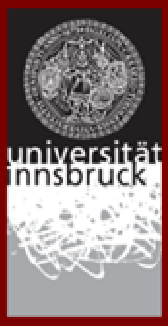


ISIPTA '11

July 25-28 2011 INNSBRUCK AUSTRIA

Proceedings of the
Seventh International Symposium on

Imprecise Probability: Theories and Applications

July 25-28 2011, Innsbruck, Austria

Unit for Engineering Mathematics
University of Innsbruck, Austria

www.sipta.org/isipta11

Edited by

Frank Coolen
Gert de Cooman
Thomas Fetz
Michael Oberguggenberger

ISIPTA '11

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

University of Innsbruck, Austria
July 25–28 2011

Edited by

Frank Coolen
Gert de Cooman
Thomas Fetz
Michael Oberguggenberger

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

Printed 2011 by
STUDIA Universitätsverlag
Herzog-Sigmund-Ufer 15
6020 Innsbruck, Austria

ISBN 978-3-902652-40-9

The book was typset using LaTeX.

Contents

Preface

The *Seventh International Symposium on Imprecise Probability: Theories and Applications* is held in Innsbruck, Austria, 25–28 July 2011.

The ISIPTA meetings are a primary forum for presenting and discussing advances in imprecise probability and are organized 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), Prague, and Durham (UK). 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 2011 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, 40 papers are presented by a short talk and poster, which guarantees ample time for discussion of each contribution. The papers are included in these proceedings and are also available on the SIPTA webpage (`www.sipta.org`). Submitted papers have undergone a high quality reviewing process by members of the Program Committee. 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. We continue with this tradition; short abstracts of these poster-only presentations will be distributed at the conference and are 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'11, engineering applications have been emphasized. In engineering, information on risk and uncertainties usually lies in the triangle spanned by probability, intervals, and expert opinion. Methods of imprecise probability thus are especially apt to modelling uncertainties in this field. This fact is increasingly acknowledged in the engineering community, as evidenced by the growing number of papers in engineering journals using methods from imprecise probability.

Two tutorial sessions are devoted to engineering applications. We thank Alberto Bernardini and Fulvio Tonon for preparing and presenting a tutorial on random set methods in civil engineering. An additional overview tutorial is given by Michael Oberguggenberger. The material is available at the SIPTA webpage.

A special historical and scientific session will be devoted to Bruno de Finetti. Bruno de Finetti, the founder of subjective probability theory, was born in Innsbruck in 1906, where he spent the first six years of his life. His father and his grandfather were engineers and both were involved in railway construction in Tyrol, the western parts of Austria and in Northern Italy at that time. The year 2011 marks the eightieth anniversary of the publication of the famous "De Finetti Theorem" in *Funzione caratteristica di un fenomeno aleatorio* (Atti della R. Academia Nazionale dei Lincei, Serie 6. Memorie, Classe di Scienze Fisiche, Mathematice e Naturale, 4:251–299, 1931). An early essay on subjective probability appeared in 1931 as well: *Sul significato soggettivo della probabilità* (Fundam. Math. 17, 298–329, 1931). The special session will be followed by a visit to Bruno de Finetti's birth place where a memorial tablet will be unveiled in the presence of representatives of the City of Innsbruck and the University of Innsbruck.

We are grateful to the speakers who agreed to contribute to the special session: Fulvia de Finetti, Bruno de Finetti's daughter, who will give a historical account on *Bruno de Finetti, an Italian on the border*, Gert de Cooman, who will speak about *Exchangeability: A case study of how Bruno de Finetti's ideas thrive in indeterminate soil*, Paolo Vicig and Teddy Seidenfeld, who will venture into *Bruno de Finetti and Imprecision*, and Reinhard Viertl, who will collect historical relations of Bruno de Finetti with Austria and also talk about *Bruno de Finetti and fuzzy probability distributions*. The contributions of Fulvia de Finetti, Paolo Vicig and

Teddy Seidenfeld as well as a short abstract of the contribution of Reinhard Viertl are gathered in a special section of this volume, together with historical photographs from the collection of Fulvia de Finetti, with her kind permission.

During the conference two prizes will be awarded: the *Best Poster Award*, sponsored by Springer-Verlag, and the *IJAR Young Researcher Award*, granted by the International Journal of Approximate Reasoning.

We believe that, in the twelve 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. We think 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'11 provides a good platform to present and discuss work, and also leads to new ideas and collaborations.

Finally, we wish to thank several people for their support. Teddy Seidenfeld, the SIPTA President, regularly supported us with useful information and cheerful encouragement, and ensured that this conference benefits from previous experiences. In addition, he volunteered to chair the IJAR Award Committee. We thank Serafín Moral for his extensive and expert help in maintaining the electronic system and webpage of the conference. Thanks also to Serena Doria for joining the IJAR Award Committee.

We thank the members of the Program Committee for their excellent reviewing activities. Special thanks also to the Local Organizing Committee, in particular, to Anna Bombasaro, Bernhard Schmelzer and Reinhard Stix, as well as to Reinhold Friedrich for advice on matters of local organization. Thanks to Anton Bodner and Klaus Marcher of Studia-Verlag for their supportive handling of the publication of the proceedings. We thank all our sponsors; we are particularly grateful to the chair of the Center for Italian Studies of the University of Innsbruck, Barbara Tasser, and to Lukas Morscher of the Cultural Office of the City of Innsbruck for their support of the memorial tablet.

Finally, we thank all who have contributed to the success of ISIPTA'11, be it by submitting their research results, presenting them at the conference, or by attending sessions and participating in discussions. We hope that these proceedings will convey the state of the art of imprecise probability, raise interest and contribute to the further dissemination of the fascinating ideas of this active and highly relevant research field.

Frank Coolen
Gert de Cooman
Thomas Fetz
Michael Oberguggenberger

Innsbruck, July 2011

Organization, Supporters and Sponsors

Steering Committee

Frank Coolen, UK
Gert de Cooman, Belgium
Thomas Fetz, Austria
Serafín Moral, Spain
Michael Oberguggenberger, Austria
Teddy Seidenfeld, USA

Program Committee Board

Frank Coolen
Gert de Cooman
Thomas Fetz
Michael Oberguggenberger

Program Committee Members

Joaquín Abellán, Spain
Alessandro Antonucci, Switzerland
Horacio Arlo-Costa, USA
Thomas Augustin, Germany
Michael Beer, United Kingdom
Yakov Ben-Haim, Israel
Alessio Benavoli, Switzerland
Salem Benferhat, France
Dan Berleant, USA
Alberto Bernardini, Italia
Cassio Campos, Switzerland
Andrea Capotorti, Italy
Marco Cattaneo, Germany
Giorgio Corani, Switzerland
Inés Couso, Spain
Fabio Cozman, Brazil
Richard Crossman, UK
Fabio Cuzzolin, UK
Thierry Denoeux, France
Sebastien Destercke, France
Serena Doria, Italy
Didier Dubois, France
Love Ekenberg, Sweden
Scott Ferson, USA
Pablo Fierens, Argentina
Terrence Fine, USA
Christel Geiss, Austria
Stefan Geiss, Austria
Angelo Gilio, Italy
Michel Grabisch, France
Robert Hable, Germany
Jim Hall, UK
Manfred Jaeger, Denmark
Radim Jirousek, Czech Republic
Cliff Joslyn, USA
Erich Peter Klement, Austria
Igor Kozine, Denmark
Vladik Kreinovich, USA
Tomas Kroupa, Czech Republic
Jonathan Lawry, UK
Isaac Levi, USA
Enrique Miranda, Spain
Ilya Molchanov, Switzerland
Serafín Moral, Spain
Renato Pelessoni, Italy
Erik Quaeghebeur, Belgium
David Rios Insua, Spain
Fabrizio Ruggeri, Italy
Bernhard Schmelzer, Austria
Teddy Seidenfeld, USA
Damjan Skulj, Slovenia
Michael Smithson, Australia
Joerg Stoye, USA
Choh M. Teng, USA
Fulvio Tonon, USA
Matthias Troffaes, United Kingdom
Barbara Vantaggi, Italy
Jirina Vejnarova, Czech Republic
Paolo Vicig, italy
Nic Wilson, UK
Marco Zaffalon, Switzerland

Local Organizing Committee

Anna Bombasaro
Thomas Fetz
Michael Oberguggenberger
Bernhard Schmelzer
Reinhard Stix

Supporters and Sponsors

University of Innsbruck

Center for Italian Studies

Springer

Elsevier

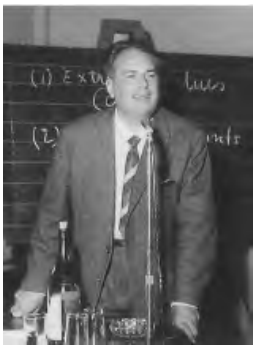
City of Innsbruck
Innsbruck and its Holiday Villages

Tiroler Sparkasse

Special Session
Bruno de Finetti

Bruno de Finetti, an Italian on the Border

Fulvia de Finetti
Rome, Italy
fulvia.definetti@teletu.it



The German translation of my work on probability means a lot to me because both my parents and grandparents were Italians but Austrian citizens. My father, engineer Walter von Finetti, planned and directed the construction of the Stubaitalbahn Innsbruck–Fulpmes, and I was born at that time in 1906 in Innsbruck where I lived for 5 years.

The first book I read on Probability was German: Czuber's "Wahrscheinlichkeitsrechnung".

Because of my attitude and my way of thinking Italians consider me a German. On the contrary Germans consider me Italian and in fact I feel so.

The conflicts between these two populations went on for many centuries and this should never be forgotten, but remembering it must never be bitter. On the contrary it must be an advice so that the tragic events of the past will not be repeated and will at most be heroically idealized like the Trojan wars. Both players: Andreas Hofer and Cesare Battisti and many others on the north and south of Brenner will not have died in vain because Independence and Rights of People were their common concern.

This is the preface written by Bruno de Finetti in 1981 for the German edition of his Theory of Probability. Probably somebody may find these words difficult to accept even today and probably it took him a whole life to arrive at writing these words.

On the border between two nations

If we analyze the 79 years of his life we discover that he spent 44 years in Innsbruck, Trento, Trieste and under Austro–Hungarian Empire, for the first 12 years of his life. The origins of the Finetti family seem to be found in Siena, but the von Finetti appears as a noble family in a draft dated 1672–1777. On December 17, 1770 Maria Theresa conferred in Vienna knighthood on one of the ancestors for merits deserved "in jure publico" and precisely for the tax reform she promoted.

When after the First World War the existing and functioning administration was changed to the inefficient Italian bureaucracy, patriots began to regret Austria in this respect.

Bruno was of course educated to love Italy and as he will recall, irredentism was especially alive in his grandmother Anna Radaelli, a niece of Carlo Alberto Radaelli, who participated in the defence of Venice in 1848–49. So the little Bruno who spoke both Italian and German started his personal war against Franz Joseph refusing to answer his German nurse when she spoke German to him.

In 1869 Anna Radaelli married Giovan Battista de Finetti, a civil engineer, member of the Association of Hungarian Engineers, working in Austria and Hungary at the railways Trieste–Fiume and Trieste–Pola. In the years 1880–1884 he worked for the Arlbergbahn. In the following years he will have worked mostly in Trieste. His first son (the father of Bruno) who was born in Fiume in 1871 studied in Innsbruck and then at the University of Graz becoming an engineer. In this way he learned a perfect German and could start working for the Ybbstalbahn. Then in 1899 he returned to Innsbruck and started working for the Stubaitalbahn. He became a friend of Francesco Menestrina a young man approximately of his same age, that had studied at Graz University and was appointed a professor of law at the newly opened Italian University in Innsbruck (1901). The day of his prolusion there were incidents caused by young

Austrians against the Italian University and confronted by Italian students coming from Graz headed by Cesare Battisti. Before being dismissed in 1904 he was visited by his sister Elvira who then met Walter de Finetti. They married in 1905.

The very day of the birth of Bruno his father started a diary. It gives us a very complete and detailed story of his physical and intellectual development but also states the attention paid by his parents to their son.

The five years spent in Innsbruck were the happiest for the family: they walked in the Hofgarten or along the Inn River to reach the theatre; sometimes they went to Trento and Trieste to meet Bruno's grandparents. In Trento Bruno was very much impressed by the big statue of Dante and he used to imitate his posture: for sure he knew the story of the statue and the meaning of the right hand pointing to Italy. In Trieste he saw the sea for the first time and easily learned how to swim. Once he was taken to Bruneck to give the first strike to the construction of one of the many railways that his father Gualtiero (Walter), an appreciated civil engineer working for the Joseph Riehl (1842–1907) enterprise operating in Tyrol, was going to build. It seems that Bruno took very seriously his job and that he would have liked to continue the excavation ... He was 4 years old when a Hungarian man travelling on the same train decided to take note of his name convinced that ... he will become a great man: "*Der wird ein großer Mann werden*".

In 1911 Gualtiero moved his family to Trieste to be near his parents who were becoming old but there he died in 1912. His wife Elvira, pregnant again, decided to move to Trento where her family lived to get their support. Bruno was admitted to the second class thanks to the many things he had learned from his father and he did very well in school.

Because of the First World War he had to leave Trento and the school and kept studying by himself. At the end of the war in 1919 he returned to Trento and was admitted to the third class of gymnasium. Owing to a very serious infection he had to be operated and he got one leg shortened by 7 centimetres. He was out of school for the whole year but kept in pace with the program by himself. Before he had just time to see the arrival in Trento of the tenth *Giro d'Italia* (Tour of Italy) with his idol Girardengo, and enrolled in the Boy Scouts Association headed by Giggino Battisti, the son of Cesare Battisti, the Italian martyr he admired both for his socialist ideals and for his fierceness at execution.

The economic situation of his family became even worse owing to the unfavourable exchange rate of Austrian crowns into liras. To gain one year Bruno studied in summer 1923 the program of the last year of high school and in October he passed the examination and immediately enrolled at *Politecnico di Milano* to become an engineer like his father and grandfather.

On the border of many branches of science

After finishing the first two years, he attended some lectures of Analysis and discovered to be more interested in the courses of the faculty of Mathematics. He immediately wrote a letter to his mother asking the permission to shift to Mathematics but he got a negative answer, she was worried about his future. Two more moving letters

> *... Mathematics is not by now a field already explored, just to learn and pass on to posterity as it is. It is always progressing, it is enriching and lightening itself, it is a lively and vital creature, in full development and just for these reasons I love it, I study it and I wish to devote my life to it ...*

did not have the desired effect. Bruno sent to his mother a very eloquent one-word cable

> *OBBEDISCO*

same answer given by Garibaldi to Vittorio Emanuele II in 1866 when ordered to stop the conquest of Trento. Sure the disappointment was the same but he stayed at the *Politecnico* for one more year.

It was during this third year that he wrote a work on population genetics that was examined by a biologist, a mathematician, a statistician and finally published in Metron in 1926. His first publication was immediately appreciated on the other side of the Atlantic Ocean:

> *I have noted with interest your important paper ...*

writes Alfred J. Lotka to "Professor" de Finetti who answered to be still a student.

The promise of a position in Rome at the Italian Central Statistical Institute founded and directed by Corrado Gini convinced his mother to give her permission, so Bruno graduated in Applied Mathematics in 1927 and immediately went to Rome accepting the promised job at the Italian Central Statistical Institute: it was too important for him to start earning to sustain his family. Rome was at that time a centre of attraction for scientific research and Bruno's hope was to have the opportunity to get in touch with it.

In fact, the three years he spent in Rome were the only ones for a long time when he could contact the big outstanding professors of the University of Rome like Enrico Fermi and his group of assistants at that time working at the experiments that would earn them the Nobel prize, like Guido Castelnuovo, who in a letter dated July 28, 1928 writes

> *I feel sure that you will be able to give important contributions to Probability Calculus and its applications*

and in September that same year Bruno would present *Funzione caratteristica di un fenomeno aleatorio* at the International Congress of Mathematicians held in Bologna. A summary of his presentation was published already in 1929 in the U.M.I. Bulletin, but the full version appeared in 1931 so this is why you celebrate this year the 80 years of his representation theorem.

This International Congress gave him the opportunity to meet many important foreign mathematicians, including, Jacques Hadamard, Maurice Fréchet, Aleksandr Khinchin, Paul Lévy, Jerzy Neyman, Octave Onicescu and George Polya. In 1929 Hadamard in a letter to Giulio Vivanti will write:

> *... je suis tout convaincu de son valeur. Je serai très heureux de le voir à Paris avec nous.*

With Fréchet the young Bruno had a polite dispute in the 30s that did not prevent him to be invited in Paris on May 1935 to give five lectures on probability at the Institut Poincaré.

In 1937 most of them will meet again in Geneva for the famous Colloquium on Probability.

Even if his job at the Central Statistical Institute did not completely satisfy him (at the end of 1929 he started to contact Assicurazioni Generali) the three years in Rome were decisive for his future ... also because there he met Renata, his future wife, and sure less important he became a fan of the Rome soccer team.

In 1931 he moved to Trieste and started working for the "Assicurazioni Generali", an insurance company. There he worked as an actuary and also on the mechanisation of some actuarial services. This probably contributed to make him one of the first mathematicians very aware of the possibilities offered by computing machinery. In the following years, he supplemented his work with several academic appointments, both in Trieste and Padua.

Then, starting from 1946, he dedicated himself to the academic activity as full professor at the University of Trieste, initially in the Faculty of Science and then in that of Economics. Even if World War II was over it was a very painful period of time for Trieste, that became a Free Territory ruled by the Allies while waiting to know the final destination. A condition particularly painful for my father worrying to become again an Italian citizen in a foreign country.

In 1950 Bruno got a Fulbright grant to visit the United States for three months. At this occasion he studied English with a young officer of the U.S. Army stationed in Trieste. He visited several places: in Cambridge, Massachusetts, at the International Congress of Mathematicians, in Berkeley at the second Berkeley Symposium to present a paper on *Recent suggestions for the reconciliation of theories of probability*. Neyman received him with great friendship and promoted his membership to the International Statistical Institute. Neyman was one of the three names; the others were Castelnuovo and Frechét who, beside Jimmy Savage, my father mentioned in his Farewell Lesson. At important occasions they gave him the possibility to explain his ideas even when in contrast to their own. This is what my father appreciated the most.

In 1954, he moved to the Faculty of Economics at *La Sapienza* University in Rome.

When in 1961 the Faculty of Science decided to resume the chair of Probability for him that had been created for Guido Castelnuovo but discontinued when he retired, the main concern of my father was that the same thing might happen when he would leave. Luckily that wasn't the case.

For his enthusiastic involvement in the teaching of mathematics he was appointed President of Mathesis and became Director of *Periodico di Matematiche* in 1972; he invited Polya for a conference and during the stay of Polya in Rome they prepared a documentary to teach mathematics at school. The protagonist was an animated pupil who got the name of *Giorgetto* (Little George) after George Polya. While Polya himself acted in the movie asking questions, Giorgetto animated by de Finetti answered by means of a succession of slides illustrating the steps to reach the solution.

Up to now I have mentioned his relationship with the mathematicians he met in Bologna, but it is time now to talk about another mathematician that I mentioned before and that he met on the occasion of the second Berkeley Symposium (1950): Jimmy Savage.

Recent suggestions for the reconciliation of theories of probability was the title of de Finetti's communication at the Symposium. I presume that Savage must have found something interesting and to better understand and deepen the ideas of Bruno he invited him to Chicago. Chicago was not a foreseen stop in Bruno's itinerary in USA, but to find somebody interested to discuss his ideas was an opportunity not to be lost because at that time there were not many people who paid attention to his view about probability. By the way this gave my father the pleasure to meet again Fermi and sadly enough that was also the last one.

That first encounter started an intense correspondence and frequent meetings. In 1957 de Finetti was again in Chicago as visiting Professor and this time also his family joined him. I remember how the Savages took care of us to make our stay as pleasant as possible. More often were the Savages to come to Europe especially for sabbatical years and Jimmy started to learn Italian to better communicate with my father. This gave rise to very amusing mistakes like for instance *carta bollata* (marked paper) becoming *carta bollita* (boiled paper). All contributed to create a very friendly

atmosphere between the two families and of course especially between Bruno and Jimmy. I remember their endless conversations and also our meeting in Bucharest in September 1971 at the Congress on Logic, Methodology and Philosophy of Science where Savage was an invited lecturer. The title of his talk was *Probability in Science: A Personalistic Account*. In Bucharest we met also Octav Onicescu, the founder of the Romanian school of probability theory and of the school of statistics. Onicescu and de Finetti first met in Rome at the beginning of their career when both lived there. Later they saw each other in 1937 in Geneva and again in Rome in the 60s.

Few months after the Congress in Bucharest the sudden news of the death of Savage came as a shock to my father, who lost the only person able to fully understand his view on probability and to adhere to it, and ended a twenty years long and fruitful correspondence.

In April 1973 my father received an invitation from the University of Michigan for the year 1973–74. I think it may be of interest to read part of the answer of my father declining the invitation:

> *. . . I am very pleased and honoured for such attracting invitation and for the interest in my research . . . and in my point of view about subjective probability. I would be surely willing to support it, especially in your University where L.J. Savage spent several years of his admirable activity . . . I am involved in many programs here, highly depending on myself (my collaborators are too young to be fully responsible for the courses).*

In the already mentioned 1976 Farewell Lesson, Bruno evaluates the importance of Savage for the acceptance of his ideas:

> *I must stress that I owe to him if my work is no longer considered a blasphemous but harmless heresy, but as a heresy with which the official statistical church is being compelled, unsuccessfully, to come to terms . . .*

It is also worth considering his vital interest in economics and social justice, as well as his struggle against bureaucracy.

Bruno de Finetti's interest in economics was innate and led him, during his first year at *Politecnico di Milano*, to attend the lectures given there by Ulisse Gobbi. These, in turn, confirmed him in his radical position, which he himself summarised as follows in an autobiographic note:

> *. . . the only directive of the whole of economics, freed from the damned game and tangle of individual and group egoisms, should always be the realisation of a collective Pareto optimum inspired by some criterion of equity.*

His longing for social justice caused him, in the 1970s, to be candidate in several elections and also arrested for his antimilitarist position. On the other hand, for his work in the field of economics in 1982 he was awarded a degree *honoris causa* in Economics by the LUISS University of Rome and received a broad international appraisal. In 1985 the Nobel Prize winner Franco Modigliani was asked which Italians would deserve the same prize, he indicated Paolo Sylos Labini and Bruno de Finetti.

More recently it came in the words of Mark Rubinstein:

> *it has recently come to the attention of economists in the English speaking world that among de Finetti's papers is a treasure trove of results in economics and finance written well before the work of the scholars that are traditionally credited with these ideas . . . de Finetti's 1940 paper anticipating much of mean variance portfolio theory later developed by Harry Markowitz.*

Markowitz himself, the 1990 Nobel Prize laureate in Economics and founder of modern finance recognized:

> *it has come to my attention that, in the context of choosing optimum reinsurance levels, de Finetti essentially proposed mean variance portfolio analysis using correlated risks.*

His last participation at an International Conference was the one on Exchangeability in Probability and Statistics, held in Rome in 1981 to honour his 75th birthday. At that occasion professor Reinhard Viertl who was born in Hall discovered that Bruno was born in Innsbruck and so devised to organize an International Symposium on Probability and Bayesian Statistics in Innsbruck to honour his 80th birthday in 1986. On January 1985 the first announcement arrived and my father filled in the form indicating he would submit a paper and he will be accompanied by *Frau* and *Tochter*. He could not maintain the promise; he died on July 20, 1985. My mother and I were there and the Symposium became in Memoriam of Bruno de Finetti.

The last time he was in Innsbruck was in 1973. He had to move to Vienna in August to present his paper *Bayesianism: its unifying role for both the foundations and the applications of statistics* at the Session of the International Statistical Institute. We decided to drive there by car and the first stop was in Trento to visit our relatives and then in Innsbruck. We saw the house in Adolf-Pichler-Straße and took the train to Fulpmes and then we were in Igls and went to Hungerburg, where at Easter 1911 Bruno got lost, and then to Hall, Salzburg, Lienz and finally Vienna, the Capital of the Austro–Hungarian Empire that for centuries had organized a fruitful synergy among multiple ethnics concurring in the commonwealth. For my father it was really a travel in the past.

Bruno de Finetti and Imprecision

Paolo Vicig
University of Trieste, Italy
paolo.vicig@econ.units.it

Teddy Seidenfeld
Carnegie Mellon University, USA
teddy@stat.cmu.edu

Abstract

We review several of de Finetti's fundamental contributions where these have played and continue to play an important role in the development of imprecise probability research. Also, we discuss de Finetti's few, but mostly critical remarks about the prospects for a theory of imprecise probabilities, given the limited development of imprecise probability theory as that was known to him.

Keywords. Coherent previsions, imprecise probabilities, indeterminate probabilities

1 Introduction

Researchers, especially members of SIPTA, approaching the theory of imprecise probabilities [IP] may easily deduce that Bruno de Finetti's ideas were influential for its development.

Consider de Finetti's foundational *Foresight* paper (1937), which is rightly included in the first volume of the series *Breakthroughs in Statistics* [16]. In that paper we find fundamental contributions to the now familiar concepts of *coherence* of subjective probabilities – having fair odds that avoid sure loss – and *exchangeable* random variables – where permutation symmetric subjective probabilities over a sequence of variables may be represented by mixtures of *iid* statistical probabilities. Each of these concepts is part of the active research agendas of many within SIPTA and have been so since the Society's inception. That is, we continue to see advances in IP that are based on novel refinements of coherence, and contributions to concepts of probabilistic independence as those relate also to *exchangeability*. For instance, 7 of 47 papers in the *ISIPTA'09 Proceedings* include at least one citation of de Finetti's work. And it is not hard to argue that another 7, at least, rely implicitly on his fundamental contributions.

Regarding origins of SIPTA, consider for instance Walley's book [42], nowadays probably the best known extensive treaty on imprecise probabilities. Key concepts like upper and lower previsions, their behavioural interpretation, the consistency notions of coherence and of previsions that avoid sure loss, appear at once as generalizations of basic ideas from de Finetti's theory. In the preface to [42], Walley acknowledges that

> *'My view of probabilistic reasoning has been especially influenced by the writings of Terrence Fine, Bruno de Finetti, Jack Good, J.M. Keynes, Glenn Shafer, Cedric Smith and Peter Williams'.*

In their turn, most of these authors knew de Finetti's theory, while Smith [36] and especially Williams [45] were largely inspired by it.

For another intellectual branch that has roots in de Finetti's work, consider contributions to SIPTA from Philosophy. For example, Levi [24, 25] generalizes de Finetti's decision–theoretic concept of *coherence* through his rule of *E–admissibility* applied with convex sets of credal probabilities and cardinal utilities.

However, a closer look at de Finetti's writings demonstrates that imprecise probabilities were a secondary issue in his work, at best. He did not write very much about them. In fact, he was rather skeptical about developing a theory based on what he understood IP to be about. To understand the incongruity between the incontrovertible fact that many SIPTA researchers recognize the origins for their work in de Finetti's ideas but that de Finetti did not think there was much of a future in IP, we must take into account the historical context in the first half of the last century, and the essentially marginal role in the scientific community of the few papers known at the time that treated imprecision by means of alternatives to precise probability.

Our note is organized as follows: In Section 2 we dis-

cuss de Finetti's viewpoint on imprecision. After reviewing some historical hints (Section 2.1), we summarize what we understand were de Finetti's thoughts on IP (Section 2.2). In Section 3 we respond to some of de Finetti's concerns about IP from the current perspective, i.e., using arguments and results that are well known now but were not so at the earlier time. We review some key aspects of the influence of de Finetti's thought in IP studies in Section 4. Section 5 concludes the paper.

2 Imprecise Probabilities in de Finetti's Theory

2.1 A Short Historical Note

De Finetti published his writings over the years 1926–1983, and developed a large part of his approach to probability theory in the first thirty years. In the first decade (1926–1936) he wrote about seventy papers, the majority on probability theory. At the beginning of his activity, measure–theoretic probability was a relatively recent discipline attracting a growing number of researchers. There was much interest in grounding probability theory and its laws (Kolmogorov's influential and measure–theoretic approach to probability was published in 1933), and few thought of other ways of quantifying uncertainty. Yet, alternatives to probability had already been explored: even in 1713, more or less at the origins of probability as a science, J. Bernoulli considered non-additive probabilities in Part IV of his *Ars Conjectandi*, but this aspect of his work was essentially ignored (with the exception of J.H. Lambert, who derived a special case of Dempster's rule in 1764 ([32], p. 76).

In the time between Bernoulli's work and the sixties of last century, some researchers were occasionally concerned with imprecise probability evaluations, but generally as a collateral problem in their approaches. Among them, de Finetti quotes (in [14], p. 133, and [15]) B.O. Koopman and I.J. Good, asserting that the introduction of numerical values for upper and lower probabilities was a specific follow–up of older ideas by J.M. Keynes [22].

Starting from the sixties, works focusing on various kinds of imprecise probabilities appeared with slowly increasing frequency. Their authors originally explored different areas, including non-additive measures (Choquet, whose monograph [2] remained virtually unknown when published in 1954 and was rediscovered several years later), Statistics [7], Philosophy [23, 24, 37, 41], robustness in statistics [20, 21], belief functions [32]. See e.g. [19] for a recent historical note.

Among these, de Finetti certainly read two papers which referred to his own approach, [36] and [45]. While Smith's paper [36] was still a transition work, Williams' [45] technical report stated a new, in-depth theory of imprecise conditional previsions, which generalized de Finetti's betting scheme to a conditional environment, proving important results like the envelope theorem. De Finetti's reaction to Smith's paper was essentially negative and, as he explained, led to the addition of two short sections in the final version of [14]. We discuss de Finetti's reactions below.

As for Williams' paper, de Finetti read it in a later phase of his activity, the mid-seventies, and we are aware of no written comments on it. However Williams commented on this very point many years later, in an interview published in *The SIPTA Newsletter*, vol. 4 (1), June 2006. In his words:

> *De Finetti himself thought the 1975 paper was too closely connected to "formal logic" for his liking, which puzzled me, though he had expressed interest and pleasure in the earlier 1974 paper linking subjective probability to the idea of the indeterminacy of empirical concepts.*

Throughout his career de Finetti proposed original ideas that were often out of the mainstream. For example, he championed the use of finite additivity as opposed to the more restrictive, received theory of countably additive probability, both regarding unconditional and conditional probability. Criticism from the prevailing measure theoretic approach to probability often dubbed finitely additive subjective probability as arbitrary. It might have been too hard to spread the even more innovative concepts of imprecise probabilities. This may be a motivation for de Finetti's caution towards imprecise probabilities. It certainly contributes to our understanding why Williams' report [45] was published [46] only in 2007, more than thirty years later. (See [40].)

2.2 Imprecision in de Finetti's Papers

In very few places in his large body of written work does de Finetti discuss imprecise probabilities, and nowhere does he do so exclusively. Discussions of some length appear in [12, 14, 15]. De Finetti's basic ideas on imprecision appear already in the philosophical, qualitative essay [12] *Probabilismo. Saggio critico sulla teoria delle probabilità e sul valore della scienza*, which de Finetti quotes in his autobiography in [17] as the first description of his viewpoint on probability. In this paper, he acknowledges that an agent's opinion on several events is often determined up to a very

rough degree of approximation, but observes that the same difficulty arises in all practical problems of measuring quantities (p. 40). He then states (p. 41) that under this perspective probability theory is actually perfectly analogous to any experimental science:

> *In experimental sciences, the world of feelings is replaced by a fictitious world where quantities have an exactly measurable value; in probability theory, I replace my vague, elusive mood with that of a fictitious agent with no uncertainty in grading the degrees of his beliefs.*

Continuing the analogy, shortly after (p. 43) he points out a disadvantage of probability theory, that

> *measuring a psychological feeling is a much more vaguely determined problem than measuring any physical quantity,*

noting however that just a few grades of uncertainty might suffice in many instances. On the other hand, he observes that the rules of probability are intrinsically precise, which allows us to evaluate the probability of various further events without adding imprecision.

In an example (p. 43, 44, abridged here), he notes that $P(A \wedge B) = P(A|B)P(B)$ is determined precisely for an agent once $P(A|B)$ and $P(B)$ are determined. By contrast, when starting from approximate evaluations like $P(B) \in [0.80, 0.95]$ and $P(A|B) \in [0.25, 0.40]$, imprecision propagates. Then $P(A \wedge B)$ can only be said to lie in the interval $[0.80 \cdot 0.25 = 0.20, 0.95 \cdot 0.40 = 0.38]$.

If B is the event: the doctor visits an ill patient at home, and A: the doctor is able to heal the ill patient, approximate evaluations – he notes – are of little use, as they do not let us conclude much more than the following merely qualitative deduction, which we paraphrase: If it is nearly sure that the doctor will come, and fairly dubious that he can heal his patient, then it is slightly more dubious that the doctor comes and heals his patient.

Further, de Finetti notes that probabilities can often be derived from mere *qualitative opinions*. For instance, in many games the atoms of a finite partition are believed to be equally likely. This remark suggests a reflection on the role of qualitative uncertainty judgements in de Finetti's work. Interestingly, he displayed a different attitude towards this definitely more imprecise tool than to imprecise probabilities. In fact, in the same year 1931 he wrote *Sul significato soggettivo della probabilità* [13], discussing rationality conditions, later known as *de Finetti's conditions*, for comparative (or qualitative) probabilities, showing their analogy with the laws of numerical probability. This paper pointed out what became an important research topic, concerning existence of agreeing or almost agreeing probabilities for comparative probability orderings. (See [18] for an excellent review.)

The ideas expressed in [12] were not substantially modified in later writings. For instance, in [14], p. 95, de Finetti and Savage quote E. Borel as sharing their thesis, that

> *the vagueness seemingly intrinsic in certain probability assessments should not be regarded as something qualitatively different from uncertainty in any quantities, numbers and data one works with in applied mathematics.*

The jointly authored 1962 paper [14], *Sul modo di scegliere le probabilità iniziali*, adds some arguments to de Finetti's ideas on imprecise probabilities while discussing Smith's then recently published paper [36]. Recall that Smith proposed a modification of de Finetti's betting scheme, introducing a one–sided lower probability $\underline{P}(A)$ and a one–sided upper probability, $\overline{P}(A) \geq \underline{P}(A)$, for an event A, rather than a single two–sided probability, as we explain next. In Smith's approach, the agent judges a bet on A (winning 1 if and only if A obtains) at a price $p < \underline{P}(A)$ to be favorable over the status quo, which has 0 payoff for sure. Such a favorable gamble has a positive lower expected value, hence greater than 0. And for the same reason the agent prefers to bet against A (paying 1 if and only if A obtains) in order to receive a price $p > \overline{P}(A)$ over the status quo. For prices p between the lower and upper probability, $\underline{P}(A) \leq p \leq \overline{P}(A)$, the agent is allowed to abstain from betting and remain with the status quo.

In de Finetti's theory, by contrast, the agent is obliged to give one two-sided probability $P(A)$ for betting on/against the event A. At the *fair* price $p = P(A)$ the de–Finetti–agent is indifferent between a gamble on/against A and abstaining, and may either accept or reject the bet. For prices $p < P(A)$ the de–Finetti–agent judges a bet on A favorable, etc. Thus, de Finetti's theory is the special case of Smith's theory when $P(A) = \overline{P}(A) = \underline{P}(A)$, modulo the interpretation of how the agent may respond to the case of a *fair* bet.

After expressing perplexity about the idea of avoiding stating *one* exact fair value $P(A)$ by introducing an indecision interval $I = [\underline{P}(A), \overline{P}(A)]$, with *two* different exact (one-sided) values as endpoints, de Finetti and Savage focus on two questions: *first*, existence of

the indecision interval I and *second*, consistency of the agent's betting using the interval I.

As for the first question, de Finetti and Savage agree that nobody is *actually* willing to accept all of the bets required according to the idealized version of de Finetti's coherence principle. They concede that the betting model introduced by de Finetti in order to give an operational meaning to *subjective probability* requires that an *idealized*, rational agent is obliged to have a real–valued probability $P(A)$ and, thus, to accept bets at favorable odds – betting on A for any price less than $P(A)$ and betting against A for any price greater than $P(A)$.[1] The real agent is committed to behave according to the idealized theory in *hypothetical circumstances* where he/she has reflected adequately on the problem. In other words, de Finetti's opinion, expressed on this point also in other papers, seems to be that the betting scheme should not be taken literally. Rather it is a way of defining the subjective probability concept in idealized circumstances. Hence, intervals of indecision exist in practice, but only where the real decision agent has not thought through the betting problem with the precision asked of the idealized agent.

As for the second question, de Finetti and Savage argue that, rather than allowing the indecision interval, from the perspective of coherence it may be better to employ the precise two–sided probability $P = (\overline{P} + \underline{P})/2$. They report the following intriguing example as evidence for their view.

Example (de Finetti and Savage, 1962, p. 139). An agent may choose whether to buy or not any combination of the following 200 tickets involving varying gambles on/against event A. The first 100 tickets are offered for prices, respectively, of 1, 2,..., 100 Euros[2] and each one pays 100 Euros if event A occurs, and 0 otherwise. The remaining 100 tickets are offered, respectively, at the same prices but on the complementary event, A^c. Each of these 100 tickets pays 100 Euros if A^c occurs and 0 otherwise. If the agent assesses a two–sided personal probability for A as in de Finetti's theory, e.g., $P(A) = 0.63$, he/she will maximize expected value by buying the first 63 tickets on A with prices 1,..., 63, for a combined price $1 + 2 + \ldots + 63 = 2016$ Euros, and buying the first 37 tickets on A^c for a combined price $1 + 2 + \ldots + 37 = 703$ Euros. (The agent is indifferent about buying the 63^{rd} ticket from the first group and, likewise, the 37^{th} ticket from the second group.) The agent's total expense for the 100 tickets, then, is 2719 Euros. The agent gains $6300 - 2719 = 3581$ Euros if A occurs; he/she gains 981 Euros otherwise, when A^c occurs.

Suppose, instead the agent fixes a lower probability $\underline{P}(A) = 0.53$ and an upper probability $\overline{P}(A) = 0.73$, as allowed by Smith's theory. De Finetti and Savage interpret this to mean that the Smith-agent will buy only the first 53 tickets for A and only the first 27 tickets for A^c – those gambles that are individually (weakly) *favorable*. Then the Smith–agent will gain only $5300 - 1809 = 3491$ Euros if A occurs, and will gain only $2700 - 1809 = 891$ Euros if A^c occurs. Their conclusion is that in this decision problem it is better for the agent to assess the real–valued, two–sided probability $0.63 = P(A) = (\overline{P}(A) + \underline{P}(A))/2$ than to use the interval $I = [0.53, 0.73]$. The decision maker's gain increases by 90 Euros, whatever happens, using this two-sided, de Finetti–styled probability. We respond to this example in the next section. □

De Finetti and Savage continue their criticism of IP theory on pp. 140 ÷ 144 of [14]. To our thinking, the most interesting argument they offer is perhaps that imprecision in probability assessments does not give rise to a new kind of uncertainty measure, but rather points out an incomplete elicitation by a third party and/or even incomplete self–knowledge. They write,

> *Even though in our opinion they are not fit for characterizing a new, weaker kind of coherent behaviour, structures and ideas like Smith's may allow for important interpretations and applications, in the sense that they elicit what can be said about a behaviour when an incomplete knowledge is available of the opinions upon which decisions are taken.*

They continue with a clarifying example.

> *What is the area of a triangle with largest side a and shortest side b? Any S such that $\underline{S} \leq S \leq \overline{S}$, with $\underline{S}$: area of the triangle with sides (a, b, b), $\overline{S}$: area of the triangle with sides (a, a, b). This does not mean: there exists a triangle whose area is indeterminate ($\underline{S}$: lower area, $\overline{S}$: upper area); every triangle has a well determined area, but we might at present be unable to determine it for lack of sufficient information.*

In the Appendix of [15], while mainly summarizing ideas on imprecise probabilities already expressed in [12, 14], de Finetti adds other examples supporting the same thesis. One is particularly interesting because it does not resort to the analogy between probabilities and other experimental measures but involves his *Fundamental Theorem of Prevision*. As well

[1] As recalled in [14], such agents were termed *Stat Rats* (by G.A. Barnard) in the discussion of [36].

[2] We introduce an anachronism, here and in later examples, updating the monetary unit to 2011.

known, that theorem ensures that, given a coherent probability function $P(\cdot)$ defined on an arbitrary set of events $\mathcal{D}$, all of its coherent extensions that include a probability for an additional event $E \notin \mathcal{D}$ belong to a non-empty closed interval $I_E = [\underline{P}(E), \overline{P}(E)]$. This interval I_E of potential (coherent) values for $P(E)$ is defined by analogy with how one may extend a measure μ to give a value for a non-measurable set using the interval of inner and outer measure values. In de Finetti's theorem, the interval I_E arises by approximations to E (from below and from above) using events from the linear span of $\mathcal{D}$. But, de Finetti argues, the fact that prior to the extension, we can only affirm about $P(E)$ that it belongs to I_E rather than having a unique value

> *does not imply that some events like E have an indeterminate probability, but only that $P(E)$ is not uniquely defined by the starting data we consider.*

De Finetti's thinking about imprecise personal probability is unchanged from his early work. In his classic ([31], p. 58) Savage quotes de Finetti's [16] view on this question.

> *The fact that a direct estimate of a probability is not always possible is just the reason that the logical rules of probability are useful.*

Revealing of Savage's subsequent thinking on this question of existence of unsure, or imprecise (personal) probabilities is the footnote on p. 58, added for the 1972 edition of [31], where Savage teases us with these guarded words.

> *One tempting representation of the unsure is to replace the person's single probability measure P by a set of such measures, especially a convex set. Some explorations of this are Dempster (1968), Good (1962), and Smith (1961).*

3 Rejoinder from the Perspective of 2011

Many of the objections raised by de Finetti (and others) towards the use of imprecise probabilities have been discussed at length elsewhere. (See especially [42], Secs. 5.7, 5.8, 5.9). Of course, some recently formulated arguments in favor of IP, e.g., some relating to group decision making [34] or IP models for frequency data [10], were not anticipated by de Finetti. Here, we offer brief comments, including responding to the challenges against IP raised in the previous section.

The first of de Finetti's arguments supporting precise rather than imprecise probabilities is roughly that – barring e.g., Quantum Mechanical issues – ordinary theoretical quantities that are the objects of experimental measurement are precise. In practice however, when the process for eliciting a precise personal probability is not sufficiently reliable, impractical, or too expensive, the use of imprecise probabilities seems appropriate. By modeling the elicitation process, e.g., by considering psychometric models of introspection, we may be able to formalize the degree of imprecision of the assessment [27]; a first, intuitive measure of imprecision is of course the difference $\overline{P}(A) - \underline{P}(A)$.

De Finetti hits the mark with his second observation, basically that inferences with imprecise probabilities may be highly imprecise. This is unquestionably true, but there are different levels: highly imprecise measures like possibilities and necessities typically ensure many vacuous inferences [44], while standard, less imprecise instruments are (now) available in other instances, e.g., the Choquet integral for 2–monotone measures [3], the imprecise Dirichlet distribution [43], etc..

De Finetti and Savage's [14] example, which we summarized in Section 2.2, merits several responses. First, it is not clear what general claim they make. Are they suggesting that a decision maker who uses Smith's lower and upper IP betting odds *always* makes inferior decisions compared with some de Finetti–styled decision maker who uses precise betting odds but has no other advantage – no other special information? Is their claim instead that *sometimes* the IP decisions will be inferior? What is their objection?

De Finetti and Savage's example uses particular values for $P, \overline{P}$, and $\underline{P}$, combined with a controversial (we think unacceptable) interpretation of how the IP decision maker chooses in their decision problem. It is not difficult to check that the same conclusion they reach may be achieved by varying the three quantities $P, \overline{P}$, and $\underline{P}$ subject to the constraint that $\underline{P} < P < \overline{P}$ and these belong to the set $\{0, 1/100, 2/100, \ldots, 1\}$ while retaining the same ticket prices, and the same seemingly myopic decision rule for determining which tickets the IP decision maker purchases. That is, it appears to us that what drives de Finetti and Savage's result in this example is the tacit use of a decision rule that is invalid with sets of probabilities but which is valid in the special case of precise probabilities.

We think they interpret Smith's lower and upper betting odds to mean that when offered a bet on or against an event A at a price between its lower and

upper values, the IP decision maker will reject that option *regardless what other (non-exclusive) options are available*. That is, we think they reason that, because at odds between the lower and upper probabilities it is not favorable to bet either way on A *compared with the one option to abstain*, therefore the IP decision maker will abstain, i.e. not buy such a ticket in their decision problem.

The familiar decision rule to reject as inadmissible any option that fails to maximize expected utility reduces to pairwise comparisons between pairs of acts when the agent uses a precise probability. That is, in the example under discussion where utility is presumed to be linear in the *numeraire* used for the gambles[3], a de Finetti–styled decision maker will maximize expected utility by buying each ticket that, *by itself*, has positive expected value: Buy each ticket that in a pairwise comparison with abstaining is a favorable gamble and only those. But this rule is not correct for a decision maker who uses sets of probabilities. De Finetti and Savage's conclusion about which tickets the IP decision maker will buy is incorrect when she/he uses an appropriate decision rule.

As members of SIPTA know, there is continuing debate about decision rules for use with an IP theory. However, for the case at hand, we think it is non-controversial that the IP decision maker will judge inadmissible any combination of tickets that is *simply dominated* in payoff by some other combination of tickets. That is, in the spirit of de Finetti's coherence condition, particularly as he formulates it with Brier score, the decision maker will not choose an option when there is a second option available that simply dominates the first. Then, in this example, it is permissible for such an IP decision maker to buy the very same combination of tickets as would any de Finetti–styled decision maker who has a precise personal probability for the event A. That is because, in this finite decision problem, all and only Bayes–admissible options are undominated. Thus, it is impermissible for the IP decision maker to buy only the $80 = (53 + 27)$ tickets that de Finetti and Savage allege will be purchased.

Call *House* the vendor of the 200 tickets. *House* is clearly incoherent. In fact, an agent can make arbitrage without needing to consider her/his uncertainty about the event A: buying the first 50 tickets for A and the first 50 for A^c produces a sure gain of 2450 Euros! See [35] for different indices for the degree of incoherence displayed by *House*, what strategies maximize the sure gains that can be achieved against *House*, and how these are related to different IP models for the events in question.

There is a related point about IP-coherence that we think is worth emphasizing. Consider making a single bet in favor of A. If the decision maker adopts a precise probability $P(A)$, her/his gain per Euro staked on a bet on A will be $G = A - P(A)$. However, if the decision maker's judgment is unsure and she/he uses Smith's lower betting odds with $\underline{P}(A) < P(A)$, her/his gain increases to $\underline{G} = A - \underline{P}(A) > G$. It is true that in this latter case the decision maker will abstain from betting when the price for A is higher than $\underline{P}$ and lower than P, *and provided there are no other options to consider*. But this results only in the loss of some additional *opportunities* for gambling. There is no loss of a *sure gain*.

The role of the Fundamental Theorem in relation to IP theory is also of worth discussing. Let us accept de Finetti's interpretation of the interval I_E as giving all coherent extensions of the decision maker's current probability $P(\cdot)$, defined with respect to events in the set $\mathcal{D}$, in order to include the new event E. Suppose, however, that we consider extending P to include a second additional event F as well. To use the Fundamental Theorem to evaluate probability extensions for both E and F we must work step–by–step. Extend $P(\cdot)$ to include only one of the two events E or F using either interval I_E or I_F defined with respect to the set $\mathcal{D}$. For instance, first extend P to include a precise value for $P(E)$ taken from I_E. Denote the resulting probability $P^E(\cdot)$ defined with respect to the set $\mathcal{D} \cup \{E\}$. Then iterate to extend $P^E(\cdot)$ to include a precise value for $P^E(F)$. Of course, the two intervals I_F and I_F^E usually are not the same. We state without demonstration that, nonetheless, if the step–by–step method allows choosing the two values $P(E) = c$ and $P^E(F) = d$, then it is possible to reverse the steps to achieve the same pair, $P(F) = d$ and $P^F(E) = c$. Then the order of extensions is innocuous.

If instead we interpret the starting coherent probability P (defined on the linear span of $\mathcal{D}$) as a special coherent lower probability, and look for a lower probability which coherently extends it, we can *avoid the step–by–step procedure*, simply by always choosing the lower endpoint from the intervals based on the common set $\mathcal{D}$ and using these as 1-sided lower probabilities. We obtain what Walley [42] calls the *natural extension* of P, interpreted as a coherent lower probability (actually, it is even n–monotone) on all additional events. The correctness of such a procedure depends also on the *transitivity* property of the natural extension.

[3] Linearity of utility is no real restriction, because coherence is equivalent to *constrained coherence*, where an arbitrary upper bound $k > 0$ is set *a priori* on the agent's gains/losses in absolute value (see [30], Sec. 3.4). Just choose k such that the utility variation is to a good approximation linear.

There is a second consideration relevant to de Finetti's preferred interpretation of the interval I_E from the *Fundamental Theorem* relating to IP theory, which is particularly relevant in the light of Levi's [26] distinction between *imprecision* and *indeterminacy* of interval–valued probabilities. Levi's distinction is illustrated by Ellsberg's well known challenge [9].

In Ellsberg's puzzle [9] the decision maker faces decisions under risk and decisions under uncertainty simultaneously. The decision maker contemplates two binary choices: *Problem I* is a choice between two options labeled *1* and *2*, and *Problem II* is a choice between two options labeled *3* and *4*. The payoffs for these options are determined by the color of a randomly drawn chip from an urn known to contain only red, black, or yellow chips.

In *Problem I*, *option 1* pays off 1,000 Euros if the chip drawn is red, 0 Euros otherwise, i.e. if it is black or yellow. *Option 2* pays off 1,000 Euros if the chip drawn is black, 0 Euros otherwise, i.e, if the chip is red or yellow. In *Problem II*, *option 3* pays off 1,000 Euros if the chip drawn is either red or yellow, 0 if it is black. *Option 4* pays off 1,000 Euros if the chip drawn is black or yellow, 0 Euros if it is red. In addition, the urn is stipulated to contain exactly $1/3^{rd}$ red chips, with unknown proportions of black and yellow other than that their total is $2/3^{rds}$ the contents of the urn. Thus, under the assumptions for the problem, options *1* and *4* have determinate risk: they are just like a Savage *gamble* with determinate (personal) probabilities for their outcomes. However Ellberg's conditions leave options *2* and *3* as ill–defined gambles: the personal probabilities for the payoffs are not determined.

Across many different audiences with varying levels of sophistication, the modal choices are option *1* from Problem I and option *4* from Problem II. Assuming that the agent prefers more money to less, that there is no *moral hazard* relating the decision maker's choices with the contents of the urn, and that the choices reveal the agent's preferences, there is no expected utility model for the modal pattern, *1* over *2* and *4* over *3*.

In a straightforward IP–de–Finetti representation of this puzzle, the decision maker has a precise probability for the events $\{red,\ black\ or\ yellow\}$: $P(red) = 1/3$, $P(black\ or\ yellow) = 2/3$. But the agent's uncertainty about black or yellow is represented by the common intervals $I_{black} = I_{yellow} = [0, 2/3]$. Under these circumstances the agent's imprecise probabilities do not dictate the choices for either problem. However, if after reflection the agent decides for option *1* over option *2* in Problem I, then (as in the *Fundamental Theorem*) this corresponds to an extension of $P(\cdot)$ where now $P(black) < 1/3$. But then $P(yellow) > 1/3$ and option *3* has greater expected utility than option *4* relative to this probability extension. Likewise, if the agent reflects first on Problem II and decides for option *4* over option *3*, this corresponds to an extension of $P(\cdot)$ where now $P(yellow) < 1/3$. Then in Problem I option *2* has greater expected utility than option *1*.

In short, under what we understand to be de Finetti's favored interpretation of the *Fundamental Theorem*, the modal Ellsberg choices are anomalous. They cannot be justified even when the agent uses the uncertainty intervals from the *Fundamental Theorem*. Levi calls this a case of *imprecise* probability intervals. Under this interpretation the agent is committed to resolving her/his uncertainty with a coherent, precise probability.

By contrast, if the agent uses the two intervals, $I_{black} = I_{yellow} = [0, 2/3]$, to identify a set of probabilities for the two events, then relative to this set neither option in either Problem is ruled out by considerations of expected utility. That is, in Problem I, for some probabilities in the set, option *1* has greater expected utility than option *2*, and for other probabilities in the set this inequality is reversed. Likewise with the two options in Problem II. If the non–comparability between options by expected utility is resolved through an appeal to lower expected utility, e.g., as a form of security, then in Problem I the agent chooses option *1* and in Problem II the agent chooses option *4*. This is what Levi means by saying that the decision maker's IP is an *indeterminate* (not an *imprecise*) probability. With indeterminate probability, the agent is not committed to resolving uncertainty with a precise probability prior to choice.

4 De Finetti's Theory in Imprecise Probabilities

Let us repeat a simple fact. Notwithstanding what we see as de Finetti's mostly unsupportive opinions on imprecise probabilities, in the sense of IP as that is used by many in SIPTA, our co-researchers in this area find it appropriate to refer to his work in the development of their own. One reason for this is that many within SIPTA use aspects of de Finetti's work on personal probability which often are in conflict with the more widely received but less general, classical theory, associated with Kolmogorov's measure theoretic approach.

Take for instance de Finetti's concept of a coherent *prevision* $P(X)$ of a (bounded) random quantity X,

which is a generalization of a coherent probability. That special case obtains when X is the indicator function for an event, and then a prevision is a probability.

A prevision may be viewed as a finitely additive *expectation* $E(X)$ of X. But there are non-trivial differences between de Finetti's concept of *prevision* and the more familiar concept of a mathematical expectation as that is developed within the classic measure theoretic account. In order to determine the classical expectation of a random variable X, we first have to assess a probability for the events $\{\omega : X(\omega) = x\}$, or at least assess a density function. In uncountable state spaces, common with familiar statistical models, the classical theory includes measurability constraints imposed by countable additivity. But this is not at all necessary for assessing a prevision, $P(X)$, which may be determined directly within de Finetti's theory free of the usual measurability constraints. The difference may seem negligible, but it becomes more appreciable when considering previsions for several random quantities at the same time, and by far more so when passing to imprecise previsions, where additivity in general no longer applies. This is an illustration of how de Finetti's foundational ideas can become *more important* in IP theory than they are even in traditional probability theory.

The problem reiterates within the theory of conditional expectations, magnified by the fact that finitely additive conditional expectations do not have to satisfy what de Finetti called *conglomerability*, first in his 1930 paper *Sulla proprietà conglomerativa delle probabilità subordinate* [11]. Assume that $P(\cdot)$ is a coherent unconditional probability. Let $\pi = \{h_1, \ldots\}$ be a denumerable partition, and let $\{P(\cdot|h_i) : i = 1, \ldots\}$ be a set of corresponding coherent conditional probability functions for P, given each element of π. With respect to an event E, define $m_E = \inf_{h\in\pi} P(E|h)$, and $M_E = \sup_{h\in\pi} P(E|h)$. These conditional probabilities for event E are *conglomerable* in π provided that $P(E) \in [m_E, M_E]$. Schervish et al. [33] establish that each finitely but not countably additive probability fails to be conglomerable for some event E and denumerable partition π. Also, they identify the greatest lower bound for the *extent of non–conglomerability* of P, where that is defined by the supremum difference between the unconditional probability $P(E)$ and the nearest point to the interval $[m_E, M_E]$, taken over all denumerable partitions π and events E.

The treatment of conglomerability in IP is still controversial. While Walley [42] imposes some conglomerability axioms to his concepts of coherence for conditional lower previsions, Williams' more general approach does not. In Walley's words ([42], p. 644)

> *Because it* [...] *does not rely on the conglomerative principle, Williams' coherence is also a natural generalization of de Finetti's (1974) definition of coherence.*

See [29], Secs. 3.4, 4.2.2 for a further discussion of [11], Williams' coherence and of some arguments in favor/against conglomerativity in IP theory.

Also de Finetti's use of a generalized betting scheme to define coherent previsions serves as an example for several subsequent variants, which underly many uncertainty measures. Examples include *coherent upper and lower previsions* [45, 42], *convex previsions* [30], and *capacities* ([1], Sec. 4). Moreover, in all such instances this approach based on de Finetti's theory of previsions provides vivid, immediate interpretations of basic concepts and often relatively simple proofs of important results.

Another issue, which was our focus in the previous section, concerns de Finetti's attention to extension problems, i.e. to the existence of at least one coherent extension of a coherent prevision, defined on an arbitrary set of (bounded) variables. Walley [42] used this idea in the realm of imprecise probabilities to define several useful notions: a *natural* extension; a *regular* extension; an *independent* extension, etc. For instance, a natural extension is the largest, i.e., "least committal" coherent IP extension.

In general, research in IP theory exposes new facets of probability concepts already discussed and sometimes not quite fixed by de Finetti. An illustration is with the notion of *stochastic independence*, which de Finetti found unconvincing in its classical identification with the factorization property, but which he left somewhat undeveloped in his own work. In [15] he gives an epistemically puzzling example of two random quantities that are functionally dependent *and* stochastically independent according to the factorization property. Problems for a theory of independence arise especially when conditioning on events of extreme (0 or 1) probability. For instance, Dubins' version [8] of de Finetti's theory leads to an asymmetric relevance relation. The situation is more complex in the IP framework, and de Finetti would perhaps be surprised at the variety of independence concepts that have been developed. (See, e.g., [5, 6, 38, 39]).

De Finetti discovered important connections between *independence* and *exchangeability* as reported in his *Representation Theorem*, 1937. IP generalizations are being developed, e.g., [4]. Soon, will we see IP generalizations of *partial exchangeability* along the same lines. In yet other settings, IP methods have been employed to achieve advances in probability problems to which de Finetti himself contributed [28].

5 Conclusions

We close our comments with this metaphor, which will be entirely familiar to any parent. You raise your children with an eye for the day when each becomes an independent agent. Sometimes, however, contrary to your advice, one embarks on what you fear is an ill conceived plan. When to your great surprise the plan succeeds, does not that offspring then make you a very proud parent?!

Acknowledgements

We thank the ISIPTA'11 Program Committee Board for the opportunity to present our views on how de Finetti saw imprecise probability theory. Paolo Vicig wishes to thank his former teachers, and colleagues, L. Crisma and A. Wedlin for many fruitful discussions on subjective probability, and to acknowledge financial support from the PRIN Project 'Metodi di valutazione di portafogli assicurativi danni per il controllo della solvibilità'. Teddy Seidenfeld thanks two of his teachers, H.E. Kyburg, Jr. and I. Levi, for having introduced him to de Finetti's *"Book"* argument concerning coherence of Bayesian previsions. He appreciates these two experts' numerous debates using, respectively, *modus tollens* and *modus ponens*, about how best to connect de Finetti's premises with his conclusions!

References

[1] P. Baroni, R. Pelessoni and P. Vicig. Generalizing Dutch risk measures through imprecise previsions. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems*, 17:153–177, 2009.

[2] G. Choquet. Theory of capacities. *Ann. Inst. Fourier* 5:131–295, 1954.

[3] G. de Cooman, M. C. M. Troffaes and E. Miranda. n-Monotone lower previsions. *Journal of Intelligent & Fuzzy Systems*, 16: 253–263, 2005.

[4] G. de Cooman, E. Quaeghebeur and E. Miranda. Exchangeable lower previsions. *Bernoulli* 15:721–735, 2009.

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

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

[7] A.P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.* 38:325–339, 1967.

[8] L.E. Dubins. Finitely Additive Conditional Probabilities, Conglomerability and Disintegrations. *Ann. Prob.* 3:89–99, 1975.

[9] D. Ellsberg. Risk, Ambiguity, and the Savage Axioms. *Quarterly J. of Economics* 75:643–669, 1961.

[10] P.I. Fierens and T.L. Fine. Towards a Frequentist Interpretation of Sets of Measures. *Proceedings of ISIPTA'01*, 179–187, 2001.

[11] B. de Finetti. Sulla proprietà conglomerativa delle probabilità subordinate. *Rendiconti R. Istituto Lombardo di Scienze e Lettere* 43:339–343, 1930.

[12] B. de Finetti. *Probabilismo. Saggio critico sulla teoria delle probabilità e sul valore della scienza.* In: *Biblioteca di filosofia*, 1–57. Libreria editrice Perrella, Napoli, 1931.

[13] B. de Finetti. Sul significato soggettivo della probabilità. *Fundamenta Mathematicae*, 17:298–329, 1931.

[14] B. de Finetti and L.J. Savage. Sul modo di scegliere le probabilità iniziali. *Biblioteca del Metron*, Ser. C, Vol. I:81–154, 1962.

[15] B. de Finetti. *Teoria delle probabilità*. Einaudi, Torino, 1970 (English translation: B. de Finetti. *Theory of Probability*, Wiley, 1974).

[16] B. de Finetti [1937] *Foresight: Its Logical Laws, Its Subjective Sources.* (H.E. Kyburg, Jr. translation.) In: *Breakthroughs in Statistics* (S. Kotz and N.J. Johnson eds.), 1:134-174, Springer–Verlag, N.Y., 1993.

[17] B. de Finetti. *Scritti (1926–1930).* Cedam, Padova, 1981.

[18] P. Fishburn. The Axioms of Subjective Probability (with discussion). *Stat. Sci.* 1:333–358, 1984.

[19] F. Hampel. *Nonadditive Probabilities in Statistics.* In: *Imprecision in Statistical Theory and Practice* (P. Coolen–Schrijner, F. Coolen, M. Troffaes, T. Augustin and S. Gupta eds.), 13–25, Grace Scientific Publ., Greenboro, USA, 2009.

[20] P.J. Huber, V. Strassen. Minimax tests and the Neyman–Pearson lemma for capacities. *Annals of Statistics* 1:251–263, 1973.

[21] P.J. Huber. *Robust Statistics.* Wiley, N.Y., 1981.

[22] J.M. Keynes. *A Treatise on Probability*. Macmillan and Co., London, 1921.

[23] H.E. Kyburg, Jr. *Probability and the Logic of Rational Belief*. Middletown, Conn., 1961.

[24] I. Levi. On Indeterminate Probabilities. *J.Phil.* 71:391–418, 1974.

[25] I. Levi. *Enterprise of Knowledge*. MIT Press, Cambridge, MA, 1980.

[26] I. Levi. Imprecise and Indeterminate Probabilities. *Proceedings of ISIPTA'99*, 258–265, 1999.

[27] D.V. Lindley, A. Tversky, and R.V. Brown. On the Reconciliation of Probability Assessments (with discussion). *J. Roy. Statist. Soc.* Ser. A, 142:146–180, 1979.

[28] E. Miranda, G. de Cooman and E. Quaeghebeur. Finitely additive extensions of distribution functions and moment sequences: The coherent lower prevision approach. *Int. J. Approx. Reasoning* 48:132–155, 2008.

[29] R. Pelessoni and P. Vicig. Williams coherence and beyond. *Int. J. Approx. Reasoning* 50(4):612–626, 2009.

[30] R. Pelessoni and P. Vicig. Uncertainty modelling and conditioning with convex imprecise previsions. *Int. J. Approx. Reasoning* 39(2–3):297–319, 2005.

[31] L.J. Savage. *The Foundations of Statistics*. John Wiley, New York, 1954.

[32] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NY, 1976).

[33] M.J. Schervish, T. Seidenfeld, and J.B. Kadane. The extent of non–conglomerability of finitely additive probabilities. *Z.f. Wahrscheinlichkeitstheorie* 66:205–226, 1984.

[34] T. Seidenfeld, J.B. Kadane, and M.J. Schervish. On the Shared Preferences of Two Bayesian Decision Makers. *J.Phil.* 86:225–244, 1989.

[35] M.J. Schervish, T. Seidenfeld, and J.B. Kadane. How sets of coherent probabilities may serve as model for degrees of incoherence. *Proceedings of ISIPTA'99*, 319–323, 1999.

[36] C.A.B. Smith. Consistency in statistical inference and decision (with discussion) *J. Roy. Statist. Soc.* Ser. B, 23:1–37, 1961.

[37] P. Suppes. The Measurement of Belief. *J. Roy. Statist. Soc.* Ser. B 36:160-175, 1974.

[38] B. Vantaggi. Graphical representation of asymmetric graphoid structures. *Proceedings of ISIPTA'03*, 560–574, 2003.

[39] P. Vicig. Epistemic independence for imprecise probabilities. *Int. J. Approx. Reasoning* 24:235–250, 2000.

[40] P. Vicig, M. Zaffalon, and F.G. Cozman. Notes on "Notes on Conditional Previsions". *Int. J. Approx. Reasoning* 44:358–365, 2007.

[41] P. Walley and T. Fine. Varieties of Modal (Classificatory) and Comparative Probability. *Synthese* 41:321–374. 1979.

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

[43] P. Walley. Inferences from multinomial data: learning about a bag of marmles. *J. Roy. Statist. Soc.* Ser. B 58:3–57, 1996.

[44] P. Walley and G. de Cooman. Coherence of rules for defining conditional possibility. *Int. J. Approx. Reasoning* 21 (1):63–107, 1999.

[45] P. M. Williams. Notes on conditional previsions. *Research Report*, School of Math. and Phys. Science, University of Sussex, 1975.

[46] P. M. Williams. Notes on conditional previsions. (Revised version of [45]). *Int. J. Approx. Reasoning* 44:366–383, 2007.

Bruno de Finetti and Fuzzy Probability Distributions

Reinhard Viertl
Technische Universität Wien, Austria
r.viertl@tuwien.ac.at

Abstract

Bruno de Finetti stated that probability does not exist in an objective sense. This is the basis for subjective Bayesian inference. For de Finetti probabilities are real numbers from the closed unit interval. Descriptive statistics for fuzzy data yield fuzzy relative frequencies. That is the starting point for modern considerations concerning probability. Recent research results are proposing a general probability concept where probabilities are special fuzzy numbers obeying a generalized form of additivity. This concept of so-called fuzzy probability distributions is explained in the paper.

1 Introduction

In his monumental and basic book *Theory of Probability* Bruno de Finetti gave a deep analysis of probability. One of his main conclusions is that probability is not an objective existing – frequently unknown – quantity, but as he says "probability does not exist, except in the mind". This idea is the basis for all neo-Bayesian statistical methods which were developed in the 20th century.

Another criticism by Bruno de Finetti about probability is concerning countable additivity of probability measures.

These and other comments on the theory of probability raise the question what mathematical model is suitable to describe probability.

2 Current probability models

There are different concepts of probability models. The most popular mathematical model for probability is the concept of probability spaces $(M, \mathcal{E}, \Pr)$, where M is a general set, $\mathcal{E}$ is a sigma field of subsets of M, and Pr a σ-additive and normalized measure on $\mathcal{E}$, i. e.

(1) $\Pr : \mathcal{E} \longrightarrow [0;1]$

(2) $\Pr(M) = 1$

(3) For every countable family $A_1, A_2, \cdots$ of pairwise disjoint events $A_i \in \mathcal{E}$ the following holds

$$\Pr\big(\bigcup_{n=1}^{\infty} A_n\big) = \sum_{n=1}^{\infty} \Pr\big(A_n\big).$$

Bruno de Finetti's concept of probability is starting with the events as elementary concept. So he is considering a family $\big(E_i, i \in I\big)$ of so-called events and defines probabilities as real numbers fulfilling finite additivity and the so-called coherence condition. For him all probabilities are conditional on the state of information H, i. e. $\Pr\big(E \mid H\big)$, where new information (for example data) is changing the probability:

(1) $0 \leq \Pr\big(E_i \mid H\big) \leq 1$ for all E_i in the event system

(2) For finitely many pairwise exclusive events $E_1, E_2, \cdots, E_n$
$\Pr\big(E_1 \vee E_2 \vee \cdots \vee E_n \mid H\big) = \sum_{i=1}^{n} \Pr\big(E_i \mid H\big)$
(finite additivity)

(3) $\Pr\big(E_1 \wedge E_2 \mid H\big) = \Pr\big(E_1 \mid E_2 \wedge H\big) \cdot \Pr\big(E_2 \mid H\big)$
(coherence)

From these axioms the so-called *Bayes' formula* follows:

For any exhaustive and pairwise exclusive finite family of events $E_1, \cdots, E_n$ and arbitrary event E_0 of the event system $\big(E_i,\, i \in I\big)$ the following holds:

$$\Pr\big(E_i \mid E_0 \wedge H\big) = \frac{\Pr\big(E_0 \mid E_i \wedge H\big) \cdot \Pr\big(E_i \mid H\big)}{\sum\limits_{j=1}^{n} \Pr\big(E_0 \mid E_j \wedge H\big) \cdot \Pr\big(E_j \mid H\big)}$$

for $i = 1(1)n$.

Proof: By $E_0 = E_0 \wedge \big(\bigvee_{i=1}^{n} E_i\big) = \bigvee_{i=1}^{n} \big(E_0 \wedge E_i\big)$ we obtain $\Pr\big(E_0 \mid H\big) = \Pr\Big(\bigvee_{i=1}^{n} \big(E_0 \wedge E_i \mid H\big)\Big) = \sum_{i=1}^{n} \Pr\big(E_0 \wedge E_i \mid H\big) = \sum_{i=1}^{n} \Pr\big(E_0 \mid E_i \wedge H\big) \Pr\big(E_i \mid H\big)$.

From the coherence condition we obtain

$\Pr\big(E_0 \wedge E_i \mid H\big) = \Pr\big(E_0 \mid E_i \wedge H\big) \cdot \Pr\big(E_i \mid H\big)$ and $\Pr\big(E_0 \wedge E_i \mid H\big) = \Pr\big(E_i \mid E_0 \wedge H\big) \cdot \Pr\big(E_0 \mid H\big)$ which concludes the proof. //

There are several other theories of probability. For more details compare [1] and [6].

3 Fuzzy probability distributions

More recently looking at histograms for fuzzy data it turns out that frequencies become fuzzy numbers. Therefore it is natural to look for more general concepts of probability, so-called *fuzzy probability distributions*. In this theory probabilities are special *fuzzy numbers*.

A fuzzy number $x^\star$ is characterized by its so-called *characterizing function* $\xi(\cdot)$ which is a generalization of an indicator function $I_A(\cdot)$ of a subset A of the set $I\!R$ of all real numbers.

A characterizing function $\xi(\cdot)$ is a real function of one real variable x obeying the following:

(1) $0 \le \xi(x) \le 1 \qquad \forall \quad x \in I\!R$

(2) $\forall\ \delta \in (0;1]$ the so-called δ-cut $C_\delta\big[\xi(\cdot)\big]$, defined by $C_\delta\big[(\cdot)\big] := \big\{x \in I\!R : \xi(x) \ge \delta\big\}$ is non-empty and a finite union of bounded closed intervals.

In case all δ-cuts are intervals the corresponding fuzzy number is called a *fuzzy interval*.

The system of all fuzzy intervals is denoted by $\mathcal{F}_I(I\!R)$. So-called fuzzy probability distributions $\Pr^\star$ on event systems $\big(E_i,\, i \in I\big)$ are defined in the following way:

A fuzzy probability distribution $\Pr^\star$ is a function $\Pr^\star : \big(E_i,\, i \in I\big) \longrightarrow \mathcal{F}_I(I\!R)$ obeying the following:

(1) $\Pr^\star\big(E_i\big)$ is a fuzzy interval $p^\star$ with characterizing function $\xi_i(\cdot)$ whose support is a subset of $[0;1]$

(2) For all finite families of pairwise exclusive events $E_1, \cdots, E_n$ the following holds true:
Let $C_\delta\big[\Pr^\star(E_i)\big] = \big[a_{i,\delta}; b_{i,\delta}\big]\ \forall\ i = 1(1)n$ and $C_\delta\big[\Pr^\star\big(\bigvee_{i=1}^{n} E_i\big)\big] = \big[c_\delta; d_\delta\big]$ be the corresponding δ-cuts then $c_\delta \ge \sum_{i=1}^{n} a_{i,\delta}$ and $d_\delta \le \sum_{i=1}^{n} b_{i,\delta}$
$\forall\ \delta \in (0;1]$

Special cases of fuzzy probability distributions are defined by so-called *fuzzy densities* $f^\star$ on measure spaces $(M, \mathcal{E}, \mu)$. A fuzzy density on $(M, \mathcal{E}, \mu)$ is a fuzzy valued function $f^\star : M \longrightarrow \mathcal{F}_I\big([0;\infty)\big)$ for which all δ-level functions $\underline{f}_\delta(\cdot)$ and $\overline{f}_\delta(\cdot)$, defined by $C_\delta\big[f^\star(x)\big] = \big[\underline{f}_\delta(x); \overline{f}_\delta(x)\big]\ \forall\, \delta \in (0;1]$, are integrable and there exists a classical probability density $f(\cdot)$ on $\big(M, \mathcal{E}, \mu\big)$, i. e.

$$\int_M f(x) d\mu(x) = 1 \quad \text{for which} \quad \underline{f}_1(x) \le f(x) \le \overline{f}_\delta(x)$$

for all $x \in M$.

Based on fuzzy densities probabilities of classical events $E \in \mathcal{E}$ are defined in the following way:
$\forall\ \delta \in (0;1]$ defining $\mathcal{D}_\delta$ to be the set of all classical probability densities $g(\cdot)$ on $\big(M, \mathcal{E}, \mu\big)$ obeying $\underline{f}_\delta(x) \le g(x) \le \overline{f}_\delta(x)\ \forall\ x \in M$, the fuzzy probability $\Pr^\star(E)$ of an event E is the fuzzy interval $p^\star$ which is generated by the following nested set of closed bounded intervals $\big[a_\delta; b_\delta\big]\ \ \forall\ \delta \in (0;1]$:

$$b_\delta := \quad \sup\Big\{ \int_E g(x) d\mu(x) : g(\cdot) \in \mathcal{D}_\delta \Big\}$$

$$a_\delta := \quad \inf\Big\{ \int_E g(x) d\mu(x) : g(\cdot) \in \mathcal{D}_\delta \Big\}$$

The characterizing function $\psi(\cdot)$ of $p^\star$ is given by its values

$$\psi(x) := \sup\Big\{\delta \cdot I_{[a_\delta; b_\delta]}(x) : \delta \in [0;1]\Big\} \qquad \forall\ x \in I\!R.$$

This definition yields a fuzzy probability distribution on the events system $\mathcal{E}$ for which the extremal events $\varnothing$ and M have precise probabilities $\Pr^\star(\varnothing) = I_{\{0\}}(\cdot)$ and $\Pr^\star(M) = I_{\{1\}}(\cdot)$. The inequalities for the endpoints of the δ-cuts follow from the integration.

References

[1] T. L. Fine. *Theories of Probability.* Academic Press, New York, 1973.

[2] B. de Finetti. *Theory of Probability.* Vol. 1 and Vol. 2, Wiley, London, 1974 and 1975.

[3] R. Viertl (editor). *Probability and Bayesian Statistics.* Plenum Press, New York, 1987.

[4] R. Viertl. *Statistical Methods for Fuzzy Data.* Wiley, Chichester, 2011.

[5] Z. Wang and G. J. Klir. *Fuzzy Measure Theory.* Plenum Press, New York, 1992.

[6] K. Weichselberger. Alternative Probabilistic Systems, in: *Encyclopedia of Life Support Systems*, published by UNESCO, Paris, 2001.

Conference Papers

Likelihood-Based Naive Credal Classifier

Alessandro Antonucci
IDSIA, Lugano
alessandro@idsia.ch

Marco E. G. V. Cattaneo
Department of Statistics, LMU Munich
cattaneo@stat.uni-muenchen.de

Giorgio Corani
IDSIA, Lugano
giorgio@idsia.ch

Abstract

The naive credal classifier extends the classical naive Bayes classifier to imprecise probabilities, substituting the imprecise Dirichlet model for the uniform prior. As an alternative to the naive credal classifier, we present a likelihood-based approach, which extends in a novel way the naive Bayes towards imprecise probabilities, by considering any possible quantification (each one defining a naive Bayes classifier) apart from those assigning to the available data a probability below a given threshold level. Besides the available supervised data, in the likelihood evaluation we also consider the instance to be classified, for which the value of the class variable is assumed missing-at-random. We obtain a closed formula to compute the dominance according to the maximality criterion for any threshold level. As there are currently no well-established metrics for comparing credal classifiers which have considerably different determinacy, we compare the two classifiers when they have comparable determinacy, finding that in those cases they generate almost equivalent classifications.

Keywords. Classification, naive credal classifier, naive Bayes classifier, likelihood-based learning.

1 Introduction

Classification, understood as the problem of assigning *class* labels to instances described by a set of *features*, is one of the major problems of AI, with lots of important applications, including pattern recognition, prediction, and diagnosis. Bayesian approaches to classification are particularly popular and effective. In particular, the *naive Bayes classifier* (NBC; e.g., see [11, Chap. 17]), assumes the conditional independence of the feature variables given the class; because of this unrealistic assumption, NBC requires the estimation of only a few parameters from the data. Yet, this assumption typically biases the probability computed by NBC which, regarding all the features as independent pieces of evidence, tends to assign a excessively high probability to the most probable class. The problem is emphasised in the presence of many features, among which could easily exist correlations [9]. However, NBC generally achieves a good accuracy under 0-1 loss; this means that, despite the biased probabilities, it produces good ranks among the competing classes [7]. The parameters are typically learned in a Bayesian way with uniform prior. Maximum-likelihood quantification has the advantage of being unbiased and independent from the prior specification, but generally leads to inferior classification performance, especially on data sets where the contingency tables, which contain the counts of the joint occurrences of specific values of the features and the class, are characterised by several zeros [8, 12] (see also Example 3).

The *naive credal classifier* (NCC, [18]), a generalisation of the NBC based on the theory of *imprecise probability* [15], attempts to make classification independent of the choice of the prior in a different way. NCC learns from data through the *imprecise Dirichlet model* (IDM, [16]); this corresponds to adopting a set of priors, which model a condition of near-ignorance about the model parameters. A NCC is equivalent to a collection of NBCs; while NBC returns the single class with highest probability according to the posterior probability mass function, NCC can in some cases suspend the judgment, by returning a set of classes rather than a single one. This provides a cautious and robust classification. A similar approach could be obtained by applying a *rejection option* to NBC, namely by returning more classes when the posterior probability estimated for the most probable class does not exceed a certain threshold. However, the rejection option requires accurate probability estimates to be effective, which is hardly the case for the NBC.

Of course, IDM is not the only technique to learn sets of distributions from data. Among others, *likelihood-based* approaches to the learning of imprecise-proba-

bilistic models from data [3, 14] can be regarded as an alternative to the IDM. Loosely speaking, the idea is to consider, instead of the single maximum-likelihood estimator, all the models whose likelihood is above a certain threshold level.

In this paper we investigate how likelihood-based techniques apply to NCC quantification. To do that, we keep the same independence assumptions of the NBC (and of the NCC), but we change the way the model is quantified. We call the resulting model *likelihood-based naive credal classifier* (LNCC). This model is associated with a classification algorithm which computes the set of unrejected classes according to the *maximality* criterion [15] (exactly as the NCC does) for any threshold level.

A notable feature of our approach is that, in the likelihood evaluation, we do not only consider the available (learning) data set, but also the instance to be classified, whose value of the class variable is assumed to be missing-at-random. This is important to obtain more accurate classification performances when coping with zero counts in the data set.

The paper is organised as follows. We first review some background material about the naive Bayes (Section 2.1) and credal (Section 2.2) classifiers and the likelihood-based approaches to the learning of imprecise-probabilistic models from data (Section 3). Then, in Section 4, we introduce the LNCC and obtain an analytic inference formula to compute the set of candidate optimal classes. Numerical tests are in Section 5. Conclusions and outlooks are finally in Section 6, while the proofs are in the appendix.

2 Naive Classifiers

In this section we review the necessary background information about classifiers developed under the naive assumption (i.e., independence between features given the class). First let us introduce the general problem of classification together with the necessary notation.

We use uppercase for the variables, lowercase for the states, calligraphic for the possibility spaces, and boldface for sets of variables. Let C denote the *class* variable, with generic value c, taking values in a finite set $\mathcal{C}$. Similarly, we have m features, $\mathbf{F} := (F_1, \ldots, F_m)$, each one taking values in the finite set $\mathcal{F}_j$, $j = 1, \ldots, m$.[1] Assume that the available data are d joint observations of these variables, say $\mathcal{D} := \{(c^{(i)}, f_1^{(i)}, \ldots, f_m^{(i)})\}_{i=1}^d$, with $c^{(i)} \in \mathcal{C}$ and $f_j^{(i)} \in \mathcal{F}_j$, for each $i = 1, \ldots, d$ and $j = 1, \ldots, m$. Information associated with the data set $\mathcal{D}$ is described by a *count function* n returning the number of elements of the data set $\mathcal{D}$ satisfying a condition to be specified in its argument. E.g., $n(C = c)$ is the number of instances where the class has value $c \in \mathcal{C}$, while $n(C = c, F_j = f_j)$ is the number of instances where C has value c and the j-th feature has value f_j. For sake of notation, we denote these counts as $n(c)$ and $n(c, f_j)$, and similarly for the others, with $n(\cdot) = d$.

Given an instance of the features $\tilde{\mathbf{f}} = (\tilde{f}_1, \ldots, \tilde{f}_m)$, classification is the problem of assigning it a single class label or, as in the case of Section 2.2, a set of them, all of which are candidates to be the correct category. A classifier always returning a single class is called *precise*, and *credal* otherwise.

2.1 Naive Bayes Classifier

A probabilistic approach to classification consists of learning from the data $\mathcal{D}$ a joint probability mass function for the whole set of variables $(C, \mathbf{F})$. Let the unknown chances of this distribution be denoted by $\theta_{c,\mathbf{f}}$ for each $(c, \mathbf{f}) \in \mathcal{C} \times \mathcal{F}_1 \times \ldots \times \mathcal{F}_m$. Once we learn these chances, we assign to the instance $\tilde{\mathbf{f}}$ the class label maximising the posterior (which is proportional to the joint) probability, i.e.,

$$\arg\max_{c \in \mathcal{C}} \theta_{c,\tilde{\mathbf{f}}}.$$

As the number of parameters specifying the joint distribution grows exponentially with the number of features, such a probabilistic approach is generally too demanding, unless we make some assumption about the independence relations between the variables. A notable example is the so-called *naive* assumption, which says that, given the class variable, the features are conditionally independent from each other.[2] This induces in the joint the following factorisation:

$$\theta_{c,\mathbf{f}} := \theta_c \cdot \prod_{j=1}^{m} \theta_{f_j|c}, \tag{1}$$

where θ_c is the (unconditional) chance for $C = c$, and similarly for the conditional ones. Equation (1) makes it possible to assess the joint distribution, and hence perform classification, by means of a number of parameters which is linear in the number of features and classes. Let θ denote the whole set of chances to be quantified on the right-hand side of (1) and Θ the corresponding set of possible assignments. The parameter θ is quantified in a Bayesian way; given a Dirichlet prior over Θ, we obtain the following poste-

[1] We focus on classification of discrete features. A discussion on the extension to continuous variables is in the conclusions.

[2] We say that A and B are conditionally independent given C if $P(a, b|c) = P(a|c) \cdot P(b|c)$, for each a, b, and c.

rior estimates:

$$\theta_c = \frac{n(c) + s\,t(c)}{n(\cdot) + s}, \quad (2)$$

$$\theta_{f_j|c} = \frac{n(c, f_j) + s\,t(c, f_j)}{n(c) + s\,t(c)}, \quad (3)$$

where Walley's parametrisation of the Dirichlet distribution is employed. In particular, s can be thought of as a number of *hidden instances*, in the usual interpretation of conjugate Bayesian priors as additional samples. The parameters $t(\cdot)$ can be interpreted as the proportion of hidden instances of a given type; for instance, $t(c)$ is the expected proportion of hidden instances for which $C = c$.

In particular, non-informative specifications can be obtained by Perks' prior, which means $t(c) := |\mathcal{C}|^{-1}$ and $t(c, f_j) := |\mathcal{F}_j|^{-1}|\mathcal{C}|^{-1}$ for each $c \in \mathcal{C}$, $f_j \in \mathcal{F}_j$, $j = 1, \ldots, m$, and $s = 1$. In the language of Bayesian networks, this is also known as BDe [11, Chap. 17].

2.2 Naive Credal Classifier

The classification performances of the NBC can be quite sensitive to the choice of the prior. In a situation where different priors return different class labels, a conservative approach consists of taking multiple priors as a model of a condition of prior (near) ignorance about the model parameters, and hence learning a posterior independently for each prior. This can be done by means of the *imprecise Dirichlet model* (IDM, [16]), for which the "precise" specification of the NBC Dirichlet prior is relaxed, and its parameters are free to vary in the following set, with minimal constraints:

$$\mathcal{T} := \left\{ \mathbf{t} \;\middle|\; \begin{array}{l} \sum_{c\in\mathcal{C}} t(c) = 1 \\ \sum_{f_j\in\mathcal{F}_j} t(c, f_j) = t(c), \forall c \in \mathcal{C}, \forall j \\ t(c, f_j) > 0, \forall (c, f_j) \in \mathcal{C} \times \mathcal{F}_j, \forall j \end{array} \right\}. \quad (4)$$

Each $\mathbf{t} \in \mathcal{T}$ corresponds to a different Dirichlet prior and hence a different NBC quantification. The collection of all these NBCs is called *naive credal classifier* (NCC, [17]), and provides a collection of posterior distributions for the class variable given the feature of the instance to be classified. In order to decide which class labels to assign to the instance, the *maximality* criterion [15] is adopted: a class is rejected if there is another class that is more probable according to every distribution. Thus, in order to perform classification with the NCC, for each $c', c'' \in \mathcal{C}$, we have to test whether or not c' *dominates* c'', i.e.,[3]

$$\inf_{\mathbf{t}\in\mathcal{T}} \frac{P_{\mathbf{t}}(c', \tilde{\mathbf{f}})}{P_{\mathbf{t}}(c'', \tilde{\mathbf{f}})} > 1, \quad (5)$$

where $P_{\mathbf{t}}$ is the NBC quantification associated to $\mathbf{t}$. From (1), (2) and (3), we can rewrite the objective function of our optimisation problem in (5) as[4]

$$\left[\frac{n(c') + s\,t(c')}{n(c'') + s\,t(c'')}\right]^{1-m} \prod_{j=1}^{m} \frac{n(c', \tilde{f}_j)}{n(c'', \tilde{f}_j) + s\,t(c'', \tilde{f}_j)},$$

and hence check dominance by solving the corresponding optimisation with the constraints in (4).

Counterintuitive behaviors of NCC take place in presence of zero counts; in particular (a) an attribute F_j such that $n(c', \tilde{f}_j) = 0$ prevents c' from dominating any other class (see Example 3); (b) a class c' such that $n(c') = 0$ is identified as non-dominated for most instances. These behaviors were first observed in [17]; a solution to these problems, which make the NCC unnecessarily imprecise, has been studied in [4], proposing an ϵ-contamination of the IDM prior with the uniform prior of the NBC: this corresponds to a slight modification of the set $\mathcal{T}$, obtained by rewriting the constraints in (4) in the form $\epsilon\,|\mathcal{C}|^{-1} \leq t(c) \leq (1-\epsilon)+\epsilon\,|\mathcal{C}|^{-1}$, and similarly for $t(c, f_j)$. Such a NCC extension is denoted as NCC_ϵ.[5]

3 Likelihood-Based Learning of Imprecise-Probabilistic Models

Coping with multiple priors as in the IDM is not the only possible approach to learn imprecise-probabilistic models from data. In a likelihood-based approach, we can simply start by considering a collection of candidate models, and then only keep those assigning to the available data a probability beyond a certain threshold. We introduce these ideas by means of an example.

Example 1. *Consider a Boolean variable X, for which N observations are available, and n of them report the state true. If $\theta \in [0, 1]$ is the chance that X is true, the likelihood induced by the observed data is $lik(\theta) := \theta^n \cdot (1-\theta)^{N-n}$ and its maximum is attained at $\hat{\theta} = \frac{n}{N}$. For each $\alpha \in [0, 1]$, we can (numerically) compute the values of θ such that $lik(\theta) \geq \alpha\, lik(\hat{\theta})$. Figure 1 depicts the behaviour of these intervals (which can be also interpreted as confidence intervals for θ; e.g., see [10]) for increasing sample size.*

The approach considered in the above example can be easily extended to the general case, and can be interpreted as a way of updating imprecise probabilities [1, 13], in the following sense. Consider a *credal*

[3] Note that the ratio between conditional probabilities can be equivalently described as a ratio between joint probabilities.

[4] Note that a partial optimisation has been already performed in the numerators of the terms in the product.

[5] Note that NCC_0 is the NCC, while NCC_1 is the NBC.

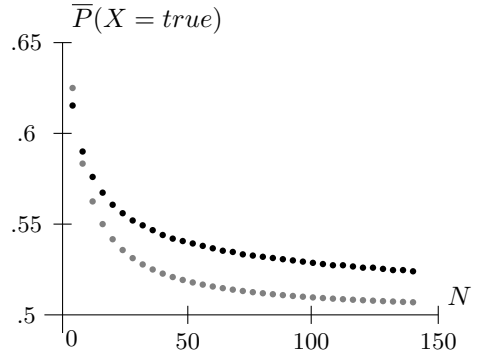


Figure 1: Comparison between probability intervals obtained by likelihood-based learning ($\alpha = .85$, black points) and IDM ($s = 2$, grey points) for Example 1. The plot shows the upper bounds of the interval probability that the variable is true as a function of the sample size N, when $\frac{n}{N} = \frac{1}{2}$. The plot for the lower bounds would be symmetric to this one.

set $\mathbf{P}$, i.e., a collection of probability distributions all over the same variable. Assume the elements of $\mathbf{P}$ are indexed by a parameter θ taking values in a set Θ, i.e., $\mathbf{P} := \{P_\theta\}_{\theta\in\Theta}$. Given the available data $\mathcal{D}$, let us consider the corresponding normalised likelihood:

$$lik(\theta) := \frac{P_\theta(\mathcal{D})}{\sup_{\theta'\in\Theta} P_{\theta'}(\mathcal{D})}. \tag{6}$$

The likelihood-based approach to learning consists of removing from $\mathbf{P}$ the distributions whose normalised likelihood is below some threshold. Thus, given $\alpha \in [0, 1]$, we consider the following (smaller) credal set:

$$\mathbf{P}_\alpha := \{P_\theta\}_{\theta\in\Theta\,:\,lik(\theta)\geq\alpha}. \tag{7}$$

Clearly, $\mathbf{P}_{\alpha=1}$ is typically a "precise" credal set including only the maximum-likelihood distribution, while $\mathbf{P}_{\alpha=0} = \mathbf{P}$. In principle, the original credal set $\mathbf{P}$ can be obtained by means of some other imprecise-probabilistic learning technique, which is indeed refined by the likelihood-based approach. Likelihood-based learning is said to be *pure*, if the credal set $\mathbf{P}$ includes all the possible distributions that can be specified over the variable under consideration (or, as in the next section, at least all those satisfying the structural judgements about symmetry and independence characterising the model under consideration).

4 Likelihood-Based Naive Credal Classifier

Let us consider a pure likelihood-based learning of the model probabilities of the naive classifier. Thus, let $\mathbf{P}$ denote the credal set associated to a NCC with vacuous quantification of the model probabilities (i.e., each chance is only required to belong to the $[0, 1]$ interval). Let the parameter θ with values in Θ denote a parametrisation of this credal set, i.e., $\mathbf{P} := \{P_\theta\}_{\theta\in\Theta}$, where θ is a NBC quantification. Given the available data $\mathcal{D}$, let us consider the normalised likelihood as in (6), and hence the credal set $\mathbf{P}_\alpha \subseteq \mathbf{P}$ as in (7).

We call *likelihood-based naive credal classifier* (LNCC, called *naive hierarchical classifier* in [3]) the collection of NBCs in the credal set $\mathbf{P}_\alpha$. This only provides an implicit specification of the model probabilities.[6] Yet, we can already describe how LNCC-based classification is intended. The same dominance criterion (i.e., maximality) as for the NCC is considered, and we say that c' dominates c'' iff

$$\inf_{\theta\in\Theta\,:\,lik(\theta)\geq\alpha} \frac{P_\theta(c', \tilde{\mathbf{f}})}{P_\theta(c'', \mathbf{f})} > 1. \tag{8}$$

In order to perform classification with the LNCC, we should discuss (8) for each pair of classes $c', c'' \in \mathcal{C}$. This task will be considered in Section 4.1. First, let us note that, when evaluating the likelihood *lik*, we do not only consider the data set $\mathcal{D}$, but also the instance under consideration $\tilde{f}$. The value of the class variable for this instance is unavailable (i.e., missing), no matter what its actual value is. Thus, the probability we should take into account for the overall likelihood evaluation is the product of $P_\theta(\mathcal{D})$ and

$$P_\theta(\tilde{\mathbf{f}}) := \sum_{c\in\mathcal{C}} \left[\theta_c \prod_{i=1}^{m} \theta_{\tilde{f}_i|c}\right]. \tag{9}$$

Note that we perform classification by means of the dominance test in (8) for each $c', c'' \in \mathcal{C}$. Thus, as we cope with the likelihood separately for each pair of classes, a simplification assumption consists of assuming that, when checking whether c' dominates c'', the instance under consideration can only be c' or c''. This basically means to restrict the sum in (9) only to c' and c''. In order to see how this kind of classification works in practice consider the following example.

Example 2. *Consider a LNCC with a Boolean class C and a single Boolean feature F. In this setup, a NBC specification is provided by the three-dimensional parameter $\theta := (\theta_c, \theta_{f|c}, \theta_{f|\neg c})$, taking values in $\Theta := [0, 1]^3$. Apart from (c, f) which appears five times, the other three possible combinations for the class/feature values appear only once in the data set. To decide whether or not $C = c$ dominates $C = \neg c$, when the instance to be classified is $F = f$, we first compute the likelihood of the available (supervised) data:*

$$lik(\theta) = \theta_c^6 \cdot (1-\theta_c)^2 \cdot \theta_{f|c}^5 \cdot (1-\theta_{f|c}) \cdot \theta_{f|\neg c} \cdot (1-\theta_{f|\neg c}).$$

[6] Note that, if regarded as a credal net [6], the LNCC (as the NCC) has non-separately specified credal sets.

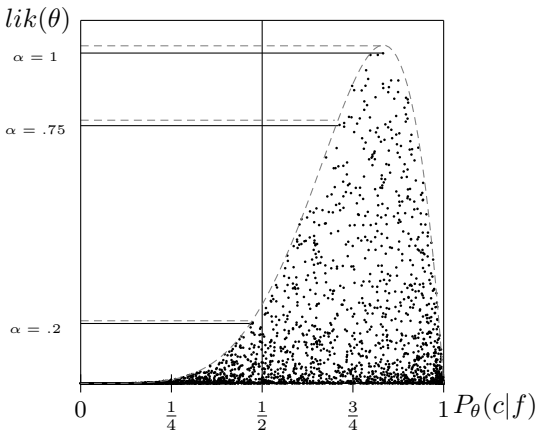


Figure 2: LNCC-based classification. The dominance test in Example 2 is solved by generating a random sample of 3000 NBC quantifications θ and depicting for each θ the posterior and the likelihood as the point $(P_\theta(c|f), lik(\theta))$. Note that in the Boolean case $P_\theta(c|f) > \frac{1}{2}$ is an equivalent dominance condition. The upper envelope of the points in the limit of an infinite sample size (see Section 4.1) is depicted in grey. Horizontal lines describe the cuts for different α-values. Black lines are based on the random sample, while those referred to the upper envelope are grey.

As we also want to consider the instance to be classified, we multiply this likelihood by the chance that $F = f$, which according to (9) is

$$\theta_c\, \theta_{f|c} + (1 - \theta_c)\, \theta_{f|\neg c}.$$

For each $\theta \in \Theta$, c dominates $\neg c$ if

$$\frac{\theta_c\, \theta_{f|c}}{(1 - \theta_c)\, \theta_{f|\neg c}} > 1.$$

To perform classification with the LNCC, we just have to check whether or not such a dominance relation is satisfied for each θ whose likelihood is not below the maximum likelihood multiplied by α. Figure 2 reports a Monte Carlo solution of this problem. Note that we have dominance for high threshold levels (e.g., $\alpha = .75$), and no dominance for low levels (e.g., $\alpha = .2$).

4.1 Statistical Inference with LNCC

In the previous section we defined the LNCC corresponding to a given α level, and described how we intend to perform inference based on this model. Yet, the sampling-based method considered in Example 2 is not necessary. In this section, we provide a classification algorithm for the LNCC based on a parametric formula for the upper envelope of the likelihood.

Let us therefore, for a generic classification problem, consider the dominance test between c' and c'' for an instance $\tilde{\mathbf{f}}$ to be classified by means of the LNCC for a given threshold α on the basis of the data $\mathcal{D}$. The idea is to parametrise the upper envelope of the likelihood (also called *profile likelihood* [2, 14]) by means of a parameter t ranging on the interval $[a, b]$, where

$$a := -\min_{j=1,\ldots,m} n(c', \tilde{f}_j) - \frac{1}{2},$$
$$b := \min_{j=1,\ldots,m} n(c'', \tilde{f}_j) + \frac{1}{2}.$$

In order to characterise the profile likelihood of the LNCC, we employ the following two results.

Theorem 1. *For each $\theta \in \Theta$ and each pair of classes $c', c'' \in \mathcal{C}$, there is a unique $t \in [a, b]$ such that*

$$\frac{P_\theta(c', \tilde{f})}{P_\theta(c'', \tilde{f})} = \frac{\left[n(c')+\frac{1}{2}+t\right] \prod_{j=1}^{m} \frac{\left[n(c', \tilde{f}_j)+\frac{1}{2}+t\right]}{\left[n(c')+\frac{1}{2}+t\right]}}{\left[n(c'')+\frac{1}{2}-t\right] \prod_{j=1}^{m} \frac{\left[n(c'', \tilde{f}_j)+\frac{1}{2}-t\right]}{\left[n(c'')+\frac{1}{2}-t\right]}}, \quad (10)$$

where $\frac{x}{0}$ is interpreted as $+\infty$ when x is positive, and as 1 when $x = 0$. Moreover, the right-hand side of (10) is a continuous, strictly increasing function of $t \in [a, b]$.

Theorem 1 defines a many-to-one relation between the elements of Θ and those of the interval $[a, b]$. For each $t \in [a, b]$, let Θ_t denote the set of all elements of Θ for which (10) is satisfied.

Theorem 2. *Let L, l', l'', p', p'' be the functions on $[a, b]$ defined by*

$$L(t) = \sup_{\theta \in \Theta_t} lik(\theta),$$
$$l'(t) = \left[n(c')+\tfrac{1}{2}+t\right]^{n(c')} \prod_{j=1}^{m} \frac{\left[n(c', \tilde{f}_j)+\frac{1}{2}+t\right]^{n(c', \tilde{f}_j)}}{\left[n(c')+\frac{1}{2}+t\right]^{n(c')}},$$
$$l''(t) = \left[n(c'')+\tfrac{1}{2}-t\right]^{n(c'')} \prod_{j=1}^{m} \frac{\left[n(c'', \tilde{f}_j)+\frac{1}{2}-t\right]^{n(c'', \tilde{f}_j)}}{\left[n(c'')+\frac{1}{2}-t\right]^{n(c'')}},$$
$$p'(t) = \left[n(c')+\tfrac{1}{2}+t\right] \prod_{j=1}^{m} \frac{\left[n(c', \tilde{f}_j)+\frac{1}{2}+t\right]}{\left[n(c')+\frac{1}{2}+t\right]},$$
$$p''(t) = \left[n(c'')+\tfrac{1}{2}-t\right] \prod_{j=1}^{m} \frac{\left[n(c'', \tilde{f}_j)+\frac{1}{2}-t\right]}{\left[n(c'')+\frac{1}{2}-t\right]},$$

for all $t \in [a, b]$, where both $\frac{0}{0}$ and 0^0 are interpreted as 1. Then

$$L \propto l'\, l''\, (p' + p''). \quad (11)$$

These two theorems can be used to perform LNCC-based classification without sampling. We first evaluate the maximum $\hat{t}$ of $L(t)$. Then, we check whether, for the values $t \in [a, b]$ such that $L(t) \geq \alpha L(\hat{t})$, the ratio on the right-hand side of (10) is always bigger than one. If so, we have that c' dominates c''. To see how this works, consider the classification task in Example 2. When testing whether or not c dominates

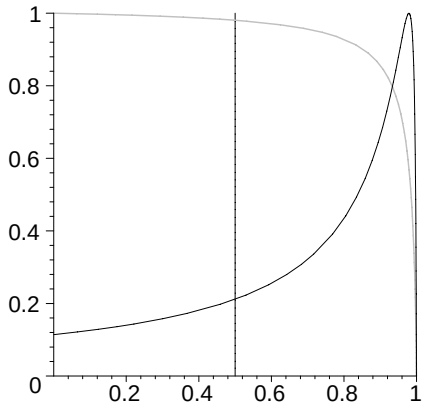


Figure 3: Profile likelihood functions for $P(c|\tilde{f}_1, \tilde{f}_2)$ in Example 3: with and without the probability of the new instance (black and grey curves, respectively).

$\neg c$, we have $[a, b] = [-\frac{11}{2}, \frac{3}{2}]$. For each $t \in [a, b]$, the right-hand side of (10) rewrites as $\frac{11+2t}{3-2t}$, while the likelihood in (11) is proportional to $(11 + 2t)^5 (3 - t)$; the resulting profile likelihood is depicted in Figure 2 (grey curve).

Example 3. *Consider a LNCC with a Boolean class C and two features F_1, F_2. We want to classify a new instance with features $\tilde{f}_1, \tilde{f}_2$, on the basis of a data set $\mathcal{D}$ containing $n(\cdot) = 100$ instances. In the data set $\mathcal{D}$, the class c has been observed $n(c) = 50$ times, always in conjunction with the feature $\tilde{f}_1$, but never with the feature $\tilde{f}_2$; that is, $n(c, \tilde{f}_1) = 50$ and $n(c, \tilde{f}_2) = 0$. Of the $n(\neg c) = 50$ observed instances with class $\neg c$, one had the feature $\tilde{f}_1$, and another one had the feature $\tilde{f}_2$; that is, $n(\neg c, \tilde{f}_1) = 1$ and $n(\neg c, \tilde{f}_2) = 1$. Figure 3 shows the profile likelihood function for $P(c|\tilde{f}_1, \tilde{f}_2)$ (compare with Figure 2) when the probability (9) of the new instance is considered in the likelihood function (black curve), and when it is not considered (grey curve).*

Hence, the LNCC classifies the new instance as c when α is sufficiently large (more precisely, when $\alpha \geq 0.22$); the same classification is obtained by the NBC with uniform prior and by the NCC$_\epsilon$ (for sufficiently large ϵ). By contrast, without using the probability (9) of the new instance in the likelihood function, the classifier would return both classes (if $\alpha \leq 0.98$), as does the standard NCC (that is, NCC$_\epsilon$ with $\epsilon = 0$), while the NBC with maximum-likelihood quantification returns the class $\neg c$ (at least when the usual likelihood function, without the probability of the new instance, is maximised). This is an example of the zero-counts issue discussed at the end of Section 2.2, which is the main reason why the ϵ-modification of NCC has been introduced and why we consider also the probability (9) of the new instance in the likelihood function.

4.2 Computational Complexity

The classification of an instance requires the iteration of the dominance test over all the possible pair of class labels, this task being clearly quadratic in $|\mathcal{C}|$. In order to perform the dominance test, the function $L(t)$ should be evaluated. This requires a number of operations which is linear in the number of attributes m. The same order of magnitude is required to compute the right-hand side of (10). In our preliminary implementation, τ equally spaced points over the interval $[a, b]$ have been considered. The numerical optimisation of the likelihood and identification of the α-cut was therefore simply performed by considering the value of the function $L(t)$ in these points. For the experiments, we adopted $\tau = 250$; empirically, increasing τ beyond this value resulted only in negligible differences in the classifications produced by LNCC. Thus, for practical purposes, we can consider τ as a constant, and we obtain $O(m|\mathcal{C}|^2)$ complexity (as for the NCC, [17]).

5 Experiments

To describe the performance of a credal classifier, we need multiple indicators. In particular, we adopt the following:

- *determinacy* (Det): the percentage of instances classified with a single class;
- *single accuracy* (Sgl-acc): the accuracy over the instances classified with a single class;
- *set-accuracy*: the accuracy over the instances classified with more classes;
- *indeterminate output size*: the average number of classes returned when the classification is indeterminate.

Note that when NCC is determinate, it returns the same class as NBC; this is due to the uniform prior being included in the IDM. This cannot be guaranteed for LNCC; however in our experiments LNCC, when precise, generally returned the same class as NBC. Thus, the single accuracy of NCC [resp. LNCC] is equivalent to the accuracy achieved by NBC on the instances determinately classified by NCC [resp. LNCC]. A credal classifier does a good job at isolating hard-to-classify instances if its Bayesian counterpart has low accuracy on the instances which are indeterminately classified. We denote as *NBC-I* the accuracy of naive Bayes on the instances indeterminately classified by the credal classifier at hand (NCC or LNCC, depending on the context). A large drop between single accuracy and NBC-I means thus that the credal

classifier is effective at isolating instances which are hard to classify.

Unfortunately, there is so far no single indicator which can reliably compare two credal classifiers. The *discounted-accuracy* (D-acc, [5]) has been proposed for this purpose; it is defined as $\frac{1}{n}\sum_{i\in acc}\frac{acc_i}{|output_i|}$, where, with reference to the i-th instance, acc_i denotes whether the set of returned classes contains or not the actual one and $|output_i|$ denotes the number of classes returned. On each instance, the classifier is thus given 0 if inaccurate or $1/|output_i|$ if accurate. Yet, discounted-accuracy sees as equivalent, in the long term, a *vacuous* classifier which returns all classes and a *random* classifier which returns a single class at random. However, the vacuous classifier should be generally preferred over the random one; this is clear if one thinks for instance of the diagnosis of a disease. In a way the vacuous, unlike the random, is aware of being ignorant; yet discounted-accuracy does not capture this point. In fact, the design of metrics to rank credal classifiers is an important *open* problem. Moreover, when dealing with credal classifiers with considerably different determinacy, discounted-accuracy favors the more determinate ones. We thus try to compare LNCC and NCC (in its NCC_ϵ generalisation) when they have the same determinacy. For this purpose we tried different values of ϵ for NCC_ϵ and α for LNCC; more precisely, denoting also the value of α as a subscript, we considered: $\text{NCC}_{0.05}$, $\text{NCC}_{0.15}$, $\text{NCC}_{0.25}$, $\text{NCC}_{0.35}$; $\text{LNCC}_{0.35}$, $\text{LNCC}_{0.55}$, $\text{LNCC}_{0.75}$, $\text{LNCC}_{0.95}$.

