and "danger" were excerpted. To determine if a statement related to judgments of threat or control, each were analyzed by the author. The data were reduced by amalgamation of similar statements and the result was checked for internal consistency (no contradictions within the statements) and integrated to form a coherent model of threat and control in military decision making.

4 Results

The results show that two things determine the level of threat in a tactical situation: i) the enemy and ii) the level of uncertainty regarding the enemy. All respondents expressed that the enemy is the major threat determinant (9 of 9). When considering the enemy, two questions occupy the participants: what forces does the enemy have and what can the enemy do? As expected, the more forces the enemy have and the more capable the forces, the higher the threat. Further, the forces can be employed differently leading to more or less threatening actions.

The other threat driver is uncertainty. The results indicate that the respondents (6 of 9) regard uncertainty almost synonymously with threat, risk or danger[3]. An uncertain situation is a threatening situation. As one of the lower experienced respondents put it "You often regarded different aspects of risk taking, what risks were acceptable, what uncertainties". When faced with uncertainty, as understood by some of the participants in this study (4 of 9), they deal with it by worst-case reasoning. This, however, gives a different bounding of risk than probability would give.

Consider the uncertainty about the enemy forces. Given no uncertainty at all, all enemy units that pose a threat, are known. Thus, the risk is equivalent to the threat posed by those units. As uncertainty increases, the more the decision maker tends towards worst-case reasoning. Consequently, risk is bound on the lower end by the threat posed by the known forces, and on the upper end by the threat posed by the worst plausible combination of enemy forces. The same reasoning goes for what the enemy can do. When uncertainty is zero then the risk is equal to what the decision maker knows the enemy is going to do. As uncertainty increases the risk approaches the threat posed by the worst plausible enemy course of action. The following statements from two of the higher-ranking respondents serve as examples:

> Let us say that you can get a decent understanding of what resources the enemy got, but what his possibilities are, how he thinks and ponders, that is not as easy. If you start to sort out, what are his resources? What kinds of ships are there, what kind of aircrafts, what other forces does he have?

> And then you lay low and wait. You know that he can approach this area, and your mission is to prevent him from entering and doing something in this area. […] Then you must keep track of where he is and what the most dangerous thing he can do is, and decide how you can counter that. And yes, the difficult part is to know how big they are, how many they are, and how strong they are. That is what you are going to think about.

In the military context, threat is controlled by employing own units and by devising an appropriate own course of action. On this point all respondents agree (9 of 9). The number of own units and the types of own units deter-

[2] About half of the respondents have participated in countering the repeated violations of Swedish territorial waters by submarines during 1980-1995, where several targets were engaged. If, and to what extent these violations took place are still causing controversy but this will not be further discussed here.

[3] This may be in part linguistic. The word 'uncertainty' has two meanings in Swedish, which can be translated to 'uncertain' and 'insecure'. However, when military personnel talk about 'uncertainties' regarding an operation they generally refer to the former meaning.

mines the control created by own units. The more own units, the higher the perceived control. The more capable the own types, the higher the perceived control. Following statements from two of the highly experienced respondents serve as examples:

> What is it that has to be done? What does the threat look like? What enemy forces are in the area? What forces will I have at my disposal? In that situation the first thought is: Do I have enough own forces or do I need support from other units? Do I need recognizance aircrafts, attack aircrafts, surveillance helicopters, or support from other surface attack forces? A first feeling; do I have enough forces, enough capability to solve this mission?

> I mean, what is level of risk you must be prepared to take? Of course there is a connection to the resources as I as tactical commander can use. And the difficulty is of course what resources I can get. What support can my mission [as tactical commander] get from the mission commander [the higher command]? There is a discussion about the supporting functions that I can get related to the level of risk. As an example: Can I get air support, costal missile batteries or something else as an additional strength. Or can I get submarine missions as support?

Control is also achieved by devising/selecting an own course of action that subjects own forces to more or less risk. The control achieved by own course of action is consequently transitive. Consider following statement from one of the high experienced respondents:

> It is embedded in this, the comparison of forces. How can I, so to say, protect my own forces and when can I strike, that is what it is all about. And if this comparison is to my advantage, which it seldom has through the years, it has always been an advantage to the enemy, both in numbers, size, resources, ranges, additional aircrafts and everything […] well yes, then I must, to protect my own forces as much as possible, utilize the protection I can get from maybe the terrain or similar, that is the archipelago, in another way than if we had an advantage of some sort in ranges. If that were the case, then you had been able to go out [on the open sea] in another way.