5.1 Artificial Data

We generated artificial data sets, considering a binary class and 10 binary features, under a naive data generation mechanism. We set the marginal chances of classes as uniform, while we drew the conditional chances of the features under the constraint $|\theta_{f_j|c'} - \theta_{f_j|c''}| \geq 0.1$ for each $c', c'' \in \mathcal{C}$ and $j = 1, \ldots, m$; the constraint forced each feature to be truly dependent on the class. We drew such chances 20 times uniformly at random and we consider the sample sizes $d \in \{25, 50, 100\}$. For each pair (θ, d) we generated 30 training sets and a huge test set of 10000 instances. For each sample size, we thus perform 20 θ × 30 trials = 600 training/test experiments. Note that, dealing with two classes, set-accuracy is fixed to 100% and indeterminate output size to 2; we do not need thus to consider these indicators.

In Figure 5 we show how the determinacy of NCC and LNCC varies with the sample size, choosing pairs $\{\alpha, \epsilon\}$ which produce reasonably comparable curves. Interestingly, NCC is more sensitive than LNCC to the sample size d; the determinacy of NCC steeply increases with d, unlike that of LNCC. In fact, NCC becomes determinate once the rank of the classes does not change under all the different priors of the IDM; but the smoothing effect of the prior decreases with d. The same is known to happen with likelihood-based methods, but convergence towards the precise model is slower, as shown by the comparison in Figure 1.

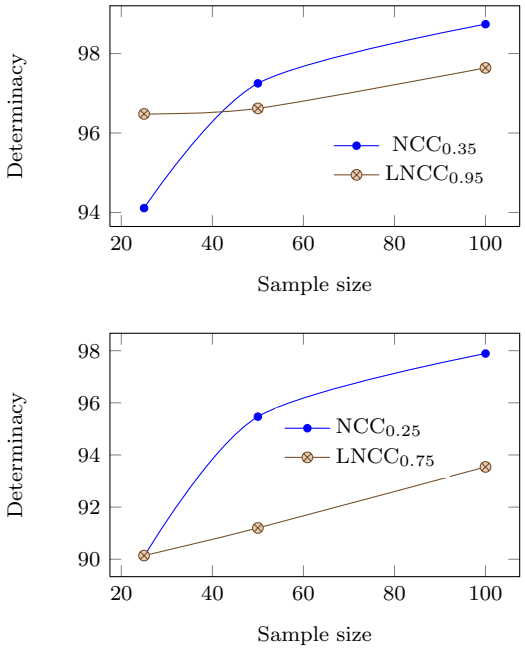


Figure 5: Determinacy of NCC and LNCC as a function of the sample size d.

Classifier	n	Det	Sgl-acc	D-acc	NBC-I
$\text{NCC}_{0.25}$	25	90.1	90.4	86.5	54.5
$\text{LNCC}_{0.75}$	25	90.1	90.4	86.6	52.8
$\text{NCC}_{0.05}$	50	91.7	91.5	88.2	57.4
$\text{LNCC}_{0.75}$	50	91.2	91.8	88.2	55.3
$\text{NCC}_{0.25}$	100	97.9	90.5	89.7	51.2
$\text{LNCC}_{0.95}$	100	97.7	90.6	89.7	51.4

Table 1: Performance indicators for NCC and LNCC, for choices of α and ϵ leading to close determinacies; each number is an average over 600 experiments.

It is interesting to compare LNCC and NCC when they have, for the same sample size, very close determinacy. This is this the case of $\text{NCC}_{0.25}$ and $\text{LNCC}_{0.75}$ for d=25; of $\text{NCC}_{0.05}$ and $\text{LNCC}_{0.75}$ for d=50; of $\text{NCC}_{0.25}$ and $\text{LNCC}_{0.95}$ for d=100. Note that in general it is not possible to predict in advance which choice of ϵ and α will allow to obtain similar determinacy from LNCC and NCC. However, when NCC and LNCC achieve the same determinacy, their performances are very similar also on the remaining indica-

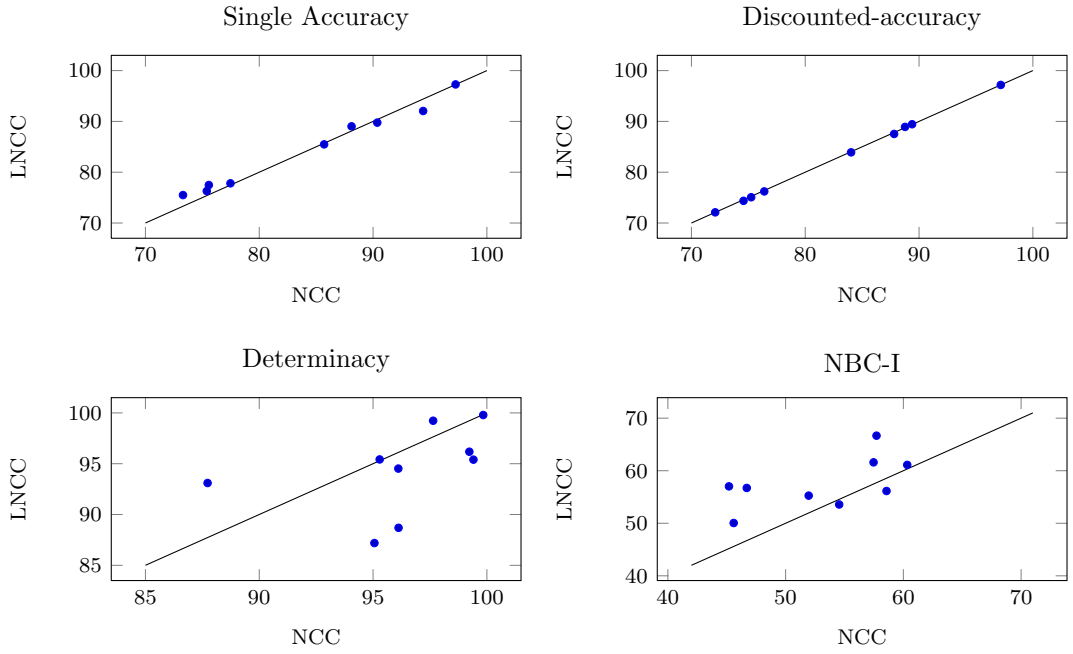


Figure 4: LNCC versus NCC: scatter plots on different UCI binary data sets.

tors, as shown in Table 1. This suggests that, for the same level of determinacy, NCC and LNCC become indeterminate on roughly the same instances, despite the different derivation of their algorithms. Note also the large drop between Sgl-acc and NBC-I for both classifiers, showing that both NCC and LNCC can be seen as extending NBC towards increased reliability.

5.2 Binary Data Sets from UCI

We then considered 9 binary data sets (containing 2 classes) from the UCI repository; the number of instances ranges from 57 to 3000 and the number of features from 8 to 60. Since the data sets are binary, set-accuracy and indeterminate output size can only be respectively 100% and 2; we do not consider thus these indicators. For each credal classifiers we report instead Sgl-acc and NBC-I, namely the accuracy of NBC when the credal classifier is respectively determinate[7] and indeterminate. If there is a large difference between these two indicators, the credal classifier is doing a good job at isolating instances which are difficult to classify for NBC. Moreover, we report determinacy and D-acc to provide a general overview of the classifiers' behavior.

The results in Table 2 show that when LNCC and NCC have close determinacy, they also have very similar performance on the remaining indicators, as in the previous experiments. Also in this case, there is in general a large drop between Sgl-acc and NBC-I, showing that both credal classifiers are effective at isolating instances that are hard to classify for NBC.

However, it is also interesting to see what happens if we set a default choice for ϵ and α. We set ϵ to 0.05 for NCC, thus considering a minimal variation over the NCC of [17], aimed at avoiding issues with zero counts. As for LNCC, we adopted a trial and error approach, from which $\alpha = 0.75$ appeared as a reasonable compromise between determinacy and reliability of the classifier. *On average*, NCC has slightly

Dataset	Classifier	Det	Sgl-acc	D-acc	NBC-I
german	$NCC_{0.05}$	96.1	75.6	74.6	58.6
german	$LNCC_{0.95}$	95.7	75.7	74.6	57.5
haberman	$NCC_{0.05}$	95.1	73.3	72.1	45.6
haberman	$LNCC_{0.95}$	93.9	73.8	71.9	50.2
hepatitis	$NCC_{0.05}$	95.3	85.7	60.3	84.0
hepatitis	$LNCC_{0.75}$	95.4	85.5	61.1	83.9

Table 2: Results for LNCC and NCC on UCI data sets, for choices of α and ϵ leading to close determinacies. We report results only for 3 out of 9 analyzed data sets, because the remaining data sets only show very similar findings: namely that when LNCC and NCC have close determinacy, their performance on all indicators is substantially identical.

[7] This follows from NBC returning the same class as the credal classifier, when the latter is determinate (this is theoretically guaranteed for NCC and only empirically verified for HNCC); Sgl-acc can be thus seen as measuring also the accuracy of NBC when the credal classifier is determinate.

higher determinacy (96.3% vs. 94.3%) and slightly lower single-accuracy (84.2% vs. 85.5%) than LNCC. Moreover, the *area of ignorance* (instances indeterminately classified) of NCC is slightly more difficult to classify for NBC than the area of ignorance of LNCC: the average NBC-I is 53.1% vs. 57.6%. In fact, NCC is slightly more determinate and thus more selective in deciding when to become indeterminate. The average discounted accuracy of the two classifiers is very close (82.8% vs 82.7%). However, averaging indicators over data sets is questionable; we thus also present in Figure 4 the scatter plots of such indicators. On each data set there is little difference between the single-accuracy of NCC and LNCC; the same holds also for the discounted-accuracy. On the other hand, there are sometimes considerable differences between NCC and LNCC as for the determinacy, which tends to be larger for NCC, and as for NBC-I, which tends to be larger for LNCC. In general, when the difference in determinacy between NCC and LNCC increases, so does the difference in NBC-I between LNCC and NCC.

6 Conclusions and Outlooks

We have presented an alternative, likelihood-based, approach to the imprecise-probabilistic quantification of a naive classifier. A numerical comparison with the naive credal classifier (in its modified formulation to cope with zero-count issues) shows that, despite their deeply different derivations, the performance of the two classifiers is very similar when they produce more or less the same amount of indeterminate classifications. When the amount of indeterminacy between the two classifiers is considerably different, a meaningful comparison is difficult: this would require modelling the trade-off between accuracy and informativeness by means of one or more performance indicators, which is currently one of the most important *open* problems in credal classification.

Extensions of the new approach to more complex independence structures (e.g., tree-augmented naive), incomplete data sets, and continuous features seem to be worth of future investigations.

Acknowledgements

The research in this paper has been partially supported by the Swiss NSF grants n. 200020-132252 and by the Hasler foundation grant n. 10030. The authors wish to thank the anonymous referees for their helpful comments.

Appendix

Proof of Theorem 1. Let g be the function assigning to each $t \in [a,b]$ the corresponding right-hand side of (10). We prove the theorem by showing that for each $x \in [0,+\infty]$ there is a unique $t \in [a,b]$ such that $g(t) = x$. When $t \in (a,b)$, all the sums of three terms (of the form $\left[n + \frac{1}{2} \pm t\right]$) in the expression of $g(t)$ are positive. In this case, each fraction

$$\frac{\left[n(c', \tilde{f}_j) + \frac{1}{2} + t\right]}{\left[n(c') + \frac{1}{2} + t\right]} \tag{12}$$

is a continuous, increasing function of t, since it is differentiable with derivative

$$\frac{n(c') - n(c', \tilde{f}_j)}{\left[n(c') + \frac{1}{2} + t\right]^2} \geq 0.$$

Therefore, the numerator of $g(t)$ is a continuous, strictly increasing function of $t \in (a,b)$, since it is the product of m continuous, increasing functions and of the continuous, strictly increasing function $\left[n(c') + \frac{1}{2} + t\right]$. Analogously, we can prove that the denominator of $g(t)$ is a continuous, strictly decreasing function of $t \in (a,b)$, and therefore g is continuous and strictly increasing on (a,b).

In order to prove Theorem 1, it now suffices to show that

$$\lim_{t \downarrow a} g(t) = g(a) = 0 \quad \text{and} \quad \lim_{t \uparrow b} g(t) = g(b) = +\infty. \tag{13}$$

We prove the first expression: the second one can be proved analogously. As t tends to a from above, the denominator of $g(t)$ tends to a positive constant, which is reached when $t = a$. To study the limit of the numerator of $g(t)$, let j_0 be such that $n(c', \tilde{f}_{j_0}) = \min_{j=1,\ldots,m} n(c', \tilde{f}_j)$. We can distinguish two cases: either $n(c', \tilde{f}_{j_0}) = n(c')$, or $n(c', \tilde{f}_{j_0}) < n(c')$. In the first case, $n(c', \tilde{f}_j) = n(c')$ for all j, and the numerator reduces to $\left[n(c') + \frac{1}{2} + t\right]$, since the fractions (12) are all equal 1. Therefore, in this case, the limit of the numerator of $g(t)$ as t tends to a from above is 0, because $a = -\frac{1}{2} - n(c')$. In the second case, $a = -\frac{1}{2} - n(c', \tilde{f}_{j_0})$, and thus the limit of the numerator of $g(t)$ as t tends to a from above is 0 as well, because the fraction (12) with $j = j_0$ tends to 0. Moreover, in both cases, the numerator of $g(t)$ is 0 when $t = a$, since $\frac{0}{0}$ is interpreted as 1. This proves the first expression of (13) and hence the theorem. □

Proof of Theorem 2. Let l_d, π', π'', r be the functions on Θ defined by

$$l_d(\theta) = \prod_{c \in \mathcal{C}} \left(\theta_c^{n(c)} \prod_{j=1}^{m} \prod_{f_j \in \mathcal{F}_j} \theta_{f_j|c}^{n(c,f_j)} \right), \quad r(\theta) = \frac{\pi'(\theta)}{\pi''(\theta)},$$

$$\pi'(\theta) = \theta_{c'} \prod_{j=1}^{m} \theta_{\tilde{f}_j|c'}, \quad \pi''(\theta) = \theta_{c''} \prod_{j=1}^{m} \theta_{\tilde{f}_j|c''},$$

for all $\theta \in \Theta$. Then, up to normalisation, the considered likelihood function lik corresponds to $l_d\,(\pi' + \pi'')$, since $l_d(\theta)$ is the probability of the observed data set $\mathcal{D}$ according to the NBC specified by θ, while $\pi'(\theta)$ and $\pi''(\theta)$ are

the probabilities of the instance under consideration (according to the NBC specified by θ), when its class is c' and c'', respectively. Therefore, in particular, $r(\theta)$ corresponds to the left-hand side of (10).

For each $t \in [a, b]$, consider now the function

$$l_d\,(\pi')^{\frac{1}{2}+t}\,(\pi'')^{\frac{1}{2}-t} = l_d\,\pi'\,r^{t-\frac{1}{2}} = l_d\,\pi''\,r^{t+\frac{1}{2}}. \qquad (14)$$

This function corresponds to the function l_d with modified counts n (which are in general not integer anymore, but still nonnegative), and can be easily maximised. Its maximum is taken in $\hat{\theta}(t)$, where $\hat{\theta}(t)$ is the maximum likelihood quantification of the NBC with respect to the modified counts: that is,

$$\hat{\theta}(t)_{c'} = \frac{n(c') + \frac{1}{2} + t}{n(\cdot) + 1}, \quad \hat{\theta}(t)_{\tilde{f}_j|c'} = \frac{n(c', \tilde{f}_j) + \frac{1}{2} + t}{n(c') + \frac{1}{2} + t},$$

$$\hat{\theta}(t)_{f_j|c'} = \frac{n(c', f_j)}{n(c') + \frac{1}{2} + t} \quad \text{for all } f_j \neq \tilde{f}_j,$$

$$\hat{\theta}(t)_{c''} = \frac{n(c'') + \frac{1}{2} - t}{n(\cdot) + 1}, \quad \hat{\theta}(t)_{\tilde{f}_j|c''} = \frac{n(c'', \tilde{f}_j) + \frac{1}{2} - t}{n(c'') + \frac{1}{2} - t},$$

$$\hat{\theta}(t)_{f_j|c''} = \frac{n(c'', f_j)}{n(c'') + \frac{1}{2} - t} \quad \text{for all } f_j \neq \tilde{f}_j,$$

$$\hat{\theta}(t)_{c} = \frac{n(c)}{n(\cdot) + 1} \quad \text{and} \quad \hat{\theta}(t)_{f_j|c} = \frac{n(c, f_j)}{n(c)} \quad \text{for all } f_j,$$

where c is any class different from c', c''. Therefore, in particular, $r(\hat{\theta}(t))$ corresponds to the right-hand side of (10): that is, $\hat{\theta}(t) \in \Theta_t$.

Since $\hat{\theta}(t)$ maximises the function (14) over all $\theta \in \Theta$, it also maximises both functions $l_d\,\pi'$ and $l_d\,\pi''$ over all $\theta \in \Theta$ such that $r(\theta) = r(\hat{\theta}(t))$. That is, $\hat{\theta}(t)$ maximises both functions $l_d\,\pi'$ and $l_d\,\pi''$ over all $\theta \in \Theta_t$, and therefore it also maximises their sum $l_d\,(\pi' + \pi'')$ over all $\theta \in \Theta_t$. Since this last function corresponds, up to normalisation, to the considered likelihood function lik, we obtain the result $L(t) = lik(\hat{\theta}(t))$.

In order to prove Theorem 2, it suffices to show that $lik(\hat{\theta}(\cdot))$ is proportional to $l'\,l''\,(p' + p'')$; that is, it suffices to show that

$$l_d(\hat{\theta}(t))\Big(\pi'(\hat{\theta}(t)) + \pi''(\hat{\theta}(t))\Big) = \gamma\,l'(t)\,l''(t)\,\big(p'(t) + p''(t)\big),$$

where the proportionality constant $\gamma \in (0, +\infty)$ may depend on anything but t. Since

$$\pi'(\hat{\theta}(t)) + \pi''(\hat{\theta}(t)) = \frac{1}{n(\cdot) + 1}\,\big(p'(t) + p''(t)\big)\,,$$

it only remains to show that $l_d(\hat{\theta}(t))$ is proportional to $l'(t)\,l''(t)$. In the expression $l_d(\hat{\theta}(t))$, we can drop all factors for classes c different from c', c'', because $\hat{\theta}(t)_c$ and $\hat{\theta}(t)_{f_j|c}$ do not depend on t when c is different from c', c''. The desired result follows easily when one considers that

$$n(c) = \sum_{f_j \in \mathcal{F}_j} n(c, f_j)$$

for all $c \in \mathcal{C}$ and all $j \in \{1, \ldots, m\}$. □

References

[1] M. Cattaneo. *Statistical Decisions Based Directly on the Likelihood Function*. PhD thesis, ETH Zurich, 2007.

[2] M. Cattaneo. A generalization of credal networks. In *ISIPTA '09*, pages 79–88. SIPTA, 2009.

[3] M. Cattaneo. Likelihood-based inference for probabilistic graphical models: Some preliminary results. In *PGM 2010*, pages 57–64. HIIT Publications, 2010.

[4] G. Corani and A. Benavoli. Restricting the IDM for classification. In *IPMU 2010*, pages 328–337. Springer, 2010.

[5] G. Corani and M. Zaffalon. Lazy naive credal classifier. In *Proc. 1st ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data*, pages 30–37. ACM, 2009.

[6] F.G. Cozman. Credal networks. *Artif. Intell.*, 120:199–233, 2000.

[7] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learn.*, 29:103–130, 1997.

[8] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian networks classifiers. *Mach. Learn.*, 29:131–163, 1997.

[9] D.J. Hand and K. Yu. Idiot's Bayes—Not so stupid after all? *Int. Stat. Rev.*, 69:385–398, 2001.

[10] D.J. Hudson. Interval estimation from the likelihood function. *J. R. Stat. Soc., Ser. B*, 33:256–262, 1971.

[11] D. Koller and N. Friedman. *Probabilistic Graphical Models*. MIT press, 2009.

[12] M.G. Madden. On the classification performance of TAN and general Bayesian networks. *Knowl.-Based Syst.*, 22:489–495, 2009.

[13] S. Moral. Calculating uncertainty intervals from conditional convex sets of probabilities. In *UAI '92*, pages 199–206. Morgan Kaufmann, 1992.

[14] Y. Pawitan. *In All Likelihood*. Oxford University Press, 2001.

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

[16] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. *J. R. Stat. Soc., Ser. B*, 58:3–34, 1996.

[17] M. Zaffalon. Statistical inference of the naive credal classifier. In *ISIPTA '01*, pages 384–393, 2001.

[18] M. Zaffalon. The naive credal classifier. *J. Stat. Plann. Inference*, 105:5–21, 2002.

The Description/Experience Gap in the Case of Uncertainty

Horacio Arlo-Costa
Carnegie Mellon University,
Philosophy,
hcosta@andrew.cmu.edu

Varun Dutt
Carnegie Mellon University,
EPP*,
varundutt@cmu.edu

Cleotilde Gonzalez
Carnegie Mellon University,
SDS†,
coty@cmu.edu

Jeffrey Helzner
Columbia University,
Philosophy,
jh2239@columbia.edu

Abstract

We present empirical evidence indicating the existence of a description/experience gap for decisions under uncertainty. The nature of the gap is different than the one arising in the case of risk but both phenomena depend essentially on the use of limited sampling in experience. While subjects are ambiguity averse in description they are robustly ambiguity seeking in experience. A probabilistic explanation of this effect is provided as well as conjectures about the possibility of studying the effect with descriptive theories like Cumulative Prospect Theory.

Keywords. uncertainty, descriptive, normative, experience, description

1 Background

Traditionally, beliefs and desires are represented by subjective probabilities and utilities, respectively, and these subjective probabilities and utilities are combined in the calculation of expectations. This expected utility tradition is dominant within the decision sciences, extending from cases of decision making under risk, where objective probabilities are available to the decision maker, to cases of decision making under uncertainty, where information about objective probabilities is scarce [17].

The familiar axiomatizations of the expected utility hypothesis, from von Neumann-Morgenstern to Anscombe-Aumann to Savage (see [15] for an introductory presentation), are usually interpreted normatively, but they also serve as a diagnostic tool in that systematic deviations from their requirements are interpreted as pathology exhibited in human behavior and in need of explanation. The following example, one among a class of examples made famous by Allais [1], serves to illustrate the point: The subject is presented with two decision problems, each consisting of a pair of risky alternatives. In the first problem the subject is given a choice between a lottery A that pays \$4000 with probability 0.8 and \$0 with probability 0.2 and a lottery B that pays \$3000 with probability 1. In the second problem the subject is given a choice between a lottery C that pays \$4000 with probability 0.2 and \$0 with probability 0.8 and a lottery D that pays \$3000 with probability 0.25 and \$0 with probability 0.75. It has been observed that a significant number of subjects choose B in the first decision problem and C in the second decision problem. Assuming that such choices reveal strict preferences they are incompatible with the expected utility hypothesis: if B is strictly preferred to A, then the expected utility hypothesis requires a strict preference for the compound lottery that rewards B with probability 0.25 and \$0 with probability 0.75 over the compound lottery that rewards A with probability 0.25 and \$0 with probability 0.75 and, moreover, the expected utility maximizer must be indifferent between the first of these compound lotteries and D as well as between the second of these compound lotteries and C.

How are these observed deviations from expected utility theory to be interpreted? More generally, what is the significance of such deviations? According to one important class of interpretations such observed deviations are evidence that the normative theories at issue are not adequate when it comes to describing the decisions of human agents – but remain nonetheless valid normatively. According to another class of interpretations such deviations can be evidence that the theory being violated is inadequate as a normative theory of decision making – this is essentially the sort of interpretation that Ellsberg took in response to the violations that he made famous in connection with Savage's theory [6]. For now we will focus on the first class of interpretations that was mentioned. Work on this class of has been dominated by two schools. Essential examples of the first of these schools can be found in the previously mentioned work of Simon [18]

* Engineering and Public Policy.
† Social and Decision Science.

and Gigerenzer [11]. A basic theme of such work is that deviations from a normative theory such as expected utility maximization are often just the result of computational limitations and that such deviations are not necessarily a sign of irrationality. Essential examples of the second of these schools is provided by the work of Kahneman and Tversky [14]. A basic theme of such work is that human decision processes, much like human senses, are subject to illusions and that these illusions lead to systematic deviations from expected utility and related norms. Work done in both these schools is potentially significant. For now we will focus on work done in the second of the two schools that were mentioned.

Let us consider what is perhaps the most well-known theory from this second school that attempts to address deviations from expected utility theory such as those associated with the Allais-type example mentioned previously. Roughly, *prospect theory* posits two phases of decision making. The first of these is an editing phase during which various operations (e.g., coding) are applied to the information that is available to the decision maker so that it can be arranged into an appropriate form. The second phase is concerned primarily with evaluation. The basic idea is that the various alternatives are assessed in terms of an index that is similar to expected utility but with "decision weights" replacing the probabilities and a "value function" replacing the utilities. The decision weights can be represented in terms of a weighting function π on the objective probabilities that are assumed to be accessible to the decision maker in the context of decision making under risk. According to Kahneman and Tversky, "decision weights measure the impact of events on the desirability of prospects, and not merely the perceived likelihood of these events." [14]. Presumably, according to this view there are a significant number of cases where differences between $\pi(p)$ and p indicate a pathology of systematic deviations from the expected utility hypothesis. Through an appeal to empirical and theoretical considerations, Kahneman and Tversky argue that these decision weights satisfy certain structural requirements, e.g., the overweighting of small probabilities. They also provide arguments, both theoretical and empirical, to show that the value function v of prospect theory, which is defined on "changes in wealth or welfare, rather than final states" satisfies certain structural requirements, e.g., concavity for gains.

We now turn to an example that illustrates the manner in which prospect theory is tested in [13]. Recall from the previous discussion that prospect theory predicts the overweighting of small probabilities, i.e., $\pi(p) > p$ for small p. Kahneman and Tversky perform the following experiment to test this prediction: Each subject in the study is asked to choose from a pair of alternatives. One of these alternatives is a lottery that pays \$5000 with probability 0.001 and pays \$0 with probability 0.999. The other alternative pays \$5 with certainty. Kahneman and Tversky [14] report that a majority of subjects have a strict preference for the first of the two alternatives just described. Consider a subject who demonstrates these preferences. Such preferences are representable in prospect theory just in case there are π and v such that $\pi(.001)v(\$5000) + \pi(.999)v(\$0) > v(\$5)$. Following Kahneman and Tversky we set $v(\$0) = 0$ so that the previous inequality simplifies to $\pi(.001)v(\$5000) > v(\$5)$, which implies that

$$\pi(.001) > \frac{v(\$5)}{v(\$5000)}. \tag{1}$$

Finally, since v is assumed to be concave for gains it follows that

$$\frac{v(\$5)}{v(\$5000)} \geq .001. \tag{2}$$

Combining inequalities (1) and (2) yields $\pi(.001) > .001$ as predicted.

In the experiment discussed in the previous paragraph subjects were presented with a menu of alternatives and a description of the relevant probabilities. Use of this sort of empirical methodology is widespread among work on the psychology of decision making. But to what extent does empirical support, such as that which was just discussed in connection with the overweighting of small probabilities, depend on this methodological choice? One might respond by maintaining that such a question presupposes that there are other plausible methodologies. An important example of an alternative methodology is the "sampling paradigm" that is used in more recent work such as [13]. For our purposes, the essential difference between this alternative methodology and the sort of approach that was taken in [14] is that in the former subjects get their information about the relevant probabilities through sampling rather than by reading a text description.

To illustrate the difference between the two approaches that were just mentioned, consider the following type of experiment from [13]: Divide the subjects into two groups. Subjects in the first group are given the previously discussed task from [13] in connection with the underweighting of small probabilities. That is, subjects in this first group are asked to choose between A, a lottery that pays \$5000 (\$0) with probability 0.001 (.999), and B, an alternative that pays \$5 with certainty. Subjects in the second group are asked to choose between pressing one of two

buttons, A and B, on a computer screen. Although the subjects in this group are never given such information, A is a chance setup that rewards either \$5000 or \$0 at the end of each trial and, furthermore, the objective probabilities (i.e., limiting frequencies) that are associated with A are .001 and .999 for \$5000 and \$0, respectively. Similarly, button B is a chance setup that rewards \$5 with probability 1. Finally, although subjects in the second group are not told the probabilities associated with A and B, they are permitted to sample both buttons as many times as they desire before making their decision between the two alternatives. It should be clear that the crucial distinction between the task that is given to the first group of subjects and the task that is given to the second is essentially the aforementioned distinction between first and second of the two empirical methodologies under consideration.

Let us assume that the subjects in the first group reveal preferences that are consistent with what Kahneman and Tversky observed in connection with the underweighting of small probabilities. Do we expect that the preferences that are revealed in the second group to essentially parallel those that are revealed in the first? Hertwig et al. have argued that we should not. Indeed, Hertwig et al., through experiments of the sort just mentioned, have shown that certain psychological effects – e.g., the overweighting of rare events – are not preserved when one changes from a description-based approach to an experience-based approach, and this lack of preservation is known as *experience-description gap* – as a matter of fact rare events are *underweighted* in experience. The gap is difficult to explain by appealing to theories like Prospect Theory.

Fox and Hadar have recently expressed criticisms in [10] about some of the claims presented in [13] concerning a possible experience-description gap. We will now consider two of the main theoretical criticisms that are discussed in [10]. First, Fox and Hadar do not believe that Hertwig et al. were sufficiently clear about what counts as experienced-based decision making [EBDM]: "The generalization that EBDM differs from DBDM is difficult to evaluate because, surprisingly, no one has yet defined 'experience-based decision making." [10] [1] Noting this lack of an adequate definition of experienced-based decision making, Fox and Hadar offer what they take to be an adequate characterization of EBDM. The upshot of their analysis is that "[...] EBDM applies to any situation in which there is uncertainty and learning through sampling." This point, which is significant, will be discussed later in this paper. For now, we turn to a matter that is more directly related to the Fox and Hadar's charge that EBDM had not been given an adequate definition.

We think that the analysis of EBDM given in [10] is not well-suited to a study of the experience-description gap as understood in [13]. In particular, the analysis that is supplied in [10] does not say anything about what it means to be an experience-based counterpart to a given description-based task, something which is crucial to the interpretation of the work that is reported in [13]. In light of this, Arló-Costa and Helzner [2] proposed the following analysis of this counterpart relation that is essential if one is to examine how well a given psychological effect travels across experience-description gap:

- In a *decision from description* the subject is presented with a specification of the type of chance mechanism.
- In a *decision from experience* the subject is not presented with such a specification but rather is allowed to observe the behavior of a chance mechanism that has the specified type.

In [2], Arló-Costa and Helzner suggest that this analysis might be useful in examining the extent to which an experience-description gap exists for certain psychological effects associated with decision making under uncertainty. The argument that was given in [2] on behalf of this suggestion is that, while classical descriptions of uncertainty – e.g., the Ellsberg urn – have no experiential counterparts, since the relevant uncertainties in such cases are epistemic, one can specify mechanisms that, at least psychologically, approximate descriptions of uncertainty and, moreover, have an experiential counterpart in the sense of Arló-Costa and Helzner's analysis of the counterpart relation. In the next section we will examine recent experimental work concerning the way effects associated with these approximations of uncertainty can vary as one moves from EBDM to DBDM. It is worth noting that the present article may be seen as building on the approach considered in [21] and [5]. In a more recent paper Yoram Halevy [12] makes a forceful case for establishing a strong correlation between ambiguity neutrality and the reduction of compound objective lotteries. Halevy concludes that his results suggest that failure to reduce compound (objective) lotteries is the underlying factor of the Ellsberg paradox. We do not want to make such a strong claim but we rely on the idea that a chance setup like B^*, as described in what follows, can be treated as an operational approximation of uncertainty.

[1] DBDM of course refers to description-based decision making.

2 Experimental Work

Example 1 (Ellsberg's two-color problem [6]) *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 \$25 (\$0) if the next random selection from Urn A results in black (white) ball? Repeat then the same question for Urn B.

It is well known that subjects tend to offer higher maximum buying prices for urn A than for urn B. This seems to be so even in non-comparative cases (see [4] and [3]) contrary to the so-called *comparative ignorance hypothesis* formulated in [9]. 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.

In a previous ISIPTA paper Arló-Costa and Helzner reported experimental results indicating that maximum buying prices for B^* are intermediate with respect to the ones for A and B. This confirms previous results reported in [21] and [5]. In a more recent paper Yoram Halevy [12] makes a forceful case for establishing a strong correlation between ambiguity neutrality and the reduction of compound objective lotteries (that would lead to treat urns A and B^* equally). He therefore concludes that his results suggest that failure to reduce compound (objective) lotteries is the underlying factor of the Ellsberg paradox. We do not want to claim something as strong as that but we rely on the idea that B^* can be treated as an operational approximation of urn B. The interest of this move is that B^* is easily implementable in experience while it is notoriously difficult to find an experiential counterpart of B. The main experimental finding reported below is that while in description subjects are averse to ambiguity (they prefer C over B^* and B) in experience this effect is reversed and subjects are ambiguity seeking (they prefer B^* over C – B has no experiential counterpart). This shows that the description-experience gap also appears (in a different form) for decisions under uncertainty.

3 Method: First Experiment

One hundred and nineteen students at Carnegie Mellon University (Pittsburgh, USA) were presented with the three decision problems presented below. Maximum buying prices for these games were requested. The options C, B^* and B described above were implemented in the following way:

C**:** A fair chance setup with possible outcomes $\{1, 2, \ldots, 99, 100\}$ has been constructed. If the outcome on the next run of this setup is less than or equal to 50, then you win \$25. Otherwise, you get \$0.

B^***:** Two fair chance setups, I and II, have been constructed. Setup I has possible outcomes $\{0, 1, \ldots, 99, 100\}$. Setup II has possible outcomes $\{1, \ldots, 99, 100\}$. The game is played by first running setup I and then running setup II. If the outcome of the run of setup II is less than or equal to the outcome from the run of setup I, then you win \$25. Otherwise, you get \$0.

B**:** An integer n has been selected from the set $\{0, 1, \ldots, 99, 100\}$. Nothing is known about the mechanism by which n has been selected. A fair chance setup with possible outcomes $\{1, \ldots, 99, 100\}$ has been constructed. If the outcome on the next run of this setup is less than or equal to n, then you win \$25. Otherwise, you get \$0.

3.1 The Description Condition

Fifty eight students from the pool of one hundred and nineteen students mentioned above faced the description condition for the first experiment. We have two types of trials. In the first type we consider gains. The subjects face a computer window with two rectangles containing the text used above to describe the options C and B^*. The subjects in this condition are asked the following question: **Which one out of the two games will you choose to play?**. They then have three possible options for a response:

Left Button: You were indifferent between the two alternatives. (A)

Middle Button: You had a strict preference for one of the alternatives. (B)

Right Button: Neither (A) nor (B) reflect my attitudes

The second type of trials involved losses. For example the loss version of the C-option is: 'A fair

chance setup with possible outcomes $\{1, \ldots, 99, 100\}$ has been constructed. If the outcome on the next run of this setup is less than or equal to 50, then you lose 25. Otherwise, you get 0.'

3.2 The Experience Condition

Sixty one students from the pool of one hundred and nineteen students mentioned above faced the experience condition for the first experiment. In the experience condition the subjects faced two rectangles containing the labels C and V. They can sample these options by clicking on them. Clicking the option C triggers a random selection from $\{1, 2, \ldots, 99, 100\}$. If the outcome is less than or equal to 50, then the subject is told that he won \$25. Otherwise, he is told that he got \$0. So, this option corresponds to option C in description. Clicking the option V yields an output obtained by triggering the double sampling procedure B^* presented in description. For example, if one clicks on V a number is selected at random in the set $\{0, 1, \ldots, 99, 100\}$ and then another from $\{1, 2, \ldots, 99, 100\}$. If the outcome of the run of second selection is less than or equal to the outcome from the first selection, then the subject is told that he won \$25. Otherwise, he gets \$0. So, button V is the experiential counterpart of option B^*. The subjects can sample as much as they want, and then they make a final selection of the C or V button. Sampling is used as follows: after the subject presses V, for example, she has the option of sampling again *the same game* selected at the second stage. The sampling option is also available for the C button. Of course, in this case one continues to sample the unique game implemented for this button (another random number will be generated from $\{1, 2, \ldots, 99, 100\}$ and if the outcome is less than or equal to 50, then the subject is told that he won \$25. Otherwise, he is told that he got \$0.

As in the description condition the subjects have three buttons at their disposal to respond to. Clicking the middle button, as before, reveals a strict preference for one of the two options. A gain and a loss version of C and B^* were implemented.

4 Results: First Experiment

At the end of the experiment subjects were asked to provide maximum buying prices for options C, B and B^* as presented in the description condition. This was done for the subjects in the experience condition and for the subjects in the description condition. So, it makes sense to pool subjects from both conditions in order to compute results. Confirming previous results presented in ISIPTA (see [4] and [3]) there is a significant difference between maximum buying prices for options C and B even when the subjects do not compare vague and clear options. And confirming results reported in [2] option B^* appears as an intermediate option between options C and B. A few remarks are in order before presenting the results from this first experiment: First, although we recognize that doing multiple comparisons to test for independent hypotheses might inflate the Type 1 error, it is important to note that we only compare experience with description (a single hypothesis) in different ways (for gain, for loss, and for pooled gain and loss). Thus, we do not consider multiple hypotheses and do not need the Bonferroni correction of $\alpha = \frac{.05}{3} = 0.02$ (to reduce the inflation in Type 1 error). Second, it is important to note that we are not using a normal approximation on small samples. We tested for normality of the data and we found the data to be non-normal in both experience and description conditions of experiments 1 and 2 using separate Shapiro-Wilk tests. For example, the data were non-normal in both the experience and description conditions (experience: $D(124) = .64$, $p < .001$; description: $D(122) = .63$, $p < .001$). Again, the data was non-normal in both experience and description conditions in experiment 2. Therefore, we used non-parametric Mann-Whitney tests to evaluate significant differences between experience and description conditions. In fact, a binomial distribution assumption, as the one anonymous referee suggested, might also not be a correct assumption. Therefore, the safest thing for us to do was to report non-parametric statistics, as we do in the current paper. The Z-score that we report as part of the statistics belongs to this Mann-Whitney non-parametric test.

The following results were obtained.

Condition	Alternative	Max. Buying Price
Exp. + Desc.	B	4.93 (n = 119)
Exp. + Desc.	B^*	6.36 (n = 119)
Exp. + Desc.	C	7.04 (n = 119)

Table 1

Alternative C is significantly greater than alternative B^*, with T=622, Z=-2.329, $p < .05$, Effect Size = -0.15. Alternative C is significantly greater than Alternative B , with T=527, Z=-4.751, $p < .001$, Effect Size = -0.31. Alternative B^* is significantly greater than Alternative B, with T=411, Z=-3.716, $p < .001$, Effect Size = -0.24. So, as we explained above, alternative B^* can be used as an operational proxy of condition B in experience.

4.1 The Description condition: Results

The main hypothesis here is that subjects will be ambiguity averse (will significantly prefer C to B^*). The numbers in the tables are the number of subjects clicking the corresponding buttons for each alternative. The boldfaced results for the pooled population show the magnitude of the effect.

Alternative	Left Button	Middle B.	Right B.
B^*	10	11	6
C	11	18	5

Table 2. Gain Trials (Description)

Alternative	Left Button	Middle B.	Right B.
B^*	8	14	5
C	7	20	7

Table 3. Loss Trials (Description)

Alternative	Left Button	Middle B.	Right B.
B^*	18	**25**	11
C	18	**38**	12

Table 4.Gain and Loss Trials (Description)

4.2 The Experience condition: Results

Here the hypothesis is the subjects will be ambiguity seekers. The hypothesis is confirmed by the following results.

Alternative	Left Button	Middle B.	Right B.
B^*	10	17	4
C	14	11	6

Table 5. Gain Trials (Experience)

Alternative	Left Button	Middle B.	Right B.
B^*	16	15	2
C	10	12	7

Table 6. Loss Trials (Experience)

Alternative	Left Button	Middle B.	Right B.
B^*	26	**32**	6
C	24	**23**	13

Table 7. Gain and Loss Trials (Experience)

One index that seems interesting is based on computing the proportion of subjects who expressed a strict preference for C, both in Description and Experience. These are the subjects who clicked the Middle Button and C in experience or the subjects who clicked the Middle Button in description expressing a strict preference for the 'clear' 50-50 lottery. We will refer indistinctly to these subjects as 'Middle & C' subjects. The analysis reveals the following:

Gain Trials: Proportion Middle & C buttons (Description) (**18/29 = .62**) = Proportion Middle & C buttons (Experience) (**11/28 = .4**), with U=314, Z=-1.705, p = .09, Effect Size = -0.23.

Loss Trials: Proportion Middle & C buttons (Description) (**20/34 = .58**) = Proportion Middle & C buttons (Experience) (**12/27 = .4**), with U=393, Z=-1.108, p = .27, Effect Size = -0.14.

Gain and Loss Trials: Proportion Middle & C buttons (Description) (**38/63 = .6**) > Proportion Middle & C buttons (Experience) (**23/55 = .4**), with U=1412, Z=-1.998, p ¡ .05, Effect Size = -0.18.

It is very interesting to notice that the proportions of Middle & C subjects in Description remains almost constant for both gain and loss trials (minimum and maximum values are, respectively, .58 and .62). By the same token the proportion of Middle & C subjects in Experience remains exactly constant with a value of .4. The constancy of the proportions across conditions is clearly depicted in Figure 1 below. Nevertheless the proportion of Middle & C subjects in Description is not significantly different from the proportion of Middle & C subjects in Experience for both the gain and loss trails takes separately. But when the two types of trials are pooled the proportion of Middle & C subjects in Description is indeed significantly greater than the proportion of Middle & C subjects in Experience.

We believe that the reason why significant results are not obtained for gain or loss trials separately is because we do not have enough subjects in these type of trials taken separately. But it seems that pooling the data for these two type of trials makes sense given that the proportions remain constant across the different types. In other words, the effect seems to have the same polarity in both types of trials.

The analysis of the pooled data reveals that subjects were more ambiguity-averse in the description condition than in the experience condition. To put this in other terms, if we compute the ratio R_D of the number of Middle & C subjects in Description divided by the number of Middle & B^*, and we compute the corresponding ratio R_E in Experience, we have that $R_D = \frac{1}{R_E}$.

We collected additional experience data in a second experiment in order to see whether we can observe significant effects not only for the pooled population

but also for gains and losses. The results are reported in the next secion.

5 Method: Second Experiment

Thirty students at Carnegie Mellon University (Pittsburgh, USA) participated in a second experiment. They faced an experiential version of the first part of the experiment. Of course in this case we can only implement C and B^*.

6 Results: Second Experiment

Since we already had data for description we selected at random thirty subjects from the first experiment. First we report the maximum buying prices for the conditions C, B^* and B in description for the selected subjects of the first experiment and for the experiential version of B^* and C in experience. In this experiential version the subjects face the V and C buttons used in experience. There is a preparatory phase where they can see results from each button and then maximum buying prices for each alternative are requested.

Condition	Alternative	Max. Buying Price
Description	C	5.38 (n = 30; SD = 3.61)
Description	B^*	4.71 (n = 30; SD = 3.81)
Description	B	3.57 (n = 30; SD = 3.39)
Experience	C	6.25 (n = 30; SD = 4.05)
Experience	B^*	5.75 (n = 30; SD = 4.75)

Table 8

Although in description the maximum buying prices for B^* also occupy an intermediate position between prices for B and C, the difference between C and B^* is not statistically significant (T=118, Z=-1.401, p = .16, Effect Size = -0.18). But this seems to be due to the fact that the effect verified in the larger population of the first experiment is not verified in this arbitrarily selected sub-population.

There is a clear effect verified in the first experiment and in a previous paper [2] according to which in description the mean maximum buying prices for C are higher than the mean maximum buying prices for B^*. Moreover the values for B^* appear as intermediate between C and B. This effect seems to disappear or suffer a complete inversion in experience.This is partly verified by considering mean maximum buying prices. In fact, mean maximum buying prices for B^* and C cannot be distinguished statistically in experience (T=172, Z=-1.037, p = .30, Effect Size = -0.13). A more clear reversal is verified in the following experiments for gains and loses. This manipulation repeats the design used in the first experiment. Rather than providing buying prices the subjects express preferences for the different buttons displayed in their screens.

6.1 The Experience condition: Results

The following results were observed for gain and loss trails for experience:

Alternative	Left Button	Middle B.	Right B.
B^*	4	11	3
C	5	4	3

Table 9. Gain Trials (Experience)

Now here we can see the first clear reversal for experience of the pattern $B^* < C$ for description. In fact we verify here that $B^* > C$ is indeed statistically significant in spite of the relatively small size of the population ($p < .001$, Effect Size =- 0.47).

Alternative	Left Button	Middle B.	Right B.
B^*	6	12	5
C	3	2	2

Table 10. Loss Trials (Experience)

It is interesting to see that the pattern gets repeated here *also for losses* and with a similar ratio. We do have as above $B^* > C$ is indeed statistically significant ($p < .01$, Effect Size =- 0.71).

Alternative	Left Button	Middle B.	Right B.
B^*	10	**23**	8
C	8	**6**	5

Table 11. Gain and Loss Trials (Experience)

And, of course, we do have the same effect verified for the pooled population. $B^* > C$ is indeed statistically significant ($p < .001$, Effect Size = -0.58). Moreover now we have an even nicer result paralleling the one obtained in the first experiment:

Gain Trials: Proportion Middle & C buttons (Description) (**17/24**) > Proportion Middle & C buttons (Experience) (**4/11**), with U=101, Z=-2.657, p < .01, Effect Size = -0.45.

Loss Trials: Proportion Middle & C buttons (Description) (**14/26**) = Proportion Middle & C buttons (Experience) (**2/14**), with U=110, Z=-2.405, p < .05, Effect Size = -0.38.

Gain and Loss Trials: Proportion Middle & C buttons (Description) (**31/50 = .62**) > Proportion Middle & C buttons (Experience) (**6/29**), with U=426, Z=-3.524, p < .001, Effect Size = -0.40.

So, the second experiment verifies the effect (the fact that subjects are ambiguity seeking in experience) seen in the first experiment for the pooled population. In addition the effect also holds for gains and loses separately and it holds with similar intensity in both conditions.

7 Discussion

It is tempting to reason as follows: a play in the chance set up B^* is equivalent to a play on chance set up C. The line of reasoning is roughly as follows: The random selection in the first stage of B^* 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 chance set up C. First note that if this line of reasoning were correct then the results presented in this paper would be rather surprising. Perhaps B^* and C can be distinguished in description (due to cognitive limitations of the players), but the two chance set ups should not be distinguishable in experience according to such an argument. However, arguments of the given sort are mistaken as they fail to account for the interaction between the subject's choices and the frequencies that are observed. For example, consider the following set up which is basically equivalent to the one we implemented. Suppose that you have a sequence of 101 possible urns with black and white balls. Each urn contains 100 balls in total but the proportion of white and black changes in each case. The subject could generate output from B^* by employing a strategy where according to which she samples from the current urn until she sees a white ball and, upon seeing a white ball, advances to the next urn in the arrangement. This strategy should eventually stabilize on the all black urn so that the observed frequencies converge to those associated with the all black urn.

7.1 Prospect Theory and its capacity to model the gap experience-description

We explained above that Fox and Hadar offered in [10] an ingenious explanation of the gap experience-description by appealing to a version of Prospect Theory applicable to decisions under uncertainty rather than risk. This version of Prospect Theory is presented in detail in the recent (and excellent) book by Peter Wakker on Prospect Theory [16] (the theory was presented first in [20]). The central idea of the theory is to use event-decision weights rather than probability-based decision weights. In fact if P is the probability used for risk we can define a function W on events by applying decision weights to P. So we have that $W(E) = w(P(E))$. Since w can be non-linear, W need not be additive. The corresponding function has the properties of a capacity.

The idea that Fox and Hadar considered in the aforementioned paper is to apply the decision weight w to *judged probabilities* rather than the objective probabilities of the lotteries considered in the case of risk. This ingenious move fits the data reasonably well. So, one can claim that decisions form experience are essentially decisions under uncertainty and one can appeal to Cumulative Prospect Theory to analyze the data. The event-decision weights are calculated by appealing to judged rather than risky probabilities. These judged probabilities are estimated in terms of observed frequencies through sampling.

Is it possible to do something similar in the case we are studying? Perhaps there is a possible strategy one can use to test the predictive power or Cumulative Prospect Theory in this setting. To see the point it is important to stress that we do agree with Hadar and Fox about the fact that decisions from experience are cases of decisions under uncertainty. Let's fist see how this applies to our experiment. For each particular play of V the subject can do some sampling and obtain a judged probability in the sense of Hadar and Fox. So, it seems that one does not have any alternative except representing the subjects playing V as entertaining a *set* of judged probabilities. The only thing that the subjects know is that there is s chance set-up that is producing a set of probabilities that he can estimate by repeated sampling. But he knows nothing about the nature of the chance set up that produces this set of probabilities. In particular the chance set up that is producing the set of probabilities need not obey the law of large numbers as in the experimental set up used by Stecher at al. [19]. In this case even the computation of wining frequencies would lead to erroneous estimation of the chances of

the chance set up that generates the probabilities.[2].

So, when one implements B_S^* in experience by fixing a sampling rule various types of indeterminacy arise. First the implemented game does not have a single objective long run frequency associated to it. Subjects can employ different playing strategies that are associated with different long run frequencies. Second, under the point of view of the subject who plays his probabilities are indeterminate also. He only knows that there is a chance set up that produces sets of probabilities that he can eventually estimate. A sophisticated player can learn that the set of probabilities associated with C are produced by a chance set up of objective probability 0.5.[3] And if the subject has a fixed playing strategy he can perhaps learn the objective probability corresponding to this strategy. But most players will not use a fixed strategy and in this case it seems that there is no learnable objective chance associated to the chance set up that produces the set of probabilities associated with B_S^*.

So, the probabilities of the subjects remain undetermined. This is so even if we guarantee that the winning frequencies of the two chance set ups converge to the same number (for example by guaranteeing fair sampling procedures for plays of B^*), in the short run agents cannot but remain uncertain about the probabilities of the two chance set ups in experience.

Is there a way of connecting this set of priors with event-decision weights? Here is a possible way of doing it. Call the set of priors $\mathbf{C}$. Then for each event E we have the interval $I_E = \{P(E) : P \in \mathbf{C}\}$. Now, one can define an event-decision weight W as $W = inf(I_E)$ or $W = sup(I_E)$. Is it possible to approximate out empirical results via this procedure? We propose a careful investigation of this issue for future work. [4]

We can point out here that a theory like Cumulative Prospect Theory will tend to predict asymmetries regarding ambiguity for gains and loses. The majority of the existing evidence seems to indicate, for example, that subjects are ambiguity seeking for losses while they are ambiguity averse for gains (see the evidence and references in [16]). This patterns does not seem to arise in our experiment. At least in the second experiment it is clear that subjects are equally ambiguity seeking for gains and losses.

The presentation of Cumulative Prospect Theory in [16] makes clear that the main idea of extending Prospect Theory to uncertainty is to avoid the representation of uncertainty via multiple priors. Wakker is quite explicit about rejecting this strategy which he sees as problematic for various reasons that have to do with measurement and elicitation. But it seems that there are experimental situations of the sort we presented in this article where the use of multiple priors seems unavoidable. In spite of this aversion to use multiple priors there might be ways of finding a connection with the way in which Cumulative Prospect Theory represents uncertainty. If this were possible a second step would consist in testing the predictive power of the extended version of Prospect Theory to uncertainty. We also propose to tackle this issue in future work.

[2]The main idea in the aforementioned paper is to use a chance set up that cannot be learned in experience. Assuming that fair sampling procedures could be used to play B^* then its objective chance could be learned by observing enough data. But in the short run the agents cannot but remain uncertain about the determinate or indeterminate chances associated with the chance set up producing the given set of probabilities.

[3]Notice that the subject does not even know that C is based on single-sampling. For each time the they play C again they do not know whether they are sampling the *same* risky lottery. So, they have to entertain as well a set of probabilities for C.

[4]See [7] and [8] for earlier discussions along these lines.

References

[1] M. Allais. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'ecole Americaine. *Econometrica* 21: 503-546, 1953.

[2] H. Arló-Costa and J. Helzner. Iterated Random Selection as Intermediate Between Risk and Uncertainty. *Proceedings of the Sixth International Symposium on Imprecise Probabilities and their applications*, eds. T. Augustin, F. P. A. Coolen, S. Moral and M. C. M. Troffaes, 2009.

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

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

[5] C.C. Chow and R. Sarin. Known, Unknown and Unknowable Uncertainties. *Theory and Decision*, 52, 127-138, 2002.

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

[7] D. Ellsberg. *Risk, Ambiguity and Decision*. Garland Publishing, 2001.

[8] H.J. Einhorn and R.M. Hogarth. Ambiguity and uncertainty in probabilistic inference. *Psychological Review*, 92, 433-461, 1985.

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

[10] C. R. Fox and L. Hadar. Decisions from experience = sampling error + prospect theory: Reconsidering Hertwig, Barron, Weber and Erev (2004). *Judgment and Decision Making*, 1, 159-161, 2006.

[11] G. Gigerenzer. *Adaptive thinking : Rationality in the real world*. New York: Oxford University Press, 2000.

[12] Y. Halevy. Ellsberg revisited. An experimental study. *Econometrica* 75(2): 503-536, 2007.

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

[14] D. Kahneman and A. Tversky. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, Vol. 47, 263-291, 1979.

[15] Kreps, D. M.: *Notes on the Theory of Choice*, Westview Press, 1988.

[16] P. Wakker. *Prospect Theory: For Risk and Ambiguity*. Cambridge University Press, 2010.

[17] L. J. Savage. *The Foundations of Statistics*. Dover Publications; 2 Revised edition, 1972.

[18] H. A. Simon, H. A. Invariants of human behavior. *Annual Review of Psychology*, 41, 1-19, 1990.

[19] J. Stecher, T. Shields and J. Dickhaut. Generating Ambiguity in the Laboratory. *Management Science*, Forthcoming.

[20] A. Tversky and D. Kahneman. Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, 5:297-323, 1992.

[21] F.J. Yates and L.G. Zukowski. Characterization of Ambiguity in Decision Making. *Behavioral Science*, 21, 19-25, 1976.

Partially identified prevalence estimation under misclassification using the Kappa coefficient

Helmut Küchenhoff
Department of Statistics
Ludwig-Maximilians-Universität
(LMU), Munich, Germany

Thomas Augustin
Department of Statistics
Ludwig-Maximilians-Universität
(LMU), Munich, Germany

Anne Kunz
Metronomia Clinical Research GmbH
Munich, Germany

Abstract

We discuss prevalence estimation under misclassification. That is we are concerned with the estimation of a proportion of units having a certain property (being diseased, showing deviant behavior, etc.) from a random sample when the true variable of interest cannot be observed, but a related proxy variable (e.g. the outcome of a diagnostic test) is available. If the misclassification probabilities were known then unbiased prevalence estimation would be possible. We focus on the frequent case where the misclassification probabilities are unknown but two independent replicate measurements have been taken. While in the traditional precise probabilistic framework a correction from this information is not possible due to non-identifiability, the imprecise probability methodology of partial identification and systematic sensitivity analysis allows to obtain valuable insights into possible bias due to misclassification. We derive tight identification intervals and corresponding confidence regions for the true prevalence, based on the often reported kappa coefficient, which condenses the information of the replicates by measuring agreement between the two measurements. Our method is illustrated in several theoretical scenarios and in an example from oral health on prevalence of caries in children.

Key words: Partial Identification; Sensitivity Analysis; Prevalence Estimation; Kappa Coefficient; Misclassification; Identification Region; Ignorance Region.

1 Introduction

Many data in social sciences, econometrics, biometrics and epidemiology are complex in the sense that the available data at hand do not exactly convey the information one is looking for. Frequently, the variables of material interest cannot be observed directly or measured correctly, and one has to be satisfied with so-called surrogates or proxies, i.e., with somehow related, but different variables. This problem of non-ascertainability of certain ideal variables is referred to as measurement error (ME in the following) if the variables are continuous and as misclassification (MC) if they are discrete variables. If one ignores the principal difference between the ideal variables and their observable counterparts and just plugs in the surrogates instead of the ideal variables ('naive estimation'), then all the parameter estimators must be suspected to be severely biased. For the distorting effects of MC in different applications, see, e.g., [8, 23, 24, 53, 55].

In the last years there has been a considerable progress how to adjust for measurement error and misclassification in statistical models. Many correction procedures are available for consistent estimation in the presence of ME or MC, see in particular the monographs [6, 17], or, e.g., [44]. Most of those procedures are based on precise information about the process of measurement (and in complex models typically on Bayesian methods with precise priors, e.g., [43]). In the case of an additive measurement error, usually the variance of measurement error has to be known or to be estimated, e.g. by replicate measurements, to enable consistent estimation. In the presence of MC, knowledge of the conditional probabilities of correct classification, in the binary case called sensitivity and specificity, allows for general estimation procedures even in complex models; see [20] and [39] for fundamental work concerned with response misclassification and, e.g., [27, 29, 30, 59] for methods handling misclassified covariates. When no such information about ME or MC is available, identification problems arise and no consistent parameter estimation is possible. Important examples include the estimation in simple linear regression with covariate ME as well as the problem of estimating probability distributions of outcomes in the presence of MC. In this paper, we examine the latter problem in the spirit of the methodology of partial identification (e.g., [32]) and systematic sensitivity analysis (e.g., [52]).

One important example for estimating probability distributions in medical and clinical research is prevalence estimation, i.e. estimating the probability that a randomly sampled person of the population has a certain property, e.g. is diseased.[1] In the presence of MC, induced, e.g., by a medical examiner or a diagnostic tool, prevalence estimation using the relative frequency ignoring MC (naive estimation) is inconsistent. In this situation, a consistent estimator is available when the conditional probabilities of correct diagnosis (sensitivity and specificity) are known or can be estimated consistently. However, estimating sensitivity and specificity using a validation study usually relies on the availability of a correct diagnostic method (gold standard) in the validation sample. If such a gold standard method is not available, then it is usual practice to replicate measurements on the same unit to get some information on the quality of the measurement procedure. In the case of the availability of three independent measurements with identical sensitivity and specificity, it is still possible to obtain consistent estimators of prevalence; for a recent discussion, see [41]. Another scenario, where the parameters are identified, is the availability of two independent measurements with identical sensitivity and specificity in two different populations, see [46].

When only two replicate measurements in one population are available, the quality of measurement can be characterized by Cohen's kappa coefficient [9], which is based on the agreement of the replicates ("inter rater reliability"). Although there is a long discussion about the problems of using the kappa coefficient (e.g. [14, 50]), it is usually reported in those studies. However, no further correction is performed, since the resulting estimation model is not well-identified, making the derivation of a precise-valued estimator impossible. In contrast, the concept of partial identification and systematic sensitivity analysis provides valuable insights into the magnitude of the misclassification bias. We derive identification regions of the misclassification probabilities and the true prevalence, and confidence regions for the latter, additionally taking sample variation into account. In our example, we use data from a validation study, which consists of a subsample of our data to estimate kappa coefficient.

We understand our contribution as a typical example where imprecise probabilistic methodology provides powerful quantitative insights into the underlying structure, while the traditional precise approach, forced to choose between the extremes 'precise solution' or 'no solution', necessarily has to surrender.

[1] For ease of argumentation and influenced by the example from oral health discussed in Section 4, we use biometric terminology throughout the paper, without limiting the application of our results to that area.

The general methodology underlying our investigation adapts recent progress in the area of partial identification and systematic sensitivity analysis for possibly deficient data, also strongly related to the conservative handling of deficient data in imprecise probability settings (e.g., [12, 49, 58]). Up to now, such methods have been mostly applied to the case of missing or coarse data with an unknown deficiency mechanism (e.g. [33, 36], for surveys), notably with regard to missingness due to counterfactuality when analysing treatment effects (see e.g. [7, 15, 25, 35, 48]). Corresponding ideas have, for instance, been proposed in general settings in [13, 18], or more specifically to handle publication bias in meta analysis [11, 21], in the reanalysis of a public opinion survey [2] or to derive tight bounds on demand responses [4], and may provide an alternative to some neighborhood models in robust statistics ([1, Section 5]). Recently partial identification has also been applied in the context of misclassification ([19, 37]).

The paper is organized as follows. In Section 2, we deduce basic formulae for the relationship between the fundamental quantities characterizing our situation, i.e. observed prevalence, sensitivity, specificity and the kappa coefficient. From that, identification regions for the true prevalence are derived. In Section 3, sampling variability is incorporated into our estimates resulting in confidence intervals. In Section 4, we apply our findings to a data set of caries research before we conclude with a brief further discussion of our approach in Section 5.

2 Prevalence Estimation under Misclassification

At the beginning of this section the basic situation is described and notation and terminology are fixed (cf. also Table 1).

We address the problem of estimating the *prevalence* of a certain disease, i.e. a probability

$$p := P(Y = 1),$$

where

$$Y = \begin{cases} 1 & \text{diseased} \\ 0 & \text{not diseased} \end{cases}$$

denotes the indicator for the (true) disease status. Due to the possible presence of MC we cannot observe Y directly, but instead the diagnosis of an examiner, which is denoted by

$$Y^* = \begin{cases} 1 & \text{diagnosis positive} \\ 0 & \text{diagnosis negative}\,. \end{cases}$$

The naive estimator $\frac{1}{n}\sum_{i=1}^{n} Y_i^*$ based on a simple random sample $Y_1^*, \ldots, Y_n^*$ of Y^* of size n is biased

and converges to $P(Y^* = 1)$. We call $p^* := P(Y^* = 1)$ the *naive prevalence* and denote the naive estimator based on the observed relative frequency by $\widehat{p^*}$.

obs. Y^*	true status Y 1	0	
1	$P(Y^* = 1\|Y = 1)$ *sens*	$P(Y^* = 1\|Y = 0)$ → false positive cases	p^*
0	$P(Y^* = 0\|Y = 1)$ → false negative cases	$P(Y^* = 0\|Y = 0)$ *spec*	
	p		

Table 1: Basic notions

The relationship between the true and the naive prevalence using sensitivity $sens := P(Y^* = 1|Y = 1)$ and specificity $spec := P(Y^* = 0|Y = 0)$ of the diagnosis is directly obtained from the law of total probability.

$$\begin{aligned} p^* &= P(Y^* = 1) \\ &= P(Y^* = 1|Y = 1) \cdot P(Y = 1) \\ &\quad + P(Y^* = 1|Y = 0) \cdot P(Y = 0) \\ &= p \cdot sens + (1 - p) \cdot (1 - spec) \end{aligned} \tag{1}$$

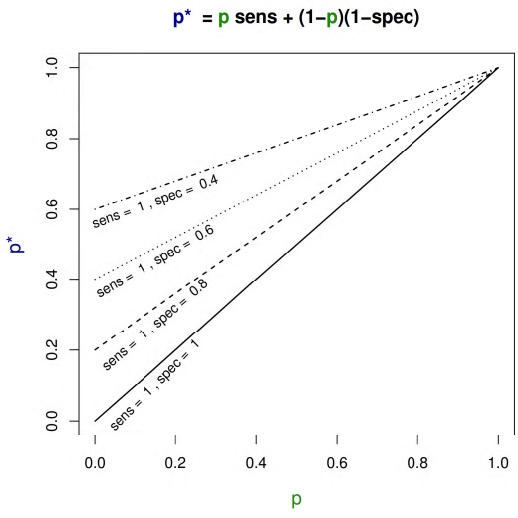


Figure 1: Illustration of misclassification bias (deviation from the angle bisector): naive (observed) prevalence p^* in dependence of the true prevalence p for different values of specificity and sensitivity = 1

Only for technical reasons we have to fix additionally the assumption that throughout the paper

$$sens + spec > 1\,. \tag{2}$$

This commonly used constraint is not a substantial restriction, since otherwise the diagnosis does not contain any useful information.

If sensitivity and specificity are known, equation (1) yields an unbiased estimator of p by

$$\hat{p} \quad = \quad \frac{\widehat{p^*} + spec - 1}{sens + spec - 1}\,. \tag{3}$$

Moreover, Equation (1) allows to illustrate the potentially rather high distorting effects of misclassification. In Figures 1 and 2 the naive prevalence p^* is plotted in dependence of the true prevalence p for different misclassification probabilities. Figure 1 shows the bias in the situation of a test with optimal sensitivity, which would detect every diseased unit, but may produce a certain amount of false positive results.

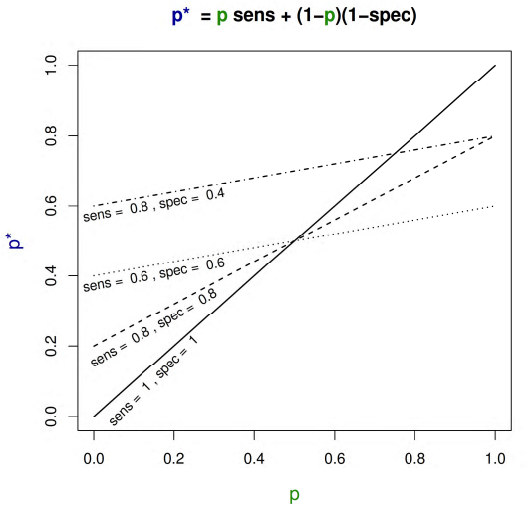


Figure 2: Illustration of misclassification bias (deviation from the angle bisector): naive (observed) prevalence p^* in dependence of the true prevalence p for different values of sensitivity and specificity

Figure 2 illustrates the more realistic situation of possibly false positive and false negative units. Note that in all situations the bias depends on the true, but unknown (!) value of p. Moreover, as in in Figure 2, the bias usually is complex in the sense that, in contrast e.g. to single-variable classical measurement error in linear regression models, even its sign can not be determined without additional knowledge. In dependence on the concrete constellation of *sense*, *spec* and p, over- and underestimation of the true prevalence is possible.

2.1 Establishing a Relationship between the Kappa Coefficient, Misclassification Probabilities and Prevalence

We now assume that we have two replicate measurements Y_1^*, Y_2^* on the same units. These replicates relate to two examiners and the data can be displayed in a 2x2 table. We define the corresponding probabilities by

$$p_{jk} := P(Y_1^* = j, Y_2^* = k), \;\; i,j = 0,1\,. \qquad (4)$$

The kappa coefficient κ as proposed by [9], see also, e.g., [42, 45] for recent developments, assesses the chance corrected agreement among the replicate measurements (inter rater agreement). The (theoretical) kappa coefficient is defined by

$$\kappa := \frac{p_o - p_e}{1 - p_e} \qquad (5)$$

$$\begin{aligned} p_o &:= p_{00} + p_{11} \\ p_e &:= (p_{00} + p_{01}) \cdot (p_{00} + p_{10}) \\ &\quad +(p_{10} + p_{11}) \cdot (p_{01} + p_{11}) \end{aligned} \qquad (6)$$

Here, p_o is the probability of the observed agreement and p_e is the probability of agreement, when both ratings are unconditionally independent. The closer κ is to 1, the better the agreement of the examiners.

Remark 2.1 *There is an explicit relation between the kappa coefficient, the prevalence and the probabilities of misclassification, which will be useful to identify regions for the prevalence. Under the assumptions*

(A1) Independent conditional distributions $Y_1^*|Y$ and $Y_2^*|Y$ for both replicates

(A2) Equal sensitivity and specificity for both replicates

the following equation holds ($p \in (0;1)$):

$$\begin{aligned} \kappa &= \frac{p\,(1-p)\,(sens + spec - 1)^2}{(spec - p\,(sens + spec - 1))} \\ &\quad \cdot \frac{1}{(1 - spec + p\,(sens + spec - 1))}\,. \end{aligned} \qquad (7)$$

Equation (7) is deduced by using the assumptions (A1) and (A2) that imply

$$\begin{aligned} p_{00} &= (1-p) \cdot spec^2 + p \cdot (1 - sens)^2 \\ p_{01} &= (1-p) \cdot spec \cdot (1 - spec) \\ &\quad +p \cdot (1 - sens) \cdot sens \\ p_{10} &= p_{01} \qquad (8) \\ p_{11} &= (1-p) \cdot (1 - spec)^2 + p \cdot sens^2 \qquad (9) \end{aligned}$$

This leads, together with (5), to formula (7). Note that the kappa coefficient can be seen as a parameter of one scoring process. It is a measurement of agreement when it is independently applied on the same subject twice. It can be estimated by a validation study, where two independent scorings are available for a (sub)sample of individuals.

Note that the assumption of conditional independence and identical sensitivity and specificity may be violated, if the two replicates correspond to two different examiners, for a further discussion we refer to Section 5. The assumption of identical sensitivity and specificity can be checked using the McNemar test, which is designed for the comparison of two probabilities for dependent data. It basically checks the identity (8), see also our example in Section 4.

2.2 Bias Correction using the Kappa Coefficient

We want to estimate the true prevalence p using the naive estimator $\widehat{p^*}$ and a given or consistently estimated kappa coefficient. The basic approach is to use equations (7) and (1) and solve them for p. Since there are three unknowns (p, *sens*, *spec*) and only two equations, there is a lack of identifiability and no direct estimator can be deduced. However, non trivial intervals $I(\vartheta \parallel p^*, \kappa)$ for the possible solutions for the three parameters $\vartheta \in \{p, sens, spec\}$ can be derived, by additionally relying on the constraint that all probabilities are in $[0;1]$. Following [32], these solutions are called identification regions. In [52] they are called ignorance regions, since they relate to ignorance in contrast to sampling error.

Theorem 2.2 *(*Identification Regions for p, *sens* and *spec* using p^* and κ*)*
Let the assumptions (A1) and (A2) hold. Additionally, let $\kappa \in (0,1]$ and $sens + spec > 1$ (see (2)). Then the identification regions for the prevalence p, the sensitivity sens and the specificity spec based on the naive prevalence $p^* \in [0,1]$, are

$$I(p \parallel p^*, \kappa) = \left[\frac{p^*}{p^* + \kappa^{-1}(1-p^*)};\;\; \frac{p^*}{p^* + \kappa\,(1-p^*)} \right], \qquad (10)$$

$$I(sens \parallel p^*, \kappa) = [p^* + \kappa\,(1-p^*);\, 1] \qquad (11)$$

$$I(spec \parallel p^*, \kappa) = [1 - p^* + p^*\kappa;\, 1]\,. \qquad (12)$$

The regions in the theorem follow directly by solving equations (7) and (1), and therefore are the best that we can learn from the given values of p^* and κ, without adding further assumptions. Details of the derivation are given in the web appendix ([26]).

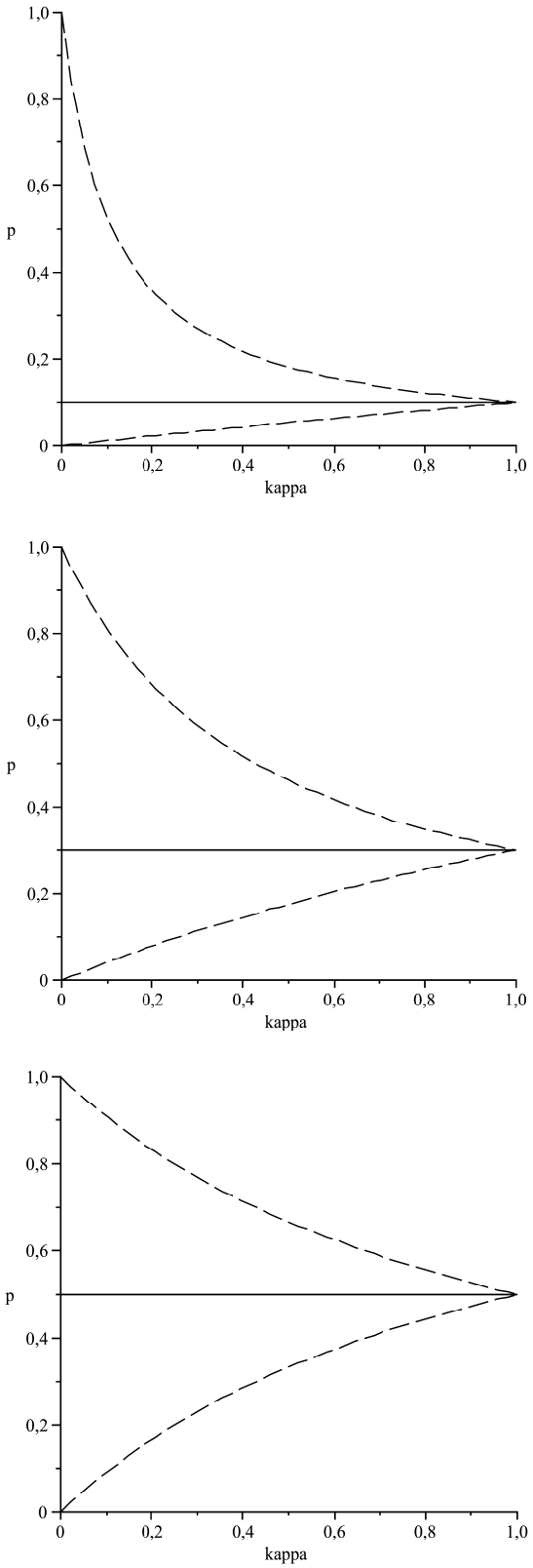


Figure 3: Identification regions (dashed lines) for the true prevalence p (solid line) in dependence of κ for different values of the naive preference $p^* \in 0.1, 0.3, 0.5$, from top to bottom.

Naturally, the width of the intervals decreases when the kappa coefficient κ increases. Indeed, considering the extreme case where the examiners' assignments are almost random, ($\kappa \to 0$) leads to the vacuous statement $I_p = [0; 1]$. On the other hand, complete agreement, and therefore $\kappa = 1$, results in point identification, where the region for p degenerates to p^* and $sens = spec = 1$. In Figure 3, the identification regions are displayed as a function of the kappa coefficient for fixed values of p^*. For reasonable agreement of the measurements, in particular, the intervals are small enough to provide valuable insight into the true prevalence.

Note that, by construction, the method is based on the data in a conservative manner. Consequently, the identification region necessarily contains p^*: By $\kappa \leq 1$,

$$\begin{aligned} \frac{p^*}{p^* + \kappa^{-1}(1-p^*)} \leq \frac{p^*}{p^* + (1-p^*)} &= p^* \\ \frac{p^*}{p^* + \kappa(1-p^*)} \geq \frac{p^*}{p^* + (1-p^*)} &= p^*. \end{aligned}$$

The regions given in Theorem 2.2 are the best we can conclude from the data alone. If we interpret them as probability assignments they describe coherent interval-valued probabilities and F-probabilities in the sense of [54] and [56, 57], for details see [26]. Note that kappa coefficient and p^* bear sufficient information for determining the probabilities $(p_{00}, p_{01}, p_{10}, p_{11})$, i.e. using those probabilities would not lead to an improvement of the bounds. Since the assumptions A1 and A2 imply $p_{01} = p_{10}$ and the probabilities add to 1, there are only two free parameters. An explicit formula is presented in [26].

Theorem 2.2 enables us to calculate identification regions for the prevalence, sensitivity and specificity from the naive estimator $\widehat{p^*}$ and an estimated kappa value $\hat{\kappa}$, by substituting p^* and κ with their estimators in equations (10) to (12). Note that these intervals correspond to point estimators and, in particular, are not confidence intervals. Strategies for finding confidence intervals, i.e. additionally taking the sampling variation into account, are given in the following section.

3 Taking Additionally Sampling Variation into Account: Confidence Intervals

We follow here the strategy from [52] and define a parameter γ, which is not identified by our data, but the other parameters of our models are identified conditional on this parameter. As a suitable choice for

this identifying parameter we propose in our context $\gamma := \frac{sens}{spec}$, which indeed would result in a point identified estimator, see (16) below. The parameter γ has an obvious interpretation relating the probabilities of the two types of misclassification. In the framework of [52] it is called a *sensitivity* parameter. We do not use this technical term here to avoid confusion with the sensitivity of the diagnosis *sens*. The parameter γ is restricted by (11) and (12). Therefore, the range of γ is given by

$$[\gamma_{min},\, \gamma_{max}] = \left[p^* + \kappa\,(1-p^*), \quad \frac{1}{1-p^*+p^*\kappa}\right]. \tag{13}$$

We now assume that a consistent estimator $(\widehat{p^*}, \hat{\kappa})$ with asymptotic covariance matrix Σ is available. If the estimator of κ is estimated by an independent validation study, Σ is diagonal. If we assume that κ is known, then the corresponding entries in Σ are 0.

To construct a confidence interval $\left[L(\widehat{p^*}, \hat{\kappa}); U(\widehat{p^*}, \hat{\kappa})\right]$ for the parameter p we have to ensure that the coverage probability exceeds the confidence level $1-\alpha$ for every $\gamma \in [\gamma_{min},\, \gamma_{max}]$, i.e.

$$\inf_{\gamma \in\, [\hat{\gamma}_{min},\, \hat{\gamma}_{max}]} P_\gamma(p \in \left[L(\widehat{p^*}, \hat{\kappa}); U(\widehat{p^*}, \hat{\kappa})\right]) \geq 1-\alpha\,. \tag{14}$$

This can be achieved by defining the confidence interval as the union of confidence intervals over the identification parameter γ

$$\begin{aligned}\left[L(\widehat{p^*}, \hat{\kappa}); U(\widehat{p^*}, \hat{\kappa})\right] :=\\ \bigcup_{\gamma \in\, [\hat{\gamma}_{min},\, \hat{\gamma}_{max}]} \left[L(\widehat{p^*}, \hat{\kappa}, \gamma); U(\widehat{p^*}, \hat{\kappa}, \gamma)\right]\end{aligned} \tag{15}$$

with $\left[L(\widehat{p^*}, \hat{\kappa}, \gamma); U(\widehat{p^*}, \hat{\kappa}, \gamma)\right]$ as suitable confidence intervals for fixed parameter γ. To calculate the latter, we apply the delta method (e.g., [3]) and use for fixed γ the point estimator for p given by

$$\hat{p}\,(\widehat{p^*}, \hat{\kappa}, \gamma) = \frac{(1-\widehat{p^*})\cdot\gamma - \widehat{p^*} - \sqrt{w}}{(\widehat{p^*}-1)\cdot\gamma^2 + (1-\sqrt{w})\cdot\gamma - \widehat{p^*} - \sqrt{w}} \tag{16}$$

with

$$\begin{aligned} w \;=\; & (\widehat{p^*}-1)^2\cdot\gamma^2 - 2\cdot\widehat{p^*}\cdot(\widehat{p^*}-1)\\ & \cdot(2\cdot\hat{\kappa}-1)\cdot\gamma + (\widehat{p^*})^2 \end{aligned}$$

derived from (7) and (1), see [26]. The asymptotic variance is given by the delta method

$$Var(\hat{p}\,(\widehat{p^*}, \hat{\kappa}, \gamma)) = D_p^T\,\Sigma\,D_p\,. \tag{17}$$

Here, D_p is the vector of derivatives of $\hat{p}\,(\widehat{p^*}, \hat{\kappa}, \gamma)$ with respect to $\widehat{p^*}$ and $\hat{\kappa}$, and Σ is the corresponding covariance matrix. Details are again given in [26]. Since the relationship (16) between γ and p is monotone, the choice of the confidence intervals in (15) can be optimized, see [52] or [22, 47]. If the local confidence intervals are small compared to the identification region, then it is actually justified to rely on the $(1-\alpha)\cdot 100\%$-quantile, instead of the $(1-\alpha/2)\cdot 100\%$-quantile. Thus the confidence interval is given by

$$\begin{aligned}&\left[\,L(\hat{p^*}; \hat{\kappa})\;, U(\hat{p}, \hat{\kappa})\right] = \\ &\Big[\hat{p}(\widehat{p^*}, \hat{\kappa}, \hat{\gamma}_{max}) - z_{1-\alpha}\cdot\sqrt{\widehat{Var}(\hat{p}(\widehat{p^*}, \hat{\kappa}, \hat{\gamma}_{max}))};\\ &\quad \hat{p}(\widehat{p^*}, \hat{\kappa}, \hat{\gamma}_{min}) + z_{1-\alpha}\cdot\sqrt{\widehat{Var}(\hat{p}\,(\widehat{p^*}, \hat{\kappa}, \hat{\gamma}_{min}))}\Big]\,.\end{aligned} \tag{18}$$

The range for γ is estimated using (13). Since the estimator of $(\widehat{p^*}, \hat{\kappa})$ is consistent, the probability that the interval $[\hat{\gamma}_{min},\, \hat{\gamma}_{max}]$ covers the true parameter γ tends to 1 as sample size n goes to infinity. Therefore, (15) is an asymptotic confidence interval. Note that we define our confidence intervals for the parameter and not for the entire identified set, see, in particular, [22] for a discussion of that distinction.

4 Example

4.1 The Signal-Tandmobiel® Study

year	n	$\widehat{p^*}$	$se(\widehat{p^*})$
1996 (age 6)	3378	0.118	0.006
1998 (age 8)	3657	0.280	0.007
2000 (age 10)	3415	0.380	0.008

Table 2: Signal-Tandmobiel® study: Estimation of $\widehat{p^*}$ per year

The Signal-Tandmobiel® study is a 6-year longitudinal oral health study, conducted in Flanders (Belgium) involving 4468 children. Data were collected on oral hygiene, gingival condition, dental trauma, prevalence and extent of enamel developmental defects, fluorosis, tooth decay, presence of restoration, missing teeth, stage of tooth eruption and orthodontic treatment need, all by using established criteria, see [51]. The children were examined annually during 1996 to 2001. Measurement of interest is the *dmft* index, which is the sum of the number of decayed, missing due to caries or filled teeth.

We use the *dmft* index as an indicator for the presence or absence of caries for each child to examine the prevalence of caries. The observed disease status Y_i^*

for child i is

$$Y_i^* = \begin{cases} 1 & \text{caries observed} \quad (dmft > 0) \\ 0 & \text{no caries observed} \quad (dmft = 0) \end{cases}.$$

For illustration of our methods, we estimate the naive prevalence and its variance for the years 1996 (age 6), 1998 (age 8) and 2000 (age 10), see Table 2. These are the years in which a calibration study was conducted. The longitudinal structure is ignored and the naive prevalence naturally increases over the years, i.e. with the age of the children, and its standard error is very low due to the high sample size n.

1996

	Rater 1		
Rater 2	78	7	85
	13	22	35
	91	29	120

$p-$value $= 0.1797$ (McNemar)
$\kappa = 0.5752(0.084)$

1998

	Rater 1		
Rater 2	85	13	98
	16	43	59
	101	56	157

$p-$value $= 0.5775$ (McNemar)
$\kappa = 0.6023(0.066)$

2000

	Rater 1		
Rater 2	89	14	103
	3	42	45
	92	56	148

$p-$value $= 0.0076$ (McNemar)
$\kappa = 0.7461(0.057)$

Table 3: Results of the validation study with two raters. Kappa indicates the kappa statistics with standard error in brackets.

In the calibration study in [38], the observations of the 16 regular examiners were compared to a gold standard examiner resulting in estimation of sensitivity and specificity. However, letting one single person be the gold standard examiner can still not guarantee correctness. For illustration of our methods and to incorporate this possibility of an error, the gold standard examiner is now considered a 'common' examiner. In the validation study we now have two observations per child. The results are presented in Table 3. However, since assumption A2 is questionable in our setting, we performed a McNemar test, which is based on the difference of the off diagonal cells of the two by two table. In case of the two by two table for 2000, the test indicates a significant deviation from the assumption. Therefore, we present results of our method only for the years 1996 and 1998. The estimated standard errors of the kappa coefficient are rather high due to the small sample size.

4.2 Correction for Misclassification

We use the methods shown in this paper to correct the estimated prevalence for misclassification. In Table 4, the corresponding identification regions based on the point estimation of p^* and κ using Theorem 2.2 are presented. The regions for the prevalence are wide. This is a consequence of the low kappa coefficient, reflecting the low agreement among the examiners. As discussed, the estimated regions include the naive estimator, but it can be seen that the naive estimator could be seriously biased. Moreover, the regions for specificity, and especially for sensitivity are wide, too.

If the kappa coefficient was considered known, the confidence intervals are only slightly smaller, indicating that the main problem is in the partial identification of our setting.

year	$\widehat{p^*}$	$\hat{\kappa}$	$I(p \parallel \widehat{p^*}, \hat{\kappa})$
1996	0.118	0.577	[0.072; 0.188]
1998	0.280	0.602	[0.190; 0.393]
year	$I(sens \parallel \widehat{p^*}, \hat{\kappa})$		$I(spec \parallel \widehat{p^*}, \hat{\kappa})$
1996	[0.627; 1.000]		[0.950; 1.000]
1998	[0.714; 1.000]		[0.889; 1.000]

Table 4: Signal-Tandmobiel® study: Estimated identification regions for p, *sens* and *spec*

In a second step, the confidence intervals for the prevalence following the strategy from Section 3 are presented in Table 5, once while incorporating the sample variability of the estimators $\widehat{p^*}$ and $\hat{\kappa}$ and, for illustration, assuming κ to be known at its estimated value. The asymptotic confidence intervals for the naive prevalence are pretty small compared to the identification regions and to the corresponding confidence intervals, which are both based on the additional information from the kappa coefficient. Consequently, the confidence regions based on naive preva-

lence estimation still suffer from a severe overprecision. Although being somewhat large, the identification region and the corresponding confidence regions still provide valuable insight into the prevalence. For example, the hypothesis $H_0 : p \geq 0.25$ could be rejected at the 5 percent-level for the 6 year old children.

year	with sampling variation of κ	fixed κ
1996	[0.057; 0.219]	[0.065; 0.205]
1998	[0.170; 0.416]	[0.179; 0.409]

Table 5: Signal-Tandmobiel® study: Confidence intervals for the prevalence with and without taking the variability of κ into account

If further nontrivial bounds on sensitivity and specificity are available by some external information, then this can be incorporated in an analogous way resulting in smaller identification regions and smaller confidence intervals based on them.

5 Discussion

The concept of using identification regions or intervals of ignorance in the case of misclassification with partial information on sensitivity and specificity provided by the kappa coefficient has been shown as a powerful tool for data analysis. It avoids the potentially substantial bias arising from simply ignoring misclassification if no direct correction method is available. The resulting identification regions are tight in the sense that they can not be improved without adding further assumptions. Thus they are the best that we can conclude from the data alone in this context. Our example shows that the possible effect of misclassification is rather high, even when the inter rater reliability is 'substantial´ in terms of [28]'s classification. Furthermore, the strategy of distinguishing between sampling error and ignorance due to non-identifiability is useful, since it highlights possible shortcomings in the sampling of the data structure, which cannot be compensated by a large sample size.

Since we use the value of the kappa coefficient from validation data or from other sources of information, one crucial assumption for our analysis is that this value is also correct for the main data set. This will be the case if our replication data are a random sample from our main study (internal validation). Otherwise this assumption could be disputable. It is well-known that the kappa coefficient depends on the prevalence when sensitivity and specificity are fixed [10]. So our procedure cannot be used when the prevalence in the validation data differs from the prevalence in the main study, even if we assume that the scoring procedure has fixed sensitivity and specificity. However, the latter assumption could also be problem, see the discussion in [50]. In our example, the validation study was part of a training program for the examiners. On the one hand the prevalence was higher for the validation but on the other hand there were possibly more children in that sample that were difficult to score. This could lead to values of sensitivity and specificity which are different in the main study. Nevertheless, the kappa coefficient could be nearly identical in both parts of the study. [50] performs some calculations and presents plausible scenarios for this assumption. Thus, our procedure can also be applied to studies where the value of the kappa coefficient can be transferred from the validation data to the main study even this is not true for sensitivity and specificity. Obviously, this issue has to be treated with great care.

Our results are a vivid illustration of the power of imprecise probability methods in statistical analysis based on misclassified data. As a topic of further research the conditional independence assumption (A1) in Remark 2.1 should be investigated further. As mentioned above, it may be violated if the assessments of two raters are used as substitutes for replication data, because then certain characteristics of the units may impose some dependence on the raters' judgements. We are currently studying the use of Frechet bounds and related methods in this setting. If no reliable information is available about the misclassification probabilities, our approach could be adopted to the case where sensitivity and specificity vary in certain ranges, closely relating our procedure to the 'direct method' of [37]. Then our identification parameter is two dimensional, which will result in larger identification regions.

The methodology underlying our work promises, mutatis mutandis, to be also powerful for other types of error-prone data, like misclassification of more than two categories and for (additive or multiplicative) measurement error with unknown variance. In the latter case, the availability of replicates would yield identification in many instances, but often no information about the measurement error is available, and then partially identified corrected estimators are again the best option available.

Acknowledgements

We thank two anonymous referees for very stimulating comments, which substantially improved our paper. We thank G. Walter, M. Cattaneo and many participants of

ISIPTA'09's poster session for helpful discussions. The data collection of Signal-Tandmobiel® study was supported by Unilever, Belgium. The Signal-Tandmobiel® project comprises following partners: D. Declerck (Dental School, Catholic Univ. Leuven), L. Martens (Dental School, Univ. Ghent), J. Vanobbergen (Working Group Oral Health Promotion and Prevention, Flemish Dental Association; Dental School, Univ. Ghent), P. Pottenberg (Dental School, Univ. Brussels), E. Lesaffre (L-Biostat, Univ. Leuven, Dep. of Biostatistics, Erasmus MC, Rotterdam), K. Hoppenbrouwers (Youth Health Dep., Catholic Univ. Leuven; Flemish Assoc. for Youth Health Care).

References

[1] T. Augustin and R. Hable. On the impact of robust statistics on imprecise probability models: A review. *Structural Safety*, 32:358–365, 2010.

[2] C. Beunckens, C. Sotto, G. Molenberghs, and G. Verbeke. A multifaceted sensitivity analysis of the Slovenian public opinion survey data. *Journal of the Royal Statistical Society: Series C*, 58(2):171–196; Corr: 575–576, 2009.

[3] P. Bickel and K. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Vol I.* Prentice Hall, 2001.

[4] R. Blundell, M. Browning, and I. Crawford. Best nonparametric bounds on demand responses. *Econometrica*, 76(6):1227–1262, 2008.

[5] C. T. Bruckner and P. Yoder. Interpreting kappa in observational research: Baserate matters. *American Journal on Mental Retardation*, 111(6):433–441, 2006.

[6] R. Carroll, D. Ruppert, L. Stefanski, and C. Crainiceanu. *Measurement Error in Nonlinear Models.* Chapman and Hall, New York, 2006. 2nd edition.

[7] J. Cheng and D. Small. Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society: Series B*, 68(5):815–836, 2006.

[8] Cheng, S., Xi, Y., and Chen, M.-H. (2008), A new mixture model for misclassification with applications for survey data, *Sociological Methods & Research*, 37, 75–104.

[9] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[10] R. Cook. Kappa and its dependence on marginal rates. In: Armitag, P., and Colton, T. (eds.) *Encyclopedia of Biostatistics.* volume 3, pp. 2166–2168. Wiley, Chichester, UK, 1998.

[11] J. Copas and D. Jackson. A bound for publication bias based on the fraction of unpublished studies. *Biometrics*, 60(1):146–153, 2004.

[12] G. De Cooman and M. Zaffalon. Updating beliefs with incomplete observations. *Artificial Intelligence*, 159(1-2):75–125, 2004.

[13] E. Diday and M. Noirhomme-Fraiture. Symbolic Data Analysis and the SODAS Software. Wiley and Sons, Chicester, 2008.

[14] M. Feuerman and A. Miller. Relationships between statistical measures of agreement: sensitivity, specificity and kappa. *Journal of Evaluation in Clinical Prctice*, 14(5):930–933, 2008.

[15] C. Gundersen and B. Kreider. Bounding the effects of food insecurity on children's health outcomes. *Journal of Health Economics*, 28(5):971–983, 2009.

[16] A. Guolo. Robust techniques for measurement error correction: a review. *Statistical Methods in Medical Research*, 17(6):555–580, 2008.

[17] P. Gustafson. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments.* Chapman and Hall, New York, 2004.

[18] P. Gustafson. Bayesian inference for partially identified models. *International Journal of Biostatistics*, 6 (2): 17, 2010.

[19] P. Gustafson and S. Greenland. Interval estimation for messy observational data. *Statistical Science*, 3 (24):328–342, 2009.

[20] J. Hausman, J. Abrevaya, and F. Scott-Morton. Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87(2):239–269, 1998.

[21] M. Henmi, J. Copas, and S. Eguchi. Confidence intervals and p-values for meta-analysis with publication bias. *Biometrics*, 63(2):475–482, 2007.

[22] G. Imbens and C. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

[23] M. Keane and R. Sauer. Classification error in dynamic discrete choice models: implications for female labor supply behavior. *Econometrica*, 77(3):975–991, 2009.

[24] D. Kenkel, D. Lillard, and A. Mathios. Accounting for misclassification error in retrospective smoking data. *Health Economics*, 13(10):1031–1044, 2004.

[25] B. Kreider and S. Hill. Partially identifying treatment effects with an application to covering the uninsured. *Journal of Human Resources*, 44(2):409–449, 2009.

[26] H. Küchenhoff, T. Augustin, and A. Kunz. Partially identified prevalence estimation under misclassification using the Kappa coefficient (Web Appendix) www.stablab.stat.uni-muenchen.de/kuechenhoff/isipta-app.

[27] H. Küchenhoff, S. Mwalili, and E. Lesaffre. A general method for dealing with misclassification in regression: The misclassification simex. *Biometrics*, 62(1):85–96, 2006.

[28] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[29] A. Lewbel. Estimation of average treatment effects with misclassification. *Econometrica*, 75(2):537–551, 2007.

[30] R. Lyles, A. Allen, W. Flanders, L. Kupper, and D. Christensen. Inference for case-control studies when exposure status is both informatively missing and misclassified. *Statistics in Medicine*, 25(23):4065–4080, 2006.

[31] R. Lyles, J. Williamson, H.-M. Lin, and C. Heilig. Extending McNemar's test: Estimation and inference when paired binary outcome data are misclassified. *Biometrics*, 61(1):287–294, 2005.

[32] C. Manski. *Partial Identification of Probability Distributions*. Springer, New York, 2003.

[33] C. Manski. Partial identification with missing data: concepts and findings. *International Journal of Approximate Reasoning*, 39(2-3):151–165, 2005.

[34] C. Manski. Partial identification of counterfactual choice probabilities. *International Economic Review*, 48(4):1393–1410, 2007.

[35] C. Manski and J. Pepper. More on monotone instrumental variables. *Econometrics Journal*, 12(1):200–216, 2009.

[36] G. Molenberghs. Incomplete data in clinical studies: analysis, sensitivity, and sensitivity analysis. *Drug Information Journal*, 43(4):409–429, 2009.

[37] F. Molinari. Partial identification of probability distributions with misclassified data. *Journal of Econometrics*, 144(1):81–117, 2008.

[38] S. Mwalili, E. Lesaffre, and D. Declerck. A Bayesian ordinal logistic regression model to correct for interobserver measurement error in a geographical oral health study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):77–93, 2005.

[39] J. Neuhaus. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, 86(4):843–855, 1999.

[40] J. Neuhaus. Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics*, 58(3):675–683, 2002.

[41] M. Pepe and H. Janes. Insights into latent class analysis of diagnostic test performance. *Biostatistics*, 8(2):474–484, 2007.

[42] C. Roberts. Modelling patterns of agreement for nominal scales. *Statistics in Medicine*, 27(6):810–830, 2008.

[43] Rummel, D., Augustin, T., & Küchenhoff, H., Correction for covariate measurement error in nonparametric longitudinal regression. *Biometrics*, 66:1209–1219, 2010.

[44] Schneeweiß, H. & Augustin, T., Some recent advances in measurement error models and methods, *ASTA Allgemeines Statistisches Archiv*, 90: 183–197, 2006.

[45] M. Shoukri and A. Donner. Bivariate modeling of interobserver agreement coefficients. *Statistics in Medicine*, 28(3):430–440, 2009.

[46] J. Stamey, D. Boese, and D. Young. Confidence intervals for parameters of two diagnostic tests in the absence of a gold standard. *Computational Statistics and Data Analysis*, 52(3):1335–1346, 2008.

[47] J. Stoye. More on confidence intervals for partially identified parameters. *Econometrica*, 77(4):1299–1315, 2009.

[48] J. Stoye. Partial identification and robust treatment choice: an application to young offenders. *Journal of Statistical Theory and Practice*, 3(1):239–254, 2009.

[49] L. Utkin and T. Augustin. Decision making under imperfect measurement using the imprecise Dirichlet model. *International Journal of Approximate Reasoning*, 44(3):322–338, 2007.

[50] W. Vach. The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology*, 58(7):655–661, 2005.

[51] J. Vanobbergen, L. Martens, E. Lesaffre, and D. Declerck. The Signal Tandmobiel project – a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results. *European Journal of Paediatric Dentistry*, 2:87–96, 2000.

[52] S. Vansteelandt, E. Goetghebeur, M. Kenward, and G. Molenberghs. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, 16(3):953, 2006.

[53] C. Vogel, H. Brenner, A. Pfahlberg, and O. Gefeller. The effects of joint misclassification of exposure and disease on the attributable risk. *Statistics in Medicine*, 24(12):1881–1896, 2005.

[54] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[55] S. Walter, C. Hsieh, and Q. Liu. Effect of exposure misclassification on the mean squared error of population attributable risk and prevented fraction estimates. *Statistics in Medicine*, 26(26):4833–4842, 2007.

[56] Weichselberger, K. (2000) The theory of interval probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24 (2-3), 149-170.

[57] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept*. Physika, Heidelberg, 2001.

[58] M. Zaffalon and E. Miranda. Conservative inference rule for uncertain reasoning under incompleteness. *Journal of Artificial Intelligence Research*, 34:757–821, 2009.

[59] D. Zucker and D. Spiegelman. Corrected score estimation in the proportional hazards model with misclassified discrete covariates. *Statistics in Medicine*, 27(11):1911–1933, 2008.

Nonparametric predictive inference for subcategory data

R.M. Baker, P. Coolen-Schrijner, F.P.A. Coolen
Durham University
Durham, UK
r.m.baker@dunelm.org.uk; frank.coolen@durham.ac.uk

T. Augustin
Ludwig-Maximilians University
Munich, Germany
thomas@stat.uni-muenchen.de

Abstract

Nonparametric predictive inference (NPI) is a framework for statistical inference in the absence of prior knowledge. We present NPI for multinomial data with subcategories, motivated by the hierarchical structure of many multinomial data sets. We consider situations with known and with unknown numbers of subcategories, and present lower and upper probabilities for general events involving one future observation. We present properties of the model and an algorithm to derive an approximation to the maximum entropy distribution.

Keywords. classification, multinomial data, nonparametric predictive inference, subcategories

1 Introduction

Nonparametric predictive inference (NPI) was presented by Coolen and Augustin [5, 7] for multinomial data in the absence of prior knowledge. A key assumption underlying the model is that the different categories are not ordered or otherwise related. The model is, therefore, not suited to multinomial data sets with a hierarchical structure in which two or more distinct categories may also be considered as subcategories of a single main category. Following the suggestion in [6], we present an extension of the NPI model for multinomial data suitable for data sets with subcategories, which we refer to as the Sub-MNPI model. As in the original NPI model for multinomial data [5, 7], we assume that there is no ordering of the main categories, and we also assume that for a single main category there is no ordering of its subcategories. Throughout the paper, categories are denoted by c_j and subcategories are denoted by s_{j,i_j}, where $s_{j,i_j} \subseteq c_j$. We assume that there are K main categories in total, and that k main categories have already been observed and are labelled $c_1, ..., c_k$. Similarly, we assume that there is a total of K_j subcategories in main category c_j, of which k_j have already been observed. Note that K and K_j may be known or unknown: these two situations are considered separately. Let n denote the total number of observations Y_1,...,Y_n in the data set, where n_j is the number of observations in main category c_j and n_{j,i_j} is the number of observations in subcategory s_{j,i_j}. Some main categories may not contain any subcategories, or may only be described at main category level, in which case we continue to denote these simply by c_j. Such categories are referred to as main-only categories, distinct from main categories which may or may not have specified subcategories.

In section 2 of this paper, we explain the probability wheel representation of the data on which the NPI model for subcategory data is based. In the following two sections, we then define the general events of interest for inference about a future observation and we present the NPI lower and upper probabilities for these events. The situation where K and K_j are known is considered in Section 3, and the situation where K and K_j are unknown is considered in Section 4. Some important properties of the model are then described in Section 5. In Section 6 we consider the application of the model to classification, and finally Section 7 provides some concluding remarks.

2 The Sub-MNPI model

The NPI approach for multinomial data is based on a variation of Hill's $A_{(n)}$ assumption [8] called circular-$A_{(n)}$ [5, 6, 7], which is an assumption of post-data exchangeability. The model uses a probability wheel representation of the data [5, 6, 7], where each of the n observations is represented by a radial line such that the wheel is partitioned into n equally-sized slices. From the circular-$A_{(n)}$ assumption we conclude that the next observation has probability $\frac{1}{n}$ of being in any given slice. The inferences made about a future observation therefore

depend upon which main category or subcategory each slice of the wheel represents, and this is determined by the key assumption that each main category and each subcategory is only allowed to be represented by one segment of the wheel, where a segment is defined as a single part of the wheel (note that the wheel is always divided radially) consisting of any number of full or partial slices. The assumption implies the following constraints:

- Two or more lines representing the same (sub)category must always be positioned next to each other on the wheel.
- Lines representing different subcategories within the same main category are always grouped together in one single segment of the wheel.
- If a slice is bordered by two lines representing the same (sub)category, it must be assigned to this (sub)category.
- A slice that is bordered by two lines representing observations in (sub)categories x and y where $x \neq y$, defined as a separating slice, may be assigned to x or to y or to an unobserved (sub)category not yet allocated to any other slice.
- Separating slices may be divided radially between multiple (sub)categories.

All possible configurations of the probability wheel are considered, and lower and upper probabilities for an event of interest are derived by respectively minimising and maximising the number of slices assigned to the event.

3 Known number of (sub)categories

When K and K_j, $j = 1, ..., K$, are known, the event of interest can be expressed generally as

$$E = \{Y_{n+1} \in \bigcup_{j \in J} c_j \cup \bigcup_{j \in J^*} \bigcup_{i_j \in I_j} s_{j,i_j}\} \qquad (1)$$

where $J \cap J^* = \emptyset$, $J \subseteq \{1, ..., K\}$, $J^* \subseteq \{1, ..., K\}$ and $I_j \subseteq \{1, ..., K_j\}$ for $j = 1, ..., K$. It should be emphasized that J is the index-set of the categories which occur in the event of interest only at main category level, while J^* is the index-set of the categories which occur in this event at subcategory level. We also define $\overline{I_j} = \{1, ..., K_j\} \backslash I_j$. This notation allows us to describe events which contain only specific subcategories of particular main categories, whilst also retaining the possibility of considering some main categories as a whole.
We define $OJ = J \cap \{1, ..., k\}$, which is the index-set of observed main-only categories in E, and $|OJ| = r_{main}$. We also define $UJ = J \cap \{k+1, ..., K\}$, which is the index-set of unobserved main-only categories in E, and $|UJ| = l_{main}$. Similarly, $OJ^* = J^* \cap \{1, ..., k\}$, where $|OJ^*| = r_{sub}$. OJ^* is the index-set of observed main categories in E which are described at subcategory level. We also define $UJ^* = J^* \cap \{k+1, ..., K\}$, where $|UJ^*| = l_{sub}$. UJ^* is the index-set of unobserved main categories in E which are described at subcategory level. Let $r = r_{main} + r_{sub}$, and let $l = l_{main} + l_{sub}$.
Let $OI_j = I_j \cap \{1, ..., k_j\}$, where $|OI_j| = r_j$, for $j = 1, ..., K$. OI_j is the index-set of observed subcategories in E. Also let $UI_j = I_j \cap \{k_j+1, ..., K_j\}$, where $|UI_j| = l_j$, for $j = 1, ..., K$. UI_j is the index-set of unobserved subcategories in E. Let $\overline{OI_j} = \overline{I_j} \cap \{1, ..., k_j\}$, where $\left|\overline{OI_j}\right| = \overline{r_j}$, and let $\overline{UI_j} = \overline{I_j} \cap \{k_j + 1, ..., K_j\}$, where $\left|\overline{UI_j}\right| = \overline{l_j}$.
We present the NPI lower and upper probabilities for E (1). A detailed derivation of these formulae is given in [4].

3.1 Lower probability

The NPI lower probability is found by constructing a configuration of the probability wheel which minimises the number of slices assigned to E. In order to construct such a configuration, we consider how many separating slices we can assign to main categories or subcategories not in E. First, separating slices on the wheel between different observed main categories in E can be assigned to main categories that are not in E. There are $(K-r-l)$ such categories. Furthermore, if we have subcategories which are not in E but which are part of a main category that appears in E, it may be possible to utilise these subcategories to separate observed main categories in E. By considering the configuration of the slices, we find that the number of separating slices which can potentially be filled in this way (with x^+ representing $\max\{x, 0\}$) is

$$S_M = \sum_{j \in OJ^*} \min\{(\overline{r_j} + \overline{l_j} - r_j + 1)^+, 2\} + \sum_{j \in UJ^*} \min\{\overline{l_j}, 1\}.$$

Minimising the number of slices that must be assigned to E results in the following general formula:

$$\underline{P}(E) = \sum_{j \in OJ} \frac{n_j - 1}{n} + \sum_{j \in OJ^*} \sum_{i_j \in OI_j} \frac{n_{j,i_j} - 1}{n} + \frac{1}{n}(2r + l - K - S_M)^+ + \frac{1}{n} \sum_{j \in OJ^*} (2r_j + l_j - K_j - 1)^+. \qquad (2)$$

Example 1 *Consider a multinomial data set with possible main categories blue (B), green (G), red (R), yellow (Y), pink (P) and orange (O). These main categories are labelled 1 to 6 respectively. Observations in B are further classified as light blue (LB), medium blue (MB), dark blue (DB) or other blue (OB), and observations in G are further classified as light green (LG), dark green (DG) or other green (OG). The data set consists of eight observations altogether, including 1 LB, 1 MB, 2 DB, 1 LG, 1 DG, 1 R and 1 Y.*
Suppose that we are interested in the event $Y_9 \in \{LB,MB,DB,LG,R,Y,P\}$. *We have* $K = 6$, $r = 4$ *and* $l = 1$. *For main categories described at subcategory level, the values of* K_j, r_j *and* l_j *are shown in Table 1. Here, we are unable to assign all*

	j	K_j	r_j	l_j
B	1	4	3	0
G	2	3	1	0

Table 1: Values of K_j, r_j and l_j for Example 1

separating slices within the B segment to subcategories not in E. Furthermore, we are unable to configure the probability wheel such that all observed main categories in E are separated by main categories not in E. We find that $2r+l-K = 3$ *in this example, and* $S_M = 2$. *While we can use some subcategories which are not in E but which are part of a main category that appears in E, there is still one separating slice between main categories which has to be assigned to E. Figure*

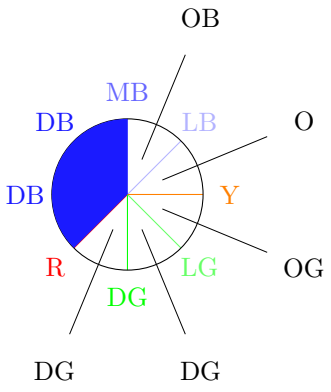


Figure 1: Probability wheel for Example 1

1 shows a possible configuration of the wheel such that O separates B and Y, OG separates Y and G, and DG separates G and R. There is then no way of separating R and B by a main category or subcategory not in E, and we are therefore forced to assign this slice to E. Looking specifically at the B segment, we see that OB separates LB and MB but the slice between MB and DB then has to be assigned to E. This leads to a NPI lower probability of $\frac{3}{8}$ *for the event E. This lower probability can be verified using (2). We see that the set* OJ *contains R and Y, the set* OJ^* *contains B and G, the set* OI_1 *contains LB, MB and DB and the set* OI_2 *contains LG. Also,* $2r + l - K = 3$, $S_M = 2$ *and* $\sum_{j \in OJ^*} \max\{2r_j + l_j - K_j - 1, 0\} = 1$, *therefore (2) gives* $\underline{P}(E) = \frac{3}{8}$.