The results indicate that the threat posed by an enemy force is a function of how large the enemy force is (how many units it contain), how capable it is (what kind of types of units it contains), what the enemy is doing (behavior), and the uncertainties regarding the number, types and behavior of the enemy. Beginning with the properties of a unit, the threat posed by a unit is determined by its ability to destroy other units. To destroy another unit it must first be able to detect the other unit, and second, have a weapon that can be used to engage the detected unit. Thus, the threat or control posed by a unit is determined by the unit's ability to detect other units, together with the weapons carried by that unit.

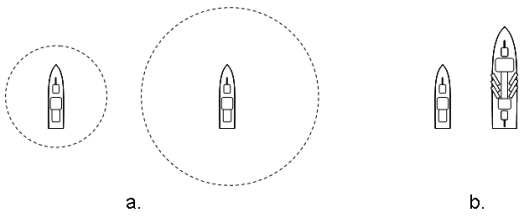


Figure 1

Looking at Figure 1a, two identical ships with regard to armament and maneuverability are depicted. In this example the right ship will be considered as more of a threat since it can detect units (and consequently fire a weapon against them) at a further distance than the left ship.

If we continue to the weapons, a unit is perceived as more of a threat if it carries more powerful weapons. Figure 1b depicts two ships: a patrol boat (to the left) and a destroyer (to the right). The patrol boat carries a single gun while the destroyer carries two guns and six surface-to-surface missiles. In this case, the destroyer will be perceived as the higher threat due to its heavier armament. Furthermore, the range of the weapons carried by a unit also determines its level of threat. A unit with long ranged weapons will be considered more of a threat than the same unit with shorter ranged weapons. The reason for this is that a unit with long ranged weapons may fire that weapon outside the detection range of friendly units.

Yet another property that increases threat or control is a unit's ability to avoid detection, its ability to stealth. If a unit has a high ability to stealth, the unit has the advantage of coming into range with its own weapons and sensors without being detected by the opposing unit.

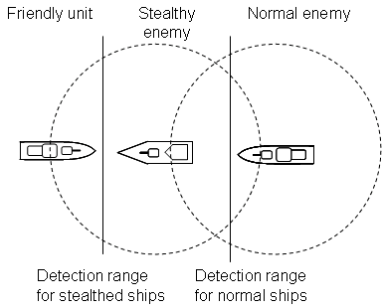


Figure 2

Looking at Figure 2, three ships are illustrated: a friendly unit (left) a stealthy enemy (middle) and a normal enemy (right). Even though the stealthy and the normal enemy have the same weapons and sensors, the stealthy enemy will be perceived as more of a threat since it can detect and fire a weapon on the friendly unit without being detected. Consequently, a unit with high ability to stealth may pose a higher threat than a normal unit, even if the normal unit is equipped with better sensors and armament.

As said earlier, the behavior of an enemy unit also affects the perceived threat. In Figure 3 an enemy ship is moving north, its weapon and sensor ranges illustrated by the dashed circle. Now suppose that the enemy unit suddenly changes course. If the course change will bring the enemy closer to the friendly unit, the perceived threat will increase since the friendly unit runs risk of coming within range of the weapons carried by the enemy. On the other hand, if the course change will bring the enemy further away from the friendly unit, the perceived threat will decrease for the opposite reasons.

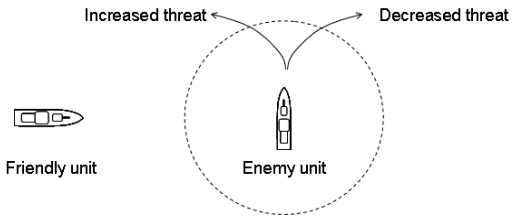


Figure 3

The capability of a force is determined in the same way as the capability of a single unit, by its ability to detect and destroy targets. But on a force level a procedure of target sharing can enhance those abilities. Once a naval operation is underway all units use their sensors to survey their immediate surroundings. All contacts are reported to designated units in the force, which compile the reports into a single, coherent view of the operation's area. This view is then distributed to the whole force. This procedure allows all units to become aware of all contacts held by the force, including contacts out of range by their own sensors.