3.2 Upper probability

The NPI upper probability is found by constructing a configuration of the probability wheel which maximises the number of slices assigned to E. We do this by considering which slices can definitely not be assigned to E and are accounted for by the $k - r$ observed main categories not in E or by the $\overline{r_j}$ observed subcategories not in E. In order to construct such a configuration, we consider the various ways in which we can separate lines or segments on the wheel representing different main categories which either are not in E or which are present in E but have neither end of their segment in E.
First, we could separate these main categories using unobserved main categories in E. There are l of these categories. Secondly, we could separate using observed main-only categories in E. There are r_{main} such categories. Finally, we could separate using the other observed main categories in E, provided that the configuration of the relevant segment is such that each end represents a subcategory in E. There are r_{sub} main categories in E that are described at subcategory level. For a segment to have the required configuration, the category must satisfy $k_j - r_j + 1 \leq r_j + l_j$. This is because we need $k_j - r_j - 1$ subcategories in E to ensure that all subcategories not in E are separated, and a further two to ensure that both ends of the segment are in E. We define the number of main categories which satisfy this condition as $\tilde{r}_{sub}$. We define the number of main categories which are present in E but have neither end of their segment belonging to E, i.e. the number which satisfy $k_j - r_j - 1 \geq r_j + l_j$, as r^0_{sub}.
By maximising the number of slices that may be assigned to E, we find that

$$\overline{P}(E) = \sum_{j \in OJ} \frac{n_j - 1}{n} + \sum_{j \in OJ^*} \sum_{i_j \in OI_j} \frac{n_{j,i_j} - 1}{n} + \frac{\min\{r + l + r_{main} + \tilde{r}_{sub} - r^0_{sub}, k\}}{n} + \sum_{j \in OJ^*} \frac{\min\{2r_j + l_j, k_j - 1\}}{n}. \quad (3)$$

Example 2 *Consider the data set described in Example 1. Suppose that we are interested in the event*

$Y_9 \in \{LB,DB,P\}$. *We have* $k = 4$, $r_{main} = 0$, $r_{sub} = 1$, $r = 1$ *and* $l = 1$. *For main categories described at subcategory level, the values of* k_j, r_j *and* l_j *are shown in Table 2. Here, we find that* $(k - r) + r^0_{sub} > l + r_{main} + \tilde{r}_{sub}$, *i.e. there is no configuration of the probability wheel such that all of the categories not in* E *are separated by a category in* E. *Also, within the* G *segment we cannot assign all separating slices to subcategories in* E. *One configuration of the wheel*

	j	k_j	r_j	l_j
B	1	3	2	0
G	2	2	0	0

Table 2: Values of k_j, r_j and l_j for Example 2

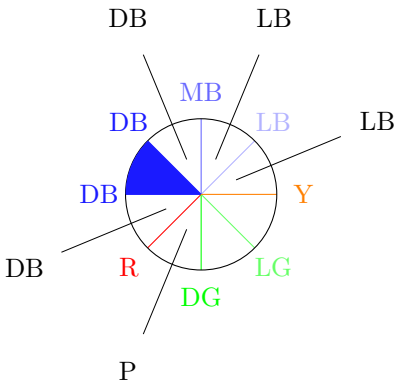


Figure 2: Probability wheel for Example 2

corresponding to the NPI upper probability is shown in Figure 2. Figure 2 shows a configuration where R and Y are separated by B, and G and R are separated by P. However, we cannot separate G and Y by a category in E. We also do not have an available subcategory in E to which we can assign the slice separating DG and LG. This leads to a NPI upper probability of $\frac{6}{8}$ *for the event E. This upper probability can be verified using (3).We see that the set OJ is empty, the set* OJ^* *contains B and the set* OI_1 *contains LB and DB. Also,* $r + l + r_{main} + \tilde{r}_{sub} - r^0_{sub} = 3$ *and* $\sum_{j \in OJ^*} \min\{2r_j + l_j, k_j - 1\} = 2$, *therefore (3) gives* $\overline{P}(E) = \frac{6}{8}$.

4 Unknown number of (sub)categories

In addition to K and K_j being unknown, it is important to note that they are not assumed to have a finite limit. In order to describe the general events of interest in this situation, we introduce some new notation. Let c_{j_s}, $s = 1, ..., r'$, be the observed main-only categories in the event of interest, let UN be the set of Unobserved New main categories, which refers to any not yet observed category, and let DN_j, $j = 1, ..., l$, be the set of Defined New main categories, which is a subset of UN and which represents categories we wish to specify in the event of interest but have not yet observed.
Also, let c_{j_s}, $s = r' + 1, ..., r$, be the observed main categories in the event of interest which are described at subcategory level, and let $s_{j_s,i_{j_s}}$, $s = r' + 1, ..., r$, $i_{j_s} = 1, ..., r_s$, be the observed subcategories in the event of interest. Let $\tilde{DN}_{j_s,i_{j_s}}$, $i_{j_s} = 1, ..., d_s$, be the set of Defined New subcategories within the observed main categories c_{j_s}, and let DN_{j,i_j}, $j = 1, ..., l$, $i_j = 1, ..., l_j$, be the set of Defined New subcategories within the Defined New main categories. Let $\tilde{UN}_{j_s}$, $s = 1, ..., r$ be the set of all Unobserved New subcategories within the observed main categories c_{j_s}, and let UN_j, $j = 1, ..., l$ be the set of all Unobserved New subcategories within the Defined New main categories. A given event can be expressed as a union involving some or all of the above terms. Let $A, B \subseteq \{1, ..., k\}$ such that $A \cap B = \emptyset$. Any event of interest can be expressed using one of the two formulae shown below. The first general event is

$$\begin{aligned} Y_{n+1} \in & \bigcup_{s=1}^{r'} c_{j_s} \cup \bigcup_{s=r'+1}^{r} \Big(\bigcup_{i_{j_s}=1}^{r_s} s_{j_s,i_{j_s}} \Big) \\ & \cup \bigcup_{s \in A} \Big(\tilde{UN}_{j_s} \setminus \bigcup_{i_{j_s}=1}^{d_s} \tilde{DN}_{j_s,i_{j_s}} \Big) \\ & \cup \bigcup_{s \in B} \Big(\bigcup_{i_{j_s}=1}^{d_s} \tilde{DN}_{j_s,i_{j_s}} \Big) \\ & \cup \bigcup_{j=1}^{l'} \Big(UN_j \setminus \bigcup_{i_j=1}^{l_j} DN_{j,i_j} \Big) \cup \bigcup_{j=l'+1}^{l} \Big(\bigcup_{i_j=1}^{l_j} DN_{j,i_j} \Big). \end{aligned} \tag{4}$$

The second general event is

$$\begin{aligned} Y_{n+1} \in & \bigcup_{s=1}^{r'} c_{j_s} \cup \bigcup_{s=r'+1}^{r} \Big(\bigcup_{i_{j_s}=1}^{r_s} s_{j_s,i_{j_s}} \Big) \\ & \cup \bigcup_{s \in A} \Big(\tilde{UN}_{j_s} \setminus \bigcup_{i_{j_s}=1}^{d_s} \tilde{DN}_{j_s,i_{j_s}} \Big) \cup \bigcup_{s \in B} \Big(\bigcup_{i_{j_s}=1}^{d_s} \tilde{DN}_{j_s,i_{j_s}} \Big) \\ & \cup UN \setminus \Big\{ \bigcup_{j=1}^{l'} \Big(UN_j \setminus \bigcup_{i_j=1}^{l_j} DN_{j,i_j} \Big) \\ & \cup \bigcup_{j=l'+1}^{l} \Big(\bigcup_{i_j=1}^{l_j} DN_{j,i_j} \Big) \Big\}. \end{aligned} \tag{5}$$

We denote these by E_1 (4) and E_2 (5). We now present formulae for the NPI lower and upper

probabilities for each of these general events. A detailed derivation of these formulae is given in [4].

4.1 Lower probability

First we consider event E_1, which includes only a finite number of unobserved main categories. By minimising the number of slices of the wheel that must be assigned to E_1, the NPI lower probability is

$$\underline{P}(E_1) = \sum_{s=1}^{r'} \frac{n_{j_s} - 1}{n} + \sum_{s \notin A} \{ \sum_{s=r'+1}^{r} (\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s,i_{j_s}} - 1}{n})\} + \sum_{s \in A} \{ \sum_{s=r'+1}^{r} (\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s,i_{j_s}} - 1}{n}) + \frac{N_s}{n} \} \tag{6}$$

where $N_s = [(r_s - 1) - d_s - (k_{j_s} - r_s)]^+$.
For E_2, which contains all except a finite number of the UN main categories, the NPI lower probability is

$$\underline{P}(E_2) = \sum_{s=1}^{r'} \frac{n_{j_s} - 1}{n} + \sum_{s \notin A} \{ \sum_{s=r'+1}^{r} (\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s i_{j_s}} - 1}{n})\} + \sum_{s \in A} \{ \sum_{s=r'+1}^{r} (\sum_{i_{j_s}=1}^{r_s} \frac{n_{j_s i_{j_s}} - 1}{n}) + \frac{N_s}{n} \} + \frac{1}{n}(2r - k - l - \sum_{s=r'+1}^{r} M_s)^+ \tag{7}$$

where $M_s = 2$ if $s \notin A$ and $M_s = \min\{[d_s + (k_{j_s} - r_s) - (r_s - 1)]^+, 2\}$ if $s \in A$.

Example 3 *Consider a multinomial data set where the set of possible main categories consists of an unknown number of different colours. We have observed the following main categories: red (R), blue (B), green (G) and pink (P). At subcategory level, we have observed DB, MB, LB, DG, MG, LG, MP and DP. In addition we define two new main categories: orange (O), with defined subcategories LO and MO, and purple (Pu) with defined subcategory DPu. We also define the new subcategory LP. We let UN_B represent all unobserved new subcategories within the main category B, including the defined new subcategory RB, and let UN_{Pu} represent the equivalent for the main category Pu. The data set consists of twenty observations including 3 R, 3 DB, 1 MB, 2 LB, 3 DG, 2 MG, 2 LG, 2 MP and 2 DP. We consider the event $Y_{21} \in \{(LB \cup MB) \cup (LG \cup MG) \cup (MP) \cup (UN_B \backslash RB) \cup (LP) \cup [UN \backslash ((UN_{Pu} \backslash DPu) \cup (LO \cup MO))]\}$. We label this event E. Let $s = 1$ correspond to B, $s = 2$ to G and $s = 3$ to P. This is an event of type E_2 (5), so (7) is used to compute the NPI lower probability for this event.*
In this example, $r = 3$. The main categories for which $s \notin A$ are G and P, and the only main category for which $s \in A$ is B. We have $N_1 = [(r_1 - 1) - d_1 - (k_{j_1} - r_1)]^+ = [(2 - 1) - 1 - (3 - 2)]^+ = 0$. We also have $M_1 = \min\{[d_1 + (k_{j_1} - r_1) - (r_1 - 1)]^*, 2\} = 1$, $M_2 = 2$ and $M_3 = 2$. Therefore $\sum_{s=1}^{3} M_s = 5$. The values of n_{j_s} and $n_{j_s,i_{j_s}}$ are shown in Tables 3 and 4.

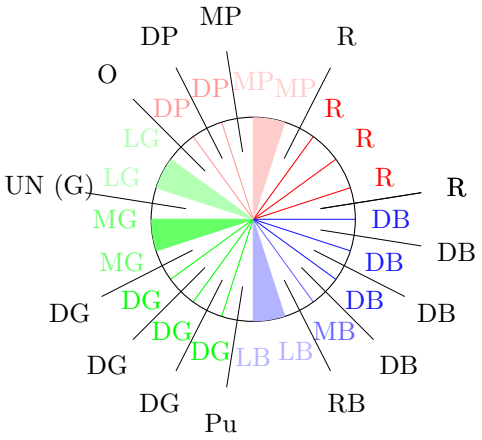


Figure 3: Probability wheel for Example 3

	B	G	P
n_{j_s}	6	7	4

Table 3: Values of n_{j_s} for Example 3

	LB	MB	LG	MG	MP
$n_{j_s,i_{j_s}}$	2	1	2	2	2

Table 4: Values of $n_{j_s,i_{j_s}}$ for Example 3

By (7), the NPI lower probability for the event E is $\frac{4}{20}$. Figure 3 shows a corresponding configuration of the probability wheel. There are four slices assigned to E, and the remaining slices are assigned to main categories or subcategories not in E and are labelled accordingly.

4.2 Upper probability

The NPI upper probabilities for events E_1 and E_2 are derived by assigning as many slices of the wheel as possible to the event of interest. The NPI upper

probability for event E_1 is

$$\overline{P}(E_1) = \sum_{s=1}^{r'} \frac{n_{j_s}-1}{n} + \sum_{s\notin A}\{\sum_{s=r'+1}^{r}(\sum_{i_{j_s}=1}^{r_s}\frac{n_{j_s,i_{j_s}}-1}{n}) + \frac{k_{j_s}-1-P_s}{n}\} + \frac{\min\{r-r^0+l+\tilde{r},k\}}{n} + \sum_{s\in A}\{\sum_{s=r'+1}^{r}(\sum_{i_{j_s}=1}^{r_s}\frac{n_{j_s,i_{j_s}}-1}{n}) + \frac{k_{j_s}-1}{n}\} \quad (8)$$

where $P_s = [(k_{j_s}-r_s-1)-r_s-d_s]^+$, r^0 denotes the number of main categories such that $s \notin A$ which satisfy $r_s + d_s - (k_{j_s} - r_s - 1) \leq 0$, and $\tilde{r}$ denotes the number of main categories c_{j_s} which satisfy either $s \in \{1, ..., r'\}$, $s \in A$ or the condition

$$s \notin A, \qquad r_s + d_s - (k_{j_s} - r_s - 1) \geq 2.$$

The NPI upper probability for event E_2 is

$$\overline{P}(E_2) = \sum_{s=1}^{r'} \frac{n_{j_s}-1}{n} + \sum_{s\notin A}\{\sum_{s=r'+1}^{r}(\sum_{i_{j_s}=1}^{r_s}\frac{n_{j_s i_{j_s}}-1}{n}) + \frac{k_{j_s}-1-P_s}{n}\} + \frac{k}{n} + \sum_{s\in A}\{\sum_{s=r'+1}^{r}(\sum_{i_{j_s}=1}^{r_s}\frac{n_{j_s i_{j_s}}-1}{n}) + \frac{k_{j_s}-1}{n}\}. \quad (9)$$

Example 4 *Consider the data set described in Example 3. Suppose that we are interested in the event* $Y_{21} \in \{(LB \cup MB) \cup (LG \cup MG) \cup (MP) \cup (UN_B \backslash RB) \cup (LP) \cup (UN_{Pu} \backslash DPu) \cup (LO \cup MO)\}$. *We label this event* E. *This is an event of type* E_1, *so (8) is used for the NPI upper probability for* E.
In this example, $r = 3$, $l = 2$ *and* $k = 4$. *Let* $s = 1$ *correspond to B,* $s = 2$ *to G and* $s = 3$ *to P. The main categories for which* $s \notin A$ *are G and P, and the only main category for which* $s \in A$ *is B. We have* $P_2 = [(k_{j_2} - r_2 - 1) - r_2 - d_2]^+ = [(3-2-1)-2-1]^+ = 0$ *and* $P_3 = [(k_{j_3} - r_3 - 1) - r_3 - d_3]^+ = [(2-2-1)-1-1]^+ = 0$. *The values of* n_{j_s}, k_{j_s} *and* $n_{j_s,i_{j_s}}$ *are shown in Tables 5 and 6.*

	B	G	P
n_{j_s}	6	7	4
k_{j_s}	3	3	2

Table 5: Values of n_{j_s} and k_{j_s} for Example 4

We have $r^0 = 0$ *and* $\tilde{r} = 3$, *as both of the main categories in E for which* $s \notin A$ *satisfy the condition* $r_s + d_s - (k_{j_s} - r_s - 1) \geq 2$. *The general formula (8) shows that the NPI upper probability for the event E is* $\frac{13}{20}$. *Figure 4 shows a corresponding configuration of the probability wheel. There are four slices of the wheel that must be assigned to E. The nine further slices that can be assigned to elements of E are labelled accordingly.*

	LB	MB	LG	MG	MP
$n_{j_s,i_{j_s}}$	2	1	2	2	2

Table 6: Values of $n_{j_s,i_{j_s}}$ for Example 4

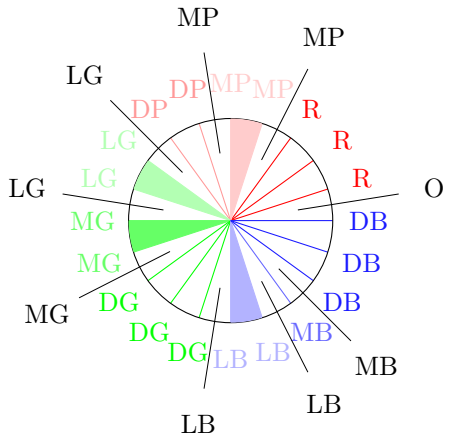


Figure 4: Probability wheel for Example 4

5 Properties of the model

We now discuss several properties of the Sub-MNPI model. We focus here on the case where K and K_j are known, but the following properties are equally applicable when these quantities are unknown.
A fundamental property for lower and upper probabilities is the conjugacy property, which states that $\overline{P}(E) = 1 - \underline{P}(E^c)$. This is implicit in the F-probability property, proven below, but can also be proven explicitly for the Sub-MNPI model [4]. It can also be shown [4] that the interval between the lower and upper probabilities always contains the relative frequency of observations in the event of interest E, i.e.

$$\underline{P}(E) \leq \sum_{j\in OJ}\frac{n_j}{n} + \sum_{j\in OJ^*}\sum_{i_j\in OI_j}\frac{n_{j,i_j}}{n} \leq \overline{P}(E). \quad (10)$$

This is an attractive property, since it shows that the Sub-MNPI model is not in conflict with the empirical probability, and one which is not always satisfied by methods such as Bayesian inferences which typically assign a positive probability to a category before it has been observed even once. A third property that can be proven [4] is that as the number of observations in the data set becomes infinitely large, the imprecision vanishes and the interval probability $P(E)$ shrinks to

a point value equal to the relative frequency. This is, in our situation, a desirable property for the model.
We now prove that the interval probabilities $[\underline{P}(E), \overline{P}(E)]$ given by the Sub-MNPI model are F-probabilities in the sense of Weichselberger [13]. F-probability is a desirable property, because it shows that none of the interval probabilities are too wide and that they could not be made any smaller given the data available to us. Also, F-probability is strongly linked to other concepts in imprecise probability theory. As stated above, conjugacy is implicit in the F-probability property. Coherence is a direct consequence of F-probability, by Walley's lower envelope theorem [11], and this can be seen as a rationality requirement. The following is based on work by Coolen and Augustin [7] that proved the F-probability property for the original NPI model for multinomial data.
For the proof we introduce some new notation in order to describe all the possible configurations of the probability wheel. Suppose that the wheel is split into K segments, and each segment is split into K_j subsegments. We move clockwise around the wheel numbering the segments as $1, ..., K$ as shown in Figure 5. We also number the subsegments within segment j as $1, ..., K_j$ as shown in Figure 6. The area of these

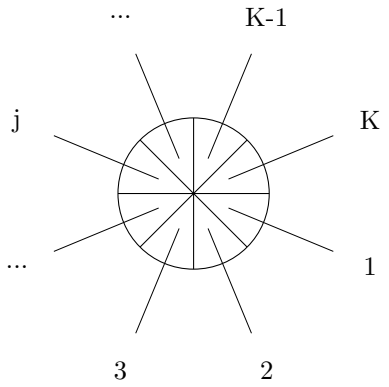


Figure 5: Numbering of segments

segments and subsegments is thus far unspecified: we allocate a different main category or subcategory to each segment or subsegment in order to describe the configuration of the wheel, but a segment assigned to an unobserved category may have area zero.
As seen in [7], we let Σ represent the set of all possible configurations σ of the wheel. Each σ can be described by a sequence

$$(\sigma(j))_{j=1...K+1}, \qquad \sigma(K+1) = \sigma(1)$$

where $\sigma(j)$ is the index of the main category assigned to segment j, and a set of sequences

$$(\sigma(i,j))_{i=1...K_j}, \qquad j \in J^*$$

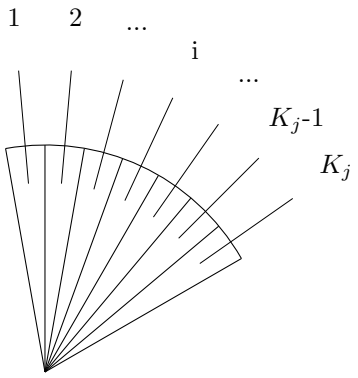


Figure 6: Numbering of subsegments

where $\sigma(i,j)$ is the index of the subcategory within main category j assigned to subsegment i.
It is also necessary to describe the position of the observed main categories and subcategories on the wheel for a given σ. Let the circular sequence

$$\sigma(i_1), ..., \sigma(i_{k+1}), \sigma(i_{k+1}) = \sigma(i_1)$$

be the indices of the observed main categories as we move around the wheel, and let the sequence

$$\sigma(i_1, j), ..., \sigma(i_{k_j}, j), \qquad j \in J^*$$

be the indices of the observed subcategories as we move through the segment representing main category j.
For $l = 1, ..., k$, we describe each separating slice between two main categories as follows:

$$J_{\sigma,l} = \{\sigma(j) | i_l \leq j \leq i_{l+1}\}$$

if categories in positions i_l and i_{l+1} are main-only,

$$J_{\sigma,l} = \{\sigma(j) | i_l \leq j < i_{l+1}\} \cup \bigcup_{x=1}^{i_1} \sigma(x, l+1)$$

if category in position i_l is main-only but category in position i_{l+1} has subcategories,

$$J_{\sigma,l} = \{\sigma(j) | i_l < j \leq i_{l+1}\} \cup \bigcup_{x=i_{k_l}}^{K_l} \sigma(x, l)$$

if category in position i_l has subcategories but category in position i_{l+1} is main-only, and

$$J_{\sigma,l} = \{\sigma(j) | i_l < j < i_{l+1}\} \cup \bigcup_{x=1}^{i_1} \sigma(x, l+1) \cup \bigcup_{x=i_{k_l}}^{K_l} \sigma(x, l)$$

if categories in positions i_l and i_{l+1} both have subcategories. $J_{\sigma,l}$ is the index set of all main categories and subcategories to which the separating

slice could be assigned. Let $c_{|J_{\sigma,l}|}$ be the set of all these main categories and subcategories.
We also describe the separating slice between two observed subcategories within the same main category using

$$B_{\sigma,j,l} = \{\sigma(b,j) | i_l \leq b \leq i_{l+1}\}, l = 1, ..., k_j - 1, j \in J^*.$$

This is the set of indices of all possible subcategories to which, for the particular configuration σ, we could assign the separating slice between the subcategories in positions i_l and i_{l+1} in the segment representing main category j. Let $s_{|B_{\sigma,j,l}|}$ be the set of these subcategories.
Now, for a given configuration σ, the Sub-MNPI model gives the following basic probability assignment [3] to the event $Y_{n+1} \in c_j$:

$$m_\sigma(Y_{n+1} \in c_j) = \max\{\frac{n_j - 1}{n}, 0\}, j = 1, ..., K.$$

Similarly, the basic probability assignment given to the event $Y_{n+1} \in s_{j,i_j}$ is

$$m_\sigma(Y_{n+1} \in s_{j,i_j}) = \max\{\frac{n_{j,i_j} - 1}{n}, 0\}, i = 1, ..., K_j.$$

With regard to distributing probability mass amongst slices separating different main categories or subcategories, we give the following basic probability assignments:

$$m_\sigma(Y_{n+1} \in c_{|J_{\sigma,l}|}) = \frac{1}{n}, \quad l = 1, ..., k.$$

$$m_\sigma(Y_{n+1} \in s_{|B_{\sigma,j,l}|}) = \frac{1}{n}, \quad l = 1, ..., k_j - 1, j \in J^*.$$

Any other event is given the basic probability assignment of zero.
Let X_E represent the index set of the event of interest E. This set contains some one-dimensional elements, corresponding to main-only categories, and some two-dimensional elements, corresponding to subcategories. We now determine the lower and upper probabilities for event E via the belief and plausibility functions [10]. For a particular configuration σ, we find that the belief function of E is

$$\begin{aligned}\underline{P}_\sigma(E) = &\sum_{j \in J} m_\sigma(\{Y_{n+1} \in c_j\}) \\ &+ \sum_{j \in J^*} \sum_{i_j \in I_j} m_\sigma(\{Y_{n+1} \in s_{j,i_j}\}) \\ &+ \sum_{J_{\sigma,l} \subseteq X_E} m_\sigma(\{Y_{n+1} \in c_{|J_{\sigma,l}|}\}) \\ &+ \sum_{B_{\sigma,j,l} \subseteq I_j} m_\sigma(\{Y_{n+1} \in s_{|B_{\sigma,j,l}|}\})\end{aligned} \tag{11}$$

and the plausibility function of E is

$$\begin{aligned}\overline{P}_\sigma(E) = &\sum_{j \in J} m_\sigma(\{Y_{n+1} \in c_j\}) \\ &+ \sum_{j \in J^*} \sum_{i_j \in I_j} m_\sigma(\{Y_{n+1} \in s_{j,i_j}\}) \\ &+ \sum_{J_{\sigma,l} \cap X_E \neq \emptyset} m_\sigma(\{Y_{n+1} \in c_{|J_{\sigma,l}|}\}) \\ &+ \sum_{B_{\sigma,j,l} \cap I_j \neq \emptyset} m_\sigma(\{Y_{n+1} \in s_{|B_{\sigma,j,l}|}\}).\end{aligned} \tag{12}$$

We therefore have a set of belief functions and a set of plausibility functions corresponding to the set Σ of possible configurations of the probability wheel. According to Theorem 3.2 of [3], and to [9], taking the lower and upper envelopes over all possible configurations leads to F-probability. Since the lower and upper probability formulae of the Sub-MNPI model are derived by considering all possible configurations $\sigma \in \Sigma$, resulting in

$$\underline{P}(E) = \min_{\sigma \in \Sigma} \underline{P}_\sigma(E)$$

and

$$\overline{P}(E) = \max_{\sigma \in \Sigma} \overline{P}_\sigma(E),$$

the interval probability $[\underline{P}(E), \overline{P}(E)]$ is an F-probability.

6 Approximate maximum entropy distribution

We present an algorithm for approximating the maximum entropy distribution consistent with the Sub-MNPI model, with a view to using this maximum entropy measure in the construction of classification trees. Further details of such classification at main category level are presented in [4]; the implementation of this method at subcategory level is ongoing research.
The process of computing the maximum entropy distribution is carried out in two stages. Initially, we work at main category level only. We apply the NPI-M algorithm presented in [1], which gives a maximum entropy probability $p_{maxE}(c_j)$ for each main category. As a second step, we share the probability mass $p_{maxE}(c_j)$ as evenly as possible between the subcategories, in such a way that the probability $\widehat{p}_{j,i_j}$ that is assigned by the algorithm to subcategory s_{j,i_j} is within the interval $[L_{j,i_j}, U_{j,i_j}]$. Let $K(i)_j$ represent the number of subcategories in main category c_j that have been observed i times. From the NPI-M algorithm [1] we have the results $p_j = p_{maxE}(c_j), j = 1, ..., K$. This means that for

each main category c_j, we have a segment consisting of np_j slices. Of these slices, $n(\sum_{i=1}^{K_j} L_{j,i_j})$ must be assigned to observed subcategories in c_j. We therefore have remaining probability mass $p_j - \sum_{i=1}^{K_j} L_{j,i_j}$ that may be assigned to any available subcategory in c_j, and this is termed optional probability mass. For each c_j, we share the optional probability mass between subcategories of c_j, beginning with subcategories with the fewest observations. This leads to the Sub-A-NPI-M algorithm, which is shown below in pseudo-code and which is similar to the A-NPI-M algorithm presented in [1] and justified in the same way.

Sub-A-NPI-M

For $j = 1$ to K

For $i = 1$ to K_j

$L_{j,i_j} \leftarrow \max\{\frac{n_{j,i_j}-1}{n}, 0\}$

$opt \leftarrow p_j - \sum_{i=1}^{K_j} L_{j,i_j}$

$\widehat{p}_{j,i_j} \leftarrow L_{j,i_j}$

$t \leftarrow 0$;

While $(opt > 0)$ do

If $(n_{j,i_j} = t$ *or* $n_{j,i_j} = t+1)$ $\widehat{p}_{j,i_j} \leftarrow \widehat{p}_{j,i_j} + \min\{\frac{opt}{K(t)_j + K(t+1)_j}, \frac{1}{n}\}$;

$opt \leftarrow opt - \min\{\frac{opt}{K(t)_j + K(t+1)_j}, \frac{1}{n}\}$;

$t \leftarrow t + 1$;

The Sub-A-NPI-M algorithm is illustrated in Example 5.

Example 5 *Consider a multinomial data set with observed main categories blue (B), green (G), red (R) and pink (P), and unobserved main category orange (O). Observations in B are further classified as light blue (LB) or dark blue (DB), and observations in G are further classified as light green (LG) or dark green (DG). The data set consists of twenty observations altogether, including 5 DB, 5 DG, 5 R and 5 P.*
First, considering the data at main category level only, we apply the NPI-M algorithm [1] and find that the maximum entropy probabilities assigned to the main categories $\{O,R,B,G,P\}$ *are* $\{\frac{1}{20}, \frac{19}{80}, \frac{19}{80}, \frac{19}{80}, \frac{19}{80}\}$. *(For further details on this, see [1] and [4].) A configuration of the wheel corresponding to this distribution is shown in Figure 7. The separating slices are shared in such a way that B, R, G and P are each assigned* $\frac{3}{4}$ *of a separating slice. We now consider the subcategories. The maximum entropy probabilities for the main categories are distributed over the subcategories using the Sub-A-NPI-M algorithm. For main category B we have* $\underline{P}(DB) = \frac{4}{20}$ *and* $\underline{P}(LB) = 0$. *For main category G we have* $\underline{P}(DG) = \frac{4}{20}$ *and* $\underline{P}(LG) = 0$. *Applying the Sub-A-NPI-M algorithm, we find that* $opt = \frac{19}{80} - \frac{4}{20} = \frac{3}{80}$ *for both of these main categories. Taking* $t = 0$ *gives*

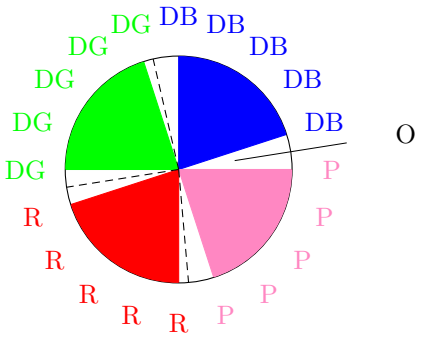


Figure 7: Probability wheel for Example 5

$$\widehat{p}(LB) = 0 + \min\{\frac{opt}{K(t)_j + K(t+1)_j}, \frac{1}{n}\} = \frac{3}{80},$$

$$\widehat{p}(LG) = 0 + \min\{\frac{opt}{K(t)_j + K(t+1)_j}, \frac{1}{n}\} = \frac{3}{80}.$$

So the probabilities assigned to the set of subcategories $\{LB,DB\}$ *are* $\{\frac{3}{80}, \frac{4}{20}\}$ *and the probabilities assigned to the set of subcategories* $\{LG,DG\}$ *are* $\{\frac{3}{80}, \frac{4}{20}\}$.

The Sub-A-NPI-M algorithm can be implemented for building classification trees using methodology similar to that shown in [2] and in [4].

7 Concluding remarks

In this paper we presented the Sub-MNPI model for inferences from multinomial data described at subcategory level as well as at main category level. NPI lower and upper probabilities were derived for the general events of interest, and some fundamental properties of the model were explained. The inferences presented here are more flexible than those given by the original NPI model for multinomial data in the sense that observations can be represented at varying levels of detail, which makes the model widely applicable to practical problems. With the view to applying the Sub-MNPI model to classification problems, an algorithm was presented for approximating the maximum entropy distribution consistent with these inferences. Implementation of this algorithm for building classification trees, and

comparison of the approach with alternative imprecise and classical methods, is ongoing. It is also of interest for future research to investigate other applications of the Sub-MNPI model.

With regard to future research, it will also be useful to compare classification trees built using the Sub-A-NPI-M algorithm presented here with classification trees constructed by ignoring the hierarchical relationship between the categories and subcategories and simply using the NPI-M algorithm presented in [1]. Note that the distinction between these two methods, and the different results they achieve, show that the Representation Invariance Principle (RIP) satisfied by Walley's IDM [12] does not generally hold for NPI. This is an issue discussed in detail by Coolen and Augustin [5, 7].

The Sub-MNPI model presented in this paper could be extended further by considering inferences about multiple future observations and by introducing further layers e.g. subsubcategories to the hierarchy. Such developments would be of theoretical and practical interest.

Acknowledgements

We thank two referees for supportive comments and suggestions to improve the presentation of this paper.

References

[1] Abellán, J., Baker, R.M. and Coolen, F.P.A. (2011) Maximising entropy on the nonparametric predictive inference model for multinomial data. *European Journal of Operational Research*, **212**, 112-122.

[2] Abellán, J. and Moral, S. (2003) Building classification trees using the total uncertainty criterion. *International Journal of Intelligent Systems*, **18**, 1215-1225.

[3] Augustin, T. (2005) Generalized basic probability assignments. *International Journal of General Systems*, **34**, 451-463.

[4] Baker, R.M. (2010) Nonparametric predictive inference: Selection, classification and subcategory data, PhD thesis, Durham University. *http://etheses.dur.ac.uk/257*

[5] Coolen, F.P.A. and Augustin, T. (2005) Learning from multinomial data: a nonparametric predictive alternative to the Imprecise Dirichlet Model. *Proceedings of the Fourth International Symposium on Imprecise Probability: Theories and Applications.*

[6] Coolen, F.P.A. and Augustin, T. (2007) Multinomial nonparametric predictive inference with subcategories. *Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications.*

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

[9] Miranda, E., de Cooman, G. and Couso, I. (2005) Lower previsions induced by multi-valued mappings. *Journal of Statistical Planning and Inference*, **133**, 173-197.

[10] Shafer, G. (1976) *A Mathematical Theory of Evidence*. Princeton University Press.

[11] Walley, P. (1991) *Statistical Reasoning With Imprecise Probabilities.* Chapman and Hall.

[12] Walley, P. (1996) Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society* B, **58**, 3-57.

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

Structural Reliability Assessment with Fuzzy Probabilities

Michael Beer
Centre for Engineering Sustainability,
School of Engineering,
University of Liverpool, UK
mbeer@liverpool.ac.uk

Zhang Mingqiang
Department of Civil & Environmental Engineering, National University of Singapore,
Singapore
mingqiang@nus.edu.sg

Quek Ser Tong
Department of Civil & Environmental Engineering, National University of Singapore,
Singapore
st.quek@nus.edu.sg

Scott Ferson
Applied Biomathematics,
Setauket, NY, USA
scott@ramas.com

Abstract

The prediction of the behavior and reliability of engineering structures and systems is often plagued by uncertainty and imprecision caused by sparse data, poor measurements and linguistic information. Accounting for such limitations complicates the mathematical modeling required to obtain realistic results in engineering analyses. The framework of imprecise probabilities provides a mathematical basis to deal with these problems which involve both probabilistic and non-probabilistic sources of uncertainty. A common feature of the various concepts of imprecise probabilities is the consideration of an entire set of probabilistic models in one analysis. But there are differences between the concepts in the mathematical description of this set and in the theoretical connection to the probabilistic models involved. This study is focused on fuzzy probabilities, which combine a probabilistic characterization of variability with a fuzzy characterization of imprecision. We discuss how fuzzy modeling can allow a more nuanced approach than interval-based concepts. The application in engineering is demonstrated by means of two examples.

Keywords. Fuzzy Probabilities, Imprecise Probabilities, Failure Probability, Reliability Analysis.

1 Introduction

The analysis and reliability assessment of engineering structures and systems involves uncertainty and imprecision in parameters and models of different type. In order to derive predictions regarding structural behavior and reliability, it is crucial to represent the uncertainty and imprecision appropriately according to the underlying real-world information which is available. To capture variation of structural parameters, established probabilistic models and powerful simulation techniques are available for engineers, which are widely applicable to real-world problems; for example, see [24]. The required probabilistic modeling can be realized via classical mathematical statistics if data of a suitable quality are available to a sufficient extent.

In civil engineering practice, however, the available data are frequently quite limited and of poor quality. These limitations create epistemic uncertainty, which can sometimes be substantial. It is frequently argued that expert knowledge can compensate for the limitations through the use of Bayesian methods based on subjective probabilities. If a subjective perception regarding a probabilistic model exists and some data for a model update can be made available, a Bayesian approach can be very powerful, and meaningful results with maximal information content can be derived. Bayesian approaches have attracted increasing attention in the recent past and considerable advancements have been reported for the solution of various engineering problems [7, 15, 23]. An important feature of Bayesian updating is that the subjective influence in the model assumption decays quickly with growing amount of data. It is then reasonable practice to estimate probabilistic model parameters based on the posterior distribution, for example, as the expected value thereof.

When less information and experience are available, greater difficulties will be faced. If the available information is very scarce or is of an imprecise nature rather than of a stochastic nature, a subjective probabilistic model description may be quite arbitrary. For example, a distribution parameter may be known merely in the form of bounds. Any prior distribution which is limited to

these bounds would then be an option for modeling. But the selection of a particular model would introduce unwarranted information that cannot be justified sufficiently. Even the assumption of a uniform distribution, which is commonly used in those cases, ascribes more information than is actually given by the bounds. This situation may become critical if no or only very limited data are available for a model update. The initial subjectivity is then dominant in the posterior distribution and in the final result. If these results, such as failure probabilities, determine critical decisions, one may wish to consider the problem from the following angle.

If several probabilistic models are plausible for the description of a problem, and no information is available to assess the suitability of the individual models or to relate their suitability with respect to one another, then it may be of interest to identify the worst case for the modeling rather than to average over all plausible model options with arbitrary weighting. The probabilistic analysis is carried out conditional on each of many particular probabilistic models out of the set of plausible models. In reliability assessment, this implies the calculation of an upper bound for the failure probability as the worst case. This perspective can be extended to explore the sensitivity of results with respect to the variety of plausible models, that is, with respect to a subjective model choice. A mathematical framework for an analysis of this type has been established with imprecise probabilities; see [28]. Applications to reliability analysis [17, 22, 26] and to sensitivity analysis [9, 13] have been reported. This intuitive view, however, is by far not the entire motivation for imprecise probabilities [16]. Imprecise probabilities are not limited to a consideration of imprecise distribution parameters. They are capable of dealing with imprecise conditions and dependencies between random variables and with imprecise structural parameters and model descriptions. Respective discussions can be reviewed, for example, in [8, 14]. Multivariate models can be constructed [11]. Imprecise probabilities also allow statistical estimations and tests with imprecise sample elements. Results from robust statistics in form of solution domains of statistical estimators can be considered directly and appropriately [1].

In this paper, the implementation of intervals and fuzzy sets as parameters of probabilistic models is discussed in the context of proposed concepts of imprecise probabilities. Structural reliability analysis is employed to illustrate the effects in examples.

2 Imprecise Probabilistic Model Parameters

In engineering analyses, parameters of probabilistic models are frequently limited in precision and are only known in a coarse manner. This situation can be approached with different mathematical concepts. First, the parameter can be considered as uncertain with random characteristics, which complies with the Bayesian approach. Subjective probability distributions for the parameters are updated by means of objective information in form of data. The result is a mix of objective and subjective information – both expressed with probability. Second, the parameter can be considered as imprecise but bounded within a certain domain, where the domain is described as a set. In this manner, only the limitation to some domain and no further specific characteristics are ascribed to the parameter, which introduces significantly less information in comparison with a distribution function as used in the Bayesian approach. Imprecision in the form of a set for a parameter does not migrate into probabilities, but it is reflected in the result as a set of probabilities which contains the true probability. Intervals and fuzzy sets can thus be considered as models for parameters of probability distributions.

An interval is an appropriate model in cases where only a possible range between crisp bounds x_l and x_r is known for the parameter x, and no additional information concerning value frequencies, preference, etc. between interval bounds is available nor any clues on how to specify such information. Interval modeling of a parameter of a probabilistic model connotes the consideration of a set of probabilistic models, which are captured by the set of parameter values

$$X_I = \left[x_l , x_r \right] . \tag{1}$$

This modeling corresponds to the p-box approach [10] and to the theory of interval probabilities [28, 29]. Events E_i are assessed with a range of probability, $\left[P_l \left(E_i \right), P_r \left(E_i \right) \right] \subseteq \left[0,1 \right]$, which is directly used for the definition of interval probability, denoted as *IP*, as follows,

$$\begin{aligned} &IP: \; E_\Omega \to I \;\; with \\ &E_\Omega = \mathfrak{P}\left(\Omega \right), \; I = \left\{ \left[a,b \right] : \; 0 \le a \le b \le 1 \right\} . \end{aligned} \tag{2}$$

In Eq. (2), $\mathfrak{P}\left(\Omega \right)$ is the power set on the set Ω of elementary events ω. This definition complies with traditional probability theory. Kolmogorov's axioms and the generation scheme of events are retained as defined in traditional probability theory, see also [30]. Traditional mathematical statistics are applicable for quantification purposes. In reliability analysis with interval probabilities, the parameter interval X_I is mapped to an interval of the failure probability,

$$X_I \to P_{fI} = \left\{ P_f \left| P_f \in \left[P_{fl}, P_{fr} \right] \right. \right\} . \tag{3}$$

Scrutinizing the modeling of parameters as intervals shows that an interval is a quite crude expression of imprecision. The specification of an interval for a parameter implies that, although a number's value is not known

exactly, exact bounds on the number can be provided. This may be criticized because the specification of precise numbers is just transferred to the bounds. Fuzzy set theory provides a suitable basis for relaxing the need for precise values or bounds. It allows the specification of a smooth transition for elements from belonging to a set to not belonging to a set. Fuzzy numbers are a generalization and refinement of intervals for representing imprecise parameters. The essence of an approach using fuzzy numbers that distinguishes it from more traditional approaches is that it does not require the analyst to circumscribe the imprecision all in one fell swoop with finite characterizations having known bounds. The analyst can now express the available information in the form of a series of plausible intervals, the bounds of which may grow, including the case of infinite limits. This allows a more nuanced approach compared to interval modeling.

Fuzzy sets provide an extension to interval modeling that considers variants of interval models, in a nested fashion, in one analysis. A fuzzy set $\tilde{X}$ of parameter values can be represented as a set of intervals X_I,

$$\tilde{X} = \left\{ \left(X_\alpha, \mu\left(x \in X_\alpha\right)\right) \middle| \begin{array}{l} X_\alpha = X_I, \\ \mu\left(x \in X_\alpha\right) = \alpha \\ \qquad \forall \alpha \in (0,1] \end{array} \right\} . \tag{4}$$

This is utilized for an approximation of $\tilde{X}$ via a series of discrete values $\alpha_i \in (0,1]$, which is referred to as α-discretization; see Figure 1 [31]. In Eq. (4), X_α denotes an α-level set of the fuzzy set $\tilde{X}$, and $\mu(.)$ is the membership function. This modeling applied to parameters of a probabilistic model corresponds to the theory of fuzzy random variables and to fuzzy probability theory according to [4, 18]. For further information on related concepts, see [6, 12, 19]. The definition of a fuzzy random variable refers to imprecise observations as outcome of a random experiment. A fuzzy random variable $\tilde{Y}$ is the mapping

$$\tilde{Y}\colon\ \Omega \to \mathfrak{F}(\mathbf{Y}) \tag{5}$$

with $\mathfrak{F}(\mathbf{Y})$ being the set of all fuzzy sets on the fundamental set $\mathbf{Y}$, whereby the standard case is $\mathbf{Y} = \mathrm{R}^n$. The pre-images of the imprecise events described by $\mathfrak{F}(\mathbf{Y})$ are elements of a traditional probability space $[\Omega, \mathcal{G}, P]$. This complies with traditional probability theory and allows statistics with imprecise data [2, 18, 27]. As a consequence of Eq. (5), parameters of probabilistic models, including descriptions of the dependencies and distribution type, and probabilities are obtained as fuzzy sets. This builds the relationship to the p-box approach and to the theory of interval probabilities. A representation of a fuzzy probability distribution function of a fuzzy random variable $\tilde{Y}$ with aid of α-discretization leads to interval probabilities $[F_{\alpha l}(y), F_{\alpha r}(y)]$ for each α-level as one plausible model variant,

$$\tilde{F}(y) = \left\{ \left(F_\alpha(y), \mu\left(F(y) \in F_\alpha(y)\right)\right) \right\} \tag{6}$$

with

$$F_\alpha(y) = [F_{\alpha l}(y), F_{\alpha r}(y)] , \tag{7}$$

$$\mu\left(F(y) \in F_\alpha(y)\right) = \alpha \forall \alpha \in (0,1] . \tag{8}$$

As depicted in Figure 1, in a reliability analysis, the fuzzy set $\tilde{X}$ of parameter values is mapped to a fuzzy set of the failure probability,

$$\tilde{X} \to \tilde{P}_f . \tag{9}$$

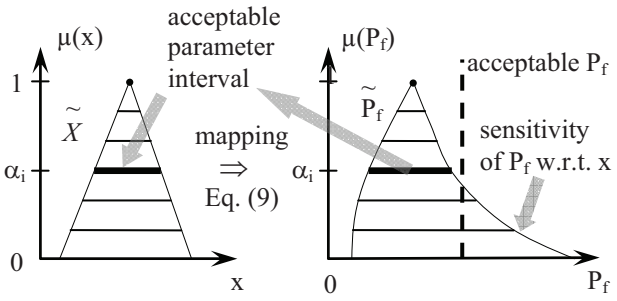


Figure 1: Relationship between fuzzy parameters and failure probability.

The membership function serves only instrumentally to summarize various plausible interval models in one embracing scheme. The interpretation of the membership value μ as epistemic possibility, which is sometimes proposed, may be useful for ranking purposes, but not for making critical decisions. The importance of fuzzy modeling lies in the simultaneous consideration of various magnitudes of imprecision at once in the same analysis.

The features of a fuzzy probabilistic analysis can be utilized to identify sensitivities of the failure probability with respect to the imprecision in the probabilistic model specification; see Figure 1. Sensitivities of P_f are indicated when the interval size of $P_{f\alpha}$ grows strongly with a moderate increase of the interval size of X_α of the parameters. If this is the case, the membership function of $\tilde{P}_f$ shows outreaching or long and flat tails. An engineering consequence would be to pay particular attention to those model options X_α, which cause large intervals $P_{f\alpha}$ and to further investigate to verify the reasoning for these options and to possibly exclude these critical cases.

A fuzzy probabilistic analysis also provides interesting features for design purposes. The analysis can be performed with coarse specifications for design parameters and for probabilistic model parameters. From the results

of this analysis, acceptable intervals for both design parameters and probabilistic model parameters can be determined directly without a repetition of the analysis; see Figure 1. Indications are provided in a quantitative manner to collect additional specific information or to apply certain design measures to reduce the input imprecision to an acceptable magnitude. This implies a limitation of imprecision to only those acceptable magnitudes and so also caters for an optimum economic effort. For example, a minimum sample size or a minimum measurement quality associated with the acceptable magnitude of imprecision can be directly identified. Further, revealed sensitivities may be taken as a trigger to change the design of the system under consideration to make it more robust. A related method is described in [5] for designing robust structures in a pure fuzzy environment. These methods can also be used for the analysis of aged and damaged structures to generate a rough first picture of the structural integrity and to indicate further detailed investigations to an economically reasonable extent–expressed in form of an acceptable magnitude of input imprecision according to some α-level.

3 Examples

3.1 Concept Demonstration: Reinforced Concrete Frame

The principle of the fuzzy probabilistic reliability analysis is illustrated by means of the reinforced concrete frame from [22] shown in Figure 2.

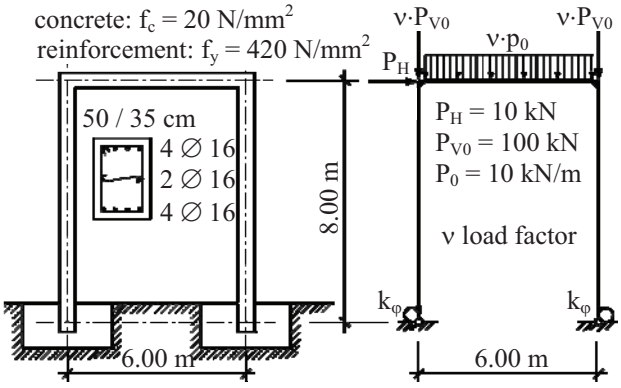


Figure 2: Reinforced concrete frame, structural model, and loading.

The structure is loaded by its dead weight, a small horizontal load P_H, and the vertical loads P_{V0} and p_0 which are increased with the factor ν until global structural failure is reached. For the purpose of demonstration, only the load factor ν is introduced as a random variable with an extreme value distribution of Ex-Max Type I (Gumbel) with mean $\tilde{m}_\nu$ and standard deviation $\tilde{\sigma}_\nu$. Imprecision of the probabilistic model is described with triangular fuzzy numbers $\tilde{m}_\nu = \langle 5.7, 5.9, 6.0 \rangle$ and $\tilde{\sigma}_\nu = \langle 0.08, 0.11, 0.12 \rangle$. In addition, the rotational stiffness of the springs at the column bases is modeled as a triangular fuzzy number $\tilde{k}_\varphi = \langle 5, 9, 13 \rangle$ MNm/rad to take account of the only vaguely known soil properties. These fuzzy parameters are considered as given for the purpose of this paper to highlight certain advantages of fuzzy probabilistic approaches in structural reliability assessment rather than to demonstrate the procedure for a specific practical case. In practical applications these fuzzy parameters need to be determined for the specific case. Although a general rule or algorithm cannot be formulated for this purpose, expert knowledge and inspection results are frequently available, which can be used together with statistical methods to determine bounds for the support of these parameters in a conservative manner. These semi-heuristic approaches can then be extended to higher α-levels in order to derive further nested intervals with an engineering meaning, e.g., to which the parameter imprecision can be reduced with certain technical efforts. Some suggestions to derive fuzzy parameters of probability distributions based on statistical data with typical characteristics as in civil engineering practice are discussed in [3]. It should be noted that the membership values are only instrumental in this approach with no specific meaning; they enable the simultaneous consideration of a variety of intervals of different size at once in the same analysis; see Section 2.

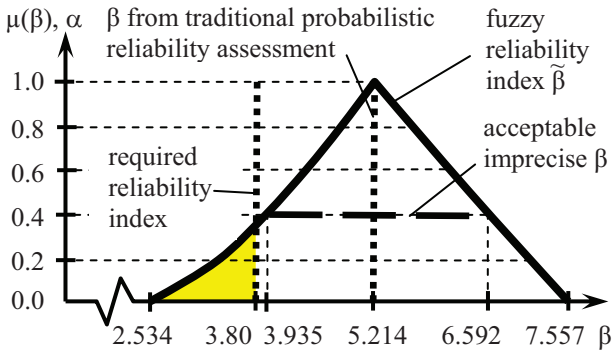


Figure 3: Fuzzy reliability index and evaluation against safety requirement.

Based on this input information, the fuzzy reliability index $\tilde{\beta}$ shown in Figure 3 is calculated. The result spreads over a large range of possible values for β. The interval bounds for each α-level are determined with the global optimization approach from [21], which is based on a modified evolution strategy. This provides advantages over a perturbation method or sensitivity investigation in view of result accuracy as the dependency between the parameters and β can be quite nonlinear, and the intervals obtained for β are quite large. The shaded part of $\tilde{\beta}$ does not comply with the safety requirements. This means that a sufficient structural reliability is not ensured when the parameters are limited to the plausible

ranges for $\alpha = 0$. In a traditional reliability analysis, using crisp assumptions for the parameters out of their plausible range such as the values associated with the membership $\mu = 1$, this critical situation is not revealed. So far, the results from p-box approach or from interval probabilities would lead to the same conclusions. As an additional feature of fuzzy probabilities, it can be observed that the left tail of the membership function of $\tilde{\beta}$ slightly tends to flatten towards small values. This indicates a slight sensitivity of β with respect to imprecision of the fuzzy input when this grows in magnitude. So one may wish to reduce the input imprecision to a magnitude which is associated with the steeper part of the membership function of β. In Figure 3, the part $\mu(\beta) \geq 0.4$ is a reasonable choice in this regard. Further, the result $\beta_{\alpha=0.4}$ = [3.935, 6.592] for $\mu(\beta) \geq 0.4 = \alpha$ (according to the definition of α-level sets) satisfies the safety requirement $\beta_{\alpha=0.4} \geq 3.8$. That is, a reduction of the imprecision of the fuzzy input parameters to the magnitude on α-level $\alpha = 0.4$ would lead to an acceptable reliability of the structure despite the remaining imprecision in the input. For example, a collection of additional information can be pursued to achieve the requirements

- $k_\varphi \in$ [6.6, 11.4] MNm/rad = $k_{\varphi,\alpha=0.4}$,
- $m_\nu \in [5.78, 5.96] = m_{\nu,\alpha=0.4}$,
- $\sigma_\nu \in [0.092, 0.116] = \sigma_{\nu,\alpha=0.4}$.

If this cannot be achieved for one or more parameters, the fuzzy analysis can be repeated with intervals for the parameters with non-reducible imprecision and with fuzzy sets for the parameters with reducible imprecision to separate the effects. The evaluation of the results then leads to a solution with proposed reduction of the imprecision only of those parameters for which this is possible. In this manner, it is also possible to explore sensitivities of the result β with respect to the imprecision of certain groups of input parameters or of individual input parameters. The repetition of the fuzzy analysis for these purposes can be avoided largely when a global optimization technique is used for the fuzzy analysis. This type of fuzzy analysis leads to a set of points distributed over the value ranges of the fuzzy input parameters and associated with results $\beta \in \tilde{\beta}$. For each construction of membership functions for the fuzzy input parameters, it is then immediately known which points belong to which α-level so that a discrete approximation of a result can be obtained directly without a repeated analysis. Repetition of the analysis is then only required for a detailed verification.

3.2 Practical Application: Offshore Structures

Reliability analysis of existing offshore structures in seawater conditions requires realistic models for corrosion. Due to scarce and imprecise information, however, the model parameters cannot be specified precisely and are merely known in form of bounds. This situation can be approached appropriately with concepts of imprecise probabilities.

3.2.1 Corrosion Model

A probabilistic model for mild steel corrosion based on results from various coupon tests and other observations is proposed in [20]. This model describes the material loss due to corrosion as a function of time; see Figure 4.

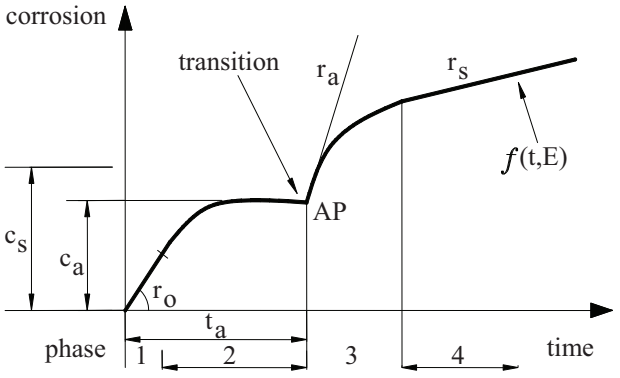


Figure 4: Corrosion model with mean value function $f(t,E)$ after [20].

Uncertainties in the corrosion process are considered with a probabilistic model for the corrosion depth $c(t,E)$, measured in mm, as

$$c(t,E) = b(t,E) \cdot f(t,E) + \varepsilon(t,E) \,, \tag{10}$$

with

- $f(t,E)$ – mean-value function,
- $b(t,E)$ – bias function,
- $\varepsilon(t,E)$ – zero-mean uncertainty function,
- E – vector of environmental (and material) parameters.

The specification of the mean-value function $f(t,E)$ requires calibration of the parameters shown in Figure 4. These parameters can be determined as a function $F(T)$ of the average seawater temperature T (contained in E),

$$\{r_0, t_a, c_a, r_a, c_s, r_s\} = F(T) \,, \tag{11}$$

see [20]. The variability of $c(t,E)$ is modeled with the zero-mean uncertainty function $\varepsilon(t,E)$ (in Eq. (10)) in the form of Gaussian white noise; $\varepsilon(t,E)$ is assumed with zero mean and a standard deviation given by

$$\sigma_\varepsilon(t,T) = (0.006 + 0.0003 \cdot T) \cdot \frac{t}{t_a} \,, \; \frac{t}{t_a} \leq 1.5 \,. \tag{12}$$

The bias function $b(t,E)$ in Eq. (10) reflects the difference of the mean value predicted by the corrosion model

and the mean values of corrosion loss derived from data. It is a function of the exposure time. Examples for bias functions based on statistical evaluations are provided in [20], see Figure 5, as functions of the non-dimensional time coordinate t/t_a with t_a as shown in Figure 4.

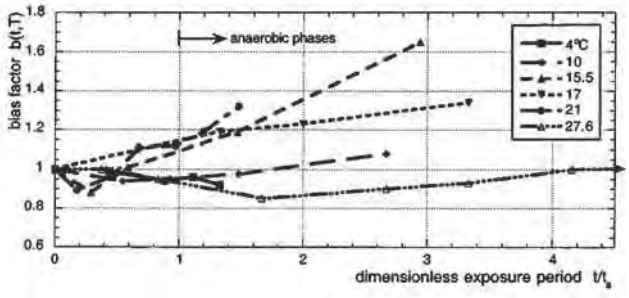


Figure 5: Bias function $b(t,E)$, dimensionless, after [20].

Before the anaerobic phases (up to the end of phase 2), the bias function lies in the range between 0.9 and 1.1. In the anaerobic phases (phases 3 and 4), the spread between the possible graphs becomes even more distinctive. A dependency between the temperature T and bias function $b(t,E)$ cannot be retrieved based on this information only. A condensation of the spread into a deterministic bias function would disregard information. On the other hand, the available information on the spread is quite sparse for the specification of a probabilistic model with sufficient confidence. A Bayesian approach would require some data for model update. If this is not available, as can be assumed for this type of data for a specific location, the model would remain subjective. Thus, one may wish to identify the worst case for the bias function $b(t,E)$ for the analysis based on the range of available information. But a simple conclusion such as "the upper bound of the bias function leads to the most critical structural behavior" may not apply. Due to the variety of members in a structural system even a uniform thickness reduction can lead to changes in kinematic failure modes. This motivates a search for the worst case under consideration of a plausible range for the bias function $b(t,E)$.

In the subsequent two examples, the uncertainty of the bias function $b(t,E)$ is accounted for with different models, and the effects on the results of a corresponding reliability analysis are investigated.

3.2.2 Steel Plate

For demonstration purposes, an example of a simple steel plate is taken from [20], and a reliability assessment is carried out under uncertain corrosion impact. The effects of different models for the uncertainty of the bias function $b(t,E)$ are investigated with respect to the failure probability P_f. The analysis is limited to the aerobic corrosion phase. It is assumed that the steel plate is exposed to seawater with a temperature of $T = 15°C$ over a period of 2.5 years.

Let d and h denote the thickness and nominal width of the uniform plate, respectively. A load is applied to cause a constant uniaxial tensile force Q in the plate. The force Q follows a normal distribution with parameters given in Table 1. It is applied at $t = 2.5$ years.

Variable	Mean	Standard deviation
Q	200 kN	23 kN
S_y	300 MPa	10 MPa
d	4 mm	0
h	250 mm	0

Table 1: Example data summary.

The resistance $R(t)$ of the plate is expressed in terms of the yield stress S_y, and the cross sectional area is reduced by the corrosion loss $c(t,E)$ on both surfaces of the plate. That is,

$$R(t) = S_y \cdot h \cdot \left[d - 2 \cdot c(t,T) \right] . \tag{13}$$

The yield stress S_y is modeled as normally distributed. The performance function is

$$G(t) = R(t) - Q . \tag{14}$$

The corrosion model is specified according to [20], which leads to a mean value $f(.) = 0.3$ mm and to a standard deviation $\sigma_\varepsilon = 0.0126$ mm for the considered $t = 2.5$ years.

The failure probability P_f is first computed with a deterministic value for the bias function, $b_{det}(.) = 1.0$. Direct Monte Carlo simulation (MCS) with a sample size of $N_{Pf} = 10^5$ leads to $P_{f,det} = 0.0126$.

The bias factor $b(.)$ is considered as merely known lying in the range between 0.9 and 1.1, which represents model uncertainty. This complies with the information provided in Figure 5. For a purely probabilistic analysis, this range is taken into account with the aid of bounded random quantities. A common probabilistic model used for those purposes in engineering is the Beta distribution with its probability density function (pdf)

$$f_X(x) = \frac{1}{B(q,r)} \frac{(x-a)^{q-1}(b-x)^{r-1}}{(b-a)^{q+r-1}} \tag{15}$$

where $B(q, r)$ is the Beta function, and the parameters a and b are the minimum and maximum value of the random variable X, respectively, with $a \le x \le b$. This model can be adjusted quite arbitrarily by means of the distribution parameters. As the available information for the modeling of the bias $b(.)$ is quite scarce, possible variants for the distribution function for $b(.)$ are considered. The following cases of parameter adjustments are investigated: Case (I): $q = r = 1$, Case (II): $q = r = 2$, and

Case (III): $q = r = 3$; see Figure 6. Case (I) represents a uniform distribution, which is frequently used when no information about the distribution is available.

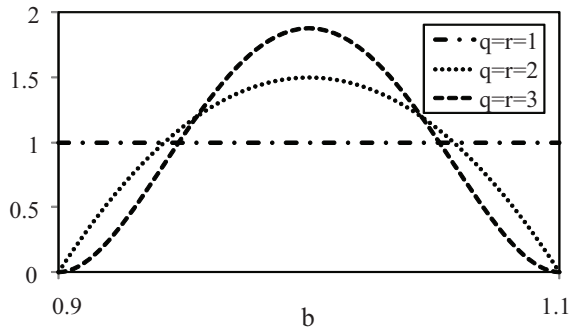


Figure 6: Variants for the pdf of the beta distribution.

The results of the subsequent reliability analysis provide extended information in comparison to the deterministic value $P_{f,det}$. To show the effects of the subjective distribution assumption on the result for P_f, a distribution for P_f is determined as dependent on the distribution of $b(.)$. An MCS is carried out for each sampling point $b(.)$ to obtain a corresponding value of $P_f(b)$, and the empirical distribution for P_f is constructed based on a sample size of $N_b = 2000$. The sample size for the determination of P_f for a given $b(.)$ is fixed at $N_{Pf} = 10^5$. The resulting plot of the distributions for the failure probability P_f in Figure 7 shows the differences between the cases considered. Since all cases represent possible models, their differences will be manifested through the distribution of P_f and their corresponding expectations $E[P_f]$ estimated in Figure 7.

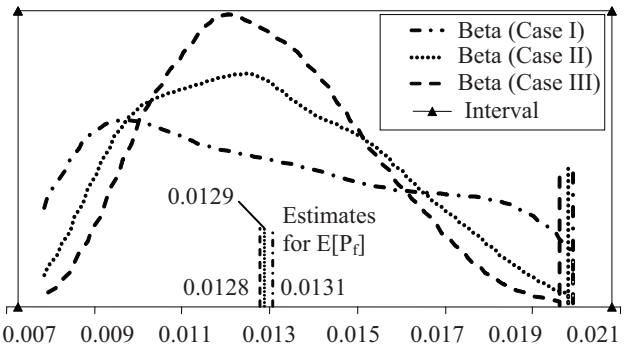


Figure 7: Failure probability, pdf's, means and upper bounds.

The modeling of $b(.)$ as a random variable involved data for various conditions and presumed variation of $b(.)$, which is reasonable for an analysis in a general context. For an analysis for a specific location, for which no data are available, one may wish to follow another approach. The bias function $b(.)$ may then be considered as given but unknown instead of showing variation. From this point of view, it is reasonable to determine the upper bound of P_f. With the stochastic parameter model, the upper bound for P_f can easily be retrieved from the sampling results shown in Figure 7, when the sampling is done conditional on $b(.)$. The results for the upper bounds in the considered cases are:

- Case (I): $P^u_{f,(I)}\left(b(.)\right) = 0.0199$,
- Case (II): $P^u_{f,(II)}\left(b(.)\right) = 0.0198$,
- Case (III): $P^u_{f,(III)}\left(b(.)\right) = 0.0196$.

The differences between these results for all three cases are quite small. The absolute values, however, are smaller than the true upper bound $P^u_{f,true}\left(b(.)\right) = 0.02082$. An improvement can be obtained by increasing the sample size N_b for $b(.)$. But a reasonable precision of $P^u_f\left(b(.)\right)$ demands a quite high numerical effort; the total number of evaluations of the limit state function is $N_b \cdot N_{Pf}$. This is hardly feasible for real structures, even when sophisticated sampling schemes are implemented.

Certainly, in a number of practical cases, including this simple example, the worst case for the imprecise parameter can be recognized in advance, so that the upper bound of P_f can be found easily. However, in a general case when the dependency between imprecise model parameters and P_f is non-monotonic, the solution is quite tedious.

A suitable approach to solve this problem is available with concepts of imprecise probabilities. The bias function is now modeled as an interval, $b_I = [0.9,1.1]$. An interval analysis is performed to map b_I to an interval for the failure probability $P_{f\_I} = \left[P^l_f\left(b(.)\right), P^u_f\left(b(.)\right)\right]$, see Eq. (3). The associated result is shown in Figure 7. This analysis is realized with the global optimization algorithm from [21]. Instead of sampling $b(.)$, a search algorithm is used to directly head for the interval bounds $P^l_f\left(b(.)\right)$ and $P^u_f\left(b(.)\right)$. Still, for each selected value $b(.) \in b_I$ an MCS needs to be carried out. The required number N_b of these simulations, however, is now significantly smaller; the exact result of the upper bound $P^u_f\left(b(.)\right)$ is approached much faster. With standard adjustments for the search algorithm, only $N_b = 45$ values of $P_f\left(b(.)\right)$ were calculated to find the true result $P^u_{f,true}\left(b(.)\right) = 0.02082$. This effort can be reduced further with an improved adjustment in the parameters of the search algorithm. The effort increases almost linearly with the number of interval input variables.

This analysis can be extended further by implementing a fuzzy probabilistic concept. This enables modeling of the bias function $b(.)$ with the aid of fuzzy sets so that a set of different intervals for $b(.)$ can be considered simultaneously. A rational approach is to assign a membership

value $\mu(b(.)) = 1.0$ to the deterministic value $b_{det}(.) = 1.0$. A reasonable interval $b_{I0}(.) = \left[b_0^l(.), b_0^u(.)\right]$ may then be specified, which is even larger than the one concluded from available information, in order to reveal effects in case that $b(.)$ takes on exceptional values. The associated membership values are assigned as $\mu\left(b_0^l(.)\right) = \mu\left(b_0^u(.)\right) = 0.0$. In the example, $b_{I0} = [0.8,1.2]$ is selected. If no further specifications for membership values are made, this leads to the fuzzy triangular number $\tilde{b}(.) = \langle 0.8, 1.0, 1.2 \rangle$ as shown in Figure 8. Of course, the interval concluded from available information should be included in the fuzzy modeling. This is provided in form of the α-level set $b_{I\alpha}(.) = b_I(.) = [0.9,1.1]$ for $\alpha = \mu(b(.)) = 0.5$; see Figure 8. The associated analysis is performed with global optimization according to [21] as a repetition of the interval analysis for various membership levels with exploitation of the nested configuration of the intervals. A fuzzy failure probability $\tilde{P}_f$ is obtained as shown in Figure 8.

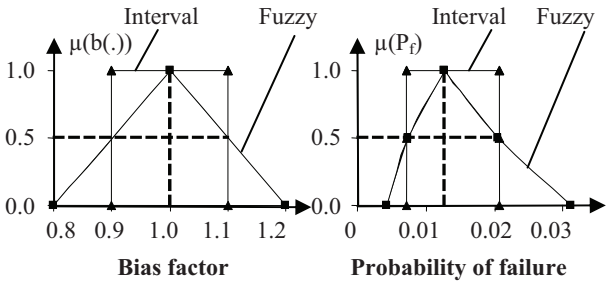


Figure 8: Fuzzy bias factor $\tilde{b}(.)$ and fuzzy failure probability $\tilde{P}_f$; interval modeling and results from Figure 7 are included for $\mu = \alpha = 0.5$.

A total of $N_b = 208$ calculations of $P_f\left(b(.)\right)$ were necessary to obtain this result. The number N_b in the fuzzy analysis is not a multiple of N_b from interval analysis according to the number of α-levels. Random elements in the optimization procedure weaken this conclusion to the statistical mean of N_b. The search domains for different α-levels are of a different and so require a different N_b. Further, the numerical procedure from [21] exploits the nested configuration of the interval to re-use all previously evaluated points inside the search domain, which leads to a significant gain in numerical efficiency for a larger number of α-levels. In the example, the significant increase of the support of the parameters in the fuzzy analysis compared to the interval possesses the governing effect, which leads to increase of N_b by a factor larger than two. But this is still a much smaller number N_b compared to a stochastic sampling of $b(.)$. Compared to interval analysis, the numerical effort is higher. But the result $\tilde{P}_f$ is much richer in information compared to P_{fI}. The fuzzy analysis contains the above interval analysis on the level $\alpha = 0.5$; see Figure 8. In addition, a series of intervals with decreasing and increasing size are analyzed, which provides information regarding sensitivities of P_{fI} with respect to the interval size of $b_I(.)$ as discussed in Sections 2 and 3.1. Again, the membership values are not of interest, they just serve as a tool in the modeling. Dependencies between the size of $b_I(.)$ and the size of P_{fI} become directly visible in the results. In the example, no particular sensitivities are obvious.

3.2.3 Offshore Platform

Deterioration of structural strength is a major factor in the safety assessment of offshore structures. The protective paints and cathodic protection may be ineffective after some years. Typically, when analyzing structural strength or structural capacity, only "uniform" corrosion is considered [20]. These issues can be addressed in an investigation as demonstrated in Section 3.2.2 applied to real structures. In the following example, a fixed offshore platform is analyzed, which is exposed to seawater with a temperature of $T = 15°C$ over a period of 5 years. All the tubular structural members beneath the seawater surface are assumed to have the same average reduction in thickness due to corrosion only on the outer side.

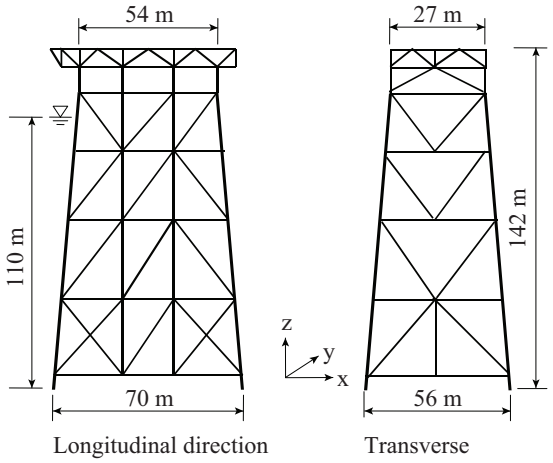


Figure 9: Structural model of the fixed jacket platform.

As an example structure, a fixed jacket platform located in the North Sea is taken from [25]. The jacket is designed for a water depth of approximately 110 m. The 8-leg jacket is arranged in a two by four rectangular grid. The overall dimensions are 27 m × 54 m at the top elevation and 56 m × 70 m at the mudline. The total height is 142 m. Horizontal bracings are installed at 5 levels. The jacket foundation consists of four corner clusters with eight skirt piles in each group and no leg piles are used. The longitudinal jacket frames are diagonal-braced, with *X*-braces between central and corner legs at the bottom bay. Transverse frames are *K*-braced, with the bottom *K* inverted to form a double *X* as shown in Figure 9.

The reliability analysis of a jacket structure involves the performance function,

$$G = \textit{Ultimate Resistance} - \textit{Environmental Loads.} \quad (16)$$

The ultimate resistance is determined through a pushover analysis of the platform. It is equal to the environmental design loads multiplied by the Reserve Strength Ratio (RSR). For this example, the environmental design loads are a 100-year wave together with a 10-year current. This is associated with a Gumbel distribution, which is implemented as a probabilistic load model in the analysis. For the structural resistance, uncertainty is considered in the yield strength of the steel and in the thickness reduction of the members due to marine corrosion. The yield strength of the steel ASTM-A7 is described with a log-normal distribution. Based on the probabilistic corrosion model discussed in Section 3.2.2, the environmental condition with $T = 15°C$ and $t = 5$ years leads to the mean value $f(.) = 0.48$ mm and the standard deviation $\sigma_\varepsilon = 0.08$ mm. The bias factor $b(t,T)$ lies in the range between 0.8 and 1.6 based on Figure 5. Implementation of these models in a structural analysis leads to the approximate performance function

$$G = \left[0.0704 \cdot F_y - 0.0887 \cdot c\left(t,T\right) - 0.0605\right] \cdot L_{100} - 0.0176 \cdot H^{2.2} \quad (17)$$

with

$$L_{100} = 0.0176 \cdot H_{100}^{\ 2.2} \ . \quad (18)$$

For the reliability analysis, the variables in Eq. (17) are described by their respective probabilistic models. These random variables are summarized in Table 2. The probability of failure is calculated as $P_f = P\left(G \le 0\right)$ via MSC. In order to calculate P_f effciently, importance sampling is utilized. A sample size of $N_{Pf} = 5000$ is used for the reliability analysis. Variants for modeling of $b(.)$ are investigated, and the results are summarized in Figure 10. Again, the interval concept shows some advantage when the bounds on the failure probability have to be found. The total number of calculations N_b of P_f using the interval concepts is 114. The accuracy of the upper bound on P_f is higher, compared to the sampling of $b(.)$.

Variable	Distribution	Parameters	
F_y	Log. Normal	$\mu = 40\ psi$	$c.o.v. = 0.087$
H	Gumbel	$\alpha_H = 21.0\ m$	$\beta_H = 1.63\ m$
$c(.)$	Normal	$\mu = 0.48\ mm$	$\sigma_\varepsilon = 0.08\ mm$

Table 2. Random variables for the reliability analysis.

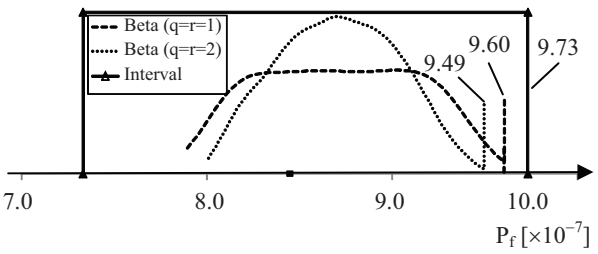


Figure 10: Failure probability; distributions, upper bounds and interval solution.

In the example, the differences in the upper bound on the failure probability are small. However, in other cases, and if more imprecision is involved in the problem, the discussed effects may become quite significant. It is obvious that the imprecision in the bias function $b(.)$ and thus, the imprecision of P_f grow dramatically with the exposure time, as can be seen in Figure 5. Further, in the example, only the annual failure probability is calculated. In a consideration of the failure probability for the entire lifetime of the structure, the imprecision in the annual failure probabilities will be accumulated accordingly. A consideration of this imprecision in a reliability analysis for the entire lifetime of an offshore structure is thus of great interest.

4 Summary and Conclusions

Different approaches were applied to describe imprecision in probabilistic models for a reliability analysis of engineering structures. The features of the models were compared with a pure probabilistic solution and with one another by means of academic and practical examples. The influence of the modeling on the prediction of structural reliability was examined. It was found that concepts of imprecise probabilities and, in particular, fuzzy probabilities, have certain advantages when bounds on the failure probability are of interest. These advantages concern the precision and the numerical effort in the calculation of these bounds and, in the case of fuzzy probabilities, some extended insight into sensitivities of the computational results with respect to the imprecision of the probabilistic input. Applicability in practice was demonstrated by means of a reliability analysis for a real offshore platform.

References

[1] Augustin, Th. and Hable, R. (2010). On the impact of robust statistics on imprecise probability models: a review, *Structural Safety*, 32(6): 358–365.

[2] Bandemer, H. and Näther, W. (1992). *Fuzzy Data Analysis*, Kluwer Academic Publishers, Dordrecht.

[3] Beer, M. (2009). Engineering Quantification of Inconsistent Information, *International Journal of Reliability and Safety*, 3(1–3), 174–200.

[4] Beer, M. (2009). Fuzzy probability theory, in: R. Meyers (Ed.), *Encyclopedia of Complexity and Systems Science*, Vol 6, Springer, New York, 4047–4059.

[5] Beer, M. and Liebscher, M. (2008). Designing robust structures – a nonlinear simulation based approach, *Computers and Structures*, 86(10), 1102–1122.

[6] Couso, I. and Dubois, D. (2009). On the variability of the concept of variance for fuzzy random variables, *IEEE Transactions on Fuzzy Systems*, 17, 1070–1080.

[7] Der Kiureghian, A. and Ditlevsen, O. (2009). Aleatory or epistemic? Does it matter?, *Structural Safety*, 31(2), 105–112.

[8] Fellin, W., Lessmann, H., Oberguggenberger, M. and Vieider, R. (eds.) (2005). *Analyzing Uncertainty in Civil Engineering*, Springer, Berlin Heidelberg New York.

[9] Ferson, S. and Tucker, W. T. (2006). Sensitivity analysis using probability bounding, *Reliability Engineering & System Safety*, 91(10–11), 1435–1442.

[10] Ferson, S. and Hajagos, J. G. (2004). Arithmetic with uncertain numbers: rigorous and (often) best possible answers, *Reliability Engineering & System Safety*, 85(1–3), 135–152.

[11] Fetz, Th. and Oberguggenberger, M. (2010). Multivariate models of uncertainty: A local random set approach, *Structural Safety*, 32(6): 417–424.

[12] Gil, M. A., López-Díaz, M. and Ralescu, D. A. (2006). Overview on the development of fuzzy random variables, *Fuzzy Sets and Systems*, 157(19), 2546–2557.

[13] Hall, J. W. (2006). Uncertainty-based sensitivity indices for imprecise probability distributions, *Reliability Engineering & System Safety*, 91(10–11), 1443–1451.

[14] Helton, J. C. and Oberkampf, W. L. (eds.) (2004). Special Issue on Alternative Representations of Epistemic Uncertainty, *Reliability Engineering and System Safety*, 85(1–3).

[15] Igusa, T., Buonopane, S. G., and Ellingwood, B. R. (2002). Bayesian analysis of uncertainty for structural engineering applications, *Structural Safety*, 24(2–4), 165–186.

[16] Klir, G.J. (2006). *Uncertainty and Information: Foundations of Generalized Information Theory*, Wiley-Interscience, Hoboken.

[17] Kozine, I. and Filimonov, Y. (2000). Imprecise reliabilities: experiences and advances, *Reliability Engineering and System Safety*, 67, 75–83.

[18] Kruse, R. and Meyer, K. (1987). *Statistics with Vague Data*, Reidel, Dordrecht.

[19] Li, S., Ogura, Y. and Kreinovich, V. (2002). *Limit Theorems and Applications of Set Valued and Fuzzy Valued Random Variables*, Kluwer Academic Publishers, Dordrecht.

[20] Melchers, R. E. (2003). Probabilistic Model for Marine Corrosion of Steel for Structural Reliability Assessment, *Journal of Structural Engineering*, 129(11), 1484–1493.

[21] Möller, B., Graf, W. and Beer, M. (2000). Fuzzy structural analysis using α-level optimization, *Computational Mechanics*, 26(6), 547–565.

[22] Möller, B., Graf, W. and Beer, M. (2003). Safety assessment of structures in view of fuzzy randomness, *Computers and Structures*, 81(15), 1567–1582.

[23] Papadimitriou, C., Beck, J. L. and Katafygiotis, L. S. (2001). Updating robust reliability structural test data, *Probabilistic Engineering Mechanics*, 16, 103–113.

[24] Schenk, C. A. and Schuëller, G. I. (2005). *Uncertainty Assessment of Large Finite Element Systems*, Springer, Berlin, Heidelberg.

[25] *USFOS, Ultimate Strength of Frame Offshore Structures 2001*, (2001). User's Manual, SINTEF.

[26] Utkin, L. V. (2004). An uncertainty model of structural reliability with imprecise parameters of probability distributions, *ZAMM - Zeitschrift für Angewandte Mathematik und Mechanik*, 84(10–11), 688–699.

[27] Viertl, R. (1996). *Statistical Methods for Non-Precise Data*, CRC Press, Boca Raton New York London Tokyo.

[28] Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*, Chapman & Hall, London.

[29] Weichselberger, K. (2000). The theory of interval-probability as a unifying concept for uncertainty, *International Journal Approximate Reasoning*, 24(2–3), 149–170.

[30] Yamauchi, Y. and Mukaidono, M. (1999). Interval and paired probabilities for treating uncertain events, *IEICE Transactions on Information and Systems*, E82_D(5), 955–961.

[31] Zimmermann, H. J. (1992). *Fuzzy Set Theory and Its Applications*, Kluwer Academic Publishers, Boston London.

Two for the Price of One: Info-Gap Robustness of the 1-Test Algorithm

Yakov Ben-Haim
Yitzhak Moda'i Chair in Technology and Economics
Faculty of Mechanical Engineering
Technion—Israel Institute of Technology
Haifa 32000 Israel
http://www.technion.ac.il/yakov
yakov@technion.ac.il

Abstract

Analysts in many domains must choose a design, a strategy, or an intervention without being able to test all relevant alternatives. We consider a situation in which one of two alternatives must be chosen, while only one alternative can be tested prior to decision. The probability of success from blind choice is 1/2. The probability of success if the distribution of the system attributes is known is 3/4. The 1-test algorithm assures probability greater than 1/2 of choosing the better system based on a single test, even without knowing the probability distribution of the system attributes. If the distribution is poorly known, then info-gap theory can robustify the 1-test algorithm. Using the info-gap robustness function we show that robust-satisficing algorithms may differ from the nominally optimal algorithm when the attribute distribution is uncertain.

Keywords. Testing, design, info-gap.

1 The 1-Test Algorithm

Consider a choice between two design concepts for a technological system. We would like to choose the system with higher reliability (or longer life or lower mean time between failure, etc.). It may be very expensive to construct and test both physical systems. It would be useful if the better system could be reliably chosen based on testing only one system.

Consider the choice between two medical interventions for a specific patient (or macro-economic interventions for a specific economy, or biological interventions in an ecosystem). We can do one or the other, but not both. Given all available information, we are epistemically indifferent between the interventions: we have no reason to believe that one intervention is better than the other, though they are different. We choose one intervention by flipping a fair coin, and we observe the result (reduction in fever, or increase in blood count, etc.). For future reference we would like to know which of the two would have been better.

Decisions such as these can be thought about generically as follows.

Two systems each have a real-valued attribute (e.g. lifetime, reliability, etc.). We would like to choose the system with the larger—better—value, but we are able to measure the attribute of only one system. We must decide if the measured attribute is the smaller or the larger of the two, where we have chosen the system to test by a throw of a fair coin. We know nothing about the distribution of the attribute values, other than that they can take any value in a specified interval.

The 1-test algorithm is stated without proof by Cover [2] and proven by Snapp [6]. The idea is also discussed in a blog [7]. We can formalize it as follows.

Two different real numbers, x_1 and x_2, are chosen by an algorithm unknown to you. One of these numbers, call it x_r, is revealed to you, where you know that the probability that $x_\mathrm{r} = x_1$ is 0.5. You must decide if x_r is the smaller or the larger of the two numbers.

The **1-test algorithm** for deciding whether x_r is the smaller or larger of the two values is as follows. Let $q(y)$ be a non-atomic probability density function (pdf) which is positive on an interval containing x_1 and x_2. The interval may be finite, half-finite, or infinite. We will refer to $q(y)$ as the "decision pdf". Decide according to the following decision rule:

1. Draw a random number, y, distributed according to $q(y)$.
2. If $y \geq x_\mathrm{r}$ then decide that x_r is the smaller of the two x_i.
3. If $y < x_\mathrm{r}$ then decide that x_r is the larger of the two x_i.

The 1-test algorithm succeeds if the number chosen by the algorithm is in fact the larger of the two numbers.

Let $P_s(x_1, x_2, q)$ denote the probability of success of the 1-test algorithm using a pdf $q(y)$ applied to real numbers x_1 and x_2. We will prove the following theorem. See Cover [2] and Snapp [6].

Theorem 1 *The probability of success of the 1-test algorithm exceeds 1/2.*

Given:

- *The two numbers, x_1 and x_2, are different.*
- *$q(y)$ is a non-atomic probability density function which is non-zero on an interval containing x_1 and x_2. $q(y)$ is zero outside this interval.*

Then:

$$P_s(x_1, x_2, q) > \frac{1}{2} \tag{1}$$

Proof of theorem 1. The two numbers are different, so one is larger. Denote the larger of the two numbers by x_1, where x_r is the number which has been revealed. Our information is:

$$\text{Prob}(x_r = x_1) = \text{Prob}(x_r = x_2) = 0.5 \tag{2}$$

If x_r is the larger of the two numbers, then the probability of success equals the probability that $y < x_r$:

$$P_s(x_r = x_1) = \int_{-\infty}^{x_1} q(y)\,dy = Q(x_1) \tag{3}$$

where $Q(y)$ is the cumulative distribution function of $q(y)$. Similarly, if x_r is the smaller of the two numbers, then the probability of success equals the probability that $y \geq x_r$:

$$P_s(x_r = x_2) = \int_{x_2}^{\infty} q(y)\,dy = 1 - Q(x_2) \tag{4}$$

$P_s(x_r = x_1)$ and $P_s(x_r = x_2)$ are illustrated in fig. 1.

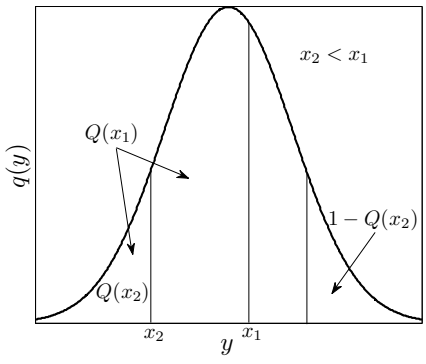


Figure 1: $Q(x_1)$ and $1 - Q(x_2)$ for $x_2 < x_1$; eqs.(3) and (4).

Recall that $x_1 > x_2$ which, since $q(y)$ is non-zero on an interval containing x_1 and x_2, implies:

$$Q(x_1) > Q(x_2) \tag{5}$$

Thus the total probability of success, with the q-based decision algorithm, is:

$$\begin{aligned} & P_s(x_1, x_2, q) \\ &= \text{Prob}(x_r = x_1)P_s(x_r = x_1) \\ &\quad + \text{Prob}(x_r = x_2)P_s(x_r = x_2) && (6) \\ &= 0.5Q(x_1) + 0.5[1 - Q(x_2)] && (7) \\ &= 0.5[1 + \underbrace{Q(x_1) - Q(x_2)}_{>0}] > 0.5 && (8) \end{aligned}$$

which completes the proof. ∎

2 Info-Gap Robustness of the 1-Test Algorithm

The system attributes, x_1 and x_2, are random variables. Let $p(x_1, x_2)$ denote their joint pdf. If we knew this distribution we could choose the 1-test decision distribution, $q(y)$, to maximize the probability of success. But suppose we only have a guess or plausible supposition of the joint pdf of x_1 and x_2. That is, we think they are drawn from a joint pdf which is something like $\widetilde{p}(x_1, x_2)$, but the true distribution may have a different shape or different moments. How should we choose $q(y)$?

In this section we introduce info-gap models to represent non-probabilistic uncertainty about the true pdf of x_1 and x_2. We then define the info-gap robustness function and illustrate its use in selecting the decision pdf $q(y)$.

2.1 Info-Gap Uncertainty and Robustness

An info-gap model [1], [4] is a family of nested sets, $\mathcal{U}(h, \widetilde{p})$, $h \geq 0$. The elements of these sets are realizations of the uncertain quantity, which is the joint pdf of x_1 and x_2 in the present case. The set-valued functions, $\mathcal{U}(h, \widetilde{p})$, of an info-gap model, have the following properties:

$$\text{Contraction:} \quad \mathcal{U}(0, \widetilde{p}) = \{\widetilde{p}\} \tag{9}$$

$$\text{Nesting:} \quad h < h' \text{ implies } \mathcal{U}(h, \widetilde{p}) \subseteq \mathcal{U}(h', \widetilde{p}) \tag{10}$$

Contraction states that, in the absence of uncertainty, only a single function—our estimate—applies, so the uncertainty set is a singleton. *Nesting* is the property that the sets become more inclusive as the horizon of uncertainty grows. An info-gap model is a non-probabilistic quantification of uncertainty. It entails no assumptions about probability distributions or about worst cases.

For instance, consider a situation where evidence supports a symmetric pdf $\widetilde{p}(x)$ for $|x| \leq d$, but no evidence is available on the far tails, $|x| > d$, and fat

tails are suspected. A simple info-gap model for this situation is:

$$\mathcal{U}(h,\widetilde{p}) = \Big\{\, p(x)\in\mathcal{P}:\ p(x)=\nu\widetilde{p}(x),\ |x|\le d \\ p(x)\le \frac{h}{x^2},\ |x|>d \,\Big\},\quad h\ge 0 \qquad (11)$$

where $\mathcal{P}$ is the set of non-negative and normalized pdfs.

The generic properties of an info-gap model are eqs.(9) and (10), for which eq.(11) is an example. An info-gap model can be a Lévy neighborhood or a contamination neighbor, as treated by Huber [3], but need not be as illustrated by eq.(11) and in [1] and [4].

2.2 Info-Gap Robustness

The unknown joint pdf of x_1 and x_2 is $p(x_1,x_2)$ where we will assume that the variables x_1 and x_2 are exchangeable: $p(x_1,x_2)=p(x_2,x_1)$. x_1 and x_2 are also exchangeable in the estimated joint pdf, $\widetilde{p}(x_1,x_2)$. $\mathcal{U}(h,\widetilde{p})$ is an info-gap model for uncertainty in $p(x)$.

Let $P_{\rm s}(p,q)$ denote the overall probability of success, regardless of the realizations of x_1 and x_2, based on the 1-test algorithm with decision pdf $q(y)$:

$$P_{\rm s}(p,q) = 2\int_{-\infty}^{\infty}\int_{x_2}^{\infty} P_{\rm s}(x_1,x_2,q)p(x_1,x_2)\,{\rm d}x_1\,{\rm d}x_2 \qquad (12)$$

In the double integral itself (without the factor 2) we assume that x_1 is greater than x_2. Multiplying by 2 accounts for the other possibility.

We aspire to choose $q(y)$ so that $P_{\rm s}(p,q)$ is no less than a "critical value", $P_{\rm c}$. We know from theorem 1 and eq.(12) that $P_{\rm s}(p,q)$ exceeds 0.5; we might aspire to exceed 0.6 or 0.7. The robustness of any choice of $q(y)$, given aspiration $P_{\rm c}$, is the greatest horizon of uncertainty in the true distribution of x_1 and x_2, up to which all distributions result in probability of success no less than $P_{\rm c}$. Large robustness implies that our estimate, $\widetilde{p}(x_1,x_2)$, can err greatly and the 1-test algorithm with $q(y)$ will still achieve a probability of success no less than $P_{\rm c}$. Small robustness implies high vulnerability to error in the estimate. Clearly, the robustness function $\widehat{h}(q,P_{\rm c})$ establishes preferences on the decision pdfs $q(y)$.

Mathematically, we define the robustness of a decision pdf, $q(y)$, as the greatest horizon of uncertainty, h, up to which the probability of success is no less than the critical value, $P_{\rm c}$, for all possible pdf's at that horizon of uncertainty:

$$\widehat{h}(q,P_{\rm c}) = \max\left\{h:\ \left(\min_{p\in\mathcal{U}(h,\widetilde{p})} P_{\rm s}(p,q)\right)\ge P_{\rm c}\right\} \qquad (13)$$

The robustness, $\widehat{h}(q,P_{\rm c})$, is the least upper bound of the set of h values which satisfice the probability of success at its critical value. We define the robustness to equal zero if the set of h values in eq.(13) is empty.

The info-gap robustness in eq.(13) is different in several respects from the concepts of robustness in robust statistics ([3], section 1.4). First of all, the info-gap model need not represent uncertainty with the neighborhoods usually treated in robust statistics, as mentioned at the end of section 2.1 and illustrated in eq.(11). Eq.(13) does not consider the bias or variance of a statistic, nor the asymptotic (large sample) properties of any statistic, nor does it assume that the statistic is consistent in the sense of converging (in probability) to an asymptotic value. For further discussion of the relation between robust statistics and info-gap robustness see [5].

2.3 Simple Example

We now examine a very simple special case. We know that x_1 and x_2 are chosen independently from an exponential distribution, $p(x)=\lambda{\rm e}^{-\lambda x}$, $x\ge 0$. Our best guess of the coefficient is $\widetilde{\lambda}$ but this guess is very uncertain. We use a fractional-error info-gap model for uncertainty in the exponential coefficient of the pdf by which the x_i are chosen:

$$\mathcal{U}(h,\widetilde{p}) = \Big\{p(x)=\lambda{\rm e}^{-\lambda x}:\ (1-h)^{+}\widetilde{\lambda}\le\lambda\le(1+h)\widetilde{\lambda}\ \Big\} \\ h\ge 0 \qquad (14)$$

where $x^{+}=x$ if $x\ge 0$ and equals zero otherwise. Furthermore, assume that the pdf used for deciding is also exponential: $q(y)=\gamma{\rm e}^{-\gamma y}$. We will derive the robustness function (actually, its inverse) and study the choice of γ.

Let $P_{\rm s}(\lambda,\gamma)$ denote the overall probability of success, eq.(12), when the true distribution is exponential with coefficient λ and the decision pdf is exponential with coefficient γ. One finds:

$$P_{\rm s}(\lambda,\gamma) = \frac{1}{2}+\frac{\lambda\gamma}{(\lambda+\gamma)(2\lambda+\gamma)} \qquad (15)$$

$$= \frac{1}{2}+\frac{\rho}{(1+\rho)(1+2\rho)},\quad \rho=\frac{\lambda}{\gamma} \qquad (16)$$

Differentiating we find:

$$\frac{\partial P_{\rm s}(\lambda,\gamma)}{\partial\lambda} = \frac{\gamma(\gamma^2-2\lambda^2)}{(\lambda+\gamma)^2(2\lambda+\gamma)^2} \qquad (17)$$

$P_s(\lambda, \gamma)$ vs. λ is a unimodal function with a maximum at $\lambda = \gamma/\sqrt{2}$, as illustrated in fig. 2.

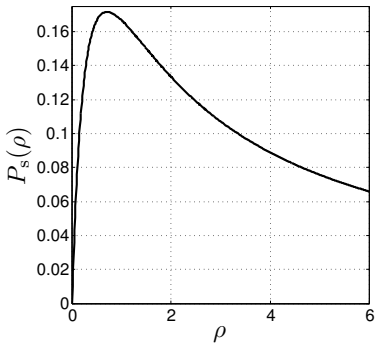


Figure 2: $P_s(\rho)$ defined in eq.(16).

Let $\mu(h, \gamma)$ denote the inner minimum in the definition of the robustness, eq.(13), which is the minimum of $P_s(\lambda, \gamma)$ as λ varies up to horizon of uncertainty h. $\mu(h, \gamma)$ is the inverse of $\widehat{h}(q, P_c)$. That is:

$$\mu(h, \gamma) = P_c \quad \text{implies} \quad \widehat{h}(q, P_c) = h \tag{18}$$

A plot of $\mu(h, \gamma)$ vs. h is the same as a plot of P_c vs. $\widehat{h}(q, P_c)$.

The minimum of $P_s(\lambda, \gamma)$, at horizon of uncertainty h, occurs when λ takes one or the other of its extreme values, which are:

$$\lambda_1(h) = (1 + h)\widetilde{\lambda} \tag{19}$$
$$\lambda_2(h) = (1 - h)^+\widetilde{\lambda} \tag{20}$$

Let us define the following two functions:

$$\mu_1(h, \gamma) = P_s[(1 + h)\widetilde{\lambda}, \gamma] \tag{21}$$
$$\mu_2(h, \gamma) = P_s[(1 - h)^+\widetilde{\lambda}, \gamma] \tag{22}$$

The inner minimum in the definition of the robustness is the lesser of these two functions:

$$\mu(h, \gamma) = \min_i \mu_i(h, \gamma) \tag{23}$$

The nominal optimal choice of γ is the value which maximizes the estimated function $P_s(\widetilde{\lambda}, \gamma)$:

$$\gamma^\star = \arg\max_\gamma P_s(\widetilde{\lambda}, \gamma) \tag{24}$$

We find $\gamma^\star$ by differentiating $P_s(\widetilde{\lambda}, \gamma)$:

$$\frac{\partial P_s(\widetilde{\lambda}, \gamma)}{\partial \gamma} = \frac{\widetilde{\lambda}(2\widetilde{\lambda}^2 - \gamma^2)}{(\widetilde{\lambda} + \gamma)^2(2\widetilde{\lambda} + \gamma)^2} \tag{25}$$

Thus we see that the nominal optimal choice of γ is:

$$\gamma^\star = \widetilde{\lambda}\sqrt{2} \tag{26}$$

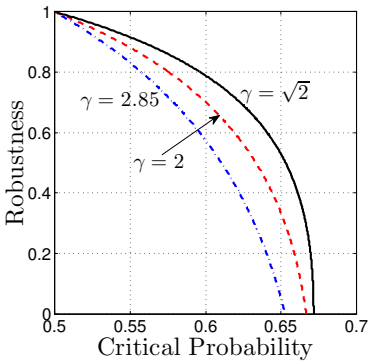


Figure 3: 3 robustness curves.

Note that the nominal optimal decision pdf, $q(y|\gamma^\star)$, differs from the estimated generating pdf, $\widetilde{p}(x|\widetilde{\lambda})$, even if the estimate is correct.

Figs. 3–5 show robustness curves, $\widehat{h}(q, P_c)$ vs P_c, for different choices of γ, which determines the decision pdf, $q(y)$. The estimated value of λ, the coefficient of the estimated distribution of x_i, is $\widetilde{\lambda} = 1$ in all cases.

The curves all converge, at the upper left, at $\widehat{h} = 1$ when $P_c = 1/2$. We understand this from eq.(15), where $P_s = 1/2$ when $\lambda = 0$.

In fig. 3 we examine values of γ for which $\mu(h, \gamma)$ in eq.(23) takes only one functional form—$\mu_1(h, \gamma)$—for all horizons of uncertainty, so no kink occurs in the curve. The peak of $P_s(\lambda, \gamma)$ vs. λ (see fig. 2 or eq.(17)) occurs when $\lambda = \gamma/\sqrt{2}$. When $\gamma = \sqrt{2}$ (solid black curve) then, since $\widetilde{\lambda} = 1$, the value of $\mu(0, \gamma)$ occurs at the peak of $P_s(\lambda, \gamma)$ vs. λ. As h increases, the value of $\mu(h, \gamma)$ moves left, down the steep positive slope illustrated in fig. 2. In the other curves of fig. 3, $\widetilde{\lambda} < \gamma/\sqrt{2}$ so the value of $\mu(0, \gamma)$ occurs on the steep positive slope of $P_s(\lambda, \gamma)$ vs λ and, as h increases, the value of $\mu(h, \gamma)$ moves left, down the steep positive slope.

From eq.(26) we see that $\gamma = \sqrt{2}$ is the nominal optimal choice since $\widetilde{\lambda} = 1$. Fig. 3 indicates that this choice is robust-dominant among the values of γ which are shown, and it is clear that this will hold for any value of γ for which $\widetilde{\lambda} \le \gamma/\sqrt{2}$.

Fig. 4 is different from fig. 3: each robustness curve in fig. 4 displays a kink when $\mu(h, \gamma)$ switches from one solution to the other as specified in eq.(23). $\widetilde{\lambda} > \gamma/\sqrt{2}$ in both cases, so $\mu(0, \gamma)$ occurs on the gentle negative-slope portion of $P_s(\lambda, \gamma)$ vs. λ. Thus, for small h, $\mu(h, \gamma)$ moves to the right down the gentle slope. However, at larger h, the value of $(1 - h)^+\widetilde{\lambda}$ occurs on the steep positive slope to the left of the peak, and now $\mu(h, \gamma)$ switches and moves left down the steep

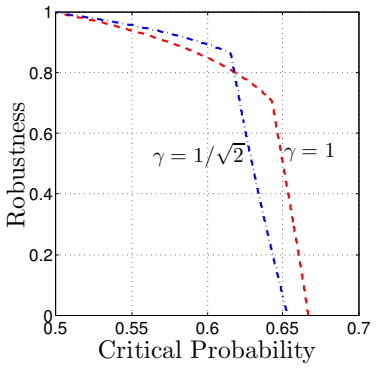


Figure 4: 2 robustness curves.

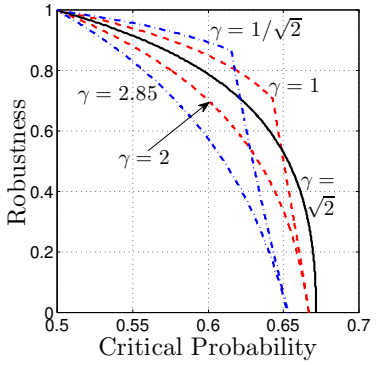


Figure 5: Figs. 3 and 4 combined.

slope. This explains the kink in the robustness curves.

Fig. 5 combines the curves of figs. 3 and 4. What is of particular interest is the intersection between the robustness curves. For instance, the curve for $\gamma = 1$ intersects the curve for $\gamma = \sqrt{2}$ at critical probability $P_c = 0.65$. For greater critical probability, $\gamma = \sqrt{2}$ is more robust (up to $P_c = 0.67$ at which its robustness vanishes). For lower probability, $\gamma = 1$ is more robust. This intersection between robustness curves entails the possibility of reversal of preferences between the corresponding choices of γ (which determines the decision pdf, $q(y)$).

3 Three Properties

We now discuss three generic properties of info-gap robustness curves—trade off, zeroing, and preference reversal—which are illustrated in the example.

Trade off between robustness and performance. Robustness curves, such as in figs. 3–5, are always monotonic, which expresses a trade off between robustness and performance: good performance entails low robustness against uncertainty. In our example, aspiring to high probability of success, P_c, entails low robustness against uncertainty in the generating pdf. This trade off is universal and results from the nesting property of info-gap models, eq.(10). The robustness function, $\widehat{h}(q, P_c)$, quantifies this trade off.

Zeroing of the robustness curve. The robustness, $\widehat{h}(q, P_c)$, will equal zero for some critical value P_c. This value is precisely the estimated performance. Using the notation of our example, the zeroing property is:

$$\widehat{h}(q, P_c) = 0 \quad \text{if} \quad P_c = P_s(\widetilde{p}, q) \tag{27}$$

This means that the robustness is zero when aspiring to a probability of success which equals the estimated probability of success. Estimated outcomes have no robustness against errors in the models that underlie the estimate. Combining this with the trade off property we conclude that only outcomes which are worse than the estimated outcome have positive robustness. This has an important implication for decision under uncertainty. Estimated outcomes are not a good basis for choosing between options because estimated outcomes have no robustness to error in the models and data underlying the estimates.

Preference reversal between options. This paper is based on the idea that more robustness against uncertainty is better than less robustness. This provides a prioritization of options—decision pdf's $q(y)$ in our example—as explained following eq.(12). Figs. 4 and 5 show several examples of intersection between robustness curves for different choices of $q(y)$. For instance, fig. 5 shows crossing between the robustness curves for the nominal optimum ($\gamma = \sqrt{2}$) and a different option ($\gamma = 1$). The former option is preferred if one requires $P_c > 0.65$, while the latter option is preferred if lower probability of success is acceptable. In short, the crossing of robustness curves entails the possibility of reversal of preference between the corresponding options.

4 Extensions of the 1-Test Algorithm

Theorem 1 and the associated decision algorithm relate to selecting a single system from two candidates based on testing only one system. We now consider three candidate systems, where either one or two systems are tested. When testing one system our aim is to select the best of the three systems. When testing two systems our aim is to select the two best systems. We prove two extensions of theorem 1, relating to these two cases, and we propose an hypothesis for more than 3 systems.

4.1 Two Tests, 3 Systems

Consider three systems, each characterized by a single real number, x_i, and assume these numbers are different. Without loss of generality we denote these numbers:

$$x_1 < x_2 < x_3 \tag{28}$$

Two of the systems are tested to reveal their attributes, x_i, where each system has the same probability of being tested. The revealed attributes are:

$$r_1 < r_2 \tag{29}$$

Let s denote the third, unrevealed, number.

Our goal is to select the two best systems, whose attributes are larger than of the third system. We do not need to identify the better of the two best; only to exclude the worst system.

The **2-test 3-system algorithm** is as follows. Let $q(y)$ be a non-atomic pdf which is positive on an interval containing x_1, x_2 and x_3. The interval may be finite, half-finite, or infinite. Select two systems according to the following decision rule:

1. Draw a random number, y, distributed according to $q(y)$.
2. If $y < r_1$, choose the two tested systems.
3. If $r_1 \le y \le r_2$, choose the systems corresponding to r_2 and s.
4. If $r_2 < y$, choose the systems corresponding to r_2 and s.