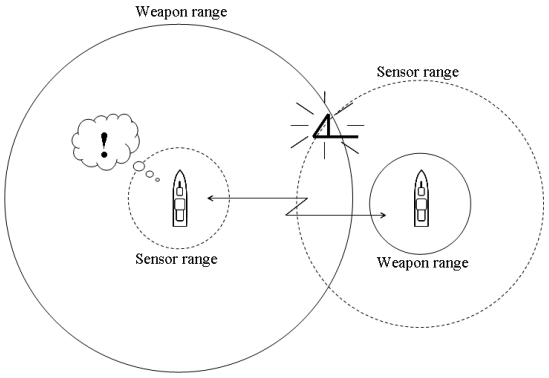


Figure 4

How this procedure can enhance the combined effect of the force is illustrated in Figure 4. The right ship with the greater sensor range detects a target with its radar. As the target is outside the range of its own weapons the right ship cannot itself destroy it. However, by sending the target to the partner to the left, the partner also becomes aware of the target. The left ship has much greater weapon range and as the target is within that range, the left ship can engage the target.

This simple scenario illustrates that the more capable a force is to detect targets, the more threatening will it appear. However, a force with superior surveillance capability is no threat at all if it does not have the capability to destroy the targets it has detected. Thus, the weapons it can employ also determine threat. The more powerful and the longer ranged they are, the more threatening the force will be perceived. On the other hand, the force is of no threat at all if it cannot detect any targets. Thus, to be a superior force it must have the upper hand both when it comes to sensor capability and weapons capability.

Figure 5 further illustrates the situation. To the left we see a force consisting of two ships of the same type. The inner zone, denoted by a dashed line, depicts the total area covered by the force's sensors. The outer zone shows the area covered by the force's weapons. The gray zone shows the area, in which this force can both detect and destroy targets; in this case it is the same as the area covered by sensors. If we now look at the right force we see that it consist of one ship and one helicopter. If we assume that this ship is of same type as the ships in the left force, we see that the area in which the right force has control is much larger that the left force's. This is due to the superior sensor range provided by the helicopter. If we now compare the threat perceived by the commanders in each force, the commander of the left force will probably perceive a higher degree of threat, despite the fact that he or she has twice as many weapons. This is quite evident since the right force can close in on the left force, use the helicopter to find the left force, fire its missiles at max range, without risking detection of the left force. Thus, the threat or control provided by a force is determined by its composition of the own force, in the same way as the threat posed by the enemy is determined by the composition of the enemy force.

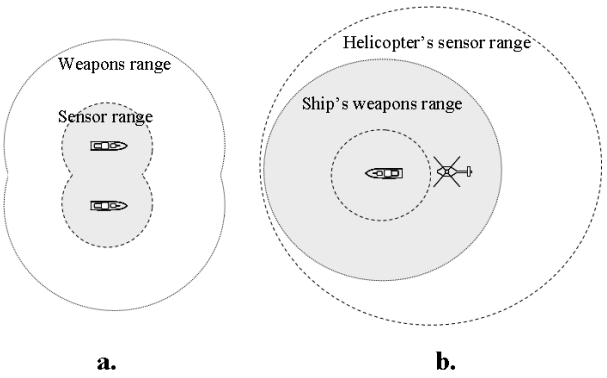


Figure 5

As have been illustrated above, the control provided by own units was determined in the same way as the threat posed by the enemy. The second way to handle the threat was to devise an appropriate own course of action. How this can be accomplished is illustrated in Figure 6. The mission is to move the ship from Port A on the mainland to Port B on the island. Intelligence has reported that during the initial phases of the operation no enemy is in the

area, but as the operation is underway the enemy will most likely try to prevent the transport. The commander concludes that if we move quickly we might get the transport to Port B without giving the enemy a chance to interfere. The plan is to move the transport ship at high speed across the open water, thus minimizing exposure time to the enemy threat. The friendly units will establish a protective screen.

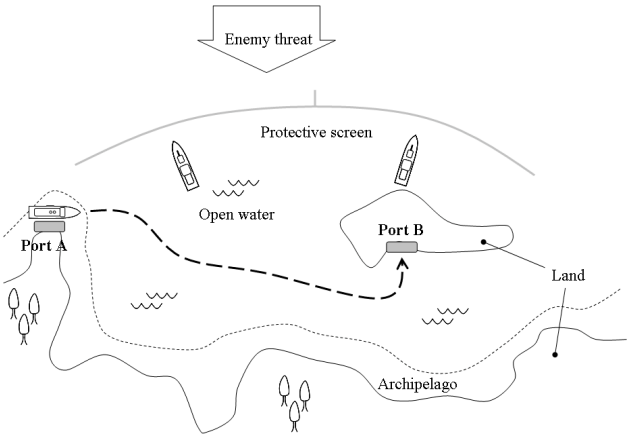


Figure 6

Now assume the operation is underway and the transport ship has reached a point on the open water between Port A and Port B. Suddenly, an enemy ship is detected and identified. Since the open sea does not provide any protection it is assumed that the enemy also has detected the transport ship. Figure 7 illustrates the situation. The enemy has a weapon range denoted by r_1 and the friendly ship a weapon's range of r_2. This means that the enemy ship cannot be allowed to get any closer than r_1 to the transport ship, or else the transport ship runs risk of being sunk.