The probability of blindly choosing the two best systems is 1/3. The following theorem asserts that the above decision algorithm successfully chooses the two best systems with probability strictly exceeding 1/3.

Theorem 2 *The probability of success of the 2-test 3-system algorithm exceeds 1/3.*

Given:

- *The three numbers, x_1, x_2 and x_3, are different.*
- *$q(y)$ is a non-atomic pdf which is positive on an interval containing x_1, x_2 and x_3.*
- *Each system has equal probability of being selected for testing.*

Then:

$$P_{\rm s}(x_1, x_2, x_3, q) > \frac{1}{3} \tag{30}$$

Proof of theorem 2. The three numbers are different, so they can be denoted as in eq.(28). The two revealed numbers are therefore also different and denoted as in eq.(29). Let $R = \{r_1, r_2\}$ denote the set of revealed values. Let s denote the third, unrevealed, number. Let $Q(\cdot)$ denote the cumulative probability distribution of $q(\cdot)$. Since the tested systems are selected with equal probability we can assert:

$$\text{Prob}(s = x_1) = \text{Prob}(s = x_2) = \text{Prob}(s = x_3) = \frac{1}{3} \tag{31}$$

The decision algorithm succeeds at step 2 if $R = \{x_2, x_3\}$ whose probability is 1/3.

The decision algorithm succeeds at step 3 if $R = \{x_1, x_2\}$ or if $R = \{x_1, x_3\}$, each of whose probabilities is 1/3.

The decision algorithm succeeds at step 4 if $R = \{x_1, x_2\}$ or if $R = \{x_1, x_3\}$, each of whose probabilities is 1/3.

Putting this together we can write the total probability of success of the decision algorithm as:

$$\begin{aligned} P_{\rm s}(x_1, x_2, x_3, q) &= \underbrace{\frac{1}{3}\int_{-\infty}^{x_2} q(y)\,{\rm d}y}_{\text{step 2}} \\ &\quad + \underbrace{\frac{1}{3}\int_{x_1}^{x_2} q(y)\,{\rm d}y + \frac{1}{3}\int_{x_1}^{x_3} q(y)\,{\rm d}y}_{\text{step 3}} \\ &\quad + \underbrace{\frac{1}{3}\int_{x_2}^{\infty} q(y)\,{\rm d}y + \frac{1}{3}\int_{x_3}^{\infty} q(y)\,{\rm d}y}_{\text{step 4}} \end{aligned} \tag{32}$$

$$\begin{aligned} &= \underbrace{\frac{1}{3}Q(x_2)}_{\text{step 2}} \\ &\quad + \underbrace{\frac{1}{3}\left[Q(x_2) - Q(x_1)\right] + \frac{1}{3}\left[Q(x_3) - Q(x_1)\right]}_{\text{step 3}} \\ &\quad + \underbrace{\frac{1}{3}\left[1 - Q(x_2)\right] + \frac{1}{3}\left[1 - Q(x_3)\right]}_{\text{step 4}} \\ &= \frac{2}{3} - \frac{1}{3}\underbrace{Q(x_1)}_{<1} + \frac{1}{3}\underbrace{\left[Q(x_2) - Q(x_1)\right]}_{>0} > \frac{1}{3} \end{aligned} \tag{33}$$

$Q(x_1) < 1$ and $Q(x_2) > Q(x_1)$ because $x_1 < x_2$ and $q(y)$ is positive on an interval containing x_1, x_2 and x_3. This completes the proof. ∎

4.2 One Test, 3 Systems

Consider three systems, each characterized by a single real number, x_i, and assume these numbers are different. Without loss of generality we denote these numbers as in eq.(28). One of the systems is tested to reveal its attribute, r, where each system has the same probability of being tested. t denote the unrevealed numbers.

Our goal is to select the best system, whose attribute is larger than of the other two systems.

The **1-test 3-system algorithm** is as follows. Let $q(y)$ be a non-atomic pdf which is positive on an interval containing x_1, x_2 and x_3. The interval may be finite, half-finite, or infinite. Select a system according to the following decision rule:

1. Draw a random number, y, distributed according to $q(y)$.
2. If $y \leq r$, choose the tested system.
3. If $y > r$, choose between the untested systems with equal probability.

The probability of blindly choosing the best system is 1/3. The following theorem asserts that the above decision algorithm successfully chooses the best system with probability strictly exceeding 1/3.

Theorem 3 *The probability of success of the 3-system 1-test algorithm exceeds 1/3.*

Given:

• *The three numbers, x_1, x_2 and x_3, are different.*

• *$q(y)$ is a non-atomic pdf which is positive on an interval containing x_1, x_2 and x_3.*

• *Each system has equal probability of being selected for testing.*

Then:

$$P_{\mathrm{s}}(x_1, x_2, x_3, q) > \frac{1}{3} \tag{34}$$

Proof of theorem 3. The three numbers are different, so they can be denoted as in eq.(28). Let r denote the revealed value. Let s and t denote the unrevealed numbers. Let $Q(\cdot)$ denote the cumulative probability distribution of $q(\cdot)$. We can assert:

$$\mathrm{Prob}(r = x_1) = \mathrm{Prob}(r = x_2) = \mathrm{Prob}(r = x_3) = 1/3 \tag{35}$$

The decision algorithm succeeds at step 2 if $r = x_3$ (with probability 1/3).

The decision algorithm succeeds at step 3 if the choice between s and t is correct (with probability 0.5), and if either $r = x_1$ or $r = x_2$ (each with probability is 1/3).

Putting this together we can write the total probability of success of the decision algorithm as:

$$P_{\mathrm{s}}(x_1, x_2, x_3, q) = \underbrace{\frac{1}{3}\int_{-\infty}^{x_3} q(y)\,\mathrm{d}y}_{\text{step 2}} + \underbrace{\frac{1}{2}\frac{1}{3}\left[\int_{x_1}^{\infty} q(y)\,\mathrm{d}y + \int_{x_2}^{\infty} q(y)\,\mathrm{d}y\right]}_{\text{step 3}} \tag{36}$$

$$= \underbrace{\frac{1}{3}Q(x_3)}_{\text{step 2}} + \underbrace{\frac{1}{6}\left[(1 - Q(x_1)) + (1 - Q(x_2))\right]}_{\text{step 3}} \tag{37}$$

$$= \frac{1}{3} + \frac{1}{6}[\underbrace{Q(x_3) - Q(x_1)}_{>0}] + \frac{1}{6}[\underbrace{Q(x_3) - Q(x_2)}_{>0}] \tag{38}$$

$$> \frac{1}{3} \tag{39}$$

which completes the proof. ∎

4.3 m Tests, n Systems

Consider n systems, each characterized by a single real number, x_i, and assume these numbers are different. Without loss of generality we denote these numbers:

$$x_1 < x_2 < \cdots < x_n \tag{40}$$

m of the systems are tested to reveal their attributes, x_i, where each system has the same probability of being tested. The revealed attributes are:

$$r_1 < r_2 < \cdots < r_m \tag{41}$$

Let $R = \{r_1, \ldots, r_m\}$ denote the set of revealed values. Let R_j denote the set R after removing the j smallest elements: $R_j = \{r_{j+1}, \ldots, r_m\}$, for $j = 0, \ldots, m$. Thus $R_0 = R$ and $R_m = \emptyset$. Define $r_0 = -\infty$ and $r_{m+1} = \infty$.

Our goal is to select the m best systems, whose attributes are larger than all the remaining systems. We do not need to identify the values of these m best systems; only to exclude the $n - m$ worst systems.

The m-test n-system algorithm takes a slightly different form depending on whether or not the number of tested systems, m, is less than the number of untested systems, $n - m$. If $m \leq n - m$ then the best m systems may be entirely in the untested set. If $m > n-m$ then at least some tested systems are among the best m systems. We specify these two realizations of the decision algorithm separately.

Let $q(y)$ be a non-atomic pdf which is positive on an interval containing x_1, x_2 and x_3. The interval may be finite, half-finite, or infinite.

If $m \leq n-m$, the m-test n-system algorithm is as follows. Select m systems according to the following decision rule:

1. Draw a random number, y, distributed according to $q(y)$.

2. For $j = 0, \ldots, m$, if $r_j \leq y < r_{j+1}$, choose the systems corresponding to R_j and choose j untested systems equi-probably from among all untested systems.

If $m > n-m$, the m-test n-system algorithm is as follows. Select m systems according to the following decision rule:

1. Draw a random number, y, distributed according to $q(y)$.
2. For $j = 0, \ldots, n-m$, if $r_j \leq y < r_{j+1}$, choose the systems corresponding to R_j and choose j untested systems equi-probably from among all untested systems.
3. For $j = n-m+1, \ldots, m$, if $r_j \leq y < r_{j+1}$, choose the systems corresponding to R_{n-m} and choose all $n-m$ untested systems.

The number of distinct subsets of m from among the n systems is the binomial coefficient $\binom{n}{m}$, which we denote γ_{nm}. Only one of these subsets contains the m best systems. Thus the probability of blindly choosing the m best systems is $1/\gamma_{nm}$. We hypothesize that one could prove, in analogy to theorems 1–3, that the above decision algorithm chooses the m best systems with probability strictly exceeding $1/\gamma_{nm}$.

5 Further Questions

The 1- and 2-test algorithms can probably be further generalized in various ways. Likewise, the info-gap analysis can be realized in many different forms, especially by using different info-gap models to represent different types of prior information about the uncertain generating pdf. Many questions remain to be explored. We mention a few possible extensions of our results.

(1) In some situations the systems are evaluated by multiple criteria, not by only one attribute as we have done. (2) One might consider adaptive testing, wherein intermediate results indicate whether or not to continue testing. (3) One would like to know what is the best possible probability of success.

References

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

[2] Cover, Thomas M., 1987, Pick the largest number, chapter 5.1 in T. Cover and B. Gopinath, 1987, *Open Problems in Communication and Computation,* Springer-Verlag, Berlin.

[3] Huber, Peter J., 1981, *Robust Statistics,* Wiley, New York.

[4] Info-gap decision theory, http://info-gap.com.

[5] Keren, Carmit, 2009, *Info Gap Bayesian Classification,* M.Sc. thesis, Technion-Israel Institute of Technology (in English).

[6] Snapp, Robert R., 2005, Tom Covers Number Guessing Game, http://www.cems.uvm.edu/~snapp/teaching/coversproblem.pdf

[7] xkcd, http://blog.xkcd.com/2010/02/09/math-puzzle.

A discussion on learning and prior ignorance for sets of priors in the one-parameter exponential family

Alessio Benavoli and Marco Zaffalon
IDSIA, Galleria 2, CH-6928 Manno (Lugano), Switzerland
email: alessio@idsia.ch, zaffalon@idsia.ch

Abstract

For a conjugate likelihood-prior model in the one-parameter exponential family of distributions, we show that, by letting the parameters of the conjugate exponential prior vary in suitable sets, it is possible to define a set of conjugate priors $\mathscr{M}$ that guarantees prior near-ignorance without producing vacuous inferences. This result is obtained following both a behavioural and a sensitivity analysis interpretation of prior near-ignorance. We also discuss the problem of the incompatibility of learning and prior near-ignorance for sets of priors in the one-parameter exponential family of distributions in the case of imperfect observations. In particular, we prove that learning and prior near-ignorance are compatible under an imperfect observation mechanism if and only if the support of the priors in $\mathscr{M}$ is the whole real axis.

Keywords. Prior near-ignorance, set of distributions, exponential family of distributions.

1 Introduction

This paper deals with the problem of modelling prior ignorance about statistical parameters through a set of prior distributions $\mathscr{M}$. There are two distinct approaches of this kind. The first approach, known as *Bayesian sensitivity analysis* [2], assumes that there is an ideal prior distribution π_0 which could, ideally, model prior uncertainty. It is assumed that we are unable to determine π_0 accurately because of limited time or resources. The criterion for including a particular prior distribution π in $\mathscr{M}$ is that π is a plausible candidate to be the ideal distribution π_0.

The second approach, known as the theory of coherent lower (and upper) previsions, was developed by Walley [11]. This approach revises Bayesian sensitivity analysis by directly emphasizing the upper and lower expectations (also called previsions) that are generated by $\mathscr{M}$. The upper and lower expectations of a bounded real-valued function (we call it a gamble) g on a possibility space, denoted by $\underline{E}(g)$ and $\overline{E}(g)$, are respectively the supremum and infimum of the expectations $E_P(g)$ over the probability measures P in $\mathscr{M}$ (if $\mathscr{M}$ is assumed to be closed and convex,[1] it is fully determined by all the upper and lower expectations). The upper and lower expectations have a behavioural interpretation (explained in Section 2), but, contrary to the sensitivity analysis approach, there is no special commitment to the individual probability distributions in $\mathscr{M}$. In choosing a set $\mathscr{M}$ to model prior near-ignorance, the main aim is to generate upper and lower expectations with the property that $\underline{E}(g) = \inf g$ and $\overline{E}(g) = \sup g$ on a specific class of gambles of interest g. This means that the only available information about $E(g)$ is that it belongs to $[\inf g, \sup g]$, which is equivalent to state a condition of complete prior ignorance about the value of g.

Modeling a state of prior ignorance about the value w of a random variable W is not the only requirement for $\mathscr{M}$, it should also lead to non-vacuous posterior inferences. Posterior inferences are vacuous if the lower and upper expectations of all gambles of interest g coincide with the infimum and, respectively, the supremum of g. This means that our prior beliefs do not change with experience (i.e., there is no learning from data).

In [1], following an approach based on the behavioural interpretation, we have defined a set of *minimal properties* that a set $\mathscr{M}$ of distributions should satisfy to be a model of prior near-ignorance that does not lead to vacuous inferences. Furthermore, in the case that the likelihood model is in the one-parameter exponential family and $\mathscr{M}$ includes the corresponding conjugate exponential priors, we have also shown that the set of priors $\mathscr{M}$ satisfying the above properties can be uniquely obtained by letting the parameters of the conjugate exponential prior vary in suitable sets.

In this paper, after reviewing the main results of [1], we show that, for the one-parameter exponential family, similar conclusions about the parametrization of $\mathscr{M}$ (which guarantee prior near-ignorance and non-vacuous in-

[1] Closed and convex in the weak* topology, see [11, Sec. 3.6] for more details.

ferences) can be derived via a sensitivity analysis of the quantities of interest to the choice of the prior parameters.

We also deal with the problem of imperfect observations. In [8], it has been proven that the imprecise Beta model yields vacuous parametric inferences in the case the observation mechanism is imperfect. It is also shown that learning and prior near-ignorance are incompatible for the imprecise Beta model in the case of imperfect observations.[2] A question is if the impossibility to learn from imperfect observations under prior near-ignorance holds in general for any prior model based on sets of distributions. Here, considering conjugate likelihood-prior models in the one-parameter exponential family, we show that learning and prior near-ignorance are compatible under an imperfect observation mechanism if and only if the support of the priors in $\mathscr{M}$ is the whole real axis.

2 A Behavioural Interpretation of Prior Near-Ignorance

The aim of this section is to define which minimal properties the set of priors $\mathscr{M}$ should satisfy in the case where there is (almost) no prior information about $w \in \mathscr{W} \subseteq \mathbb{R}$. Before listing these properties, we discuss the behavioural interpretation of upper and lower expectations.

By regarding a gamble $g : \mathscr{W} \to \mathbb{R}$ as a random reward, which depends on the a priori unknown value of w, the expectation (also called prevision) of g w.r.t. w, i.e., $E(g)$, represents a subject's fair price for the function g. This means that he should be disposed to accept the uncertain rewards $g - E(g) + \varepsilon$ (i.e., to *buy* g at the price $E(g) - \varepsilon$) and $E(g) - g + \varepsilon$ (i.e., to *sell* g at the price $E(g) + \varepsilon$) for every $\varepsilon > 0$. More generally, the supremum acceptable buying price and the infimum acceptable selling prices for g need not coincide, meaning that there may be a range of prices $[a,b]$ for which our subject is neither disposed to buy nor to sell g at a price $k \in [a,b]$. His supremum acceptable buying price for g is then his lower expectation $\underline{E}(g)$, and it holds that the subject is disposed to accept the uncertain reward $g - \underline{E}(g) + \varepsilon$ for every $\varepsilon > 0$; and his infimum acceptable selling price for g is his upper prevision $\overline{E}(g)$, implying that he is disposed to accept the reward $\overline{E}(g) - g + \varepsilon$ for every $\varepsilon > 0$. A consequence of this interpretation is that $\underline{E}(g) = -\overline{E}(-g)$ for every gamble g.

Under this behavioural interpretation, a state of ignorance about a gamble g is modelled by setting $\underline{E}(g) = \inf g$ and $\overline{E}(g) = \sup g$. This means that our subject is neither disposed to buy nor to sell g at any price $k \in [\inf g, \sup g]$. In other words, our subject is disposed to buy (sell) g only at a price strictly less (greater) than the minimum (maximum) reward that he would gain from g. This means that the available information on w does not allow our subject to set any meaningful buying or selling price for g, which is equivalent to stating that our subject is in a state of ignorance.

In [11], it is proven that a closed and convex set of probability distributions can be equivalently characterized by the lower (or upper) expectation functional that it generates as the lower (upper) envelope of the expectations obtained from the distributions in such a set. Vice versa, given a functional $\underline{E}(\cdot)$ that satisfies some regularity properties [11, Ch. 2], it is possible to define a family $\mathscr{M}$ of probability distributions that generates the lower expectation $\underline{E}(g)$ for any g. This establishes a one-to-one correspondence between closed convex sets of probability distributions and lower expectations.

In case the available prior information is scarce, it therefore seems more natural to define $\mathscr{M}$ according to the behavioural interpretation, i.e., in terms of the upper and lower expectations it generates [7]. For instance, in problems where there is (almost) no prior information one would expect the set $\mathscr{M}$ to be "large" in the sense that its generated upper and lower expectations are relatively far apart (vacuous or almost vacuous).

Modelling a state of prior ignorance about w is not the only requirement for $\mathscr{M}$, it must also produce non-vacuous posterior inferences (otherwise it is useless in practice). Hereafter, inspired by the work in [7], we define a set of minimal properties that $\mathscr{M}$ or, equivalently, the lower and upper expectations it generates, should satisfy to be a model of prior ignorance and produce consistent and meaningful posterior inferences. The first requirement for $\mathscr{M}$ is coherence.

(A.1) Coherence. Prior and posterior inferences based on $\mathscr{M}$ should be strongly coherent [11, Sec. 7.1.4(b)]. Under the behavioural interpretation, this means that we should not be able to raise the lower expectation (supremum acceptable buying price) of a given gamble g taking into account the acceptable transactions implicit in the other lower expectation models.

In practice, strong coherence imposes joint constraints on the prior, likelihood and posterior lower expectation models, in the sense that, when considered jointly, they should not imply inconsistent assessments. In [11, Sec. 7.8.1], it is proven that, in the case the prior and likelihood lower expectation models are obtained as lower envelopes of standard expectations w.r.t. sets of proper density functions and the posterior set of densities is obtained from these sets by element-wise application of Bayes' rule for density functions, then strong coherence of the

[2] Actually the results in [8] are more general and hold for a multivariate prior near-ignorant model defined on a compact set. However, since the present paper deals with the one-parameter exponential family, in the following we focus our attention on the restriction of [8] to the imprecise Beta model.

respective lower expectation models is satisfied.[3]

Besides coherence, other requirements for the set $\mathscr{M}$ are that it should represent the state of prior ignorance about w, but without producing vacuous posterior inferences. Thus, $\mathscr{M}$ should be large enough to model a state of prior ignorance w.r.t. a set of suitable gambles (i.e., a set of gambles of interest $\mathscr{G}_0$ w.r.t. which we assess our state of prior ignorance), but not too large to prevent learning from taking place. These two contrasting requirements are captured by the following two properties for $\mathscr{M}$.

(A.2) $\mathscr{G}_0$-prior ignorance. The prior upper and lower expectations of some suitable set of gambles $\mathscr{G}_0$ under $\mathscr{M}$ are vacuous, i.e., $\underline{E}[g] = \inf g(w)$ and $\overline{E}[g] = \sup g(w)$ for all $g \in \mathscr{G}_0$.

(A.3) $\mathscr{G}$-learning. For a chosen set of gambles $\mathscr{G} \supseteq \mathscr{G}_0$ and for each $g \in \mathscr{G}$ satisfying $\overline{E}[g] - \underline{E}[g] > 0$, there exists a finite $\delta > 0$ (possibly dependent on g) such that for each $n \geq \delta$ and non-empty sequence of observations $y^n = (y_1, \ldots, y_n)$, at least one of these two conditions is satisfied:

$$\underline{E}[g|y^n] \neq \underline{E}[g], \quad \overline{E}[g|y^n] \neq \overline{E}[g], \tag{1}$$

where $\underline{E}[\cdot|y^n]$ and $\overline{E}[\cdot|y^n]$ denote the posterior lower and upper expectations of g after having observed $y_1, \ldots, y_n$. Furthermore, for each $g \in \mathscr{G}_0$, (1) must hold for any $n > 0$.

Property (A.2) states that $\mathscr{M}$ should be vacuous a priori w.r.t. some set of gambles $\mathscr{G}_0$, i.e., the lower and upper expectations of $g \in \mathscr{G}_0$ respectively coincide with the infimum and the supremum of g. In case $\mathscr{M}$ includes all possible distributions then (A.2) holds for any function g. Here, conversely, we require that (A.2) is satisfied for some subset of gambles $\mathscr{G}_0$. The subset of gambles $\mathscr{G}_0$ used in (A.2) should include the gambles g w.r.t. which we state our condition of prior near-ignorance. Furthermore, the set $\mathscr{G}_0$ should be as large as possible to guarantee that also $\mathscr{M}$ is as large as possible, but no too large to be incompatible with the requirement (A.3) of learning. In fact, property (A.3) states that $\mathscr{M}$ should be non-vacuous a posteriori for any gamble $g \in \mathscr{G} \supseteq \mathscr{G}_0$, which is a condition for learning from the observations. The set of gambles $\mathscr{G}$ used in (A.3) should include the gambles g w.r.t. which we are interested in computing expectations (i.e., making inferences). The fact that $\mathscr{G}$ must include $\mathscr{G}_0$ is the only constraint on $\mathscr{G}$, meaning that (A.3) requires that $\mathscr{M}$ is not vacuous w.r.t. all these gambles for which the prior near-ignorance has been imposed. Moreover, for these gambles, it is required that (1) holds for any $n > 0$, i.e., after one observation the condition of prior-ignorance must already be left.

[3] This holds under standard assumptions about the existence of density functions and the applicability of Bayes' rule.

Since $\mathscr{M}$ is a model of prior near-ignorance, it is also desirable that the influence of $\mathscr{M}$ on the posterior inferences vanishes with increasing numbers of observations n. This is captured by the following property.

(A.4) Convergence. For each gamble $g \in \mathscr{G}$ and non-empty sequence of observations $y^n = (y_1, \ldots, y_n)$, the following conditions are satisfied for $n \to \infty$:

$$\begin{array}{rcl} \underline{E}[g|y^n] & \to & \underline{E}^*[g|y^n], \\ \overline{E}[g|y^n] & \to & \overline{E}^*[g|y^n], \end{array} \tag{2}$$

where $\underline{E}^*[g|y^n], \overline{E}^*[g|y^n]$ are the posterior lower and upper expectations obtained as lower envelopes of standard expectations w.r.t. the posterior densities derived, via Bayes' rule, from the likelihood model and the improper prior density $p(w) = 1$ for all $w \in \mathscr{W}$.

Property (A.4) states that, for $n \to \infty$, $\mathscr{M}$ should give the same lower and upper expectations of $g \in \mathscr{G}$ as those obtained from the improper prior density $p(w) = 1$. The fact that $\underline{E}^*[g|y^n] < \overline{E}^*[g|y^n]$ accounts for the general case in which the likelihood model is described by a set of likelihoods (for a single likelihood it would be $\underline{E}^*[g|y^n] = \overline{E}^*[g|y^n] = E^*[g|y^n]$). Although improper priors produce posteriors which are often incoherent with the likelihood model, (A.4) does not conflict with the requirement of coherence in (A.1). In fact (A.4) is a limiting property that holds only for $n \to \infty$ (furthermore, incoherence usually vanishes at the limit). In order to better understand properties (A.1)–(A.4), we show their instantiation for the case of the exponential family in Section 4. Before discussing these results, in the next section we introduce the exponential families of densities and review their main properties [4, Ch. 5].

3 Exponential Families

Consider a sampling model where i.i.d. samples of a random variable Z are taken from a sample space $\mathscr{Z}$.

Definition 1. A probability density $p(z|x)$, parametrized by $x \in \mathscr{X} \subseteq \mathbb{R}$, is said to belong to the one-parameter exponential family if it is of the form

$$p(z|x) = f(z)[g(x)]^{-1} \exp\left(c\phi(x)h(z)\right), \; z \in \mathscr{Z} \tag{3}$$

where, given f, h, ϕ and c, it results that $g(x) = \int_{z \in \mathscr{Z}} f(z) \exp\left(c\phi(x)h(z)\right) dz < \infty$. ■

Sometimes it is more convenient to rewrite (3) in a different form.

Definition 2. The probability density

$$p(y|w) = k(y) \exp(yw - b(w)), \; y \in \mathscr{Y}_m, \tag{4}$$

derived from (3) via the transformations $y = h(z)$, $\mathscr{Y}_m = h(\mathscr{Z})$, $w = c\phi(x)$, $b(w) = \ln(g(x))$ and $k(y) = f(z)$, is

called the canonical form of representation of the exponential family; w is called the natural (or canonical) parameter. ■

The canonical form has some useful properties. The mean and variance of Y are given by

$$E[Y|w] = \frac{db}{dw},\ E[(Y - E_Y[Y|w])^2|w] = \frac{d^2b}{dw^2}, \quad (5)$$

where it has been assumed that $\frac{d^2b}{dw^2}(w) > 0$; from (5) it follows that $\frac{db}{dw}(w) \in Int(\mathscr{Y})$ (i.e., interior of $\mathscr{Y}$) [5], where $\mathscr{Y} \subseteq \mathbb{R}$ is the smallest closed or semi-closed set that includes the sample mean of Y (if it exists, otherwise $\mathscr{Y} = Int(\mathscr{Y})$). Notice that the domain of the observations $\mathscr{Y}_m$ can be discrete or continuous, while $\mathscr{Y}$ is always continuous. In the case of n i.i.d. observations $y_i = h(z_i)$, it follows that

$$p(y^n|w) = \prod_{i=1}^{n} p(y_i|w) = \prod_{i=1}^{n} k(y_i)\exp(n(\hat{y}_n w - b(w))), \quad (6)$$

where $\hat{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean of the y_i which, together with n, is a sufficient statistic of y^n for inference about w under the i.i.d. assumption. Furthermore, by interpreting the density function in (6) as a likelihood function $L(w)$, with $y^n = (y_1, \ldots, y_n)$, we can define the corresponding conjugate prior.

Definition 3. A probability density $p(w|n_0, y_0)$, parametrized by $n_0 \in \mathbb{R}^+$ and $y_0 \in Int(\mathscr{Y})$, is said to be the canonical prior of (4) if

$$p(w|n_0, y_0) = k(n_0, y_0)\exp(n_0(y_0 w - b(w))), \quad (7)$$

where $w \in \mathscr{W}$, n_0 is the so-called number of pseudo-observations, y_0 is the so-called pseudo-observation and $k(n_0, y_0)$ is the normalization constant. ■

When $\mathscr{W} = \mathbb{R}$, $0 < n_0 < \infty$ and $y_0 \in Int(\mathscr{Y})$, (7) is a proper density [5]. Some examples of densities conjugate to a one-parameter exponential (canonical) family and defined in $\mathscr{W} = \mathbb{R}$ follow.

Gaussian with known variance: $y \in \mathscr{Y} = \mathbb{R}$, $x \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$,

$$p(y|x, \sigma^2) \propto \exp\left(-\tfrac{1}{2\sigma^2}(y-x)^2\right) \propto \exp\left(\tfrac{1}{\sigma^2}\left(yx - \tfrac{x^2}{2}\right)\right),$$

with $w = x$ and $b(w) = x^2/2$. The conjugate prior (7) transformed back to the original domain $\mathscr{X}$ is:

$$p(x|n_0, y_0) \propto \exp\left(-\tfrac{n_0}{2}(x - y_0)^2\right),$$

which is a Gaussian with mean y_0 and variance $1/n_0$.

Binomial-Beta: $x \in \mathscr{X} = (0,1)$, $y \in \{0,1\}$,

$$\begin{aligned} p(y|x) &\propto x^y(1-x)^{(1-y)} \\ &= (1-x)\exp\left(y\ln\left(\frac{x}{1-x}\right)\right) \\ &= \exp(yw - b(w)), \end{aligned}$$

$w = \ln(x/(1-x))$, $b(w) = -\ln(1-x) = \ln(1+\exp(w))$. Considering the change of variable $dx = \exp(w)/(1+\exp(w))^2 dw$, the conjugate prior (7) transformed back to the original domain $\mathscr{X}$ is:

$$p(x|n_0, y_0) \propto x^{n_0 y_0 - 1}(1-x)^{n_0(1-y_0)-1}$$

which is a Beta density with $n_0 = s > 0$ and $y_0 = t \in (0,1)$.

The pair likelihood and conjugate prior in the canonical exponential family satisfies a set of interesting properties, most of them are particularly useful to represent the nature of the Bayesian "learning" process. A list of such properties is given in the following lemmas, whose proof is omitted (see [4, Ch. 5]).

Lemma 1. *For a pair of likelihood and conjugate prior in the canonical exponential family, it holds that:*

(i) *the posterior density for w is:*

$$p(w|n_p, y_p) = k(n_p, y_p)\exp(n_p(y_p w - b(w))), \quad (8)$$

where $n_p = n + n_0$ and $y_p = \frac{n_0 y_0 + n\hat{y}_n}{n+n_0}$;

(ii) *the predictive density for future observations $(y_{n+1}, \ldots, y_{n+m})$ is*

$$p(y_{n+1}, \ldots, y_{n+m}|y_1, \ldots, y_n) = \prod_{j=1}^{m} k(y_{n+j}) \frac{k\left(n_0 + n, \frac{n_0 y_0 + n\hat{y}_n}{n+n_0}\right)}{k\left(n_0 + n + m, \frac{n_0 y_0 + (n+m)\hat{y}_{n+m}}{n+m+n_0}\right)}. \quad (9)$$

■

Lemma 2. *Suppose that the canonical conjugate prior family is such that $p(w|n_0, y_0) \to 0$ for $w \to \sup\mathscr{W}$ and $w \to \inf\mathscr{W}$. Then the prior mean of the function $\frac{db}{dw}$ is $E\left[\frac{db}{dw}\middle| n_0, y_0\right] = y_0$ and the posterior mean is:*

$$E\left[\frac{db}{dw}\middle| n_p, y_p\right] = \frac{n_0 y_0 + n\hat{y}_n}{n+n_0}. \quad (10)$$

■

Notice that $p(w|n_0, y_0) \to 0$ for $w \to \sup\mathscr{W}$ and $w \to \inf\mathscr{W}$ holds for any canonical priors such that $\mathscr{W} = \mathbb{R}$, but in general it is not true for truncated priors, i.e., in the case $\mathscr{W} \subset \mathbb{R}$. This is one of the reasons why it has been assumed that $\mathscr{W} = \mathbb{R}$. In (5), it has been shown that $\frac{d}{dw}b(w)$ is the mean of Y. Hence, $\frac{d}{dw}b(w)$ is the quantity about which we will have prior beliefs before seeing the data y and posterior beliefs after observing the data. Hence, the results in Lemma 2 are particularly important, because they provide us with a closed formula for the prior and posterior mean of $\frac{d}{dw}b(w)$. For sampling models such that $\frac{d}{dw}b(w) = x$, i.e., linear exponential form (e.g., Gaussian, Beta and Gamma density), Lemma 2 gives thus a closed formula for the prior and posterior mean of x.

4 Sets of Conjugate Priors for Exponential Families

Consider the problem of statistical inference about the real-valued parameter w from noisy measurements $(y_1,\dots,y_n)$ and assume that the likelihood is completely described by the following probability density function (PDF) belonging to the exponential family:

$$\prod_{i=1}^{n} p(y_i|w) = \prod_{i=1}^{n} k(y_i)\exp(n(\hat{y}_n w - b(w))), \qquad (11)$$

where the parameters of the likelihood, i.e., sample mean $\hat{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $n \in \mathbb{R}^+$, are known (the likelihood can be modelled by a single PDF). By conjugacy and following a Bayesian approach, as prior for w we may consider the PDF $p(w|n_0,y_0)$ defined in (7) for a given value of the parameter y_0 and n_0. In the case there is not enough information about w to uniquely determine the values of the parameters y_0 and n_0, we can consider the family of priors $p(w|n_0,y_0)$ obtained by letting y_0 vary in $\mathscr{Y}' \subseteq Int(\mathscr{Y})$ and n_0 in some set $\mathscr{A}_{y_0} \subseteq \mathbb{R}^+$, which could depend on y_0. The question to be addressed is whether such family of priors satisfies the properties (A.1)–(A.4) discussed in Section 2. The answer to this question is given in the next theorem.

Theorem 1. *Consider as set of priors $\mathscr{M}$ the family of conjugate priors $p(w|n_0,y_0)$ with y_0 spanning the set $\mathscr{Y}' \subseteq Int(\mathscr{Y})$, n_0 spanning the set $\mathscr{A}_{y_0} \subseteq \mathbb{R}^+$ (with $\mathscr{A}_{y_0}$ possibly dependent on y_0), under the assumptions: $\mathscr{Y}$ convex and $\mathscr{W} = \mathbb{R}$. If and only if the following conditions hold:*

(a) For each $y_0 \in \mathscr{Y}'$ and $n_0 \in \mathscr{A}_{y_0}$, it holds that $p(w|n_0,y_0) \to 0$ for $w \to \sup\mathscr{W}$ and $w \to \inf\mathscr{W}$;

(b) $\mathscr{Y}' = Int(\mathscr{Y})$;

(c) $\mathscr{A}_{y_0}$ satisfies the following constraints: $0 < \inf\mathscr{A}_{y_0}$, $\sup\mathscr{A}_{y_0} \leq \min(\overline{n}_0, \frac{c}{|y_0|})$ for each $y_0 \in Int(\mathscr{Y})$ and given parameters $\overline{n}_0, c > 0$;

then, given the parameters $\overline{n}_0$ and c, $\mathscr{M}$ is the largest set which satisfies properties (A.1)–(A.4), with $\mathscr{G}_0 = \{\frac{db}{dw}\}$ and $\mathscr{G}$ including sufficiently smooth gambles.[4] ■

The proof of the theorem can be found in [1, Sec. 4]. Hereafter, we illustrate the intuition behind the theorem. We distinguish three cases $\mathscr{Y} = \mathbb{R}$, $\mathscr{Y} = [a,\infty)$ (or $\mathscr{Y} = (-\infty,a]$) with $a \in \mathbb{R}$, and $\mathscr{Y} \subset \mathbb{R}$ bounded. In the last two cases w.l.o.g. it can be assumed that $\mathscr{Y} = [0,\infty)$ (or $\mathscr{Y} = (-\infty,0]$) and, respectively, $\mathscr{Y} = [0,1]$ (by shifting and scaling $\mathscr{Y}$); since $\mathscr{Y}$ has been assumed to be convex, these three cases account for all the possibilities.

[4] With sufficiently smooth gambles, we mean integrable w.r.t. the exponential family density functions with support in $\mathscr{W}$ and continuous on a neighborhood of the point where the posterior relative to the improper prior $p(w) = 1$ concentrates for $n \to \infty$.

Consider the case in which the observations belong to $\mathbb{R}$ and the likelihood is a Gaussian density with known variance, so that $\mathscr{Y} = (-\infty,+\infty)$. The conjugate model under considerations is thus a Gaussian-Gaussian model. In this case, the set of priors $\mathscr{M}$ is equal to:

$$\Big\{\mathscr{N}\left(w;y_0,\sigma_0^2\right):\ y_0 \in (-\infty,+\infty),\ \max(1/\overline{n}_0,|y_0|/c) < \sigma_0^2 < \infty\Big\}, \qquad (12)$$

where y_0 is the prior mean and $\sigma_0^2 = 1/n_0$ the prior variance. Hence, $\mathscr{M}$ includes all the Gaussian densities with mean free to vary in $\mathbb{R}$ and variance lower bounded by $1/\overline{n}_0$ but linearly increasing with $|y_0|$. Notice, in fact, that if $|y_0| > c/\overline{n}_0$, then $\sigma_0^2 \geq |y_0|/c$. Hence, considering the likelihood $\mathscr{N}(y_i;w,\sigma^2)$ for $i = 1,\dots,n$, the corresponding set of posteriors is equal to:

$$\Big\{\mathscr{N}\left(w;y_p,\sigma_p^2\right):\ y_p = \sigma_p^2\left(\frac{y_0}{\sigma_0^2} + \frac{n\hat{y}_n}{\sigma^2}\right),\ \sigma_p^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1},\ y_0 \in (-\infty,+\infty),\ \max(1/\overline{n}_0,|y_0|/c) < \sigma_0^2 < \infty\Big\}, \qquad (13)$$

where y_p is the posterior mean. Since $y_p = (n_0 y_0 + n\hat{y}_n)/(n+n_0)$ then, fixed $n_0 = 1/\sigma_0^2$, for $|y_0| \to \infty$ it follows that $|y_p| = |n_0 y_0 + n\hat{y}_n|/(n+n_0) = |y_0| \to \infty$. Similarly, fixed y_0, for $n_0 \to \infty$ it follows that $|y_p| = |y_0|$. In other words, $n_0|y_0| = \infty$ implies a vacuous posterior mean and, thus, no learning and no convergence. Theorem 1 states that a necessary and sufficient condition to guarantee near-ignorance without preventing learning and convergence to take place is by imposing the constraint:

$$|n_0 y_0| < c < \infty,$$

which means that n_0 must in general depend on y_0. In this case in fact for $|y_0| \to \infty$, it follows that $|y_p| = |n_0 y_0 + n\hat{y}_n|/(n+n_0) < \infty$. That is, the contribution of y_0 to y_p must decrease as $|y_0| \to \infty$, otherwise the observations do not contribute to y_p (learning cannot take place). This is essentially the meaning of the constraint $|y_0|/c < \sigma_0^2$ in (13), i.e., the variance of the Gaussians in $\mathscr{M}$ must be greater than $|y_0|/c$. Furthermore, $n_0 < \infty$ or, equivalently, the variance must also be greater than zero otherwise the Gaussian density would coincide with a Dirac delta; this is the reason of the constraint $\sigma_0^2 > 1/\overline{n}_0 > 0$. Under these constraints, it can be verified that y_p satisfies:

$$\min\left(\frac{-c+n\hat{y}_n}{n+\overline{n}_0}, \frac{-c+n\hat{y}_n}{n}\right) \leq y_p = \frac{n_0 y_0 + n\hat{y}_n}{n+n_0} \leq \max\left(\frac{c+n\hat{y}_n}{n+\overline{n}_0}, \frac{c+n\hat{y}_n}{n}\right), \qquad (14)$$

and converges to $\hat{y}_n$ (maximum likelihood estimate) for $n \to \infty$ (convergence property (A.4)). Observe that, for

$\overline{n}_0$ suitably small, the set of priors $\mathscr{M}$ reduces to the family of Gaussian priors with infinite variance discussed in [7, Section 3.3] and the bounds in (14) become approximately equal to:

$$\frac{-c+n\hat{y}_n}{n} \leq \frac{n_0 y_0 + n\hat{y}_n}{n+n_0} \leq \frac{c+n\hat{y}_n}{n}. \tag{15}$$

The main difference is that the family of priors defined in Theorem 1 has been proved to be strongly coherent, while no proof of coherence is given for the model in [7, Section 3.3]; the coherence of this model is still an open problem.

Consider now the case in which the observations are counts, i.e., the likelihood is a Poisson distribution, $\mathscr{Y}_m = \mathbb{N}$ and $\mathscr{Y} = [0,\infty)$. The conjugate model under consideration is now a Poisson-Gamma model. The set of priors $\mathscr{M}$ transformed back to the original parameter space $\mathscr{X}$ reduces to a set of Gamma densities:

$$\mathscr{M} = \Big\{ g(x|\alpha,\beta) : \ 0 < \alpha = n_0 y_0 \leq c, \\ 0 < \beta = n_0 \leq \min(\overline{n}_0, c/|y_0|) \Big\}, \tag{16}$$

where $x, y_0 \in (0,+\infty)$ and $g(x|\alpha,\beta) \propto x^{\alpha-1}\exp(-\beta x)$ is the Gamma density with parameters α and β. The set of posteriors resulting from (16) is:

$$\mathscr{M}_p = \Big\{ g(x|\alpha,\beta) : \ \alpha = n_0 y_0 + n\hat{y}_n, \ \beta = n+n_0, \\ y_0 \in (0,+\infty), \ 0 < n_0 \leq \min(\overline{n}_0, c/|y_0|) \Big\} \tag{17}$$

and the posterior mean is equal to $y_p = (n_0 y_0 + n\hat{y}_n)/(n+n_0)$. Notice again that, because of the constraint $n_0 \leq \min(\overline{n}_0, c/|y_0|)$ it results that y_p is always finite, satisfies[5]

$$\frac{n\hat{y}_n}{n+\overline{n}_0} \leq y_p = \frac{n_0 y_0 + n\hat{y}_n}{n+n_0} \leq \frac{c+n\hat{y}_n}{n},$$

and converges to $\hat{y}_n$ (maximum likelihood estimate) for $n \to \infty$.

Consider the case in which the observations are binary, i.e., the likelihood is a binomial distribution $\mathscr{Y}_m = \{0,1\}$ and $\mathscr{Y} = [0,1]$. The conjugate model under considerations is thus a Binomial-Beta model. It can be easily verified that in this case the set of priors $\mathscr{M}$ transformed back to the original parameter space $\mathscr{X}$ reduces to the general Imprecise Beta Model (IBM) discussed in [11, Section 5.4.3]:

$$\mathscr{M} = \Big\{ B(x; st, s(1-t)) : t \in (0,1), 0 < s < \overline{n}_0 \Big\}, \tag{18}$$

where $x \in (0,1)$, $y_0 = t$, $n_0 = s$ and $B(x;\alpha,\beta)$ is the Beta density with parameters α and β. In this case, it follows from Theorem 1 that $y_0 \in (0,1)$ and $0 < n_0 \leq \min(\overline{n}_0, c)$. Hence, if $\overline{n}_0 < c$ the set of priors in Theorem 1 reduces to (18). In this case, near-ignorance and learning/convergence are compatible even if n_0 does not depend on y_0. In fact, being $|y_0| < 1 < \infty$, the product $n_0 y_0$ is always bounded provided that $n_0 < \overline{n}_0 < \infty$.[6] Finally notice that in the special case $s = \overline{n}_0$, we obtain the IBM discussed in [11, Section 5.3.1] and [3].

Observe that the family of priors $\mathscr{M}$ in Theorem 1 is completely determined by the two parameters $c > 0$ and $\overline{n}_0 > 0$. The larger these parameters are the larger the family of priors $\mathscr{M}$ is and, thus, the more conservative are the posterior inferences. The choice of these parameters is discussed in [1, Sec. 5].

It is also interesting to compare the set of priors $\mathscr{M}$ in Theorem 1 with another model for near-ignorance, the Bounded Derivative Model (BDM) [12]. In the BDM, $\mathscr{M}_{BDM}$ includes all continuous proper probability density functions for which the derivative of the log-density is bounded by a positive constant. It can be verified that BDM satisfies all the properties (A1)–(A4), with $\mathscr{G}_0$ and $\mathscr{G}$ defined as in Theorem 1. BDM is a non-parametric model and, in this sense, is more general than the model resulting from Theorem 1 that is restricted to the one-parameter exponential family only. A drawback of this generality is that inferences with BDM can in general be difficult to compute [12, Sec. 6], while this is often not the case for the model resulting from Theorem 1 because of conjugacy.

Conversely, a model for statistical inferences based on a set of densities belonging to the exponential family is presented in [9, Ch.4], [10]. The main difference w.r.t. the present work is that the model in [10] is not a model of prior near-ignorance, as pointed out by the authors, i.e., the set $\mathscr{Y}'$ in Theorem 1 is chosen in [10] to reflect the prior information on y_0 and, thus, the posterior inferences depend on this information. Since no constraint between n_0 and y_0 is assumed, the model in [10] can also violate (A.3)–(A.4) in the case $\mathscr{Y}' = Int(\mathscr{Y})$, and hence it can produce vacuous inferences.

5 A Sensitivity Analysis Interpretation of Prior Near-Ignorance

In Section 2, we have considered an interpretation of prior near-ignorance in terms of lower and upper expectations, i.e., behavioural dispositions to buy and sell gambles. In particular, with the properties (A1)–(A4), we have given general conditions for coherence, prior near-ignorance, learning and convergence, which hold for any set of distributions $\mathscr{M}$. Then, in Section 4, we have specialized these

[5] Since $\hat{y}_n \geq 0$, it results that $\frac{c+n\hat{y}_n}{n} \geq \frac{c+n\hat{y}_n}{n+\overline{n}_0}$ and, thus, $\frac{c+n\hat{y}_n}{n}$ is a right bound for y_p.

[6] In [13] the authors propose a functional relationship between n_0 and y_0 in the exponential families with a different aim w.r.t. that of the present paper; that is highlighting prior-data conflict in the case of inference drawn from a set of informative priors, i.e., near-ignorance is not satisfied. In this case, n_0 may depend on y_0 also in the IBM.

conditions to the case in which $\mathscr{M}$ includes densities belonging to the one-parameter exponential family and, for this set of densities, we have shown that (A1)–(A4) are equivalent to a special choice of the domains for the parameters of the exponential priors.

An alternative approach is to start directly from the set of priors $\mathscr{M}$ in the one-parameter exponential family and then to perform a sensitivity analysis of the quantities of interest (posterior inferences) to the choice of the prior parameters. This is typically done by deriving the quantities of interest w.r.t. the parameters of the conjugate priors, and looking for a set of parameters that sharply changes the inferences.

In this respect, consider a function $g(\frac{db}{dw})$[7] and its Taylor series expansion around the posterior parameter y_p, i.e.:

$$g\left(\frac{db}{dw}\right) = g(y_p) + \left(\frac{db}{dw} - y_p\right) g'(y_p) + \frac{1}{2}\left(\frac{db}{dw} - y_p\right)^2 g''(y_p) + \dots \quad (19)$$

where $g'(y_p) = \frac{dg}{d(\frac{db}{dw})}|_{y_p}$ and so on for higher order derivatives. In statistical inference, we are interested in computing the expectation of g or, equivalently, of (19) w.r.t. the posterior density $k(n_p, y_p)\exp(n_p(y_p w - b(w)))$, i.e.:

$$\begin{aligned} E[g|y^n] &= \int g\left(\frac{db}{dw}\right) k(n_p, y_p)\exp(n_p(y_p w - b(w)))dw \\ &= g(y_p) + \frac{1}{2}g''(y_p)E\left[\left(\frac{db}{dw} - y_p\right)^2 \middle| y^n\right] \\ &\quad + \frac{1}{3!}g'''(y_p)E\left[\left(\frac{db}{dw} - y_p\right)^3 \middle| y^n\right] + \dots \end{aligned} \quad (20)$$

where, for short notation, $\{y_1, \dots, y_n\} = y^n$ has been introduced. The posterior expectation $E[g|y^n]$ depends on $y_p = (n_0 y_0 + n\hat{y}_n)/(n + n_0)$ which, in turn, depends on the prior parameters n_0 and y_0. The sensitivity of $E[g|y^n]$ to the prior parameters can be obtained by differentiating $E[g|y^n]$ w.r.t. n_0 and y_0. However, since the value of n_0 may depend on the value of y_0 and vice versa, it is more interesting to compute the sensitivity of $E[g|y^n]$ to variations of $n_0 y_0$. Define $n_0 y_0 = r$ and $n_0 = n_0(r)$, then

$$\frac{dy_p}{dr} = \frac{n + n_0 - (r + n\hat{y}_n)\frac{dn_0}{dr}}{(n + n_0)^2}. \quad (21)$$

where n_0 depends on r. Thus, it follows that $\dfrac{dE[g|y^n]}{dr}$ is equal to

$$\begin{aligned} &\frac{dy_p}{dr}\frac{dg(y_p)}{dy_p} + \frac{1}{2}\frac{dy_p}{dr}\frac{dg''(y_p)}{dy_p}E\left[\left(\tfrac{db}{dw} - y_p\right)^2 \middle| y^n\right] \\ &+ \frac{1}{2}\frac{dy_p}{dr}g''(y_p)\frac{dE\left[\left(\frac{db}{dw} - y_p\right)^2 \middle| y^n\right]}{dy_p} + \dots \end{aligned} \quad (22)$$

From the relationship between a derivative and its difference quotient, one gets

$$|E_{r+\Delta}[g|y^n] - E_r[g|y^n]| \leq \left|\frac{dE[g|y^n]}{dr}\right| |\Delta| \quad (23)$$

where $E_{r+\Delta}[g|y^n]$ is the expected value of g computed at $n_0 y_0 = r + \Delta$, $E_r[g|y^n]$ is the expected value of g computed at $n_0 y_0 = r$ and Δ is a scalar such that $r + \Delta \in [\min n_0 y_0, \max n_0 y_0]$.

Theorem 2. *There exists a finite $\delta > 0$ (possibly dependent on g) such that, for each $n \geq \delta$ and non-empty set of observations $y_1, \dots, y_n$, the difference $\max_{r,\Delta}|E_{r+\Delta}[g|y^n] - E_r[g|y^n]|$ is bounded and converges to zero for $n \to \infty$, if $\max|n_0 y_0| < \infty$ and $n_0 < \infty$.* ■

Proof: *If $\max|n_0 y_0| < \infty$, then it is also true that $\max|\Delta| = |\max n_0 y_0 - \min n_0 y_0| < \infty$. With $\max|\Delta|$ being bounded, a condition for $\max_{r,\Delta}|E_{r+\Delta}[g|y^n] - E_r[g|y^n]|$ to be bounded is that $|dE[g|y^n]/dr| < \infty$. Thus, also being $n_0 < \infty$, for $n \to \infty$ it follows that $y_p \to \hat{y}_n$, $n_p \to n$ and the posterior density $p(w|n_p, y_p)$ becomes a Dirac delta in $\hat{y}_n$. Then it results that $\lim E\left[\left(\frac{db}{dw} - y_p\right)^m \middle| y^n\right] = 0$ for any $m = 1, 2, \dots$ and $\lim dy_p/dr = 0$ (since $y_p = \hat{y}_n$, the derivative of y_p w.r.t. r is null). Thus, $|dE[g|y^n]/dr|$ converges to zero for $n \to \infty$. Furthermore, because $p(w|n_p, y_p)$ is always a well-defined PDF if $|n_0 y_0| < \infty$ and $n_0 < \infty$, by continuity arguments we can also conclude that there exists a finite $\delta > 0$ such that $|dE[g|y^n]/dr|$ is bounded for any $n > \delta$.* ■

Thus, we have again proven that $\max|n_0 y_0| < c$ and $n_0 \leq \overline{n}_0 < \infty$ are sufficient conditions for learning and convergence,[8] but now following an approach based on sensitivity analysis. Consider the case $g\left(\frac{db}{dw}\right) = \frac{db}{dw}$, assume that $p(w|n_p, y_p)$ is a Beta density and $n_0 = s > 0$. Then, from (21)–(22) it follows that $dy_p/dr = dE[g|y^n]/dr = 1/(n + s)$ (because $n_0 = s$ is constant). Since $y_0 \in (0, 1)$, then $0 < n_0 y_0 = r < s$ and, thus, $\max|\Delta| = s$, we conclude that $\max_{r,\Delta}|E_{r+\Delta}[g|y^n] - E_r[g|y^n]| \leq \frac{s}{n+s}$, which is exactly the imprecision (i.e., the difference between the upper and lower mean) of the IBM.

Consider the Gaussian case and assume $n_0 \approx 0$. For $g\left(\frac{db}{dw}\right) = \frac{db}{dw}$ it results that $dy_p/dr = dE[g|y^n]/dr = 1/n$. In this case the boundedness of $\max|\Delta|$ is ensured if $|n_0 y_0| \leq c < \infty$, which implies $\max|\Delta| = 2c$. Therefore,

[7] To simplify the derivations, we have assumed that g is an analytic function. Although not general, this holds for many gambles g.

[8] Theorem 1 is more general than Theorem 2, since it holds for more general functions g. Furthermore, the conditions derived there are not only sufficient but also necessary for (A.1)–(A.4).

(23) becomes $\max_{r,\Delta}|E_{r+\Delta}[g|y^n]-E_r[g|y^n]|\leq\frac{2c}{n}$, which is the imprecision of (15). Therefore, we have arrived at similar conclusions of those in Theorem 1 but via a sensitivity analysis. This approach allows to give another interpretation of the imprecision, e.g., $s/(n+s)$ and $2c/n$, in terms of the maximum value of the product $|dE[g|y^n]/dr||\Delta|$.

6 Imperfect observations

In real world applications, there is always a probability of making mistakes during the observation process. Often, if this probability is small, one assumes that the data are perfectly observable in order to use a simple likelihood model (e.g., a density belonging to the exponential family); doing so, one implicitly assumes that there is a sort of continuity between models with perfectly observable data and models with small probability of errors in the observations. In other words, one expects that a small error in the modelling of the observation mechanism leads to a small error in the inference. However, as observed in [8], this may be not true for inferences derived from a prior near-ignorance model based on set of distributions. To better understand this aspect, we introduce the imperfect observation mechanism described in [8]. An imperfect observation mechanism can be modelled as a two step process: (i) ideal observations $y'_1,\dots,y'_n$ are generated according to the likelihood $L(y'_1,\dots,y'_n|w)$; (ii) $y'_1,\dots,y'_n$ are perturbed based on a distribution $p(y_1,\dots,y_n|y'_1,\dots,y'_n)$ and imperfect observations $y_1,\dots,y_n$ are produced. Hence, the likelihood of imperfect observations can be modelled as:

$$p(y^n|w)=\int_{\mathscr{Y}_m^n}p(y^n|y'^n)L(y'^n|w)\,dy'^n, \tag{24}$$

where, for the sake of space, the notation $y^n=(y_1,\dots,y_n)\in\mathscr{Y}_m^n$ and $y'^n=(y'_1,\dots,y'_n)\in\mathscr{Y}_m^n$ has been introduced; $p(y^n|y'^n)=\prod_{i=1}^n p(y_i|y'_i)$ is any PDF such that $p(y_i|y'_i)>0$ for all $y_i,y'_i\in\mathscr{Y}_m$; $L(y'^n|w)=\prod_{i=1}^n L(y'_i|w)$ is the likelihood corresponding to the ideal unknown observations y'_i (we assume that it belongs to one-parameter canonical exponential family of distributions). Since the observations can also be discrete, $p(y^n|y'^n)$ and $L(y'^n|w)$ can also be probability mass functions and the integral in (24) becomes a sum. For the sake of notation, we use the integral notation for both continuous and discrete case, but in the latter case (24) becomes:

$$p(y^n|w)=\sum_{y'\in\mathscr{Y}_m^n}p(y^n|y'^n)L(y'^n|w).$$

Assume we have no prior information about w and we use the model in Theorem 1 to represent our state of ignorance. Since $p(y_1,\dots,y_n|w)$ might not belong to the exponential family of distributions, a question to be addressed is if properties (A3)–(A4) continue to hold also in this case. The answer is in general negative as shown in [8]. In fact, assuming the imperfect observation mechanism (24), the authors prove that, for the Imprecise Beta model (as discussed in the Introduction, the results in [8] are more general), property (A.3) does not hold (no learning from data takes place) and, consequently also (A.4) does not hold (no convergence). In this case, the only way to satisfy (A.3)–(A.4) is to not allow $y_0\to 0,1$; this means that y_0 must vary in $[\varepsilon,1-\varepsilon]$ with $0<\varepsilon<0.5$. That is, (A.3)–(A.4) can be satisfied if and only if (A.2) (prior near-ignorance) does not hold [8]. A similar conclusion is derived in [6] using more general arguments. This has an important consequence, namely that in this case, the amount of imperfection introduced by $p(y^n|y'^n)$ (as long as it is positive) does not matter, we cannot be ignorant a priori without also being vacuous a posteriori.

A further question to be addressed is if this is true for any conjugate model (e.g., Gaussian-Gaussian, Poisson-Gamma etc.), whose likelihood is perturbed as described in (24). In order to prove that, we will use the following results.

Lemma 3. *Consider the prior* $p(w|n_0,y_0)=k(n_0,y_0)\exp(n_0(y_0w-b(w)))$. *For* $y_0\to\sup\mathscr{Y}$ *or* $y_0\to\inf\mathscr{Y}$ *and* $n_0<\infty$, *it holds that* $k(n_0,y_0)\to 0$ *and* $\exp(n_0(y_0w-b(w)))$ *concentrates on the value* w^* *such that* $db(w)/dw|_{w=w^*}=y_0$. ■

This can be proven by using the same arguments in the proof of [1, Cor. 1] (notice that w^* is a maximum of $p(w|n_0,y_0)$).

Lemma 4. *Consider the observational mechanism (24) and assume that:* $p(y^n|y'^n)>0$ *for each* $y^n,y'^n\in\mathscr{Y}_m^n$, $L(y'^n|w)$ *belongs to the exponential family of distributions and* $\mathscr{W}=\mathbb{R}$. *Define* $Lg_n(w)=\ln p(w|n,y^n,y_0,n_0)=\ln(p(y^n|w)p(w|n_0,y_0)/p(y^n))$ *and assume that for any well-defined prior* $p(w|n_0,y_0)$, *with* $0<n_0<\infty$ *and* $y_0\in Int(\mathscr{Y})$, *and for every n there is a strict local maximum* m_n *of* $p(w|n,y^n,y_0,n_0)$ *satisfying:*

$$\frac{dLg_n}{dw}(m_n)=0,\ \sigma_n^2=-\left(\frac{d^2Lg_n}{dw^2}(m_n)\right)^{-1}>0 \tag{25}$$

and that m_n *converges when* $n\to\infty$. *Define* $B_\rho(w^*)=\{w:|w-w^*|<\rho\}$ *and assume also that the posterior satisfies:*

(c1) $\sigma_n^2\to 0$ *for* $n\to\infty$.

(c2) *For any* $\varepsilon>0$ *there exists* $\delta>0$ *and* $\rho>0$ *such that, for any* $n>\delta$ *and* $w\in B_\rho(m_n)$, *it holds that:*

$$1-a(\varepsilon)\leq\frac{\frac{d^2Lg_n}{dw^2}(w)}{\frac{d^2Lg_n}{dw^2}(m_n)}\leq 1+a(\varepsilon), \tag{26}$$

where $a(\varepsilon)>0$ *and tends to zero for* $\varepsilon\to 0$.

(c3) *For any* $\rho>0$

$$\int_{B_\rho(m_n)}p(w|n,y^n,y_0,n_0)\,dw\to 1,\ \text{for } n\to\infty.$$

Let ϕ_n be equal to $(w_n - m_n)/\sigma_n$, with $w_n \sim p(w|n,y^n,y_0,n_0)$. Then, given (c1) and (c2), (c3) is a necessary and sufficient condition for ϕ_n to converge in distribution to ϕ, where $p(\phi) = \mathcal{N}(\phi;0,1)$. ■

The proof of this lemma can be found in [4, Sec. 5.1]. Essentially, Lemma 4 states that, for large n, (c1),(c2) together ensure that inside a small neighborhood of m_n the function $p(w|n,y^n,y_0,n_0)$ becomes highly peaked and behaves as a normal density. Condition (c3) ensures that the probability outside any neighborhood of m_n becomes negligible for $n \to \infty$. Under these conditions, w has an asymptotic posterior limit $\mathcal{N}(w;m_n,\sigma_n^2)$.

Theorem 3. *Assume conditions in Lemma 4 hold*[9] *and that the gambles $g \in \mathcal{G}$ defined in Theorem 1 are integrable w.r.t. $p(w|n,y^n,y_0,n_0)$. Then, for the set of priors $\mathcal{M}$ in Theorem 1, (A.1) and (A.2) are always satisfied, while (A.3) and (A.4) hold if and only if $\mathcal{Y} = \mathbb{R}$ and, thus, $\inf \mathcal{Y} = -\infty$ and $\sup \mathcal{Y} = \infty$.* ■

Proof: *Since coherence and $\mathcal{G}_0$-prior ignorance properties do not depend on the likelihood (for coherence this holds since $p(y^n|w)$ is separately coherent), the fact that (A.1) and (A.2) are still verified is a direct consequence of Theorem 1.*[10] *First we prove the necessity of the conditions of the theorem, by showing that in the case $\inf \mathcal{Y} \neq -\infty$ or $\sup \mathcal{Y} \neq \infty$, (A.3)–(A.4) do not hold. Consider a gamble $g \in \mathcal{G}$ and the posterior $p(w|n,y^n,y_0,n_0)$ obtained in correspondence of the prior $p(w|n_0,y_0)$, which is equal to*

$$p(w|n,y^n,y_0,n_0) = \frac{\int_{\mathcal{Y}^n} p(y^n|y'^n)p(y'^n|w)p(w|n_0,y_0)\,dy'^n dw}{\int_{\mathcal{W}}\int_{\mathcal{Y}^n} p(y^n|y'^n)p(y'^n|w)p(w|n_0,y_0)\,dy'^n dw} \tag{27}$$

and can be rewritten as:

$$\frac{\int_{\mathcal{Y}^n} p(y^n|y'^n)\frac{\prod_{j=1}^n k(y'_j)k(n_0,y_0)}{k(n_p,y'_p)}p(w|n_p,y'_p)\,dy'^n dw}{\int_{\mathcal{W}}\int_{\mathcal{Y}^n} p(y^n|y'^n)\frac{\prod_{j=1}^n k(y'_j)k(n_0,y_0)}{k(n_p,y'_p)}p(w|n_p,y'_p)\,dy'^n dw}. \tag{28}$$

by using the fact that $L(y'^n|w)p(w|n_0,y_0) = p(y'^n)p(w|n_p,y'_p)$, with[11]

$$p(y'^n) = p(y'^n|n_0,y_0) = \prod_{j=1}^n k(y'_j)\frac{k(n_0,y_0)}{k(n_p,y'_p)}, \tag{29}$$

where $n_p = n+n_0$ and $y'_p = (n_0y_0 + \sum_{i=1}^n y'_i)/(n+n_0)$.
Consider the case in which $\mathcal{Y} = [0,1]$ (i.e, $\mathcal{Y}_m = \{0,1\}$ or $\mathcal{Y}_m = [0,1]$). Because of Lemma 3, for $y_0 \to 0$ ($y_0 \to 1$) and $y'_1,\cdots,y'_n \neq 0$ ($y'_1,\cdots,y'_n \neq 1$), it holds that $k(n_0,y_0)/k(n_p,y'_p) \to 0$ and, thus, that $p(y'^n) \to 0$ apart from the case in which $y'_1 = \cdots = y'_n = 0$ ($y'_1 = \cdots = y'_n = 1$) where the ratio $k(n_0,y_0)/k(n_p,y'_p) > 0$. Therefore, for $y_0 \to 0$, $p(y'^n)$ concentrates on $y'_1,\cdots,y'_n = 0$. From Lemma 3, it also follows that $p(w|n_p,y'_p)$ concentrates on $\underline{w}^*$ such that $db(w)/dw|_{w=\underline{w}^*} = 0$ when $y'_p \to 0$. Thus, for any choice of $\varepsilon > 0$, by continuity arguments, it is possible to find a $\underline{y}_0 \in Int(\mathcal{Y})$ and $\delta > 0$ such that

$$\int_{B_\varepsilon(\underline{w}^*)} p(w|n,y^n,y_0,n_0)\,dw > 1-\varepsilon,$$

for any $0 < y_0 \leq \underline{y}_0$ and $n > \delta$.[12] *In other words, for $y_0 \to 0$, the posterior $p(w|n,y^n,y_0,n_0)$ concentrates on $\underline{w}^*$. Similarly, for $y_0 \to 1$, the posterior $p(w|n,y^n,y_0,n_0)$ concentrates on $\overline{w}^*$ such that $db(w)/dw|_{w=\overline{w}^*} = 1$. Under continuity conditions for $g \in \mathcal{G}$ in a neighborhood of $\underline{w}^*$ ($\overline{w}^*$), this implies that, for $y_0 \to 0$ ($y_0 \to 1$), the posterior expectation of g, i.e., $E[g|n,y^n,n_0,y_0]$, concentrates on $g(\underline{w}^*)$ (on $g(\overline{w}^*)$).*[13] *Hence, for the continuous function $g = db(w)/dw$, since $g(w) = y_0$ and, thus, $g(\underline{w}^*) = 0$ and $g(\overline{w}^*) = 1$, it follows that $\underline{E}[g|n,y^n,y_0,n_0] = 0 = \underline{E}[g]$ for $y_0 \to 0$ and $\overline{E}[g|n,y^n,y_0,n_0] = 1 = \overline{E}[g]$ for $y_0 \to 1$, i.e., prior and posterior lower and upper expectations coincide. It can thus be concluded that (A.3) does not hold (no learning from data) and, consequently also (A.4) does not hold (no convergence).*
Consider now the case $\mathcal{Y} = [0,+\infty)$ (or $\mathcal{Y} = (-\infty,0]$), then if $y_0 \to 0$ the ratio $k(n_0,y_0)/k(n_p,y'_p) \to 0$ apart from the case in which also $\hat{y}'_n = 0$, where $y'_p \to 0$ and $k(n_0,y_0)/k(n_p,y'_p) > 0$. Therefore, for the same arguments of the case $\mathcal{Y} = [0,1]$, it follows that for $g = db(w)/dw$, $E[g|n,y^n,n_0,y_0] = g(\underline{w}^*) = 0$. This means that $\underline{E}[g|n,y^n,n_0,y_0] = 0$, it does not matter the value of y^n. Therefore, we conclude that (A.4) does not hold. (A.3) holds for some gambles. For instance, for the gamble $g = db(w)/dw$, (A.3) holds, since the upper expectation differs from its prior value for any $n > 0$ (but the lower expectation is always zero). Hence, the validity of (A.3) depends on the choice of the set $\mathcal{G}$. In particular, if $\mathcal{G}$ includes a function g which gets its infimum and supremum for $y_0 \to 0$ and, respectively, $y_0 \to \lim_{n\to\infty} \hat{y}'_n \neq 0$, then for $n \to \infty$ the prior lower and upper expectations coincide respectively with the posterior lower and upper expectations and, thus, (A.3) does not hold.
Finally assume that $\inf \mathcal{Y} = -\infty$, $\sup \mathcal{Y} = \infty$ and, thus, $\mathcal{Y} = (-\infty,\infty)$. Consider the parameters $n_p = n+n_0$ and $y'_p = (n_0y_0 + \sum_{i=1}^n y'_i)/(n+n_0)$ of the posterior density $p(w|n_p,y'_p)$. Under the conditions of Theorem 1, i.e., $y_0 \in Int(\mathcal{Y})$ and $0 < n_0 < \min(\overline{n}_0, \frac{c}{|y_0|})$, it results that y'_p is bounded as in (14). From this fact it follows that conditions (c1) holds for any $y_0 \in Int(\mathcal{Y})$ and $0 < n_0 < \min(\overline{n}_0, \frac{c}{|y_0|})$ since $y'_p \to y'_n$ for $n \to \infty$. For (c2), by continuity arguments is always possible to find an ε in the definition of (c2), for which (26) is satisfied for any $y_0 \in Int(\mathcal{Y})$ and $0 < n_0 < \min(\overline{n}_0, \frac{c}{|y_0|})$ and, thus, for any prior in $\mathcal{M}$. It is in fact sufficient to consider the largest δ for which (26) holds for any y'_p in (14). This upper δ must exist finite, otherwise Lemma 4 cannot hold. Same considerations hold for (c3). Thus, for any prior in $\mathcal{M}$ satisfying hypotheses of Theorem (1), asymptotic normality holds. Under continuity conditions for $g \in \mathcal{G}$ in a small neighborhood of m_n, this implies that also (A.4) and, consequently, (A.3)

[9] This means that the imperfect observation mechanism still allows asymptotic normality to hold for any prior $p(w|n_0,y_0)$ with fixed $0 < n_0 < \infty$ and $y_0 \in Int(\mathcal{Y})$.

[10] More precisely, from Theorem 1, it can be derived that the likelihood $p(y^n|w)$, the set of priors $\mathcal{M}$ in the exponential family and the corresponding set of posteriors are strongly coherent.

[11] Equation (29) can be derived form (9).

[12] In the case $w^* = -\infty$, $B_\varepsilon(\underline{w}^*)$ must be intended as the open interval, e.g., $(-\infty,\underline{w}-1/\rho)$ for some $\underline{w} \in \mathcal{W}$.

[13] This was also proven in [8, Ths. 11–12].

hold. This proves that $\inf \mathscr{Y} = -\infty$ *and* $\sup \mathscr{Y} = \infty$ *are necessary and sufficient conditions for (A.3)–(A.4).* ■

The theorem states that for a set of Gaussian priors near-ignorance and learning/convergence are compatible even in the case of imperfect observations while this is for instance not the case for a set of Beta priors. The main point is that for the latter, when $\hat{y}'_n = 0$, $y'_p = (n_0 y_0 + n\hat{y}'_n)/(n_0 + n)$ can be made as close as desired to the left boundary of $Int(\mathscr{Y})$ and, thus, from Lemma 3 the posterior $p(w|n'_p, y'_p)$ can be made as closer as desired to a Dirac delta. Thus, in the integration in (27) the only meaningful term is the one relative to the case $\hat{y}'_n = 0$ and, therefore, $p(w|n, \hat{y}_n, n_0, y_0 = 0) = p(w|n'_p, y'_p = 0)$. Conversely, in the Gaussian case, since $|n_0 y_0| < \infty$ it follows that $|y'_p| = |n_0 y_0 + n\hat{y}'_n|/(n_0 + n) = \infty$ only if $|\hat{y}'_n| \to \infty$, but this case must have probability zero otherwise Lemma 4 would not be satisfied. This ensures that $p(w|n, \hat{y}_n, n_0, y_0)$ converges in distribution to $\mathscr{N}(w; m_n, \sigma_n^2)$ for any value of n_0, y_0 in Theorem 1. To better understand the peculiarity of the Gaussian density, assume that $p(y_i|y'_i) = \mathscr{N}(y_i; y'_i, \sigma_r^2)$[14], $L(y'_i|x) = \mathscr{N}(y'_i; x, \sigma^2)$ and consider

$$p(y^n|x) = \int_{y'^n \in \mathscr{Y}^n} \prod_{i=1}^{n} \mathscr{N}(y_i; y'_i, \sigma_r^2)\mathscr{N}(y'_i; x, \sigma^2)\, dy'^n. \tag{30}$$

Since $\mathscr{N}(y_i; y'_i, \sigma_r^2)\mathscr{N}(y'_i; x, \sigma^2)$ is equal to

$$\mathscr{N}(y_i; x, \sigma^2 + \sigma_r^2)\mathscr{N}(y'_i; \sigma_s^2(y_i/\sigma^2 + x/\sigma_r^2), \sigma_s^2),$$

where $\sigma_s^2 = \sigma^2\sigma_r^2/(\sigma^2 + \sigma_r^2)$, (30) becomes $p(y^n|x) = \prod_{i=1}^{n} \mathscr{N}(y_i; x, \sigma^2 + \sigma_r^2)$. Therefore, we can see that in this case the effect of the imperfect observation mechanism is just that of increasing the variance of the measurement noise.

7 Conclusions

This paper has discussed the problem of learning and prior near-ignorance for sets of priors in the one-parameter exponential family. In particular, for conjugate likelihood-prior models in the one-parameter exponential family of distributions, we show that, by letting the parameters of the conjugate exponential prior vary in suitable sets, it is possible to define a set of conjugate priors $\mathscr{M}$ which guarantees prior ignorance without producing vacuous inferences. This result is obtained following both a behavioural and a sensitivity analysis interpretation of prior near-ignorance. We have also discussed the incompatibility of learning and prior near-ignorance for sets of priors in the one-parameter exponential family of distributions in the case of imperfect observations. In particular, we have shown that learning and prior near-ignorance are compatible under an imperfect observation mechanism provided that the support of the priors in $\mathscr{M}$ is the whole real axis. Future work will address the following issues: extension of the model to the multivariate case; extension to more general family of densities.

[14] This satisfies the hypotheses of Theorem 2.

Acknowledgements

This work has been partially supported by the Swiss NSF grants n. 200020-121785/1, 200020-134759/1 and by the Hasler Foundation grant n. 10030.

References

[1] A. Benavoli and M. Zaffalon. A model of prior ignorance for inferences in the one-parameter exponential family. Available at http://www.idsia.ch/∼alessio/TR2011.pdf.

[2] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics, New York, 1985.

[3] J.M. Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *Int. Journal of Approximate Reasoning*, pages 123–150, 2005.

[4] J.M. Bernardo and A.F.M. Smith. *Bayesian theory*. John Wiley & Sons, 1994.

[5] P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *The Annals of statistics*, 7(2):269–281, 1979.

[6] Serafin Moral. Imprecise probabilities for representing ignorance about a parameter. *International Journal of Approximate Reasoning*, In Press, Corrected Proof, 2010.

[7] L.R. Pericchi and P. Walley. Robust Bayesian credible intervals and prior ignorance. *International Statistical Review*, pages 1–23, 1991.

[8] A. Piatti, M. Zaffalon, F. Trojani, and M. Hutter. Limits of learning about a categorical latent variable under prior near-ignorance. *Int. Journal of Approximate Reasoning*, 50(4):597–611, 2009.

[9] E. Quaeghebeur. Learning from samples using coherent lower previsions. PhD thesis, Ghent University, 2009.

[10] E. Quaeghebeur and G. De Cooman. Imprecise probability models for inference in exponential families. In *Proc. of ISIPTA'05*, pages 287–296, 2005.

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

[12] P. Walley. A bounded derivative model for prior ignorance about a real-valued parameter. *Scandinavian Journal of Statistics*, 24(4):463–483, 1997.

[13] G. Walter and T. Augustin. Imprecision and prior-data conflict in generalized Bayesian inference. *Journal of Statistical Theory and Practice*, 3:255–271, 2009.

Dirichlet model *versus* expert knowledge.

Diogo de Carvalho Bezerra
Core Management
University Federal of Pernambuco
Caruaru, Brazil
Email: dicbezerra@hotmail.com

Fernando Menezes Campello de Souza
Electronic and Systems Department
University Federal of Pernambuco
Recife, Brazil
Email: fmcs@hotlink.com.br

Abstract

Decision theory is used to choose a portfolio. Elicitation methods was used based on the utility function and from expert opinion thus, enabling the creation of a utility function for the investor and another for the a priori distribution on economic indicators. The model chosen for an investment portfolio was formulated based on decision theory, incorporating aspects of systematic and unsystematic risk. The model was developed so as to structure an efficient way to understand the application of decision theory in the financial market as well as the application of the Imprecise Dirichlet Model-IDM. The IDM allows the use of imprecise probability. Finally, the IDM was compared to the Markowitz method and also, to the decision model, using only expert opinion, considering an allocation over time to verify which of the three models was the best one. The final conclusion is that expert opinion should not be neglected in her compiling a portfolio.

Keywords. Linear Programming, Elicitation, Portfolio Selection, Financial.

1 Introduction

In the financial market, the portfolio selection problem consists of distributing the total amount available for investment among the financial "products" in the market. Hitherto, the Markowitz portfolio selection procedures, in [10], use ad hoc procedure. One of the numbers used most frequently as a guide, was the average value of the investment payback, usually estimated from past data. The Markowitz procedure is essentially a trade-off between the average and the standard deviation of the (future) payback. It is implemented as a quadratic programming problem: either one minimizes the standard deviation (risk, in the jargon) subject to the constraint that the average must be greater than some previously determined value (usually taken to be zero), or one which maximizes the average payback, subject to an upper bound constraint on the risk. This article suggests using decision theory in the portfolio selection problem. It is divided into five sections. Introduction sets the context and present of the other sections. The second is a brief review of articles related to portfolio selection and imprecise probability. The third presents a decision model that incorporates elements of the economy, such as indicators of economic scenarios that result in the compiling the portfolio. The fourth section presents methods to elicit the utility function and expert knowledge. the measures are used in comparison with the Imprecise Dirichlet Model – IDM. Finally, some conclusions are drawn from the main results.

According to [8] "*Developments in portfolio are stimulated by two basic requirements: (1) adequate modeling of utility functions, risks and constratints; (2) efficiency, i.e., ability to handle large numbers of instruments and scenarios.*" This paper presents a model that satisfies both conditions.

2 A Review of the Literature

Markets in which the price reflects the available information are called efficient markets. The idea of efficient markets is the premise for the Markowitz method. The estimated average return, $\overline{R(A)}$ and the estimated risk $\hat{\sigma}$ of an asset, are expressed by the mathematical expectation of past returns and its standard deviation. The equations below represent the estimate of the expected return and risk of an asset:

$$\overline{R(A)} = \frac{\sum_{t=1}^{n} R_t}{n} \tag{1}$$

$$\hat{\sigma}(A) = \frac{\sum_{t=1}^{n} (R_t - \overline{R(A)})^2}{n-1} \tag{2}$$

The number of observations is represented by n, and

R_t represents the return at time t.
Markowitz method is based on the formation of an asset portfolio so that the risk attributed to each asset can be minimized. This risk is called unsystematic risk. In the Markowitz method, the risk that is not being considered is the market risk, known as systematic risk. Markowitz idea consists is to diversify risk. Thus, the portfolio comprises assets with a negative correlation. Therefore, to the extent that one asset generates losses for the portfolio, another will generate earnings. The average return $R(P)$ and average risk $\sigma(P)$ of a portfolio are expressed by the following equations:

$$R(P) = \sum_{j=1}^{n} R_j W_j \tag{3}$$

$$\sigma(P) = \left[\sum_{i=1}^{n} \sum_{j=1}^{n} W_i W_j \rho_{i,j} \sigma_i \sigma_j \right]^{1/2} \tag{4}$$

where

- The percentage of investment in each asset is W_j;
- σ_j represents the risk of each asset;
- $\rho_{i,j}$ are the coefficients of correlation between the return of two assets.

To obtain the percentage of investment in each asset the nonlinear programming method is used, in which the variables of choice are: the percentages of application. The functional objective is the risk of the Portfolio and the restrictions are quite logical. Given that the percentage of implementation is a probability, it will be positive and the sum of the percentages will be equal to one. The problem is expressed as follows:

$$\min_{W_j} \sigma(P)$$

$$s.a$$

$$\sum W_j = 1 \ , \ W_j \geqslant 0$$

2.1 Probability in Finance Theory

An increasing number of studies are being developed in order to apply imprecise probability to *portfolio* models. At first, the models attempt to introduce the concept of *fuzziness* into the necessary measures for implementing Markowitz model. Examples of *fuzzy* being applied to the development of a *portfolio* are [13], [6] and [2]. Another application of imprecise probability in portfolio management is to seek conditions for separations of the investment fund. In [7] there is an introduction of classical conditions in order to divide funds, and in [12] there is an application subadditive probabilities, where the possibility of inertia in the choice of optimal portfolios is proved. The studies by applying imprecise probability to the economy, but the ideas are going in the direction of finding coherent risk measures and/or price arbitrage of assets.

3 The Decision Model

The model was proposed in [1], which used a simple characterization of the economic scenario, by reducing it to a unique economic indicator $\theta \in [0, 1]$. The observations x (time series data) were modeled in the same way as economic scenarios. For example, if four economic indicators were used, one of them would have 16 scenarios. These scenarios were ordered from worst to best, and an integer number was attached to each of them. The better the scenario, the larger the integer. So, the likelihood function is, for that model, a binomial distribution.

$$P(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

The prior distribution of θ is the Beta distribution

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

3.1 The elements of the problem

The notation is as follows:

$\xi_i = i^{th}$ financial product (i^{th}asset);

a_i = fraction of the available initial capital to be invested in asset ξ_i;

p = net return (payback) of the portfolio , $p \in [-M, M]$, $M > 0$;

GIP_t = gross internal product in period t;

IR_t = inflation rate in period t;

PR_t = prime rate in period t;

UN_t = Unemployment in period t;

The states of nature are defined as follows. First, one defines an intermediate variable ω_i:

Let X_t be an economic indicator, if X_{t+1} is better for the economy than X_t, one then writes $\omega_{t+1} = 1$;

otherwise, $\omega_{t+1} = 0$. Since there are four economic indicators (GIP_t,IR_t, PR_t, UR_t) there will be 16 economic scenarios for each period (one month). Table 1 shows the 16 scenarios which will constitute the states of nature in the decision theory model.

Table 1: The Possible 16 Scenarios.

Scenarios	ω_1	ω_2	ω_3	ω_4
θ_1	0	0	0	0
θ_2	0	0	0	1
θ_3	0	0	1	0
θ_4	0	0	1	1
θ_5	0	1	0	0
θ_6	0	1	0	1
θ_7	0	1	1	0
θ_8	0	1	1	1
θ_9	1	0	0	0
θ_{10}	1	0	0	1
θ_{11}	1	0	1	0
θ_{12}	1	0	1	1
θ_{13}	1	1	0	0
θ_{14}	1	1	0	1
θ_{15}	1	1	1	0
θ_{16}	1	1	1	1

So, $\Theta = \{\theta_1, \theta_2 \dots \theta_{16}\}$.

Scenarios θ_1 and θ_{16} are the worst and the best, respectively, for economy. The remaining ones are not naturally orderable, since the effects they have in the economy will depend upon a series of other characteristics of the specific country. Thus, the θ_js are essentially categorical.

3.2 Data

A time series of the 100 months is available for each of the four economic indicators, as well as for the financial assets to be used in the portfolio. It was thus possible to establish the evolution of the scenarios. These observations, x_j, correspond to a sample of a multinomial probability distribution:

$$P(x|\theta) = \frac{n!}{\prod_{j=1}^{16}(x_j!)}\theta_j^{x_j}$$

Table 2 shows the number of times that each of the scenarios occurred.

3.3 Dirichlet Prior Distribution

To incorporate expert opinion in this model, it is natural to use the conjugate prior distribution of the multinomial, which is the Dirichlet prior. The Dirichlet prior density then is:

Table 2: scenarios occurring.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	4	8	7	4	6	9	4
x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
6	6	4	13	6	5	6	11

$$\gamma(\theta) = \frac{\Gamma(\nu)}{\prod_{j=1}^{16}\Gamma(\alpha_j)}\prod_{j=1}^{16}\theta_j^{\alpha_j - 1}$$

where $\nu = \sum \alpha_j$, $\alpha_j > 0$, $\sum \theta_j = 1$.

The parametrization used in [15] will also be used here:

$$\pi(\theta) = \frac{\Gamma(\nu)}{\prod_{j=1}^{16}\Gamma(st_j)}\prod_{j=1}^{16}\theta_j^{st_j - 1}$$

where $\nu = s\sum t_j$, $\sum \theta_j = 1$, $s > 0$; s is called a hyperparameter.

3.4 Dirichlet Posterior Distribution

When combined, by Bayes rule, with the multinomial likelihood function $P(x|\theta)$, the Dirichlet prior density generates density function a posteriori

$$\pi(\theta|x) = \frac{\Gamma(\upsilon)}{\prod_{j=1}^{k}\Gamma(\alpha_j + x_j)}\prod_{j=1}^{k}\theta_j^{\alpha_j - 1 + x_j},$$

where $\upsilon = \sum \alpha_j + x_j$. The set of all distributions a posteriori is defined by:

$$t^* = \frac{n_j + st_j}{N + s}. \tag{5}$$

3.5 The Action Space

The investment alternatives constitute the action space $A = \{a\}$. Each a is a mix of financial assets, and is a vector of nonnegative numbers that add up to one. The following assets were used:

- Bank Certified Deposit (CDB) (30 days rentability average);

- Gold - percentage of monthly variation;
- Ibovespa - Sao Paulo Stock Monthly Average Growth Rate;
- Financial Assets Fund (FAF) - Accumulated monthly rentability.

Table 3 shows a descriptive statistics of those assets, as well as the result of applying of the Markowitz portfolio (MP) selection procedure. In this procedure, the optimal action corresponds to CDB -0.9638, Gold - 0.0106, Ibovespa - 0.0049 and FAF - 0.0205.

The available series corresponds to the same period as those of the economic indicators, and were obtained from the Brazilian Central Bank.

Table 3: Descriptive Statistics.

Assets	**Mean**	**Min**	**Max**	**Std. Dev.**
CDB	2.06	1.15	5.20	0.91
GOLD	1.43	-16.40	70.00	8.95
IBOVESPA	1.73	-39.55	28.02	11.54
FAF	1.61	-17.98	17.07	5.67
MP	2.04	0.89	5.24	0.92

3.6 The Consequence Function

In [3], the choice of the analytical expression of the consequence function considers some aspects:

- States of nature and actions are merged in the right sense; θ and a work independently of each other, but they are merged to make up the probability distribution of p;
- It represents the behavior which is usually observed in the investment payback: unimodality, and bounded variance and asymmetry; some robustness is desirable, i.e., the persistence of a distribution's characteristic behavior under perturbations in the paraments;
- It should be analytically tractable when in association with the other analytical expressions the decision rule when calculating.

In the portfolio selection model [3] the following consequence function was suggested:

$$f(p|\theta,a) = [M(1+R(a))\theta(1-\theta)]^{-1} \text{ if} \qquad (6)$$

$$M(1+R(a))\left[\frac{\theta}{2}(3+\theta)-1\right]+\mu(a) \leq p \text{ and}$$

$$p \leq M(1+R(a))\left[\frac{\theta}{2}(5-\theta)-1\right]+\mu(a);$$

$$f(p|\theta,a) = 0, \text{ otherwise,}$$

where $a = [a_j]$ is the vector of fractions attributed to each asset; this corresponds to an action; $\mu(a) =$ average value of the portfolio, $R(a) = \left(1 - \frac{\mu(a)}{\mu(a)+\sigma(a)}\right)$ a measure of the risk of the portfolio. It is important to look at the consequence function (equation 6). A closer look will shed some light in the behavior of this function: the larger the value of θ, the better the economy. For $\theta = \frac{1}{2}$ one has:

$$\frac{\theta}{2}(3+\theta)-1 = -\frac{1}{8} \text{ and}$$

$$\frac{\theta}{2}(5-\theta)-1 = \frac{1}{8}.$$

If $\mu(a) = 0$ then one has a uniform distribution between $-(1/4)M$ and $(1/4)M$. For any portfolio, a has a uniform distribution between $-(1/8)M(1+R(a))+\mu(a)$ and $(1/8)M(1+R(a))+\mu(a)$

In this model, the generalization of the consequence function is:

$$f(p|\theta,a) = \left[2M(1+R)\tau\prod\theta_j\right]^{-1} \text{ if}$$

$$M(1+R)\left[\sum n_j\theta_j + \tau\prod\theta_j\right]+\mu \geq p \text{ and}$$

$$M(1+R)\left[\sum n_j\theta_j - \tau\prod\theta_j\right]+\mu \leq p \ ;$$

$$f(p|\theta,a) = 0 \quad \text{otherwise,}$$

where τ is a proportionality constant and n_j represents the impact of each θ_j in the consequence function.

3.7 Loss Function

Consider the quadratic utility function:

$$v(p) = k_0 + k_1 p - k_2 p^2.$$

The loss function is denoted by $L(\theta,a)$. It is defined as:

$$L(\theta,a) = -k_0 - k_1\left[M(1+R)\left[\sum n_j\theta_j\right]+\mu\right]$$
$$+k_2\left[M(1+R)\left[\sum n_j\theta_j\right]+\mu\right]^2 +$$
$$+\frac{1}{3}k_2\left[M(1+R)\tau\prod\theta_j\right]^2$$

3.8 The Bayes risk

To apply of the Bayes rule the following calculations are necessary:

1. $$u(f(p|\theta,a_j)) = \int u(p)f(p|\theta,a_j)dp.$$
2. $$L(\theta,a_j) = -u(f(p|\theta,a_j)).$$
3. $$R_d(\theta) = \sum_x P(x|\theta)L(\theta,d(x)).$$
4. $$r_d = \int_0^1 \pi(\theta)R_d(\theta)d\theta \text{ (Bayes risk).}$$
5. $$r_d = \int_0^1 \left[\sum_x \pi(\theta)P(x|\theta)L(\theta,d(x))\right]d\theta.$$
6. $$r_d = \int_0^1 \left[\sum_x \pi(\theta|x)P(x)L(\theta,d(x))\right]d\theta.$$
7. $$rd = \sum_x P(x)\int_0^1 \pi(\theta|x)L(\theta,d(x))d\theta.$$
8. To minimize r_d by a choice of d, which is the same as to minimize, for each x, the term
$$\int_0^1 \pi(\theta|x)L(\theta,d(x))d\theta,$$
by a choice of $d(x)$.

To facilitate the calculations one denotes $\frac{\Gamma(\nu)}{\prod_{j=1}^k \Gamma(\alpha_j+x_j)} = \omega$

$$r_d = \int_0^1 -\omega[\prod_{i=1}^k \theta_i^{\alpha_i-1+x_i}]\times$$
$$\times[k_0 + k_1\left[M(1+R)\left[\sum n_j\theta_j\right]+\mu\right]$$
$$-k_2\left[M(1+R)\left[\sum n_j\theta_j\right]+\mu\right]^2\times$$
$$\times\frac{1}{3}k_2\left[M(1+R)\tau\prod\theta_j\right]^2 d\theta$$

$$\therefore r_d = -[k_0\omega\int_0^1 \prod\theta_j^{\alpha_j-1+x_j}d\theta +$$

$$-\int_0^1 k_1[M(1+R)\sum n_j\theta_j+\mu]\omega\prod\theta_j^{\alpha_j+x_j-1}d\theta -$$

$$-\int_0^1 k_2[M(1+R)\sum n_j\theta_j+\mu]^2\omega\prod\theta^{\alpha_j-1+x_j}d\theta +$$

$$+\frac{1}{3}k_2\omega M^2(1+R)^2\int_0^1 \tau\prod\theta_j^2\prod\theta_j^{\alpha_j-1+x_j}d\theta$$

One thus obtains the expression of the risk of adopting a decision rule:

$$rd = -\{k_0 + k_1M(1+R)\omega\times$$

$$\sum_{i\neq j}^k\left[\frac{n_j}{(\alpha_j+x_j+1)\Pi(\alpha_i+x_i)}\right]+$$

$$k_1\mu - k_2[M^2(1+R)^2\omega\times$$

$$\left[\sum_{i\neq j}^k\frac{n_j^2}{(\alpha_j+x_j+1)\Pi(\alpha_i+x_i)}\right]+$$

$$2\sum_{j=1;i<j}^k\frac{n_jn_i(\prod\limits_{t:t\neq j\neq i}(\alpha_t+x_t))^{-1}}{(\alpha_j+x_j+1)(\alpha_i+x_i+1)}+$$

$$M(1+R)\omega\sum_{j=1\ :j\neq i}^k\frac{n_j(\prod(\alpha_i+x_i))^{-1}}{(\alpha_j+x_j+1)}+\mu]+$$

$$\frac{1}{3}\tau\omega M^2(1+R)^2k_2\prod_{j=1}^k\frac{1}{(\alpha_j+x_j+1)}\}$$

4 The Expert *Versus* The IDM

4.1 Utility

The elicitation of the utility by the original method developed by Von Neumann and Morgenstern occurs when an individual responds to only one question about the likelihood such that he becomes such individual indifferent between a consequence, P, or about a game with a probability λ to win $\overline{P}$ or $(1-l\lambda)$ to get $\underline{P}$. The questions are put in the form game or lottery. Game layout can also vary depending on operational convenience to applied method.

An elicitation protocol (some questions) was applied to the individual in order for him to declare the value of λ for which he feels indifferent between a certain amount and a game (lottery). It should be noted that there is no "right answer" for each question. However, it is necessary to be careful about obtaining a good insight in order to obtain good accuracy. The answers are individual and must be tailored to the individual psychology of risk. There will never be perfect accuracy; one must not confuse rationality with perfection.

The assumption for use of a von Neumann-Morgenstern weak cardinal utility function is that these are two goods, one of them more desirable, $\overline{P}$, and the other one is less desirable, $\underline{P}$, which assigns two arbitrary utilities. When these values $\overline{P}$ and $\underline{P}$ are distant from each other, it is very difficult to choose the value of λ for a given value P, where $\underline{P} < P < \overline{P}$. Thus, we must ask what is the value of λ which makes P indifferent to a lottery between $\underline{P}$ and $\overline{P}$ in different overlapping limits. Later, as the utility function is an interval measure, λ values must be passed to the same. It is intended to elicit the utility function of money in a range from - R$ 95,000.00 (minus ninety-five thousand reais) to R$ 95,000.00 (ninety-five thousand reais). After the questions, a regression is used to infer the error of the the decision-maker when Like answered the questions. A quadratic function expression was used. In which the were parameters $k_0 = 0.7025$, $k_1 = 0.0047$, and $k_2 = 1.7608 \times 10^{-5}$, for one individual (an investor).

4.2 The Expert

Keynes, at the beginning of his book, *Treatise on Probability*, cites Leibniz, who is already tired of saying that there is a new logic that deals with degrees of probability. Keynes advocates the hypothesis that in the long term, we'll all be dead and that a historical series, that would make predictions about our future, would when Like answered exist. When there are few data or no data, the *a priori* knowledge of the expert should be used. A new elicitation procedure of *a priori* knowledge of the expert was presented in [5] and [11].

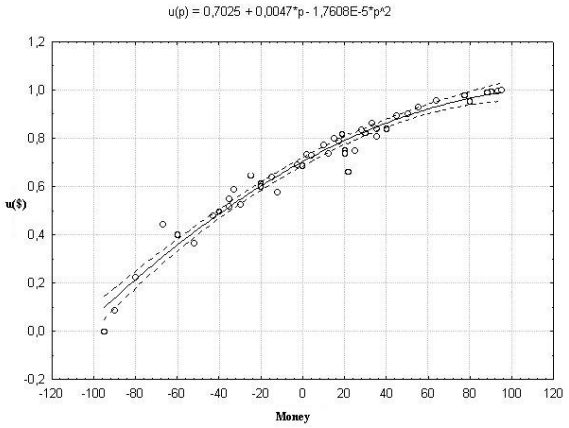


Figure 1: The Decision Maker's utility function.

The method used to elicit of the expert's prior distribution has the basic assumption that the expert has a vague knowledge about the state of nature, θ; It is assumed he can only make a finite amount of comparative probabilistic assertions when answering questions about the likelihood of the event belonging to one of two given ranges, I_A or I_B. For example, the expert will respond if it is more likely theta belongs to $I_A = [\theta_1, \theta_2, \theta_3]$ or $I - B = [\theta_4, \theta_5]$. This method expresses the expert's knowledge using a family of probability. Using this method allows, among other things, to make inferences about facts that cannot be presented by a historical series, but the facts there are and the probability of their occurrence is very high. These events change the decision on whether to invest or not in a financial asset. However, how significant is this change?

The model for solving two linear programming problems, mathematically, is expressed as follows: First of all, it is necessary to solve the maximization problem and later the minimization problem. Both are subject to the same set of constraints. These problems can be expressed as follows:

$$\max(\min)_{\theta_j} \sum_{j=1}^{2n} c_j \theta_j \tag{7}$$

subject to:

$$a_{ik} \sum_{j=i}^{k} \theta_j - a_{lm} \sum_{j=l}^{m} \theta_j \leq b_s \tag{8}$$

$$\alpha_j \theta_j \leq \theta_{j+1}, j = 1, 2, \ldots, 2n-1, \alpha > 0 \quad (9)$$

$$\beta_j \theta_{j+1} \leq \theta_j, j = 1, 2, \ldots, 2n-1, \beta > 0 \quad (10)$$

$$\theta_j \geq 0, j = 1, 2, \ldots, 2n \quad (11)$$

$$\sum_{j=1}^{2n} \theta_j = 1 \quad (12)$$

The values of c_j were randomly determined in order to allow a random search process. The restriction 8 becomes a questions in a questionnaire where the expert must answer them, and depending on the responses the signal may be $\leq$ or $\geq$.

Depending on the combination of parameters a_{ik}, a_{lm} e b_s, the expert's opinion can be collected in various ways. Constraints 9 and 10 are used when one wants to use a *a priori* distribution, so that such distribution may be as informative as possible. Otherwise, a good option is to suppress these restrictions. The remaining restrictions are considered the basic ones acceded to obtain a probability distribution. To obtain the probability distributions from an expert's opinion, he/she must be consistent in his/her responses. If a response is not consistent with all the other ones, the feasible set of restrictions will be empty. The expert must not answer these questions. The expert does not answer the questions when he/she cannot say anything about the fact of the likelihood of θ belonging or not belonging to one of the existing intervals. The questions which the expert does not answer will not enter the constraints of the linear programming problem. Questions will be displayed according to the indicators shown in [5]. The model defines new constructs such as vagueness, precision, concordance, overall vagueness, conflicts, decidability, harmony, quality of inference and amount of information. The elicitation method for linear programming also allows the combination of bodies of evidence.

After analysing a questionnaire referring to the 16 scenarios presented in Table 1 and solving the linear programming problem above, for different values of c_j, the result shown in Table 4 was obtained. This result can be interpreted as a convex set of probabilities within a range with an upper and lower probability for each state of nature. Any combination of values within the ranges can be used as a prior distribution of the expert.

Table 4: Expert opinion.

Scenarios	$\underline{\pi(\theta)}$	$\overline{\pi(\theta)}$
θ_1	0,00%	6,25%
θ_2	0,00%	1,67%
θ_3	5,00%	5,00%
θ_4	4,58%	5,00%
θ_5	1,67%	5,00%
θ_6	1,67%	3,33%
θ_7	3,33%	3,33%
θ_8	0,00%	2,50%
θ_9	3,33%	4,17%
θ_{10}	7,08%	8,33%
θ_{11}	10,83%	12,50%
θ_{12}	3,33%	6,67%
θ_{13}	5,83%	6,67%
θ_{14}	6,67%	7,50%
θ_{15}	7,50%	9,17%
θ_{16}	25,83%	26,25%

4.3 Imprecise Dirichlet Model

One of the hypotheses of the model is the existence of a prior distribution, $\pi(\theta)$. In [15] a way is presented for obtaining posterior distributions without having a prior distribution. This model is known as the imprecise Dirichlet model (IDM). From the set of posterior Dirichlet distributions, one obtains upper and lower probabilities for the event θ_j. The lower probability is obtained by making $t_j \to 0$ and the upper probability is obtained by making $t_j \to 1$ in Equation 5. One will then get:

$$\overline{P}(\theta_j|x) = \frac{n_j + s}{N + s}, \quad \text{and}$$

$$\underline{P}(\theta_j|x) = \frac{n_j}{N + s}.$$

where N is the number of observations about θ. In the example, $N = 100$. As discussed in [15], $s = 1$ corresponds to a frequentist outlook, and $s = 2$ to a cautious Bayesian. Table 5 shows the results that were obtained by the IDM in these two cases.

4.4 Comparisons

The comparison among the three forms of selecting a portfolio is considering the time in which information is available. The criterion of minimizing the maximum risk will be used; This result is obtained by calculating the lowest upper risk. Using the upper posterior distributions and the upper probabilities of the expert, the Monte Carlo method was used to calculate which is the action of lowest and highest risk. During

Table 5: Upper and Lower Probability.

s	$s=1$		$s=2$	
$P(\theta\|x)$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$
$P(\theta_1\|x)$	0.0099	0.0198	0.0098	0.0294
$P(\theta_2\|x)$	0.0396	0.0495	0.0392	0.0588
$P(\theta_3\|x)$	0.0792	0.0891	0.0784	0.0980
$P(\theta_4\|x)$	0.0693	0.0792	0.0686	0.0882
$P(\theta_5\|x)$	0.0396	0.0495	0.0392	0.0588
$P(\theta_6\|x)$	0.0594	0.0693	0.0588	0.0784
$P(\theta_7\|x)$	0.0891	0.0990	0.0882	0.1078
$P(\theta_8\|x)$	0.0396	0.0495	0.0392	0.0588
$P(\theta_9\|x)$	0.0594	0.0693	0.0588	0.0784
$P(\theta_{10}\|x)$	0.0594	0.0693	0.0588	0.0784
$P(\theta_{11}\|x)$	0.0396	0.0495	0.0392	0.0588
$P(\theta_{12}\|x)$	0.1287	0.1386	0.1275	0.1471
$P(\theta_{13}\|x)$	0.0594	0.0693	0.0588	0.0784
$P(\theta_{14}\|x)$	0.0495	0.0594	0.0490	0.0686
$P(\theta_{15}\|x)$	0.0594	0.0693	0.0588	0.0784
$P(\theta_{16}\|x)$	0.1089	0.1188	0.1078	0.1275

the data series, the portfolio with lowest upper risk was calculated while the information was obtained. The return that an investor would obtain over the 100 months by using the Markowitz method for compiling a portfolio was calculated as well. The cumulative return by the IDM during the period analyzed was the following: for $s = 1$ it was 557.71%, for $s = 2$, it was 519%. If the investor had used the Markowitz method the cumulative return would be 754.98%. The return would have been 820 % if the expert's opinion had been used.

5 Conclusions

An interesting point regarding the model is that the formulation is general, broad and flexible. Thus, there is the option of using other analytical expressions. Another more general observation is that the better the economic theory being used in the preparation of constructs, the better the results should be. The main conclusions of this article are the following:

- Subjective aspects can be used such as: the utility of the investor and expert's opinion can be measured and used to guide the decision-making in the financial markets;
- The expert's opinion about uncertain states of the world can be used as a measure of systematic risk. Thus, uncertainty about events like the presidential election, agreements and international wars are measured and incorporated into the problem of choosing the investment;
- The imprecise Dirichlet model presents an important advanced in making the decisions with insufficient information. Besides, this model should be used in problems involving the choice of investment portfolios. It is possible to incorporate the expert's opinion. Moreover, information from these bodies of evidence should be used together, since the result of the application shows that it is not correct to disregard the expert opinion;
- The use of analytical models can lead to theoretical conclusions about the investor's behavior;
- Analytical models are also easy to implement: they can be used in a spreadsheet or a calculator.

The Consequence Function is perfect for the implementation of models based on Condicional Value-at-Risk (CVAR) [14] and [8]. Comparisons between the model presented in this article and CVAR are objects of future works.

Acknowledgements

We express our gratitude to Bruno Tadeu for comments and corrections of the text. Any mistake remaining are the full responsability of the authors.

References

[1] Bezerra, D. C., Wanderley A. L. and Campello de Souza, F. M. *Portfolio Selection Using Imprecise Probabilities.* The 15th INFORMS Applied Probability Society Conference Cornell University, Ithaca, New York, U.S.A. July 12 - 15, 2009.

[2] Bilbao-Terol, A., Prez-Gladish, B., Arenas-Parra, M. and Rodrguez-Ura, M. V. *Fuzzy compromise programming for portfolio selection* Applied Mathematics and Computation 173. 2006. p. 251-264

[3] Campello de Souza, F.M. *Decises Racionais em Situaes de Incerteza.* Recife, PE, Brazil, Vade Mecum Ltda, ISBN-85-7315-178-1, 2007.

[4] Ferreira, R. J. P. and Almeida Filho, A. T. and Campello de Souza, F. M. *A Decision Model For Portifolio Selection* Pesquisa Operacional, v.29, n.2, p.403-417, Maio a Agosto de 2009

[5] Nadler Lins, G. and Campello de Souza, F. M. *A Protocol for the Elicitation of Prior Distributions.* In: ISIPTA'01, Ithaca. 2001. p. 265-272.

[6] Gupta, P. Mehlawat, M. K. and Saxena, A. *Asset portfolio optimization using fuzzy mathematical*

programming Information Sciences 178. 2008. p. 1734-1755.

[7] Hung, Chi-fu and Litzenberger, R. H. *Foundations For Financial Economics* Prentice Hall, 1988.

[8] Krokhmal, P., Palmquist, J. and Uryasev, S. *Portfolio Optimization With Conditional Value-at-Risk Objective and Constraints* The Journal of Risk. 2001/02. v.4, n.2.

[9] Lucas, A. D. P.; Cezario, A.; Bezerra, D. C. *Modelo Impreciso de Dirichelet: Uma Aplicao a Ativos Financeiros.* In: XIII Conferencia Latino-Ibero-Americana de Investigacin de Operaciones, 2006.

[10] Markowitz, H. M. *Portfolio Selection* Journal of Finance, 1952, vol. VII, pp. 77-91.

[11] Silva, A. A. and Campello de Souza, F. M. *A Protocol for the Elicitation of Imprecise Probabilities.* In: ISIPTA'01, Pittsburgh. 2005. p.315-321, 2005.

[12] Simonsen, M. H., Werlang, S. R. C. *Subadditive probabilities and portfolio inertia.* Revista de Econometria, v. 11, p. 1-19, 1991

[13] Smimou, K., Bector, C.R. and Jacoby G. *Portfolio selection subject to Expert' judgments* International Review of Financial Analysis 17. 2008. p. 1036-1054.

[14] Uryasev, S. *Conditional Value-at-Risk: Optimization Algorithms and Applications* Financial Engineering News. February 2000. 14.

[15] Walley, P. *Inferences from Multinomial Data: Learning about a Bag of Marbles* Journal of the Statistical Society, Ser. B, 58, 3-57.

The Description of Least Favorable Pairs in Huber-Strassen Theory, Finite Case

Andrey G. Bronevich
JSC "Research, Development and Planning Institute for Railway Information Technology, Automation and Telecommunication"
brone@mail.ru

Abstract

In this paper we provide the algebraic description of the minmax problem solutions, which are considered in Huber-Strassen theory providing effective algorithms of searching least favorable pairs. This investigation gives also new insights to understanding well-known algorithms for maximizing Shannon entropy and other functionals.

Keywords. 2-monotone capacities, least favorable pairs, Huber-Strassen theory, Kullback–Leibler distance.

1 Introduction

In 1973 Huber and Strassen [13] have published their prodigious paper showing that the optimal test between composite hypotheses described by 2-alternative capacities can be reduced to testing two simple hypotheses described by usual probability measures called a least favorable pair. This result was derived for Polish spaces and supplied with other remarkable results. In particular, the case of 2-alternative capacities cannot be extended for a wider class of coherent upper probabilities, the likelihood ratio does not depend on the chosen least favorable pair, i.e. this likelihood ratio is unique; any pair of probability measures minimizing a functional of a certain type has to be least favorable. In addition, they have shown the way of constructing the optimal test for independent experiments with observations described by 2-alternative capacities. After this famous work, there were some works generalizing Huber-Strassen results by using other topological assumptions [7,17], for special neighborhood models [12,16], or even for more general theories of imprecise probabilities [2,3,11] (see the overview of the results in [3]). However, most of the results are not constructive: they are based on the theorems establishing existence of least favorable pairs without showing how to obtain them. However, for some special neighborhood models there are explicit solutions for finding least favorable pairs (see results obtained by Österreicher [15], Rieder [16], and Bednarski [4]). Augustin [2] proposed a method for finding least favorable pairs for models on finite spaces, based on linear programming.