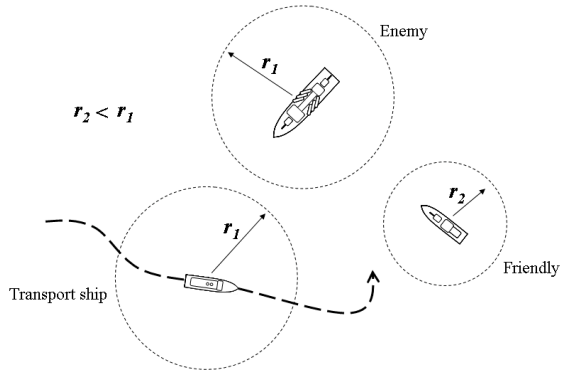


Figure 7

The commander can handle the threat in two ways. One alternative is to order the transport ship to head south and hide in the archipelago. This makes the transport ship difficult to detect and consequently difficult to destroy. The other option is try to sink the enemy ship, removing the threat altogether. However, attacking the enemy is dangerous since the own ship is inferior when it comes to weapon ranges ($r_2 < r_1$). On the other hand, it may be worth the risk since a successful attack will lower the overall threat for the rest of the operation.

In this case the commander orders the transport ship to head south and seek cover in the archipelago. The idea is to let the transport ship move in the archipelago to the point on the mainland where the distance to the island is minimal. Once there, it will lay low and wait until the friendly units have cleared the route to Port B, as shown in Figure 8. Using same reasoning as before, the area that must be cleared is obtained by measuring the range of the enemy's longest ranged weapon and apply that distance perpendicular to the planned route. When the area is cleared the transport ship will rush out at maximum speed, giving the enemy minimum amount of time to act before the transport ship reaches Port B.

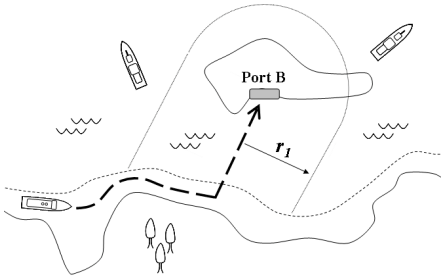


Figure 8

As pointed out, one of the most difficult aspects of military decision-making is the analysis of the enemy. Such analysis is made difficult because all information regarding the enemy is afflicted with uncertainty. The uncertainty regards three aspects of the enemy forces: (i) the number of units, (ii) the types of units and (iii) the behavior of the units. All these aspects affect the perceived threat.

This can be modeled in a tree structure (see Figure 9). The root node (S) represents the current scenario, i.e., the context in which the naval operation should be conducted. The intermediate nodes consist of the three aspects describing the enemy, where the first level represents the number of enemy units (n), the second level the types of enemy units (t), and the third level the behavior of the enemy units (b). The value nodes (v) quantify the perceived threat of each branch in the tree.

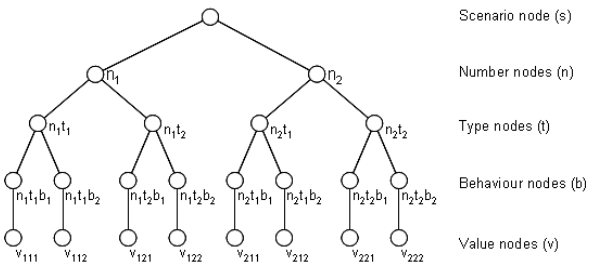


Figure 9

But as we saw above, the threat posed by a naval force could not be obtained by just adding the threat values of

the single units. It was the composition of the force that created the actual threat value. The tree structure accommodates this situation. Earlier it was illustrated that a force consisting of one surface ship and one helicopter posed a different threat than a force consisting of two surface ships (all surface ships are of the same type). A tree representation of this situation is presented in Figure 10. The two forces consists of the same number of units, hence the number node is n=2. The types are however different giving two type-nodes: t_1=2 surface ships; t_2=1 helicopter and one surface ship. If we assume the same behavior of each force, b_1=Attacking, then different threat values are assigned to the value nodes, $v_{n_1t_1b_1}$=4 and $v_{n_1t_2b_1}$=8.

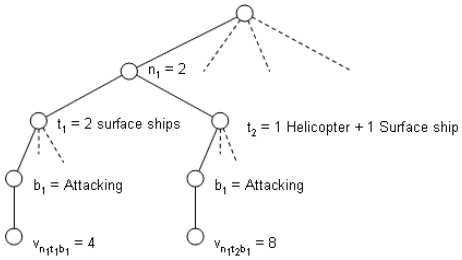


Figure 10

When analyzing the own forces, the commander considers the same aspects as those of the enemy, the number of units, the types of units, and the behavior of the units. It is consequently tempting to model the own forces in a tree structure, similar to the enemy. There is, however, a difference. There is hardly any uncertainty at all regarding the own forces. When an operation is initiated the commander receives a mission statement from higher command. This statement contains the task to be solved, a roster of the forces assigned to the commander, and information about the enemy. When planning begins all these pieces are fixed. The commander can neither influence the mission assigned, nor the forces, nor the intelligence about the enemy. Representing the own force could be quite straight forward, as illustrated in Figure 11:

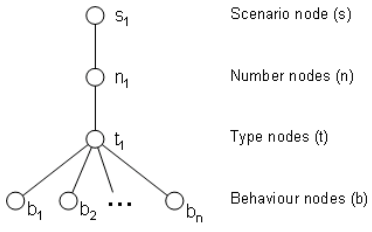


Figure 11

Nevertheless, using a tree structure in the case presented here brings along two problems: (i) it is hard to determine the value nodes since the control provided by the own force depend on a comparison between the own force and the enemy, (ii) the probability assignments of the own behavior has no meaning because the commander decides on a course of action given the own force and the threat. So, how should the own force be represented taking these constraints in mind?

As we saw earlier, the roster of the own forces made both the numbers of ships (n) and the types of ships (t) fixed. The only thing the commander can influence is the behavior of the own forces. As a consequence, the own force can represented similar to the enemy, as a single type-node that is then used as an argument when deciding how to solve the mission. Thus, the own behavior can be seen as a threat-altering function that given the own force influence the enemy's opportunity to pose threat to the own operation. Consider the situation described in Figure 7. When the transport ship heads south to take cover in the archipelago the negative value of being sunk is the same, however the probability that the enemy will sink the ship has been reduced. The alternative behavior, attacking the enemy ship and trying to sink it, will lead to that the probabilities of the number of enemy ships are altered.

To this point we have looked at how the commander analyses the threat and how the commander's own course of action alters that threat. However, the problem facing the commander is of course how to devise a proper course of action, taking in to account all uncertainties inherent in the information about the enemy. As the results indicated, the commander copes with this situation by employing worst-case reasoning. Even if this strategy might reduce the cognitive load it brings along at least two problems. First, the commander may have to design a very specific course of action to deal with the worst possible threat. The risk of that is of course that the commander may stretch the own resources towards the specific case so much that the solution might be fragile to other cases: by optimizing to solve a single case the robustness of the solution is lost. A second problem is that given limited resources the commander may end up in a situation where no solution can be found. In any case, if we want to analyze the situation beyond what is done intuitively a more systematic approach is required.

5 Representation and Evaluation

The commander's decision consists of selecting one of several scenarios. In such a scenario tree, the decision is represented in tree form as a sequence of probabilities leading to some final outcomes described by the end nodes. All decision trees consists of a root node, representing the decision, a set of intermediary nodes, representing the scenarios and uncertainty regarding the scenarios, and the outcome nodes describing the consequences of the scenarios. For each intermediate node, there is a probability associated with the node (number node, type node, or behavior node). In real planning situations, there is uncertainty inherent in the input data to the planning process. In the model, this is represented by probabilities and outcome values being in the form of

interval variables, i.e. the variables having a lower and an upper bound. For example, the decision-maker statement that probability p_i is between a_1 and a_2 is denoted $p_i \in [a_1, a_2]$ and translated into $p_i > a_1$ and $p_i < a_2$ in the model. Similarly, the value of the outcome i (v_i) is between a_1 and a_2 is denoted $v_i \in [a_1, a_2]$ and translated into $v_i > a_1$ and $v_i < a_2$. In this way, sets of statements (inequalities) are formed.