On the other hand, we can observe that recently developed algorithms [1,14] for computing the maximum entropy functional for 2-monotone capacities are evidently based on recovering the likelihood ratio between 2-monotone capacity and equiprobable probability distribution. But this fact has not been recognized yet.

In the paper we try to get more explicit expressions of Huber-Strassen results for a finite case. This allows us to get the description of all possible least favorable pairs and to construct the algorithm for searching them. This algorithm generalizes the procedure for computing the maximal entropy functional considered in [1].

2 Technical preliminaries

Let X be a finite universal set and $\mathfrak{A} = 2^X$ is the algebra consisting of all subsets of X. A set function $\mu : \mathfrak{A} \to [0,1]$ is called a *monotone measure* [9] or *capacity* [8] 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 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* [18] 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 $\mu \le P$;

4) M_{coh} is the set of all *coherent lower probabilities* [18] 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* [8] 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 $A,B\in\mathfrak{A}$.

For any $\mu\in M_{mon}$ we define the set $core(\mu)=\{P\in M_{pr} \mid P\ge\mu\}$, Clearly $core(\mu)\ne\varnothing$ if $\mu\in M_{low}$. We also remind that $M_{mon}\supset M_{low}\supset M_{coh}\supset M_{2-mon}\supset M_{pr}$. In the sequel we use also the upper probability measures that can be got from lower probability measures using dual relation. A dual monotone measure μ^d of μ is computed by $\mu^d=1-\mu(A^c)$, where $A\in\mathfrak{A}$ and $A^c=X\setminus A$ is the complement of A. Let us remind also that measures, which are dual to coherent lower probabilities, are called coherent upper probabilities [18], and also if $\mu\in M_{2-mon}$ then μ is 2-alternative monotone measure [8], i.e. it characterizes by the following inequality: $\mu^d(A)+\mu^d(B)\ge\mu^d(A\cup B)+\mu^d(A\cap B)$.

3 Huber-Strassen theory, finite case

In this section we consider the Huber-Strassen theory for the finite case: we establish connections between Huber-Strassen theory and canonical sequences of monotone measures [5] and provide an effective algorithm for finding least favorable pairs.

Let us remind that the Huber-Strassen theory solves the problem of the Neymann-Pearson testing between two hypotheses H_0 and H_1 described by 2-monotone measures μ_0 and μ_1 on an algebra $\mathfrak{A}$. We assume here that $\mathfrak{A}$ is the powerset of some nonempty set X. According to this theory the testing problem can be reduced to the classical case, i.e. there exist probability measures (a *least favorable pair*) $P_0\in core(\mu_0)$ and $P_1\in core(\mu_1)$ such that the optimal test for any level of significance can be obtained by using P_0 and P_1. The searching of P_0 and P_1 is closely connected to the following optimization problem:

$$q_{\mu_0^d,\mu_1^d}(t)=\min_{A\in 2^X}\{(1-t)\mu_0^d(A)+t\mu_1^d(A^c)\},$$

where $t\in[0,1]$. Obviously, the value $q_{\mu_0^d,\mu_1^d}(t)$ gives us the exact upper probability of error if we use the Bayesian classifier and the prior probability of H_0 is $(1-t)$ and the prior probability of H_1 is t. Hence, the expression for $q_{\mu_0^d,\mu_1^d}$ can be rewritten as

$$q_{\mu_0^d,\mu_1^d}(t)=\min_{A\in 2^X}\max_{\substack{P_0\in core(\mu_0),\\ P_1\in core(\mu_1)}}(1-t)P_0(A)+tP_1(A^c).$$

This optimization problem can be considered also for coherent lower probabilities, but for this case it is impossible to choose $P_0\in core(\mu_0), P_1\in core(\mu_1)$ for any $\mu_0,\mu_1\in M_{coh}$, such that

$$q_{\mu_0^d,\mu_1^d}(t)=\min_{A\in 2^X}\{(1-t)\mu_0^d(A)+t\mu_1^d(A^c)\}=$$
$$\min_{A\in 2^X}\{(1-t)P_0(A)+tP_1(A^c)\},$$

in general. Let us denote

$$\mathcal{L}_{\mu_0,\mu_1}(t)=\{A\in 2^X \mid (1-t)\mu_0^d(A)+t\mu_1^d(A^c)=q(t)\}$$

for $t\in[0,1]$ and $\mathcal{L}_{\mu_0,\mu_1}=\bigcup_{t\in(0,1)}\mathcal{L}_{\mu_0,\mu_1}(t)$.

We analyze first the properties of $\mathcal{L}_{\mu_0,\mu_1}$. We will show that $\mathcal{L}_{\mu_0,\mu_1}$ is a lattice and measures μ_0^d and μ_1 are additive on it.

Proposition 1. *Let* $\mu_0,\mu_1\in M_{2-mon}$ *and assume that* $A\in\mathcal{L}_{\mu_0,\mu_1}(t)$, $B\in\mathcal{L}_{\mu_0,\mu_1}(s)$, *where* $t\le s$. *Then* $A\cap B\in\mathcal{L}_{\mu_0,\mu_1}(t)$ *and* $A\cup B\in\mathcal{L}_{\mu_0,\mu_1}(s)$. *In addition,* $\mu_0^d(A)=\mu_0^d(A\cap B)$ *and* $\mu_1(B)=\mu_1(A\cup B)$ *if* $t<s$.

Proof. Let $A\in\mathcal{L}_{\mu_0,\mu_1}(t)$, $B\in\mathcal{L}_{\mu_0,\mu_1}(s)$ and $t\le s$. Then

$$q_{\mu_0^d,\mu_1^d}(t)+q_{\mu_0^d,\mu_1^d}(s)=$$
$$(1-t)\mu_0^d(A)+t\mu_1^d(A^c)+(1-s)\mu_0^d(B)+s\mu_1^d(B^c)=$$
$$(1-s)\left(\mu_0^d(A)+\mu_0^d(B)\right)+(s-t)\mu_0^d(A)+$$
$$t\left(\mu_1^d(A^c)+\mu_1^d(B^c)\right)+(s-t)\mu_1^d(B^c).$$

Because μ_0^d,μ_1^d are 2-alternative, we get the following inequality:

$$q_{\mu_0^d,\mu_1^d}(t)+q_{\mu_0^d,\mu_1^d}(s)\ge$$
$$(1-s)\left(\mu_0^d(A\cap B)+\mu_0^d(A\cup B)\right)+(s-t)\mu_0^d(A)+$$
$$t\left(\mu_1^d\left((A\cap B)^c\right)+\mu_1^d\left((A\cup B)^c\right)\right)+(s-t)\mu_1^d(B^c)=$$
$$\left[(1-s)\mu_0^d(A\cup B)+s\mu_1^d\left((A\cup B)^c\right)\right]+$$
$$\left[(1-t)\mu_0^d(A\cap B)+t\mu_1^d\left((A\cap B)^c\right)\right]+$$
$$(s-t)\left[\mu_1^d(B^c)-\mu_1^d\left((A\cup B)^c\right)\right]+$$
$$(s-t)\left[\mu_0^d(A)-\mu_0^d(A\cap B)\right].$$

Since $\mu_1^d(B^c)-\mu_1^d\left((A\cup B)^c\right)\ge 0$, $\mu_0^d(A)-\mu_0^d(A\cap B)\ge 0$, and, by our assumption,

$$(1-s)\mu_0^d(A\cup B)+ s\mu_1^d\left(\left(A\cup B\right)^c\right)\ge q_{\mu_0^d,\mu_1^d}(s),$$

$$(1-t)\mu_0^d(A\cap B)+ t\mu_1^d\left(\left(A\cap B\right)^c\right)\ge q_{\mu_0^d,\mu_1^d}(t),$$

we get that there is the only possibility that

$$(1-s)\mu_0^d(A\cup B)+s\mu_1^d\left(\left(A\cup B\right)^c\right)= q_{\mu_0^d,\mu_1^d}(s),$$

$$(1-t)\mu_0^d(A\cap B)+t\mu_1^d\left(\left(A\cap B\right)^c\right)= q_{\mu_0^d,\mu_1^d}(t),$$

and if $s>t$, then

$$\mu_1^d(B^c)-\mu_1^d\left(\left(A\cup B\right)^c\right)=0,$$

$$\mu_0^d(A)-\mu_0^d(A\cap B)=0,$$

i.e. $A\cap B\in\mathcal{L}_{\mu_0,\mu_1}(t)$ and $A\cup B\in\mathcal{L}_{\mu_0,\mu_1}(s)$.

Corollary 1. *$\mathcal{L}_{\mu_0,\mu_1}$ is a lattice, and monotone measures μ_0^d and μ_1 are additive on $\mathcal{L}_{\mu_0,\mu_1}$.*

Proof. It is clear that $\mathcal{L}_{\mu_0,\mu_1}$ is a lattice. It follows from Proposition 1. Let $A\in\mathcal{L}_{\mu_0,\mu_1}(t)$, $B\in\mathcal{L}_{\mu_0,\mu_1}(s)$ and $t<s$. Then, by Proposition 1, $\mu_0^d(A)=\mu_0^d(A\cap B)$ and $\mu_1(B)=\mu_1\left(A\cup B\right)$. Since μ_0^d is 2-alternative,

$$\mu_0^d(A)+\mu_0^d(B)\ge\mu_0^d(A\cap B)+\mu_0^d(A\cup B),$$

i.e. $\mu_0^d(B)\ge\mu_0^d(A\cup B)$, and this is possible if $\mu_0^d(B)=\mu_0^d(A\cup B)$, i.e.

$$\mu_0^d(A)+\mu_0^d(B)=\mu_0^d(A\cap B)+\mu_0^d(A\cup B).$$

Analogously, since μ_1 is 2-monotone,

$$\mu_1(A)+\mu_1(B)\le\mu_1(A\cap B)+\mu_1(A\cup B),$$

i.e. $\mu_1(A)\le\mu_1(A\cap B)$, and this is possible if $\mu_1(A)=\mu_1(A\cap B)$, i.e.

$$\mu_1(A)+\mu_1(B)=\mu_1(A\cap B)+\mu_1(A\cup B).$$

Consider the case, when $s=t\in(0,1)$. Then

$$(1-t)\left(\mu_0^d(A)+\mu_0^d(B)\right)+t\left(\mu_1^d(A^c)+\mu_1^d(B^c)\right)=$$
$$(1-t)\left(\mu_0^d(A\cap B)+\mu_0^d(A\cup B)\right)+$$
$$t\left(\mu_1^d\left((A\cap B)^c\right)+\mu_1^d\left((A\cup B)^c\right)\right).$$

Because $\mu_0^d(A)+\mu_0^d(B)\ge\mu_0^d(A\cap B)+\mu_0^d(A\cup B)$ and $\mu_1^d(A^c)+\mu_1^d(B^c)\ge\mu_1^d((A\cap B)^c)+\mu_1^d((A\cup B)^c)$, this is possible if $\mu_0^d(A)+\mu_0^d(B)=\mu_0^d(A\cap B)+\mu_0^d(A\cup B)$ and $\mu_1^d(A^c)+\mu_1^d(B^c)=\mu_1^d((A\cap B)^c)+\mu_1^d((A\cup B)^c)$, i.e. monotone measures μ_0^d and μ_1 are additive on $\mathcal{L}_{\mu_0,\mu_1}$.

Further we will consider sublattices $\mathcal{L}$ of $\mathcal{L}_{\mu_0,\mu_1}$ such that $\mathcal{L}\cap\mathcal{L}_{\mu_0,\mu_1}(t)\ne\varnothing$ for any $t\in(0,1)$. The notable examples of these lattices are $\left\{\underline{A}_t\right\}_{t\in(0,1)}$ and $\left\{\overline{A}_t\right\}_{t\in(0,1)}$, where $\underline{A}_t=\bigcap_{A\in\mathcal{L}_{\mu_0,\mu_1}(t)}A$, $\overline{A}_t=\bigcup_{A\in\mathcal{L}_{\mu_0,\mu_1}(t)}A$. Notice that $\underline{A}_t,\overline{A}_t\in\mathcal{L}_{\mu_0,\mu_1}(t)$ by Proposition 1.

Consider now the case, when $\mu_0,\mu_1\in M_{pr}$.

Proposition 2. *Let $P_0,P_1\in M_{pr}$. Then for any $t\in[0,1]$*

$$\mathcal{L}_{P_0,P_1}(t)=\left\{A\in 2^X\mid\underline{A}_t\subseteq A\subseteq\overline{A}_t\right\},$$

where $\underline{A}_t=\left\{x\in X\mid(1-t)P_0\left(\{x\}\right)<tP_1\left(\{x\}\right)\right\}$ *and* $\overline{A}_t=\left\{x\in X\mid(1-t)P_0\left(\{x\}\right)\le tP_1\left(\{x\}\right)\right\}$.

Proof. According to the definition $A\in\mathcal{L}_{P_0,P_1}(t)$ iff it minimizes the value

$$(1-t)P_0(A)+t\left(1-P_1(A)\right)=t+\sum_{x\in A}\left[(1-t)P_0\left(\{x\}\right)-tP_1\left(\{x\}\right)\right],$$

and we get the minimum if $(1-t)P_0\left(\{x\}\right)-tP_1\left(\{x\}\right)\le 0$ for all $x\in A$ and $(1-t)P_0\left(\{x\}\right)-tP_1\left(\{x\}\right)\ge 0$ for all $x\notin A$, and the above condition implies the statement of the proposition.

Remark 1. We can express the statement of Proposition 2 using the likelihood ratio of probability measures P_0 and P_1. For this reason, let us introduce two functions $\underline{\pi}:X\to[0,+\infty]$ and $\overline{\pi}:X\to[0,+\infty]$ by

1) $\underline{\pi}(x)=\overline{\pi}(x)=P_0\left(\{x\}\right)/P_1\left(\{x\}\right)$ if at least one of the values $P_0\left(\{x\}\right)$ and $P_1\left(\{x\}\right)$ is greater than zero (we define $\underline{\pi}(x)=\overline{\pi}(x)=+\infty$ if $P_0\left(\{x\}\right)>0$ and $P_1\left(\{x\}\right)=0$);

2) $\underline{\pi}(x)=0$ and $\overline{\pi}(x)=+\infty$ if $P_0\left(\{x\}\right)=0$ and $P_1\left(\{x\}\right)=0$.

Then

$$\underline{A}_t=\left\{x\in X\mid\overline{\pi}(x)<t/(1-t)\right\}$$

and

$$\overline{A}_t=\left\{x\in X\mid\underline{\pi}(x)\le t/(1-t)\right\}.$$

Our next problem is to find sufficient and necessary conditions, under which $q_{\mu_0^d,\mu_1^d}(t)=q_{P_0,P_1}(t)$, $t\in[0,1]$, for probability measures $P_0\in core\left(\mu_0\right)$ and $P_1\in core\left(\mu_1\right)$. The solution of this problem is presented below.

Lemma 1. *Let $\mu_0,\mu_1\in M_{2-mon}$. Assume also that $P_0\in core\left(\mu_0\right)$ and $P_1\in core\left(\mu_1\right)$ such that*

$q_{\mu_0^d,\mu_1^d}(t) = q_{P_0,P_1}(t)$ *for all* $t \in [0,1]$. *Then* $\mathcal{L}_{\mu_0,\mu_1}(t) \subseteq \mathcal{L}_{P_0,P_1}(t)$ *for all* $t \in [0,1]$.

Proof. Assume that the conditions of the lemma are fulfilled, and $B \in \mathcal{L}_{\mu_0,\mu_1}(t)$. Then

$$q_{\mu_0^d,\mu_1^d}(t) = (1-t)\mu_0^d(B) + t\mu_1^d(B^c) \geq (1-t)P_0(B) + tP_1(B^c).$$

Because $q_{\mu_0^d,\mu_1^d}(t) = q_{P_0,P_1}(t)$ by our assumption, we have

$$q_{P_0,P_1}(t) = (1-t)P_0(B) + tP_1(B^c),$$

i.e. $B \in \mathcal{L}_{P_0,P_1}(t)$. The lemma is proved.

Obviously, there are some cases, when $\mathcal{L}_{\mu_0,\mu_1}(t) = \mathcal{L}_{P_0,P_1}(t)$ for probability measures from Lemma 1. In these cases it is possible to recover the likelihood ratio of (P_0, P_1) as shown in the following lemma.

Lemma 2. *Let the conditions of Lemma 1 be fulfilled and* $\mathcal{L}_{P_0,P_1}(t) = \mathcal{L}_{\mu_0,\mu_1}(t)$ *for all* $t \in [0,1]$. *Then the likelihood ratio of* (P_0, P_1) *is uniquely defined on* X *by*

1) $\underline{\pi}(x) = 0$ *and* $\overline{\pi}(x) = +\infty$ *if* $x \in \overline{A}_0 \setminus \underline{A}_1$;

2) $\underline{\pi}(x) = \overline{\pi}(x) = \sup\left\{t/(1-t) \mid x \in \overline{A}_t, t \in [0,1]\right\}$ *if* $x \in \underline{A}_1$;

3) $\underline{\pi}(x) = \overline{\pi}(x) = +\infty$ *if* $x \in X \setminus \left(\overline{A}_0 \cup \underline{A}_1\right)$

Proof. Let us show that the formulas for $\underline{\pi}$ is valid. We see that $\overline{A}_0 = \left\{x \in X \mid P_0(\{x\}) = 0\right\}$ and $\underline{A}_1 = \left\{x \in X \mid P_1(\{x\}) > 0\right\}$. Therefore, formulas 1) and 3) are valid. Let $t \in (0,1)$, then

$$\left\{x \in X \mid \underline{\pi}(x) = t/(1-t)\right\} = \bigcap_{t>s}\left(\overline{A}_t \setminus \overline{A}_s\right),$$

i.e. the formula 2) is also valid. The lemma is proved.

Remark 2. Let us notice that functions $\underline{\pi}(x)$ and $\overline{\pi}(x)$, considered in Lemma 2 can be computed for every pair of 2-monotone measures $\mu_0, \mu_1 \in M_{2-mon}$. We call these functions as in Huber-Strassen theory, a likelihood ratio of 2-monotone measures μ_0, μ_1.

Lemma 3. *Let* $\mu_0, \mu_1 \in M_{2-mon}$ *and let* $P_0 \in core(\mu_0)$ *and* $P_1 \in core(\mu_1)$ *be such that* $q_{\mu_0^d,\mu_1^d}(t) = q_{P_0,P_1}(t)$ *for all* $t \in [0,1]$. *Then the likelihood ratio of* (P_0, P_1) *is equal to the likelihood ratio of* (μ_0, μ_1) *in all points, where at least* $P_0(\{x\}) > 0$ *or* $P_1(\{x\}) > 0$.

Proof. For these points the likelihood ratio of (P_0, P_1) can be computed by

$$\underline{\pi}(x) = \overline{\pi}(x) = \sup\left\{t/(1-t) \mid x \in \overline{B}_t, t \in [0,1]\right\},$$

where $\overline{B}_t$ is the maximal element of $\mathcal{L}_{P_0,P_1}(t)$. Consider also maximal elements $\overline{A}_t$ of lattices $\mathcal{L}_{\mu_0,\mu_1}(t)$, $t \in [0,1]$. Because $P_0\left(\overline{B}_t\right) = P_0\left(\overline{A}_t\right)$, $P_0\left(\overline{B}_t\right) = P_0\left(\overline{A}_t\right)$, and $\overline{A}_t \subseteq \overline{B}_t$ by Lemma 1 and Proposition 2, we find that $P_0(\{x\}) = P_1(\{x\}) = 0$ for all $x \in \overline{B}_t \setminus \overline{A}_t$. Therefore,

$$\underline{\pi}(x) = \overline{\pi}(x) = \sup\left\{t/(1-t) \mid x \in \overline{A}_t, t \in [0,1]\right\}.$$

The proposition is proved.

Proposition 3. *Let* $\mu_0, \mu_1 \in M_{2-mon}$, *and let* $\underline{A}_t$ *be the minimal elements of* $\mathcal{L}_{\mu_0,\mu_1}(t)$, $t \in [0,1]$. *Assume also that* $P_0 \in core(\mu_0)$ *and* $P_1 \in core(\mu_1)$. *Then* $q_{\mu_0^d,\mu_1^d}(t) = q_{P_0,P_1}(t)$ *for all* $t \in [0,1]$ *iff*

1) $P_0(\underline{A}_t) = \mu_0^d(\underline{A}_t)$, $P_1(\underline{A}_t) = \mu_1(\underline{A}_t)$ *for all* $t \in [0,1]$;

2) *the likelihood ratio of* (P_0, P_1) *is equal to the likelihood ratio of* (μ_0, μ_1) *in all points, where at least* $P_0(\{x\}) > 0$ *or* $P_1(\{x\}) > 0$.

Proof. The necessary statement of the proposition follows from Lemma 1 and Lemma 3. Let us show sufficiency. Let $\underline{\pi}(x)$ and $\overline{\pi}(x)$ define the likelihood ratio of (P_0, P_1), then the maximal and minimal elements of $\mathcal{L}_{P_0,P_1}(t)$ are defined by

$$\underline{B}_t = \left\{x \in X \mid \overline{\pi}(x) < t/(1-t)\right\},$$

$$\overline{B}_t = \left\{x \in X \mid \underline{\pi}(x) \leq t/(1-t)\right\}.$$

Obviously, 2) implies that $\underline{B}_t \subseteq \underline{A}_t \subseteq \overline{B}_t$, i.e. $\underline{A}_t \in \mathcal{L}_{P_0,P_1}(t)$ by Proposition 2, and

$$q_{P_0,P_1}(t) = (1-t)P_0(\underline{A}_t) + t(1 - P_1(\underline{A}_t) =$$

$$(1-t)\mu_0^d(\underline{A}_t) + t(1 - \mu_1(\underline{A}_t)) = q_{\mu_0^d,\mu_1^d}(t).$$

The proposition is proved.

The existence of probability measures, considered in Proposition 3, is shown in the next proposition.

Proposition 4. *Let* $\mu_0, \mu_1 \in M_{2-mon}$, $\overline{A}_t$ *and* $\underline{A}_t$ *be maximal and minimal elements of* $\mathcal{L}_{\mu_0,\mu_1}(t)$, *respectively. Then there are* $P_0 \in core(\mu_0)$ *and* $P_1 \in core(\mu_1)$ *such that*

1) $P_0(\underline{A}_t) = \mu_0^d(\underline{A}_t)$, $P_1(\underline{A}_t) = \mu_1(\underline{A}_t)$ *for all* $t \in [0,1]$;

2) $q_{\mu_0^d,\mu_1^d}(t) = q_{P_0,P_1}(t)$ *for all* $t \in [0,1]$.

Proof. Let functions $\underline{\pi}(x)$ and $\overline{\pi}(x)$ define the likelihood ratio of (μ_0,μ_1). Because we consider the finite case, there is an increasing sequence $\{t_1,t_2,...,t_{m-1}\}$, such that $0<t_1<t_2<...<t_{m-1}=1$ and $\underline{\pi}(x)=t_k/(1-t_k)$ if $x\in \underline{A}_{t_{k+1}}\setminus \underline{A}_{t_k}$, $k=1,...,m-2$. Let us denote $B_k=\underline{A}_{t_k}$, $k=1,...,m-1$, $B_m=\left(X\setminus \overline{A}_0\right)\cup \underline{A}_1$. Observe that $\underline{\pi}(x)=\overline{\pi}(x)=t_k/(1-t_k)$ if $x\in B_{k+1}\setminus B_k$, $k=1,...,m-1$, and also $\underline{\pi}(x)=\overline{\pi}(x)=0$ if $x\in B_1$, and $\underline{\pi}(x)=0$ and $\overline{\pi}(x)=+\infty$ if $x\in X\setminus B_m$. According to the condition 1), we should find a pair of probability measures (P_0,P_1) satisfying the conditions:

$$P_0\in\left\{P\in M_{pr}\mid P\le \mu_0^d, P(B_k)=\mu_0^d(B_k), k=1,...,m\right\}, \quad (1)$$

$$P_1\in\left\{P\in M_{pr}\mid P\ge \mu_1, P(B_k)=\mu_1(B_k), k=1,...,m\right\}. \quad (2)$$

These families of probability measures can be obviously described by canonical sequences of monotone measures[1]. For this purpose, let us consider a sequence of sets $\Gamma=\{B_k\}_{k=1}^m$ and limit measures $(\mu_0^d)_\Gamma$ and $(\mu_1)_\Gamma$ generated by canonical sequences of monotone measures by the sequence Γ. Then obviously, the conditions (1) and (2) are equivalent to $P_0\in\left\{P\in M_{pr}\mid P\le (\mu_0^d)_\Gamma\right\}$ and $P_1\in\left\{P\in M_{pr}\mid P\ge (\mu_1)_\Gamma\right\}$.

Because $\Gamma=\{B_k\}_{k=1}^m$ is an increasing sequence, we can use the explicit expressions for limit measures $(\mu_0^d)_\Gamma$ and $(\mu_1)_\Gamma$ as follows:

$$(\mu_0^d)_\Gamma(A)=\sum_{k=1}^{m+1}\left[\mu_0^d((A\cap B_k)\cup B_{k-1})-\mu_0^d(B_{k-1})\right],$$

$$(\mu_1)_\Gamma(A)=\sum_{k=1}^{m+1}\left[\mu_1((A\cap B_k)\cup B_{k-1})-\mu_1(B_{k-1})\right],$$

where $B_0=\varnothing$ and $B_{m+1}=X$. Let us analyze what kind of additional conditions for a pair of probability measures P_0 and P_1 should be fulfilled. According to condition 2) of Proposition 3 a pair (P_0,P_1) should have the same likelihood ratio as (μ_0,μ_1) in all points, where at least $P_0(\{x\})>0$ or $P_1(\{x\})>0$. In other words, P_0 and P_1 have the constant positive likelihood ratio on sets $B_k\setminus B_{k-1}$, $k=2,...,m-1$, excluding points, where $P_0(\{x\})=0$ or $P_1(\{x\})=0$. This means that conditional probability measures $(P_0)_{B_k\setminus B_{k-1}}$ and $(P_1)_{B_k\setminus B_{k-1}}$ defined by

$$(P_0)_{B_k\setminus B_{k-1}}(A)=\frac{P_0((A\cap B_k)\cup B_{k-1})-P_0(B_{k-1})}{P_0(B_k)-P_0(B_{k-1})},$$

$$(P_1)_{B_k\setminus B_{k-1}}(A)=\frac{P_1((A\cap B_k)\cup B_{k-1})-P_1(B_{k-1})}{P_1(B_k)-P_1(B_{k-1})}$$

are the same. Introduce also into consideration monotone measures

$$(\mu_0^d)_{B_k\setminus B_{k-1}}(A)=\frac{\mu_0^d((A\cap B_k)\cup B_{k-1})-\mu_0^d(B_{k-1})}{\mu_0^d(B_k)-\mu_0^d(B_{k-1})},$$

$$(\mu_1)_{B_k\setminus B_{k-1}}(A)=\frac{\mu_1((A\cap B_k)\cup B_{k-1})-\mu_1(B_{k-1})}{\mu_1(B_k)-\mu_1(B_{k-1})}.$$

Assume that measures $(\mu_0^d)_{B_k\setminus B_{k-1}}$, $(\mu_1)_{B_k\setminus B_{k-1}}$ are defined for $k=1,...,m$ if the corresponding divisor $\mu_0^d(B_k)-\mu_0^d(B_{k-1})$ or $\mu_1(B_k)-\mu_1(B_{k-1})$ is not equal to zero. Then using expressions for $(\mu_0^d)_\Gamma$ and $(\mu_1)_\Gamma$, we have $(P_0)_{B_k\setminus B_{k-1}}\le(\mu_0^d)_{B_k\setminus B_{k-1}}$ and $(P_1)_{B_k\setminus B_{k-1}}\ge(\mu_1)_{B_k\setminus B_{k-1}}$, or combining with $(P_0)_{B_k\setminus B_{k-1}}$ and $(P_1)_{B_k\setminus B_{k-1}}$, we get

$$(\mu_1)_{B_k\setminus B_{k-1}}\le(P_1)_{B_k\setminus B_{k-1}}=(P_0)_{B_k\setminus B_{k-1}}\le(\mu_0^d)_{B_k\setminus B_{k-1}},$$

where $k=2,...,m-1$. To prove that a probability measure $P\in M_{pr}$ with $(\mu_1)_{B_k\setminus B_{k-1}}\le P\le(\mu_0^d)_{B_k\setminus B_{k-1}}$ exists, let us show first that the inequality $(\mu_1)_{B_k\setminus B_{k-1}}\le(\mu_0^d)_{B_k\setminus B_{k-1}}$ holds, i.e.

$$\frac{\mu_1(A\cup B_{k-1})-\mu_1(B_{k-1})}{\mu_1(B_k)-\mu_1(B_{k-1})}\le\frac{\mu_0^d(A\cup B_{k-1})-\mu_0^d(B_{k-1})}{\mu_0^d(B_k)-\mu_0^d(B_{k-1})} \quad (3)$$

for any $A\subseteq B_k\setminus B_{k-1}$. Let us notice that the choice of sets B_k implies that

$$(1-t_{k-1})\mu_0^d(B_{k-1})+t_{k-1}(1-\mu_1(B_{k-1}))=$$
$$(1-t_{k-1})\mu_0^d(B_k)+t_{k-1}(1-\mu_1(B_k)).$$

Or equivalently,

$$\frac{\mu_0^d(B_k)-\mu_0^d(B_{k-1})}{\mu_1(B_k)-\mu_1(B_{k-1})}=\frac{t_{k-1}}{1-t_{k-1}}.$$

Therefore, if the inequality (3) is not fulfilled for some $A\subseteq B_k\setminus B_{k-1}$, then

$$\frac{\mu_0^d(A\cup B_{k-1})-\mu_0^d(B_{k-1})}{\mu_1(A\cup B_{k-1})-\mu_1(B_{k-1})}<\frac{t_{k-1}}{1-t_{k-1}},$$

but this contradicts to the choice of B_{k-1}, because in this case

$$(1-t_{k-1})\mu_0^d(B_{k-1})-\mu_1(B_{k-1})>$$
$$(1-t_{k-1})\mu_0^d(A\cup B_{k-1})-\mu_1(A\cup B_{k-1}).$$

[1] A reader can find the main results on canonical sequences of monotone measures in [5] and a brief description of them in [6].

Because $(\mu_1)_{B_k \setminus B_{k-1}}$ is 2-monotone and $(\mu_0^d)_{B_k \setminus B_{k-1}}$ is 2-alternative, the existence of P follows from [10, Lemma 4.3].

It remains to show how to define values of probability measures $P_0(\{x\})$ and $P_1(\{x\})$ if $x \in B_1$, $x \in B_m \setminus B_{m-1}$, or $x \in X \setminus B_m$. (It is worth to keep in mind that it is often $B_1 = \varnothing$ or $B_m \setminus B_{m-1} = \varnothing$.)

If $x \in B_1$, then $P_0(\{x\}) = 0$ and values $P_1(\{x\})$ should be chosen such that $(\mu_1)_{B_1} \le (P_1)_{B_1}$.

If $x \in B_m \setminus B_{m-1}$, then $P_1(\{x\}) = 0$ and values $P_0(\{x\})$ should be chosen such that $(P_0)_{B_m \setminus B_{m-1}} \le (\mu_0^d)_{B_m \setminus B_{m-1}}$.

If $x \in X \setminus B_m$, then $P_0(\{x\}) = 0$ and $P_1(\{x\}) = 0$.

Let us notice that the constructed pair of probability measures (P_0, P_1) has the same likelihood ratio as (μ_0, μ_1) in all points, where at least $P_0(\{x\}) > 0$ or $P_1(\{x\}) > 0$. It means that $q_{\mu_0^d, \mu_1^d}(t) = q_{P_0, P_1}(t)$ for all $t \in [0,1]$ by Proposition 3. The proposition is proved.

Remark 3. Let us notice that Proposition 4 establishes the existence of a least favorable pair, because the optimal test for any level of significance can be obtained by probability measures P_0 and P_1, considered in Proposition 4. In some sense, a least favorable pair (P_0, P_1) gives an approximation of (μ_0, μ_1) and its exactness depends on the chosen (P_0, P_1). The exact approximation should give the best approximation of sets $\mathcal{L}_{\mu_0, \mu_1}(t)$, $t \in [0,1]$ in a sense that the cardinality of $\mathcal{L}_{P_0, P_1}(t) \setminus \mathcal{L}_{\mu_0, \mu_1}(t)$ should be minimal. It happens if $\mathcal{L}_{P_0, P_1}(t) = \{A \mid \underline{A}_t \subseteq A \subseteq \overline{A}_t\}$ for all $t \in [0,1]$, where $\overline{A}_t$ and $\underline{A}_t$ are maximal and minimal elements of $\mathcal{L}_{\mu_0, \mu_1}(t)$, respectively. In this case the likelihood ratios of (P_0, P_1) and (μ_0, μ_1) coincide.

Corollary 2. *Let $\mu_0, \mu_1 \in M_{2-mon}$ and we use the notation from Proposition 4 and its proof. Then every least favorable pair of probability measures P_0 and P_1 can be represented as*

$$P_0 = \sum_{k=2}^{m+1} \left(\mu_0^d(B_k) - \mu_0^d(B_{k-1})\right)(P_0)_{B_k \setminus B_{k-1}},$$

$$P_1 = \sum_{k=1}^{m} \left(\mu_1(B_k) - \mu_1(B_{k-1})\right)(P_1)_{B_k \setminus B_{k-1}},$$

where conditional probability measures satisfy the following inequalities:

$$(\mu_1)_{B_1} \le (P_1)_{B_1};$$

$$(\mu_1)_{B_k \setminus B_{k-1}} \le (P_1)_{B_k \setminus B_{k-1}} = (P_0)_{B_k \setminus B_{k-1}} \le (\mu_0^d)_{B_k \setminus B_{k-1}},$$

$$k = 2, \ldots, m-1;$$

$$(P_0)_{B_m \setminus B_{m-1}} \le (\mu_0^d)_{B_m \setminus B_{m-1}}.$$

The algorithm for searching sets B_k, $k = 2, \ldots, m-1$, is based on the following lemma.

Lemma 4. *The construction of sets B_k, $k = 1, 2, \ldots, m$, can be based on the following:*

a) B_1 is the set with the smallest cardinality such that

$$\mu_1(B_1) = \max\{\mu_1(B) \mid \mu_0^d(B) = 0\};$$

b) Let us assume that sets $B_0 = \varnothing, B_1, \ldots, B_{k-1}$, $k \ge 2$, have been constructed. Then if $\mu_1(B_{k-1}) < 1$ the next B_k should be chosen from the set Ω of possible solutions of the following optimization problem

$$\min_{B \left| \substack{B_{k-1} \subset B \\ \mu_1(B) > \mu_1(B_{k-1})} \right.} \frac{\mu_0^d(B) - \mu_0^d(B_{k-1})}{\mu_1(B) - \mu_1(B_{k-1})}.$$

If the set Ω is not singleton, then we should choose set B_k with the smallest cardinality such that

$$\mu_1(B_k) = \max_{B \in \Omega} \mu_1(B).$$

c) the set B_m ($\mu_1(B_{m-1}) = 1$) is the set with the smallest cardinality from the family

$$\{B \in \mathfrak{A} \mid B \supseteq B_{m-1}, \mu_0^d(B) = 1\}.$$

The above conditions define sets B_k, $k = 1, 2, \ldots, m$, uniquely.

Proof. The conditions a) and c) follow easily from the definition of sets B_1 and B_m. Let us show that b) is true. We prove first that the set B_k should be chosen from Ω. Assume to the contrary $B_k \notin \Omega$. Then there is a $B \in 2^X$ such that

$$\frac{\mu_0^d(B_k) - \mu_0^d(B_{k-1})}{\mu_1(B_k) - \mu_1(B_{k-1})} > \frac{\mu_0^d(B) - \mu_0^d(B_{k-1})}{\mu_1(B) - \mu_1(B_{k-1})}.$$

Let us notice that $\frac{\mu_0^d(B_k) - \mu_0^d(B_{k-1})}{\mu_1(B_k) - \mu_1(B_{k-1})} = \frac{t_{k-1}}{1 - t_{k-1}}$, i.e. we can rewrite the last inequality as

$$\frac{\mu_0^d(B) - \mu_0^d(B_{k-1})}{\mu_1(B) - \mu_1(B_{k-1})} < \frac{t_{k-1}}{1 - t_{k-1}},$$

or

$$(1 - t_{k-1})\mu_0^d(B) - t_{k-1}\mu_1(B) <$$

$$(1 - t_{k-1})\mu_0^d(B_{k-1}) - t_{k-1}\mu_1(B_{k-1}).$$

However, the last inequality contradicts to the choice of $B_{k-1} = \underline{A}_{t_{k-1}}$.

Let us show that we should choose set $B \in \Omega$ with the maximal value $\mu_1(B) - \mu_1(B_{k-1})$ and with the smallest cardinality. Assume that $B'_k, B''_k \in \Omega$. Then

$$\frac{\mu_0^d(B'_k) - \mu_0^d(B_{k-1})}{\mu_1(B'_k) - \mu_1(B_{k-1})} = \frac{\mu_0^d(B''_k) - \mu_0^d(B_{k-1})}{\mu_1(B''_k) - \mu_1(B_{k-1})} = \frac{t_{k-1}}{1-t_{k-1}},$$

and, obviously, $B'_k, B''_k \in \mathcal{L}_{\mu_0,\mu_1}(t_{k-1})$.

Let us check the sign of the following difference:

$$\Delta = (1-t_k)\mu_0^d(B'_k) + t_k(1-\mu_1(B'_k)) -$$
$$\left[(1-t_k)\mu_0^d(B''_k) - t_k(1-\mu_1(B''_k))\right] =$$
$$(1-t_k)\left(\mu_0^d(B'_k) - \mu_0^d(B_k)\right) - t_k\left(\mu_1(B'_k) - \mu_1(B_k)\right) -$$
$$\left[(1-t_k)\left(\mu_0^d(B''_k) - \mu_0^d(B_k)\right) - t_k\left(\mu_1(B''_k) - \mu_1(B_k)\right)\right].$$

Assuming that $\mu_1(B'_k) - \mu_1(B_k) > 0$, we get

$$\Delta = \left[\frac{(1-t_k)t_{k-1}}{1-t_{k-1}} - t_k\right]\left(\mu_1(B'_k) - \mu_1(B''_k)\right),$$

i.e. $\Delta < 0$, and we should choose the set B in Ω with the largest value $\mu_1(B)$ and, of course, with the smallest cardinality, because this follows from the definition of the set $\underline{A}_{t_k}$. The set B_k is defined uniquely by above conditions, because Ω coincides with the lattice $\mathcal{L}_{\mu_0,\mu_1}(t_{k-1})$. The lemma is proved.

Example 1. Consider 2-monotone measures μ_0 and μ_1 defined on the algebra 2^X, where $X = \{x_1, x_2, x_3, x_4\}$, with values given in Table 1. Let us describe the algorithm proposed for searching favorable pairs of probability measures if the first hypothesis is described by μ_0, and the second hypothesis is described by μ_1. Applying the algorithm we get $B_1 = \varnothing$; the set B_1 should be chosen to minimize the value $\frac{\mu_0^d(B)}{\mu_1(B)}$. Clearly, $B_2 = \{x_1\}$ and $\frac{\mu_0^d(B_2)}{\mu_1(B_2)} = \frac{0.02}{0.3} = \frac{1}{15}$. Then minimizing the value $\frac{\mu_0^d(B) - \mu_0^d(B_2)}{\mu_1(B) - \mu_1(B_2)}$ for $B \supset B_2$, we get that $B_3 = \{x_1, x_3\}$ and $\frac{\mu_0^d(B_3) - \mu_0^d(B_2)}{\mu_1(B_3) - \mu_1(B_2)} = \frac{0.3 - 0.02}{0.7 - 0.3} = \frac{7}{10}$. By analogy, $B_4 = X$ and $\frac{\mu_0^d(B_4) - \mu_0^d(B_3)}{\mu_1(B_4) - \mu_1(B_3)} = \frac{1 - 0.3}{1 - 0.7} = \frac{7}{3}$. Now it is easy to calculate the likelihood ratio. In our case, $\underline{\pi}(x) = \overline{\pi}(x) = \pi(x)$, and $\pi(x_1) = 1/15$; $\pi(x_3) = 7/10$, and $\pi(x_2) = \pi(x_4) = 7/3$. We see that in our case $(P_0)_{B_k \setminus B_{k-1}}$, $(P_1)_{B_k \setminus B_{k-1}}$, $k = 2,3$, are Dirac measures. For searching $(P_0)_{B_4 \setminus B_3}$, $(P_1)_{B_4 \setminus B_3}$, we need to solve the following inequalities:

x_1	x_2	x_3	x_4	μ_0	μ_0^d	μ_1
0	0	0	0	0	0	0
1	0	0	0	0	0.02	0.3
0	1	0	0	0	0.2	0
1	1	0	0	0	0.2	0.3
0	0	1	0	0	0.3	0
1	0	1	0	0	0.3	0.7
0	1	1	0	0	0.45	0
1	1	1	0	0	0.45	0.7
0	0	0	1	0.55	1	0
1	0	0	1	0.55	1	0.3
0	1	0	1	0.7	1	0
1	1	0	1	0.7	1	0.6
0	0	1	1	0.8	1	0
1	0	1	1	0.8	1	0.7
0	1	1	1	0.98	1	0
1	1	1	1	1	1	1

Table 1: Values of monotone measures.

$$(\mu_1)_{B_4 \setminus B_3} \le (P_1)_{B_4 \setminus B_3} = (P_0)_{B_4 \setminus B_3} \le (\mu_0^d)_{B_4 \setminus B_3}.$$

This system of inequalities can be rewritten as

$$\begin{cases} 0 \le (P_0)_{B_4 \setminus B_3}(\{x_2\}) \le \dfrac{15}{70}, \\ 0 \le (P_0)_{B_4 \setminus B_3}(\{x_4\}) \le 1, \\ (P_0)_{B_4 \setminus B_3}(\{x_2\}) + (P_0)_{B_4 \setminus B_3}(\{x_4\}) = 1. \end{cases}$$

It is easy to find any solution of this system of inequalities can be represented as

$$\begin{pmatrix} (P_0)_{B_4 \setminus B_3}(\{x_2\}) \\ (P_0)_{B_4 \setminus B_3}(\{x_4\}) \end{pmatrix} = a\begin{pmatrix} 0 \\ 1 \end{pmatrix} + (1-a)\begin{pmatrix} 15/70 \\ 55/70 \end{pmatrix},$$

where $a \in [0,1]$. Thus, we have the following solution for favorable pairs:

$$\begin{pmatrix} P_0(\{x_1\}) \\ P_0(\{x_2\}) \\ P_0(\{x_3\}) \\ P_0(\{x_4\}) \end{pmatrix} = a\begin{pmatrix} 0.02 \\ 0 \\ 0.28 \\ 0.7 \end{pmatrix} + (1-a)\begin{pmatrix} 0.02 \\ 0.15 \\ 0.28 \\ 0.55 \end{pmatrix},$$

$$\begin{pmatrix} P_1(\{x_1\}) \\ P_1(\{x_2\}) \\ P_1(\{x_3\}) \\ P_1(\{x_4\}) \end{pmatrix} = a\begin{pmatrix} 0.3 \\ 0 \\ 0.4 \\ 0.3 \end{pmatrix} + (1-a)\begin{pmatrix} 0.3 \\ 45/700 \\ 0.4 \\ 165/700 \end{pmatrix},$$

where $a \in [0,1]$. Let us notice that if $a \in [0,1)$ the least favorable pair (P_0, P_1) has the same likelihood ratio as (μ_0, μ_1), but $P_0(\{x_2\}) = P_1(\{x_2\}) = 0$, when $a = 1$, this pair does not give us sufficient information how to construct the optimal test, when we observe the outcome x_2.

For illustration, we can also calculate sets $\underline{A}_t$:

$$\underline{A}_t = B_1 = \varnothing \text{ if } \frac{t}{1-t} \le \frac{1}{15} \text{ (or } 0 \le t \le \frac{1}{16}\text{)};$$

$$\underline{A}_t = B_2 \text{ if } \frac{1}{16} < t \le \frac{7}{17};$$

$$\underline{A}_t = B_3 \text{ if } \frac{7}{17} < t \le \frac{7}{10};$$

$$\underline{A}_t = B_4 = X \text{ if } \frac{7}{10} < t \le 1.$$

4 Characterization of least favorable pairs by functionals

The next theorem is the analog of the result proved by Huber and Strassen [13]. Its proof gives us also some corollaries that are useful for computing functionals using least favorable pairs.

Theorem 1. *Let $\mu_0, \mu_1 \in M_{2-mon}$ and let Φ be any twice continuously differentiable function on $[0,1]$, such that $\Phi'' > 0$. Then the pair $(Q_0, Q_1) \in core(\mu_0) \times core(\mu_1)$ minimizes the functional*

$$H(P_0, P_1) = \int_X \Phi\left(\frac{dP_0}{dP_0 + dP_1}\right) d(P_0 + P_1)$$

among all $(P_0, P_1) \in core(\mu_0) \times core(\mu_1)$ iff $q_{P_0,P_1}(t) \le q_{Q_0,Q_1}(t)$ for all $t \in [0,1]$.

Proof. Let us denote by $y = \frac{\underline{\pi}(x)}{\underline{\pi}(x)+1} = \frac{dP_0}{dP_0 + dP_1}$, where $\pi(x)$ is the likelihood ratio of probability measures P_0 and P_1. We derive first the explicit expression for $q_{P_0,P_1}(t)$ using the result formulated in Proposition 2:

$$q_{P_0,P_1}(t) = (1-t)P_0(\overline{A}_t) + t\left(1 - P_1(\overline{A}_t)\right) =$$

$$(1-t)\int_{y \le t} dP_0 + t\left(1 - \int_{y \le t} dP_1\right) =$$

$$(1-t)\int_{y \le t} \left(\frac{dP_0}{dP_0 + dP_1}\right) d(P_0 + P_1) -$$

$$t\int_{y \le t} \left(\frac{dP_1}{dP_0 + dP_1}\right) d(P_0 + P_1) + t =$$

$$(1-t)\int_{y \le t} y\, d(P_0 + P_1) - t\int_{y \le t} (1-y) d(P_0 + P_1) + t =$$

$$\int_{y \le t} (y - t) d(P_0 + P_1) + t.$$

Therefore, $q_{P_0,P_1}(t) \le q_{Q_0,Q_1}(t)$ for all $t \in [0,1]$ iff

$$\int_{\frac{dQ_0}{dQ_0+dQ_1} \le t} \left[\frac{dQ_0}{dQ_0 + dQ_1} - t\right] d(Q_0 + Q_1) \ge$$

$$\int_{\frac{dP_0}{dP_0+dP_1} \le t} \left[\frac{dP_0}{dP_0 + dP_1} - t\right] d(P_0 + P_1) \qquad (4)$$

for all $t \in [0,1]$. Introduce into consideration an arbitrary positive integrable function φ on $[0,1]$. Then the condition (4) can be equivalently transformed to

$$\int_0^1 \varphi(t) \left(\int_{\frac{dQ_0}{dQ_0+dQ_1} \le t} \left[t - \frac{dQ_0}{dQ_0 + dQ_1}\right] d(Q_0 + Q_1) \right) dt \le$$

$$\int_0^1 \varphi(t) \left(\int_{\frac{dP_0}{dP_0+dP_1} \le t} \left[t - \frac{dP_0}{dP_0 + dP_1}\right] d(P_0 + P_1) \right) dt.$$

Let us denote

$$H(P_0, P_1) = \int_0^1 \varphi(t) \left(\int_{y \le t} (t - y) d(P_0 + P_1) \right) dt.$$

We transform next the functional H to the form, which is used in the theorem. After changing the order of integration, we get

$$H(P_0, P_1) = \int_X \left(\int_y^1 \varphi(t)(t - y)\, dt \right) d(P_0 + P_1).$$

Let $\Phi(y) = \int_y^1 \varphi(t)(y - t)\, dt$. Then $\Phi'(y) = -\int_y^1 \varphi(t) dt$ and $\Phi''(y) = \varphi(y) > 0$. Notice that for this function Φ: $\Phi(1) = 0$ and $\Phi'(1) = 0$. Let $\Phi_1(y) = \Phi(y) + by + a$, where $a, b \in \mathbb{R}$. Observe that this is the general possible choice of twice differentiable function with $\Phi'' = \varphi > 0$. Then

$$\int_X \Phi_1\left(\frac{dP_0}{dP_0 + dP_1}\right) d(P_0 + P_1) = H(P_0, P_1) +$$

$$b\int_X dP_0 + a\int_X d(P_0 + P_1) = H(P_0, P_1) + 2a + b.$$

The theorem is proved.

Corollary 3. *Let us use assumptions and notations from Theorem 1. Then*

$$H(P_0,P_1)=\int_X \Phi\left(\frac{dP_0}{dP_0+dP_1}\right)d(P_0+P_1)=$$
$$\int_0^1 \left(t-q_{P_0,P_1}(t)\right)\Phi''(t)dt-\Phi'(1)-\Phi(1).$$

Obviously we can use Theorem 1 for solving optimization problems using least favorable pairs. This result is formulated below.

Corollary 4. *Let* $\mu_0,\mu_1\in M_{2-mon}$ *and let* $\Phi:[1]\to(-\infty,+\infty]$ *be any twice continuously differentiable function on* $(0,1)$, *such that* $\Phi''(y)\ge 0$ *for all* $y\in(0,1)$; *in addition* $\Phi(0)=\lim_{y\to+0}\Phi(y)$ *and* $\Phi(1)=\lim_{y\to 1-0}\Phi(y)$. *Then any least favorable pair* $(Q_0,Q_1)\in core(\mu_0)\times core(\mu_1)$ *minimizes the functional*

$$H(P_0,P_1)=\int_X \Phi\left(\frac{dP_0}{dP_0+dP_1}\right)d(P_0+P_1)$$

among all $(P_0,P_1)\in core(\mu_0)\times core(\mu_1)$.

Proof. Let Φ be any twice continuously differentiable function on $[0,1]$ such that $\Phi''(y)\ge 0$ for all $y\in[0,1]$. Then this result obviously follows from Corollary 3, namely, from the formula:

$$H(P_0,P_1)=\int_0^1 \left(t-q_{P_0,P_1}(t)\right)\varphi(t)dt+C,$$

If $(Q_0,Q_1),(P_0,P_1)\in core(\mu_0)\times core(\mu_1)$ and (Q_0,Q_1) is a least favorable pair, then $q_{P_0,P_1}(t)\le q_{Q_0,Q_1}(t)$ for all $t\in[0,1]$ and obviously $H(P_0,P_1)\ge H(Q_0,Q_1)$. If (P_0,P_1) is also a favorable pair, then $q_{P_0,P_1}(t)=q_{Q_0,Q_1}(t)$ for all $t\in[0,1]$ and $H(P_0,P_1)=H(Q_0,Q_1)$.

Consider now the general case, formulated in the corollary. For any $\varepsilon>0$ introduce the functional

$$H_\varepsilon(P_0,P_1)=\int_X \Phi_\varepsilon(y)\,d(P_0+P_1),$$

where

1) $\Phi_\varepsilon(y)=\Phi(\varepsilon)+\Phi'(\varepsilon)(y-\varepsilon)+0.5\Phi''(\varepsilon)(y-\varepsilon)^2$ if $y\in[0,\varepsilon]$;

2) $\Phi_\varepsilon(y)=\Phi(y)$ if $y\in(\varepsilon,1-\varepsilon)$;

3) $\Phi_\varepsilon(y)=\Phi(1-\varepsilon)+\Phi'(1-\varepsilon)(y+\varepsilon-1)+$
$0.5\Phi''(1-\varepsilon)(y+\varepsilon-1)^2$ if $y\in[1-\varepsilon,1]$.

Then Φ_ε is a twice continuously differentiable function on $[0,1]$ such that $\Phi_\varepsilon''(y)\ge 0$ for all $y\in[0,1]$. This implies that $H_\varepsilon(Q_0,Q_1)\le H_\varepsilon(P_0,P_1)$ if $(Q_0,Q_1),(P_0,P_1)\in core(\mu_0)\times core(\mu_1)$ and (Q_0,Q_1) is a least favorable pair. Clearly that

$$H(Q_0,Q_1)=\lim_{\varepsilon\to 0}H_\varepsilon(Q_0,Q_1)\le$$
$$\lim_{\varepsilon\to 0}H_\varepsilon(P_0,P_1)=H(P_0,P_1).$$

The corollary is proved.

Because the value $H(P_0,P_1)$ does not depend on a chosen favorable pair $(P_0,P_1)\in core(\mu_0)\times core(\mu_1)$, it can be expressed through the values of measures ${\mu_0}^d$ and μ_1 on the chain $\{B_0,B_1,...,B_m\}$. This result is given in the next corollary.

Corollary 5. *Assume that we use notations from Corollary 2 and* $(P_0,P_1)\in core(\mu_0)\times core(\mu_1)$ *be a least favorable pair. Let* $\nu={\mu_0}^d+\mu_1$. *Then*

$$H(P_0,P_1)=\sum_{k=1}^m \Phi\left(\frac{{\mu_0}^d(B_k)-{\mu_0}^d(B_{k-1})}{\nu(B_k)-\nu(B_{k-1})}\right)\left(\nu(B_k)-\nu(B_{k-1})\right).$$

Proof. Notice that $P_0(\{x\})=P_1(\{x\})=0$ if $x\in X\setminus B_m$. Therefore,

$$H(P_0,P_1)=\sum_{x\in B_m}\Phi\left(\frac{P_0(\{x\})}{P_0(\{x\})+P_1(\{x\})}\right)\left(P_0(\{x\})+P_1(\{x\})\right),$$

and by Corollary 2 $\frac{P_0(\{x\})}{P_0(\{x\})+P_1(\{x\})}=$
$\frac{{\mu_0}^d(B_k)-{\mu_0}^d(B_{k-1})}{{\mu_0}^d(B_k)-{\mu_0}^d(B_{k-1})+\mu_1(B_k)-\mu_1(B_{k-1})}$ if $x\in B_k\setminus B_{k-1}$ and $P_0(\{x\})+P_1(\{x\})\ne 0$; $P_0(B_k\setminus B_{k-1})={\mu_0}^d(B_k)-{\mu_0}^d(B_{k-1})$ and $P_1(B_k\setminus B_{k-1})=\mu_1(B_k)-\mu_1(B_{k-1})$. Hence, the formula in the corollary is true.

Example 2. The Kullback–Leibler divergence (distance) between probablity measures P_0 and P_1 is defined as

$$D_{KL}(P_1,P_0)=\int_X \ln\left(\frac{dP_1}{dP_0}\right)dP_1.$$

In applications, we need to minimize $D_{KL}(P_1,P_0)$ if $(P_0,P_1)\in core(\mu_0)\times core(\mu_1)$. In particular, if $X=\{x_1,...,x_n\}$ and $P_0(\{x_i\})=1/n$, then

$$D_{KL}(P_1,P_0)=\sum_{i=1}^n \ln\left(P_1\{x_i\}\right)P_1\{x_i\}+\ln(n)=$$
$$-S(P_1)+\ln(n),$$

where $S(P_1)=-\sum_{i=1}^n \ln\left(P_1\{x_i\}\right)P_1\{x_i\}$ is the Shannon entropy. Let us transform the functional D_{KL} to the form used in Theorem 1.

$$D_{KL}(P_1,P_0)=\int_X \ln\left(\frac{dP_1}{dP_0}\right)\frac{dP_1}{dP_1+dP_0}d\left(P_1+P_0\right).$$

Let $y=\dfrac{dP_0}{dP_0+dP_1}$. Then $D_{KL}(P_1,P_0)=\int_X \Phi(y)d\left(P_1+P_0\right)$, where $\Phi(y)=(1-y)\ln\left(\dfrac{1-y}{y}\right)$. Notice that in this case $\varphi(y)=\Phi''(y)=\dfrac{1}{y^2(1-y)}\geq 0$ for all $y\in(0,1)$, i.e. by Corollary 4 any least favorable pair $(Q_0,Q_1)\in core(\mu_0)\times core(\mu_1)$ minimizes the functional $D_{KL}(P_1,P_0)$ among all $(P_0,P_1)\in core(\mu_0)\times core(\mu_1)$. It is remarkable, that we can use the algorithm for finding least favorable pairs in the problem of maximizing the Shannon entropy functional. In this case, we get explicitly the same algorithm proposed firstly for belief measures [14] and then justified for 2-monotone measures [1].

5 Concluding remarks

This work gives a new look on Huber-Strassen results, presented here in the explicit form. Some of them are even strengthened (see Proposition 4 and Corollary 2) or clarified (see Theorem 1 and Corollaries 4-6). As a result we have an effective algorithm for searching least favorable pairs and also the way for minimizing functionals on 2-monotone measures described in Theorem 1 and its corollaries. As shown in [5], it is possible to generalize canonical sequences of 2-monotone measures generated by any chain of sets. This generalization can be useful for describing least favorable pairs for the general case of 2-monotone measures.

Acknowledgements

I am very grateful to the referees for many very helpful and detailed remarks. I express also my sincere thanks to Russian Foundation of Basic Research for support by grants 10-07-00135-a, 10-07-00478-a, 11-07-00591-a, and 11-07-08026-з.

References

[1] J. Abellán, S. Moral. An algorithm to compute the upper entropy for order-2 capacities. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems* 14: 141 – 154, 2006.

[2] T. Augustin. *Optimale Tests bei Intervallwahrscheinlichkeit.* Vandenhoeck & Ruprecht, Göttingen, 1998.

[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. Bednarski. On solutions of minimax test problems for special capacities. *Z. Wahrsch. Verw. Gebiete* 58: 397–405, 1981.

[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, T. Augustin. Approximation of coherent lower probabilities by 2-monotone measures. In Proc. of the 6th International Symposium on Imprecise Probability: Theories and Their Applications, Durham, United Kingdom, 2009, 9 pp.

[7] A. Buja. On the Huber-Strassen theorem. *Probability Theory and Related Fields* 73: 149 – 152, 1986.

[8] G. Choquet. Theory of capacities. *Ann. Inst. Fourier 5*: 131 – 295, 1954.

[9] D. Denneberg. Non-additive Measure and Integral. Dordrecht, Kluwer, 1997.

[10] D. Denneberg. Conditional expectation for monotone measures, the discrete case. *Journal of Mathematical Economics* 37: 105 – 121, 2002.

[11] R. Hable. Data-based decisions under imprecise probability and least favorable models. *International Journal of Approximate Reasoning* 50: 642 – 654, 2009.

[12] R. Hafner. Construction of Minimax-tests for Bounded Families of Probabilities-densities. *Metrika* 40:1 – 23, 1993.

[13] P. J. Huber, V. Strassen. Minimax tests and the Neyman-Pearson lemma for capacities. *Ann. Statist.* 1: 251–263, 1973.

[14] G. J. Klir. *Uncertainty and Information: Foundations of Generalized Information Theory*. Hoboken, NJ: Wiley-Interscience, 2006.

[15] F. Österreicher. On the construction of least favorable pairs of distributions. *Z. Wahrsch. Verw. Gebiete* 43: 49–55, 1978.

[16] H. Rieder. Least favorable pairs for special capacities. *Ann. Statist.* 5: 909 – 921, 1977.

[17] H. Schwarte, J.S. Sadowsky. Revisiting the Huber-Strassen minimax theorem for capacities. In Proc. of IEEE International Symposium on Information Theory, Whistler, BC, Canada, 1995, 8 pp.

[18] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman & Hall, London, 1991.

Comparing Binary and Standard Probability Trees in Credal Networks Inference

Andrés Cano, Manuel Gómez-Olmedo, Andrés R. Masegosa, Serafín Moral
Department of Computer Science and Artificial Intelligence
University of Granada, Spain
{acu,mgomez,andrew,smc}@decsai.ugr.es

Abstract

This paper proposes the use of Binary Probability Trees in the propagation of credal networks. Standard and binary probability trees are suitable data structures for representing potentials because they allow to control the accuracy of inference algorithms by means of a threshold parameter. The choice of this threshold is a trade-off between accuracy and computing time. Binary trees enable the representation of finer-grained independences than probability trees. This leads to more efficient algorithms for credal networks with variables with more than two states. The paper shows experiments comparing binary and standard probability trees in order to demonstrate their performance.

Keywords. Bayesian and Credal networks, Inference algorithms, Imprecise probabilities, Variable elimination, Probability trees

1 Introduction

A Bayesian network (BN) is a probabilistic graphical model where precise assessments for the conditional probability mass functions of the variables in the network given the values of their parents are used. A credal network (CN) is also a graphical structure (a directed acyclic graph (DAG) [13]) which is similar to a BN [17], but now the conditional mass functions belong to convex sets of mass functions (*credal sets*).

There has been an increasing interest in propagation algorithms for CNs in the last years. Different algorithms have been proposed for propagation in CNs using standard probability trees (SPTs) [11, 10, 7]. In this paper we propose to apply binary probability trees (BPTs) [8] to propagate in CNs with the variable elimination (VE) algorihtm.

The remainder of this paper is organized as follows: In Section 2 we introduce Bayesian and credal networks, and the problem of inference on them. Section 3 explains the use of standard and binary probability trees to obtain compact representations of potentials and presents how they can be approximated by pruning them. Section 4 explains how to use the VE algorithm to propagate in CNs using BPTs. Section 5 provides details of the experimental work. Finally, Section 6 gives the conclusions.

2 Inference in Credal Networks

Bayesian and credal networks are based on a set of random variables $\mathbf{X} = \{X_1, \ldots, X_n\}$ and a directed acyclic graph (DAG) $\mathcal{G}$, whose nodes are associated with the variables of $\mathbf{X}$. Let us assume that each variable X_i takes values on a finite set of states Ω_{X_i} (the domain of X_i). We shall use x_i to denote one of the values of X_i, $x_i \in \Omega_{X_i}$. If I is a set of indices, we shall write $\mathbf{X}_I$ for the set $\{X_i | i \in I\}$. The Cartesian product $\times_{i \in I} \Omega_{X_i}$ will be denoted by $\Omega_{\mathbf{X}_I}$. The elements of $\Omega_{\mathbf{X}_I}$ are called configurations of $\mathbf{X}_I$ (represented as $\mathbf{x}_I$). We use $|\Omega|$ to denote the cardinality of a set Ω. We denote by $\mathbf{x}_I^{\downarrow \mathbf{X}_J}$ the *projection* of the configuration $\mathbf{x}_I$ to the set of variables $\mathbf{X}_J$, $\mathbf{X}_J \subseteq \mathbf{X}_I$. We denote by Π_i the set of parents of X_i in $\mathcal{G}$ and $\pi_i \in \Omega_{\Pi_i}$ a configuration for the variables in Π_i. $P(X_i)$ is the mass function for X_i and $P(x_i)$ the probability that $X_i = x_i$. $P(X_i|\pi_i)$ denotes the probability mass function for X_i conditional on $\Pi_i = \pi_i$. A mapping from a set $\Omega_{\mathbf{X}_I}$ into $\mathbb{R}_0^+$ will be called a *potential* p for $\mathbf{X}_I$. The process of inference in probabilistic graphical models requires the definition of two operations on potentials: *combination* $p_1 \otimes p_2$ and *marginalization* $p^{\downarrow \mathbf{X}_J}$. If p_1 and p_2 are potentials for $\mathbf{X}_I$ and $\mathbf{X}_J$ respectively then $p_1 \otimes p_2$ is a potential for $\mathbf{X}_{I \cup J}$ that can be obtained by pointwise multiplication. If p is a potential for $\mathbf{X}_I$, and $J \subseteq I$ then $p^{\downarrow \mathbf{X}_J}$ is a potential for $\mathbf{X}_J$ that can be obtained by summing out all the variables not in X_J.

In a BN, each node labelled with a variable X_i has attached a conditional probability distribution

$P(X_i|\Pi_i)$, that defines a conditional mass function $P(X_i|\pi_i)$ for X_i given each $\pi_i \in \Omega_{\Pi_i}$. A BN determines the following joint probability distribution:

$$P(\mathbf{x}) = \prod_{i=1}^{n} P(x_i|\pi_i) \quad \forall \mathbf{x} \in \Omega_{\mathbf{X}} \tag{1}$$

where x_i and π_i are the projections of $\mathbf{x}$ to X_i and Π_i respectively. Let be $\mathbf{E} \subset \mathbf{X}$ the set of observed variables and $\mathbf{e} \in \Omega_{\mathbf{E}}$ the instantiated value. Each observation, $X_i = e_i$, can be represented by means of a Dirac function defined as $\delta_{X_i}(x_i; e_i) = 1$ if $e_i = x_i$, $x_i \in \Omega_{X_i}$, and $\delta_{X_i}(x_i; e_i) = 0$ if $e_i \neq x_i$. An algorithm that computes the a posteriori distribution $P(x_q|\mathbf{e})$ for each $x_q \in \Omega_{X_q}$, $X_q \in \mathbf{X} \setminus \mathbf{E}$, ($X_q$ is a queried variable) by making local computations is called a *propagation algorithm*. This distribution verifies:

$$P(x_q|\mathbf{e}) \propto \sum_{\mathbf{X}_R} \prod_{X_i \in \mathbf{X}} P(x_i|\pi_i) \prod_{X_i \in \mathbf{E}} \delta_{X_i}(x_i; e_i) \tag{2}$$

where $\mathbf{X}_R = \mathbf{X} \setminus \{\{X_q\}, \mathbf{E}\}$. In fact, the previous formula is the expression for $P(x_q, \mathbf{e})$. $P(x_q|\mathbf{e})$ can be obtained from $P(x_q, \mathbf{e})$ by normalization.

CNs relax the precise probability assessments of BNs. In this work we suppose that the conditional mass functions of a CN are required to belong to a credal set defined as follows. A credal set for a variable X_i is a convex, closed set of probability distributions and shall be denoted by $K(X_i)$. We assume that every credal set has a finite number of extreme points (also called *vertices*), although it may contain an infinite number of mass functions. A credal set can be identified by enumerating its vertices.

An *extensive conditional credal set* [14] about X_i given the set of parent variables Π_i will be a closed, convex set $K(X_i|\Pi_i)$ of mappings $P : X_i \times \Pi_i \longrightarrow [0, 1]$, verifying $\sum_{x_i \in \Omega_{X_i}} P(x_i, \pi_\mathbf{i}) = 1$, $\forall \pi_i \in \Omega_{\Pi_i}$. Again, an extensive conditional credal set can be determined by its set of extreme points which we assume to be finite: $\text{Ext}[K(X_i|\Pi_i)] = \{P_1, \ldots, P_l\}$. In a CN each variable is associated with an extensive conditional credal set $K(X_i|\Pi_i)$. In this paper, we suppose that a *local* credal set $K(X_i|\Pi_i = \pi_i)$ is given for each π_i of Π_i. This is described by Rocha and Cozman [19] as *separately specified credal sets*. For example Fig. 1 shows a CN with two variables (X and Y). Conditional information for X is given by two separately specified credal sets ($K(X|Y = y_1)$ and $K(X|Y = y_2)$). From the separately specified credal sets, we obtain the extensive conditional credal set with:

$$K(X_i|\Pi_i) = \{P|P(x_i, \pi_i) \in K(X_i|\Pi_i = \pi_i), \quad \forall \pi_i \in \Omega_{\Pi_i}\} \tag{3}$$

$K(Y)$	y_1	y_2
r_1	0.1	0.9
r_2	0.2	0.8

$K(X\|Y = y_1)$	x_1	x_2
p_1	0.2	0.8
p_2	0.3	0.7

$K(X\|Y = y_2)$	x_1	x_2
q_1	0.4	0.6
q_2	0.6	0.4

Figure 1: A simple credal network

Table 1 shows the extensive conditional credal set $K(X|Y)$ obtained from the separately specified credal sets $K(X|Y = y_1)$ and $K(X|Y = y_2)$ of Figure 1.

$K(X\|Y)$	x_1, y_1	x_2, y_1	x_1, y_2	x_2, y_2
p_1, q_1	0.2	0.8	0.4	0.6
p_1, q_2	0.2	0.8	0.6	0.4
p_2, q_1	0.3	0.7	0.4	0.6
p_2, q_2	0.3	0.7	0.6	0.4

Table 1: An extensive conditional credal set

As in BNs, the topology $\mathcal{G}$, of a CN represents independence relations between variables using the d-separation criterion. The meaning of such independences depends on which concept of independence for credal sets is adopted. This paper uses the concept of *strong independence* [13, 12]. The *strong extension* $K(\mathbf{X})$ of a CN is the largest joint credal set such that every variable is strongly independent [13, 12] of its nondescendants nonparents given its parents. It is the joint credal set that contains every possible combination of vertices for all credal sets in the network, where the vertices are combined by multiplication as in Expression 1 [13]. That is, the strong extension $K(\mathbf{X})$ of the CN is the convex hull (CH) of the collection of joint mass functions that can be obtained with every possible combination of the vertices of the separately specified credal sets $K(X_i|\pi_i)$:

$$K(\mathbf{X}) = \text{CH}\{P(\mathbf{X}) : P(\mathbf{x}) = \prod_{i=1}^{n} P(x_i|\pi_i), \quad \forall \mathbf{x} \in \Omega_{\mathbf{X}}, \forall \pi_i \in \Omega_{\Pi_i},\ P(X_i|\pi_i) \in K(X_i|\pi_i)\} \tag{4}$$

A CN can be regarded as a collection of BNs [1] where the topology is given by $\mathcal{G}$. The joint probability of each BN is defined by one of the vertices of $K(\mathbf{X})$. So, the CN defines the following collection of joint probabilities:

$$\mathbf{P}(\mathbf{X}) = \{P_k(\mathbf{X})\}_{k=1}^{n_v} \tag{5}$$

where n_v is the number of vertices in $Ext[K(\mathbf{X})]$.

This paper is dedicated to inference in the strong extension of a CN, in particular, to the computation of

tight bounds for the probability values of a queried variable X_q given a set of observed variables $\mathbf{E}$.

The *combination* of two credal sets is the convex hull of the set obtained by multiplying a mapping of the first credal set with a mapping of the second credal set (repeating the probabilistic combination for all pairs of vertices of the two credal sets). The *marginalization* of a credal set is defined by marginalizing each mapping of the credal set. A more detailed description of these operations can be found for example in [9]. With these operations, we can carry out the same propagation algorithms as in the probabilistic case.

$K(\mathbf{X})$ can also be defined as the multiplication (combination) of all the (extensive) conditional credal sets $K(X_i|\Pi_i)$ in the credal network:

$$K(\mathbf{X}) = \prod_{i=1}^{n} K(X_i|\Pi_i) \tag{6}$$

The computation of the a posteriori credal set $K(X_q|\mathbf{E})$ for a queried variable X_q given some evidence $\mathbf{E}$ can be done in similar way as in Bayesian networks (expression 2) by calculating $K(X_q, \mathbf{E})$.

$$K(X_q, \mathbf{E}) = (K(\mathbf{X}) \prod_{X_i \in \mathbf{E}} \delta_{X_i}(x_i; e_i))^{\downarrow X_q} \tag{7}$$

The vertices in $K(X_q, \mathbf{E})$ are mappings from Ω_{X_q} in $[0, 1]$. $K(X_q|\mathbf{E})$ can be calculated by normalizing the vertices in $K(X_q, \mathbf{E})$. If $Ext[K(X_q, \mathbf{E})] = \{P_k(X_q)\}_{k=1}^{n_v}$ is the set of vertices of $K(X_q, \mathbf{E})$, then the computation of tight bounds for the a posteriori probabilities of X_q given the evidence $\mathbf{E}$ can be done with:

$$\begin{aligned} \underline{P}(x_q|\mathbf{e}) &= \min_{k=1,\ldots,n_v} \frac{P_k(x_q)}{\sum_{x_q} P_k(x_q)} \\ \overline{P}(x_q|\mathbf{e}) &= \max_{k=1,\ldots,n_v} \frac{P_k(x_q)}{\sum_{x_q} P_k(x_q)} \end{aligned} \tag{8}$$

Exact computation in CNs has a high complexity [5], much more than in BNs. It could be done by propagating in the n_v BNs defined by the CN.

3 Standard and Binary Trees

Probability trees [20] and binary probability trees [8] have been used as flexible data structures that enables the specification of *context-specific independences* (see [4]) and provides exact or approximate representations of probability potentials. SPTs and BPTs are usually a more compact representation of potentials than tables, because they allow inference algorithms to take advantage of context-specific independences. In previous works we have defined detailed algorithms [20, 8] for making the basic operations (*combination*, *marginalization* and *restriction*) on potentials, directly over SPTs and BPTs.

3.1 Probability Trees

A *standard probability tree* $\mathcal{T}$ is a directed labelled tree, in which each internal node represents a variable and each leaf represents a non-negative real number. Each internal node has one outgoing arc for each state of the variable that labels that node; each state labels one arc. The *size* of a tree $\mathcal{T}$, denoted by $size(\mathcal{T})$, is defined as its nodes count.

A subtree of $\mathcal{T}$ is a *terminal tree* if it contains only one node labelled with a variable name, and all the children are numbers (leaf nodes).

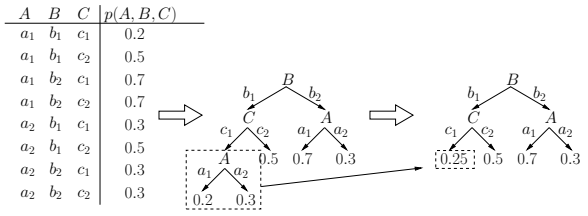


Figure 2: Potential p, its representation as a probability tree and its approximation after pruning

Figure 2 displays a potential p and its representation, using a SPT. This tree shows that the potential is independent of the value of A in the context $\{B = b_1, C = c_2\}$ (the value in the potential is 0.5 for $\{A = a_1, B = b_1, C = c_2\}$ and $\{A = a_2, B = b_1, C = c_2\}$). The tree contains the same information as the table, but only requires five values, while the table contains eight values. Furthermore, SPTs enable even more compact representations. This is achieved by pruning certain leaves, replacing them with the average value, as shown in the second tree shown in Fig. 2. The trade-off is a loss of accuracy.

3.2 Binary Probability Trees

A *binary probability tree* $\mathcal{BT}$ is similar to a SPT. It can also be defined as a directed labelled tree, where each internal node is labelled with a variable, and each leaf is labelled with a non-negative real number. But in this case, each internal node has always two outgoing arcs, and a variable can label several nodes in the path from the root to a leaf node. Another

difference is that, for an internal node labelled with X_i, the outgoing arcs can generally be labelled with more than one state of the domain of X_i, Ω_{X_i}. The size of a BPT (i.e., the number of nodes) is equal to twice the number of leaves minus one.

For example, Fig. 3 (ii) shows a BPT for the table in (i). In the figure, we use a superscript number at each node of the tree, in order to easily identify it. The domain of A, Ω_A, is $\{a_1, a_2, a_3\}$, and the domain of B, Ω_B, is $\{b_1, b_2, b_3\}$. This potential can also be represented with the SPT shown in Fig. 3 (iii). It can be seen that the BPT contains only five leaves, whereas the SPT contains seven. The SBT shown in Fig. 3 (iii) is able to capture a context-specific independence: *the potential does not depend on B when $A = a_1$*. The BPT in Fig. 3 (ii) captures the previous independence, but it is also able to capture other *fine-grained independences*. For example, *the potential does not depend on B when $A = a_2$ and $B \neq b_3$* ($B = b_1$ or $B = b_2$). This independence cannot be represented with the SPT of Fig. 3 (iii).

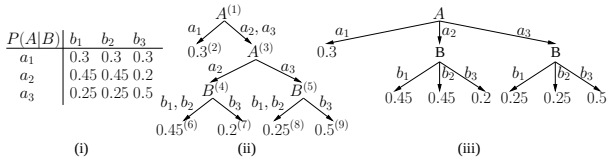


Figure 3: Potential $P(A|B)$ as table, BPT, and SPT

3.3 Constructing standard and binary trees

In [20] and [8] we have proposed a methodology for constructing a SPT $\mathcal{T}$ or a BPT $\mathcal{BT}$ from a potential p. These methods were inspired by the methods for inducing classification trees, such as Quinlan's ID3 algorithm [18], which builds a *decision tree* from a set of examples. But the measure used as the *splitting criterion* in SPTs and BPTs was specifically adapted to probabilities. Here we summarize the procedure for BPTs. For SPTs, a similar procedure is used.

Let p be a potential for a set of variables $\mathbf{X}_I$. It is generally possible to obtain several BPTs for p, depending on the order assigned to the variables of $\mathbf{X}_I$ in the internal nodes of the tree, and the distribution at each internal node of the available variable states over its outgoing arcs.

The process begins with a BPT $\mathcal{BT}_0$ with only one node (a leaf node) labelled with the average of the potential values: $L_t = \sum_{\mathbf{x}_I \in \Omega_{\mathbf{X}_I}} p(\mathbf{x}_I)/|\Omega_{\mathbf{X}_I}|$.

A greedy step is then applied successively until we obtain an exact BPT, or until a given *stop criterion* is satisfied. At each step, a new $\mathcal{BT}_{j+1}$ is obtained from the previous one, $\mathcal{BT}_j$. The greedy step requires the choice of a *splitting criterion*. It consists of expanding one of the leaf nodes t in $\mathcal{BT}_j$ with a terminal tree (with t rooting the terminal tree, and two new nodes t_l and t_r as children of t). Node t will be labelled with one of the *candidate variables*. Suppose $\Omega^t_{X_i}$, $\Omega^t_{X_i} \subseteq \Omega_{X_i}$, is the set of availabe states of X_i at node t. It is also necessary to distribute the set of available states $\Omega^t_{X_i}$ of the chosen candidate variable X_i into two subsets, $\Omega^{t_l}_{X_i}$ and $\Omega^{t_r}_{X_i}$, to label the two outgoing arcs (left and right) of t. This process is illustrated in Fig. 4, where the terminal node t in tree $\mathcal{BT}_j$ is expanded using variable B. The set of available states of B at node t, $\Omega^t_B = \{b_1, b_2, b_3\}$ was partitioned into the sets $\Omega^{t_l}_B = \{b_1\}$ and $\Omega^{t_r}_B = \{b_2, b_3\}$. After applying this process, we say that the leaf node t has been expanded with variable X_i and the sets of states $\Omega^{t_l}_{X_i}$ and $\Omega^{t_r}_{X_i}$.

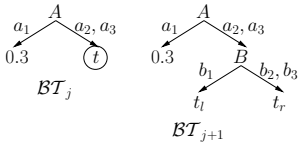


Figure 4: Expansion of the terminal tree t with B

The choice of the splitting criterion requires a distance to measure the goodness of the approximation of a BPT $\mathcal{BT}$ for a given potential p. If we denote by $\overline{\mathcal{BT}}$ and $\overline{p}$ the probability distributions (normalized potentials) proportional to $\mathcal{BT}$ and p, respectively, then the *distance* from a BPT $\mathcal{BT}$ to a potential p is measured using the Kullback-Leibler divergence [16]:

$$D(p, \mathcal{BT}) = \sum_{\mathbf{x}_I \in \Omega_{\mathbf{X}_I}} \overline{p}(\mathbf{x}_I) \log \frac{\overline{p}(\mathbf{x}_I)}{\overline{\mathcal{BT}}(\mathbf{x}_I)} \qquad (9)$$

Kullback-Leibler's divergence is always positive or zero. It is equal to zero if $\mathcal{BT}$ provides an exact representation of the potential p. It is a standard divergence used in information theory to measure the difference between two probability distributions. Here we use it to measure differences between potentials that are not really probability distributions (they represents conditional credal sets containing transparent variables), but experiments show that its use is a good heuristic procedure applied when reordering the variables of a tree or when pruning leaf nodes.

In [8] we proposed as *splitting criterion* to choose the partition that maximizes the *information gain* obtained for the current BPT $\mathcal{BT}_j$ after performing the mentioned expansion on leaf node t. For SPTs the information gain is calculated with:

$$I(t, X_i) = D(p, \mathcal{T}_j) - D(p, \mathcal{T}_j(t, X_i)) \quad (10)$$

where $\mathcal{T}_j(t, X_i)$ is the SPT $\mathcal{T}_j$ after expanding node t with the variable X_i.

For BPTs the information gain obtained after expanding node t is calculated with:

$$I(t, X_i, \Omega_{X_i}^{t_l}, \Omega_{X_i}^{t_r}) = D(p, \mathcal{BT}_j) - D(p, \mathcal{BT}_j(t, X_i, \Omega_{X_i}^{t_l}, \Omega_{X_i}^{t_r})) \quad (11)$$

where $\mathcal{BT}_j(t, X_i, \Omega_{X_i}^{t_l}, \Omega_{X_i}^{t_r})$ is $\mathcal{BT}_j$ after expanding node t with variable X_i and a partition of its available states $\Omega_{X_i}^{t}$ into sets $\Omega_{X_i}^{t_l}$ and $\Omega_{X_i}^{t_r}$.

It is immediate to see that $I(t, X_i, \Omega_{X_i}^{t_l}, \Omega_{X_i}^{t_r}) \geq 0$. By maximizing $I(t, X_i, \Omega_{X_i}^{t_l}, \Omega_{X_i}^{t_r})$ in the current greedy step, we manage to minimize Kullback-Leibler's distance to potential p in that step.

The information gain (expressions 10 and 11) obtained by expanding node t, can be efficiently calculated in SPTs and BPTs (see Proposition 1 in [20, 8]).

The methodology explained in this section for building a SPT or BPT can also be used to reorder the variables (or the split sets) of a SPT or BPT resulting from an operation of combination or marginalization. This enables us to move the most informative variables to the upper levels of the tree. So, if a pruning operation is applied, only the less informative variables will be removed. The process to reorder a BPT $\mathcal{BT}$ is the same as the one for building a BPT from a potential p. Here, p is the potential that $\mathcal{BT}$ represents. So, we can build a new BPT applying the same procedure explained in this section.

3.4 Pruning standard and binary trees

During the inference process it is possible that some trees have a large size, making it impossible to obtain any result with the available memory of our computer. Pruning of SPTs [20] was proposed as a way to control the size of trees during the propagation process. This operation has also been extended to BPTs [8]. In this way, we can obtain a result from an inference algorithm although it will be approximate. Basically, a *pruning* in a SPT or BPT consists of replacing a terminal tree by the average of values that it represents. For example, if we wish to prune the terminal tree rooted by node (4) in the BPT of Fig. 3 (ii), we must replace it by $(0.45 + 0.45 + 0.2)/3$. In [6] we demonstrated that the pruned tree obtained with the previous procedure is the tree that minimizes the Kullback-Leibler divergence between the exact potential and all the trees with the same structure as that pruned tree.

In [20, 8] it is proposed to repeat the pruning process until the tree contains no terminal tree which *information loss* is under a given threshold Δ. The information loss is also calculated with the difference of the Kullback-Leibler's distances, before and after pruning (expressions 10 and 11). The goal of the pruning of a tree involves detecting leaves that can be replaced by one value without a big increment in Kullback-Leibler's divergence of the potential represented by that tree, before and after pruning.

Again, the information loss can be locally computed at node t in the current SPT or BPT.

4 Propagating credal sets using binary probability trees

The simpler approximate algorithm for propagating credal sets using SPTs is based on the *Variable Elimination* algorithm [11]. VE is one of the most popular algorithms for computing *a posteriori* information in probabilistic graphical models using local computations. It was independently proposed by Shafer and Shenoy [21], Zhang and Poole [22] and Dechter [15]. The input of this algorithm is a set of potentials and a queried variable. It iteratively eliminates variables from the set of potentials by using combination and marginalization until only the queried variable remains in the set of potentials.

In this paper, we propose to use also the VE algorithm to propagate in CNs, but using BPTs (see Algorithm 1) to represent the credal sets $K(X_i|\Pi_i)$. In CNs, all the variables should be removed (by marginalization) except the queried variable and the transparent variables (see bellow for an explanation of transparent variables). Here, the set of potentials is the set $\{K(X_i|\Pi_i)\}$ of extensive conditional credal sets in the CN.

For each X_i, we originally have a collection of m separately specified credal sets $\{K(X_i|\pi_1), \ldots, K(X_i|\pi_m)\}$, where m is the number of configurations of Π_i. The problem is transformed into an equivalent one by using a *transparent variable* T_{π_i} for each configuration of the parents of X_i ($\pi_i \in \Omega_{\Pi_i}$). T_{π_i} will have as many cases as the number of vertices in the separately specified credal set $K(X_i|\pi_i)$. Each vertex of the extensive conditional credal set $K(X_i|\Pi_i)$ can be obtained by fixing each transparent variable T_{π_i} to one of its values. This transformation is equivalent in size to the one proposed by Antonucci et al. in [1], although

that one requires modifications in the graph of the CN.

SBTs and BPTs enable an extensive conditional credal set $K(X_i|\Pi_i)$ to be represented efficiently when it comes from m separately specified credal sets $\{K(X_i|\pi_1),\ldots,K(X_i|\pi_m)\}$ and with a single data structure (the necessary space for the tree is proportional to the sum of the necessary spaces for the m local trees). In Fig. 5, we can see one example where a BPT represents the extensive conditional credal set $K(X|Y)$ associated to the two separately specified credal sets $K(X|Y=y_1)$ and $K(X|Y=y_2)$. In the BPT in Fig. 5, we can obtain the extreme points of $K(X|Y)$ by fixing T_{y_1} and T_{y_2} to each one of its values. For example, if the BPT is restricted to $T_{y_1}=t^1_{y_1}$ and $T_{y_2}=t^2_{y_2}$, we obtain a new BPT that gives us the extreme point of $K(X|Y)$ associated to p_1 and q_2. The tree avoids repetition of probability values, reducing the space necessary with respect to the table representation.

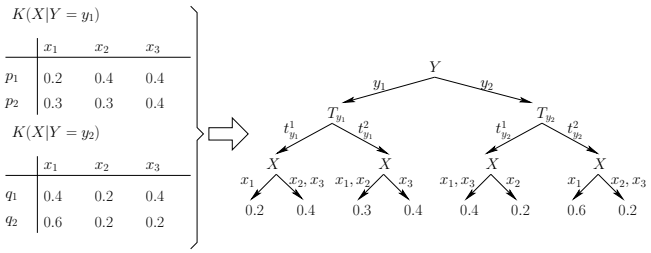


Figure 5: A binary probability tree for $K(X|Y)$

Algorithm 1: Variable Elimination

Input : $\mathbf{K}=\{K(X_i|\pi_i): i=1,\ldots,n\}$ the set of separately specified credal sets in the CN; $\mathbf{e}$ the set of observed values, $\mathbf{e}\in\Omega_{\mathbf{E}}$; a variable of interest X_q, $X_q\in\mathbf{X}\setminus\mathbf{E}$; and Δ the threshold for pruning

Output: $\underline{P}(x_q|\mathbf{e})$ and $\overline{P}(x_q|\mathbf{e})$ for each $x_q\in\Omega_{X_q}$, $X_q\in\mathbf{X}\setminus\mathbf{E}$

1 Get the set $S_{\mathcal{BT}}$ of binary trees, building each binary tree $\mathcal{BT}_i$ from the credal sets $K(X_i|\pi_i),\forall\pi_i\in\Omega_{\Pi_i}$ (as in Figure 5)
2 Transform each $\mathcal{BT}_i$ into $\mathcal{BT}_i^{R(\mathbf{e})}$ (restrict to evidence)
3 Reorder variables and split sets in every $\mathcal{BT}_i$
4 Prune each $\mathcal{BT}_i$ with the Δ threshold
5 **foreach** $Y\in\mathbf{X}\setminus(\mathbf{E}\cup\{X_q\})$ **do**
6 Let $\mathbf{S}_Y=\{\mathcal{BT}_i|Y\in s(\mathcal{BT}_i)\}$
7 Calculate $\mathcal{BT}_{prod}=\prod_{\mathcal{BT}_i\in\mathbf{S}_Y}\mathcal{BT}_i$
8 Calculate $\mathcal{BT}_{sum}=\mathcal{BT}_{prod}^{\downarrow s(\mathcal{BT}_{prod})\setminus Y}$
9 Reorder variables and split sets in $\mathcal{BT}_{sum}$
10 Prune $\mathcal{BT}_{sum}$ using the Δ threshold
11 $S_{\mathcal{BT}}=\{(S_{\mathcal{BT}}\setminus\mathbf{S}_Y)\}\cup\mathcal{BT}_{sum}$
12 Calculate $\mathcal{BT}_q=\prod_{\mathcal{BT}_i\in S_{\mathcal{BT}}}\mathcal{BT}_i$
13 Get $\underline{P}(x_q|\mathbf{e})$ and $\overline{P}(x_q|\mathbf{e})$ by normalizing the vertices in $\mathcal{BT}_q$

In our version of the VE algorithm (Algorithm 1), each conditional credal set $K(X_i|\Pi_i)$ is represented with a BPT as in Fig. 5 from the set of separately specified credal sets $\{K(X_i|\pi_i),\forall\pi_i\in\Omega_{\Pi_i}\}$. This is done in step 1 of the algorithm. The evidence (if available) is incorporated with restriction operations (step 2). Then each tree is reordered (step 3) so that the most informative variables appear in the upper levels of the tree, using the procedure described in Section 3.3. Step 4 consists of pruning the trees using a given Δ threshold in order to reduce their sizes as much as possible. The loop (step 5) deletes a variable (non-transparent) in each iteration. The combination of trees containing the variable to be removed is performed in step 7. This operation is made directly over trees (see [8]). The resulting tree is marginalized to discard the variable to be removed using marginalization (step 8). Again, this operation is made directly over the tree (see [8]). Steps 9 and 10 reorder the variables of the tree (see Section 3.3) and prune it respectively. The pruning operation can select any variable (normal or transparent one) in the tree. Finally, the resulting trees (all of them will be defined only on the queried variable and on transparent variables) are combined to produce a single tree (step 12). Finally the upper and lower bounds for the probability of the queried variable can be obtained by normalizing each one of the vertices in $\mathcal{BT}_q$ (step 13) using expression 8. The pruning reduces the complexity of posterior operations. The more transparent variables are pruned the less vertices appears in the final credal set obtained with the BPT in step 12 of the algorithm. When a $\Delta=0.0$ threshold is used, no variable will be pruned unless there are context specific independences in the potentials. In the worst case, using $\Delta=0.0$, the BPT obtained in step 12 corresponds to a credal set with n_v possible vertices.

With respect to the complexity of Algorithm 1, using $\Delta=0.0$, if the potentials do not contain any context specific independence, no pruning will be done, and so inference is equivalent to make n_v propagations in a BN. This is the worst case. Using values of Δ greater than 0.0 we can reduce the size of potentials and so computing times. A theoretical evaluation of the computational complexity is out of the scope of this paper.

5 Experiments

In order to compare the performance of SPTs and BPTs we have used two classical BNs (*Alarm* [2] and *Insurance* [3]). The number of states for the variables in these networks is maintained as in their original specifications. These networks contain variables with more than two states. For each model, we obtained a CN by randomly generating separately specified conditional credal sets for each variable X_i and each configuration of the parents of X_i. The number of vertices

at each $K(X_i|\pi_i)$ is selected as follows: For a given percentage of the configurations in Ω_{Π_i} we associated a given number of vertices in the credal sets $K(X_i|\pi_i)$. For the rest of configurations we used only one vertex. This allows us to control the potential size of the strong extension of the CN, so that exact inference is not too difficult to be done in our computers, in order to allow the comparison of the error of approximate inference with respect to the exact one. The process to randomly select the probabilities for the vertices at each separately specified credal set $K(X_i|\pi_i)$ is as follows. When only one vertex must be used we take the probability values in the original BN. When several vertices are used we take as basis the probability distribution in the original BN ($P(X_i|\pi_i)$). If a value equal to 0.0 is found for a given configuration of $P(X_i|\pi_i)$, it will be kept for that configuration. If a value equal to v, $v > 0.0$, is found, we select a new uniform random value in the interval $[-v, v]$ (negative values are converted into positive). The resulting vertex is then normalized. This procedure do not produce too much context specific independences in the resulting potentials, but we must take into account that these kind of independences are present in our representation of extensive conditional credal sets by means of trees. For example, in Fig. 5 the potential do not depend on T_{y_2} when $Y = y_1$.

Several experiments have been done using different variables for each network. In some cases we have considered that some of the variables of the network are observed. In Table 2 we show for each experiment (Ex), the chosen variable (Var), the name of the network, the number of observed variables ($|\mathbf{E}|$), the number of vertices per credal set (nvpc), the percentage of configurations (per) of Ω_{Π_i} that will contain nvpc vertices, and the potential size of the strong extension (n_v) of the CN. In the calculus of nv we suppose that the barren nodes for the given query have been removed from the network.

Ex	Var	Network	$\lvert\mathbf{E}\rvert$	nvpc	per	n_v
1	Venttube	Alarm	0	3	90	354294
2	Expco2	Alarm	0	3	17	177147
3	RiskAversion	Insurance	0	3	70	177147
4	DrivHist	Insurance	0	3	31.5	177147
5	Venttube	Alarm	6	3	12.25	354294
6	DrivHist	Insurance	9	3	12	944784

Table 2: Experiments we have done

We have measured the maximum required size of SPTs and BPTs during the propagation (biggest tree used in the computations), the *mean square error* for the a posteriori bounds of the queried variable and the running time used by the propagation algorithm. The mean square error for a queried variable X_q is measured using the following expression:

$$\sqrt{\frac{\sum_{x_q \in \Omega_{X_q}} ((\underline{P}^*(x_q|\mathbf{e}) - \underline{P}(x_q|\mathbf{e}))^2 + (\overline{P}^*(x_q|\mathbf{e}) - \overline{P}(x_q|\mathbf{e}))^2)}{2 \cdot |\Omega_{X_q}|}} \tag{12}$$

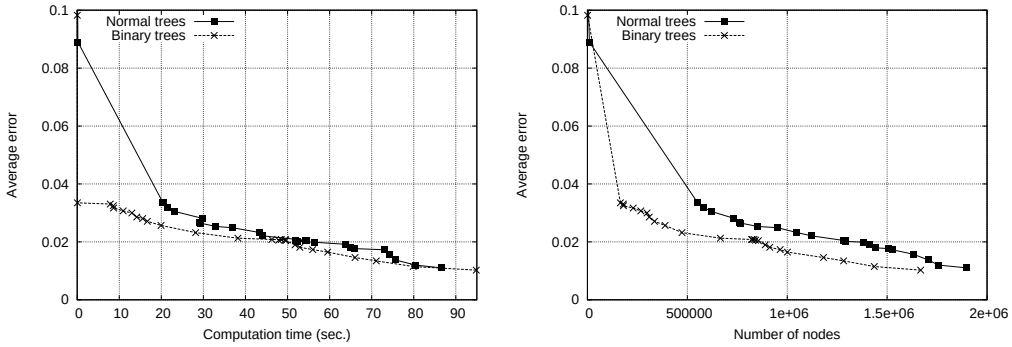
where $\underline{P}^*(x_q|\mathbf{e}, \overline{P}^*(x_q|\mathbf{e})$ are the approximate lower and upper bounds and $\underline{P}(x_q|\mathbf{e}, \overline{P}(x_q|\mathbf{e})$ the exact ones.

These parameters (mean square error, maximum size and time) are measured running the Algorithm 1 with several values for the Δ threshold using SPTs and BPTs. We have used values for Δ in the interval $[10^{-7}, 10^{-2}]$. Each experiment was run ten times. Each time, we began randomly generating the probabilities for each credal set. So, average of mean square error, maximum size and time (in seconds) are calculated and reported in figures 6 to 11 for the different experiments. For each experiment, we show the average mean square error versus largest tree size required in the two versions of the propagation algorithm (using SPTs and BPTs) and the average mean square error versus average time required in the two versions of the propagation algorithm (using SPTs and BPTs).

As expected with both kind of trees, high values of Δ cause large errors but require lower computing time and smaller trees. Small values of Δ give small errors but require a high computing time and large trees.

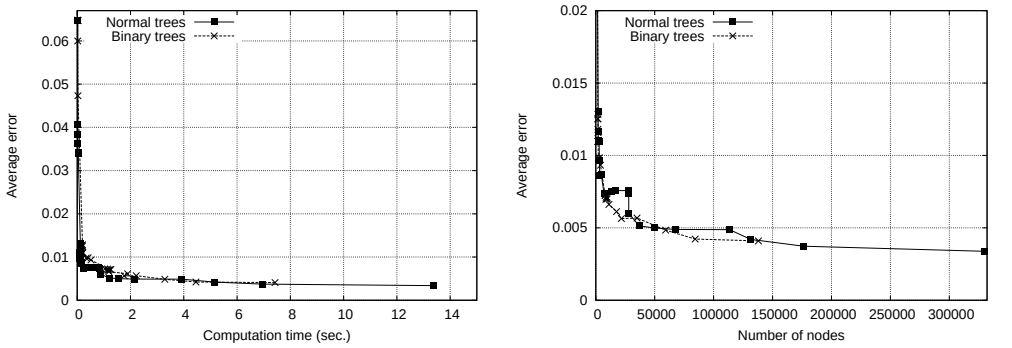
The figures allow to compare propagation with SPTs and BPTs for each experiment. In some cases, we can see a noticeable reduction in the size and required time using BPTs with respect to SBTs: that is, the same level of error can be achieved with BPTs, but with a very important reduction in size and time. This is the case of Experiment 1 for VENTTUBE variable (4 states) in Alarm network (Fig. 6), Experiment 3 for RiskAversion (4 states) in Insurance network (Fig. 8), Experiment 5 for VENTTUBE variable in Alarm network using 6 observed variables (Fig. 10). There are also cases where the performance of SPTs and BPTs is quite similar. For example, see Experiment 2 for EXPCO2 variable (4 states) in Alarm network (Fig. 7) or Experiment 6 for DrivHist in Insurance network using 9 observed variables (Fig. 11).

We have also tried to propagate using tables for representing the extensive conditional credal sets, like the one in Table 1, but our computer run out of memory in all the experiments in about 18 minutes. This is because a table does not allow to capture the context specific independences for transparent variables, and so the size of potentials increases quicker for tables in the propagation process, even if we do not use pruning in trees. We have also compared the maximum tree size and computing time. Obviously computing time increases when bigger trees are used (figures are



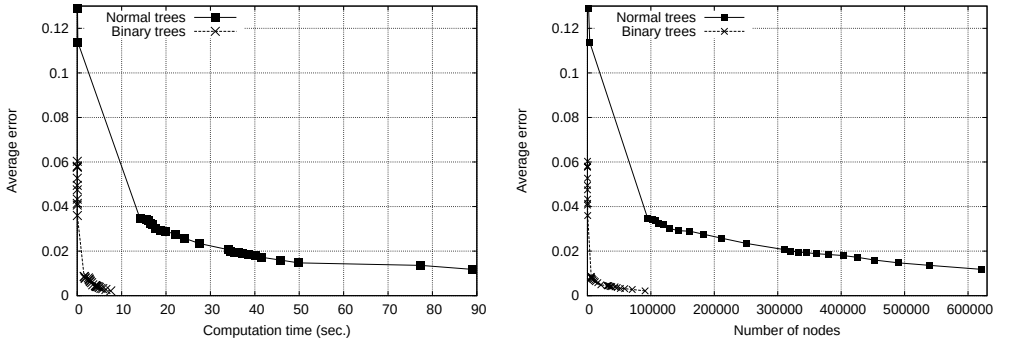
(a) MSE versus computing time (b) MSE versus maximum tree size

Figure 6: Inference for VENTTUBE in Alarm network (no evidence)



(a) MSE versus computing time (b) MSE versus maximum tree size

Figure 7: Inference for EXPCO2 in Alarm network (no evidence)



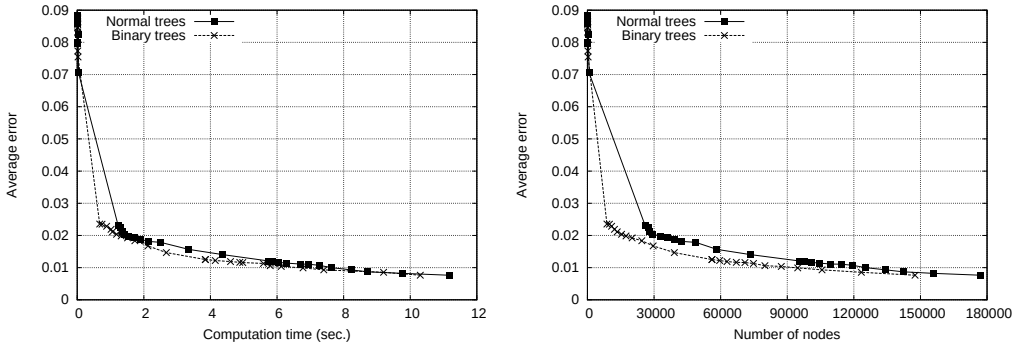
(a) MSE versus computing time (b) MSE versus maximum tree size

Figure 8: Inference for RiskAversion in Insurance network (no evidence)

not includes because of the space).

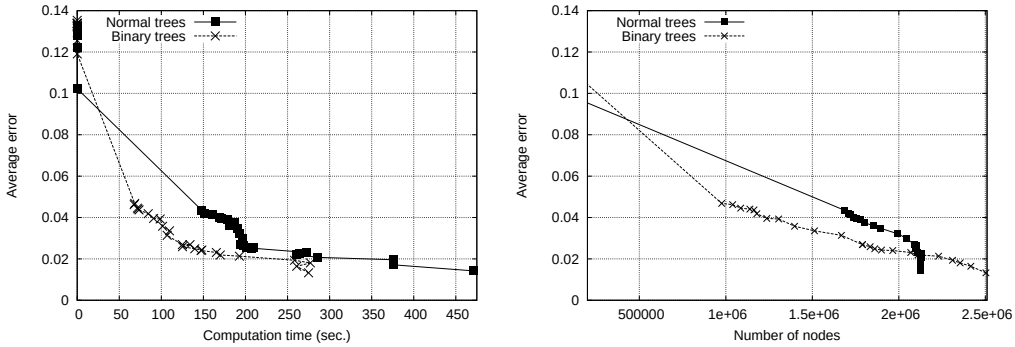
6 Conclusions

In this paper we have proposed the use of BPTs to propagate in CNs. BPTs and SPTs make possible to control the accuracy of the propagation by means of a given threshold Δ used for pruning the trees. The choice of Δ is a trade-off between accuracy and computing time. The experiments show that BPTs offer better performance than SPTs in some cases, and similar one in other cases. So, we think that BPTs is a better representation for the potentials of a CN.



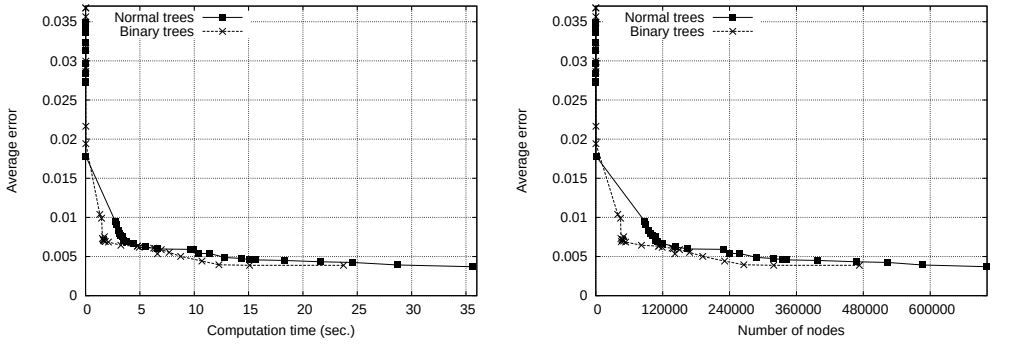
(a) MSE versus computing time (b) MSE versus maximum tree size

Figure 9: Inference for DrivHist in Insurance network (no evidence)



(a) MSE versus computing time (b) MSE versus maximum tree size

Figure 10: Inference for VENTTUBE in Alarm network (evidence in 6 variables)



(a) MSE versus computing time (b) MSE versus maximum tree size

Figure 11: Inference for DrivHist in Insurance network (evidence in 9 variables)

In the future we intend to perform more exhaustive experiments so we can characterize the situations where BPTs will be better than SPTs. In this way we will check the complete list of unobserved variables in these networks and in other classical BNs. We will also analyze the impact of the number of vertices in the conditional credal sets in the performance of BPTs with respect to SPTs.

Acknowledgements

This work has been supported by the Spanish Ministry of Science and Innovation, under projects TIN2007-67418-C03-03, TIN2010-20900-C04-01 and

by the European Regional Development Fund (FEDER). We are also very grateful to the anonymous reviewers for their valuable comments and suggestions.

References

[1] A. Antonucci, Y. Sun, C. P. de Campos, and M. Zaffalon. Generalized loopy 2u: A new algorithm for approximate inference in credal networks. *Int. J. Approx. Reasoning*, 51(5):474–484, 2010.

[2] I. Beinlich, G. Suermondt, R. Chavez, and G. Cooper. The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Second European Conference on AI and Medicine*, Berlin, 1989. Springer-Verlag.

[3] J. Binder, D. Koller, S. Russell, and K. Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29, 1997.

[4] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in Bayesian networks. In *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence (UAI–96)*, pages 115–123, Portland, Oregon, 1996.

[5] C. P. De Campos and F. G. Cozman. The inferential complexity of Bayesian and credal networks. In *In Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1313–1318, 2005.

[6] A. Cano. *Propagación aproximada de intervalos de probabilidad en grafos de dependencias*. PhD thesis, Universidad de Granada, Dpt. Computer Sciences and Artificial Intelligence, June 1999.

[7] A. Cano, M. Gómez, S. Moral, and J. Abellán. Hill-climbing and branch-and-bound algorithms for exact and approximate inference in credal networks. *International Journal of Approximate Reasoning*, 44:261–280, 2007.

[8] A. Cano, M. Gómez-Olmedo, and S. Moral. Approximate inference in Bayesian networks using binary probability trees. *International Journal of Approximate Reasoning*, 52:49–62, 2011.

[9] A. Cano and S. Moral. A review of propagation algorithms for imprecise probabilities. In *Proceedings of the First International Symposium on Imprecise Probabilities and their Applications (ISIPTA'99)*, Ghent, 1999.

[10] A. Cano and S. Moral. Computing probability intervals with simulated annealing and probability trees. *Journal of Applied Non-Classical Logics*, 12(2):151–171, 2002.

[11] A. Cano and S. Moral. Using probabilities trees to compute marginals with imprecise probabilities. *International Journal of Approximate Reasoning*, 29:1–46, 2002.