The collection of probability statements in a decision situation is called the node constraint set. A constraint set is said to be consistent if it can be assigned at least one real number to each variable so that all inequalities are simultaneously satisfied. The probability and value constraint sets are collections of linear inequalities. A minimal requirement for such a system of inequalities to be meaningful is that it is consistent, i.e., there must exist some vector of variable assignments that simultaneously satisfies each inequality in the system. In other words, a consistent constraint set is a set where the constraints are not contradictory.

Definition: Given a tree T, let N be a constraint set in the variables $\{ n_{...i...j...} \}$. Substitute the intermediary node labels $x_{...i...j...}$ with $n_{...i...j...}$. N is a *node constraint set* for T if for all sets $\{n_{...i1},\ldots,n_{...im}\}$ of all sub-nodes of nodes $n_{...i}$ that are not leaves, the statements $n_{...ij} \in [0,1]$ and $\sum_j n_{...ij} = 1$, $j \in [1,\ldots,m]$ are in N.

Thus, a node constraint set relative to a tree can be seen as characterizing a set of discrete probability distributions after a certain level (the *probability constraint set*). The core of these can be thought of as an attempt to estimate a class of mass functions by estimating the individual discrete function values. The normalization constraints ($\sum_j x_{ij} = 1$) require the probabilities of sets of exhaustive and mutually exclusive nodes to sum to one.

Requirements similar to those for node variables can be found for value variables. However, no dimension reducing normalization constraints (variables summing to one) exist for the value variables.

Definition: Given a tree T, let L be a constraint set in $\{t_{...1}\}$. Substitute the leaf labels $x_{...1}$ with $c_{...1}$. Then L is a *value constraint set* for T.

Similar to probability constraint sets, a value constraint set can be seen as characterizing a set of value functions. The elements above constitute a command frame, which constitutes a complete description of the probabilistic threat situation.

Definition: A *command frame* is a structure $\langle T,N,V \rangle$, where T is a scenario tree, N is a node constraint set for T and V is a threat constraint set for T.

While an evaluation of a consequence set may result in an acceptable expected value, the consequences of selecting it might be so dire that it should nevertheless be avoided. The commander may want to exclude particular alternative courses of action that are, in some sense, too risky. It might, for example, endanger the entire purpose of the operation, and in that case even a consequence with a low probability is too risky to neglect.

The intuition behind security levels is that they express when a scenario is undesirable. Thus, a decision-maker might regard a scenario as undesirable if it has consequences with too low a value, and with too high a probability to occur. This means that if several consequences of a strategy are too dire (w.r.t. a certain value parameter), their total probability should be considered even if their individual probability is too low to render the scenario undesirable. Such exclusions can be dealt with by specifying a security level for the probability and a threshold for the value. Then a consequence set would be undesirable if it violates both of these settings. The security level has the following basic form

$$S(C_i, r, s) = (\sum_{v_{ij} \le r} p_{ij} \le s)$$

where r is the minimally tolerable value threshold and s is the maximally acceptable probability for events below the threshold to occur. This is a boolean function sorting out unwanted consequence sets.

The remaining scenarios are selected according to a decision rule, usually by maximizing the expected value of an alternative. Looking at Figure 9 the expected threat in the situation s, T(s), is calculated using the following formula:

$$T(s) = \sum_{i=1}^{2} n_i \sum_{j=1}^{2} t_j \sum_{k=1}^{2} b_k v_{ijk}$$

This structure is generalized into the following formula for calculating the generalized expected threat:

Definition: Given a scenario S_i for i=1,…,r the expected threat of that scenario is given by the expression

$$T(S_i) = \sum_{i=1}^{n_i} n_i \sum_{j=1}^{n_j} t_j \sum_{k=1}^{n_k} b_k v_{ijk}$$

where n_i denotes the probability that the enemy has n_i number of ships, t_j denotes the probability that the enemy has t_j types of ships, b_k denotes the probability that the enemy will use behavior b_k, and v_{ijk} denotes value of the perceived threat of the combination n_i t_j b_k.