[12] I. Couso, S. Moral, and P. Walley. Examples of independence for imprecise probabilities. In *Proceedings of the First International Symposium on Imprecise Probabilities and Their Applications (ISIPTA'99)*, 1999.

[13] F. G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.

[14] F. G. Cozman. Graphical models for imprecise probabilities. *International Journal of Approximate Reasoning*, 39:167–184, 2005.

[15] R. Dechter. Bucklet elimination: A unifying framework for probabilistic inference. In E. Horvitz and F.V. Jennsen, editors, *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, pages 211–219, 1996.

[16] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:76–86, 1951.

[17] J. Pearl. *Probabilistic Reasoning with Intelligent Systems*. Morgan & Kaufman, San Mateo, 1988.

[18] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–105, 1986.

[19] J.C. F. Rocha and F.G. Cozman. Inference with separately specified sets of probabilities in credal networks. In A. Darwiche and N. Friedman, editors, *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 2002.

[20] A. Salmerón, A. Cano, and S. Moral. Importance sampling in Bayesian networks using probability trees. *Computational Statistics and Data Analysis*, 34:387–413, 2000.

[21] P. P. Shenoy and G. Shafer. Axioms for probability and belief-function propagation. In Shachter et al., editors, *Uncertainty in Artificial Intelligence, 4*, pages 169–198. North-Holland, 1990.

[22] N. L. Zhang and D. Poole. Exploiting causal independence in Bayesian network inference. *International Journal of Intelligent Research*, 5:301–328, 1996.

Incoherence correction strategies in statistical matching

Andrea Capotorti
Dip. di Matematica e Informatica,
University of Perugia, Italy
capot@dmi.unipg.it

Barbara Vantaggi
Dip. Scienze di Base e Applicate per l'Ingegneria,
University La Sapienza, Roma, Italy
vantaggi@dmmm.uniroma1.it

Abstract

We deal with the statistical matching problem and in particular we study the problem related to the managing of inconsistencies. In fact, when logical relations among the variables are present incoherence can arise in the probability evaluations. The aim of this paper is to remove such incoherences by using different methods. Specific precise distances minimization or least committal imprecise probability extensions are adopted. We compare these methods using a practical example that brings to light the peculiarities of the statistical matching problem.

Keywords. Statistical matching, incoherence, inference, specialized discrepancy measure.

1 Introduction

In several economic applications there is a need to consider different data sources and to integrate the information coming from them [3, 13, 23, 25, 26]. In particular, we deal with the so called statistical matching problem, that can be represented by the following simple situation: there are two different sources, A and B, with some overlapping variables and some variables collected only in one source. Let X represent the common variables, Y denotes the variables collected only in A, and Z those only in B. Thus, the data consist of a first sample (X, Y) and a second sample on (X, Z). In this context data are missing by design since they have been already collected separately, and to get jointly data on Y and Z would be expensive and time-consuming.

Traditionally, to cope with these problems the available data are combined with assumptions strong enough to point-identify the joint probability distribution (see references in [26]): we recall, for example, those based on a conditional independence assumption, i.e. the variables Y and Z are independent conditional on X.

However, in several situations the independence assumption is not adequate, as first raised by Sims [31] (see also [25, 28, 29, 32]). Other methods aim at incorporating auxiliary information about relationships between Y and Z to avoid or to relax conditional independence assumption (see, e.g. [32]). Although this is an important case, it is not always feasible because the required external knowledge may not be available.

Actually, since there are many distributions on (X, Y, Z) compatible with the available partial information on (X, Y) and (X, Z), it is too restrictive to consider just one of the compatible distributions, obtained perhaps by taking a specific assumption (as already noted in [14, 17, 30] and for the missing data problem [11, 22, 34]).

This problem has been faced in a coherent conditional probability setting in [35, 36]: coherence allows us to check the compatibility of partial (conditional) assessments, to manage further available knowledge, for example coming from field experts; moreover it allows us to draw inferences by considering all the compatible distributions.

A further remarkable advantage of using this approach is that we are able to consider multiple integration, that is important for real applications (for instance, see [33] for some economic Hungarian applications based on the combination of three different surveys).

Moreover, this approach [36] allows to manage logical constraints characterizing the relevant links among variables describing the phenomenon. In particular, in [36] it is proved that when there is no logical constraint among the variables, coherence is always satisfied by also requiring conditional independence, then this hypothesis is legitimate from a syntactical point of view (even if it is useful to look for all compatible coherent extensions). On the other hand, when logical constraints are present it is necessary to check global coherence of the relevant partial assessments

drawn from the different sources and if coherence is not satisfied we need to remove incoherences. In [35] this is done by looking for the "minimal" incoherent assessments and to remove them in order to restore coherence by using the $L1$ norm.

The aim of this paper is to deal with incoherences and to look for the coherent assessment "closest" to the given one with respect to different distances ($L1, L2$, Kulback-Leibler divergence, discrepancy). Then, when coherence is restored we can draw inference: for each (conditional) event we can directly build the interval of all coherent probability values solely on the base of a partial assessment, i.e. it is not needed to artificially fulfill the missing values of the data base. It is important to remark that the interval bounds are computed analytically.

Actually, our aim is in the same line of those based on multiple imputation [30] and its extension [26], which aims at approximating the lower and upper bounds for the quantities of interest in the multinormal setting. A similar approximation for these bounds is carried out in [14] on the base of maximum likelihood approach.

To let this paper be as much as possible self-contained, in Section 2 we introduce the basic notions and characteristic of coherent conditional partial assessments, either based on precise **p** or on imprecise **lub** evaluations given on a finite domain $\mathcal{E}$. Coherence of an assessment is required to perform a sound inference that for partial assessments coincide with a coherent extension. Hence also basic extension notions, both for the precise and imprecise context, are given. Afterwards in Section 3 the main (pseudo)distances between conditional assessments are introduced. It is in fact thanks to their minimization that consistent correction of incoherent assessments will be possible. Such (pseudo)distances can be based on geometrical properties, e.g. $L1$ and $L2$ norms, or on information theoretic foundation, e.g. KL divergence, or can derive from proper scoring rules, e.g. discrepancy Δ, suitably tailored for partial conditional assessments. In Subsection 3.1 it is sketched an alternative way of restoring consistency: whenever it is possible to identify a coherent sub-assessment $(\mathcal{G}, \mathbf{p}_{|\mathcal{G}})$, it can be coherently extended to the rest of the initial domain $\mathcal{F} = \mathcal{E} \setminus \mathcal{G}$. This inevitably produces an imprecise conditional probability assessment. Subsequently, in Section 4, the statistical matching problem is reformulated inside a conditional probability assessment framework and conditions are given that guarantee the coherence of the whole assessment. On the contrary, whenever there are logical constraints among the variables under investigation, even starting from separately coherent sources of information, the whole assessment could result incoherent. In Section 5 this is well described by a simple example. This section is the core of our contribution, where the previous concepts are merged together and the two main approaches for inconsistency correction, the minimization of (pseudo)distances or the extension of a coherent sub-assessment, are specialized to the statistical matching problem. It is also shown how the peculiarity of the statistical matching suggests a specialization of the general discrepancy Δ into a peculiar one Δ_{mix}. This new discrepancy is a mixture of the original one applied to the different scenarios and the consequent inconsistency correction, obtained by its minimization, leaves unchanged the marginal distribution of the common variables X. To better show the advantages and drawbacks of the proposed methods, in Section 6 we introduce an example built from data taken from [14]. The final short concluding Section 7 sums up the proposed methodologies.

2 Preliminaries about coherent conditional probability

Whenever several sources of information, that could represent expert's opinions and/or knowledge bases, are merged together, we can generally start to deal with an overall domain $\mathcal{E} = [E_1|H_1, \ldots, E_n|H_n]$.

The events E_i's 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" ($E_i \wedge H_i$), E_i^c will indicate "not E_i", the contrary of E_i, and the event $H^0 = \bigvee_{i=1}^n H_i$ will represent the whole set of contexts.

Starting with 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 a generic atom that is the elementary element in the algebra generated by the E_i and H_i. Note that the sample space Ω, together with H^0, are not part of the assessment but only auxiliary tools.

Every probability mass function $\alpha : \mathcal{P}(\Omega) \to \mathbb{R}$ 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 results $\alpha(E) = \sum_{\omega_j \subseteq E} \alpha_j$.
We need to introduce a nested hierarchy among probability distributions sets:

- $\mathcal{A} = \left\{\boldsymbol{\alpha}, \sum_1^k \alpha_i = 1, \alpha_j \geq 0, j = 1, \ldots, k\right\}$;
- $\mathcal{A}_0 = \left\{\boldsymbol{\alpha} \in \mathcal{A} | \alpha(H^0) = 1\right\}$;
- $\mathcal{A}_1 = \{\boldsymbol{\alpha} \in \mathcal{A}_0 | \alpha(H_i) > 0, i = 1, \ldots, n\}$.

It is easy to see that the set $\mathcal{A}_1$ is a convex set and $\mathcal{A}_0$ is the closure of $\mathcal{A}_1$ in the usual topology.

We focus our attention on coherent (conditional) probability assessments $\mathbf{p}$, that can be reduced to the compatibility with a conditional probabilities, as introduced by Dubins [15] and De Finetti [12] (see also Krauss [18] and Rényi [27]).

Definition 1 *Let $\mathcal{E} = [E_1|H_1, \ldots, E_n|H_n]$ be an arbitrary set of conditional events, an assessment $\mathbf{p}$ on $\mathcal{E}$ is said to be a coherent conditional probability if there exists a conditional probability $P(\cdot|\cdot)$ defined on $\mathcal{B} \times (\mathcal{B} \setminus \emptyset)$ (with $\mathcal{B}$ the algebra spanned by $E_1, H_1, ..., E_n, H_n$) which restriction to $\mathcal{E}$ coincides with $\mathbf{p}$.*

Every probability distribution $\boldsymbol{\alpha} \in \mathcal{A}_1$ generates a coherent conditional probability assessment $\mathbf{q}_{\boldsymbol{\alpha}}$ on $\mathcal{E}$ through the usual formula

$$q_{\alpha_i} = \sum_{\omega_j \subseteq E_i H_i} \alpha_j / \sum_{\omega_j \subseteq H_i} \alpha_j \text{ for all } i = 1, \ldots, n. \quad (1)$$

Note that $\mathbf{q}_{\boldsymbol{\alpha}}$ is a continuous function of $\boldsymbol{\alpha}$ when $\boldsymbol{\alpha} \in \mathcal{A}_1$. When $\boldsymbol{\alpha} \in \mathcal{A}_0$, the previous formula (1) defines $\mathbf{q}_{\boldsymbol{\alpha}}$ only on

$$\mathcal{E}_{\boldsymbol{\alpha}} := \{E_i|H_i \in \mathcal{E}, \alpha(H_i) > 0\}. \quad (2)$$

To cover the case of conditioning events with null probability, in fact we need to resort to a suitable class of probability distributions $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_l$ agreeing with $P(\cdot|\cdot)$ (for more details refer to the characterization theorem reported e.g. in [9, 10]).

Coherence is crucial since it is a prerequisite for a sound inference, that means extension of the given assessment to any new conditional event. In fact the following theorem, essentially due to [12], holds:

Theorem 1 *Let $\mathbf{p}$ be an assessment on an arbitrary family $\mathcal{E}$; then there exists a (possibly not unique) coherent extension of $\mathbf{p}$ to any family $\mathcal{K} \supset \mathcal{E}$* if and only if *$\mathbf{p}$ is a coherent conditional probability on $\mathcal{E}$.*

Moreover, if $\mathbf{p}$ is a coherent conditional probability on $\mathcal{E}$, then the coherent probability values for any conditional event $F|K \in \mathcal{K} \setminus \mathcal{E}$ belong to a closed interval $[\underline{p_{F|K}}, \overline{p_{F|K}}]$.

The aforementioned coherent interval $[\underline{p_{F|K}}, \overline{p_{F|K}}]$ can be obtained by solving specific linear optimization problems (for details refer again to [10]) based on suitable classes of probability distributions $\{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_l\}$ agreeing with $\mathbf{p}$.

The notion of coherence also apply to imprecise conditional assessments, i.e. whenever the numerical part of the assessment is elicited through interval values

$$\mathbf{lub} = ([lb_1, ub_1], \ldots, [lb_n, ub_n]). \quad (3)$$

Of course, some of the intervals $[lb_i, ub_i]$ could degenerate to a precise value p_i.

For assessments such as $(\mathcal{E}, \mathbf{lub})$, although defined on finite spaces, there could be different kinds of consistency requirements (for a detailed exposition, among others, refer to [24]). The basic consistency notion is the so called *avoiding of partial loss*, while in this paper we focus on the most stringent one: *(strong) coherence*. By taking into account a Bayesian sensitivity analysis interpretation, coherent lower-upper conditional probability assessments $(\mathcal{E}, \mathbf{lub})$ are such that intervals' lower (upper) extremes lb_i (ub_i) can be obtained as lower (upper) envelopes of sets of coherent precise conditional probability assessments on $\mathcal{E}$. It follows that to have a coherent lower-upper assessment $(\mathcal{E}, \mathbf{lub})$, there should exist a *set* of probability *classes* $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_l$ such that they induce probabilities for the $E_i|H_i$ inside the ranges $[lb_i, ub_i]$, and moreover each lower (lbi_s) and upper (ubi_s) bound on a conditional event is attained in at least one distribution.

Also, starting from a coherent lower-upper assessment $(\mathcal{E}, \mathbf{lub})$, it is possible to infer coherent bounds $[\underline{p_{F|K}}, \overline{p_{F|K}}]$ for the probability of any target conditional event $F|K$ through specific sequences of linear optimization problems and/or satisfiability of logical configurations (for details refer to [4]).

3 Coherent adjustments

Given an incoherent conditional probability assessment, for example, on a domain arising from the merging of separately coherent partial probability assessments, we need to restore coherence in a way to preserve, as much as possible, the information on the initial assessments, without introducing exogenous information. This goal is obtained generally by minimizing some kind of distance among partial conditional assessments.

(Pseudo)distances among probability distributions are usually defined through divergencies (e.g. Euclidean distance, Kulback-Leibler divergence, Csiszár f-divergences, etc.). Some of them can be applied only among unconditional probability distributions; others could be applied in our context of partial conditional

assessments, but could have no probabilistic justification, being purely geometrical tools.

Given two conditional assessments $\mathbf{p} = [p_1, \ldots, p_n]$ and $\mathbf{q} = [q_1, \ldots, q_n]$ on the same set of conditional events $\mathcal{E}$, the most widely adopted divergencies among them are:

1. $L1(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} |q_i - p_i|$;
2. $L2(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} (q_i - p_i)^2$;
3. $KL(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} (q_i \ln(q_i/p_i) - q_i + p_i)$.

$L1$ and $L2$ are usual metric distances, endowed with all their geometric properties, but until now remain without an intuitive probabilistic interpretation for conditional assessments. Moreover, their use in conditional context could lead to numerical troubles due to non-convexity of coherent assessments, as the following simple example borrowed from [2] shows:

Example 1 *Consider $\mathcal{E} = [A|H, B|AH, AB|H]$ with A, B, H logically independent. Hence the sample space is composed by 8 atoms, 4 of them inside $H^0 \equiv H$. The set of coherent assessments $\mathcal{Q}_\mathcal{E}$ is formed by the triples $[q_1, q_2, q_3] \in [0, 1]^3$ with $q_3 = q_1\, q_2$.*

Then, the set $\mathcal{Q}_\mathcal{E}$ is evidently non-convex.

KL is the so called logarithmic Bregman divergence. In the unconditional framework, such divergence is the most frequently adopted, because of its information theoretic properties. In fact, it generalizes the well known Kulback-Leibler divergence [19] to partial assessments. Anyhow, it is known that this Bregman divergence is generated by a logarithmic scoring rule that has a peculiarity that in some cases it is better to avoid: it evaluates only the events that occur, without considering those that turn out to be false.

To overcome this characteristic and to encompass the need of considering the conditional framework where the assessment is given, recently in [6, 8] for partial conditional assessments $\mathbf{v} = [v_1, \ldots, v_n] \in (0, 1)^n$ over $\mathcal{E} = [E_1|H_1, \ldots, E_n|H_n]$, the following random variable has been proposed as scoring rule:

$$S(\mathbf{v}) := \sum_{i=1}^{n} |E_i H_i| \ln v_i + \sum_{i=1}^{n} |E_i^c H_i| \ln(1 - v_i) \quad (4)$$

with $|\cdot|$ the indicator function of unconditional events.

The motivation of such a score is that the assessor "loses less" the higher the probabilities are of occurring events, and at the same time, the lower the probabilities of events are, which do not occur. The values assessed on events that turn out to be undetermined do not influence the score. Such a score $S(\mathbf{v})$ is an extension to partial and conditional probability assessments of the "total-log proper scoring rule" for probability distributions proposed by Lad in [20, pag. 355].

By considering the difference between the expected penalties suffered by the two evaluations $\mathbf{p}$ and $\mathbf{q}_{\boldsymbol{\alpha}}$ as distance criterion, it is possible to define the "discrepancy" $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ between a partial conditional assessment $\mathbf{p}$ over $\mathcal{E}$ and a distribution $\boldsymbol{\alpha} \in \mathcal{A}_0$ through the expression

$$\sum_{i|\alpha(H_i)>0} \alpha(H_i) \left(q_i \ln \frac{q_i}{p_i} + (1 - q_i) \ln \frac{(1 - q_i)}{(1 - p_i)} \right) \quad (5)$$

taking the convention $0 \ln(0) = 0$. Note that in $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ each term is weighted by $\alpha(H_i)$, which reflects the "relevance" of each context H_i with respect to all the assessments.

The main idea is to take as coherent correction of $\mathbf{p}$ the assessment $\mathbf{q}_\mathbf{p} \equiv \mathbf{q}_{\widetilde{\boldsymbol{\alpha}}}$ generated by the distribution $\widetilde{\boldsymbol{\alpha}}$ solution of the nonlinear optimization program

$$\min_{\boldsymbol{\alpha} \in \mathcal{A}^0} \Delta(\mathbf{p}, \boldsymbol{\alpha}). \quad (6)$$

The motivation for this choice is that (intuitively) the assessor of $\mathbf{p}$ would expect to suffer the penalty $S(\mathbf{p})$, hence we select the coherent assessment $\mathbf{q}_\mathbf{p}$ that has a (probabilistic) expected score as close as possible.

In [8] it is formally proved that $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ is a non negative function on $\mathcal{A}_0$ and that $\Delta(\mathbf{p}, \boldsymbol{\alpha}) = 0$ if and only if $\mathbf{p} = \mathbf{q}_{\boldsymbol{\alpha}}$; moreover $\Delta(\mathbf{p}, \cdot)$ admits a minimum on $\mathcal{A}_0$. Finally if $\boldsymbol{\alpha}, \boldsymbol{\alpha}^0 \in \mathcal{A}_0$ are distributions that minimize $\Delta(\mathbf{p}, \cdot)$, then for all $i \in \{1, \ldots, n\}$ such that $\alpha(H_i) > 0$ and $\alpha^0(H_i) > 0$ we have $(\mathbf{q}_\alpha)_i = (\mathbf{q}_{\alpha^0})_i$; in particular if $\Delta(\mathbf{p}, \cdot)$ attains its minimum value on $\mathcal{A}_1$ then there is a unique coherent assessment $\mathbf{q}_{\underline{\alpha}}$ such that $\Delta(\mathbf{p}, \underline{\boldsymbol{\alpha}})$ is minimum. On the contrary, if the minimum is attained in $\mathcal{A}_0 \setminus \mathcal{A}_1$, i.e. there exists some conditioning event forced to have null probability, the optimization program (6) can be iterated by restricting the assessment only on the different "zero layers" (for details refer again to [8]). Moreover, the discrepancy measure $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ can be used to correct incoherent assessments and to aggregate expert opinions [6, 8]. $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ can even be applied to correct incoherent assessments and to aggregate conflicting opinions based on imprecise conditional probabilities [7] but this feature will not be used here since the statistical matching analysis will be based on a precise initial assessment $\mathbf{p}$. Imprecise probabilities can appear whenever a consistent sub-assessment is selected

and it is coherently extended to the rest of the domain, as it is shown in the next sub-section.

3.1 Coherent Extension

Another possibility to adjust the initially incoherent assessment $(\mathcal{E}, \mathbf{p})$ could be to determine a coherent sub-assessment $(\mathcal{G}, \mathbf{p}_{|\mathcal{G}})$ and coherently extend it to the rest $\mathcal{F} = \mathcal{E} \setminus \mathcal{G}$ as prescribed by the generalized Bayesian updating scheme (see e.g. [9, 10, 36] among others). Since, in general, coherent extension produces intervals of plausible values, with this approach the whole assessment turns out to be imprecise due to the interval values $((\mathcal{F}, [\underline{\mathbf{p}_{\mathcal{F}}}, \overline{\mathbf{p}_{\mathcal{F}}}]))$. Also in such a situation, inference can pe performed again through the generalized Bayesian updating scheme but applied to imprecise evaluations (see e.g. [1, 4] among others). Whenever such inferences are too vague, i.e. when the intervals are very wide (close to [0,1]), they can be eventually reduced by a procedure proposed in [5] that enucleates coherent cores, i.e. surely coherent subintervals with highest degree of support. This is motivated by the fact that, in general, not all the subintervals of the extensions are coherent, whereas this is guaranteed by the choice of such coherent cores since they are *total* coherent (for this stronger consistency notion refer e.g. to [16]).

The choice of the coherent sub-assessment $(\mathcal{G}, \mathbf{p}_{|\mathcal{G}})$ should follow some criterion, since it could not be uniquely determined. Anyhow, for the specific application to statistical matching that is the scope of the present paper, such a choice comes quite naturally since in [35] it has been shown that it is possible to detect the incoherent sub-assessment $(\mathcal{F}, \mathbf{p}_{|\mathcal{F}})$ with minimal cardinality.

4 Integration of sources in a coherent setting

We briefly describe how the problem of integration of sources, named statistical matching, can be formalized in the coherent conditional probability setting. In particular here we refer to the case of two sources as already described in [35], while the case of more sources has been studied in [36].

Let us denote by $(X_1, Y_1), ..., (X_{n_A}, Y_{n_A})$ and by $(X_{n_A+1}, Z_{n_A+1}), ..., (X_{n_A+n_B}, Z_{n_A+n_B})$ two random samples (with a finite range) related to two sources A and B. We suppose that the two samples both concern the same population of interest and are drawn according to the same sampling scheme. We can regard, under the above conditions, $(X_1, Y_1), ..., (X_{n_A}, Y_{n_A})$ (analogously $(X_{n_A+1}, Z_{n_A+1}), ..., (X_{n_A+n_B}, Z_{n_A+n_B})$) exchangeable, as well as the sequence $X_1, ..., X_{n_A}$, $X_{n_A+1}, ..., X_{n_A+n_B}$.

We can elicit from the two files the relevant probability values: from file A the conditional probabilities

$$\mathcal{Y}_{j|i} = P_{Y|(X=x_i)}(Y = y_j), \tag{7}$$

that the next unit has $Y = y_j$ on the hypothesis that $(X = x_i)$ (for any x_i taken by X), and analogously from file B the conditional probability values

$$\mathcal{Z}_{k|i} = P_{Z|(X=x_i)}(Z = z_k). \tag{8}$$

Moreover, from data on both files we can evaluate

$$\mathcal{X}_i = P_X(X = x_i). \tag{9}$$

Given $\mathcal{Y}_{j|i}, \mathcal{Z}_{k|i}, \mathcal{X}_i$, for any i, j, k, one needs to check coherence of the whole assessment $(\mathcal{E}, \mathbf{p})$, that is

$$\begin{array}{l} \mathcal{E} = \\ \left\{ (X = x_i),\ (Y = y_j)|(X = x_i),\ (Z = z_k)|(X = x_i) \right\}, \\ \qquad\qquad\qquad\qquad \text{for any } x_i, y_j, z_k \\ \mathbf{p} = \{\mathcal{X}_i, \mathcal{Y}_{j|i}, \mathcal{Z}_{k|i}\}_{i,j,k} \quad . \end{array} \tag{10}$$

Now we recall the result proved in [36], that claims that when the partitions $\mathcal{E}_X, \mathcal{E}_Y, \mathcal{E}_Z$ associated to the variables are logically independent (i.e. for any $A \in \mathcal{E}_X$, $B \in \mathcal{E}_Y$, $C \in \mathcal{E}_Z$, $A \wedge B \wedge C \neq \emptyset$) coherence is assured.

Theorem 2 *Let X, Y, Z be three finite random variables and consider the following three coherent assessments $\{P_X(X = x_i)\}_i$, $\{P_{Y|X=x_i}(Y = y_j)\}_j$ and $\{P_{Z|X=x_i}(Z = z_k)\}_k$.*

Then the assessment

$$\{P_X(X = x_i)\,,\ P_{Y|X=x_i}(Y = y_j) :\ \textit{for any}\ x_i, y_j\}$$

(analogously $\{P_X(X = x_i), P_{Z|X=x_i}(Z = z_k) :$ for any $x_i, z_k\}$) is coherent.

Moreover, if the partitions $\mathcal{E}_Y, \mathcal{E}_Z$ are logically independent with respect to $\mathcal{E}_X$ (i.e. $(X = x_i, Y = y_j, Z = z_k)$ is possible for any value x_i of X, y_j of Y, z_k of Z s.t. the events $(X = x_i, Y = y_j)$ and $(X = x_i, Z = z_k)$ are possible), then the whole assessment (10) is coherent.

On the other hand, when there are some logical constraints among the variables Y and Z, the coherence of the whole assessment (10) is not assured by coherence of the single assessments (7-9) (see [35]). Notice that the need of managing logical constraints arises from practical applications [14].

5 Removing inconsistencies in statistical matching

We have now all the elements to specialize the general approaches for inconsistencies correction described in

Section 3 for the specific setting of the statistical matching as depicted in the previous Section 4.

The starting point is that the whole assessment (10) is not coherent, then inconsistencies must be detected in order to restore coherence. This kind of problem has already been studied (e.g. see [21]) in combining assessments given by different experts: the approach to the identification and reconciliation of incoherence uses an external observer equipped with a prior distribution and likelihood function. Actually, this approach does not seem suitable in the context of statistical matching because of the lack of information on the variables not jointly observed, so that the prior distribution cannot be updated and the likelihood function has a flat ridge (as already noted in [30]). Hence we propose a different method: to restore coherence we can easily find the minimal restriction of the whole assessment which is not coherent (as proposed in [36]) and adjust it by a specialization of the techniques presented in Section 3. Let us see it into details.

As claimed by Theorem 2, in statistical matching incoherences are related to conditional events with the same conditioning event $(X = x_i)$. Hence the check of coherence of the whole assessments (10) can be reduced to the check of coherence for the sub-assessments

$$\left\{\mathcal{Y}_{j|i},\, \mathcal{Z}_{k|i} : \text{ for fixed } i \text{ and any } j,k\right\}. \tag{11}$$

Once not coherent sub-assessments of type (11) have been disclosed, they can be adjusted by finding coherent values that minimize some of the (pseudo)distances presented in Section 3.

Whereas classical distances - $L1$, $L2$ and KL - can be directly applied to such minimal incoherent restriction since their arguments are directly the conditional probabilities, for the discrepancy $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ a "reformulation" is required. In fact, we require that its expression (5) specifically acts on values for any conditioning events $(X = x_i)$. This is possible by considering the following mixture of discrepancies $\Delta_{mix}(\mathbf{p}, \{\boldsymbol{\alpha}_i\}_i)$:

$$\sum_i \mathcal{X}_i \left[\sum_j \left(q_{j|i}^{\alpha_i} \ln \frac{q_{j|i}^{\alpha_i}}{\mathcal{Y}_{j|i}} + (1 - q_{j|i}^{\alpha_i}) \ln \frac{(1 - q_{j|i}^{\alpha_i})}{(1 - \mathcal{Y}_{j|i})}\right) + \right.$$
$$\left. + \sum_k \left(q_{k|i}^{\alpha_i} \ln \frac{q_{k|i}^{\alpha_i}}{\mathcal{Z}_{k|i}} + (1 - q_{k|i}^{\alpha_i}) \ln \frac{(1 - q_{k|i}^{\alpha_i})}{(1 - \mathcal{Z}_{k|i})}\right)\right] \tag{12}$$

where each distribution α_i works just on the sample space spanned by the conditional events $\{(Y = y_j)|(X = x_i), (Z = z_k)|(X = x_i)\}$, it is constrained to fulfill the normalizing condition

$$\alpha_i(X = x_i) = \mathcal{X}_i, \tag{13}$$

and generates the conditional probabilities

$$q_{j|i}^{\alpha_i} = \frac{\alpha_i(Y = y_j)}{\alpha_i(X = x_i)} \quad q_{k|i}^{\alpha_i} = \frac{\alpha_i(Z = z_k)}{\alpha_i(X = x_i)}. \tag{14}$$

As already mentioned, coherence of the overall assessment $(\mathcal{E}, \mathbf{q})$, with

$$\mathbf{q} = \{\mathcal{X}_i, q_{j|i}^{\alpha_i}, q_{k|i}^{\alpha_i}\}_{i,j,k}$$

is guaranteed by Theorem 2.

Since the specialized discrepancy defined in equation (12) is a mixture of discrepancies, each one working on a specific scenario $(X = x_i)$, its use in an optimization program like (6) allows to adjust only the values inside sub-domains of $\mathcal{E}$ conditioned to scenarios $(X = x_i)$ where some incoherence appear, without changing the other values. This characteristic differentiates the specialized discrepancy (12) from the original discrepancy (5), as the following simple example shows:

Example 2 *Let $\{x_1, x_2\}$, $\{y_1, y_2, y_3\}$, $\{z_1, z_2, z_3\}$ be the sample space of three r.v. X, Y, Z with constraints*

$$(Z = z_1) \wedge ((Y = y_1) \vee (Y = y_2)) = \emptyset$$

and

$$(Z = z_2) \wedge (Y = y_1) = \emptyset.$$

Consider the following conditional assessment $\mathbf{p}$:

$$\begin{array}{lll} \mathcal{X}_1 = \frac{1}{3} & \mathcal{X}_2 = \frac{2}{3} & \\ \mathcal{Y}_{1|1} = \frac{387}{1111} & \mathcal{Y}_{2|1} = \frac{102}{1111} & \mathcal{Y}_{3|1} = \frac{622}{1111} \\ \mathcal{Y}_{1|2} = \frac{2}{3} & \mathcal{Y}_{2|2} = 0 & \mathcal{Y}_{3|2} = \frac{1}{3} \\ \mathcal{Z}_{1|1} = \frac{179}{1108} & \mathcal{Z}_{2|1} = \frac{443}{1108} & \mathcal{Z}_{3|1} = \frac{486}{1108} \\ \mathcal{Z}_{1|2} = \frac{2}{3} & \mathcal{Z}_{2|2} = \frac{1}{9} & \mathcal{Z}_{3|2} = \frac{2}{9} \end{array} .$$

It is easy to check that $\mathbf{p}$ on events $(X = x_i)$ is coherent, as well as $\mathcal{Y}_{j|i} = P(Y = y_j|X = x_i)$ (and analogously $\mathcal{Z}_{k|i} = P(Z = z_k|X = x_i)$) for any $(X = x_i)$. However, the whole assessment is not coherent, and incoherence is localized on events conditioned to $(X = x_2)$.

By applying either $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ or $\Delta_{mix}(\mathbf{p}, \{\boldsymbol{\alpha}_i\}_i)$ the same correction on those values is induced (see Table 1), whereas with the former also the unconditional values for $P(X = x_i)$ are modified, even if they are coherent.

Note that, with such specialized discrepancy, the subdomains, where incoherence must be removed, are implicitly detected, without the need of a preliminary

$\mathcal{E}$	P	Δ	Δ_{mix}
$X = x_1$	0.3333	**0.3726**	-
$X = x_2$	0.6667	**0.6274**	-
$Y = y_1 \mid X = x_1$	0.3483	0.3483	0.3483
$Y = y_2 \mid X = x_1$	0.0918	0.0918	0.0918
$Y = y_3 \mid X = x_1$	0.5599	0.5599	0.5599
$Z = z_1 \mid X = x_1$	0.1616	0.1616	0.1616
$Z = z_2 \mid X = x_1$	0.3998	0.3998	0.3998
$Z = z_3 \mid X = x_1$	0.4386	0.4386	0.4386
$Y = y_1 \mid X = x_2$	0.6667	*0.4156*	*0.4156*
$Y = y_2 \mid X = x_2$	0	*0.0996*	*0.0996*
$Y = y_3 \mid X = x_2$	0.3333	*0.4848*	*0.4848*
$Z = z_1 \mid X = x_2$	0.6667	*0.4848*	*0.4848*
$Z = z_2 \mid X = x_2$	0.1111	*0.0996*	*0.0996*
$Z = z_3 \mid X = x_2$	0.2222	*0.4156*	*0.4156*

Table 1: Correction comparison for Example 2. In boldface changes associated to unconditional events, while in italic changes associated to conditional ones

inspection of the assessment $(\mathcal{E}, \mathbf{p})$. Moreover the adjustments are weighted by the relevance of the scenarios expressed through the $\varkappa_i$'s in (12).

From these data we can also get a comparison between Δ and Δ_{mix}: actually Δ also changes the probability distribution on $(X = x_i)$'s in order to reduce the minimum value taken from Δ even if Theorem 2 assures the separate coherence of the probability assessments $(\varkappa_i, \mathcal{Y}_{j|i})$ and $(\varkappa_i, \mathcal{Z}_{j|i})$, for any $i = 1, 2$ and $j = 1, 2, 3$. Then, we can stress that for the statistical matching problem Δ_{mix} seems to be more appropriate than Δ.

Another criterion (further than the quoted ones based on $L1, L2, KL$ minimizations) for restoring coherence could be based on the maximum likelihood criterion: when the evaluations are obtained through the maximum likelihood criterion, we can maximize the "partial likelihood function" on the set of events generating incoherence. Also in this situation we have an optimization problem with a non-linear objective function and a set of linear constraints.

Note that if we apply this criterion to data in Example 2 the marginal distribution of X does not change and the adjustment is localized on the assessment over $(X = x_2)$, analogously to what happens with Δ_{mix}. We have not reported these values on Table 1 because the aim of the example is just to stress the difference between Δ and Δ_{mix}. Explicit results of the maximum likelihood criterion will appear in the next section.

6 A practical example

In order to show our proposal we develop an example with data taken from [14] and studied also in [36]. The data are a subset of 2313 employees (people at least 15 years old) extracted from 2000 pilot survey of the Italian Population and Household Census. Three categorical variables have been analyzed: Age, Educational Level and Professional Status. In file A, containing 1148 units, the variables Age and Professional Status are observed, while file B, consisting of 1165 observations, the variables Age and Educational Level are considered. The variables are grouped in homogeneous response categories as follows: A_1=15-17 years old, A_2=18-22 years old, A_3=23-64 years old, A_4=more than 65 ; E_1=None or compulsory school, E_2=Vocational school, E_3=Secondary school, E_4=Degree; S_1=Manager, S_2=Clerk, S_3=Worker.

Logical constraints between the variables Age and Educational level (Age and Professional Status) are denoted by the symbol "–" (to be distinguished from the zero frequencies) in Table 2 (Table 3): for example, in Italy a 17 years old person cannot have a University degree. Tables 2 and 3 show, respectively, the distribution of Age and Professional Status in file A, and in file B that related to Age and Educational Level.

	Prof. Status			
Age	S_1	S_2	S_3	Tot.
A_1	–	–	9	9
A_2	–	5	17	22
A_3	179	443	486	1108
A_4	6	1	2	9
Tot.	185	449	514	1148

Table 2: Distribution of Age and Professional Status in file A.

	Educ. level				
Age	E_1	E_2	E_3	E_4	Tot.
A_1	6	0	–	–	6
A_2	14	6	13	–	33
A_3	387	102	464	158	1111
A_4	10	0	3	2	15
Tot.	417	108	480	160	1165

Table 3: Distribution of Age and Educational level in file B.

Additional logical constraints involving both the variables Professional Status and Educational level are

the following ones:

$$S_1 \wedge (E_1 \vee E_2) = \emptyset \text{ and } S_2 \wedge E_1 = \emptyset.$$

By considering the frequencies (that, whenever coherent, correspond also to the maximum likelihood estimations) as evaluation of the relevant conditional probabilities, we get the assessment reported in Table 4. Such conditional probability assessment is not

	A_1	A_2	A_3	A_4
$P(\cdot)$	0.0065	0.0238	0.9594	0.0104
$P(S_1\|\cdot)$	$--$	$--$	0.1616	0.6667
$P(S_2\|\cdot)$	$--$	0.2273	0.3913	0.1111
$P(S_3\|\cdot)$	1	0.7727	0.4293	0.2222
$P(E_1\|\cdot)$	1	0.4242	0.3419	0.6667
$P(E_2\|\cdot)$	0	0.1818	0.0918	0
$P(E_3\|\cdot)$	$--$	0.3940	0.4176	0.2
$P(E_4\|\cdot)$	$--$	$--$	0.1422	0.1333

Table 4: Conditional probability assessment elicited from frequencies of Tab.2 and Tab.3.

coherent as shown in [36]. The incoherencies need to be identified and removed. It comes out that $P(\cdot|A_4)$ is not coherent since from logical constraints between Educational Level and Professional Status it follows $E_1 \wedge S_1 = \emptyset$ and $E_1 \subseteq S_3$, respectively, while from Table 4 result $P(E_1|A_4) + P(S_1|A_4) + P(S_3|A_4) > 1$ and $P(E_1|A_4) > P(S_1|A_4)$.

Then, we could either identify, as proposed in [36], the minimal set of conditional events involved in incoherencies that is $\mathcal{F} = \{E_1|A_4, S_1|A_4, S_3|A_4\}$, or adjust, with respect to a suitable distance, the whole distribution on Professional Status and Educational Level conditioned to A_4.

Different corrections are considered and the results are shown in Table 5, where

- $L1_{|\mathcal{F}}$ gives the solution proposed in [36] by minimizing $L1$ distance only among $\mathcal{F}$, the minimal incoherent subset of $\mathcal{E}$;
- $L1_{|A4}$, $L2_{|A4}$, $KL_{|A4}$ gives the solutions obtained by minimizing usual distances discussed in Section 3 only among events conditioned to A_4;
- Δ_{MIX} gives the solution obtained by minimizing the specific discrepancy (12);
- ML gives the maximum likelihood estimation;
- $IP_{\mathcal{E}\setminus\mathcal{F}}$ gives the coherent lower-upper extension induced by the given assessment on $\mathcal{E} \setminus \mathcal{F}$;
- $IP_{\mathcal{E}\setminus\{\cdot|A_4\}}$ gives the coherent lower-upper extension induced by the given assessment on $\mathcal{E} \setminus \{S_i|A_4, E_j|A_4 : i = 1, 2, 3\,;\, j = 1, ..., 4\}$;
- the last column gives the extensions of the respective corrections on the inference target $S_3|E_4$ with the respective "core" rows showing the total coherent sub-interval extension with maximum support in line with [5].

Note that only the values conditioned to A_4 are reported, those involved in the incoherence (the other 18 values remaining the same as the given assessment $\mathbf{p}$).

Firstly, we compare the rows related to remove the minimal set of incoherence, and it seems that $L1_{|\mathcal{F}}$ and $IP_{\mathcal{E}\setminus\mathcal{F}}$ perform similarly. Even though we can observe a drastic change on the probability values, mainly induced by removing not all the set of conditioning events with conditioning A_4 but just a subset (a minimal subset), they induce quite reasonable inference bounds. In particular, the imprecise adjustment $IP_{\mathcal{E}\setminus\mathcal{F}}$ performs quite well. In fact it induces inference bounds for $S_3|E_4$ similar to the precise corrections with the advantage of having the possibility to enucleate the "core" sub-interval. This sub-interval, even though it remains quite vague, has the positive aspect of bounding away from zero the lower probability, and this is seen very often as a positive aspect.

Note that $L1_{|A4}$ and ML give similar results and in particular they leave to 0 the probability of $E_2|A_4$ since the absence of observations in the original data. And the impossibility to change null values is one of the peculiarities of maximum likelihood criterion.

On the other hand, we observe that precise adjustments on the whole assessment conditioned to A_4 have all quite similar behaviors for the other distances taken into consideration, and in particular they also modify the assessment related to $E_2|A_4$, where there is no observation.

The advantage of Δ_{mix} correction is its automatic localization of the scenarios (in this specific example A_4) where the adjustment can be performed and their relative importance expressed by the unconditional probabilities x_i. Note that we apply Δ_{mix}, instead of Δ, in order to avoid any change on the probability distribution of X, that is coherent with any conditional probability on $Y|(X = x)$ (or equivalently $Z|(X = x)$), for any x, as shown in Theorem 2. In fact, Δ tighten to change also the distribution of X (through the weights) in order to reduce the inconsistencies, as shown in Example 2.

On the other hand, the wider imprecise correction $IP_{\mathcal{E}\setminus\{\cdot|A_4\}}$, being the one with less assumption requirement, surely performs worst. Its vagueness on the values conditioned on A_4 is due to freedom induced by the coherence characterization, and this re-

	$S_1\|A_4$	$S_2\|A_4$	$S_3\|A_4$	$E_1\|A_4$	$E_2\|A_4$	$E_3\|A_4$	$E_4\|A_4$	$S_3\|E_4$
p	0.6667	0.1111	0.2222	0.6667	0	0.2000	0.1333	$\emptyset$
$L1_{\|\mathcal{F}}$	0.2222	-	0.6667	0.6667	-	-	-	[0,0.6285]
$L1_{\|A4}$	0.5266	0	0.4734	0.4734	0	0.2836	0.2431	[0,0.6234]
$L2_{\|A4}$	0.5333	0.0389	0.4278	0.4278	0.0389	0.3	0.2333	[0,0.6238]
$KL_{\|A4}$	0.4856	0.1179	0.3965	0.3965	0.1179	0.2914	0.1942	[0,0.6257]
Δ_{mix}	0.4985	0.0939	0.4077	0.4077	0.0939	0.2943	0.2042	[0,0.6252]
ML	0.4286	0.0714	0.5000	0.5000	0	0.3000	0.2000	[0,0.6254]
$IP_{\mathcal{E}\setminus\mathcal{F}}$	[0 , 0.2222]	-	[0.6667 0.8889]	-	-	-	-	[0,0.6386]
core								[0.0017,0.6286]
$IP_{\mathcal{E}\setminus\{\cdot\|A_4\}}$	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0,0.6607]
core								[0,0.6349]

Table 5: Several incoherence correction with associated inference results for the target $S3|E4$

flects also on the inference performances.

Note that we report, just as an example, only the extension values for a conditional event, however we could compute all the values of the (conditional) events of interest, as for example on the partition generated by the three random variables.

7 Conclusion

Checking coherence and removing incoherences in the data is a long debated problem in literature, we have studied it by focusing on statistical matching applications. In fact, in this kind of application the incoherence can arise when the variables are linked by logical relations.

We have applied several incoherence adjustment procedures in this specific ambit. From this study some differences among these adjustments come out. Due to peculiarities of source integration and lack of information on the variables not jointly observed, usual divergences techniques can be specialized. In particular, a specific adjustment of a discrepancy, originally introduced for general conditional probability assessment, shows the advantage of an automatic and weighted localization of the sub-domains where incoherence must be removed.

We have analyzed also a very simple practical application and we have shown that better results are obtained not simply focusing on the minimal number of incoherent values, but involving all the elements conditioned to the same scenarios, where incoherence arises. On the other hand, coherent imprecise adjustment performs better focusing on the minimal number of incoherent values. This entail a minimal number of changes with respect the original assessment, but has as counterpart obvious vaguer inference conclusions. Vagueness that can however be reduced by the aforementioned "maximally supported" sub-intervals detection.

References

[1] V. Biazzo and A. Gilio: A generalization of the fundamental theorem of de Finetti for imprecise conditional probability assessments. *Int. J. of Approximate Reasoning*, 24: 251-272, 2000.

[2] V. Biazzo, A. Gilio. Some Theoretical Properties of Conditional Probability Assessments. In *LNAI (Lecture Notes in Computer Science)* 3571: 775–787, 2005.

[3] E.C. Budd. The creation of a microdata file for estimating the distribution of income. *Review of Income and Wealth*, 17: 317–333, 1971.

[4] A. Capotorti, L. Galli and B. Vantaggi. How to use locally strong coherence in an inferential process based on upper-lower probabilities. *Soft Computing*, 7(5): 280–287, 2003.

[5] A. Capotorti and M. Zagoraiou. Implicit Degree of Support for Finite Lower-Upper Conditional Probabilities Extensions, in *Proceed. of Information Processing and Management of Uncertainty in Knowledge-based Systems, IPMU'06*, EDK Paris - France, vol III: 2331– 2338, 2006.

[6] A. Capotorti and G. Regoli. Coherent correction of inconsistent conditional probability assessments. Proc. of IPMU'08 - Malaga (Es), 2008.

[7] A. Capotorti, G. Regoli and F. Vattari. On the use of a new discrepancy measure to correct incoherent assessments and to aggregate conflicting opinions based on imprecise conditional probabilities. Proc. of ISIPTA'09 - Durham (UK), 2009.

[8] A. Capotorti, G. Regoli and F. Vattari. Correction of incoherent conditional probability assessments. *International Journal of Approximate Reasoning*, 51(6): 718–727, 2010.

[9] G. Coletti. Coherent numerical and Ordinal probabilistic assessments. *IEEE Transaction on Systems, Man, and Cybernetics*, 24: 1747-1754, 1994.

[10] G. Coletti, R. Scozzafava. *Probabilistic Logic in a Coherent Setting*, Dordrecht: Kluwer, Series "Trends in Logic", 2002.

[11] G. De Cooman, and M. Zaffalon. Updating beliefs with incomplete observations. *Artificial Intelligence*, 159: 75-125, 2004.

[12] B. de Finetti. Sull'impostazione assiomatica del calcolo delle probabilità. *Annali Univ. Trieste*, 19:3-55, 1949. - Engl. transl. in: Ch.5 of *Probability, Induction, Statistics*. Wiley, London, 1972.

[13] M. D'Orazio, M. Di Zio and M. Scanu. Statistical Matching for Categorical Data: displaying uncertainty and using logical constraints. *Journal of Official Statistics*, in press.

[14] M. D'Orazio, M. Di Zio and M. Scanu. *Statistical Matching: Theory and Practice*. Wiley, 2006.

[15] L.E. Dubins. Finitely additive conditional probabilities, conglomerability and disintegration. *The Annals of Probability*, 3:89–99, 1975.

[16] A. Gilio and S. Ingrassia. Totally coherent set-valued probability assessments. *Kybernetika*, 34(1): 3–15, 1998.

[17] J. B. Kadane. Some statistical problems in merging data files. *Journal of Official Statistics*, 17: 423–433, 2001.

[18] P.H. Krauss. Representation of Conditional Probability Measures on Boolean Algebras. *Acta Math. Acad. Scient. Hungar.*, 19:229–241, 1968.

[19] S. Kulback,*Information Theory and Statistics*, New York: John Wiley, 1957.

[20] F. Lad, *Operational Subjective Statistical Methods: a mathematical, philosophical, and historical introduction*, New York: John Wiley, 1996.

[21] D.V. Lindley, A. Tversky, R.V. Brown. On the reconciliation of probability assessments. *Journal of Royal Statistical Society* Ser. A 142(2): 146–180, 1979.

[22] C.F. Manski, *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press, 1995.

[23] B. A. Okner. Data matching and merging: an overview. *Annals of Economic and Social Measurement*, 3(2): 347–352, 1974.

[24] E. Miranda, Updating coherent previsions on finite spaces, *Fuzzy Sets and Systems*, 160(9): 1286–1307, 2009.

[25] G. Paass. Statistical match: evaluation of existing procedures and improvements by using additional information. In *Microanalytic Simulation Models to Support Social and Financial Policy*, (Eds. G.H. Orcutt and H. Quinke) Elsevier Science, Amsterdam, 401–422, 1986.

[26] S. Rässler. Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches. *Lecture Notes in Statistics*, Springer Verlag, 2002.

[27] A. Rényi. On a new axiomatic theory of probability. *Acta mathematica Academiæ Scientiarium Hungaricæ*, 6:285–335. 1955.

[28] R. H. Renssen. Use of statistical matching techniques in calibration estimation. *Survey Methodology*, 24: 171–183, 1998.

[29] W. L. Rodgers. An evaluation of statistical matching. *Journal of Business and Economic Statistics*, 2(1): 91–102, 1984.

[30] D.B. Rubin. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 2(1): 87–94, 1986.

[31] C. A. Sims. Comments (on Okner). *Annals of Economic and Social Measurement*, 1(3): 343–345, 1972.

[32] A. C. Singh, H. J.Mantel, M. D. Kinack and G. Rowe. Statistical matching: use of auxiliary information as an alternative to conditional independence assumption. *Survey Methodology*, 19(1): 59–79, 1993.

[33] P. Szivós, T. Rudas, I.G. Tóth. A tax-benefit microsimulation model for Hungary. *Workshop on Microsimulation in the New Millennium: Challenges and Innovations*, Cambridge, 1998.

[34] J.L. Schafer, *Analysis of incomplete multivariate data*. Chapman & Hall, London, 1997.

[35] B. Vantaggi. The role of coherence for the integration of different sources. Proc. 4th International Symposium on Imprecise Probabilities and their Applications ISIPTA'05 (Pittsburgh, USA), 369–378, 2005.

[36] B. Vantaggi. Statistical matching of multiple sources: A look through coherence, *International Journal of Approximate Reasoning*, 49(3):701–711, 2008.

Regression with Imprecise Data: A Robust Approach

Marco E. G. V. Cattaneo
Department of Statistics, LMU Munich
cattaneo@stat.uni-muenchen.de

Andrea Wiencierz
Department of Statistics, LMU Munich
andrea.wiencierz@stat.uni-muenchen.de

Abstract

We introduce a robust regression method for imprecise data, and apply it to social survey data. Our method combines nonparametric likelihood inference with imprecise probability, so that only very weak assumptions are needed and different kinds of uncertainty can be taken into account. The proposed regression method is based on interval dominance: interval estimates of quantiles of the error distribution are used to identify plausible descriptions of the relationship of interest. In the application to social survey data, the resulting set of plausible descriptions is relatively large, reflecting the amount of uncertainty inherent in the analyzed data set.

Keywords. Robust regression, imprecise data, nonparametric statistics, likelihood inference, imprecise probability distributions, survey data, informative coarsening, complex uncertainty, interval dominance, identification regions.

1 Introduction

Data are often available only with limited precision. However, only few general methods for analyzing the relationships between imprecisely observed variables have been proposed so far. These approaches seem to fall in two categories. One of them consists of approaches suggesting to apply standard regression methods to all possible precise data compatible with the observations, and to consider the range of outcomes as the imprecise result: see for example [8]. The approaches in the second category consist in representing the imprecise observations by few precise values (for example, intervals by center and width), and in applying standard regression methods to those values: see for instance [7].

In the present paper, we follow another line of approach and suggest a new regression method directly applicable to the imprecise data. This method combines likelihood inference with imprecise probability. It allows to take into account different kinds of uncertainty, that are also reflected in the imprecise results of the regression. The suggested method imposes only very weak assumptions and yields extremely robust results. In particular, it is nonparametric, in the sense that no assumptions about the error distribution are necessary, in contrast, for instance, to the approach of [20]. We describe the regression method in Section 3, which is based on the general methodology for inference with imprecise data introduced in Section 2.

In addition to the theoretical results, in Section 4 we apply the method to analyze an interesting question in the social sciences. We investigate the relationship between age and income on the basis of survey data. The source of data used in this paper is "Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) — German General Social Survey" of 2008. The data is provided by GESIS — Leibniz Institute for the Social Sciences.

2 Imprecise Data

Let $V_1, \dots, V_n$ be n random objects taking values in a set $\mathcal{V}$, and let $V_1^*, \dots, V_n^*$ be n random sets taking values in a set $\mathcal{V}^* \subseteq 2^{\mathcal{V}}$, such that the events $V_i \in V_i^*$ are measurable. We are actually interested in the data V_i, but we can only observe the imprecise data V_i^*. The connection between precise and imprecise data is established by the following assumptions about the probability measures considered as models of the situation.

For each $\varepsilon \in [0,1]$, let $\mathcal{P}_\varepsilon$ be the set of all probability measures[1] P such that the n random objects $(V_1, V_1^*), \dots, (V_n, V_n^*)$ are independent and identically distributed and satisfy

$$P(V_i \in V_i^*) \geq 1 - \varepsilon. \tag{1}$$

[1] Probability measures and random objects are defined on an underlying measurable space.

We assume that the precise and imprecise data can be modeled by a probability measure P included in a particular set $\mathcal{P} \subseteq \mathcal{P}_\varepsilon$, for some $\varepsilon \in [0, 1]$. Each $P \in \mathcal{P}$ can be identified with a particular joint distribution for V_i and V_i^* (that is, the precise and imprecise data, respectively) satisfying condition (1). In particular, $\mathcal{P} = \mathcal{P}_\varepsilon$ corresponds to the fully nonparametric assumption that any joint distribution for V_i and V_i^* satisfying condition (1) is a possible model of the situation (this is the assumption we consider in Sections 3 and 4). The usual choice for the value of ε is 0 (see for example [6, 17]), which corresponds to an assumption of correctness of the imprecise data: $V_i^* = A$ implies $V_i \in A$ (a.s.). However, this assumption is often too strong: some imprecise data can be incorrect, in the sense that $V_i^* = A$, but $V_i \notin A$. This is for example the case, when the imprecise data represent the classification of the precise data into categories, and some observations are misclassified. By choosing a positive value for ε, we allow each imprecise observation to be incorrect with probability at most ε.

The set $\mathcal{V}^*$ describes which imprecise data $V_i^* = A$ are considered as possible. As extreme cases we have the actually precise data (when A is a singleton) and the missing data (when $A = \mathcal{V}$). In general, the fully nonparametric assumption $\mathcal{P} = \mathcal{P}_\varepsilon$ does not exclude informative coarsening (see for example [23]): parametric models or uninformative coarsening can be imposed by a stronger assumption $\mathcal{P} \subset \mathcal{P}_\varepsilon$. However, it is important to note that the set $\mathcal{P}_\varepsilon$ depends strongly on the choice of $\mathcal{V}^*$. For example, when $\varepsilon = 0$, the choice of a set $\mathcal{V}^*$ such that its elements build a partition of $\mathcal{V}$ implies the assumption that the coarsening is deterministic and uninformative, because each possible precise data value is contained in exactly one possible imprecise observation $A \in \mathcal{V}^*$.

2.1 Complex Uncertainty

In general, we are uncertain about which of the probability measures in $\mathcal{P}$ is the best model of the reality under consideration. Our uncertainty is composed of two parts. On the one hand, we are uncertain about the distribution of the imprecise data V_i^*: this uncertainty decreases when we observe more and more (imprecise) data. On the other hand, even if we (asymptotically) knew the distribution of the imprecise data V_i^*, we would still be uncertain about the distribution of the (unobserved) precise data V_i: this uncertainty is unavoidable. To formulate this mathematically, let P_V and P_{V^*} be the marginal distributions of V_i and V_i^*, respectively, corresponding to the probability measure $P \in \mathcal{P}$. There is uncertainty about P_{V^*} in the set[2] $\mathcal{P}_{V^*} := \{P'_{V^*} : P' \in \mathcal{P}\}$, but even if P_{V^*} were known, there would still be an unavoidable uncertainty about P_V in the set

$$[P_{V^*}] := \{P'_V : P' \in \mathcal{P},\, P'_{V^*} = P_{V^*}\}.$$

The sets $[P_{V^*}]$ with $P_{V^*} \in \mathcal{P}_{V^*}$ are the identification regions for P_V in the terminology of [12]. Each of them consists of all the distributions for the precise data V_i compatible with a particular distribution for the imprecise data V_i^*. Hence, each set $[P_{V^*}]$ can be interpreted as an imprecise probability distribution on $\mathcal{V}$. By observing the realizations of the imprecise data V_i^*, we learn something about which of the imprecise probability distributions $[P_{V^*}]$ is the best model for the (unobserved) precise data V_i.

Example 1 *Let $\mathcal{V} = \{0, 1\}$ and $\mathcal{V}^* = 2^{\{0,1\}}$, and assume $\mathcal{P} = \mathcal{P}_\varepsilon$ for some $\varepsilon \in [0, 1]$. Then $\mathcal{P}_{V^*}$ is the set of all probability distributions on $2^{\{0,1\}}$ such that the probability of $\varnothing$ is at most ε. For each $P_{V^*} \in \mathcal{P}_{V^*}$, the identification region $[P_{V^*}]$ is the set of all probability distributions on $\{0, 1\}$ such that the probability of 1 lies in the interval $\left[\underline{P}_{V^*}\{1\},\, \overline{P}_{V^*}\{1\}\right]$, with*

$$\underline{P}_{V^*}\{1\} = \max\left(P_{V^*}\left\{\{1\}, \varnothing\right\} - \varepsilon,\, 0\right)$$
$$\overline{P}_{V^*}\{1\} = \min\left(P_{V^*}\left\{\{1\}, \{0, 1\}\right\} + \varepsilon,\, 1\right).$$

In particular, when $\varepsilon = 0$, the imprecise probability distribution $[P_{V^*}]$ corresponds to the belief function on $\{0, 1\}$ with basic probability assignment P_{V^*} (see for example [16]), in the sense that $[P_{V^*}]$ is the set of all probability distributions on $\{0, 1\}$ dominating that belief function.

2.2 Likelihood

The likelihood function is a central concept in statistical inference. For parametric probability models, it is usually expressed as a function of the parameters: here we consider the more general formulation (as a function of the probability measures), which is applicable also to nonparametric models (see for example [14]). The observed (imprecise) data $V_1^* = A_1, \ldots, V_n^* = A_n$ induce the (normalized) likelihood function $lik : \mathcal{P} \to [0, 1]$ defined by

$$lik(P) = \frac{P(V_1^* = A_1, \ldots, V_n^* = A_n)}{\sup_{P' \in \mathcal{P}} P'(V_1^* = A_1, \ldots, V_n^* = A_n)} = \frac{\prod_{i=1}^n P_{V^*}\{A_i\}}{\sup_{P' \in \mathcal{P}} \prod_{i=1}^n P'_{V^*}\{A_i\}}$$

for all $P \in \mathcal{P}$. The likelihood function describes the relative ability of the probability measures P in predicting the observed (imprecise) data. Therefore, the value $lik(P)$ depends only on the marginal distribution P_{V^*} of the imprecise data V_i^*. The likelihood

[2] The symbol := denotes "is defined to be".

function can be interpreted as the second level of a hierarchical model for imprecise probabilities, with $\mathcal{P}$ as first level (see for example [4, 5]). In particular, for any $\beta \in (0,1)$, the likelihood function can be used to reduce $\mathcal{P}$ to the set

$$\mathcal{P}_{>\beta} := \{P \in \mathcal{P} : lik(P) > \beta\}$$

of all the probability measures that were sufficiently good in predicting the observed (imprecise) data.

Let g be a multivalued mapping[3] from $\mathcal{P}$ to a set $\mathcal{G}$, describing a particular characteristic (in which we are interested) of the models considered. For example, g can be the multivalued mapping from $\mathcal{P}$ to $\mathbb{R}$ assigning to each probability measure P the p-quantile of the distribution of $h(V_i)$ under P, for some $p \in (0,1)$ and some measurable function $h : \mathcal{V} \to \mathbb{R}$. This is the kind of mapping g we consider in Sections 3 and 4: it is multivalued, because in general quantiles are not uniquely defined[4]. For each $\beta \in (0,1)$, the set

$$\mathcal{G}_{>\beta} := \bigcup_{P \in \mathcal{P}_{>\beta}} g(P)$$

is called likelihood-based confidence region with cutoff point β for the values of the multivalued mapping g. This confidence region consists of all values that the characteristic described by g takes on the set $\mathcal{P}_{>\beta}$ of all the probability measures that were sufficiently good in predicting the observed (imprecise) data.

The unique function $lik_g : \mathcal{G} \to [0,1]$ describing these confidence regions, in the sense that

$$\mathcal{G}_{>\beta} = \{\gamma \in \mathcal{G} : lik_g(\gamma) > \beta\}$$

for all $\beta \in (0,1)$, is called (normalized) profile likelihood function induced by the multivalued mapping g. It can be easily checked that[5] for all $\gamma \in \mathcal{G}$,

$$lik_g(\gamma) = \sup_{P \in \mathcal{P} : \gamma \in g(P)} lik(P).$$

Example 2 *In the situation of Example 1, let $\varepsilon = 0$, and consider the mapping[6] g from $\mathcal{P}$ to $[0,1]$ assigning to each probability measure P the probability $P_V\{1\}$ that a precise data value V_i is 1 (before observing the corresponding imprecise data value V_i^*). The induced profile likelihood function[7] lik_g on $[0,1]$ is plotted in Figure 1 for the cases in which the imprecise data $\{0\}$, $\{1\}$, and $\{0,1\}$ have been observed 11, 21, and 6 times, respectively (solid line), and 213, 651, and 98 times, respectively (dashed line).*

In these two cases, the likelihood-based confidence regions with cutoff point $\beta = 0.15$ for the probability $P_V\{1\}$ are approximately the intervals $[0.39, 0.84]$ and $[0.65, 0.80]$, respectively (the cutoff point $\beta = 0.15$ is represented by the dotted line in Figure 1). They are (conservative) confidence intervals of approximate level 95% (see for example [11]).

[3] Mathematically, $g : \mathcal{P} \to 2^{\mathcal{G}} \setminus \{\varnothing\}$, but g is interpreted as an "imprecise" mapping from $\mathcal{P}$ to $\mathcal{G}$.

[4] A p-quantile of the distribution of $h(V_i)$ is any value $q \in \mathbb{R}$ such that $P(h(V_i) < q) \le p \le P(h(V_i) \le q)$.

[5] In this paper, $\sup \varnothing = 0$.

[6] As a multivalued mapping, g is defined by $g(P) = \{P_V\{1\}\}$ for all $P \in \mathcal{P}$.

[7] The details of the calculation of lik_g are not of primary interest at this point.

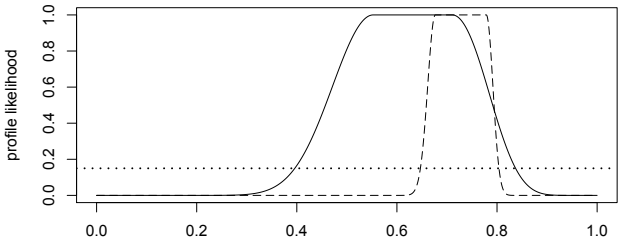


Figure 1: Profile likelihood functions from Examples 2 and 3.

2.3 Likelihood for Imprecise Data Models

In the situation we consider, we are actually interested in the (unobserved) precise data V_i. In this case, the characteristic of interest (described by g) depends only on the marginal distribution P_V of the precise data V_i; that is, we can write $g(P) =: g'(P_V)$ for all $P \in \mathcal{P}$. For example, the p-quantile of the distribution of $h(V_i)$ depends only on the distribution of V_i. By contrast, as noted at the beginning of Subsection 2.2, the value $lik(P)$ depends only on the marginal distribution P_{V^*} of the imprecise data V_i^*. By writing $lik(P) = lik^*(P_{V^*})$ for all $P \in \mathcal{P}$, we define a function $lik^* : \mathcal{P}_{V^*} \to [0,1]$, which can be interpreted as the likelihood function on $\mathcal{P}_{V^*}$.

In order to obtain the profile likelihood function lik_g, it can be useful to consider the multivalued mapping g^* from $\mathcal{P}_{V^*}$ to $\mathcal{G}$ defined by

$$g^*(P_{V^*}) = \bigcup_{P_V \in [P_{V^*}]} g'(P_V)$$

for all $P_{V^*} \in \mathcal{P}_{V^*}$. The multivalued mapping g^* assigns to each P_{V^*} all the values that the characteristic described by g' takes on the set $[P_{V^*}]$ of all distributions for the precise data V_i compatible with the distribution P_{V^*} for the imprecise data V_i^*. That is, g^* can be interpreted as an imprecise version of g', assigning to each imprecise probability distribution $[P_{V^*}]$ the corresponding imprecise value of g'.

The multivalued mapping g^* can be useful to obtain the profile likelihood function lik_g because, as can be

easily checked,

$$lik_g(\gamma) = \sup_{P_{V^*} \in \mathcal{P}_{V^*} : \gamma \in g^*(P_{V^*})} lik^*(P_{V^*})$$

for all $\gamma \in \mathcal{G}$. The right-hand side of this expression can be interpreted as the value $lik^*_{g^*}(\gamma)$ of the profile likelihood function $lik^*_{g^*}$ induced by the multivalued mapping g^*, when lik^* is considered as the likelihood function on $\mathcal{P}_{V^*}$.

The profile likelihood function $lik^*_{g^*}$ is particularly interesting, because lik^* describes the uncertainty about the distribution P_{V^*} of the imprecise data V_i^*, which decreases when we observe more and more (imprecise) data, while g^* describes the unavoidable uncertainty about the values of the multivalued mapping g'. In the terminology of [12], the values of g^* are the identification regions for the values of the multivalued mapping g.

Example 3 *The imprecise version g^* of the mapping g of Example 2 is the multivalued mapping from $\mathcal{P}_{V^*}$ to $[0,1]$ assigning to each P_{V^*} the interval*

$$\left[\underline{P}_{V^*}\{1\},\, \overline{P}_{V^*}\{1\}\right] = \left[P_{V^*}\left\{\{1\}\right\},\, P_{V^*}\left\{\{1\},\{0,1\}\right\}\right].$$

That is, $g^*(P_{V^*})$ is the interval probability that a precise data value V_i is 1 (before observing the corresponding imprecise data value V_i^*) according to the imprecise probability distribution $[P_{V^*}]$ (i.e., the belief function on $\{0,1\}$ with basic probability assignment P_{V^*}).

The profile likelihood function $lik_g = lik^*_{g^*}$ on $[0,1]$ is plotted in Figure 1 for the two cases considered in Example 2. In the case with 38 data (solid line) there is uncertainty also about the distribution P_{V^*} of the imprecise data V_i^*, while in the case with 962 data (dashed line) almost only the unavoidable uncertainty described by g^* remains, in the sense that $lik^*_{g^*}$ is almost equal to the indicator function of an identification region for $P_V\{1\}$ (i.e., of a probability interval $\left[\underline{P}_{V^*}\{1\},\, \overline{P}_{V^*}\{1\}\right]$).

3 Regression

Now consider that the (unobservable) precise data are pairs $V_i = (X_i, Y_i)$, where $X_1, \ldots, X_n$ are n random objects taking values in a set $\mathcal{X}$, and $Y_1, \ldots, Y_n$ are n random variables, with $\mathcal{V} = \mathcal{X} \times \mathbb{R}$. For some $\mathcal{V}^* \subseteq 2^{\mathcal{X}\times\mathbb{R}}$ and some $\varepsilon \in [0,1]$, we consider the fully nonparametric assumption $\mathcal{P} = \mathcal{P}_\varepsilon$. In the remainder of the paper, we focus on this setting.

We want to describe the relation between X_i and Y_i by means of a function $f \in \mathcal{F}$, where $\mathcal{F}$ is a particular set of (measurable) functions $f : \mathcal{X} \to \mathbb{R}$. In order to assess the quality of the description by means of f, we define the (absolute) residuals

$$R_{f,i} := |Y_i - f(X_i)|\,.$$

The n random variables $R_{f,1}, \ldots, R_{f,n} \in [0, +\infty)$ are independent and identically distributed: the more their distribution is concentrated near 0, the better is the description by means of f.

In order to compare the quality of the descriptions by means of different functions $f \in \mathcal{F}$, we need to compare the concentration near 0 of the distributions of the corresponding residuals $R_{f,i}$. Usual choices of measures for this concentration are the second and first moments $E(R_{f,i}^2)$ and $E(R_{f,i})$, respectively. However, the moments of the distribution of the residuals cannot be really estimated in the fully nonparametric setting we consider, because moments are too sensitive to small variations in the distribution (see also Subsection 4.2). In fact, if $\varepsilon > 0$ or the set

$$\mathcal{R}_f := \{|y - f(x)| : (x,y) \in A,\, A \in \mathcal{V}^*\}$$

is unbounded, then the likelihood-based confidence region for any particular moment of the distribution of the residuals is unbounded (even when only the distributions with finite moments are considered), independently of the cutoff point and of the observed (imprecise) data.

By contrast, the quantiles of the distribution of the residuals can in general be estimated even in the fully nonparametric setting we consider. Therefore, we propose to use the p-quantile of the distribution of the residuals $R_{f,i}$ as a measure of the concentration near 0 of this distribution, for some $p \in (0,1)$. The technical details of the estimation of such quantiles are given in Subsections 3.1 and 3.2.

The minimizations of the second and first moments of the distribution of the residuals can be interpreted as the theoretical counterparts of the methods of least squares and least absolute deviations, respectively. In the same sense, the minimization of the p-quantile of the distribution of the residuals can be interpreted as the theoretical counterpart of the method of least quantile of squares (or absolute deviations), introduced in [15] as a generalization of the method of least median of squares (corresponding to the choice $p = 0.5$). The method of least quantile of squares leads to robust regression estimators, with breakdown point $\min\{p, 1-p\}$ (that is, the highest possible breakdown point 50% is reached when $p = 0.5$). By contrast, the methods of least squares and least absolute deviations lead to regression estimators with breakdown point 0, since they cannot even handle a single outlier (including leverage points).

In the location problem (that is, when $\mathcal{F}$ is the set of all constant functions $f : \mathcal{X} \to \mathbb{R}$), the values of the constant functions f minimizing the second and first moments of the distribution of the residuals $R_{f,i}$ are the mean and median of the distribution of Y_i, respectively (when these exist and are unique). The value of the constant function f minimizing the p-quantile of the distribution of the residuals $R_{f,i}$ is the p-center of the distribution of Y_i (that is, the center of the shortest interval containing Y_i with probability at least p), when this exists and is unique. The p-center can be interpreted as a generalization of the mode of a distribution, since under some regularity conditions the mode corresponds to the limit of the p-center when p tends to 0. The p-center of a symmetric, strictly unimodal distribution corresponds to its median and mean (when this exists), independently of p. Therefore, the minimizations of the p-quantile, first moment, and second moment of the distribution of the residuals lead to the same (correct) regression function, under the usual assumptions for the error distribution: see for example [18].

3.1 Determination of Profile Likelihood Functions for Quantiles of Residuals

We want to determine the likelihood-based confidence regions for the quantiles of the distribution of the residuals: to this purpose, we calculate the profile likelihood function for such quantiles. Let $p \in (0,1)$, and for each function $f \in F$, let $\mathcal{Q}_f := \mathcal{L}_f \cap \mathcal{U}_f$, with

$$\mathcal{L}_f = \bigcup_{r \in \mathcal{R}_f} [r, +\infty)$$

when $p > \varepsilon$ and $\mathcal{L}_f = [0, +\infty)$ otherwise, while

$$\mathcal{U}_f = \bigcup_{r \in \mathcal{R}_f} [0, r]$$

when $p < 1-\varepsilon$ and $\mathcal{U}_f = [0, +\infty)$ otherwise. It can be easily checked that $\mathcal{Q}_f$ is the set of all possible values for the p-quantile of the distribution of the residuals $R_{f,i}$, since $P(R_{f,i} \notin \mathcal{R}_f) \leq \varepsilon$. In particular, if $\varepsilon < p < 1-\varepsilon$, then $\mathcal{Q}_f$ is the smallest interval containing $\mathcal{R}_f$.

For each $f \in F$, let Q_f be the multivalued mapping from $\mathcal{P}$ to $\mathcal{Q}_f$ assigning to each probability measure P the p-quantile of the distribution of the residuals $R_{f,i}$ under P. As noted in Subsection 2.2, the mapping Q_f is multivalued, because in general quantiles are not uniquely defined. We want to determine the profile likelihood function $lik_{Q_f} : \mathcal{Q}_f \to [0,1]$ induced by the multivalued mapping Q_f. It is important to note that we would obtain the same results by considering only the distributions for which the p-quantile is unique (that is, the vagueness in the definition of quantiles has no influence on the resulting likelihood-based confidence regions).

Assume that the (imprecise) data $V_1^* = A_1, \ldots, V_n^* = A_n$ are observed, where $A_1, \ldots, A_n \in \mathcal{V}^* \backslash \{\varnothing\}$. In order to obtain the profile likelihood function lik_{Q_f} for the p-quantile of the distribution of the residuals $R_{f,i}$, we define for each function $f \in \mathcal{F}$ and each distance $q \in [0, +\infty)$ the bands

$$\overline{B}_{f,q} := \{(x,y) \in \mathcal{V} : |y - f(x)| \leq q\}$$
$$\underline{B}_{f,q} := \{(x,y) \in \mathcal{V} : |y - f(x)| < q\}$$

and the functions $\overline{k}_f, \underline{k}_f$ on $[0, +\infty)$ such that[8]

$$\overline{k}_f(q) = \#\left\{i \in \{1, \ldots, n\} : A_i \cap \overline{B}_{f,q} \neq \varnothing\right\}$$
$$\underline{k}_f(q) = \#\left\{i \in \{1, \ldots, n\} : A_i \subseteq \underline{B}_{f,q}\right\}$$

for all $q \in [0, +\infty)$. That is, $\overline{k}_f(q)$ is the number of imprecise data intersecting $\overline{B}_{f,q}$, while $\underline{k}_f(q)$ is the number of imprecise data completely contained in $\underline{B}_{f,q}$. Therefore, in particular, $\underline{k}_f(q) \leq \overline{k}_f(q)$ for all $q \in [0, +\infty)$.

Thanks to the results of Subsection 2.3 and the above definitions, we can now express the profile likelihood function lik_{Q_f} for the p-quantile of the distribution of the residuals $R_{f,i}$ as follows (a sketch of the proof is given in the Appendix):

$$lik_{Q_f}(q) = \begin{cases} \left[\lambda\left(\frac{\overline{k}_f(q)}{n}, p - \varepsilon\right)\right]^n & \text{if } \overline{k}_f(q) < (p-\varepsilon)\, n \\ \left[\lambda\left(\frac{\underline{k}_f(q)}{n}, p + \varepsilon\right)\right]^n & \text{if } \underline{k}_f(q) > (p+\varepsilon)\, n \\ 1 & \text{otherwise} \end{cases}$$

for all $q \in \mathcal{Q}_f$, where λ is the function on $[0,1] \times (0,1)$ defined by[9]

$$\lambda(s,t) = \left(\frac{s}{t}\right)^{-s} \left(\frac{1-s}{1-t}\right)^{s-1}$$

for all $s \in [0,1]$ and all $t \in (0,1)$. Hence, lik_{Q_f} is a piecewise constant function, which can take at most $n+2$ different values.

Example 4 *Consider the (imprecise) data described in Subsection 4.1 and depicted in Figure 4, and the regression function f represented by the upper curve (blue) in Figure 5. The corresponding profile likelihood function lik_{Q_f} for the 0.5-quantile of the distribution of the residuals $R_{f,i}$ is plotted in Figure 2 for the cases with $\varepsilon = 0$ (solid line) and $\varepsilon = 0.05$ (dashed line).*

[8] The cardinality of a set A is denoted by $\#A$.

[9] In this paper, $0^0 = 1$.

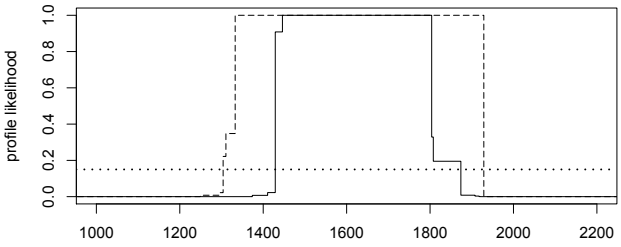


Figure 2: Profile likelihood functions from Examples 4 and 5.

3.2 Determination of Confidence Intervals for Quantiles of Residuals

Thanks to the above expression for the profile likelihood function lik_{Q_f}, we can now calculate the likelihood-based confidence regions for the quantiles of the distribution of the residuals $R_{f,i}$. Choose $\beta \in (0,1)$ and assume that

$$(\max\{p, 1-p\} + \varepsilon)^n \leq \beta \tag{2}$$

(that is, $\varepsilon < p < 1-\varepsilon$, and n is sufficiently large). Let $\mathcal{K} := \{0, \ldots, n\}$, and define

$$\underline{k} := \max\left\{k \in \mathcal{K} : k < (p-\varepsilon)\, n,\ \lambda(\tfrac{k}{n}, p-\varepsilon) \leq \sqrt[n]{\beta}\right\}$$
$$\overline{k} := \min\left\{k \in \mathcal{K} : k > (p+\varepsilon)\, n,\ \lambda(\tfrac{k}{n}, p+\varepsilon) \leq \sqrt[n]{\beta}\right\}.$$

Then $\underline{k} < \overline{k}$, and for each $f \in \mathcal{F}$, the interval

$$\mathcal{C}_f := \left\{q \in [0, +\infty) : \underline{k} < \overline{k}_f(q),\ \underline{k}_f(q) < \overline{k}\right\}$$

is the likelihood-based confidence region with cutoff point β for the p-quantile of the distribution of the residuals $R_{f,i}$. The interval $\mathcal{C}_f$ consists of all $q \in [0, +\infty)$ such that the band $\overline{B}_{f,q}$ intersects at least $\underline{k}+1$ imprecise data, and the band $\underline{B}_{f,q}$ contains at most $\overline{k}-1$ imprecise data. When $\varepsilon = 0$, the interval $\mathcal{C}_f$ is asymptotically a (conservative) confidence interval of level $F_{\chi^2}(-2\log\beta)$ for the p-quantile of the distribution of the residuals $R_{f,i}$, where F_{χ^2} is the cumulative distribution function of the chi-square distribution with 1 degree of freedom (see for example [13]). The exact level of the (conservative) confidence interval $\mathcal{C}_f$ can be obtained directly from its definition, by means of simple combinatorial arguments (also when $\varepsilon > 0$).

It is important to note that the confidence intervals $\mathcal{C}_f$ do not depend on the choice of the set $\mathcal{V}^*$ of possible imprecise data (as far as the observed ones, $A_1, \ldots, A_n$, are contained in it). This can be surprising, since the set $\mathcal{P} = \mathcal{P}_\varepsilon$ of probability measures considered depends strongly on $\mathcal{V}^*$, as noted at the beginning of Section 2. However, the independence of the confidence intervals $\mathcal{C}_f$ from the choice of the set $\mathcal{V}^*$ is not so surprising when one considers that the intervals $\mathcal{C}_f$ are likelihood-based confidence regions, and that likelihood inference is always conditional on the data (that is, independent of considerations about which other data could have been observed). This can be considered as a sort of robustness against misspecification of the set $\mathcal{V}^*$ of possible imprecise data. The practical advantage is that it is not necessary to think about which other imprecise data could have been observed, besides the ones that were actually observed (that is, $A_1, \ldots, A_n$).

Example 5 *In the situation of Example 4, the confidence interval $\mathcal{C}_f$ with $\beta = 0.15$ is approximately $[1429, 1874]$ when $\varepsilon = 0$, and $[1304, 1929]$ when $\varepsilon = 0.05$ (the cutoff point $\beta = 0.15$ is represented by the dotted line in Figure 2).*

3.3 Regression as a Decision Problem

The problem of minimizing the p-quantile of the distribution of the residuals $R_{f,i}$ can be described as a statistical decision problem: the set of probability measures considered is $\mathcal{P} = \mathcal{P}_\varepsilon$, the set of possible decisions is $\mathcal{F}$, and the loss function $L : \mathcal{P} \times \mathcal{F} \to [0, \infty)$ is defined by

$$L(P, f) = Q_f(P)$$

for all $P \in \mathcal{P}$ and all $f \in \mathcal{F}$. That is, the p-quantile of the distribution of the residuals $R_{f,i}$ is interpreted as the loss we incur when we choose the function f. In fact, the loss function L is multivalued, since in general the p-quantile is not unique: $L(P, f)$ could be reduced to a single value by taking for example the upper p-quantile of the distribution of the residuals $R_{f,i}$.

The information provided by the observed (imprecise) data is described by the likelihood function lik on $\mathcal{P}$. A very simple way of using this information consists in reducing $\mathcal{P}$ to the set $\mathcal{P}_{>\beta}$ for some cutoff point $\beta \in (0,1)$. The resulting set $\mathcal{P}_{>\beta}$ can be interpreted as an imprecise probability measure, on which we can base our choice of f. For each $f \in \mathcal{F}$, the set of all possible values of the loss $L(P, f)$ when P varies in $\mathcal{P}_{>\beta}$ can be interpreted as the imprecise p-quantile of the residuals $R_{f,i}$ under the imprecise probability measure $\mathcal{P}_{>\beta}$. It corresponds to the interval $\mathcal{C}_f$, when condition (2) is satisfied.

Assume that condition (2) is satisfied. In order to choose a function f, we can minimize the supremum of $\mathcal{C}_f$. This approach is similar to the Γ-minimax decision criterion with respect to the imprecise probability measure $\mathcal{P}_{>\beta}$, and is called LRM (likelihood-based region minimax) criterion in [4]. When there

is a unique $f \in \mathcal{F}$ minimizing $\sup \mathcal{C}_f$, it can be denoted by f_{LRM}, and $\sup \mathcal{C}_f$ can be denoted by $\overline{q}_{LRM}$. In this case, f_{LRM} is characterized geometrically by the fact that $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ is the thinnest band of the form $\overline{B}_{f,q}$ containing at least $\overline{k}$ imprecise data, for all $f \in \mathcal{F}$ and all $q \in [0, +\infty)$. Finding the function f_{LRM} is an interesting computational problem: see for example [2, 15, 22].

An interesting description of the uncertainty about the optimal choice of $f \in \mathcal{F}$ is obtained by considering interval dominance for the imprecise p-quantiles of the residuals $R_{f,i}$ under the imprecise probability measure $\mathcal{P}_{>\beta}$. When f_{LRM} exists, the undominated functions $f \in \mathcal{F}$ are those such that $\mathcal{C}_f$ intersects $\mathcal{C}_{f_{LRM}}$. In particular, when $\overline{q}_{LRM} \in \mathcal{C}_{f_{LRM}}$ (that is, $\mathcal{C}_{f_{LRM}}$ is right-closed), the undominated functions $f \in \mathcal{F}$ are characterized geometrically by the fact that $\overline{B}_{f,\overline{q}_{LRM}}$ intersects at least $\underline{k}+1$ imprecise data. In general, the set of undominated functions f tends to get smaller when we observe more and more (imprecise) data, but it does not necessarily tend to reduce to a singleton, because of the unavoidable uncertainty discussed in Subsection 2.1.

3.4 Prediction

Consider the case in which (instead of n) we have $n+1$ pairs (V_i, V_i^*) of precise and imprecise data $V_i = (X_i, Y_i)$ and V_i^*, respectively. We want to predict the realization of the precise data value V_{n+1} on the basis of the realization of the n imprecise data $V_1^*, \ldots, V_n^*$. Choose $k \in \{1, \ldots, n\}$, and assume that for each possible realization of the $n+1$ imprecise data $V_1^*, \ldots, V_{n+1}^*$, there is a distance $q' \in [0, +\infty)$ such that for some $f' \in \mathcal{F}$ (not necessarily unique), $\overline{B}_{f',q'}$ is a thinnest band of the form $\overline{B}_{f,q}$ containing at least k of the $n+1$ imprecise data, for all $f \in \mathcal{F}$ and all $q \in [0, +\infty)$. Because of symmetry, the probability that V_{n+1}^* is included in a band $\overline{B}_{f,q'}$ containing at least k of the $n+1$ imprecise data (for some $f \in \mathcal{F}$) is at least $\frac{k}{n+1}$. Hence, when $\overline{B}_{f'',q''}$ is a thinnest band of the form $\overline{B}_{f,q}$ containing at least k of the n imprecise data $V_1^*, \ldots, V_n^*$ (for all $f \in \mathcal{F}$ and all $q \in [0, +\infty)$), the probability that V_{n+1}^* is included in the union $\mathcal{B}$ of all bands $\overline{B}_{f,q''}$ containing at least $k-1$ of the n imprecise data $V_1^*, \ldots, V_n^*$ (for all $f \in \mathcal{F}$) is at least $\frac{k}{n+1}$. That is, $\mathcal{B}$ is a (conservative) prediction region of level $\frac{k}{n+1} - \varepsilon$ for the precise data value V_{n+1}.

In particular, when condition (2) is satisfied and f_{LRM} exists, the union $\mathcal{B}$ of all bands $\overline{B}_{f,\overline{q}_{LRM}}$ containing at least $\overline{k}-1$ of the n imprecise data $V_1^*, \ldots, V_n^*$ (for all $f \in \mathcal{F}$) is a (conservative) prediction region of level $\frac{\overline{k}}{n+1} - \varepsilon$ for the precise data value V_{n+1}. Prediction regions of this form can sometimes be reduced to smaller regions thanks to the assumption that V_{n+1}^* takes values in $\mathcal{V}^*$. When besides the realization of the n imprecise data $V_1^*, \ldots, V_n^*$, also the (precise or imprecise) realization of X_{n+1} has been observed, the realization of Y_{n+1} can be predicted for example by using the idea of conformal prediction (see [21]), but this goes beyond the scope of the present paper.

4 Example of Application

In this section, we apply the proposed regression method to socioeconomic data from the ALLBUS (German General Social Survey). Data collection in surveys is subject to many different influences that may cause various biases in the data set (see for example [3]). Therefore, it is often reasonable to assume that the actual value lies rather in some interval around the observed value. Furthermore, data on sensitive quantities is sometimes only available in categories that form a partition of the space of possible values. A simple approach to analyze this kind of data is to reduce the intervals to their central values and to apply usual regression methods to the reduced, precise data. In contrast to this, we suggest to analyze directly the interval-valued data by means of the regression method proposed in Section 3.

We want to investigate the age-income profile, which is a fundamental relationship in the social sciences and a typical example in textbooks on social research methods (see for example [1]).

Income is a key demographic variable for socioeconomic research questions. But asking for income in an interview is a sensitive question that some respondents refuse to answer. Thus, survey data on income often include missing values. One way to make the question less sensitive is to present predefined income categories according to which the income of the respondent shall be classified. In the ALLBUS, income data is collected with a two-step design with the open question for income as first step and the presentation of a category scheme as second step. As a result, the data set contains at the same time precise values for some individuals and interval-valued observations for others. Yet, even if the respondents are willing to give their exact income, limited remembrance usually prevents them from doing so. Instead, they will give rounded and heaped values (see [9]), where heaping refers to irregular rounding behavior (see for example [10]). Therefore, it is more reliable to regard also the precise income values as interval-valued observations.

Data on the age of respondents is more easily obtained, but it is always measured with limited precision, e.g. in years. In this case, it might be useful to

consider intervals $[age, age+1)$ instead. Furthermore, age data might be available as age classes only.

4.1 ALLBUS Data and Regression Model

We analyze the ALLBUS data set of 2008 containing 3 247 interviews. The considered variables are *personal income* (on average per month) and *age*. Here, we consider the worst case, where both variables are available in categories only ($v389$ and $v155$ of the data set with 22 possible income categories and six age classes; see [19]), although the proposed regression method could be applied to the data set with some precise and some imprecise observations, too. Thus, for each individual $i \in \{1, \ldots, n\}$ we consider observations $V_i^* = X_i^* \times Y_i^*$, where $X_i^* = [\underline{x}_i, \overline{x}_i)$ is the corresponding age class and $Y_i^* = [\underline{y}_i, \overline{y}_i)$ is the category into which the income of respondent i falls. In the given data set, there are 620 missing income values and 11 missing age values. Missing values are replaced by intervals that cover the entire observation space of each variable. In this case, $X_i^* = [18, 100)$ or $Y_i^* = [0, +\infty)$, respectively. A two-dimensional histogram of the data set is given in Figure 4.

The relationship between age and income is usually modeled by a quadratic function in age (see for example [1]). Thus, the set of regression functions we consider here is

$$\mathcal{F} = \{f_{a,b_1,b_2} : \ a, b_1, b_2 \in \mathbb{R}\},$$

where each function f_{a,b_1,b_2} is defined by

$$f_{a,b_1,b_2}(x) = a + b_1\, x + b_2\, x^2$$

for all $x \in \mathcal{X} := [18, 100)$. We choose to minimize the 0.5-quantile of the distribution of the residuals (i.e., $p = 0.5$), and we take the cutoff point $\beta = 0.15$. Furthermore, we want to compare the results obtained by the proposed method with those from an ordinary least squares (OLS) regression based on the interval centers. Since the latter implies the assumption $P(V_i \in V_i^*) = 1$, we also set $\varepsilon = 0$ here.

We conduct the regression analysis as follows: First, the likelihood-based confidence regions $\mathcal{C}_{f_{a,b_1,b_2}}$ are computed for reasonable parameter values (a, b_1, b_2). Then, we identify the parameter combination among these that minimizes the upper bound of $\mathcal{C}_{f_{a,b_1,b_2}}$. The function corresponding to this parameter combination is the function f_{LRM} which is optimal according to the LRM criterion (see Subsection 3.3). Finally, the upper bound $\overline{q}_{LRM}$ of $\mathcal{C}_{f_{LRM}}$ is used to determine the set of undominated functions.

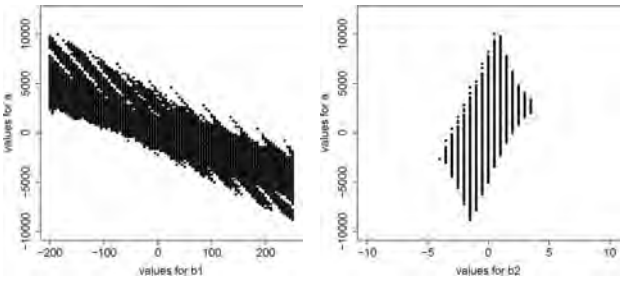


Figure 3: Two-dimensional projections of the set of undominated parameter values.

4.2 Results

We considered a grid of combinations of parameter values where $a \in [-10\,000, 12\,000]$, $b_1 \in [-200, 250]$, and $b_2 \in [-10, 10]$. Corresponding to the set of undominated functions, we find the set of undominated parameter combinations displayed in Figure 3. This set is clearly not convex. Moreover, in the case considered here, the parameters are not independent from each other, in the sense that many different combinations of parameter values (a, b_1, b_2) may lead to very similar functions f_{a,b_1,b_2} over $\mathcal{X}$. Thus, there are actually infinitely many undominated parameter combinations, but the associated curves are similar to those we find within the considered grid.

The parameter combination implying the smallest upper endpoint of the confidence interval for the 0.5-quantile of the residuals is (850, 0, 0) with $\mathcal{C}_{f_{850,0,0}} = [525, 650]$. The function f_{LRM} is thus a constant line: this is due to the rectangular shape and the locations of the observations in our data set. Hence, the value 850 can be interpreted as an estimate of the p-center (with $p = 0.5$) of the income distribution (see the beginning of Section 3). A further interpretation of the function f_{LRM} is given by the band $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ limited by the functions $f_{LRM} - \overline{q}_{LRM}$ and $f_{LRM} + \overline{q}_{LRM}$: Among all bands constructed around all considered functions, this band is the thinnest one that contains at least $\overline{k} = 1\,679$ imprecise observations (see Subsection 3.3).

The function f_{LRM} and the band $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ are presented in Figure 5, besides the undominated functions. It can be seen that within the set of undominated functions there is a large variety of shapes of the age-income profile, including straight lines, convex parabolic curves as well as concave ones. From a social scientist's point of view this result may be unsatisfying because it doesn't support only one form of the relationship between age and income. However, given the imprecision of the data, it is reasonable to consider all shapes consistent with the data as possi-

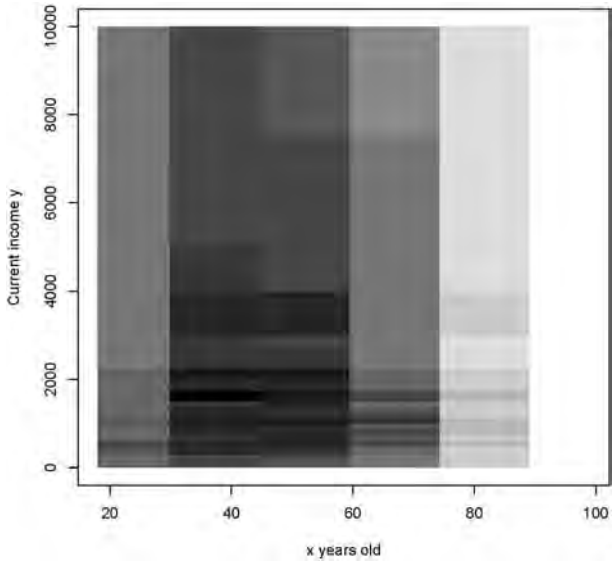


Figure 4: Two-dimensional histogram of the data set.

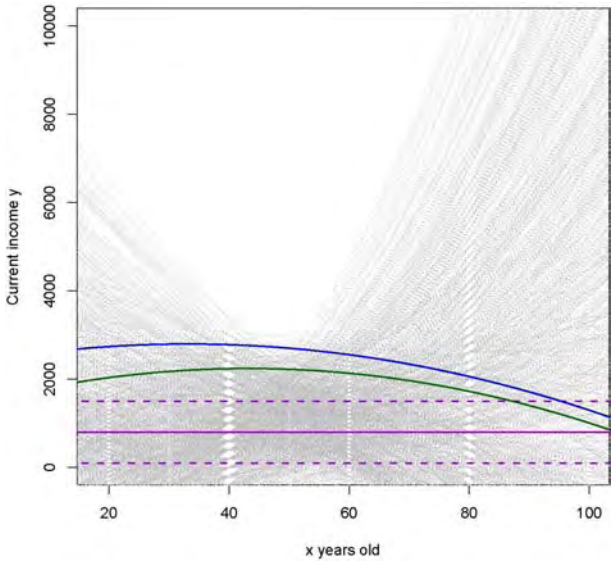


Figure 5: Undominated functions (dotted curves, gray), interval data-based f_{LRM} (solid line, violet) and band $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ (dashed lines, violet) versus OLS regressions on interval centers with upper income limit 15 000 (upper curve, blue) and upper income limit 10 000 (lower curve, green).

ble age-income profiles. If the observed intervals were overlapping or if they constituted a finer partition of the space of possible observations, the set of undominated functions would be smaller. Hence, the set of undominated functions can be interpreted as the set of plausible descriptions of the age-income profile that reflects at the same time the uncertainty inherent in the imprecise data.

The usual method to analyze this kind of interval data is to conduct a quadratic OLS regression based on the interval centers ignoring the imprecision of the data. In this case, one has to give an upper limit for the highest income class $[7\,500, +\infty)$ in order to compute the interval centers. Of course, the choice of this upper limit has an impact on the estimates of the OLS regression. The effect of two different choices of the upper income limit is illustrated in Figure 5. The OLS curves displayed there are based on interval centers with upper income limits 15 000 and 10 000, respectively. In contrast to the OLS approach, the regression method proposed in this paper is not sensitive to the extremes, since the regression functions are evaluated on the basis of confidence regions for the 0.5-quantile of the residuals' distribution.

The proposed regression method permits to identify plausible descriptions of the relationship between the socioeconomic characteristics *age* and *income*. Given the imprecise data, many different shapes of the age-income profile are plausible. Further computations indicated that our findings hold for transformed income data on the logarithmic scale, too. The results are not very informative, but reliable. To obtain more informative, but less reliable results, it suffices to increase the cutoff point β (that is, to decrease the confidence level of the intervals $\mathcal{C}_{f_{a,b_1,b_2}}$). One idea to obtain more informative results without sacrificing reliability could be to use many different category schemes during the income data collection and thereby obtain a data set with overlapping categories.

5 Conclusion

In this paper, we introduced a robust approach to regression with imprecise data, in which the error distribution is not constrained to a particular parametric family. The method was presented within a very general framework and it can be adapted to a wide range of practical settings, since it can be applied to all kinds of imprecise data covering e.g. interval data, precise data, and missing data. In our method, the imprecise data are interpreted as the result of a coarsening process which can be informative, and even wrong with a certain probability.

In future work, the statistical properties of the proposed regression method shall be studied in more detail. In particular, we plan to investigate the impact of stronger assumptions about the error distribution and the coarsening process. Moreover, the performance of the regression method shall be compared to those of alternative approaches to regression with imprecise data, also with regard to computational aspects.

Acknowledgements

The authors wish to thank Thomas Augustin and the anonymous referees for their helpful comments.

Appendix

The expression for the profile likelihood function lik_{Q_f} given in Subsection 3.1 can be proved as follows. In Subsection 2.3, we have seen that $lik_{Q_f} = lik^*_{Q^*_f}$, where lik^* and Q^*_f are defined on the set $\mathcal{P}_{V^*}$ of all possible distributions P_{V^*} for the imprecise data V_i^*. The function lik^* assigns to each P_{V^*} the corresponding likelihood value: in particular, it has a unique maximum in the empirical distribution (of the imprecise data) $\widehat{P}_{V^*}$. The multivalued mapping Q^*_f assigns to each P_{V^*} all p-quantiles of the residuals $R_{f,i}$ for all distributions of the precise data V_i compatible with P_{V^*}. Consider in particular $Q^*_f(\widehat{P}_{V^*})$: if $\varepsilon = 0$, then $q \in \mathcal{Q}_f$ is a p-quantile of the residuals $R_{f,i}$ for some distribution of the precise data V_i compatible with $\widehat{P}_{V^*}$ if and only if $\underline{k}_f(q) \leq p\,n \leq \overline{k}_f(q)$. The case with $\varepsilon > 0$ corresponds to the case with $\varepsilon = 0$ when $Q^*_f(\widehat{P}_{V^*})$ is enlarged to all p'-quantiles of the residuals $R_{f,i}$ such that $p - \varepsilon \leq p' \leq p + \varepsilon$. This proves the "otherwise" part of the expression for lik_{Q_f} given in Subsection 3.1, since $lik^*(\widehat{P}_{V^*}) = 1$.

Now assume that $q \in \mathcal{Q}_f$ satisfies $\overline{k}_f(q) < (p - \varepsilon)\,n$. Let $P'_{V^*} \in \mathcal{P}_{V^*}$ be the empirical distribution obtained when only the $n - \overline{k}_f(q)$ imprecise data not intersecting $\overline{B}_{f,q}$ are considered, and let $P''_{V^*} \in \mathcal{P}_{V^*}$ be the empirical distribution obtained when only the $\overline{k}_f(q)$ imprecise data intersecting $\overline{B}_{f,q}$ are considered. The latter is not well-defined when $\overline{k}_f(q) = 0$: in this case, let $P''_{V^*} \in \mathcal{P}_{V^*}$ be the Dirac distribution assigning probability 1 to a set $A \in \mathcal{V}^*$ intersecting $\overline{B}_{f,q}$ (such a set A exists, since $q \in \mathcal{Q}_f$). Then $q \in Q^*_f(P'''_{V^*})$ with $P'''_{V^*} = (p - \varepsilon)\,P''_{V^*} + (1 - p + \varepsilon)\,P'_{V^*}$, and it can be easily checked that

$$lik^*_{Q^*_f}(q) = lik^*(P'''_{V^*}) = \left[\lambda\left(\frac{\overline{k}_f(q)}{n},\, p - \varepsilon\right)\right]^n .$$

This proves the first case of the expression for lik_{Q_f} given in Subsection 3.1, and the second one can be proved analogously.

References

[1] Allison, P. D. (1998). *Multiple Regression*. Pine Forge Press.

[2] Bernholt, T. (2005). Computing the least median of squares estimator in time $O(n^d)$. In *Computational Science and Its Applications — ICCSA 2005*. Springer, 697–706.

[3] Biemer, P. P., and Lyberg, L. E. (2003). *Introduction to Survey Quality*. Wiley.

[4] Cattaneo, M. (2007). *Statistical Decisions Based Directly on the Likelihood Function*. PhD thesis, ETH Zurich. doi:10.3929/ethz-a-005463829.

[5] Cattaneo, M. (2008). Fuzzy probabilities based on the likelihood function. In *Soft Methods for Handling Variability and Imprecision*. Springer, 43–50.

[6] de Cooman, G., and Zaffalon, M. (2004). Updating beliefs with incomplete observations. *Artif. Intell.* 159, 75–125.

[7] Domingues, M. A. O., de Souza, R. M. C. R., and Cysneiros, F. J. A. (2010). A robust method for linear regression of symbolic interval data. *Pattern Recognit. Lett.* 31, 1991–1996.

[8] Ferson, S., Kreinovich, V., Hajagos, J., Oberkampf, W., and Ginzburg, L. (2007). *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*. Technical Report SAND2007-0939. Sandia National Laboratories.

[9] Hanisch, J. U. (2005). Rounded responses to income questions. *Allg. Stat. Arch.* 89, 39–48.

[10] Heitjan, D. F., and Rubin, D. B. (1991). Ignorability and coarse data. *Ann. Stat.* 19, 2244–2253.

[11] Hudson, D. J. (1971). Interval estimation from the likelihood function. *J. R. Stat. Soc. B* 33, 256–262.

[12] Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Springer.

[13] Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC.

[14] Pawitan, Y. (2001). *In All Likelihood*. Oxford University Press.

[15] Rousseeuw, P. J., and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley.

[16] Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

[17] Strassen, V. (1964). Meßfehler und Information. *Z. Wahrscheinlichkeitstheorie* 2, 273–305.

[18] Tasche, D. (2003). Unbiasedness in least quantile regression. In *Developments in Robust Statistics*. Physica-Verlag, 377–386.

[19] Terwey, M., and Baltzer, S. (2009). *ALLBUS Datenhandbuch 2008*. GESIS.

[20] Utkin, L., Zatenko, S., and Coolen, F. (2009). Combining imprecise Bayesian and maximum likelihood estimation for reliability growth models. In *ISIPTA '09*. SIPTA, 421–430.

[21] Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

[22] Watson, G. A. (1998). On computing the least quantile of squares estimate. *SIAM J. Sci. Comput.* 19, 1125–1138.

[23] Zaffalon, M., and Miranda, E. (2009). Conservative inference rule for uncertain reasoning under incompleteness. *J. Artif. Intell. Res. (JAIR)* 34, 757–821.

Building Imprecise Classification Trees With Entropy Ranges

Richard J. Crossman
University of Warwick
r.j.crossman@warwick.ac.uk

Joaquin Abellan
University of Granada
jabellan@decsai.ugr.es

Thomas Augustin
University of Munich
augustin@stat.uni-muenchen.de

Frank P.A. Coolen
Durham University
frank.coolen@dur.ac.uk

Abstract

One method for building classification trees is to choose split variables by maximising expected entropy. This can be extended through the application of imprecise probability by replacing instances of expected entropy with the maximum possible expected entropy over credal sets of probability distributions.

Such methods may not take full advantage of the opportunities offered by imprecise probability theory. In this paper, we change focus from maximum possible expected entropy to the full range of expected entropy. We then choose one or more potential split variables using an interval comparison method.

This method is presented with specific reference to the case of ordinal data, and we present algorithms that maximise and minimise entropy within the credal sets of probability distributions which are generated by the NPI method for ordinal data.

Keywords. Imprecise probability, classification trees, nonparametric predictive inference

1 Introduction

The process of classification can be summarised as follows. In a data set D each element is described by n *attribute variables* (or *features*) $X_1, \ldots, X_n$, and a single *class variable* (or *variable in study*) C. The variable X_i takes some value a_i from the set $\mathcal{A}_i$, and the variable C takes some value, or *category*, from $\mathcal{C} = \{c_1, \ldots, c_K\}$. The aim is to take a given vector $\boldsymbol{a} = (a_1, \ldots, a_n)$ and determine the associated category.

One such method is the classification tree. This is a hierarchical graph in which each parent node represents an attribute variable (called the *split variable* of the node), the edges represent the values of that variable, and the leaves represent categories. A data vector $\boldsymbol{a}$ is categorised by starting at the root node and following the appropriate edges until a leaf is reached. The category given at that leaf is the prediction for the associated category of the data point.

Such a method requires finding an order for considering the attribute variables. We base our method upon the one given in [2], summarised as follows:

1. Using information measure IM, calculate $IM(R)$ and $IM(R|X_i)$ (the information measure following splitting on X_i) for each unassigned attribute variable X_i, where R is the data relevant to the current node (i.e. the subset of D which matches the values given to the attribute variables already assigned);
2. If $IM(R|X_{i^*}) := \max_i IM(R|X_i) \leq IM(R)$, go to step 3. Otherwise, split data by the value of X_{i^*}. If the data relevant to the current node is R, then the relevant data for each child node will be $\{\boldsymbol{v} \in R | v_{i^*} = j\}$ where the edge between this node and the child node is labelled $j \in \mathcal{A}_{i^*}$. Return to step 1 for each of these child nodes;
3. This is a leaf node, labelled with the most common class in R. If more than one class is equally common, choose the class most common at the leaf's parent (this approach is due to [4]).

Step 1 will be adapted to make use of imprecise probability, but consider first the information measure when imprecision is not applied. Both $IM(R)$ and each $IM(R|X_i)$ are functions of an associated probability distribution. These distributions are estimated using relative frequencies. For the current data set R, consider each unassigned attribute variable X_i as follows. Define $n^R := |R|$ and denote by n_j^R the number of data points in R with class c_j. Define

$$p_j^R := \frac{n_j^R}{n^R}, \; p_j^{\hat{a}_i} := \frac{n_j^{\hat{a}_i}}{n^{\hat{a}_i}} \tag{1.1}$$

where $\hat{a}_i := \{\boldsymbol{v} \in R | X_i = a_i\}$. We also define

$$I(R, X_i) := \sum_{a_i \in \mathcal{A}_i} p(X_i = a_i) H(\boldsymbol{p}^{\hat{a}_i}) \tag{1.2}$$

where $p(X_i = a_i)$ is also estimated using relative frequencies, and where $H(\cdot)$ is Shannon's entropy

$$H(\boldsymbol{p}) = -\sum_{i=1}^{n} p_i \ln(p_i) \tag{1.3}$$

for probability distribution $\boldsymbol{p}$. $H(\boldsymbol{p})$ is maximum when $\boldsymbol{p}$ is uniform, and minimum when $p_i = 1$ for some i.

Define an information measure as follows:

$$IM(R, X_i) := H(\boldsymbol{p}^R) - I(R, X_i). \tag{1.4}$$

Only the $I(R, X_i)$ value determines the next split variable at each node. Further, maximising the information measure is equivalent to minimising $I(R, X_i)$, which in turn requires minimising entropy.

Instead of using the relative frequencies for each X_i to generate a distribution for the categories, we can generate a credal set of probabilites, referred to as a *structure*. This is done in [2] by using the *imprecise Dirichlet model* (IDM) [12], giving the following intervals for $p_j^{\hat{a}_i}$

$$\left[\frac{n_j^{\hat{a}_i}}{n^{\hat{a}_i} + s}, \frac{n_j^{\hat{a}_i} + s}{n^{\hat{a}_i} + s}\right] \tag{1.5}$$

for some value of s, commonly chosen to be 1 or 2.

Alternatives to the IDM exist. In [6] the IDM is replaced with the NPI method for categorical data (requiring a modification to the algorithm, which is contained in that reference). In this paper, the NPI method for ordinal data [7] replaces the IDM. This method takes account of the ordering amongst the categories, resulting (in general) in smaller credal sets than would otherwise be generated. In this set-up categories c_1 and c_K ($c_i,\ i = 2, \ldots, K-1$) are referred to as *boundary* (*internal*) categories. The corresponding component in a probability distribution for a category is referred to as a boundary (internal) component.

A summary of the ordinal NPI method follows. For $X_1, \ldots, X_n, X_{n+1}$ real-valued absolutely continuous and exchangeable random quantities, assume that the first n ordered observed values are denoted by $x_1 < x_2 < \ldots < x_n$, and let $x_0 = -\infty$ and $x_{n+1} = \infty$. We use Hill's assumption $A_{(n)}$ [8] that for a future observation X_{n+1} and for all $j = 1, \ldots, n+1$

$$P(x_{n+1} \in I_j = (x_{j-1}, x_j)) = \frac{1}{n+1}. \tag{1.6}$$

$A_{(n)}$ assumes nothing else, and can be used to define a lower (upper) probability vector $\boldsymbol{L}$ ($\boldsymbol{U}$) for the category of X_{n+1} by a latent variable representation. Assume n observations, with n_j in category c_j. Let Y_{n+1} denote the random quantity representing the category a future observation will belong to. We assume that category c_j is represented by interval Ic_j, where $\cup_{j=1,\ldots,k} Ic_j = \mathbb{R}$ and $Ic_j \cap Ic_i = \emptyset$ for all $i \neq j$. The ordering is such that Ic_j has neighbouring intervals Ic_{j-1} to the left and Ic_{j+1} to the right on the real line, with only one such neighbour when $j \in \{1, k\}$. Assume further that n_j values of $x_1 < x_2 < \ldots < x_n$ are in interval Ic_j. We therefore assume that the event $X_{n+1} \in Ic_j$ is equivalent to the event $Y_{n+1} = c_j$. See [7] for more detail.

The lower probability of a category c_i is therefore equal to the number of intervals I_j entirely contained within Ic_i, and the upper probability of that category is equal to the number of intervals I_j with a non-empty intersection with Ic_i. Hence, the ordinal NPI model replaces (1.5) with:

$$\left[\max\left(\frac{n_j^{\hat{a}_i} - 1}{n^{\hat{a}_i} + 1}, 0\right), \frac{n_j^{\hat{a}_i} + 1}{n^{\hat{a}_i} + 1}\right] \tag{1.7}$$

when $1 < j < K$, and otherwise

$$\left[\frac{n_j^{\hat{a}_i}}{n^{\hat{a}_i} + 1}, \frac{n_j^{\hat{a}_i} + 1}{n^{\hat{a}_i} + 1}\right]. \tag{1.8}$$

Therefore all intervals I_j lying entirely within an interval Ic_i are assigned to category c_i, and intervals overlapping both Ic_i and Ic_{i+1} can be assigned entirely to either c_i or c_{i+1}, or split between them.

We can thus talk about the probability mass "on either side" of categories. From this point on the *available mass* to the left (right) of internal category c_j is defined as the probabilty mass that has not been assigned to c_j or c_{j-1} (c_{j+1}) whilst calculating the lower probabilities L_j and L_{j-1} (L_{j+1}). This mass is therefore described as being *available* to $\hat{\boldsymbol{p}}_j$ and $\hat{\boldsymbol{p}}_{j-1}$ ($\hat{\boldsymbol{p}}_{j+1}$). Any distribution $\hat{\boldsymbol{p}}$ within the ordinal NPI structure has the property $\hat{p}_j \geq L_j$; $\hat{p}_j - L_j$ is described as the mass *assigned* to $\hat{p}_j$.

When using either IDM or ordinal NPI, we take the category distribution from each credal set that maximises entropy (note that the distributions of attribute variables are still generated using relative frequencies). This will generate a maximum expected value of each $I(R, X_i)$ and of $H(\boldsymbol{p}^R)$, and so determine our next split variable. However, for $K > 2$ the algorithm given in [2] for maximising entropy does not work within the structure of ordinal NPI (see Section 4), necessitating the algorithm described in this paper.

Once all possible values of (1.4) are non-positive, we do not split further, and decide on the class value to assign by choosing the most common class in R, just as is done in the case without imprecise probability.

However, this application of the IDM or ordinal NPI excludes one of the most fundamental justifications for using imprecise probability: the possibility of circumstances in which we are unable to choose between two options. We therefore describe an alternative method in this paper, which rather than comparing maximum entropies compares the ranges of entropies, and chooses between those ranges only when our method for comparing intervals allows it.

Section 2 defines and explores entropy intervals, which are then used to describe our method, given in Section 3. Section 4 and 5 describe the algorithms by which entropy over the ordinal NPI structure is maximised and minimised, respectively. Section 6 contains conclusions and ideas for future work.

2 Entropy Intervals

Considering entropy intervals requires the following two definitions.

Definition 2.1 For a closed structure $\mathcal{M}$, a vector is defined as a *potential* of $\mathcal{M}$, denoted $\boldsymbol{v}^*$, if $\boldsymbol{v}^* \in \mathcal{M}$ and

$$H(\boldsymbol{v}^*) = \max_{\boldsymbol{w}\in\mathcal{M}} H(\boldsymbol{w}). \tag{2.1}$$

A vector is defined as the *guarantee* of $\mathcal{M}$, denoted $\boldsymbol{v}_*$, if $\boldsymbol{v}_* \in \mathcal{M}$ and

$$H(\boldsymbol{v}_*) = \min_{\boldsymbol{w}\in\mathcal{M}} H(\boldsymbol{w}). \tag{2.2}$$

If $\frac{1}{k}(1,\ldots,1) \in \mathcal{M}$, then $v^* = \frac{1}{k}(1,\ldots,1)$. Any probability vector in $\mathcal{M}$ with a component equal to 1 is a guarantee of $\mathcal{M}$.

The names of these properties are justified as follows: for a given X_i the entropy of the potential and guarantee generate respectively the maximum and minimum value of $I(R, X_i)$. Thus we can guarantee a minimum value for this function, but also talk of the potential maximum. This is also true for $H(\boldsymbol{p}^R)$.

In a convex structure, the potential is unique (see the algorithm in [2]). This is not necessarily true of the guarantee; when $\mathcal{M} = [0,1] \times [0,1]$, both $(1,0)$ and $(0,1)$ are guarantees.

Because entropy is a continuous function, and the ordinal NPI structure is connected, we can define an entropy interval as follows.

Definition 2.2 The *entropy interval* of a connected structure $\mathcal{M}$ is defined as

$$\{H(\boldsymbol{v}) : \boldsymbol{v} \in \mathcal{M}\} = [H(\boldsymbol{v}_*), H(\boldsymbol{v}^*)]$$

where $\boldsymbol{v}_*$ and $\boldsymbol{v}^*$ are the guarantee and the potential of $\mathcal{M}$ respectively.

Example 4.1 Consider $K = 8$ classes, and six observations (1,0,0,2,0,3,0,0). From [7] we have that the structure is contained within the following set:

$$\frac{1}{7}([1,2],[0,1],[0,1],[1,3],[0,1],[2,4],[0,1],[0,1]). \tag{2.3}$$

The maximum entropy algorithm adapted for ordinal data (see Section 4) gives the following vectors at each stage

1. $\frac{1}{7}(1,0,0,1,0,2,0,0)$, 2. $\frac{1}{7}(1,0,0,1,1,2,0,0)$
3. $\frac{1}{14}(2,1,1,2,2,4,0,0)$, 4. $\frac{1}{14}(2,1,1,2,2,4,1,1)$

and the minimum entropy algorithm (see Section 5) gives

1. $\frac{1}{7}(1,0,0,1,0,2,0,0)$, 2. $\frac{1}{7}(1,0,0,1,0,4,0,0)$
3. $\frac{1}{7}(2,0,0,1,0,4,0,0)$

The resulting entropy interval is $[0.9557, 1.9459]$. For comparison, note that the full entropy range for an 8 element probability distribution is $[0, 2.079]$, and that had we used the algorithm given in [2] without taking into account the structure of the model, we would have incorrectly generated a potential with entropy 1.9668. This concludes the example.

Instead of a single value $I(R, X_i)$, consider an interval $[\underline{I}(R, X_i), \overline{I}(R, X_i)] := I_i$ where the bounds of the interval are calculated using guarantees and potentials in the obvious way. Further, replace $H(\boldsymbol{p}^R)$ with the interval I_R, generated by the guarantee and potential of the current data set R.

These intervals provide an alternative method for choosing the split variables. Define a set of intervals $\mathcal{I} = \{I_{a_1}, \ldots, I_{a_n}\}$, where each a_i corresponds to a potential split variable X_{a_i}. Remove from $\mathcal{I}$ any interval that is dominated by another interval in the set. There are various methods by which one can determine dominance, but in this paper we use the simplest: *interval dominance*. Under this method, interval $I_i = [c_i, d_i]$ dominates interval $I_j = [c_j, d_j]$, denoted $I_i >_{\mathrm{d}} I_j$, iff $c_i \geq d_j$. The use of alternative methods for comparing intervals [11] can be explored, this is left as a topic for future research.

Once all dominated intervals have been removed, we say we cannot choose between each of variables corresponding to the remaining elements of $\mathcal{I}$ as the next choice for the split variable.

3 Imprecise classification trees

Just as imprecise probabilities are expressed as sets rather than single values, an imprecise classification tree is expressed as a forest.

Consider node P at the end of a path, length l, from the root node. There are $n - l$ choices for the next split variable, denoted $X_{P_1}, X_{P_2}, \ldots, X_{P_{n-l}}$. If no interval I_{P_j}

dominates I_R, then there is no split (as no split variable can be considered superior to no split at all). Otherwise, create a set $\mathcal{S}$ as follows

$$\mathcal{S} := \{X_{P_j} | \exists \text{ no } i \neq j \text{ s.t. } I_{P_i} >_{\mathrm{d}} I_{P_j}\}. \qquad (3.1)$$

Therefore $\mathcal{S}$ is the set of all potential split variables for which a superior choice of split variable cannot be found. Let $m := |\mathcal{S}|$. We create $m-1$ identical copies of the current tree. This produces m trees, each of which uses a different split variable, chosen from $\mathcal{S}$ to continue the current path.

If it is determined that no split variable is preferable to not splitting, the node becomes a leaf. The method in [2] uses the relative frequency of the current data set to choose the most likely category (if two or more categories are equally likely, the most likely category at the parent node is chosen). One alternative would be to use the NPI method for ordinal data to construct a structure for the category probabilities.

A method by which the structures of the trees can be combined is now required. All trees should not be given equal weight, or some choices for split variables may dominate others. For example, consider three Boolean attribute variables X_1, X_2, X_3. Variables X_1 and X_2 are chosen for the initial split, so X_i is chosen as the root node for Tree i, with $i = 1, 2$. In Tree 1, whatever the value of X_1, we split on X_2 next. In Tree 2, when $X_2 = 0$, we split on X_1 next, but when $X_2 = 1$, we cannot choose between splitting next on X_1 or X_3. Therefore Tree 2 splits upon X_1, and a new tree is generated, Tree 3, which splits upon X_3. No further splits are made (see Figure 1).

Giving each of these three trees equal weight would imply that the initial choice of X_2 is twice as desirable as the choice of X_1. There is no justification for this.

There are various ways to tackle this issue. We could weight each tree according to the nature of its relevant entropy interval: how wide it is, and how far it lies from the interval for the full data set which it is dominating. For now, however, each tree is given weight one, which decreases each time there is a split after the root node. In the situation shown in Figure 1, Tree 1 would be given a weight of 1, and Trees 2 and 3 a weight of $\frac{1}{2}$ each, ensuring each choice of root node is given equal weight. This method is equivalent to creating duplicate trees. For example, in the situation shown in Figure 1, rather than weight each tree, Tree 1 could be duplicated.

There are many potential methods by which such a forest can be used to classify a data point. Denote by $\mathcal{X}$ the set of all categorisations given by the forest. The most conservative approach would be to simply return $\mathcal{X}$, making the imprecise tree a credal classifier (see e.g. [13]). Alternatively, for each c_i we can sum the weights of the trees which returned c_i, denoted Σ_i and choose c_* where $\Sigma_* = \max_i \Sigma_i$, selecting randomly amongst all categories for which $\Sigma_i = \Sigma_*$ if the maximum is non-unique. A third option is to return each category c_i for which $\Sigma_i \geq C$, for some constant C. Setting $C = 0$ ($C = \Sigma_*$) reduces to the first (second) method. These approaches will be compared in a later paper.

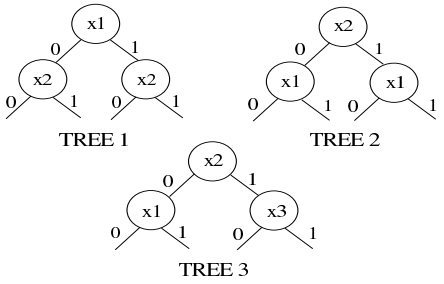


Figure 1: Forest generated by this method

Example 5.1 A small table of data, shown in Table 2, is used to draw an imprecise classification tree. Note that this example is included to illustrate our method, not to consider its efficacy. Indeed, the binary categories in this data set are categorical, not ordinal. Since $K = 2$, we can use the maximum entropy algorithm given in [2] (see Section 5 for the minimum entropy algorithm). We use the first forty data points as a training set, and the final ten as a test set. The binary nature of the category set makes it easy to calculate entropies:

$$H(\boldsymbol{v}) = -\frac{n_1^R}{a}\ln(\frac{n_1^R}{a}) - \frac{n_2^R}{a}\ln(\frac{n_2^R}{a}) \qquad (3.2)$$

where n_i^R is the number of instances of category i and $a := n_1^R + n_2^R$.

The method generates three trees, displayed in Figures 3 to 5. Tree 1 has weight 1, the others have weight $\frac{1}{2}$. Each leaf is labelled with the category assigned to it. Note that it appears in some cases that the set of possible values of the attribute variables changes from tree to tree. This is because, depending on previous attribute variables, the set of data R under consideration may not contain any instances of one or more values of the variable chosen as the split variable.

It is simple to compare this method with the one given in [2], as that method generates only one tree, which is in fact Tree 2. Tree 2 is correct 8 times out of 10. The imprecise tree is correct every time, though it returns both categories on two occasions; this is true irrespective of the choice of C. There is a one-to-one correspondence between the data points incorrectly classified by Tree 2, and the data points for which the imprecise tree gives both categories.

4 Maximum entropy algorithm

In this section the algorithm in [2] is adapted for the ordinal NPI method. We will provide an example later in the section demonstrating why that algorithm cannot be applied to the ordinal NPI case directly, but in short, it fails because it requires that the structure be convex, which is not the case here.

Our algorithm is too complex to be described in full, instead an overview is presented, including the relevant lemmas and proofs. This complexity is required despite the

fact that the structure and the function to be maximised are both simple in form, because the constraints upon the maximisation problem are complicated by the conditions imposed by the ordinal NPI approach. For example, whilst we have that $L_i \leq p_i \leq U_i$ and $L_{i+1} \leq p_{i+1} \leq U_{i+1}$, we also have that $L_i + L_{i+1} + \frac{1}{n+1} \leq p_i + p_{i+1} \leq U_i + U_{i+1} - \frac{1}{n+1}$. Moreoever, the sum of three adjacent elements will have its own constraints, and so on.

Maximum Entropy Algorithm

This algorithm is broadly similar to one presented in [6], which utilises NPI for the categorical case. Our consideration of the probability mass being available "on either side" of a component is based on the approach in that reference.

The algorithm requires two K-vectors $\boldsymbol{v}^L := (0, 1, \ldots, 1)$ and $\boldsymbol{v}^R := (1, \ldots, 1, 0)$. The j-th component of $\boldsymbol{v}^L$ ($\boldsymbol{v}^R$) represents the amount of mass available to $\hat{p}_j$ to the left (right). These vectors are updated after each mass assignment.

The algorithm can be broken down into two processes. The first process assigns the mass between two components in situations in which only those two components need be considered. For example, for $j \in \{1, \ldots, K-2\}$, let

$$|\hat{p}_j - \hat{p}_{j-1}| \geq \frac{2}{n+1}, \tag{4.1}$$

and assume without loss of generality (WLOG) that $\hat{p}_j > \hat{p}_{j-1}$. The convex nature of the entropy function means entropy is maximised by adding mass to the smallest components of $\hat{p}$ possible. Even if all mass to the left of c_{j-1} is assigned to $\hat{p}_{j-1}$, then that component will be no larger than $\hat{p}_j - \frac{1}{n+1}$. This means we must assign all mass between c_j and c_{j-1} to $\hat{p}_{j-1}$. An exception is the case where $n_{j-1} = n_{j-2} = 0$. In this case we cannot be sure that assigning all mass to $\hat{p}_{j-1}$ is justified. We do however know that no mass to the right of c_j can be assigned to $\hat{p}_j$. An extension of this argument can be applied to the boundary components and their neighbours.

Further, if there exists any adjacent components for which $|\hat{p}_j - \hat{p}_{j-1}| = \frac{1}{n+1}$, v^R_{j-1} and v^L_j are non-zero, there may be another assignment to be made. If we assume WLOG that $\hat{p}_{j-1} < \hat{p}_j$, then if $v^L_{j-1} = 0$, we must assign the mass between c_j and c_{j-1} to $\hat{p}_{j-1}$. Again, in situations with consecutive categories with zero observations, we may at this stage be only able to decide that some components *cannot* be assigned available mass, without being able to decide which components *should* be assigned that mass. Lastly, if $\hat{p}_j = \hat{p}_{j-1} = 0$ and $v^R_j = v^L_{j-1} = 0$, the mass is shared equally between the components.

The second process can consider three or more components simultaneously, using the concept of *strings*.

Definition 4.1 The categories $c_a, c_{a+1}, \ldots, c_b$ form a string if $v^R_j + v^L_j > 0$ for all $a \leq j \leq b$ and further $v^R_{a-1} = 0$ when $c_a \neq c_1$ and $v^L_{b+1} = 0$ when $c_b \neq c_K$.

We define $\mathcal{S} := \{c_a, c_{a+1}, \ldots, c_b\}$, and refer to the vector $(\hat{p}_a, \hat{p}_{a+1}, \ldots, \hat{p}_b)$ as the *string vector*. The *length* of a string equals the number of classes in the string.

The algorithm finds a string within the vector $\hat{\boldsymbol{p}}$ (which might be the entire vector), assigns mass to either reduce the length of the string or split it in two, and then finds another string, until none remain. By definition there cannot be more than $[\frac{K}{2}]$ strings, each of maximum length K, so the number of iterations required to assign all available mass must be less than $\frac{K(K+1)}{2}$.

The algorithm makes use of the following result.

Lemma 4.1 Let $\{c_1, \ldots, c_K\}$ contain the set of strings $\zeta := \{S_i,\ i = 1, \ldots, r\}$. String S_i has length l_i and contains categories c_{a_i} to c_{b_i}; $a_{i+1} > b_i$. Categories with their mass assignments already determined will belong to no S_i. The vector maximising entropy in the structure is uniquely determined by the vector maximising entropy over ζ. Let $\boldsymbol{w}^i$ represent the observations $(n_{a_i}, \ldots, n_{b_i})$. Consider $\boldsymbol{w}^i$ as a complete observation vector, and generate $\boldsymbol{v}^i$ as the corresponding vector maximising entropy. Entropy over ζ is maximised by the vector $d(t_1\boldsymbol{v}^1, t_2\boldsymbol{v}^2, \ldots, t_r\boldsymbol{v}^r)$ for $t_i = \frac{l_i+1}{K+1}$ and normalising constant d.

Proof.

$$H(\hat{\boldsymbol{p}}) = -\sum_{j=1}^{K} \hat{p}_j \ln(\hat{p}_j) = -\sum_{j:c_j \in \zeta} \hat{p}_j \ln(\hat{p}_j) + C$$

where $C = -\sum_{j:c_j \notin \zeta} \hat{p}_j \ln(\hat{p}_j)$ is constant. Therefore maximising $H(\hat{\boldsymbol{p}})$ is equivalent to maximising $-\sum_{j:c_j \in \zeta} \hat{p}_j \ln(\hat{p}_j)$. Moreover, when considering the maximum algorithm for $\tilde{S}_i$,

$$\begin{aligned} \frac{H(\hat{\boldsymbol{p}})}{m} + \frac{\ln(m)}{m} &= -\sum_{k=1}^{l_i} \left(\frac{\hat{p}_k}{m}(\ln(\hat{p}_k) - \ln(m))\right) \\ &= -\sum_{k=1}^{l_i} \left(\frac{\hat{p}_k}{m}(\ln(\frac{\hat{p}_k}{m}))\right) \\ &= H(\frac{\hat{\boldsymbol{p}}}{m}) \end{aligned} \tag{4.2}$$

where $m > 0$ is constant. The final expression in (4.2) is a slight abuse of notation, since entropy is generally only considered in terms of probability distributions, but there is no mathematical problem with considering it as a function over all l_i-vectors with non-negative elements. We define by V^i_a the set of all l_i-vectors with non-negative elements which sum to a (hence V^K_1 is the set of all possible probability distributions over the categories), and define by H_{S_i} the contribution to the overall entropy supplied by the string S_i. This means that 4.2 leads to

$$\begin{aligned} H(\boldsymbol{v}^*) &= \max_{\boldsymbol{v} \in V^i_1} H(\boldsymbol{v}) = t_i(\max_{\boldsymbol{v} \in V^i_{t_i^{-1}}} H(\boldsymbol{v})) - \ln(t_i) \\ &= t_i(H(\frac{\boldsymbol{v}^*}{t_i})) - \ln(t_i) \end{aligned} \tag{4.3}$$

so the vector which maximises entropy over $\tilde{S}_i$ is a positive multiple of the vector which maximises the contribution

of string S_i to the whole vector. Hence

$$\begin{aligned}\max H(\boldsymbol{p}) &= \max(-\sum_{j:c_j\in\zeta} p_j \ln(p_j)) + C \\ &= \sum_i^r (t_i(\max_{\boldsymbol{v}\in V_{t_i^{-1}}^i} H_{S_i}(\boldsymbol{v})) - \ln(t_i)) + C_2 \\ &= \sum_i^r t_i(\max_{\boldsymbol{v}\in V_{t_i^{-1}}^i} H_{S_i}(\boldsymbol{v})) + C \\ &= \sum_i^r t_i(H_{S_i}(\boldsymbol{v}^i)) + C_2 \qquad (4.4)\end{aligned}$$

where the inclusion of the t_i is justified by the need to rescale each vector $\boldsymbol{v}^i$, and C_2 is constant. This completes the proof. □

From Lemma 4.1 each string can be considered independently, as though its corresponding observation vector was the only one under consideration, with c_{a_i} and c_{b_i} as boundary categories. Let $x_1 := \min_{j\in\mathcal{S}} \hat{p}_j$ and $x_2 := \max_{j\in\mathcal{S}} \hat{p}_j$. In each case, a mass assignment is made and the vectors $\boldsymbol{v}^L$ and $\boldsymbol{v}^R$ are updated accordingly.

There are nine different types of string, all of which the algorithm deals with differently. A full description of these methods will be presented in a later paper; the methods described below are by no means exhaustive and are merely intended as a demonstration of the ideas involved.

We aim to ensure the components of a string are as close to being equal as is possible. If $x_2 > x_1 + \frac{1}{n+1}$, equality is impossible. However, in this case we can denote by y_1 and $y_2 > y_1$ the smallest integers for which $\hat{p}_{y_1} \in \{x_1, x_2\}$, $y_2 = I_{[\hat{p}_{y_1}=x_2]}x_1 + I_{[\hat{p}_{y_1}=x_1]}x_2$, and for which $\hat{p}_j \notin \{x_1, x_2\}$ for all $y_1 < j < y_2$ (where I_A is the indicator function). We have that $\min(\hat{p}_{y_1}, \hat{p}_{y_2}) = x_1$ and $\max(\hat{p}_{y_1}, \hat{p}_{y_2}) = x_2$. Moreover, for any $\min(y_1, y_2) < j < \max(y_1, y_2)$, $x_1 < \hat{p}_j < x_2$.

Assume WLOG that $p_{y_1} = x_1$, and that $x_1 > 0$ (if $x_1 = 0$ we require a different approach, not described here). The mass between c_{y_1} and c_{y_1-1} must be available, and even if this mass is assigned entirely to $\hat{p}_{y_1}$, $\min\{\hat{p}_{y_1}, \ldots, \hat{p}_{y_2}\} = \hat{p}_{y_1}$ holds. Each component between $\hat{p}_{y_1}$ and $\hat{p}_{y_2}$ therefore requires at least $\frac{1}{n+1}$ mass to reach the value of $\hat{p}_{y_2}$, as does $\hat{p}_{y_1}$ itself, so all available mass will be assigned before $\hat{p}_{y_2}$ is eligible to receive any of it. The mass between c_{y_2-1} and c_{y_2} is therefore assigned to $\hat{p}_{y_2-1}$.

In the case where $x_2 = x_1 + \frac{1}{n+1}$, we denote by y_1 the number of components equal to x_1, and by y_2 the number equal to x_2. Therefore, in this situation, we would want to use a total mass $\frac{y_1}{n+1}$ to increase all minima from x_1 to x_2. We would then share the remaining $\frac{y_2}{n+1}$ mass equally between all components.

This mass assignment is not always allowed in our NPI structure, as shown by the example below.

Consider the situation defined by the observation vector $(2, 2, 3, 2, 2, 2, 3, 2, 2, 2, 2)$, with associated lower probability vector $\frac{1}{25}(2, 1, 2, 1, 1, 1, 2, 1, 1, 1, 2)$. Entropy would be maximised by increasing each element equal to $\frac{1}{25}$ by $(\frac{1}{25})(\frac{14}{11}) = \frac{14}{275}$, and each increasing element equal to $\frac{2}{25}$ by $(\frac{1}{25})(\frac{3}{11}) = \frac{3}{275}$. Note that this is also the assignment given by the algorithm in [2]. Figure 1 shows this assignment is not possible.

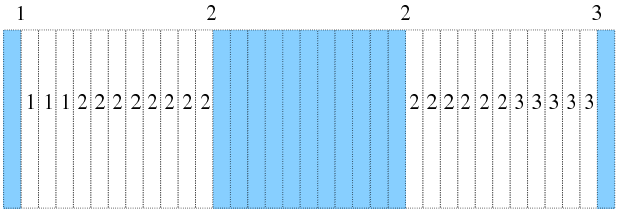


Figure 2: Inability to apply Abellan/Moral algorithm

Figure 1 shows the available mass between the pairs of categories c_1 and c_2, and c_2 and c_3. Each of these pieces of available mass are of size $\frac{1}{25}$, and each of the smaller rectangles has mass $\frac{1}{275}$. It is clear from the diagram that assigning enough probability mass to $\hat{p}_1$ and $\hat{p}_2$ to make them both equal to $\frac{25}{275}$ forces the value of $\hat{p}_3$ to become at least $\frac{27}{275}$.

Therefore the available probability mass cannot be spread across $\hat{\boldsymbol{p}}$ so as to produce a uniform distribution. The available mass that must be assigned to categories c_1, c_2 and c_3 is too great. We therefore require not only a method for assigning mass that is compatible with our model, but a proof demonstrating that the resulting distribution does indeed give the maximum entropy possible.

This approach is summarised as follows. If $x_2 = x_1 + \frac{1}{n+1}$, we try to assign the uniform distribution $\frac{1}{n}(1, \ldots, 1)$, category by category, from left to right, where n is a function of the number of categories within the string with associated components equal to x_1 (when $x_1 = x_2$, this assignment is automatically successful). This assignment will fail if either all available mass to some category c_i is assigned without $\hat{p}_i$ reaching $\frac{1}{n}$, or $\hat{p}_i$ reaches $\frac{1}{n}$ and yet mass remains available which must be assigned to category c_i.

Once either event occurs (if neither does, then the uniform distribution can in fact be reached), we start again from $\hat{p}_{c_i}$ and attempt the same mass assignment. This may occur several times. As a result, we will have divided the components in the string vector into two or more sets. Each set has either too much or too little available mass for the desired assignment. Moreover, there must exist sets $\{c_{a_1}, \ldots, c_{b_1}\}$ and $\{c_{b_1+1}, \ldots, c_{b_2}\}$ for which one has too much mass, and the other too little. In such a situation, the algorithm can create two new strings, one ending with c_{b_1} and another beginning with c_{b_1+1}. The justification for this follows from Lemma 4.2.

Lemma 4.2 Let $0 < l_1 < n$ and $l_2 = n - l_1$. Let $\mathcal{M}$ be a closed credal set of n-vectors for which $\boldsymbol{u} \in \mathcal{M} \Leftrightarrow \boldsymbol{u} = (\boldsymbol{v}, \boldsymbol{w})$, where $\boldsymbol{v}$ is a l_1-vector and $\boldsymbol{w}$ is a l_2-vector with the following properties: $\sum_{i=1}^{l_1} v_i > k_1$ and $\sum_{j=1}^{l_2} w_j < k_2$, and $\frac{k_1}{l_1} > \frac{k_2}{l_2}$. Then if $(\frac{k_1}{l_1}(1, \ldots, 1), \frac{k_2}{l_2}(1, \ldots, 1))$ lies within $\mathcal{M}$, it is the distribution which maximises entropy.

Proof. From Lemma 4.1 entropy orderings are invariant

-	O-NPI	IDM	IGR	IG	W-L
O-NPI	- -	(6,14) (0,4)	(11,9) (1,4)	(11,9) (2,4)	-4 -9
IDM	(14,6) (4,0)	- -	(14,6) (3,2)	(11,9) (1,1)	18 5
IGR	(9,11) (4,1)	(6,14) (2,3)	- -	(10,10) (0,0)	-10 2
IG	(9,11) (4,2)	(9,11) (1,1)	(10,10) (0,0)	- -	-4 2

Table 1: Comparison of methods

over multiplication by a constant, so we can assume $k_1 + k_2 = 1$. Also by Lemma 4.1, following the assignment of mass α to $\boldsymbol{v}$ and mass $1-\alpha$ to $\boldsymbol{w}$, the entropy is maximised by setting $v_i = \frac{\alpha}{l_1}$, $\forall i$ and $w_j = \frac{1-\alpha}{l_2}$, $\forall j$. Clearly, $\alpha \geq k_1$ and $1-\alpha \leq k_2$. The proof, then, reduces to demonstrating that, indeed, entropy is maximised by setting $\alpha = k_1$.

Construct a structure for which $\boldsymbol{L} = (\boldsymbol{v}_L, \boldsymbol{w}_L)$ and $\boldsymbol{U} = (\boldsymbol{v}_U, \boldsymbol{w}_U)$, where $(\boldsymbol{v}_L)_i = \frac{k_1}{l_1}$ and $(\boldsymbol{v}_U)_i = 1$ $\forall i$, and $(\boldsymbol{w}_L)_j = 0$ and $(\boldsymbol{w}_U)_j = \frac{k_2}{l_2}$ $\forall j$. By the algorithm in [2], the elements corresponding to $\boldsymbol{w}$ must all reach their upper bound before the elements of $\boldsymbol{v}$ are considered at all. Therefore, giving more mass to $\boldsymbol{v}$ than the minimum requirement violates the algorithm, and so the given assignment must, indeed, maximise entropy. □

Lemma 4.2 proves that when $\{c_{a_1}, \ldots, c_{b_1}\}$ has too much (too little) mass and $\{c_{a_2}, \ldots, c_{b_2}\}$ has too little (too much) mass for the uniform distribution to be assigned, entropy cannot be increased by taking mass from the set of categories with less mass, so long as both sets can separately be assigned mass in such a way as to make all components equal. However, if such an assignment is not possible, we can simply split the set up again, and once again apply Lemma 4.2, and so on. This justifies splitting the string as described.

We now compare our method with the IDM imprecise method, along with the Info Gain [9] and Info Gain Ratio methods [9] over 21 data sets in which the categories can plausibly be argued to be ordinal. All these methods were run using the computer package known as WEKA. The results are given in Table 1. The first pair of numbers in each cell represent the number of wins and losses for the row classifier with respect to the column classifier, in terms of percentage of correct classification. For example, the pair (12,6) in the second row tells us that the IDM classifier outperformed the O-NPI classifier 12 times, and was outperformed 6 times. The second pair of numbers also represent wins and losses, this time using a paired t-test at the 5% level. The final row of the table gives the total number of wins minus the total number of losses for both tests.

These results do not show any obvious improvement in replacing the IDM structure with that of ordinal NPI. Indeed, it seems to be performing worse than the IDM method, and roughly equivalently to the IGR and IG methods.[1]

We now describe the algorithm for minimising entropy.

5 Minimum entropy algorithm

Minimising entropy is difficult in general. However, in the specific case of ordinal NPI, it is quite simple. We begin with three lemmas.

Lemma 5.1 When minimising entropy, all available mass between observed categories is assigned entirely to one of the associated components.

Proof. Let $H_2(v_1, v_2) = -v_1 \ln(v_1) - v_2 \ln(v_2)$ be the contribution of two components to the entropy. The entropy function is concave, so if $b \leq c$ we have

$$H_2(a+c, a) \leq H_2(a+c-b, a+b). \tag{5.1}$$

Therefore for any values a and c, $H_2(\cdot,\cdot)$ is minimised when either $b = 0$ or $b = c$. Let the adjacent observed categories have components $\hat{p}_i = a$ and $\hat{p}_j = a + c$, and let the mass between them be denoted by $m > 0$. Then from (5.1) we have two inequalities

$$\begin{aligned} H_2(a+c+m, a) &\leq H_2(a+c+m-b, a+b) \\ H_2(a+c, a+m) &\leq H_2(a+c-b, a+b+m), \end{aligned}$$

meaning the minimum entropy occurs when the entirety of m is assigned either $\hat{p}_i$ or $\hat{p}_j$. □

Lemma 5.2 When minimising entropy, no unobserved category is assigned mass.

Proof. Setting $a = 0$ in (5.1) shows any assignment of mass to a zero component leads to an increase in entropy compared to assigning that mass to a non-zero component. Therefore this should never be done if an alternative is available, which is always the case for unobserved categories in the ordinal NPI case. □

Therefore the minimum entropy algorithm can operate simply by assigning all unassigned mass between observed categories c_i and c_j entirely to $\hat{p}_i$ or to $\hat{p}_j$.

Lemma 5.3 When minimising entropy, for any two components $\hat{p}_i$ or $\hat{p}_j$ corresponding to adjacent observed internal categories, the mass between c_i and c_j is assigned to $\hat{p}_i$ if $\hat{p}_i > \hat{p}_j$.

Proof. Consider any pair of adjacent observed internal categories c_j, c_{j+1} for which $\hat{p}_j \neq \hat{p}_{j+1}$. Assume WLOG that $\hat{p}_j < \hat{p}_{j+1}$, and set $\hat{p}_j =: r_2$ and $\hat{p}_{j+1} =: r_3 = r_2 + \frac{1}{n+1} + \alpha$ for some $\alpha \in \mathbb{N}$. Since both categories are internal, we can also consider the components $\hat{p}_{j-1} =: r_1$ and $\hat{p}_{j+2} =: r_4$. Between these four categories is available mass $\frac{3}{n+1}$, and from Lemma 5.1 we have that three (not necessarily distinct) components must receive $\frac{1}{n+1}$ mass.

[1] Note that the WEKA code for the exact O-NPI algorithm is still in development.

There are eight ways that this mass assignment can be carried out. These can be divided into four pairs. In each pair one assignment gives the mass between $\hat{p}_j$ and $\hat{p}_{j+1}$ to $\hat{p}_j$, and in the other it gives it to $\hat{p}_{j+1}$; the other two mass assignments are identical. The eight mass assignments are given in the table below, along with the resulting size of each component. Each pair is presented together, and the first case in the pair is always the one in which the smaller component is assigned the mass.

-	r_1	r_2	r_3	r_4
1	r_1	$r_2 + \frac{2}{n+1}$	$r_2 + \frac{1}{n+1} + \alpha$	$r_4 + \frac{1}{n+1}$
2	r_1	$r_2 + \frac{1}{n+1}$	$r_2 + \frac{2}{n+1} + \alpha$	$r_4 + \frac{1}{n+1}$
3	r_1	$r_2 + \frac{2}{n+1}$	$r_2 + \frac{2}{n+1} + \alpha$	r_4
4	r_1	$r_2 + \frac{1}{n+1}$	$r_2 + \frac{3}{n+1} + \alpha$	r_4
5	$r_1 + \frac{1}{n+1}$	$r_2 + \frac{1}{n+1}$	$r_2 + \frac{1}{n+1} + \alpha$	$r_4 + \frac{1}{n+1}$
6	$r_1 + \frac{1}{n+1}$	r_2	$r_2 + \frac{2}{n+1} + \alpha$	$r_4 + \frac{1}{n+1}$
7	$r_1 + \frac{1}{n+1}$	$r_2 + \frac{1}{n+1}$	$r_2 + \frac{2}{n+1} + \alpha$	r_4
8	$r_1 + \frac{1}{n+1}$	r_2	$r_2 + \frac{3}{n+1} + \alpha$	r_4

We now prove that for all four pairs, the second assignment has lower entropy than the first. Therefore, irrespective of how the mass between $\hat{p}_{j-1}$ and $\hat{p}_j$ and between $\hat{p}_{j+1}$ and $\hat{p}_{j+2}$ is assigned, we must assign the mass between $\hat{p}_j$ and $\hat{p}_{j+1}$ to the larger component.

This is immediately clear for the second, third and fourth pairs by the concave nature of the entropy function; entropy is minimised by setting two values as far apart as possible. For the first pair, the two cases are equivalent when $\alpha = 0$, and we still lose nothing by assigning the mass to the larger component. If $\alpha \geq 1$, we can define $\alpha - \frac{1}{n+1} =: \alpha_1 \geq 0$ and re-write the pair

-	r_1	r_2	r_3	r_4
1	r_1	$2_1 + \frac{2}{n+1}$	$r_2 + \frac{2}{n+1} + \alpha_1$	$r_4 + \frac{1}{n+1}$
2	r_1	$2_1 + \frac{1}{n+1}$	$r_2 + \frac{3}{n+1} + \alpha_1$	$r_4 + \frac{1}{n+1}$

and once again from the fact that the entropy function is concave we see that minimising entropy requires assigning the mass to the larger component.

We have then that for all mass assignments between the category pairs c_{j-2}, c_{j-1} and c_j, c_{j+1}, entropy is minimised by adding the available mass to the larger component. Therefore every individual slice of available mass between c_j and c_{+1} can be considered separately, so long as both neighbouring categories are internal, and $\hat{p}_j \neq \hat{p}_{j+1}$. □

Boundary categories are handled in a similar way. If a boundary category is unobserved, no mass will be assigned to it. Otherwise, we can consider, say, $\hat{p}_1$ as being equivalent to an internal category for which the mass on the left has already been assigned elsewhere. This allows us to make use of Lemma 5.3.

Lastly we deal with situations in which there are consecutive equal components. Suppose there are n consecutive components all of size m. If $n = 2$, we can assign the mass to either component. If $n = 3$, we assign all mass to the central component, as $H(m + a, m, m + a) > H(m, m + 2a, m)$. This leaves the third component unchanged, and so we can use this more generally to reduce the number of consecutive components from n to $n - 2$. This means we can continually re-apply this assignment until either all mass has been assigned, or we are left with just two equal components with available mass between them. We then simply assign that mass to either side.

Note that our algorithm runs from left to right, assigning mass to the larger component each time, and once finished, returns and deals with each sequence of equal components. Running the algorithm from right to left might result in a different vector being returned. In other words, the vector we find results in a global minimum for entropy, but it does not follow that no other vector could not also produce a global miniminum for entropy. As an obvious example, consider $K = 3$ and observation vector (1,0,1). Clearly $\boldsymbol{L} = \frac{1}{3}(1, 0, 1)$ and $\boldsymbol{U} = \frac{1}{3}(2, 1, 2)$. Our minimisation algorithm returns the vector $\frac{1}{3}(2, 0, 1)$, but clearly the vector $\frac{1}{3}(1, 0, 2)$ will have an identical entropy value.

6 Conclusions and Further Work

At each stage of tree construction the method presented here allows for the possibility that we cannot choose between potential split variables. This has been considered previously regarding choice of root node [4], but we are aware of no method in which this idea is applied to the construction of the whole tree, or one which compares entropy ranges. Clearly, it remains to test this method against others - at the time of writing the WEKA code for our method has not yet been written - but by expanding focus beyond the root nodes and by utilising comparisons between intervals, this method combines classification and imprecise probability in an attractive way, by recognising situations in which it is unreasonable to consider one split variable choice as clearly superior to another.

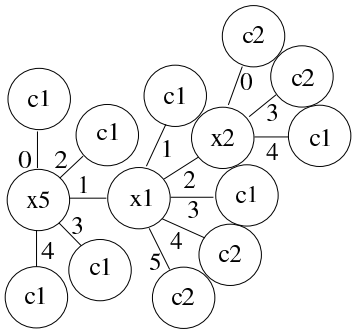


Figure 3: Example 5.1: Tree 1

Further work is required on considering how best to collate the set of categories given by the imprecise decision tree for each data point. It would also be of value to consider more thoroughly the implications of minimising entropy, particularly with regard to unobserved categories. One of the criticisms regarding attempts to maximise entropy is that it invariably gives as much mass as possible to categories that were unobserved. This is difficult to justify

	A_1	A_2	A_3	A_4	A_5	C
1	1	3	0	0	2	1
2	1	3	0	1	1	1
3	1	2	0	1	0	1
4	1	2	0	1	0	1
5	1	3	0	1	0	1
6	1	3	0	1	3	2
7	1	0	1	0	1	1
8	1	0	1	0	1	2
9	1	1	0	1	0	1
10	1	1	0	1	1	1
11	5	0	0	0	1	2
12	5	0	0	0	1	2
13	5	0	1	0	0	1
14	5	0	0	0	0	1
15	2	3	0	1	1	2
16	2	4	1	0	1	1
17	1	1	1	1	1	2
18	2	4	0	1	1	1
19	2	0	1	0	0	1
20	2	2	0	1	2	1
21	2	0	0	0	1	2
22	3	1	0	1	0	1
23	3	3	1	1	1	1
24	3	1	0	1	0	1
25	1	0	1	1	4	1
26	2	1	0	0	3	1
27	3	1	0	0	1	1
28	3	0	0	0	2	1
29	3	0	1	1	1	2
30	3	1	0	1	0	1
31	4	2	0	0	0	1
32	4	1	1	1	0	1
33	4	0	1	0	1	2
34	3	1	0	1	1	1
35	4	0	0	0	1	2
36	4	0	0	0	0	1
37	4	0	1	0	1	1
38	4	0	0	0	1	2
39	5	0	0	0	0	1
40	4	0	0	1	1	2
41	3	0	0	1	1	1
42	3	0	1	0	0	1
43	3	1	1	0	1	1
44	5	0	0	0	1	2
45	4	0	0	1	0	1
46	5	0	0	0	1	2
47	5	0	0	0	0	1
48	5	0	0	0	0	1
49	2	1	0	1	2	1
50	4	1	0	1	0	1

Table 2: Data set for imprecise tree

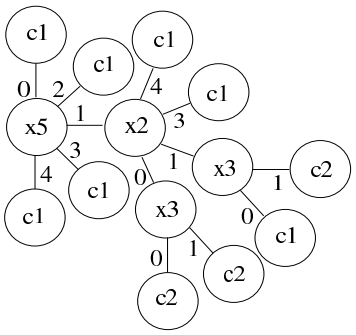


Figure 4: Example 5.1: Tree 2

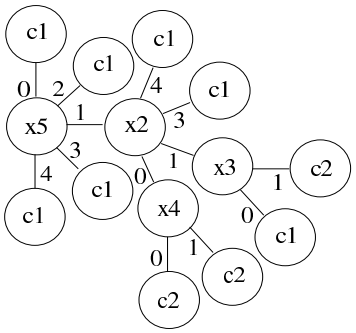


Figure 5: Example 5.1: Tree 3

theoretically. Moreover, the amount of mass given to each unobserved category depends on how the unobserved categories are described, and how many there are, which may cause problems. In contrast, minimising entropy guarantees that no unobserved category will be given any mass, side-stepping the issue of how to label and quantify unobserved categories.

Finally, we should also consider using alternative information measures (one such alternative is the Gini index [10]) to generate imprecise decision trees.

Acknowledgements

This work was supported by the UK National Institute of Health Research, and by the Spanish Consejería de Economía, Innovación y Ciencia de la Junta de Andalucía, under project TIC-06016, which supported the first and second authors respectively. We would like to thank our two reviewers for their comments and suggestions.

References

[1] Abellan, J. (2006) Uncertainty measures on probability intervals from the imprecise Dirichlet model, *International Journal of General Systems*, Vol 35, **5**, 509-528.

[2] Abellan, J. & Moral, S. (2003) Building classification trees using the total uncertainty criterion, *International Journal of Intelligent Systems* 18 (12), 1215-1225.

[3] Abellan, J. & Moral, S. (2005) Difference of entropies as a non-specifity function on credal sets, *Interna-*

tional Journal of General Systems, Vol 34, **3**, 201-214.

[4] Abellan, J. & Masegosa, A. (2010) An ensemble method using credal decision trees, *European Journal of Operations Research*, 205 (1), 218-226.

[5] Abellan, J., Baker, R.M. & Coolen, F.P.A. (2011) Maximising entropy on the nonparametric predictive inference model for multinomial data, *European Journal of Operational Research* , 212(1) 112-122.

[6] Baker, R.M. (2010) *Multinomial Nonparametric Predictive Inference: Selection, Classification and Subcategory Data*, PhD thesis, www.etheses.dur.ac.uk/257/

[7] Coolen, F.P.A., Coolen-Schrijner, P. & Maturi, T.A. (2010) On nonparametric predictive inference for ordinal data. In: E. Hullermeier *et al* (eds), *Computational Intelligence for Knowledge-Based Systems Design*, Springer, pp 188-197.

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

[9] Quinlan, J.R. (1986) Induction of decision trees, *Machine Learning 1*, 81-106.

[10] Strobl, C. & Augustin, T. (2009) Adaptive selection of extra cutpoints - an approach towards reconciling robustness and interpretability in classification trees, *Journal of Statistical Theory and Practice*, Vol 3, 119-135.

[11] Troffaes, M.C.M. (2007) Decision making under uncertainty using imprecise probabilities, *International Journal of Approximate Reasoning*, Vol 45, 17-29.

[12] Walley, P. (1996) Inferences from multinomial data: learning about a bag of marbles, *Journal of the Royal Statistical Society* B 58, 3-57.

[13] Zaffalon, M. (2002) The naive credal classifier. *Journal of Statistical Planning and Inference*, 105 (i1), 5-21.

L_p consonant approximation of belief functions in the mass space

Fabio Cuzzolin
Oxford Brookes University, Oxford, United Kingdom
Fabio.Cuzzolin@brookes.ac.uk

Abstract

In this paper we pose the problem of approximating an arbitrary belief function (b.f.) with a consonant one, in a geometric framework in which belief functions are represented by the vectors of their basic probabilities, or "mass space". Given such a vector $\vec{m}_b$, the consonant b.f. which minimizes an appropriate distance function from $\vec{m}_b$ can be sought. We consider here the classical L_1, L_2 and L_p norms. As consonant belief functions live in a collection of simplices in the mass space, partial approximations on each individual simplex have to be computed in order to find the overall approximation. Interpretations of the obtained approximations in terms of basic probabilities are proposed, and the results compared with those of previous approaches, in particular outer consonant approximation.

Keywords. Consonant belief functions, (outer) consonant approximation, mass space, Lp norms.

1 Introduction

The theory of evidence (ToE) [22] is a popular approach to uncertainty description. Probabilities are there replaced by *belief functions* (b.f.s), which assign values between 0 and 1 to subsets of the sample space Θ instead of single elements. Possibility theory [10], on its side, is based on *possibility measures*, i.e., functions $Pos : 2^\Theta \to [0,1]$ on Θ such that $Pos(\bigcup_i A_i) = \sup_i Pos(A_i)$ for any family $\{A_i | A_i \in 2^\Theta, i \in I\}$ where I is an arbitrary set index. Given a possibility measure Pos, the dual *necessity* measure is defined as $Nec(A) = 1 - Pos(A^c)$.
Necessity measures have as counterparts in the theory of evidence *consonant* b.f.s, i.e., belief functions whose focal elements are nested [22]. The problem of approximating a belief function with a necessity measure is then equivalent to approximating a belief function with a consonant b.f. [1, 11, 15, 16]. As possibilities are completely determined by their values on the singletons $Pos(x)$, $x \in \Theta$, they are less computationally expensive than b.f.s, making the approximation process interesting for many applications. Several authors, such as Yager [25] and Romer [21] amongst others, have studied the connection between fuzzy numbers and Dempster-Shafer theory. Klir *et al* have published an excellent discussion [20] on the relations among fuzzy and belief measures and possibility theory. Heilpern [13] has also presented the theoretical background of fuzzy numbers connected with the possibility and Dempster-Shafer theories, describing some types of representation of fuzzy numbers and studying the notions of distance and order between fuzzy numbers based on these representations. Caro and Nadjar [2], instead, have suggested a generalization of the Dempster-Shafer theory to a fuzzy valued measure. The links between transferable belief model and possibility theory have been briefly investigated by Ph. Smets in [24].
Dubois and Prade [11], more specifically, have extensively worked on consonant approximations of belief functions. As belief functions are computationally expensive to work on (at least in a naive way), mapping them to necessity or possibility measures, which only depends on their values on singletons, can greatly reduce the complexity of making inferences or decisions under uncertainty. Dubois and Prade's work has been later considered in [15, 16]. In particular, the notion of "outer consonant approximation" has received considerable attention in the past. Indeed, belief functions admit the following order relation: $b \leq b' \Leftrightarrow b(A) \leq b'(A)\ \forall A \subseteq \Theta$, called "weak inclusion". It is then possible to introduce the notion of "outer consonant approximations" [11] of a belief function b, i.e., those co.b.f.s such that $\forall A \subseteq \Theta$ $co(A) \leq b(A)$. Dubois and Prade's work has been later extended by Baroni [1] to capacities. In [7] the author has indeed provided a comprehensive description of the geometry of the set of outer consonant approximations.

In recent times the opportunity of seeking probabil-

ity or consonant approximations/transformations of belief functions by minimizing appropriate distance functions has been explored. The author has himself introduced the notion of orthogonal projection $\pi[b]$ of a belief function onto the probability simplex [3], and studied consistent approximations of belief functions induced by classical L_p norms [8] in the space of belief functions [4]. In [6] he has shown that norm minimization can also be used to define families of geometric conditional b.f.s. Jousselme et al [17] have recently conducted a nice survey of the similarity measures between belief functions introduced so far. Other similarity measures between belief functions have been proposed by Shi et al [23], Jiang et al [14], and others [9, 14, 19]. Many of these measures could be in principle employed to define conditional belief functions, or approximate b.f.s by necessity measures.

Paper outline. In this paper we derive the expressions of all the consonant approximations of belief functions induced by minimizing L_p distances in the mass space (with respect to the counting measure on 2^Θ). After providing the necessary background on consonant b.f.s and the approximation problem (Section 2), we compute the approximations induced by L_1 (3.1), L_2 (3.2) and L_∞ (3.3) norms, respectively. Their interpretation in terms of mass re-assignment and their relation with outer consonant approximations are discussed in Section 4, and illustrated in the significant ternary case.

2 Consonant approximation

Consonant belief functions. We briefly recall here a few basis definitions. A *basic probability assignment* (b.p.a.) over a finite set *(frame of discernment [22])* Θ is a function $m_b : 2^\Theta \to [0,1]$ on its power set $2^\Theta = \{A \subseteq \Theta\}$ such that $m_b(\emptyset) = 0$ and $\sum_{A\subseteq\Theta} m_b(A) = 1$. Subsets of Θ associated with non-zero values of m_b are called *focal elements*. The *belief function* $b : 2^\Theta \to [0,1]$ associated with a basic probability assignment m_b on Θ is defined as: $b(A) = \sum_{B\subseteq A} m_b(B)$. The *plausibility function* (pl.f.) $pl_b : 2^\Theta \to [0,1]$, $A \mapsto pl_b(A)$, where $pl_b(A) \doteq 1 - b(A^c) = 1 - \sum_{B\subseteq A^c} m_b(B) = \sum_{B\cap A\neq\emptyset} m_b(B)$, expresses the amount of evidence *not against* A. A probability measure is simply a special belief function assigning non-zero masses to singletons only (*Bayesian* b.f.): $m_b(A) = 0$ $|A| > 1$. A belief function is said to be *consonant* if its focal elements are nested.

Mass vector representations. Given a frame Θ, each belief function $b : 2^\Theta \to [0,1]$ is completely specified by its $N-2$ belief values $\{b(A), \emptyset \subsetneq A \subsetneq \Theta\}$, $N \doteq 2^n$ ($n \doteq |\Theta|$), (as $b(\emptyset) = 0$, $b(\Theta) = 1$ for all b.f.s) and can therefore be represented as a point of $\mathbb{R}^{N-2}$ [4]. In the same way, each belief function is uniquely associated with the related set of mass values $\{m(A), \emptyset \subsetneq A \subseteq \Theta\}$ (Θ this time included). It can therefore be seen also as a point of $\mathbb{R}^{N-1}$, the vector $\vec{m}_b$ of its $N-1$ mass components:

$$\vec{m}_b = \sum_{\emptyset\subsetneq B\subseteq\Theta} m_b(B)\vec{m}_B, \tag{1}$$

where $\vec{m}_B$ is the vector of mass values associated with the ("categorical") mass function $\vec{m}_B$ assigning all the mass to a single event B: $\vec{m}_B(B) = 1$, $\vec{m}_B(A) = 0$ $\forall A \neq B$. Note that in $\mathbb{R}^{N-1}$ $\vec{m}_\Theta = [0, ..., 0, 1]'$ and cannot be neglected. However, since the mass of Θ is determined by all the other masses in virtue of the normalization constraint, we can also choose to represent mass vectors as vectors of $\mathbb{R}^{N-2}$ of the form $\vec{m}_b = \sum_{\emptyset\subsetneq B\subsetneq\Theta} m_b(B)\vec{m}_B$, in which this time the component $\bar{\Theta}$ is neglected. We will consider both representations in the following. The collection $\mathcal{M}$ of points which are valid basic probability assignments is a *simplex*[1], which we call *mass space*. $\mathcal{M}$ is the convex closure[2] $\mathcal{M} = Cl(\vec{m}_A, \emptyset \subsetneq A \subseteq \Theta)$.

The consonant complex. In this framework the geometry of consonant belief functions can be described in terms of *simplicial complexes* [12], i.e., collections Σ of simplices of arbitrary dimensions such that: 1. if a simplex belongs to Σ, then all its faces of any dimension belong to Σ; 2. the intersection of any two simplices is a face of both. Now, the region $\mathcal{CO}$ of consonant belief functions in the belief space is a simplicial complex [7]. Namely, $\mathcal{CO}$ is the union of a collection of (maximal) simplices, each of them associated with a maximal chain $\mathcal{C} = \{A_1 \subset \cdots \subset A_n\}$, $|A_i| = i$ of subsets of Θ. When the mass of some element of the maximal chain is zero, the simplicial coordinate of the associated b.f. is also zero. Analogously, the region of consonant belief functions in the mass space $\mathcal{M}$ will be the simplicial complex:

$$\mathcal{CO}_\mathcal{M} = \bigcup_{\mathcal{C}=A_1\subset\cdots\subset A_n} Cl(\vec{m}_{A_1}, \cdots, \vec{m}_{A_n}).$$

Binary example. In the case of a frame of discernment containing only two elements, $\Theta_2 = \{x, y\}$, each b.f. $b : 2^{\Theta_2} \to [0,1]$ is completely determined by its mass values $m_b(x)$, $m_b(y)$, as $m_b(\Theta) =$

[1] An n-dimensional *simplex* is the convex closure $Cl(x_1, ..., x_{n+1})$ of $n+1$ affinely independent points $x_1, ..., x_{n+1}$ of the Euclidean space $\mathbb{R}^n$. An *affine combination* of k points $v_1, ..., v_k \in \mathbb{R}^m$ is a sum $\alpha_1 v_1 + \cdots + \alpha_k v_k$ such that $\sum_i \alpha_i = 1$. The affine subspace generated by the points $v_1, ..., v_k \in \mathbb{R}^m$ is the set $\{v \in \mathbb{R}^m : v = \alpha_1 v_1 + \cdots + \alpha_k v_k, \sum_i \alpha_i = 1\}$. If $v_1, ..., v_k$ generate an affine space of dimension k they are said to be *affinely independent*.

[2] Here Cl denotes convex closure: $Cl(\vec{m}_1, ..., \vec{m}_k) = \{\vec{m} \in \mathcal{M} : \vec{m} = \alpha_1\vec{m}_1 + \cdots + \alpha_k\vec{m}_k, \sum_i \alpha_i = 1,\ \alpha_i \geq 0\ \forall i\}$.

$1 - m_b(x) - m_b(y)$ and $m_b(\emptyset) = 0$. We can therefore collect them in a vector of $\mathbb{R}^{N-2} = \mathbb{R}^2$ (since $N = 2^2 = 4$): $\vec{m}_b = [m_b(x), m_b(y)]' \in \mathbb{R}^2$. In this example we adopt therefore the $N-2$-dimensional version of the mass space. Since $m_b(x) \geq 0$, $m_b(y) \geq 0$, and $m_b(x)+m_b(y) \leq 1$ we can easily infer that the set $\mathcal{M}_2$ of all the possible basic probability assignments on Θ_2 can be depicted as the triangle in the Cartesian plane of Figure 1, whose vertices are the points $\vec{m}_\Theta = [0,0]'$, $\vec{m}_x = [1,0]'$, $\vec{m}_y = [0,1]'$, 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 Bayesian b.f.s on Θ_2 is the diagonal line segment $Cl(\vec{m}_x, \vec{m}_y)$.

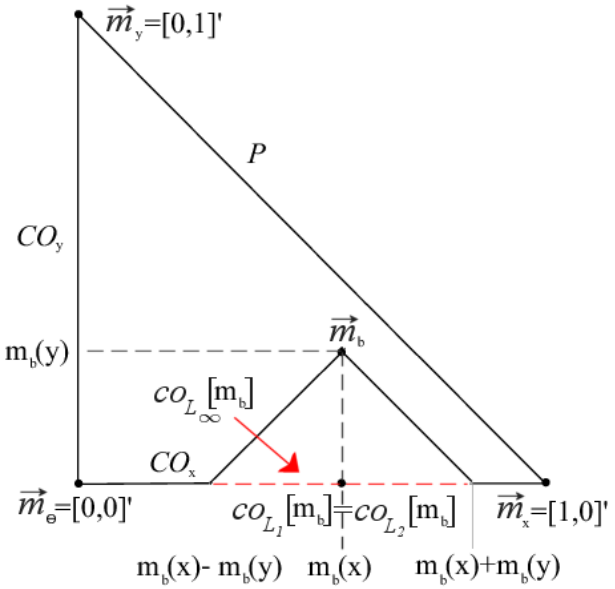


Figure 1: The belief space $\mathcal{M}_2$ for a binary frame is a triangle in $\mathbb{R}^2$ whose vertices are the mass vectors associated with the categorical belief functions focused on $\{x\}, \{y\}$ and Θ: $\vec{m}_x, \vec{m}_y, \vec{m}_\Theta$. Consonant b.f.s live in the union of the two segments $\mathcal{CO}_x = Cl(\vec{m}_\Theta, \vec{m}_x)$ and $\mathcal{CO}_y = Cl(\vec{m}_\Theta, \vec{m}_y)$. The unique $L_1 = L_2$ consonant approximation and the set of L_∞ consonant approximations (dashed) on $\mathcal{CO}_x$ are also shown.

Consonant approximations in the binary case. On $\Theta_2 = \{x, y\}$ consonant belief functions can have as chain of focal elements either $\{\{x\}, \Theta_2\}$ or $\{\{y\}, \Theta_2\}$. Therefore the region $\mathcal{CO}_2$ of all the co.b.f.s on Θ_2 is the union of two segments (see Figure 1): $\mathcal{CO}_2 = \mathcal{CO}_x \cup \mathcal{CO}_y = Cl(\vec{m}_\Theta, \vec{m}_x) \cup Cl(\vec{m}_\Theta, \vec{m}_y)$.

Figure 1 illustrates the L_p consonant approximations of a given $\vec{m}_b$ as well. We can notice that the L_1 and L_2 (partial) approximations coincide, and are located in the barycenter of the set of L_∞ approximations, which form instead a whole interval. Such L_1/L_2 approximations leave the mass of $\{x\}$ unchanged, and re-assign the mass of $\{y\}$ (which is not in the chain $\{\{x\}, \{x, y\}\}$) to Θ. Such features are retained in the general case (Section 4).

The consonant approximation problem. Given a belief function b with basic probability assignment m_b, we call (metric) *consonant approximation of a belief function b induced by a distance function d* in $\mathcal{M}$ the b.f.(s) $co_d[m_b]$ which minimize(s) the distance $d(\vec{m}_b, \mathcal{CO})$ between the mass vector $\vec{m}_b$ representing m_b and the consonant simplicial complex

$$co_d[m_b] = \arg \min_{\vec{m}_{co} \in \mathcal{CO}} d(\vec{m}_b, \vec{m}_{co}), \tag{2}$$

under the condition that such minima exist.

Why use L_p norms. A close relation exists between consonant belief functions and L_p norms, in particular the L_∞ one. Consonant b.f.s are the counterparts of necessity measures in the theory of evidence, so that their plausibility functions are possibility measures. Possibility measures Pos, in turn, are inherently related to L_∞ as $Pos(A) = \max_{x \in A} Pos(x)$. It makes therefore sense to conjecture that a consonant transformation obtained by picking as distance function in the problem (2) one of the classical norms

$$\begin{aligned} \|\vec{m}_b - \vec{m}_{b'}\|_{L_1} &= \sum_{A \subseteq \Theta} |m_b(A) - m_{b'}(A)|, \\ \|\vec{m}_b - \vec{m}_{b'}\|_{L_2} &= \sqrt{\textstyle\sum_{A \subseteq \Theta} (m_b(A) - m_{b'}(A))^2}, \\ \|\vec{m}_b - \vec{m}_{b'}\|_{L_\infty} &= \max_{A \subseteq \Theta} \Big\{ |m_b(A) - m_{b'}(A)| \Big\} \end{aligned} \tag{3}$$

would be meaningful. In the probabilistic case, in the belief space $\mathcal{B}$ ($p[b] = \arg\min_{p \in \mathcal{P}} dist(b, p)$), the use of L_p norms leads indeed to quite interesting results. On one side, the L_2 approximation induces the so-called "orthogonal projection" of b onto $\mathcal{P}$ [3]. On the other, the set of L_1/L_∞ probabilistic approximations of b (in the belief space) coincides with the set of probabilities dominating b: $\{p : p(A) \geq b(A)\}$ (at least in the binary case).

Other norms. The L_p family of norms is important and useful also in classical probability theory. Clearly, however, a number of other norms can be introduced in the framework of belief functions and used to define consonant (or Bayesian) approximations. For instance, generalizations to belief functions of the classical Kullback-Leibler divergence of two probability distributions P, Q ($D_{KL}(P|Q) = \int_{-\infty}^{\infty} p(x) \log(\frac{p(x)}{q(x)}) dx$) or other measures based on information theory such as fidelity and entropy-based norms [18] can be studied. Many other similarity measures have indeed been proposed [9, 14, 19, 23]. The application of similarity measures more specific to belief functions or inspired by classical probability to the approximation problem

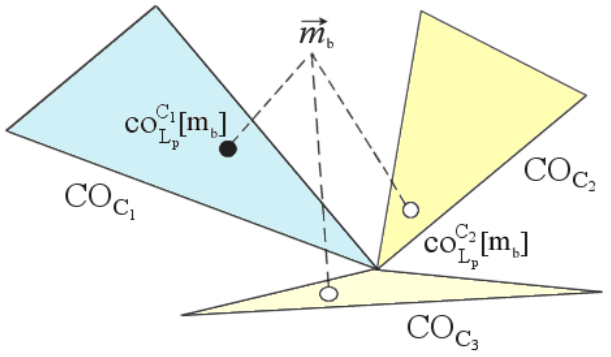


Figure 2: To minimize the distance of a point from a simplicial complex, we need to find all the partial solutions (4) on all the maximal simplices of the complex (empty circles), to later compare these partial solutions to select a global optimum (black circle).

is an enormous task, of which this paper can be seen as just a first step.

Distance of a point from a simplicial complex. As the consonant complex $\mathcal{CO}$ is a *collection* of simplices, solving the consonant approximation problem involves finding a number of partial solutions

$$co^{\mathcal{C}}_{L_p}[m_b] = \arg \min_{\vec{co} \in \mathcal{C}\mathcal{O}_{\mathcal{C}}} \|\vec{m}_b - \vec{co}\|_{L_p} \tag{4}$$

(see Figure 2), one for each maximal chain $\mathcal{C}$ of subsets of Θ. Then, the distance of $\vec{m}_b$ from all such partial solutions has to be assessed in order to select a global optimal approximation. Figure 1 shows the obtained (partial) L_p consonant approximations onto $\mathcal{CO}_x$ in the binary case. In such a toy example, $co_{L_1}[m_b] = co_{L_2}[m_b]$ coincide and are unique, lying on the barycenter of the set $co_{L_\infty}[m_b]$ of L_∞ approximations, which instead form a whole interval. Some of these features are retained in the general case, others are not. Note also that, in the binary case, consonant and consistent [8] approximations coincide, and there is no difference between belief and mass space [6] representation. In the rest of the paper we will explicitly compute the L_1, L_2, and L_∞ consonant approximations in the mass space and discuss the results.

3 Consonant approximation in $\mathcal{M}$

If we choose the $N-1$-dimensional version of the mass space (see Equation (1)), the mass vector associated with an arbitrary consonant b.f. *co* with maximal chain of focal elements $\mathcal{C}$ reads as $\vec{m}_{co} = \sum_{A \in \mathcal{C}} m_{co}(A) \vec{m}_A$, so that the difference vector is

$$\vec{m}_b - \vec{m}_{co} = \sum_{A \in \mathcal{C}} (m_b(A) - m_{co}(A)) \vec{m}_A + \sum_{A \notin \mathcal{C}} m_b(A) \vec{m}_A. \tag{5}$$

If we instead pick the $N-2$-dimensional version of the mass space, the mass vector associated with the same, arbitrary consonant b.f. *co* with maximal chain $\mathcal{C}$ reads as $\vec{m}_{co} = \sum_{A \in \mathcal{C}, A \neq \Theta} m_{co}(A) \vec{m}_A$, and the difference vector is

$$\sum_{A \in \mathcal{C}, A \neq \Theta} (m_b(A) - m_{co}(A)) \vec{m}_A + \sum_{A \notin \mathcal{C}} m_b(A) \vec{m}_A. \tag{6}$$

3.1 L_1 approximation

3.1.1 $\mathbb{R}^{N-1}$ representation

Consider first the $\mathbb{R}^{N-1}$ representation of mass vectors. Given the difference vector (5) its L_1 norm is $\|\vec{m}_b - \vec{m}_{co}\|_{L_1} = \sum_{A \in \mathcal{C}} |m_b(A) - m_{co}(A)| + \sum_{A \notin \mathcal{C}} m_b(A) = \sum_{A \in \mathcal{C}} |\beta(A)| + \sum_{A \notin \mathcal{C}} m_b(A)$, where $\beta(A) \doteq m_b(A) - m_{co}(A)$ and

$$\sum_{A \in \mathcal{C}} \beta(A) = \sum_{A \in \mathcal{C}} (m_b(A) - m_{co}(A)) = \sum_{A \in \mathcal{C}} m_b(A) - 1 \tag{7}$$

so that $\beta(\Theta) = \sum_{A \in \mathcal{C}} m_b(A) - 1 - \sum_{A \in \mathcal{C}, A \neq \Theta} \beta(A)$.
The above norm reads therefore as, as a function of the variables $\{\beta(A), A \in \mathcal{C}, A \neq \Theta\}$,

$$\begin{aligned} \|\vec{m}_b - \vec{m}_{co}\|_{L_1} = &\Big| \sum_{A \in \mathcal{C}} m_b(A) - 1 - \sum_{A \in \mathcal{C}, A \neq \Theta} \beta(A) \Big| \\ &+ \sum_{A \in \mathcal{C}, A \neq \Theta} |\beta(A)| + \sum_{A \notin \mathcal{C}} m_b(A). \end{aligned} \tag{8}$$

Partial approximation. This function has the form

$$\sum_i |x_i| + \Big| - \sum_i x_i - k \Big|, \quad k \geq 0 \tag{9}$$

which has an entire simplex of minima, namely: $x_i \leq 0 \ \forall i$, $\sum_i x_i \geq -k$ (see [6] for a similar optimization problem in the geometric conditioning context). The minima of the L_1 norm (8) are therefore given by the following system of constraints:

$$\begin{cases} \beta(A) \leq 0 & \forall A \in \mathcal{C}, A \neq \Theta, \\ \sum_{A \in \mathcal{C}, A \neq \Theta} \beta(A) \geq \sum_{A \in \mathcal{C}} m_b(A) - 1. \end{cases} \tag{10}$$

The solution in terms of the mass of the consonant approximation reads as:

$$\begin{cases} m_{co}(A) \geq m_b(A) & \forall A \in \mathcal{C}, A \neq \Theta, \\ \sum_{A \in \mathcal{C}, A \neq \Theta} (m_b(A) - m_{co}(A)) \geq \sum_{A \in \mathcal{C}} m_b(A) - 1 \end{cases} \tag{11}$$

where the last constraint reduces to

$$\begin{aligned} &\sum_{A \in \mathcal{C}, A \neq \Theta} (m_b(A) - m_{co}(A)) = \\ &= \sum_{A \in \mathcal{C}, A \neq \Theta} m_b(A) - \Big(1 - m_{co}(\Theta)\Big) \geq \sum_{A \in \mathcal{C}} m_b(A) - 1, \end{aligned}$$

i.e., $m_{co}(\Theta) \geq m_b(\Theta)$. Therefore the solution is $m_{co}(A) \geq m_b(A)\ \forall A \in \mathcal{C}$.

Vertices and barycenter of the partial approximation. The vertices of the set of approximations which are the solutions of (10) are given by the vectors $\{\vec{\beta}_A, A \in \mathcal{C}\}$ such that

$$\vec{\beta}_A(B) = \begin{cases} -\sum_{A \notin \mathcal{C}} m_b(A) & B = A, \\ 0 & B \neq A \end{cases}$$

when $A \neq \Theta$, while $\vec{\beta}_\Theta = \vec{0}$. In terms of masses the vertices of the set of partial L_1 approximations are the vectors $\{\vec{m}_A^{L_1}, A \in \mathcal{C}\}$ such that

$$\vec{m}_A^{L_1}(B) = \begin{cases} m_b(B) + \sum_{A \notin \mathcal{C}} m_b(A) & B = A, \\ m_b(B) & B \neq A \end{cases} \tag{12}$$

whose barycenter is $co_{\overline{L_1},N-1}[m_b](B) = m_b(B) + \frac{\sum_{A \notin \mathcal{C}} m_b(A)}{n}$.

Global approximation. To find the *global* L_1 approximation on the consonant complex, we need to find out which component is associated with the minimal L_1 distance. The partial approximations (11) onto $\mathcal{CO}^{\mathcal{C}}$ have L_1 distance from $\vec{m}_b$ given by:

$$\sum_{A \notin \mathcal{C}} m_b(A) = 1 - \sum_{A \in \mathcal{C}} m_b(A). \tag{13}$$

Therefore, the component of the consonant complex at minimal distance is that one associated with the chain that has maximal mass in the original b.f.

3.1.2 $\mathbb{R}^{N-2}$ representation

In the $\mathbb{R}^{N-2}$ representation of mass vectors, the L_1 norm of the difference vector (6) is

$$\|\vec{m}_b - \vec{m}_{co}\|_{L_1} = \sum_{A \in \mathcal{C}, A \neq \Theta} |m_b(A) - m_{co}(A)| + \sum_{A \notin \mathcal{C}} m_b(A)$$

which is obviously minimized by

$$m_{co}(A) = m_b(A) \quad \forall A \in \mathcal{C}, A \neq \Theta. \tag{14}$$

Again, to find the global L_1 approximation on the consonant complex in $\mathbb{R}^{N-2}$, we need to find the closest simplicial component. As the partial approximation (14) onto $\mathcal{CO}^{\mathcal{C}}$ has L_1 distance from $\vec{m}_b$ given as before by (13), we have the following.

Theorem 1. *Given a belief function $b : 2^\Theta \to [0, 1]$ with b.p.a. m_b, the global L_1 consonant approximations of b in the mass space $\mathcal{M}$ of dimension $\mathbb{R}^{N-1}$ is the set of partial approximations*

$$\begin{aligned} co^{\mathcal{C}^*}_{L_1,\mathcal{M},N-1}[m_b] &= \left\{ m_{co}(A) \geq m_b(A)\ \forall A \in \mathcal{C} \right\} \\ &= Cl(\vec{m}_A^{L_1}, A \in \mathcal{C}), \end{aligned}$$

with vertices given by Equation (12), associated with the maximal chain of focal elements which maximizes the total original mass of the chain

$$\mathcal{C}^* = \arg\max_{\mathcal{C}} \sum_{A \in \mathcal{C}} m_b(A).$$

Its global L_1 consonant approximations in the mass space $\mathcal{M}$ of dimension $\mathbb{R}^{N-2}$ is the (unique) partial approximation $co^{\mathcal{C}^*}_{L_1,\mathcal{M},N-2}[m_b]$ such that

$$\begin{cases} m_{co}(A) = m_b(A) & \forall A \in \mathcal{C}, A \neq \Theta, \\ m_{co}(\Theta) = m_b(\Theta) + 1 - \sum_{A \in \mathcal{C}} m_b(A) \end{cases}$$

associated with the same chain of focal elements.

Not only the two approximations are consistent in the sense that they have the same chain of focal elements, but the set of L_1 consonant approximations in $\mathbb{R}^{N-1}$ is convex and forms a polytope, one of whose vertices is indeed the L_1 approximation in $\mathbb{R}^{N-2}$.

3.2 L_2 approximation

In order to find the L_2 consonant approximation(s) it is convenient to recall that the minimal L_2 distance between a point and a vector space is attained by the point of the vector space such that the difference vector is orthogonal to all the generators $\vec{g}_i$ of the vector space:

$$\arg\min_{\vec{q} \in V} \|\vec{p} - \vec{q}\|_{L_2} = \hat{q} \in V : \langle \vec{p} - \hat{q}, \vec{g}_i \rangle = 0 \quad \forall i$$

whenever $\vec{p} \in \mathbb{R}^m$, $V = span(\vec{g}_i, i)$. Hence, instead of minimizing the L_2 norm of the difference vector $\|\vec{m}_b - \vec{m}_{co}\|_{L_2}$ we can just impose a condition of orthogonality between the difference vector itself $\vec{m}_b - \vec{m}_{co}$ and each component $\mathcal{CO}^{\mathcal{C}}$ of the consonant complex. In the two cases $\mathbb{R}^{N-1}$ and $\mathbb{R}^{N-2}$ we will therefore have two different difference vectors and two different orthogonality conditions. In the both cases we need to write:

$$\langle \vec{m}_b - \vec{m}_{co}, \vec{m}_A - \vec{m}_\Theta \rangle = 0 \quad \forall A \in \mathcal{C}, A \neq \Theta. \tag{15}$$

3.2.1 $\mathbb{R}^{N-1}$ representation

In the $N-1$ dimensional mass space, however, the vector $\vec{m}_A - \vec{m}_\Theta$ is such that $\vec{m}_A - \vec{m}_\Theta(B) = 1$ if $B = A$, $\vec{m}_A - \vec{m}_\Theta(B) = -1$ if $B = \Theta$, 0 otherwise. Hence, the orthogonality condition becomes

$$\beta(A) - \beta(\Theta) = 0 \quad \forall A \in \mathcal{C}, A \neq \Theta.$$

Partial approximation. By Equation (7) $\beta(\Theta) = \sum_{A \in \mathcal{C}} m_b(A) - 1 - \sum_{A \in \mathcal{C}, A \neq \Theta} \beta(A)$ and the orthogonality condition becomes

$$\left\{ 2\beta(A) + 1 - \sum_{B \in \mathcal{C}} m_b(B) + \sum_{B \in \mathcal{C}, B \neq A, \Theta} \beta(B) = 0 \right.$$

for all focal elements A in the maximal chain $\mathcal{C}$, $A \neq \Theta$. By substitution it can be proven that the solution is $\beta(A) = \frac{\sum_{B \in \mathcal{C}} m_b(B) - 1}{n}$. The mass of the partial L_2 consonant approximation is therefore, $\forall A \in \mathcal{C}$:

$$m_{co}(A) = m_b(A) + \frac{1 - \sum_{B \in \mathcal{C}} m_b(B)}{n}. \qquad (16)$$

Global approximation. To find the global approximation, we need to compute the L_2 distance of b from the closest such partial solution. We have:

$$\|\vec{m}_b - \vec{m}_{co}\|_{L_2}^2 = \sum_{A \subseteq \Theta} (m_b(A) - m_{co}(A))^2 = \frac{(\sum_{B \notin \mathcal{C}} m_b(B))^2}{n} + \sum_{A \notin \mathcal{C}} (m_b(A))^2,$$

which is minimized by the component $\mathcal{C}\mathcal{O}^{\mathcal{C}}$ that minimizes $\sum_{A \notin \mathcal{C}} (m_b(A))^2$.

3.2.2 $\mathbb{R}^{N-2}$ representation

In the $\mathbb{R}^{N-2}$ representation, as $\vec{m}_\Theta = \vec{0}$, the orthogonality condition reads as:

$$\langle \vec{m}_b - \vec{m}_{co}, \vec{m}_A \rangle = \beta(A) = 0 \quad \forall A \in \mathcal{C}, A \neq \Theta$$

so that the L_2 partial approximation is given by

$$\begin{cases} m_{co}(A) = m_b(A) & A \in \mathcal{C}, A \neq \Theta \\ m_{co}(\Theta) = m_b(\Theta) + \sum_{B \notin \mathcal{C}} m_b(B). \end{cases} \qquad (17)$$

The optimal distance is, in this case, $\|\vec{m}_b - \vec{m}_{co}\|_{L_2}^2 = \sum_{A \subseteq \Theta} (m_b(A) - m_{co}(A))^2 = \sum_{A \notin \mathcal{C}} (m_b(A))^2 + (\sum_{A \notin \mathcal{C}} m_b(A))^2$, which is once again minimized by the maximal chain $\mathcal{C}^* = \arg\min_{\mathcal{C}} \sum_{A \notin \mathcal{C}} (m_b(A))^2$.

Theorem 2. *Given a belief function $b : 2^\Theta \to [0, 1]$ with b.p.a. m_b, the global L_2 consonant approximations of b in the mass space $\mathcal{M}$ of dimension $\mathbb{R}^{N-1}$ is the set of partial approximations $co^{\mathcal{C}^*}_{L_2, \mathcal{M}, N-1}[m_b] =$*

$$= \left\{ m_{co}(A) = m_b(A) + \frac{1 - \sum_{B \in \mathcal{C}^*} m_b(B)}{n} \right\}$$

associated with the maximal chain of focal elements which minimizes the sum of square masses outside the chain: $\mathcal{C}^* = \arg\min_{\mathcal{C}} \sum_{A \notin \mathcal{C}} (m_b(A))^2$.
Its global L_2 consonant approximations in the mass space $\mathcal{M}$ of dimension $\mathbb{R}^{N-2}$ is the (unique) partial approximation $co^{\mathcal{C}^*}_{L_1, \mathcal{M}, N-2}[m_b] =$

$$= \left\{ \begin{array}{l} m_{co}(A) = m_b(A) \quad \forall A \in \mathcal{C}, A \neq \Theta, \\ m_{co}(\Theta) = m_b(\Theta) + 1 - \sum_{A \in \mathcal{C}} m_b(A) \end{array} \right\}$$

associated with the same chain of focal elements, and coincides with the global L_1 consonant approximation in the mass space $\mathcal{M}$ of dimension $\mathbb{R}^{N-2}$.

Indeed, in virtue of (17) and (14) all partial L_1 and L_2 consonant approximations coincide in the mass space of dimension $N - 2$.

3.3 L_∞ approximation

3.3.1 $\mathbb{R}^{N-1}$ representation

In the $N - 1$ representation, the L_∞ norm of the difference vector is

$$\|\vec{m}_b - \vec{m}_{co}\|_{L_\infty} = \max \left\{ \max_{A \in \mathcal{C}} |\beta(A)|, \max_{B \notin \mathcal{C}} m_b(B) \right\},$$

$\beta(\Theta) = \sum_{B \in \mathcal{C}} m_b(B) - 1 - \sum_{B \in \mathcal{C}, B \neq \Theta} \beta(B)$, so that

$$|\beta(\Theta)| = \Big| \sum_{B \notin \mathcal{C}} m_b(B) + \sum_{B \in \mathcal{C}, B \neq \Theta} \beta(B) \Big|$$

and the norm to minimize becomes

$$\|\vec{m}_b - \vec{m}_{co}\|_{L_\infty} = \max \Big\{ \max_{A \in \mathcal{C}, A \neq \Theta} |\beta(A)|, \Big| \sum_{B \notin \mathcal{C}} m_b(B) + \sum_{B \in \mathcal{C}, B \neq \Theta} \beta(B) \Big|, \max_{B \notin \mathcal{C}} m_b(B) \Big\}. \qquad (18)$$

This is a function of the form

$$\max \Big\{ |x_1|, |x_2|, |x_1 + x_2 + k_1|, k_2 \Big\} \qquad (19)$$

with $0 \leq k_2 \leq k_1 \leq 1$. If $|\mathcal{C}| = 2$, for instance, $x_1 = \beta(A_1)$, $x_2 = \beta(A_2)$, $k_1 = \sum_{B \notin \mathcal{C}} m_b(B)$ and $k_2 = \max_{B \notin \mathcal{C}} m_b(B)$. Such a function has two possible behaviors in terms of its minimal region in the plane x_1, x_2.

Case 1. If $k_1 \leq 3k_2$ its contour function has the form rendered in Figure 3. The set of minimal points is given by $x_i \geq -k_2$, $x_1 + x_2 \leq k_2 - k_1$. In the more

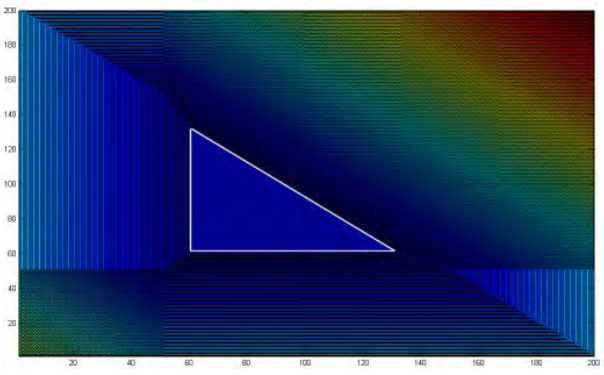


Figure 3: Contour function (level sets) and minimal points (white triangle) of a function of the form (19), when $k_1 \leq 3k_2$. In the example $k_2 = 0.4$ and $k_1 = 0.5$.

general case of an arbitrary number $m-1$ of variables $x_1, ..., x_{m-1}$ such that $x_i \geq -k_2$, $\sum_i x_i \leq k_2 - k_1$, the set of minimal points is a simplex with m vertices:

each vertex v^i is such that $v^i(j) = -k_2 \; \forall j \neq i$; $v^i(i) = -k_1 + (m-1)k_2$ (obviously $v^m = [-k_2, \cdots, -k_2]$). For the norm (18), in the first case

$$\max_{B \notin \mathcal{C}} m_b(B) \geq \frac{1}{n} \sum_{B \notin \mathcal{C}} m_b(B) \tag{20}$$

the set of partial L_∞ approximations is given by

$$\begin{cases} \beta(A) \geq -\max_{B \notin \mathcal{C}} m_b(B) \quad A \in \mathcal{C}, A \neq \Theta \\ \sum_{B \in \mathcal{C}, B \neq \Theta} \beta(B) \leq \max_{B \notin \mathcal{C}} m_b(B) - \sum_{B \notin \mathcal{C}} m_b(B) \end{cases}$$

This is a simplex $Cl(\vec{m}_{\bar{A}}^{L_\infty}, \bar{A} \in \mathcal{C})$ with vertices

$$\begin{cases} \beta_{\bar{A}}(A) = -\max_{B \notin \mathcal{C}} m_b(B) \quad A \in \mathcal{C}, A \neq \bar{A} \\ \beta_{\bar{A}}(\bar{A}) = -\sum_{B \notin \mathcal{C}} m_b(B) + (n-1) \max_{B \notin \mathcal{C}} m_b(B) \end{cases}$$

or, in terms of their basic probability assignments,

$$\begin{cases} \vec{m}_{\bar{A}}^{L_\infty}(A) = m_b(A) + \max_{B \notin \mathcal{C}} m_b(B) \quad A \in \mathcal{C}, A \neq \bar{A} \\ \vec{m}_{\bar{A}}^{L_\infty}(\bar{A}) = m_b(\bar{A}) + \sum_{B \notin \mathcal{C}} m_b(B) + \\ \qquad -(n-1) \max_{B \notin \mathcal{C}} m_b(B). \end{cases} \tag{21}$$

Note that such quantity is not guaranteed to be positive, as, for instance, when there exists a single subset B s.t. $m_b(B) \neq 0$ outside $\mathcal{C}$, $\vec{m}_{\bar{A}}^{L_\infty}(\bar{A})$ is negative unless $n \leq 2$. The barycenter of this simplex can be computed as follows:

$$m^{\overline{L_\infty}}(A) = \frac{\sum_{\bar{A} \in \mathcal{C}} \vec{m}_{\bar{A}}^{L_\infty}(A)}{n} = m_b(A) + \frac{\sum_{B \notin \mathcal{C}} m_b(B)}{n},$$

i.e., the L_2 partial approximation. The corresponding minimal L_∞ norm of the difference vector is, according to (18), equal to $\max_{B \notin \mathcal{C}} m_b(B)$.

Case 2. In the second case $k_1 > 3k_2$, or for us

$$\max_{B \notin \mathcal{C}} m_b(B) < \frac{1}{n} \sum_{B \notin \mathcal{C}} m_b(B), \tag{22}$$

the contour function of (19) is as in Figure 4. There is a single minimal point, located in $[-1/3k_1, -1/3k_1]$. For an arbitrary number $m-1$ of variables the minimal point is located in $[(-1/m)k_1, \cdots, (-1/m)k_1]'$, i.e., for system (18), $\beta(A) = -\frac{1}{n} \sum_{B \notin \mathcal{C}} m_b(B)$ for all $A \in \mathcal{C}, A \neq \Theta$ or, in terms of b.p.a.s,

$$m_{co_{L_\infty}[m_b]}(A) = m_b(A) + \frac{1}{n} \sum_{B \notin \mathcal{C}} m_b(B) \quad \forall A \in \mathcal{C}.$$

The mass of Θ is obtained by normalization.
The corresponding minimal L_∞ norm of the difference vector is $\frac{1}{n} \sum_{B \notin \mathcal{C}} m_b(B)$.

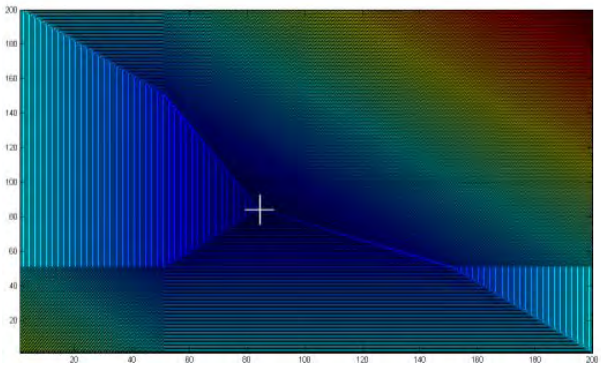


Figure 4: Contour function (level sets) and minimal point (white cross) of a function of the form (19), when $k_1 \geq 3k_2$. In the example $k_2 = 0.1$ and $k_1 = 0.5$.

3.3.2 $\mathbb{R}^{N-2}$ representation

In $\mathbb{R}^{N-2}$ the L_∞ norm of the difference vector is

$$\begin{aligned} \|\vec{m}_b - \vec{m}_{co}\|_{L_\infty} &= \max_{\emptyset \subsetneq A \subsetneq \Theta} |m_b(A) - m_{co}(A)| \\ &= \max \left\{ \max_{A \in \mathcal{C}, A \neq \Theta} |\beta(A)|, \max_{B \notin \mathcal{C}} m_b(B) \right\} \end{aligned} \tag{23}$$

which is minimized by

$$|\beta(A)| \leq \max_{B \notin \mathcal{C}} m_b(B) \quad \forall A \in \mathcal{C}, A \neq \Theta \tag{24}$$

i.e., in the original mass coordinates,

$$\begin{aligned} m_b(A) - \max_{B \notin \mathcal{C}} m_b(B) \leq m_{co}(A) \leq \\ \leq m_b(A) + \max_{B \notin \mathcal{C}} m_b(B) \quad \forall A \in \mathcal{C}, A \neq \Theta. \end{aligned} \tag{25}$$

According to (23) the corresponding minimal L_∞ norm is equal to $\max_{B \notin \mathcal{C}} m_b(B)$.
Clearly, the vertices of the set (24) are all the vectors of β variables such that $\beta(A) = +/- \max_{B \notin \mathcal{C}} m_b(B)$ for all $A \in \mathcal{C}$, $A \neq \Theta$. Its barycenter is clearly given by $\beta(A) = 0$ for all $A \in \mathcal{C}$, $A \neq \Theta$, i.e.:

$$m_{co}(B) = \begin{cases} m_b(B) & B \in \mathcal{C}, B \neq \Theta \\ m_b(B) + \sum_{B \notin \mathcal{C}} m_b(B) & B = \Theta. \end{cases} \tag{26}$$

Summarizing:

Theorem 3. *Given a belief function $b : 2^\Theta \to [0,1]$ with b.p.a. m_b, the partial L_∞ consonant approximations of b in the mass space $\mathcal{M}$ of dimension $\mathbb{R}^{N-1}$ can form either a simplex*

$$co^{\mathcal{C}^*}_{L_\infty, \mathcal{M}, N-1}[m_b] = Cl(\vec{m}_{\bar{A}}^{L_\infty}, \bar{A} \in \mathcal{C})$$

with vertices (21) when $\max_{B \notin \mathcal{C}} m_b(B) \geq \frac{1}{n} \sum_{B \notin \mathcal{C}} m_b(B)$, or a reduce to a single belief function when the opposite is true, the barycenter of the above simplex, located on the partial L_2 approximation (16). In both cases, the global L_∞ consonant

approximation is associated with the maximal chain of focal elements:

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \max_{B \notin \mathcal{C}} m_b(B). \quad (27)$$

The partial L_∞ consonant approximations of b in the mass space $\mathcal{M}$ of dimension $\mathbb{R}^{N-2}$ form the set $co^{\mathcal{C}^*}_{L_\infty,\mathcal{M},N-2}[m_b]$ given by Equation (25). Its barycenter reassigns all the mass outside the chain to Θ, leaving the masses of the other elements untouched. The related global approximations of b are associated with the same optimal chain (27).

4 Semantics

Let us interpret the results we obtained in terms of basic probability assignments of the various consonant approximations, and compare those results with the outer consonant approximations [11] whose geometry has been described in [7].

Summary of approximations in $\mathcal{M}$. We can summarize all the results obtained here in the following tables. In the $\mathbb{R}^{N-1}$ mass representation the partial L_p approximations are:

$$\begin{array}{rcl} co^{\mathcal{C}}_{L_1,N-1}[m_b] & = & Cl(\vec{m}^{L_1}_A, A \in \mathcal{C}) \\ & : & m_{co}(A) \geq m_b(A)\ \forall A \in \mathcal{C}; \\ co^{\mathcal{C}}_{\overline{L_1},N-1}[m_b] & = & co^{\mathcal{C}}_{L_2,N-1}[b] \\ & : & m_{co}(A) = m_b(A) + \frac{\sum_{B \notin \mathcal{C}} m_b(B)}{n}. \end{array} \quad (28)$$

Concerning the L_∞ approximation, if (20) holds

$$\begin{array}{rcl} co^{\mathcal{C}}_{L_\infty,N-1}[m_b] & = & Cl(\vec{m}^{L_\infty}_{\bar{A}}, \bar{A} \in \mathcal{C}); \\ co^{\mathcal{C}}_{\overline{L_\infty},N-1}[m_b] & = & co^{\mathcal{C}}_{L_2,N-1}[m_b], \end{array}$$

while if (22) holds: $co^{\mathcal{C}}_{L_\infty,N-1}[m_b] = co^{\mathcal{C}}_{L_2,N-1}[m_b]$.
We can observe the following facts:

1. the set of L_1 partial approximation is the set of inner consonant approximations of b according to the order relation: $b \geq b'$ iff $m_b(A) \geq m_{b'}(A)$;
2. this set is a simplex, whose vertices are obtained by re-assigning all the mass outside the desired chain to a single focal element of the chain itself (see (12));
3. its barycenter coincides with the L_2 partial approximation;
4. such approximation redistributes the mass of focal elements outside the chain on an equal basis to all the elements of the chain;
5. when the partial L_∞ approximation is unique, it coincides with the L_2 approximation and the barycenter of the L_1 approximations;
6. when it is not unique, it is a simplex whose vertices assign to each element of the chain but one the maximal mass outside the chain, with barycenter still in the L_2 approximation.

In particular, points 2. and 4. (and 5.) remind us of the behavior of geometric conditional belief functions in the mass space [6]. There,

Proposition 1. *Given a belief function $b : 2^\Theta \to [0,1]$ and an arbitrary non-empty focal element $\emptyset \subsetneq A \subseteq \Theta$, the unique L_2 conditional belief functions $b_{L_2,\mathcal{M}}(.|A)$ with respect to A in $\mathcal{M}$ is the b.f. whose b.p.a. redistributes the mass $1 - b(A)$ to each focal element $B \subseteq A$ in an equal way.*
The set of L_1 conditional belief functions $b_{L_1,\mathcal{M}}(.|A)$ with respect to A in $\mathcal{M}$ is a simplex whose vertices re-assign the mass $1 - b(A)$ of focal elements not in the conditioning event A to a specific subset of A.

It is tempting to speculate that this is a consistent behavior of L_1 and L_2 minimization in the $\mathbb{R}^{N-1}$ representation of the mass space.
In the $\mathbb{R}^{N-2}$ mass representation the corresponding partial L_p approximations are:

$$\begin{array}{rcl} co^{\mathcal{C}}_{L_\infty,N-2}[b] & : & |m_{co}(A) - m_b(A)| \leq \max_{B \notin \mathcal{C}} m_b(B) \\ & & \forall A \in \mathcal{C}, A \neq \Theta; \\ co^{\mathcal{C}}_{\overline{L_\infty},N-2}[b] & = & co^{\mathcal{C}}_{L_1,N-2}[b] = co^{\mathcal{C}}_{L_2,N-2}[b] \\ & : & \begin{cases} m_{co}(A) = m_b(A),\ A \in \mathcal{C}, \neq \Theta \\ m_{co}(\Theta) = m_b(\Theta) + \sum_{B \notin \mathcal{C}} m_b(B). \end{cases} \end{array} \quad (29)$$

We can notice a number of facts here too:

1. the L_∞ (partial) approximation is not unique, and it falls entirely inside the simplex of admissible consonant b.f. only if each focal element in the desired chain has mass greater then all focal elements outside the chain: $m_b(A) \leq \max_{B \notin \mathcal{C}} m_b(B)$;
2. it forms a polytope in the mass space $\mathcal{M}$, whose size is determined by the largest mass outside the desired maximal chain;
3. the L_1 and L_2 partial approximations are uniquely determined, and coincide with the barycenter of the set of L_∞ partial approximations;
4. their semantic is straightforward: all the mass outside the chain is re-assigned to Θ, increasing the overall uncertainty of the belief state.

Clearly, approximations in the mass space do not take into account the contributions of focal elements outside the chain to the plausibility of elements of the chain. A similar phenomenon has been observed in the case of geometric conditioning [6].

Relation with outer consonant approximations. Let us recall the main results on the geometry of outer consonant approximations [7].

Proposition 2. *For each simplicial component $\mathcal{CO}_\mathcal{C}$ of the consonant space associated with any maximal chain of focal elements $\mathcal{C} = \{A_1 \subset \cdots \subset A_n, |A_i| = i\}$ the set of outer consonant approximation of any b.f.*

b is the convex closure $O_{\mathcal{C}}[b] = Cl(o^{\vec{B}}[b], \forall \vec{B})$ *of the co.b.f.s with basic probabilities*

$$m_{o^{\vec{B}}[b]}(B_i) = \sum_{A \subseteq \Theta : \vec{B}(A) = A_i} m_b(A), \qquad (30)$$

each associated with an "assignment function" $\vec{B} : 2^\Theta \to \mathcal{C}$, $A \mapsto \vec{B}(A) \supseteq A$ *which maps each event A to one of the elements of the chain containing it.*

The points (30) are not guaranteed to be proper vertices of the polytope $O_{\mathcal{C}}[b]$, as some of them can be obtained as a convex combination of the others. The outer approximation produced by the permutation ρ of singletons which generates the desired chain $A_i = \{x_{\rho(1)}, \cdots, x_{\rho(i)}\}$, $i = 1, ..., n$, i.e.

$$m_{co^\rho}(A_i) = \sum_{B \subseteq A_i, B \not\subset A_{i-1}} m_b(B), \qquad (31)$$

is an actual vertex of $O_{\mathcal{C}}[b]$, and corresponds to the *maximal* outer consonant approximation with maximal chain $\mathcal{C}$.
Indeed, by Equation (11), the partial L_1 approximations in $\mathbb{R}^{N-1}$ are such that $m_{co}(A) \geq m_b(A)$ for all $A \in \mathcal{C}$: *they are the opposite of outer consonant approximations, using the natural order relation between basic probabilities* (rather than belief values).
It can be seen in Figure 1 that, in the binary case, such maximal outer approximation coincides with the (partial) $L_1 = L_2 = \overline{L}_\infty$ approximation in the $N-2$ representation. It looks unclear what the relationship should be in the general case.

Comparison on a ternary example. It can therefore be useful to compare the different approximations in the toy case of a ternary frame, $\Theta = \{x, y, z\}$, to look for insights. Let us assume that we want the consonant approximation to have maximal chain $\mathcal{C} = \{\{x\}, \{x, y\}, \Theta\}$.
Figure 5 illustrates the different partial consonant approximations in the simplex of consonant belief functions with focal element $\{\{x\}, \{x, y\}, \Theta\}$, for a belief function with masses

$$m_b(x) = 0.2, m_b(y) = 0.3, m_b(x, z) = 0.5 \qquad (32)$$

We notice that the different simplices of L_p consonant approximations are distinct, with the $L_{1,N-1}$ one (red simplex) falling entire in the consonant simplex $Cl(\vec{m}_x, \vec{m}_{x,y}, \vec{m}_\Theta)$, while most of $L_{\infty,N-2}$ (green quadrangle) does not. It is interesting to note, though, they are not unrelated to each other: indeed, the $L_1/L_2/\overline{L_\infty}$ consonant approximation in $\mathbb{R}^{N-2}$ (green little square) is a vertex of the simplex of L_1 approximation in $N-1$.
Even though the case for the unique $L_{1,N-2}/L_{2,N-2}/\overline{L_{\infty,N-2}}$ and $\overline{L_{1,N-1}}$ approximations seems compelling, it will be worth exploring in the near future the behavior of the intersection of the set of approximations not entirely admissible with the consonant complex.

According to the formulae at page 8 of [5], the set of outer consonant approximations of (32) with chain $\{\{x\}, \{x, y\}, \Theta\}$ is the convex closure of the points:

$$\begin{aligned}
\vec{m}_{B_1,B_2} &= [m_b(x), m_b(y), 1 - m_b(x) - m_b(y)]', \\
\vec{m}_{B_3,B_4} &= [m_b(x), 0, 1 - m_b(x)]', \\
\vec{m}_{B_5,B_6} &= [0, m_b(x) + m_b(y), 1 - m_b(x) - m_b(y)]', \\
\vec{m}_{B_7,B_8} &= [0, m_b(x), 1 - m_b(x)]', \\
\vec{m}_{B_9,B_{10}} &= [0, m_b(y), 1 - m_b(y)]', \\
\vec{m}_{B_{11},B_{12}} &= [0, 0, 1]',
\end{aligned} \qquad (33)$$

These points are plotted as light blue squares in Figure 5. We can notice many interesting things.

1. the set $O^{\mathcal{C}}[b]$ of outer consonant approximations with chain $\mathcal{C}$ is a subset of (the admissible part of) the set of $L_{\infty,N-2}$ partial approximations; actually, the barycenter of the latter is a vertex of $O^{\mathcal{C}}[b]$;
2. on the contrary, outer approximations and $L_{1,N-1}$ approximations are mutually exclusive, as it can be inferred by Equation (11);
3. the maximal outer approximation co^ρ lies on the border between the two, where $m_{co}(x, y) = m_b(x, y)$.

Several other intriguing facts can be noticed there: they surely deserve further analysis.

5 Conclusions

In this paper we computed all the consonant approximations of a belief function induced by minimizing its L_p distances to the consonant complex, in the mass space of basic probability vectors. Interpretations for such approximations are rather natural in terms of mass redistribution. We compared them with each other and related them with classical outer consonant approximations, with the help of an example.
The nature of L_p-induced consonant approximations in the belief space remains an open problem, as is a comprehensive analysis of consonant and consistent approximations induced by distance minimization.

References

[1] P. Baroni. Extending consonant approximations to capacities. *Proceedings of IPMU*, 1127–1134, 2004.

[2] L. Caro and A. B. Nadjar. Generalization of the Dempster-Shafer theory: a fuzzy-valued measure, *IEEE Transactions on Fuzzy Systems*, 7, 255–270, 1999.

[3] F. Cuzzolin. Two new Bayesian approximations of belief functions based on convex geometry. *IEEE Transactions on SMC - Part B*, 37:4, 993–1008, 2007.

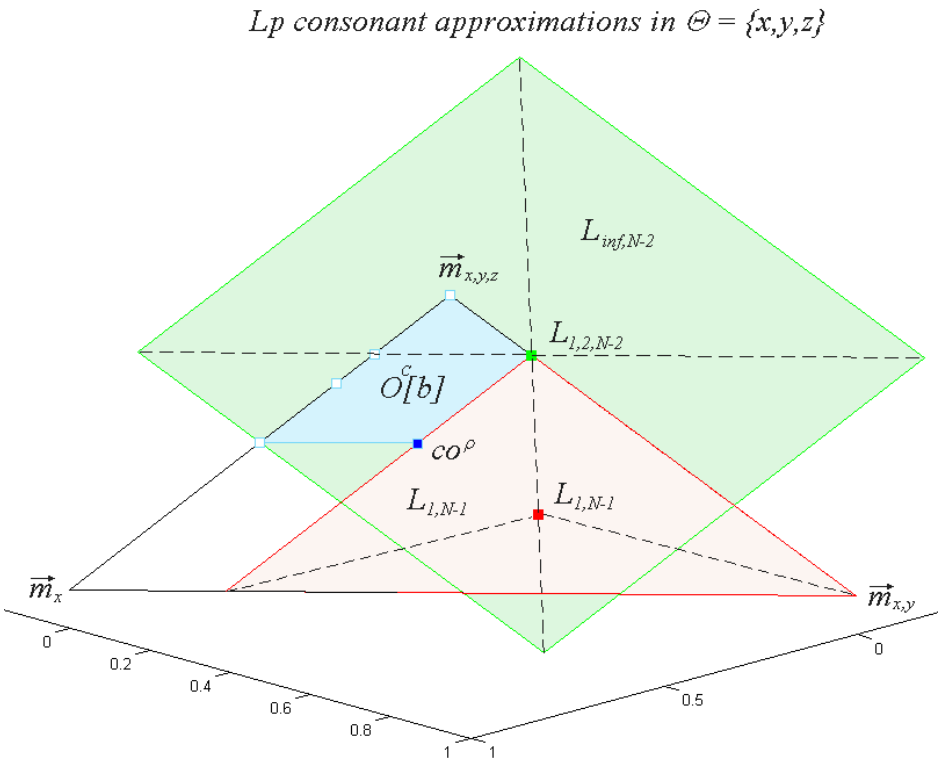


Figure 5: The simplex (solid black triangle) of consonant belief functions with maximal chain $\{\{x\}, \{x, y\}, \Theta\}$, and the L_p partial consonant approximations of the belief function with basic probabilities (32) on $\Theta = \{x, y, z\}$. The related set of outer consonant approximations (33) is also shown.

[4] F. Cuzzolin. A geometric approach to the theory of evidence. *IEEE Transactions on SMC - Part C*, 38:4, 522–534, 2008.

[5] F. Cuzzolin. Complexes of outer consonant approximations. *Proceedings of ECSQARU'09*, Verona, Italy, 2009.

[6] F. Cuzzolin. Geometric conditioning of belief functions. *Proceedings of BELIEF'10*, Brest, France, 2010.

[7] F. Cuzzolin. The geometry of consonant belief functions: simplicial complexes of necessity measures. *Fuzzy Sets and Systems*, 161:10, 1459–1479, 2010.

[8] F. Cuzzolin. Consistent approximation of belief functions. *Proceedings of ISIPTA'09*, Durham, UK, June 2009.

[9] J. Diaz, M. Rifqi and B. Bouchon-Meunier. A similarity measure between basic belief assignments. *Proceedings of FUSION'06*, 2006.

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

[11] D. Dubois and H. Prade. Consonant approximations of belief functions. *International Journal of Approximate Reasoning*, 4, 419–449, 1990.

[12] B. A. Dubrovin, S. P. Novikov, and A. T. Fomenko. *Sovremennaja geometrija. metody i prilozenija*. Nauka, Moscow, 1986.

[13] S. Heilpern. Representation and application of fuzzy numbers. *Fuzzy Sets and Systems*, 91, 259–268, 1997.

[14] W. Jiang, A. Zhang and Q. Yang. A new method to determine evidence discounting coefficient. *Lecture Notes in Computer Science*, 5226/2008, 882–887, 2008.