Given the threat in a scenario, the own course of action was regarded as a threat-altering function, taking the own force and the threat as arguments:

Definition: Given a scenario S_i with the expected threat $T(S_i)$ and the own forces F(n,t) where n=number of ships and t=types of ships. Behavior B_j is a function such as:

$$B : B\big(F(n,t), T(S_i)\big) \rightarrow T(S_i)'$$

Faced with many possible own course of action the question arises of which one to choose. What rule the commander uses have not been established but we suggest that the commander should *devise and select a behavior that given the own force solves the mission and minimizes the expected threat.*

Definition: Given a scenario S_i with the expected threat $T(S_i)$ and a set of own behaviors B_j, j=1...r such that $B_j : B_j(T(S_i)) \rightarrow T_j(S_i)$ giving the set of expected threats $T'(S_i) = \{T_1(S_i), \ldots, T_r(S_i)\}$. Minimizing the expected threat means selecting B_j such that $T_j(S_i) = \min T'(S_i)$

Here we use expected utility, but the framework allows for other methods to be used. As an example, quantiles can be implemented using security levels. It seems however somewhat reasonable to use the mean as an initial assumption because this will distribute the own forces according to the 'center of gravity' of the threat. Nevertheless, if the commander uses such an approach has to be established empirically.

Often, however, the expected value by itself is unable to discriminate between the scenarios. In such cases, a further analysis is called for in the form of an automated analysis called contraction. Contraction is a generalized sensitivity analysis that can be carried out in any number of dimensions. In complex decision situations, when an information frame contains numerically imprecise information, the different principles suggested above are often too weak to yield a conclusive result and will often yield a far too crowded set of candidates. One way to handle this could be to determine the stability of the relation between the considered consequence sets. As interval statements are deliberately imprecise, a natural way to investigate this is to consider values near the boundaries of the intervals as being less reliable than more central values. Using contractions we take this into account by indirectly measuring the dominated regions.

The principle of contraction is justified by the difficulties of performing simultaneous sensitivity analysis in several dimensions at the same time. If one uses only one-dimensional analyses, it can be hard to gain real understanding of the solutions to large decision problems because different combinations of dimensions can be critical to the evaluation results. Exploring all possible such combinations would lead to a highly complex procedure regarding the number of cases to investigate. Using contractions circumvents this difficulty. By co-varying the contractions of a set of intervals, it is possible to gain a much better insight into the influence of the structure of the information frame on the solutions. Both the set of intervals under investigation and the scale of individual contractions can be controlled. Further, contractions are measures of the strength of statements when original solutions sets are modified in controlled ways, rather than measures of the solution sets as given by volume estimates. Consequently, a contraction can be regarded as a focus parameter that zooms in on central sub-intervals of the full statement intervals.

Definition: X is a base with the variables $x_1,\ldots,x_n$, $\pi \in [0,1]$ is a real number, and $\{\pi_i \in [0,1] : i = 1,\ldots,n\}$ is a set of real numbers. $[a_i, b_i]$ is the interval corresponding to the variable x_i in the solution set of the base, and $\bar{k} = (k_1,\ldots,k_n)$ is a consistent point in X. A *π-contraction* of X is to add the interval statements $\{x_i \in [a_i+\pi\cdot\pi_i\cdot(k_i-a_i), b_i-\pi\cdot\pi_i\cdot(b_i-k_i)] : i = 1,\ldots,n\}$ to the base X. $\bar{k}$ is called the *contraction point*.

By varying π from 0 to 1, the intervals are decreased proportionally using the gain factors in the π_i-set, thereby facilitating the study of co-variation among the variables.

6 Discussion and further work

We have presented a model of how a commander estimates the threat in a tactical situation and how an own course of action is selected to control that threat. In a tactical situation the information about the enemy is almost always afflicted with uncertainty and the results indicated that the commander coped with this situation by worst-case reasoning. This work is part of the groundwork for further study of how a decision support system for tactical decision-making could look like. If such system should be realized as automatic quantitative support or as verbal heuristics remains to be determined.

Just considering alternatives and choosing in accordance with our like or dislike of risk can be considered a quite passive way of treating risk [23]. As we saw in this study, the own course of action was treated as a threat-altering function, which points to a more active stance towards risk: When facing a risky situation the respondents want to take action to influence and modify the risky situation. This is what [23] calls "adjusting the risks" and means gaining time, information or control. Time allows for information to be gathered, and information may resolve the uncertainty that makes the situation appear risky. Gaining control means taking actions to reduce the magnitude or the chance of loss. It would not be too surprising to find similar strategies employed by the participants in this study.

We suggested that a course of action should be selected that minimized the expected threat. It can be argued that a solution that tries to solve all 'possible threats' risk to end up being multi-useless instead of multi-purpose. However, statements like "…have enough width [in your COA]…" indicate a desire to devise a course of action that is easily adaptable so it can handle several developments of events.

To enable automatic reasoning the necessary information must be extracted from the commander or the staff and

structured rapidly. Populating the threat constraint set could be time consuming but a solution would be to find a formula that given the enemy forces and the own forces automatically can calculate the threat posed by any combination of own and enemy forces.

This study was based on the assumption that determining acceptable risk means making a decision that strikes a balance between the factors that increase risk, the factors that decrease risk and the factors that justify risk. Having dealt with the former two, our next work will focus on how a military decision maker judge if a risk is worth taking.

7 References

[1] Z. Lanir, B. Fischhoff and S. Johnson, "Military risk-taking: C^3I and the cognitive functions of boldness in war," *Journal of Strategic Studies,* vol. 11, pp. 96-114, 1988.

[2] Y. Vertzberger, *Risk Taking and Decisionmaking: Foreign Military Intervention Decisions.* Stanford, California: Stanford University Press, 1998,

[3] L. J. Savage, *The Foundations of Statistics.* ,2nd ed.New York: Dover Publications, 1972,

[4] J. von Neumann and O. Morgenstern, "Theory of Games and Economic Behavior," 1944.

[5] D. R. Luce and H. Raiffa, *Games and Decisions.* New York: Dover Publications, 1957,

[6] F. Knight, *Risk, Uncertainty, and Profit.* ,Reprint ed.Washington, D.C.: Beard Books, 2002,

[7] D. Kahneman, P. Slovic and A. Tversky, *Judgement Under Uncertainty: Heuristics and Biases.* US: Cambridge University Press, 1982,

[8] T. Gilovich, D. Griffin and D. Kahneman, *Heuristics and Biases.* New York, US: Cambridge University Press, 2002,

[9] D. Kahneman and A. Tversky, "Prospect theory: an analysis of decision under risk," *Econometrica,* vol. 47, pp. 263-291, 1979.

[10] Z. Shapira, *Risk Taking a Managerial Perspective.* New York, US: Russel Sage Foundation, 1995,

[11] H. Einhorn and R. Hogarth, "Ambiguity and uncertainty in probabilistic intference," *Psychological Review,* vol. 92, pp. 433-461, 1985.

[12] D. Ellsberg, "Risk, ambiguity and the Savage axioms," *Quarterly Journal of Economics,* vol. 75, pp. 643-669, 1962.

[13] C. Waldenström, "What is difficult in naval sensemaking," in *Proceedings of the 13Th International Command and Control Research and Technology Symposium,* 2008,

[14] US Army, "Field Manual 5-0 Army Planning and Orders Production." 2005.

[15] H. Simon, "Rationality as process and product of thought," *The American Economic Review,* vol. 68, pp. 1-16, 1978.

[16] P. Thunholm, "Military decision making and planning: towards a new prescriptive model," 2003.

[17] J. Schmitt and G. Klein, "A recognitional planning model," in *Proceedings of the 4Th ICCRTS,* 1999,

[18] M. van Creveld, *Command in War.* US: Harvard University Press, 1985,

[19] D. Dörner, *The Logic of Failure: Recognizing and Avoiding Error in Complex Situations (Originally Published in Germany 1989 Under the Title Die Logik Des Misslingens by Rowholt Verlag).* ,3rd ed.NY: Metropolitan Books, 1996,

[20] G. Klein, J. Orasanu, R. Calderwood and C. E. Zsambok, *Decision Making in Action: Models and Methods.* Norwood, NJ: Ablex Publishing Corporation, 1993,

[21] R. Flin, E. Salas, M. Strub and L. Martin, *Decision Making Under Stress: Emerging Themes and Applications.* Ashgate, 1997,

[22] R. Lipshitz and O. Strauss, "Coping with Uncertainty: A Naturalistic Decision Making Analysis," *Organizational Behavior and Human Decision Processes,* vol. 69, pp. 149-163, 1997.

[23] K. R. MacCrimmon and D. A. Wehrung, *Taking Risks.* New York, US: The Free Press, 1986,

[24] R. J. Hutton, "Types of "uncertainty about" and strategies for uncertainty management: An integrative approach," Klein Associates Inc., Fairborn, OH, Tech. Rep. #04TA2-SP1-RT2, 2004.

[25] Swedish Armed Forces, *Taktikreglemente För Flottan (Tactical Regulations for the Navy).* Stockholm: Chefen för Marinen (Head of Navy), 1987,

[26] R. Selten, "What is bounded rationality?" in *Bounded Rationality: The Adaptive Toolbox* G. Gigerenzer and R. Selten, Eds. US: First MIT Press, 2002, pp. 13-36.

Index