[15] C. Joslyn. Possibilistic normalization of inconsistent random intervals. *Advances in Systems Science and Applications*, 44–51, 1997.

[16] C. Joslyn and G. Klir. Minimal information loss possibilistic approximations of random sets, 1992.

[17] A.-L. Jousselme and P. Maupin. On some properties of distances in evidence theory. *Proceedings of BELIEF'10*, Brest, France, 2010.

[18] A.-L. Jousselme and P. Maupin. Distances in evidence theory: Comprehensive survey and generalizations. *International Journal of Approximate Reasoning*, 2011 (under review).

[19] V. Khatibi and G. A. Montazer. A new evidential distance measure based on belief intervals. *Scientia Iranica - Transactions D: Computer Science and Engineering and Electrical Engineering*, 17:2, 119–132, 2010.

[20] G. J. Klir, W. Zhenyuan and D. Harmanec. Constructing fuzzy measures in expert systems. *Fuzzy Sets and Systems*, 92, 251–264, 1997.

[21] C. Roemer and A. Kandel. Applicability analysis of fuzzy inference by means of generalized Dempster-Shafer theory. *IEEE Transactions on Fuzzy Systems*, 3:4, 448–453, 1995.

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

[23] C. Shi, Y. Cheng, Q. Pan and Y. Lu. A new method to determine evidence distance. *Proceedings of the 2010 International Conference on Computational Intelligence and Software Engineering (CiSE)*, 1–4, 2010.

[24] Ph. Smets. The transferable belief model and possibility theory. *Proceedings of NAFIPS-90*, 215–218, 1990.

[25] R. R. Yager. Class of fuzzy measures generated from a Dempster-Shafer belief structure. *International Journal of Intelligent Systems*, 14, 1239–1247, 1999.

Non-conflicting and Conflicting Parts of Belief Functions

Milan Daniel
Institute of Computer Science, Academy of Sciences of the Czech Republic
milan.daniel@cs.cas.cz

Abstract

Non-conflicting and conflicting parts of belief functions are introduced in this study. The unique decomposition of a belief function defined on a two-element frame of discernment to non-conflicting and indecisive conflicting belief function is presented. Several basic statements about algebra of belief functions on a general finite frame of discernment are introduced and unique non-conflicting part of a BF on an n-element frame of discernment is presented here.

Keywords. belief function, Dempster-Shafer theory, Dempster's semigroup, conflict between belief functions, uncertainty, non-conflicting part of belief function, conflicting part of belief function.

1 Introduction

Belief functions are one of the widely used formalisms for uncertainty representation and processing that enables representation of incomplete and uncertain knowledge, belief updating, and combination of evidence. They were originally introduced as a principal notion of the Dempster-Shafer Theory or the Mathematical Theory of Evidence [17].

When combining belief functions (BFs) by the conjunctive rules of combination, conflicts often appear, which are assigned to $\emptyset$ by un-normalized conjunctive rule $\textcircled{\cap}$ or normalized by Dempster's rule of combination $\oplus$. Combination of conflicting BFs and interpretation of conflicts is often questionable in real applications, thus a series of alternative combination rules was suggested and a series of papers on conflicting belief functions was published, e.g. [2, 5, 16, 19].

In [9], new ideas concerning interpretation, definition, and measurement of conflicts of BFs were introduced. We presented three new approaches to interpretation and computation of conflicts: combinational conflict, plausibility conflict, and comparative conflict. Differences were made between conflicts between BFs and internal conflicts of single BF; a conflict between BFs was distinguished from the difference between BFs.

When analyzing mathematical properties of the three approaches to conflicts of BFs in [10], there appears a possibility of expression of a BF Bel as Dempster's sum of non-conflicting BF Bel_0 with the same plausibility decisional support as the original BF Bel has and of indecisive BF Bel_S which does not prefer any of the elements of frame of discernment. The presented contribution analyses existence and uniqueness of such BFs Bel_0 and Bel_S.

The study starts with belief functions and algebraic preliminaries in Section 2. The situation on 2-element frame (Section 3) is followed by a study of a/the case of general finite frames of discernment (Section 4). Some comments on alternative rules of belief combination are presented in Section 5.

2 Preliminaries

2.1 General Primer on Belief Functions

We assume classic definitions of basic notions from theory of *belief functions* (BFs) [17] on finite frames of discernment $\Omega_n = \{\omega_1, \omega_2, ..., \omega_n\}$, see also [4–9]; for illustration or simplicity, we often use 2- or 3-element frames Ω_2 and Ω_3. A *basic belief assignment (bba)* is a mapping $m : \mathcal{P}(\Omega) \longrightarrow [0,1]$ such that $\sum_{A\subseteq\Omega} m(A) = 1$; the values of the bba are called *basic belief masses (bbm)*. $m(\emptyset) = 0$ is usually assumed, then we speak about *normalized bba*. A *belief function (BF)* is a mapping $Bel : \mathcal{P}(\Omega) \longrightarrow [0,1]$, $Bel(A) = \sum_{\emptyset\neq X\subseteq A} m(X)$. A *plausibility function* $Pl(A) = \sum_{\emptyset\neq A\cap X} m(X)$. There is a unique correspondence among m and corresponding Bel and Pl thus we often speak about m as about belief function.

A *focal element* is a subset X of the frame of discernment, such that $m(X) > 0$. If all the focal elements are *singletons* (i.e. one-element subsets of Ω), then we speak about a *Bayesian belief function* (BBF), it

is a probability distribution on Ω in fact. If all the focal elements are either singletons or whole Ω (i.e. $|X| = 1$ or $|X| = |\Omega|$), then we speak about a *quasi-Bayesian belief function* (qBBF), it is something like 'un-normalized probability distribution'. If all focal elements are nested, we speak about *consonant belief function.*

Dempster's (conjunctive) rule of combination $\oplus$ is given as $(m_1 \oplus m_2)(A) = \sum_{X\cap Y=A} Km_1(X)m_2(Y)$ for $A \neq \emptyset$, where $K = \frac{1}{1-\kappa}$, $\kappa = \sum_{X\cap Y=\emptyset} m_1(X)m_2(Y)$, and $(m_1 \oplus m_2)(\emptyset) = 0$, see [17]; putting $K = 1$ and $(m_1 \oplus m_2)(\emptyset) = \kappa$ we obtain the *un-normalized conjunctive rule of combination* $\textcircled{\cap}$, see e. g. [18]. *The disjunctive rule of combination* is given by the formula $(m_1 \textcircled{\cup} m_2)(A) = \sum_{X\cup Y=A} m_1(X)m_2(Y)$, see [12].

Yager's rule of combination $\textcircled{Y}$, see [21], is given as $(m_1 \textcircled{Y} m_2)(\emptyset) = 0$, $(m_1 \textcircled{Y} m_2)(A) = \sum_{X,Y\subseteq\Theta,\, X\cap Y=A} m_1(X)m_2(Y)$ for $\emptyset \neq A \subset \Theta$, and $(m_1 \textcircled{Y} m_2)(\Theta) = m_1(\Theta)m_2(\Theta) + \sum_{X,Y\subseteq\Theta,\, X\cap Y=\emptyset} m_1(X)m_2(Y)$;

Dubois-Prade's rule of combination $\textcircled{D}$ is given as $(m_1 \textcircled{D} m_2)(A) = \sum_{X,Y\subseteq\Theta,\, X\cap Y=A} m_1(X)m_2(Y) + \sum_{X,Y\subseteq\Theta,\, X\cap Y=\emptyset, X\cup Y=A} m_1(X)m_2(Y)$ for $\emptyset \neq A \subseteq \Theta$, and $(m_1 \textcircled{D} m_2)(\emptyset) = 0$, see [11].

We say that BF *Bel* is *non-conflicting* when conjunctive combination of *Bel* with itself does not produce any conflicting belief masses (when $(Bel \textcircled{\cap} Bel)(\emptyset) = 0$, i.e., $Bel \textcircled{\cap} Bel = Bel \oplus Bel$), i.e. whenever $Pl(\omega_i) = 1$ for some $\omega \in \Omega_n$. Otherwise, BF is *conflicting*, i.e., it contains some internal conflict [9].

Let us recall U_n the *uniform Bayesian belief function*[1] [9], i.e., the uniform probability distribution on Ω_n, and *normalized plausibility of singletons*[2] of *Bel*: the BBF (probability distribution) $Pl\_P(Bel)$ such, that $(Pl\_P(Bel))(\omega_i) = \frac{Pl(\{\omega_i\})}{\sum_{\omega\in\Omega} Pl(\{\omega\})}$ [3, 7].

Let us define an *indecisive (indifferent) BF* as a BF, which does not prefer any $\omega_i \in \Omega_n$, i.e., BF which gives no decisional support for any ω_i, i.e., BF such that $h(Bel) = Bel \oplus U_n = U_n$, i.e., $Pl(\{\omega_i\}) = const.$, i.e., $(Pl\_P(Bel))(\{\omega_i\}) = \frac{1}{n}$.

2.2 Belief Functions on 2-Element Frame of Discernment; Dempster's Semigroup

Let us suppose, that the reader is slightly familiar with basic algebraic notions like *a semigroup* (an algebraic structure with an associative binary operation), *a group* (a structure with an associative binary operation, with a unary operation of inverse, and with a neutral element), *a neutral element* n $(n * x = x)$, *an absorbing element* a $(a * x = a)$, *a homomorphism* f $(f(x * y) = f(x) * f(y))$, etc. (Otherwise, see e.g., [4, 6, 14, 15].)

We assume $\Omega_2 = \{\omega_1, \omega_2\}$, in this subsection. There are only three possible focal elements $\{\omega_1\}, \{\omega_2\}, \{\omega_1, \omega_2\}$ and any normalized *basic belief assignment (bba)* m is defined by a pair $(a, b) = (m(\{\omega_1\}), m(\{\omega_2\}))$ as $m(\{\omega_1, \omega_2\}) = 1-a-b$; this is called *Dempster's pair* or simply *d-pair* in [4, 6, 14, 15] (it is a pair of reals such that $0 \leq a, b \leq 1, a + b \leq 1$).

Extremal d-pairs are the pairs corresponding to BFs for which either $m(\{\omega_1\}) = 1$ or $m(\{\omega_2\}) = 1$, i.e., $(1, 0)$ and $(0, 1)$. The set of all non-extremal d-pairs is denoted as D_0; the set of all non-extremal *Bayesian d-pairs* (i.e. d-pairs corresponding to Bayesian BFs, where $a + b = 1$) is denoted as G; the set of d-pairs such that $a = b$ is denoted as S (set of indecisive[3] d-pairs), the set where $b = 0$ as S_1, and analogically, the set where $a = 0$ as S_2 (simple support BFs). Vacuous BF is denoted as $0 = (0, 0)$ and there is a special BF (d-pair) $0' = (\frac{1}{2}, \frac{1}{2})$, see Figure 1.

The *(conjunctive) Dempster's semigroup* $\mathbf{D}_0 = (D_0, \oplus, 0, 0')$ is the set D_0 endowed with the binary operation $\oplus$ (i.e. with the Dempster's rule) and two distinguished elements 0 and $0'$. Dempster's rule can be expressed by the formula $(a, b) \oplus (c, d) = (1 - \frac{(1-a)(1-c)}{1-(ad+bc)},\ 1 - \frac{(1-b)(1-d)}{1-(ad+bc)})$ for d-pairs [14]. In D_0 it is defined further: $-(a, b) = (b, a)$, $h(a, b) = (a, b) \oplus 0' = (\frac{1-b}{2-a-b},\ \frac{1-a}{2-a-b})$, $h_1(a, b) = \frac{1-b}{2-a-b}$, $f(a, b) = (a, b) \oplus (b, a) = (\frac{a+b-a^2-b^2-ab}{1-a^2-b^2},\ \frac{a+b-a^2-b^2-ab}{1-a^2-b^2})$; $(a, b) \leq (c, d)$ iff $[h_1(a, b) < h_1(c, d)$ or $h_1(a, b) = h_1(c, d)$ and $a \leq c]$ [4].

The principal properties of $\mathbf{D}_0$ are summarized by the following theorem:

Theorem 1 *(i) The Dempster's semigroup* $\mathbf{D}_0$ *with the relation* $\leq$ *is an ordered commutative (Abelian) semigroup with the neutral element* 0*;* $0'$ *is the only non-zero idempotent of* $\mathbf{D}_0$*.*
(ii) $\mathbf{G} = (G, \oplus, -, 0', \leq)$ *is an ordered Abelian group, isomorphic to the group of reals with the usual ordering. Let us denote its negative and positive cones as* $G^{\leq 0'}$ *and* $G^{\geq 0'}$*.*
(iii) The sets S, S_1, S_2 *with the operation* $\oplus$ *and the ordering* $\leq$ *form ordered commutative semigroups with neutral element* 0*; they are all isomorphic to the*

[1] U_n which is idempotent w.r.t. Dempster's rule $\oplus$, and moreover neutral on the set of all BBFs, is denoted as ${}_nD0'$ in [7], $0'$ comes from studies by Hájek & Valdés.

[2] Plausibility of singletons is called *contour function* by Shafer in [17], thus $Pl\_P(Bel)$ is a normalization of contour function in fact.

[3] BFs (a, a) from S are called *indifferent* BFs by Haenni [13].

[4] Note, that $h(a, b)$ is an abbreviation for $h((a, b))$, similarly for $h_1(a, b)$ and $f(a, b)$.

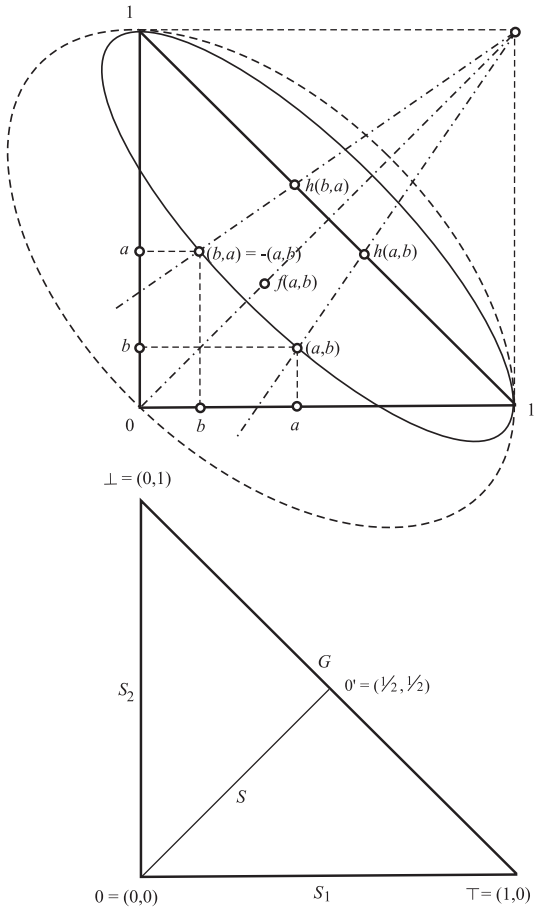


Figure 1: Dempster's semigroup D_0. Homomorphism h is in this representation a projection to group G along the straight lines running through the point $(1,1)$. All the Dempster's pairs lying on the same ellipse are mapped by homomorphism f to the same d-pair in semigroup S.

positive cone of the group of reals.
(iv) h is ordered homomorphism: $(D_0, \oplus, -, 0, 0', \leq) \longrightarrow (G, \oplus, -, 0', \leq)$*;* $h(Bel) = Bel \oplus 0' = Pl\_P(Bel)$*, i.e., the normalized plausibility probabilistic transformation.*
(v) f is homomorphism: $(D_0, \oplus, -, 0, 0') \longrightarrow (S, \oplus, -, 0)$*; (but, not an ordered one).*

For proofs see [14, 15, 20]. Let us denote $h^{-1}(a) = \{x \,|\, h(x) = a\}$ and similarly $f^{-1}(a) = \{x \,|\, f(x) = a\}$. Using the theorem, see (iv) and (v), we can express $\oplus$ as:

$$(a \oplus b) = h^{-1}(h(a) \oplus h(b)) \cap f^{-1}(f(a) \oplus f(b)).$$

Let us denote $D_0^{\geq 0} = \{(a,b) \in D_0 \,|\, (a,b) \geq 0\}$ and analogically $D_0^{\leq 0'} = \{(a,b) \leq 0'\}$.

2.3 BFs on n-Element Frames of Discernment

Analogically to the case of Ω_2, we can represent a BF on any n-element frame of discernment Ω_n by an enumeration of its m values (bbms), i.e., by a $(2^n - 2)$-tuple $(a_1, a_2, ..., a_{2^n-2})$, or as a $(2^n - 1)$-tuple $(a_1, a_2, ..., a_{2^n-2}; a_{2^n-1})$ when we want to explicitly mention also the redundant value $m(\Omega) = a_{2^n-1} = 1 - \sum_{i=1}^{2^n-2} a_i$. For BFs on Ω_3 we use $(a_1, a_2, ..., a_6; a_7) = (m(\{\omega_1\}), m(\{\omega_2\}), m(\{\omega_3\}), m(\{\omega_1, \omega_2\}), m(\{\omega_1, \omega_3\}), m(\{\omega_2, \omega_3\}); m(\{\Omega_3\}))$.

Unfortunately, no algebraic analysis of BFs on Ω_n for $n > 2$ has been presented till now.

3 Non-conflicting and Conflicting Parts of Belief Functions on 2-Element Frames of Discernment

For BFs on Ω_2 the following holds true:

Proposition 1 *BF Bel on Ω_2 is non-conflicting iff $Bel \in S_1 \cup S_2$.*

Proof. Obviously the simple support elements of S_1, S_2 are non-conflicting. $Pl(\{\omega_i\}) = m(\{\omega_i\}) + m(\{\omega_1, \omega_2\}) = 1 - m(\{\omega_j\})$, where $i \neq j$. Thus $Pl(\{\omega_i\}) = 1$ iff $m(\{\omega_j\}) = 0$ iff $Bel \in S_1 \cup S_2$. □

We will use the important property of Dempster's sum, which is respecting the homomorphisms h and f, i.e., respecting the h-lines and f-ellipses, when two BFs are combined on two-element frame of discernment [4, 14, 15]. Using this property we obtain the following statement.

Proposition 2 *Any belief function $(a,b) \in \Omega_2$ is the result of Dempster's combination of BF $(a_0, b_0) \in S_1 \cup S_2$ and a BF $(s,s) \in S$, such that (a_0, b_0) has the same plausibility decision support (same normalized plausibility) for the elements of Ω_2 as (a,b) does. (Trivially, $(s,s) = (0,0) \oplus (s,s)$ for $(s,s) \in S$, and $(a_0, b_0) = (a_0, b_0) \oplus (0,0)$ for elements of S_1 and S_2).*

$(a_0, b_0) \in S_1 \cup S_2$ has no internal conflict, and (s,s) does not prefer any of the elements of Ω_2. Let us call (a_0, b_0) a non-conflicting part *of (a,b). There is $(a_0, b_0) = (\frac{a-b}{1-b}, 0)$ for $a \geq b$ and $(a_0, b_0) = (0, \frac{b-a}{1-a})$ for $a \leq b$.*

Proof. (a_0, b_0) is the intersection of h-line containing (a,b) with $S_1 \cup S_2$. Semigroup S is a part of h-line containing 0 and $0'$, thus the result of combination of any element $(s,s) \in S$ with (a_0, b_0), i.e, $(s,s) \oplus (a_0, b_0)$ lies on the same h-line as both (a_0, b_0) and (a,b).

$Pl\_P(a,b) = Pl\_P(a_0,b_0)$, thus $\frac{1-b}{2-a-b} = \frac{1-b_0}{2-a_0-b_0}$ and $\frac{1-a}{2-a-b} = \frac{1-a_0}{2-a_0-b_0}$. For $a \geq b$ there is $b_0 = 0$ and $\frac{1-b}{2-a-b} = \frac{1}{2-a_0}$, thus $\frac{2-a-b}{1-b} = \frac{2-a_0}{1}$, and $a_0 = 2 - \frac{2-a-b}{1-b} = \frac{a-b}{1-b}$. And similarly for $a \leq b$ there is $a_0 = 0$ and $\frac{1-a}{2-a-b} = \frac{1}{2-b_0}$, thus $b_0 = \frac{b-a}{1-a}$. □

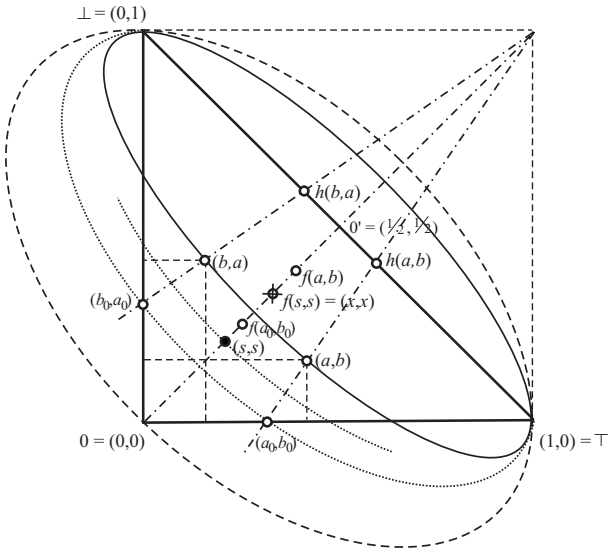


Figure 2: Conflicting and non-conflicting parts of BF on 2-element frame of discernment.

Let us look for (s,s) from the proposition now. It holds true that $(a,b) = (a_0,b_0) \oplus (s,s)$, thus it also holds true $f(a,b) = f(a_0,b_0) \oplus f(s,s)$. Let us denote $f(a_0,b_0) = (u,u), f(a,b) = (v,v), f(s,s) = (x,x)$ for a moment, thus we have $(u,u) \oplus (x,x) = (v,v)$, where $v = 1 - \frac{(1-u)(1-x)}{1-2ux} = \frac{u+x-3ux}{1-2ux}$, hence $u + x - 3ux = v - 2vux$ and $x = \frac{v-u}{1-3u+2uv}$. We can express this as Lemma 1 (i).

The existence of (x,x), thus also a possibility of its computation from (v,v) and (u,u) follows the fact, that S is isomorphic to the positive cone of group of reals, or a property subtraction in S as a substructure of algebraic structure dempsteroid [14, 15].

We already can compute value $f(s,s)$, the rest is computation of (s,s) as S-preimage of $f(s,s) = (s,s) \oplus (s,s) = (x,x)$. Similarly as before we have $x = 1 - \frac{(1-s)(1-s)}{1-2ss} = \frac{2s-3s^2}{1-2s^2}$ now, thus $2s - 3s^2 = x - 2s^2x$ and $0 = (3-2x)s^2 - 2s + x = 0$, hence $s_{1,2} = \frac{2 \pm \sqrt{4-4(3-2x)x}}{2(3-2x)} = \frac{1 \pm \sqrt{(1-x)(1-2x)}}{3-2x}$.

We know that $0 \leq s \leq x \leq \frac{1}{2}$, thus $0 \leq \sqrt{(1-x)(1-2x)} \leq 1$, $0 \leq 1 \pm \sqrt{(1-x)(1-2x)}$, $2 \leq 3 - 3x$. Thus $0 \leq \frac{1 \pm \sqrt{(1-x)(1-2x)}}{3-2x}$ always holds true. It should further hold true that $\frac{1 \pm \sqrt{(1-x)(1-2x)}}{3-2x} \leq \frac{1}{2}$, thus $2 \pm 2\sqrt{(1-x)(1-2x)} \leq 3 - 2x$ and $\pm 2\sqrt{(1-x)(1-2x)} \leq 1 - 2x$. It always hods true that $-\sqrt{(1-x)(1-2x)} \leq 0 \leq 1 - 2x$ for $0 \leq x \leq \frac{1}{2}$. On the other hand, from $2\sqrt{(1-x)(1-2x)} \leq 1 - 2x$, $4(1-x)(1-2x) \leq (1-2x)(1-2x)$, $4(1-x) \leq (1-2x)$, $3 \leq (2x)$ and $\frac{3}{2} \leq x$; this is in contradiction with $x \leq \frac{1}{2}$, hence it must be $s = \frac{1-\sqrt{(1-x)(1-2x)}}{3-2x}$.

We can formulate this as Lemma 1(ii). Finally, we obtain a summarization in Theorem 2.

Lemma 1 *(i) For any BFs (u,u), (v,v) on S, such that $u \leq v$, we can compute their Dempster's 'difference' (x,x) such that $(u,u) \oplus (x,x) = (v,v)$, as $(x,x) = (\frac{v-u}{1-3u+2uv}, \frac{v-u}{1-3u+2uv})$.*

(ii) For any BF (w,w) on S, we can compute its Dempster's 'half' (s,s) such that $(s,s) \oplus (s,s) = (w,w)$, as $(s,s) = (\frac{1-\sqrt{1-3w+2w^2}}{3-2w}, \frac{1-\sqrt{1-3w+2w^2}}{3-2w}) = (\frac{1-\sqrt{(1-w)(1-2w)}}{3-2w}, \frac{1-\sqrt{(1-w)(1-2w)}}{3-2w})$.

(iii) There is no Dempster's 'difference' on D_0 in general.

Proof. Parts (i) and (ii) were already proved by deriving of formulas for computing of (x,x) and (s,s). Nevertheless, we can alternatively verify the formulas is it follows.

$(a,b) \oplus (c,d) = (1 - \frac{(1-a)(1-c)}{1-(ad+bc)},\ 1 - \frac{(1-b)(1-d)}{1-(ad+bc)})$ in general, for $a = b$ and $c = d$ we obtain a special case of the formula: $(a,a) \oplus (c,c) = (1 - \frac{(1-a)(1-c)}{1-(2ac)},\ 1 - \frac{(1-a)(1-c)}{1-(2ac)})$.

$(u,u) \oplus (\frac{v-u}{1-3u+2uv}, \frac{v-u}{1-3u+2uv}) = (1 - \frac{(1-u)(1-\frac{v-u}{1-3u+2uv})}{1-(2u\frac{v-u}{1-3u+2uv})},$
$1 - \frac{\frac{(1-u)(v-2u+2uv-v)}{1-3u+2uv}}{\frac{1-3u+2uv-2u(v-u)}{1-3u+2uv}}) = (\frac{-3uv+v+2u^2v)}{1-3u+2u^2},\ \frac{v(1-3u+2u^2)}{1-3u+2u^2}) = (v,v)$.

$(s,s) \oplus (s,s) = (1 - \frac{(1-s)^2}{1-(2s^2)},\ \frac{2s-3s^2}{1-(2s^2)}) = (\frac{2\frac{1-\sqrt{(1-w)(1-2w)}}{3-2w}}{1-(2s^2)} +$
$\frac{-3(\frac{1-\sqrt{(1-w)(1-2w)}}{3-2w})^2}{1-(2(\frac{1-\sqrt{(1-w)(1-2w)}}{3-2w})^2)},\ \frac{2\frac{1-\sqrt{\ }}{3-2w} - 3\frac{1-2\sqrt{\ }+(1-w)(1-2w)}{(3-2w)^2}}{1-(2\frac{1-2\sqrt{\ }+(1-w)(1-2w)}{(3-2w)^2})}) =$
$(\frac{5w+4w\sqrt{\ }-6w^2)}{5-6w+4\sqrt{\ }},\ \frac{w(5+4\sqrt{\ }-6w)}{5-6w+4\sqrt{\ }}) = (w,w)$.

(iii) There is a lot of counter-examples, e.g., BFs Bel_1 and Bel_2 on the same f-ellipse: when combining any BF different from $0 = (0,0)$ with any of them, the result is on a narrower ellipse closer to G. □

Theorem 2 *Any BF (a,b) on 2-element frame of discernment Ω_2 is Dempster's sum of its unique* non-conflicting part $(a_0,b_0) \in S_1 \cup S_2$ *and of its unique* conflicting part $(s,s) \in S$, *which does not prefer any element of Ω_2, i.e. $(a,b) = (a_0,b_0) \oplus (s,s)$. It holds true that $s = \frac{b(1-a)}{1-2a+b-ab+a^2} = \frac{b(1-b)}{1-a+ab-b^2}$ and $(a,b) = (\frac{a-b}{1-b}, 0) \oplus (s,s)$ for $a \geq b$; and similarly that $s = \frac{a(1-b)}{1+a-2b-ab+b^2} = \frac{a(1-a)}{1-b+ab-a^2}$ and $(a,b) = (0, \frac{b-a}{1-a}) \oplus (s,s)$ for $a \leq b$.*

Proof. The existential part of the statement simply follows proposition 2 and both parts of Lemma 1. Uniqueness follows proposition 1, uniqueness of the h-line containing (a, b) and of its intersection with $S_1 \cup S_2$, and uniqueness of f-ellipse containing (a, b) and of its intersection with S. The rest is direct computation or verification. A verification for $a \geq b$ follows:
$(a_0, b_0) \oplus (s, s) = (\frac{a-b}{1-b}, 0) \oplus (\frac{b(1-b)}{1-a+ab-b^2}, \frac{b(1-b)}{1-a+ab-b^2}) = (1 - \frac{(1-\frac{a-b}{1-b})(1-\frac{b(1-b)}{1-a+ab-b^2})}{1-\frac{a-b}{1-b}\cdot\frac{b(1-b)}{1-a+ab-b^2}}, 1 - \frac{(1-\frac{b(1-b)}{1-a+ab-b^2})}{\frac{(1-b)(1-a+ab-b^2)}{(1-b)(1-a+ab-b^2)} - \frac{(a-b)b(1-b)}{(1-b)(1-a+ab-b^2)}}) = (\frac{a(1-b)}{(1-b)}, \frac{b(1-a)}{(1-a)}) = (a, b)$.
For $a \leq b$ we have:
$(a_0, b_0) \oplus (s, s) = (0, \frac{b-a}{1-a}) \oplus (\frac{a(1-a)}{1-b+ab-a^2}, \frac{a(1-a)}{1-b+ab-a^2}) = ...$, a and b and components of the couple are mutually substituted w.r.t. the case $a \geq b$, thus the result is (a, b) again. For equality of both formulas for s see [10]. □

An alternative proof is a derivation of formulas which is based on a similar idea as the derivation of formulas in Lemma 1. As we know the existence of (s, s) and that $a_0 = \frac{a-b}{1-b}$ for $a \geq b$, we know that $(a, b) = (a_0, 0) \oplus (s, s) = (1 - \frac{(1-a_0)(1-s)}{1-(a_0 s+0)}, 1 - \frac{(1-0)(1-s)}{1-a_0 s})$. Thus $a = 1 - \frac{(1-a_0)(1-s)}{1-(a_0 s+0)} = \frac{s+a-b-2as+bs}{1-b-as+bs}$. Hence $a - ab - a^2 s + abs = s + a - b - 2as + bs$ and $s = \frac{b(1-a)}{1-2a+b-ab+a^2}$. Similarly we have $b = 1 - \frac{(1-0)(1-s)}{1-a_0 s} = \frac{s-as}{1-b-as+bs}$. Hence $s = \frac{b(1-b)}{1-a+ab-b^2}$. Analogically, we can compute both versions of s for the case where $a \leq b$, see [10]. □

We can summarize formulas from the theorem as $(a, b) = (a_0, b_0) \oplus (s, s) = (max(\frac{a-b}{1-b}, 0), max(\frac{b-a}{1-a}, 0)) \oplus (\frac{min(a,b)(1-min(a,b))}{1+ab-max(a,b)-min^2(a,b)}, \frac{min(a,b)(1-min(a,b))}{1+ab-max(a,b)-min^2(a,b)})$. And analogically for the second expression of s [10].

Proof. Just a verification for $a \geq b$, and that for $a \leq b$. □

4 Non-conflicting Part of BFs on General Finite Frames of Discernment

Let us turn our attention to a question of non-conflicting and conflicting parts of BFs defined on an n-element frame of discernment $\Omega_n = \{\omega_1, ..., \omega_n\}$. We start with a characterization of the set of non-conflicting BFs.

Proposition 3 *The set of non-conflicting BFs is just the set of all BFs such, that all focal elements of a BF have non-empty intersection.*

Consonant BFs are a special case of non-conflicting BFs.

Proof. $Pl(\{\omega_i\}) = 1$ for some $\omega_i \in \Omega$ iff $\omega_i \in X$ for all X such that $m(X) > 0$ iff $\omega_i \in \bigcup_{m(X)>0} X$ iff $\bigcup_{m(X)>0} X \neq \emptyset$.
The least focal element of a consonant BF is intersection of its focal elements; there are many non-conflicting BFs which are not consonant on Ω_n, $n > 2$, e.g., $(0, 0, 0, 0.7, 0.3, 0; 0)$ on Ω_3, i.e., $m(\{\omega_1, \omega_2\}) = 0.7, m(\{\omega_1, \omega_3\}) = 0.3$. □

We would like to verify that Theorem 2 holds true also for BFs defined on general finite frames, i.e., to verify the following hypothesis:

Hypothesis 1 *We can represent any BF Bel on n-element frame of discernment $\Omega_n = \{\omega_1, ..., \omega_n\}$ as Dempster's sum $Bel = Bel_0 \oplus Bel_S$ of non-conflicting BF Bel_0 and of indecisive conflicting BF Bel_S which has no decisional support, i.e. which does not prefer any element of Ω_n to the others, see Figure 3.*

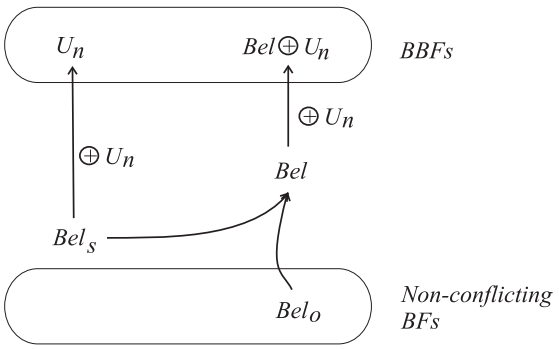


Figure 3: Schema of Hypothesis 1.

Similarly to two-element frames, we have simple trivial examples $Bel_N = Bel_N \oplus 0$ for all non-conflicting BFs Bel_N and $Bel_I = 0 \oplus Bel_I$ for all indecisive BFs Bel_I, where $0 = (0, 0, ..., 0; 1)$.

We would like to follow the idea from the case of two-element frames, see Figure 4. Unfortunately, there was not presented any algebraic description of BFs defined on n-element frames till now. We have nothing like Dempster's semigroup for n-element frames, we have no n-versions of $-Bel$ and of homomorphisms f and h, neither group properties of a set of indecisive BFs.

An issue of homomorphism h is quite promising: $h(Bel) = Bel \oplus U_n = Pl\_P(Bel)$. From results on probabilistic transformations presented in [7] it can be concluded that, $Pl\_P(Bel) = Bel \oplus U_n$, for proof see [8]. From [3] we know that Pl_P commutes with $\oplus$, i.e. $Pl\_P(Bel_1 \oplus Bel_2) = Pl\_P(Bel_1) \oplus Pl\_P(Bel_2)$,

thus we have homomorphism h for BFs on an n-element frame of discernment. To generalize all homomorphic properties of h we have also to verify a general versions of $h(0) = 0'$ and $h(0') = 0'$. It really holds true that $h(0,0,...,0) = 0 \oplus U_n = (0,0,...,0) \oplus (\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}, 0, 0, ..., 0) = (\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}, 0, 0, ..., 0) = U_n$. And similarly $h(U_n) = U_n \oplus U_n = U_n$. Hence the following theorem is proved. As there is no ordering of either BFs or elements of a frame of discernment, we cannot speak of ordered homomorphism as in two-element case.

Theorem 3 *The mapping $h(Bel) = Bel \oplus U_n = Pl\_P(Bel)$ is an homomorphism of an algebra of BFs on an n-element frame of discernment with the binary operation of Dempster's sum $\oplus$ and two nulary operations (constants) 0 and U_n.*

Thus, we can apply h with its homomorphic properties also in a general case. We have Bel and $h(Bel) = Pl\_P(Bel)$ which is BBF, i.e., BF which has upto n positive m-values (bbms). $h(Bel) = (h_1(Bel), h_2(Bel), ..., h_n(Bel), 0, 0, ...; 0)$; when interpreting $h(Bel)$ as a probability distribution on Ω, we have $h(Bel)(\omega_i) = h_i(Bel)$. We can use the following procedure to compute a related unique consonant BF Bel_0 to any $h(Bel)$.

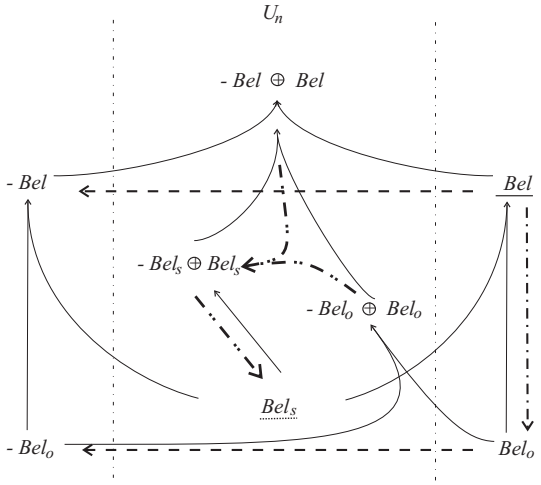


Figure 4: Schema of a decomposition of BF Bel.

Let there are k different values $h_i(Bel)$ for $i = 1, ..., n$, thus $1 \leq k \leq n$. According to this, we have splitting of the frame Ω into k disjoint subsets $\Omega = \Omega_1 \cup \Omega_2 \cup ... \cup \Omega_k$, such the the elements of the same subset have the same value $h(Bel)(\omega)$. Let $\Omega_1 = \{\omega_{11}, ..., \omega_{1j_1}\}$ be a set of elements of the frame with the highest m-value (bbm) $(h(Bel)(\omega_{11}) = h(Bel)(\omega_{12}) = ... = h(Bel)(\omega_{1j_1})$, where $1 \leq j_1 \leq n - k + 1)$, and $\Omega_2 = \{\omega_{21}, ..., \omega_{2j_2}\}$ be a set of elements with the 2nd highest bbm $(h(Bel)(\omega_{21})$; $1 \leq j_2 \leq n - j_1 - k + 2)$, then we define $m_w(\Omega_1) = h(Bel)(\omega_{11}) - h(Bel)(\omega_{21})$, further we define $m_w(\Omega_1 \cup \Omega_2) = h(Bel)(\omega_{21}) - h(Bel)(\omega_{31})$, where $h(Bel)(\omega_{31})$ is the 3rd largest m-value of $h(Bel)$. We continue similarly defining $m_w(\bigcup_{i=1}^{m} \Omega_i) = h(Bel)(\omega_{m1}) - h(Bel)(\omega_{(m+1)1})$, where $\Omega_i = \{\omega_{i1}, ..., \omega_{ij_i}\}$ is the set of elements with the i-th highest m-value of $h(Bel)$, until $m_w(\Omega) = h(Bel)(\omega_{k1})$ is defined, where $\Omega_k = \{\omega_{k1}, ..., \omega_{kj_k}\}$ is the set of elements with the least (possibly zero), m-value $h(Bel)(\omega_{k1})$, $j_k = n - \sum_{i=1}^{k-1} j_i$. $m_w(\bigcup_{i=1}^{m} \Omega_i) > 0$ for all $m < k$ because less value is always decreased, $m_w(\Omega_k) \geq 0$, $\sum_{m=1}^{k} m_w(\bigcup_{i=1}^{m} \Omega_i) = h(Bel)(\omega_{11}$. Then m_0 is a normalization of working bba m_w, thus focal elements of m_0 are nested and $Pl(\omega) = 1$ for $\omega \in \Omega_1$, hence Bel_0 is normalized consonant, i.e., non-conflicting BF. For detail and verification that, $Bel_0 \oplus U_n = h(Bel)$ and that $m_0 = (\frac{a-b}{1-b}, 0)$ is a special case of general m_0, see [10].

Finally, we can simplify the construction of Bel_0 in the following way: there is one normalization in computation of $Bel \oplus U_n = Pl\_P(Bel)$ and the following normalization in the transformation of m_w to m_0. Normalization commutes with the construction of m_w from $Pl\_P(Bel)$, thus when computing Bel_0, we can use $Pl(Bel)$ instead of $h(Bel) = Pl\_P(Bel)$ and apply only one normalization in the end, where normalization factor is the multiple of the original ones. Thus we obtain $m'_w(\{\omega_{11}, ..., \omega_{1j_1}\}) = Pl(Bel)(\omega_{11}) - Pl(Bel)(\omega_{21})$, etc. This computational simplification is important also from the theoretical point of view, because it removes Dempster's rule $\oplus$ hidden in h from the construction of Bel_0. Hence any Bel_0 has defined its non-conflicting part independently of any belief combination rule.

Lemma 2 *For any BF Bel defined on Ω_n there exists unique consonant BF Bel_0 such that,*

$$h(Bel_0 \oplus Bel_S) = h(Bel) \tag{1}$$

for any BF Bel_S such that $Bel_S \oplus U_n = U_n$.

Proof. The existence follows the construction of Bel_0 when replacing (1) with $Bel_0 \oplus Bel_s \oplus U_n = Bel \oplus U_n$. For uniqueness we will also follow the construction of Bel_0: $h(Bel)$ is unique, thus also set of its m-values $h_i(Bel)$ is unique, k of them are different, $h_i(Bel)$ are real values from $[0, 1]$, thus their order is also unique, hence splitting of Ω into k disjoint subsets is unique as well, i.e. set of focal elements of m_w and m_0 is unique. Computation of differences is also unique thus we have unique m_w values and also their normalization m_0 values, hence m_0 is unique consonant bba such that $h(m_0) = h(Bel)$.
Futher it holds true that, $h(Bel_0 \oplus Bel_S) = h(Bel_0) \oplus h(Bel_S) = h(Bel) \oplus h(Bel_S) = h(Bel) \oplus U_n = h(Bel)$. □

Let us notice, that the stronger statement for a general non-conflicting BFs does not hold true on Ω_n. There could be several different non-conflicting BFs Bel_i such that $h(Bel_i \oplus Bel_S) = h(Bel)$ for any indecisive BF B_S. See, the following example.

Example 1 *To BF* $Bel = (0.25, 0.175, 0.075, 0.35, 0.15, 0)$ *with* $h(Bel) = (0.25, 0.175, 0.075, 0.35, 0.15, 0) \oplus (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0) = (0.50, 0.35, 0.15, 0, 0, 0)$ *there are following non-conflicting BFs:* $Bel_0 = (0.3, 0, 0, 0.4, 0, 0; 0.3)$, $Bel_1 = (0, 0, 0, 0.7, 0.3, 0; 0)$, $Bel_2 = (0.2, 0, 0, 0.5, 0.1, 0; 0.2)$; $Pl_i(\{\omega_1\}) = 1$, *thus* Bel_i*s are all non-conflicting, we can simply verify that* $h(Bel_i) = h(Bel)$, *thus* $(Bel_i \oplus Bel_S) \oplus U_3 = Bel_i \oplus (Bel_S \oplus U_3) = Bel_i \oplus U_3 = h(Bel)$.

Let us turn our attention to $f(Bel)$ and $-Bel$ now. $f(a, b) = -(a, b) = (b, a)$ on Ω_2, thus we will try to generalize $-Bel$ to BFs on Ω_n now. We have nothing like S defined for BFs on Ω_n, thus we suppose $h(Bel \oplus -Bel) = U_n$ for $-Bel$. On Ω_2 it holds true that $-m(\{\omega_1\}) = m(\{\omega_2\}) = m(\Omega_2 \setminus \{\omega_1\})$, $-m(\{\omega_2\}) = m(\Omega_2 \setminus \{\omega_2\})$, and $-m(\Omega_2) = m(\Omega_2)$[5]. Unfortunately, the simple idea to define $-m$ as $-m(X) = m(\Omega_n \setminus X)$ does not work in general, not even for general consonant BFs, e.g., for $Bel = (0.5, 0, 0, 0.2, 0, 0; 0.3)$ and $\sim Bel = (0, 0, 0.2, 0, 0, 0.5; 0.3)$ we have $Bel \oplus \sim Bel = (\frac{15}{61}, \frac{10}{61}, \frac{6}{61}, \frac{6}{61}, \frac{0}{61}, \frac{15}{61}; \frac{9}{61})$, $(\frac{15}{61}, \frac{10}{61}, \frac{6}{61}, \frac{6}{61}, \frac{0}{61}, \frac{15}{61}; \frac{9}{61}) \oplus (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0; 0) = (\frac{30}{70}, \frac{40}{70}, \frac{30}{70}, 0, 0, 0; 0) = (\frac{3}{7}, \frac{4}{7}, \frac{3}{7}, 0, 0, 0; 0) \neq U_3$. Thus $h(Bel \oplus \sim Bel) \neq U_n$, hence $\sim Bel \neq -Bel$. The idea of complements $(\Omega \setminus X)$ works only in some special cases, e.g., for $(0.7, 0, 0, 0, 0, 0) \oplus (0, 0, 0, 0, 0, 0.7) = (21/51, 0, 0, 0, 0, \frac{21}{51}) \doteq (0.41, 0, 0, 0, 0, 0.41)$, $h(0.41, 0, 0, 0, 0, 0.41) = U_3$ on Ω_3 and for other simple support BFs in general.

To simplify the investigated situation, we will start with qBBFs on 3-element frame of discernment Ω_3 (i.e., with BFs such that $m(X) = 0$ for $|X| = 2$). The set of qBBFs on Ω_3 can be represented by a three dimensional triangle which simply generalizes the triangle of Dempster's pairs, see Figure 5. Unfortunately, the only consonant, i.e. non-conflicting, BFs are singleton simple support functions as $(a, 0, 0, 0, 0, 0; 1\text{–}a)$, thus only a small part of the triangle is mapped to non-conflicting BFs within the triangle (Bel_0 is outside of the triangle for a majority of qBBFs). Thus, this is not a good domain to search for $-Bel_0$.

Let us look at BBFs now, i.e. BFs as $(a, b, c, 0, 0, 0; 0) = (a, b, 1 - a - b, 0, 0, 0; 0)$. Let $-(a, b, 1-a-b, 0, 0, 0) = (x, y, 1-x-y, 0, 0, 0)$, thus

[5]Note that $-m(X)$ is an abbreviation for $(-m)(X)$, thus both $m(X)$ and $-m(X)$ may be positive in general. Specially $-m(\Omega_2)$ is an abbreviation for $(-m)(\Omega_2)$, thus $-m(\Omega_2) = m(\Omega_2)$, where both sides of the equation are positive in general.

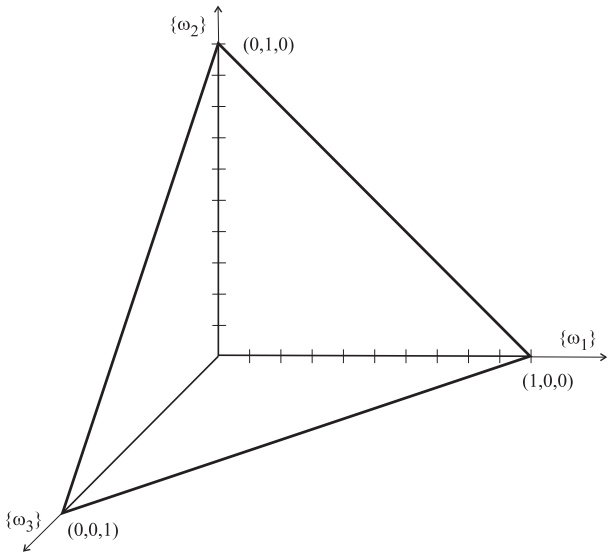


Figure 5: Quasi Bayesian BFs on 3-element frame Ω_3.

$-(a, b, 1-a-b, 0, 0, 0) \oplus (x, y, 1-x-y, 0, 0, 0) = U_3$ should hold true.

Thus $ax = by = (1 - a - b)(1 - x - y)$, $y = \frac{a}{b}x, (1 - x - y) = \frac{a}{1-a-b}x$, hence $1 - x - \frac{a}{b}x = \frac{a}{1-a-b}x$. Solving the previous equation we obtain $x = \frac{b(1-a-b)}{a+b-a^2-b^2-ab}$ and further $y = \frac{a(1-a-b)}{a+b-a^2-b^2-ab}$. Using $c = 1 - a - b$, we obtain $x = \frac{bc}{ab+ac+bc}$, $y = \frac{ac}{ab+ac+bc}$ and $1 - x - y = z = 1 - \frac{bc+ac}{ab+ac+bc} = \frac{ab}{ab+ac+bc}$. E.g. $(a, b, c, 0, 0, 0) = (0.5, 0.3, 0.2)$, $x = \frac{0.3 \cdot 0.2}{0.5 \cdot 0.3 + 0.5 \cdot 0.2 + 0.3 \cdot 0.2}$, $y = \frac{5 \cdot 2}{5 \cdot 3 + 5 \cdot 2 + 3 \cdot 2}$ $z = \frac{3 \cdot 2}{5 \cdot 3 + 5 \cdot 2 + 3 \cdot 2}$, thus $-(0.5, 0.3, 0.2, 0, 0, 0) = (\frac{6}{31}, \frac{10}{31}, \frac{15}{31}, 0, 0, 0)$.

Thus we have $-Bel$ for any BBF $(a, b, 1-a-b, 0, 0, 0)$ on Ω_3 such that $0 < a, b < 1, a + b < 1$.

Analogically to the case of Ω_3, we can generalize the $-Bel$ to BBFs on Ω_n, to BFs $(a_1, a_2, ..., a_n, 0, 0, ..., 0; 0)$ such that $0 < a_i < 1$, for $i = 1, ..., n$ and $a_n = 1 - \sum_{i=1}^{n-1} a_i$. Let us denote $-(a_1, a_2, ..., a_n, 0, 0, ..., 0; 0) = (x_1, x_2, ..., x_n, 0, 0, ..., 0; 0)$ (where $x_n = 1 - \sum_{i=1}^{n-1} x_i$), thus we obtain $x_1 = 1/(1 + \sum_{i=2}^{n} \frac{a_1}{a_i})$, $x_i = \frac{a_1}{a_i} x_1$, or similarly to x_1: $x_i = 1/(1 + \sum_{i \neq j} \frac{a_i}{a_j})$.

An alternative expression for x_i is $x_i = \frac{\prod_{i \neq j} a_j}{\sum_{k=1}^{n} \prod_{j \neq k} a_j}$, for detail see [10].

Lemma 3 *For any BBF* $(a_1, a_2, ..., a_n, 0, 0, ..., 0; 0)$ *such that,* $a_i > 0$ *for* $i = 1, ..., n$, *there exists uniquely defined* $-(a_1, a_2, ..., a_n, 0, 0, ..., 0; 0) = (x_1, x_2, ..., x_n, 0, 0, ..., 0; 0) = (1/(1 + \sum_{i=2}^{n} \frac{a_1}{a_i}), \frac{a_1}{a_2} x_1, \frac{a_1}{a_3} x_1, ..., \frac{a_1}{a_n} x_1, 0, 0, ..., 0; 0)$ *such that,*

$$(a_1, a_2, ..., a_n, 0, 0, ..., 0) \oplus -(a_1, a_2, ..., a_n, 0, 0, ..., 0) = U_n.$$

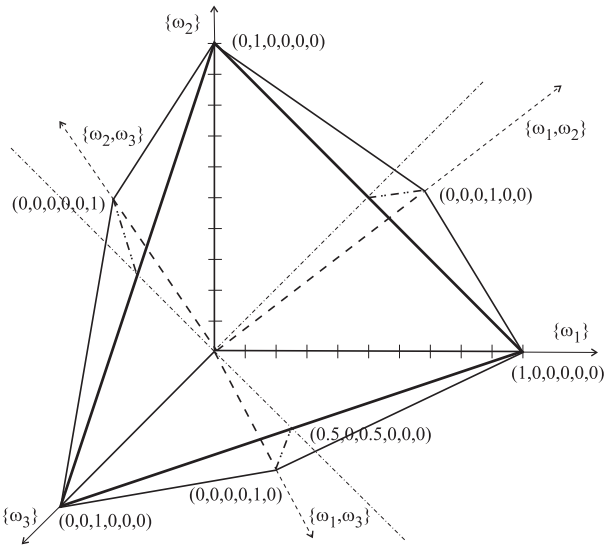


Figure 6: General BF on 3-element frame Ω_3.

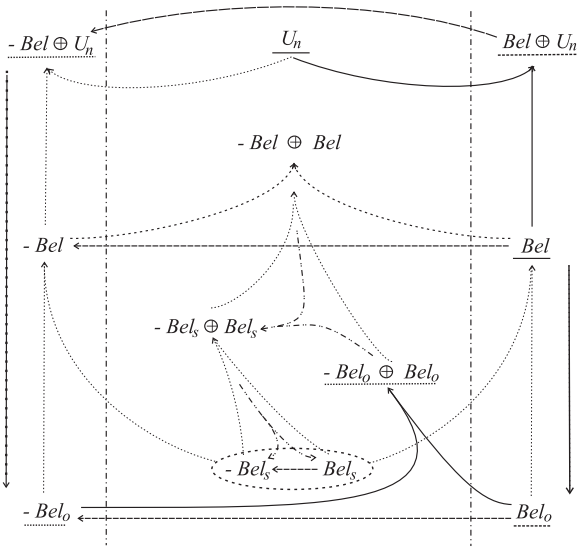


Figure 7: Detailed schema of a decomposition of BF Bel.

We have already observed, that $-Bel$ for a simple support function (SSF) is another SSF with a complementary focal element such that, $-m(\Omega_n \setminus X) = m(X)$; similarly we can define $-Bel$ also for simple support BBFs (i.e. categorical BBFs), see e.g., $-(1,0,0,0,0,0) = (0,0,0,0,0,1)$, but we have to notice that $(1,0,0,0,0,0) \oplus (0,0,0,0,0,1)$ is not defined (similarly to $(1,0) \oplus (0,1)$ on Ω_2). A definition of $-Bel$ for BBFs like $(a, 1-a, 0, 0, ..., 0)$ remains still open for more-element frames Ω_n, $n > 2$.

Summarising the previous results, we can step by step compute $h(Bel)$, $-h(Bel)$ and $(-h(Bel))_0$ from any Bel such that $Pl(\{\omega_i\}) > 0$ for all $\omega_i \in \Omega_n$, see Figure 7. Thus the following theorem holds true:

Theorem 4 *For any BF Bel defined on Ω_n there exists unique consonant BF Bel_0 such that,*

$$h(Bel_0 \oplus Bel_S) = h(Bel)$$

for any BF Bel_S such that $Bel_S \oplus U_n = U_n$. If for $h(Bel) = (h_1, h_2, ..., h_n, 0, 0, ..., 0)$ holds true that, $0 < h_i < 1$, then further exists unique BF $-Bel_0$ such that,

$$h(-Bel_0 \oplus Bel_S) = -h(Bel) \text{ and } h(Bel_0) \oplus -h(Bel_0) = U_n.$$

Proof. For existence and uniqueness of Bel_0 see Lemma 2. Existence of $-Bel_0$ follows its construction, $h(Bel)$ is unique according to its definition, for uniqueness of $-h(Bel)$ see Lemma 3 and final uniqueness of $-Bel_0$ follows Lemma 2 again. □

Corollary 1 *(i) For any consonant BF Bel such that $Pl(\{\omega_i\}) > 0$ there exist a unique BF $-Bel$; $-Bel$ is consonant in this case.*

(ii) There is one-to-one correspondence between Bayesian BFs and consonant BFs.

Proof. (i) Just take a consonant BF Bel, due to uniqueness of Bel_0 we have $Bel = Bel_0$, and also $-Bel = -Bel_0$. $Pl(\{\omega_i\}) > 0$ for all ω_i in the case of a consonant BF implies that $m(\Omega) > 0$, thus also $m_h(\{\omega_i\}) > 0$ for all ω_i, where $Bel_h = h(Bel) = Bel \oplus U_n$, thus we have $-Bel_h$ and $(-Bel_h)_0 = -Bel$; according its construction $-Bel$ is consonant and unique. If $Pl(\{\omega_i\}) = 0$ for some $\omega_i \in \Omega$, then $m(\Omega) = 0$, thus there exists ω_j such, that $m_h(\{\omega_j\}) = 0$, hence we have not defined either $-Bel_h$ or $-Bel$. (ii) Taking any BBF Bel, we obtain unique consonant Bel_0; $h(Bel_0)$ is also unique. □

We observed that $-m(X) = m(\Omega \setminus X)$ for $X \subset \Omega$ and SSF m. We can verify that the definition of $-Bel$ using $-h(Bel)$ agree with this observation. E.g., $Bel = Bel_0 = (\frac{2}{3}, 0, 0, 0, 0, 0; \frac{1}{3})$, $h(Bel) = (\frac{3}{5}, \frac{1}{5}, \frac{1}{5}, 0, 0, 0; 0)$, $-h(Bel) = (\frac{1}{7}, \frac{3}{7}, \frac{3}{7}, 0, 0, 0; 0)$, and $-Bel = (-h(Bel))_0 = (0, 0, 0, 0, 0, \frac{2}{3}; \frac{1}{3})$. In general we have $m(X) = a$ and $m(\Omega) = 1 - a$, where $|X| = k, |\Omega| = n$. Thus $h(m)(\omega_i) = \frac{a+(1-a)}{k(a+(1-a))+(n-k)(1-a)} = \frac{1}{n-(n-k)a}$ for $\omega_i \in X$ and $h(m)(\omega_j) = \frac{1-a}{k(a+(1-a))+(n-k)(1-a)} = \frac{1-a}{n-(n-k)a}$ for $\omega_j \in \Omega \setminus X$. Further, $-h(m)(\omega_i) = \frac{1^{k-1}(1-a)^{n-k}}{k \cdot 1^{k-1}(1-a)^{n-k}+(n-k)1^k(1-a)^{n-k-1}} = \frac{1-a}{n-ka}$, $-h(m)(\omega_j) = \frac{1^k(1-a)^{n-k-1}}{k \cdot 1^{k-1}(1-a)^{n-k}+(n-k)1^k(1-a)^{n-k-1}} = \frac{1}{n-ka}$, hence $-m(\Omega \setminus X) = (-h(m))_0(\Omega \setminus X) = (-h(m))(\omega_j) - (-h(m))(\omega_i) = \frac{\frac{1}{n-ka} - \frac{1-a}{n-ka}}{\frac{1}{n-ka} - \frac{1-a}{n-ka} + \frac{1-a}{n-ka}} = a$, and $-m(\Omega) = (-h(m))_0(\Omega) = (-h(m))(\omega_i) =$

$\frac{\frac{1-a}{n-ka}}{\frac{1}{n-ka}-\frac{1-a}{n-ka}+\frac{1-a}{n-ka}} = 1 - a$. Thus really $-m(\Omega \setminus X) = m(X)$ and $-m(\Omega) = m(\Omega)$. □

For completion of the diagram in Figure 7, we need a definition of $-Bel$ for general BFs on Ω to compute $Bel \oplus -Bel$ and analysis of indecisive BFs (i.e. BFs Bel such that, $h(Bel) = U_n$) to compute $Bel_S \oplus -Bel_S$ and resulting Bel_S and specify conditions under which Bel_S is defined and unique. Hence an algebraic analysis of BFs on a general finite frame of discernment is required.

5 Comments on other belief combination rules

There arises an interesting question about similar kind of decomposition of belief functions with another combination rules.

As it was already mentioned, the non-conflicting part Bel_0 of a belief function Bel defined above is independent from Dempster's rule of combination, as we can use the representation of homomorphism h using normalized plausibility of singletons $Pl\_P(Bel)$ instead of the original $h(Bel) = Bel \oplus U_n$. Thus Bel_0 can be computed without any relation to Dempster's rule and $Pl\_P(Bel_0) = Pl\_P(Bel)$ independently from any combination rule.

On the other hand $Pl\_P(Bel) \neq Bel_0 \textcircled{Y} U_n$, $Pl\_P(Bel) \neq Bel_0 \textcircled{D} U_n$, $Pl\_P(Bel) \neq Bel_0 \textcircled{\cup} U_n$, see Example 2. Even $Pl\_P(Bel) \neq Pl\_P(Bel_0 \textcircled{\odot} U_n)$, where $\textcircled{\odot}$ is either $\textcircled{Y}$, $\textcircled{D}$, $\textcircled{\cup}$ or some other combination rule. The equality holds true only for Dempster's rule: $Pl\_P(Bel) = Bel_0 \oplus U_n$; in the case of un-normalized conjunctive rule $\textcircled{\cap}$ we can apply additional normalization to obtain the equality, thus we have normalized conjuntive rule, i.e., Dempster's rule $\oplus$ again.

Example 2 *Let us take* $Bel = (0.3, 0.2, 0.1, 0.2, 0.1, 0.0; 0.1)$*, thus there is* $Pl = (0.7, 0.5, 0.3, 0.9, 0.8, 0.7; 1.0)$*,* $Pl\_P(Bel) = (\frac{7}{15}, \frac{5}{15}, \frac{3}{15})$*, and* $Bel_0 = (\frac{2}{7}, 0, 0, \frac{2}{7}, 0, 0; \frac{3}{7})$*. Hence we obtain* $(\frac{2}{7}, 0, 0, \frac{2}{7}, 0, 0; \frac{3}{7}) \oplus (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0; 0) = (\frac{7}{15}, \frac{5}{15}, \frac{3}{15}, 0, 0, 0; 0)$*; but* $(\frac{2}{7}, 0, 0, \frac{2}{7}, 0, 0; \frac{3}{7}) \textcircled{Y} (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0; 0) = (\frac{7}{21}, \frac{5}{21}, \frac{3}{21}, 0, 0, 0; \frac{6}{21}) \neq (\frac{7}{15}, \frac{5}{15}, \frac{3}{15}, 0, 0, 0; 0)$*,* $(\frac{2}{7}, 0, 0, \frac{2}{7}, 0, 0; \frac{3}{7}) \textcircled{D} (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0; 0) = (\frac{7}{21}, \frac{5}{21}, \frac{3}{21}, \frac{2}{21}, \frac{2}{21}, 0; \frac{2}{21}) \neq (\frac{7}{21}, \frac{5}{21}, \frac{3}{15}, 0, 0, 0; 0)$*, and similarly* $(\frac{2}{7}, 0, 0, \frac{2}{7}, 0, 0; \frac{3}{7}) \textcircled{\cup} (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0; 0) = (\frac{2}{21}, 0, 0, \frac{6}{21}, \frac{2}{21}, 0; \frac{11}{21}) \neq (\frac{7}{15}, \frac{5}{15}, \frac{3}{15}, 0, 0, 0; 0)$*. Further* $Pl\_P(\frac{7}{21}, \frac{5}{21}, \frac{3}{21}, 0, 0, 0; \frac{6}{21}) = (\frac{13}{33}, \frac{11}{33}, \frac{9}{33}) \neq (\frac{7}{15}, \frac{5}{15}, \frac{3}{15})$*,* $Pl\_P(\frac{7}{21}, \frac{5}{21}, \frac{3}{15}, \frac{2}{21}, \frac{2}{21}, 0; \frac{2}{21}) = (\frac{13}{29}, \frac{9}{29}, \frac{7}{29}) \neq (\frac{7}{15}, \frac{5}{15}, \frac{3}{15})$*, and* $Pl\_P((\frac{2}{21}, 0, 0, \frac{6}{21}, \frac{2}{21}, 0; \frac{11}{21}) = (\frac{21}{51}, \frac{17}{51}, \frac{13}{51}) \neq (\frac{7}{15}, \frac{5}{15}, \frac{3}{15})$.

Nevertheless, if there is a couple of homomorphisms for any combination rule $\textcircled{\odot}$ analogic to morphisms f and h from Dempster's semigroup, then there exists an analogy of Bel_0 also for the combination rule $\textcircled{\odot}$.

When expressing h using $Pl\_P(Bel)$ there arises another interesting question about similar kind of non-conflicting part and decomposition of belief functions using a different probabilistic transformations.

Considering Smets' pignistic transformation $BetT$ for computing pignistic probability $BetP$ we obtain non-conflicting BF $Bel_{0\text{–}BetP}$, where $m_{w\text{–}BetP}(\bigcup_{i=1}^{m} \Omega_i) = |\bigcup_{i=1}^{m} \Omega_i|(h(Bel)(\omega_{m1}) - h(Bel)(\omega_{(m+1)1}))$, which is normalized, hence $m_{0\text{–}BetP} = m_{w\text{–}BetP}$. $BetT$ does not commute either with Dempster's rule nor with other rules defined for belief combination, thus we cannot use $Bel_{0\text{–}BetP}$ for decomposition of belief functions to conflicting and non-conflicting parts. For counter-examples see [10].

The most perspective pignistic transformation is normalized belief of singletons Bel_P which is compatible with disjunctive rule of combination [7], unfortunately, the reverse transformation maps any Bel and $Bel\_P(Bel)$ to the vacuous belief function $0 = (0, 0, ..., 0; 1)$, which is really non-conflicting, but it does not reprezent non-conflicting part of the belief function Bel. In this case it represents zero conflicting part, as the disjunctive rule is completely non-conflicting; thus it holds true $Bel = Bel \textcircled{\cup} 0$, where Bel is trivial 'disjunctive non-conflicting' part of itself and 0 is trivial 'disjunctive conflicting' part of any BF Bel.

Moreover, it is possible to show that there is no similar decomposition of belief functions for $\textcircled{Y}$, $\textcircled{D}$, $\textcircled{\cup}$ and a for a series of other combination rules, see [10]. Any Bayesian BF serves as counter-example there.

6 Conclusion

Decomposition of a belief function (BF) defined on a two-element frame of discernment to Dempster's sum of its unique non-conflicting and unique indecisive conflicting part is defined and presented here.

Homomorphic properties of mapping $h(Bel) = Bel \oplus U_n$ which corresponds to normalized plausibility of singletons were verified for BFs defined on a general finite frame of discernment. $-Bel$ was generalized to Bayesian BFs and for consonant BFs on a general n-element frame.

Further a unique consonant non-conflicting part Bel_0 of a general BF Bel on a finite frame was defined. For specification of a corresponding conflicting part of Bel and its uniqueness/existence properties, an algebraic analysis of BFs on a general finite frame of discernment is required.

The presented topic is finally discused also from the point of view of alternative rules of combination and alternative probabilistic transformations.

The presented results improve general understanding of belief functions and their combination, especially in conflicting cases. They can be used as one of corner-stones to further study of conflicts between belief functions.

Acknowledgments

This research is supported by the grant P202/10/1826 of the Grant Agency of the Czech Republic.

Partial support by the Institutional Research Plan AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications" is also acknowledged.

The author is grateful to Eva Pospíšilová for creation of useful and illuminative figures.

References

[1] R. G. Almond: Graphical Belief Modeling. Chapman & Hall, London, 1995.

[2] A. Ayoun, Ph. Smets: Data association in multi-target detection using the transferable belief model. *International Journal of Intelligent Systems* **16** (10): 1167–1182, 2001.

[3] B. R. Cobb, P. P. Shenoy: A Comparison of Methods for Transforming Belief Functions Models to Probability Models. In: Nielsen, T. D., Zhang, N. L. (eds.) ECSQARU 2003. LNAI 2711: 255–266. Springer, Heidelberg, 2003.

[4] M. Daniel: Algebraic structures related to Dempster-Shafer theory. In B. Bouchon-Meunier, R. R. Yager, and L. A. Zadeh (eds.), *Advances in Intelligent Computing - IPMU'94*. LNCS 945: 51–61. Springer-Verlag, Berlin Heidelberg, 1995.

[5] M. Daniel: Distribution of Contradictive Belief Masses in Combination of Belief Functions. In: B. Bouchon-Meunier, R. R. Yager, L. A. Zadeh (eds.) *Information, Uncertainty and Fusion*, pp. 431–446. Kluwer Acad. Publ., Boston, 2000.

[6] M. Daniel: Algebraic Structures Related to the Combination of Belief Functions. *Scientiae Mathematicae Japonicae* **60** (2): 245–255, 2004. *Sci. Math. Jap. Online* **10**: 501–511, 2004.

[7] M. Daniel: Probabilistic Transformations of Belief Functions. In: Godo, L. (ed.) Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2005). LNAI 3571: 539–551. Springer, Heidelberg, 2005.

[8] M. Daniel: New Approach to Conflicts within and between Belief Functions. Technical report V-1062, ICS AS CR, Prague, 2009.

[9] M. Daniel: Conflicts within and between Belief Functions. In: E. Hüllermeier, R. Kruse, E. Hoffmann (eds.) *IPMU 2010*. LNAI 6178: 696–705. Springer-Verlag, Berlin Heidelberg, 2010.

[10] M. Daniel: *Conflicts of Belief Functions.* Technical report V-1108, ICS AS CR, Prague, 2011.

[11] D. Dubois, H. Prade: Representation an combination of uncertainty with belief functions and possibility measures. *Computational Intelligence*, **4**: 244–264, 1988.

[12] D. Dubois, H. Prade: Consonant Approximations of Belief Functions. *International Journal of Approximate Reasoning* **4**: 419–449, 1990.

[13] R. Haenni: Aggregating Referee Scores: an Algebraic Approach. *COMSOC'08, 2nd International Workshop on Computational Social Choice*, Liverpool, UK, 2008.

[14] P. Hájek, T. Havránek, R. Jiroušek: *Uncertain Information Processing in Expert Systems.* CRC Press, Boca Raton, Florida, 1992.

[15] P. Hájek, J. J. Valdés: Generalized algebraic foundations of uncertainty processing in rule-based expert systems (dempsteroids). *Computers and Artif. Intell.* **10** (1): 29–42, 1991.

[16] W. Liu: Analysing the degree of conflict among belief functions. Artificial Intelligence **170**: 909–924, 2006.

[17] G. Shafer: *A Mathematical Theory of Evidence.* Princeton University Press, New Jersey, 1976.

[18] Ph. Smets: The combination of evidence in the transferable belief model. IEEE-Pattern analysis and Machine Intelligence **12**: 447–458, 1990.

[19] Ph. Smets: Analyzing the combination of conflicting belief functions. *Information Fusion* **8**: 387–412, 2007.

[20] J. J. Valdés: *Algebraic and logical foundations of uncertainty processing in rule-based expert systems of Artificial Intelligence.* PhD Thesis, Czechoslovak Academy of Sci., Prague, 1987.

[21] R. R. Yager: On the Demspter-Shafer framework and new combination rules. *Information Sciences*, **41**: 93–138, 1987.

State sequence prediction in imprecise hidden Markov models

Jasper De Bock and Gert de Cooman
SYSTeMS, Ghent University, Belgium
{jasper.debock,gert.decooman}@UGent.be

Abstract

We present an efficient exact algorithm for estimating state sequences from outputs (or observations) in imprecise hidden Markov models (iHMM), where both the uncertainty linking one state to the next, and that linking a state to its output, are represented using coherent lower previsions. The notion of independence we associate with the credal network representing the iHMM is that of epistemic irrelevance. We consider as best estimates for state sequences the (Walley–Sen) maximal sequences for the posterior joint state model (conditioned on the observed output sequence), associated with a gain function that is the indicator of the state sequence. This corresponds to (and generalises) finding the state sequence with the highest posterior probability in HMMs with precise transition and output probabilities (pHMMs). We argue that the computational complexity is at worst quadratic in the length of the Markov chain, cubic in the number of states, and essentially linear in the number of maximal state sequences. For binary iHMMs, we investigate experimentally how the number of maximal state sequences depends on the model parameters.

Keywords. Imprecise hidden Markov model, optimal state sequence, maximality, coherent lower prevision, credal network, epistemic irrelevance.

1 Introduction

In a recent paper on inference in credal networks [5], De Cooman et al. developed the so-called MePiCTIr[1] algorithm for coherently updating beliefs about a single node in the tree after instantiating any number of other nodes. The local uncertainty models associated with the nodes of the network are coherent lower previsions [10, 14], and the independence notion used to interpret the graphical structure is that of epistemic irrelevance [2, 14]. This algorithm is quite efficient—it is essentially linear in the number of nodes—but it has a number of limitations. First of all, it only works for very special graphical structures: trees. While this is a serious limitation, there are, nevertheless quite a number of models and applications that involve a tree structure. Amongst these, hidden Markov models (HMMs) are definitely the simplest, and perhaps also the most popular ones. But this brings us to the second limitation: MePiCTIr only allows updating of beliefs about a *single* node. Whereas one of the most important applications for, say, HMMs, involves finding the *sequence* of (hidden) states with the highest posterior probability after observing a sequence of outputs [11]. For HMMs with precise local transition and emission probabilities, there are quite efficient dynamic programming algorithms, such as Viterbi's [11, 13], for performing this task. For imprecise-probabilistic local models, such as coherent lower previsions, we know of no algorithm in the literature for which the computational complexity comes even close to that of Viterbi's.

In this paper, we take the first steps towards remedying this situation. We describe imprecise hidden Markov models as special cases of credal trees (a special case of credal networks) under epistemic irrelevance in Section 2. We show in particular how we can use the ideas underlying the MePiCTIr algorithm (independent natural extension and marginal extension) to construct a most conservative joint model from imprecise local transition and emission models, and derive a number of interesting and useful formulas from that construction. In Section 3 we explain how a sequence of observations leads to (a collection of) so-called maximal state sequences. Finding all of them seems a daunting task at first: it has a search space that grows exponentially in the length of the Markov chain. However, in Section 4 we use the basic formulas found in Section 2 to derive an appropriate version of Bellman's [1] Principle of Optimality, which allows for an exponential reduction of the search space. By using a number of additional tricks, we are able in Section 5 to devise an algorithm for finding all maximal state sequences that is essentially linear in the number of such maximal sequences, quadratic in the length of the chain, and cubic in the number of states. We perceive this complexity to be comparable to that of the Viterbi algorithm, especially after realising that the latter makes the simplifying step of resolving ties more or less arbitrarily in order to produce

[1] MePiCTIr: Message Passing in Credal Trees under Irrelevance.

only a single optimal state sequence. This is something we will not allow our algorithm to do, for reasons that should become clear further on. In Section 6, we consider the special case of binary iHMMs, and investigate experimentally how the number of maximal state sequences depends on the model parameters. We comment on the very interesting structures that emerge, and give an heuristic explanation for them. We show off the algorithm's efficiency in Section 7 by calculating the maximal sequences for a specific iHMM of length 100.

We assume that the reader has a good working knowledge of the theory of coherent lower previsions; see Ref. [14] for an in-depth study, and Ref. [10] for a recent survey.

2 Basic notions

A hidden Markov model can be depicted using the following probabilistic graphical model:

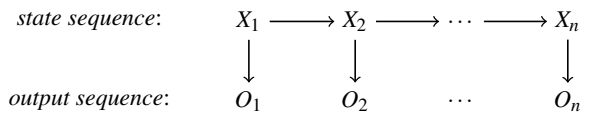


Figure 1: Tree representation of a hidden Markov model

Here n is some natural number. The *state variables* $X_1, \ldots, X_n$ assume values in the respective finite sets $\mathscr{X}_1, \ldots, \mathscr{X}_n$, and the *output variables* $O_1, \ldots, O_n$ assume values in the respective finite sets $\mathscr{O}_1, \ldots, \mathscr{O}_n$. We denote generic values of X_k by x_k, $\hat{x}_k$ or z_k, and generic values of O_k by o_k.

Local uncertainty models. We assume that we have the following local uncertainty models for these variables. For X_1, we have a *marginal* lower prevision $\underline{Q}_1$, defined on the set $\mathscr{G}(\mathscr{X}_1)$ of all real-valued maps (or *gambles*) on $\mathscr{X}_1$. For the subsequent states X_k, with $k \in \{2, \ldots, n\}$, we have a conditional lower prevision $\underline{Q}_k(\cdot|X_{k-1})$ defined on $\mathscr{G}(\mathscr{X}_k)$, called a *transition model*. In order to maintain uniformity of notation, we will also denote the marginal lower prevision $\underline{Q}_1$ as a conditional lower prevision $\underline{Q}_1(\cdot|X_0)$, where X_0 denotes a variable that may only assume a single value, and whose value is therefore certain. For any gamble f_k in $\mathscr{G}(\mathscr{X}_k)$, $\underline{Q}_k(f_k|X_{k-1})$ is interpreted as a gamble on $\mathscr{X}_{k-1}$, whose value $\underline{Q}_k(f_k|z_{k-1})$ in any $z_{k-1} \in \mathscr{X}_{k-1}$ is the lower prevision (or lower expectation) of the gamble $f_k(X_k)$, conditional on $X_{k-1} = z_{k-1}$.

In addition, for each output O_k, with $k \in \{1, \ldots, n\}$, we have a conditional lower prevision $\underline{S}_k(\cdot|X_k)$ defined on $\mathscr{G}(\mathscr{O}_k)$, called an *emission model*. For any gamble g_k in $\mathscr{G}(\mathscr{O}_k)$, $\underline{S}_k(g_k|X_k)$ is interpreted as a gamble on $\mathscr{X}_k$, whose value $\underline{S}_k(g_k|z_k)$ in any $z_k \in \mathscr{X}_k$ is the lower prevision (or lower expectation) of the gamble $g_k(O_k)$, conditional on $X_k = z_k$.

We take all these local (marginal, transition and emission) uncertainty models to be separately coherent; see for instance Ref. [5] for more details about such local uncertainty models and their separate coherence.

Interpretation of the graphical structure. We will assume that the tree in Fig. 1 represents the following irrelevance assessments: *conditional on its mother variable, the non-parent non-descendants of any variable in the tree are epistemically irrelevant to this variable and its descendants.* This is a weaker condition than the one usually associated with credal networks [3], which imposes strong independence rather than epistemic irrelevance. Recent work [5] has shown that using this weaker condition guarantees that an efficient algorithm exists for updating a credal *tree*, that is essentially linear in the number of nodes in the tree.

A joint uncertainty model. By applying the general analysis in Ref. [5] to the special case considered here, we find that the local uncertainty models can always be extended to a point-wise smallest (most conservative or least committal) coherent family of *joint* lower previsions $\underline{P}_k(\cdot|X_{k-1})$ on $\mathscr{G}(\mathscr{X}_{k:n} \times \mathscr{O}_{k:n})$, where $k \in \{1, \ldots, n\}$, $\mathscr{X}_{k:n} := \times_{i=k}^{n} \mathscr{X}_i$ and $\mathscr{O}_{k:n} := \times_{i=k}^{n} \mathscr{O}_i$. Again, for $k = 1$ the joint lower prevision $\underline{P}_1 = \underline{P}_1(\cdot|X_0)$ is effectively an unconditional lower prevision, defined on $\mathscr{G}(\mathscr{X}_{1:n} \times \mathscr{O}_{1:n})$. These joint lower previsions are given by the following recursion equations:

$$\underline{E}_k(\cdot|X_k) := \begin{cases} \underline{S}_n(\cdot|X_n) & k = n \\ \underline{S}_k(\cdot|X_k) \otimes \underline{P}_{k+1}(\cdot|X_k) & k = n-1, \ldots, 1 \end{cases} \tag{1}$$

and

$$\underline{P}_k(\cdot|X_{k-1}) := \underline{Q}_k(\underline{E}_k(\cdot|X_k)|X_{k-1}) \text{ for } k = n, \ldots, 1. \tag{2}$$

Eq. (1) states that, for $k = n-1, \ldots, 1$, the conditional lower prevision $\underline{E}_k(\cdot|X_k)$ on $\mathscr{G}(\mathscr{X}_{k+1:n} \times \mathscr{O}_{k:n})$ is the so-called (conditionally) *independent natural extension* [14, Chapter 9] of the conditional lower previsions $\underline{S}_k(\cdot|X_k)$ and $\underline{P}_{k+1}(\cdot|X_k)$, which was studied in detail in Ref. [6]. For our present purposes, it will suffice to recall from that study that such independent natural extensions are *factorising*, which implies in particular that

$$\begin{aligned} \underline{E}_k(fg|z_k) &= \underline{E}_k(g\underline{E}_k(f|z_k)|z_k) \\ &= \begin{cases} \underline{S}_k(g|z_k)\underline{P}_{k+1}(f|z_k) & \text{if } \underline{P}_{k+1}(f|z_k) \geq 0 \\ \overline{S}_k(g|z_k)\underline{P}_{k+1}(f|z_k) & \text{if } \underline{P}_{k+1}(f|z_k) \leq 0 \end{cases} \\ &= \overline{\underline{S}}_k(g|z_k) \odot \underline{P}_{k+1}(f|z_k), \end{aligned} \tag{3}$$

for all $z_k \in \mathscr{X}_k$, all $f \in \mathscr{G}(\mathscr{X}_{k+1:n} \times \mathscr{O}_{k+1:n})$ and all non-negative $g \in \mathscr{G}(\mathscr{O}_k)$, where $k \in \{1, \ldots, n-1\}$ (we call a gamble non-negative if all its values are). In this expression, we have used the shorthand notation $\overline{a} \odot b := \underline{a}\max\{0, b\} + \overline{a}\min\{0, b\}$.

Interesting lower and upper probabilities. Without too much trouble, we can use Eqs. (1)–(3) to derive the follow-

ing expressions for a number of interesting lower and upper probabilities:

$$\underline{P}_k(\{z_{k:n}\}|z_{k-1}) = \prod_{i=k}^{n} \underline{Q}_i(\{z_i\}|z_{i-1})$$
$$\overline{P}_k(\{z_{k:n}\}|z_{k-1}) = \prod_{i=k}^{n} \overline{Q}_i(\{z_i\}|z_{i-1}),$$

and

$$\underline{P}_k(\{z_{k:n}\} \times \{o_{k:n}\}|z_{k-1}) = \prod_{i=k}^{n} \underline{S}_i(\{o_i\}|z_i)\underline{Q}_i(\{z_i\}|z_{i-1}) \tag{4}$$

$$\overline{P}_k(\{z_{k:n}\} \times \{o_{k:n}\}|z_{k-1}) = \prod_{i=k}^{n} \overline{S}_i(\{o_i\}|z_i)\overline{Q}_i(\{z_i\}|z_{i-1}), \tag{5}$$

for $k = \{1,\dots,n\}$. We will assume throughout that

$\underline{P}_1(\{z_{1:n}\} \times \{o_{1:n}\}) > 0$ for all $z_{1:n} \in \mathscr{X}_{1:n}$ and $o_{1:n} \in \mathscr{O}_{1:n}$

or equivalently, that all *local lower previsions are positive* [5], in the sense that

$$\underline{Q}_k(\{z_k\}|z_{k-1}) > 0 \text{ and } \underline{S}_k(\{o_k\}|z_k) > 0$$

for all $z_{k-1} \in \mathscr{X}_{k-1}$, $z_k \in \mathscr{X}_k$ and $o_k \in \mathscr{O}_k$, $k \in \{1,\dots,n\}$. This implies in particular that $\underline{P}_k(\{o_{k:n}\}|z_{k-1}) > 0$ for all $k \in \{1,\dots,n\}$, $z_{k-1} \in \mathscr{X}_{k-1}$ and $o_{k:n} \in \mathscr{O}_{k:n}$.

We have good reason to believe that our results remain valid, *mutatis mutandis*, on the weaker condition that all local *upper* previsions should be positive, and we intend to deal with this issue in further work.

3 Estimating states from outputs

In a hidden Markov model, the states are not directly observable, but the outputs are, and the general aim is to use the outputs to estimate the states. In the present paper, we concentrate on the following problem: *Suppose we have observed the output sequence $o_{1:n}$, estimate the state sequence $x_{1:n}$.* We will use an essentially Bayesian approach to do so, but need to allow for the fact that we are working with imprecise rather than precise probability models.

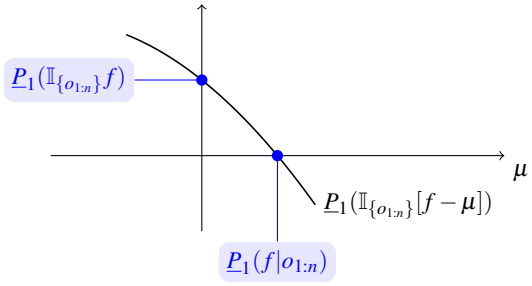
Updating the iHMM. The first step in our approach consists in updating (or conditioning) the joint model $\underline{P}_1$ on the observed outputs $O_{1:n} = o_{1:n}$. Given our positivity assumptions on the local lower prevision, we see that the lower probability $\underline{P}_1(\{o_{1:n}\})$ of the conditioning event $\{o_{1:n}\}$ is strictly positive. This implies [5] that there is only one coherent way to perform this updating, namely using the Generalised Bayes Rule [14], which reduces to Bayes's Rule when all local models are precise. We are thus led to consider the updated lower prevision $\underline{P}_1(\cdot|o_{1:n})$ on $\mathscr{G}(\mathscr{X}_{1:n})$, given by

$$\underline{P}_1(f|o_{1:n}) := \max\left\{\mu \in \mathbb{R} \colon \underline{P}_1(\mathbb{I}_{\{o_{1:n}\}}[f-\mu]) \geq 0\right\}, \tag{6}$$

for all gambles f on $\mathscr{X}_{1:n}$. Using the coherence of $\underline{P}_1$, it is not too hard to prove that when $\underline{P}_1(\{o_{1:n}\}) > 0$, $\underline{P}_1(\mathbb{I}_{\{o_{1:n}\}}[f-\mu])$ constitutes a strictly decreasing and continuous function of μ, which therefore has a unique zero. As a consequence, we have for any $f \in \mathscr{G}(\mathscr{X}_{1:n})$ that

$$\begin{aligned}\underline{P}_1(f|o_{1:n}) \leq 0 &\Leftrightarrow (\forall \mu > 0)\underline{P}_1(\mathbb{I}_{\{o_{1:n}\}}[f-\mu]) < 0\\ &\Leftrightarrow \underline{P}_1(\mathbb{I}_{\{o_{1:n}\}}f) \leq 0.\end{aligned} \tag{7}$$

In fact, it is not hard to infer from the strictly decreasing and continuous character of $\underline{P}_1(\mathbb{I}_{\{o_{1:n}\}}[f-\mu])$ that $\underline{P}_1(f|o_{1:n})$ and $\underline{P}_1(\mathbb{I}_{\{o_{1:n}\}}f)$ have the same sign. They are either both negative, both positive or both equal to zero; see also the illustration below.



Maximal state sequences. The next step consists in using the posterior model $\underline{P}_1(\cdot|o_{1:n})$ to find best estimates for the state sequence $x_{1:n}$. On the Bayesian approach, this is usually done by solving a decision-making, or optimisation, problem: we associate a gain function $\mathbb{I}_{\{x_{1:n}\}}$ with every candidate state sequence $x_{1:n}$, and select as best estimates those state sequences $\hat{x}_{1:n}$ that maximise the expected gain, resulting in state sequences with maximal posterior probability.

Here we generalise this decision-making approach towards working with imprecise probability models. The criterion we use to decide which estimates are optimal for the given gain functions is that of (Walley–Sen) *maximality* [12, 14]. Maximality has a number of very desirable properties that make sure it works well in optimisation contexts [7, 9], and it is well-justified from a behavioural point of view, as we shall see presently.

We can express a strict preference $\succ$ between two state sequence estimates $\hat{x}_{1:n}$ and $x_{1:n}$ as follows:

$$\hat{x}_{1:n} \succ x_{1:n} \Leftrightarrow \underline{P}_1(\mathbb{I}_{\{\hat{x}_{1:n}\}} - \mathbb{I}_{\{x_{1:n}\}}|o_{1:n}) > 0.$$

On a behavioural interpretation, this expresses that a subject with lower prevision $\underline{P}_1(\cdot|o_{1:n})$ is disposed to pay some strictly positive amount of utility to replace the (gain associated with the) estimate $x_{1:n}$ with the (gain associated with the) estimate $\hat{x}_{1:n}$; see Ref. [14, Section 3.9]. This induces a strict partial order $\succ$ [an irreflexive and transitive binary relation] on the set of state sequences $\mathscr{X}_{1:n}$, and we consider an estimate $\hat{x}_{1:n}$ to be optimal when it is undominated, or maximal, in this strict partial order:

$$\begin{aligned}
&\hat{x}_{1:n} \in \operatorname{opt}(\mathscr{X}_{1:n}|o_{1:n}) \\
&\quad \Leftrightarrow (\forall x_{1:n} \in \mathscr{X}_{1:n}) x_{1:n} \not> \hat{x}_{1:n} \\
&\quad \Leftrightarrow (\forall x_{1:n} \in \mathscr{X}_{1:n}) \underline{P}_1(\mathbb{I}_{\{x_{1:n}\}} - \mathbb{I}_{\{\hat{x}_{1:n}\}}|o_{1:n}) \leq 0 \qquad (8) \\
&\quad \Leftrightarrow (\forall x_{1:n} \in \mathscr{X}_{1:n}) \underline{P}_1(\mathbb{I}_{\{o_{1:n}\}}[\mathbb{I}_{\{x_{1:n}\}} - \mathbb{I}_{\{\hat{x}_{1:n}\}}]) \leq 0,
\end{aligned}$$

where the last equivalence follows from Eq. (7). *In summary then, the aim of this paper is to develop an efficient algorithm for finding the set of maximal estimates* $\operatorname{opt}(\mathscr{X}_{1:n}|o_{1:n})$.

Another approach, which we will not consider here, could consist in trying to find the so-called *maximin* state sequences $\overline{x}_{1:n}$, which maximise the posterior lower probability:

$$\overline{x}_{1:n} \in \operatorname{argmax}_{x_{1:n} \in \mathscr{X}_{1:n}} \underline{P}_1(\{x_{1:n}\}|o_{1:n})$$

While it is well known that any such maximin sequence is in particular guaranteed to also be a maximal sequence, finding such maximin sequences seems to be a much more complicated affair.[2]

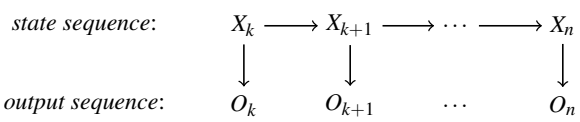
More general optimality operators. We shall see below that in order to find the set of maximal estimates, it is useful to consider a more general collection of 'optimality operators': for any $k \in \{1,\ldots,n\}$ and $z_{k-1} \in \mathscr{X}_{k-1}$, we define the optimality operator

$$\operatorname{opt}(\cdot|z_{k-1}, o_{k:n}) : \mathscr{P}(\mathscr{X}_{k:n}) \to \mathscr{P}(\mathscr{X}_{k:n})$$

such that for all $S \in \mathscr{P}(\mathscr{X}_{k:n})$, or in other words $S \subseteq \mathscr{X}_{k:n}$, and all $\hat{x}_{k:n} \in S$:

$$\begin{aligned}
&\hat{x}_{k:n} \in \operatorname{opt}(S|z_{k-1}, o_{k:n}) \\
&\Leftrightarrow (\forall x_{k:n} \in S) \underline{P}_k(\mathbb{I}_{\{o_{k:n}\}}[\mathbb{I}_{\{x_{k:n}\}} - \mathbb{I}_{\{\hat{x}_{k:n}\}}]|z_{k-1}) \leq 0. \qquad (9)
\end{aligned}$$

The interpretation of these operators is immediate: consider the following part of the original iHMM:



where we take $\underline{Q}_k(\cdot|z_{k-1})$ as the marginal model for the first state X_k. Then the corresponding joint lower prevision on $\mathscr{G}(\mathscr{X}_{k:n} \times \mathscr{O}_{k:n})$ is precisely $\underline{P}_k(\cdot|z_{k-1})$, and if we have a sequence of outputs $o_{k:n}$, then $\operatorname{opt}(\cdot|z_{k-1}, o_{k:n})$ selects from a set $S \subseteq \mathscr{X}_{k:n}$ those state sequence estimates that are undominated by any other estimate in S. It should be clear that the set $\operatorname{opt}(\mathscr{X}_{1:n}|o_{1:n})$ we are eventually looking for, can also be written as $\operatorname{opt}(\mathscr{X}_{1:n}|z_0, o_{1:n})$.

[2] Private communication from Cassio de Campos. Of course, once we know all maximal solutions, we could determine which of them are the maximin solutions by comparing their posterior lower probabilities. As far as we can see now, calculating these does not seem a trivial task.

Useful recursion equations. Fix any k in $\{1,\ldots,n\}$. If we look at Eq. (9), we see that it will be useful to derive a manageable expression for $\underline{P}_k(\Delta[x_{k:n}, \hat{x}_{k:n}]|z_{k-1})$, where $\Delta[x_{k:n}, \hat{x}_{k:n}]$ is the gamble on $\mathscr{X}_{k:n} \times \mathscr{O}_{k:n}$ given by:

$$\Delta[x_{k:n}, \hat{x}_{k:n}] := \mathbb{I}_{\{o_{k:n}\}}[\mathbb{I}_{\{x_{k:n}\}} - \mathbb{I}_{\{\hat{x}_{k:n}\}}].$$

Using Eqs. (1)–(5) together with a few algebraic manipulations, we can derive the following equations for $\underline{P}_k(\Delta[x_{k:n}, \hat{x}_{k:n}]|z_{k-1})$:

If $k \in \{1,\ldots,n-1\}$ and $\hat{x}_k = x_k$ then, with some fairly obvious abuse of notation:

$$\begin{aligned}
\underline{P}_k(\Delta[x_{k:n}, \hat{x}_{k:n}]|z_{k-1}) = \overline{\underline{Q}}_k(\{x_k\}|z_{k-1})\overline{\underline{S}}_k(\{o_k\}|x_k) \qquad (10) \\
\odot \underline{P}_{k+1}(\Delta[x_{k+1:n}, \hat{x}_{k+1:n}]|x_k).
\end{aligned}$$

If $\hat{x}_n = x_n$ then

$$\underline{P}_n(\Delta[x_n, \hat{x}_n]|z_{n-1}) = 0. \qquad (11)$$

If $k \in \{1,\ldots,n\}$ and $\hat{x}_k \neq x_k$ then

$$\begin{aligned}
&\underline{P}_k(\Delta[x_{k:n}, \hat{x}_{k:n}]|z_{k-1}) \\
&\qquad = \underline{Q}_k(\mathbb{I}_{\{x_k\}}\beta(x_{k:n}) - \mathbb{I}_{\{\hat{x}_k\}}\alpha(\hat{x}_{k:n})|z_{k-1}), \qquad (12)
\end{aligned}$$

where we define, for any $z_{k:n} \in \mathscr{X}_{k:n}$:

$$\begin{aligned}
\beta(z_{k:n}) &:= \underline{S}_k(\{o_k\}|z_k) \prod_{i=k+1}^{n} \underline{S}_i(\{o_i\}|z_i)\underline{Q}_i(\{z_i\}|z_{i-1}) \\
\alpha(z_{k:n}) &:= \overline{S}_k(\{o_k\}|z_k) \prod_{i=k+1}^{n} \overline{S}_i(\{o_i\}|z_i)\overline{Q}_i(\{z_i\}|z_{i-1}).
\end{aligned}$$

For any given sequence of states $z_{k:n} \in \mathscr{X}_{k:n}$, the $\alpha(z_{k:n})$ and $\beta(z_{k:n})$ can be found by simple backward recursion:

$$\alpha(z_{k:n}) = \alpha(z_{k+1:n})\overline{S}_k(\{o_k\}|z_k)\overline{Q}_{k+1}(\{z_{k+1}\}|z_k) \qquad (13)$$

$$\beta(z_{k:n}) = \beta(z_{k+1:n})\underline{S}_k(\{o_k\}|z_k)\underline{Q}_{k+1}(\{z_{k+1}\}|z_k), \qquad (14)$$

for $k \in \{1,\ldots,n-1\}$, and starting from:

$$\begin{aligned}
\alpha(z_{n:n}) &= \alpha(z_n) = \overline{S}_n(\{o_n\}|z_n) \\
\beta(z_{n:n}) &= \beta(z_n) = \underline{S}_n(\{o_n\}|z_n).
\end{aligned}$$

4 The Principle of Optimality

Determining the state sequences in $\operatorname{opt}(\mathscr{X}_{1:n}|o_{1:n})$ directly using Eq. (8) clearly has exponential complexity (in the length of the chain). We are now going to take a dynamic programming approach [1] to reducing this complexity by deriving a recursion equation for the optimality operators $\operatorname{opt}(\cdot|z_{k-1}, o_{k:n})$.

Theorem (Principle of Optimality)**.** *For $k \in \{1,\ldots,n-1\}$, all $z_{k-1} \in \mathscr{X}_{k-1}$ and all $\hat{x}_{k:n} \in \mathscr{X}_{k:n}$:*

$$\begin{aligned}
&\hat{x}_{k:n} \in \operatorname{opt}(\mathscr{X}_{k:n}|z_{k-1}, o_{k:n}) \\
&\qquad \Rightarrow \hat{x}_{k+1:n} \in \operatorname{opt}(\mathscr{X}_{k+1:n}|\hat{x}_k, o_{k+1:n}).
\end{aligned}$$

Proof. Fix $k \in \{1,\dots,n-1\}$, $z_{k-1} \in \mathscr{Z}_{k-1}$ and $\hat{x}_{k:n} \in \mathscr{X}_{k:n}$. Assume that $\hat{x}_{k+1:n} \notin \operatorname{opt}(\mathscr{X}_{k+1:n}|\hat{x}_k, o_{k+1:n})$, then we show that $\hat{x}_{k:n} \notin \operatorname{opt}(\mathscr{X}_{k:n}|z_{k-1}, o_{k:n})$. It follows from the assumption that there is some $x_{k+1:n} \in \mathscr{X}_{k+1}$ such that $\underline{P}_{k+1}(\Delta[x_{k+1:n}, \hat{x}_{k+1:n}]|\hat{x}_k) > 0$. Now prefix the state sequence $x_{k+1:n}$ with the state $\hat{x}_k$ to form the state sequence $x_{k:n}$, implying that $\hat{x}_k = x_k$. We then infer from Eq. (10) that

$$\begin{aligned}&\underline{P}_k(\Delta[x_{k:n}, \hat{x}_{k:n}]|z_{k-1})\\ &= \underline{Q}_k(\{\hat{x}_k\}|z_{k-1})\underline{S}_k(\{o_k\}|\hat{x}_k)\underline{P}_{k+1}(\Delta[x_{k+1:n}, \hat{x}_{k+1:n}]|\hat{x}_k)\\ &> 0,\end{aligned}$$

which tells us that indeed $\hat{x}_{k:n} \notin \operatorname{opt}(\mathscr{X}_{k:n}|z_{k-1}, o_{k:n})$. □

As an immediate consequence, we find that

$$\operatorname{opt}(\mathscr{X}_{k:n}|z_{k-1}, o_{k:n}) \subseteq \bigcup_{z_k \in \mathscr{X}_k} z_k \oplus \operatorname{opt}(\mathscr{X}_{k+1:n}|z_k, o_{k+1:n}), \tag{15}$$

where $\oplus$ denotes concatenation of state sequences. From this we can infer that

$$\begin{aligned}&\operatorname{opt}(\mathscr{X}_{k:n}|z_{k-1}, o_{k:n})\\ &= \operatorname{opt}\left(\bigcup_{z_k \in \mathscr{X}_k} z_k \oplus \operatorname{opt}(\mathscr{X}_{k+1:n}|z_k, o_{k+1:n}) \,\middle|\, z_{k-1}, o_{k:n}\right),\end{aligned} \tag{16}$$

since the optimality operator selecting the maximal elements in a strict partial order is insensitive to the omission of non-optimal elements; see Ref. [7] for a detailed discussion. While Eq. (16) clearly exhibits the reduction in computational complexity that the Principle of Optimality allows for, it is perhaps useful to point out here that we will not use this specific form for it in our algorithm.

5 An algorithm for finding maximal state sequences

Instead, we use Eq. (15) to devise an algorithm for constructing the set $\operatorname{opt}(\mathscr{X}_{1:n}|o_{1:n})$ of maximal state sequences in a recursive manner.

Initial set-up using backward recursion. We begin by defining a few auxiliary notions. First of all, we consider the thresholds:

$$\begin{aligned}&\theta_k(\hat{x}_k, x_k|z_{k-1})\\ &\quad := \min\left\{a \in \mathbb{R}\colon \underline{Q}_k(\mathbb{I}_{\{x_k\}} - a\mathbb{I}_{\{\hat{x}_k\}}|z_{k-1}) \leq 0\right\}\end{aligned} \tag{17}$$

for all $k \in \{1,\dots,n\}$, $z_{k-1} \in \mathscr{Z}_{k-1}$ and $x_k, \hat{x}_k \in \mathscr{X}_k$. Observe that it follows from the positivity assumptions on the $\underline{Q}_k(\cdot|X_{k-1})$ that $\theta_k(\hat{x}_k, x_k|z_{k-1}) > 0$.

Next, we define

$$\alpha_k^{\max}(x_k) := \max_{\substack{z_{k:n} \in \mathscr{X}_{k:n}\\ z_k = x_k}} \alpha(z_{k:n}) \tag{18}$$

and

$$\beta_k^{\max}(x_k) := \max_{\substack{z_{k:n} \in \mathscr{X}_{k:n}\\ z_k = x_k}} \beta(z_{k:n}) \tag{19}$$

for all $k \in \{1,\dots,n\}$ and $x_k \in \mathscr{X}_k$. Using Eq. (13)–(14), these can be calculated efficiently using the following backward recursive (dynamic programming) procedure:

$$\begin{aligned}&\alpha_k^{\max}(x_k)\\ &= \max_{z_{k+1} \in \mathscr{X}_{k+1}} \alpha_{k+1}^{\max}(z_{k+1})\overline{S}_k(\{o_k\}|x_k)\overline{Q}_{k+1}(\{z_{k+1}\}|x_k)\\ &= \overline{S}_k(\{o_k\}|x_k) \max_{z_{k+1} \in \mathscr{X}_{k+1}} \alpha_{k+1}^{\max}(z_{k+1})\overline{Q}_{k+1}(\{z_{k+1}\}|x_k),\end{aligned} \tag{20}$$

and

$$\begin{aligned}&\beta_k^{\max}(x_k)\\ &= \max_{z_{k+1} \in \mathscr{X}_{k+1}} \beta_{k+1}^{\max}(z_{k+1})\underline{S}_k(\{o_k\}|x_k)\underline{Q}_{k+1}(\{z_{k+1}\}|x_k)\\ &= \underline{S}_k(\{o_k\}|x_k) \max_{z_{k+1} \in \mathscr{X}_{k+1}} \beta_{k+1}^{\max}(z_{k+1})\underline{Q}_{k+1}(\{z_{k+1}\}|x_k),\end{aligned} \tag{21}$$

for $k \in \{1,\dots,n-1\}$, starting from

$$\alpha_n^{\max}(x_n) = \alpha(x_n) = \overline{S}_n(\{o_n\}|x_n) \tag{22}$$

and

$$\beta_n^{\max}(x_n) = \beta(x_n) = \underline{S}_n(\{o_n\}|x_n). \tag{23}$$

Finally, we let

$$\alpha_k^{\operatorname{opt}}(\hat{x}_k|z_{k-1}) := \max_{\substack{x_k \in \mathscr{X}_k\\ x_k \neq \hat{x}_k}} \beta_k^{\max}(x_k)\theta_k(\hat{x}_k, x_k|z_{k-1}), \tag{24}$$

for all $k \in \{1,\dots,n\}$, $z_{k-1} \in \mathscr{X}_{k-1}$ and $\hat{x}_k \in \mathscr{X}_k$.

Reformulation of the optimality condition. First, we consider $k = n$. For every $z_{n-1} \in \mathscr{X}_{n-1}$, we determine $\operatorname{opt}(\mathscr{X}_n|z_{n-1}, o_n)$ as the set of those elements $\hat{x}_n$ of $\mathscr{X}_n$ for which

$$(\forall x_n \in \mathscr{X}_n \setminus \{\hat{x}_n\})\underline{Q}_n(\mathbb{I}_{\{x_n\}}\beta(x_n) - \mathbb{I}_{\{\hat{x}_n\}}\alpha(\hat{x}_n)|z_{n-1}) \leq 0,$$

as this condition is equivalent to condition (9) for $k = n$, considering Eqs. (11) and (12). But this condition is also equivalent to

$$(\forall x_n \in \mathscr{X}_n \setminus \{\hat{x}_n\})\frac{\alpha(\hat{x}_n)}{\beta_n^{\max}(x_n)} \geq \theta_n(\hat{x}_n, x_n|z_{n-1}),$$

considering Eqs. (23) and (17). Eq. (24) now tells us that this is equivalent to $\alpha(\hat{x}_n) \geq \alpha_n^{\operatorname{opt}}(\hat{x}_n|z_{n-1})$. In summary,

$$\operatorname{opt}(\mathscr{X}_n|z_{n-1}, o_n) = \left\{\hat{x}_n \in \mathscr{X}_n\colon \alpha(\hat{x}_n) \geq \alpha_n^{\operatorname{opt}}(\hat{x}_n|z_{n-1})\right\}. \tag{25}$$

Next, we consider any $k \in \{1,\dots,n-1\}$. Fix $z_{k-1} \in \mathscr{Z}_{k-1}$, then we must determine $\operatorname{opt}(\mathscr{X}_{k:n}|z_{k-1},o_{k:n})$. We know from the Principle of Optimality (15) that we can limit the candidate optimal sequences $\hat{x}_{k:n}$ to the set

$$\bigcup_{z_k \in \mathscr{Z}_k} z_k \oplus \operatorname{opt}(\mathscr{X}_{k+1:n}|z_k,o_{k+1:n}). \tag{26}$$

Consider any such $\hat{x}_{k:n}$, then we must check for any $x_{k:n} \in \mathscr{X}_{k:n}$ whether $\underline{P}_k(\Delta[x_{k:n},\hat{x}_{k:n}]|z_{k-1}) \leq 0$; see Eq. (9). But if $x_{k:n}$ is such that $x_k = \hat{x}_k$, then it follows from Eq. (10) that $\underline{P}_k(\Delta[x_{k:n},\hat{x}_{k:n}]|z_{k-1}) \leq 0$, because the fact that $\hat{x}_{k+1:n} \in \operatorname{opt}(\mathscr{X}_{k+1:n}|\hat{x}_k,o_{k+1:n})$ also guarantees that $\underline{P}_{k+1}(\Delta[x_{k+1:n},\hat{x}_{k+1:n}]|\hat{x}_k) \leq 0$. So we can limit ourselves to checking the inequality for $x_{k:n}$ for which $x_k \neq \hat{x}_k$.

So fix any $x_k \neq \hat{x}_k$, then we must check whether

$$(\forall x_{k+1:n} \in \mathscr{X}_{k+1:n}) \quad \underline{Q}_k(\mathbb{I}_{\{x_k\}}\beta(x_{k:n}) - \mathbb{I}_{\{\hat{x}_k\}}\alpha(\hat{x}_{k:n})|z_{k-1}) \leq 0;$$

see Eq. (12). Considering Eq. (19), this is equivalent to

$$\underline{Q}_k(\mathbb{I}_{\{x_k\}}\beta_k^{\max}(x_k) - \mathbb{I}_{\{\hat{x}_k\}}\alpha(\hat{x}_{k:n})|z_{k-1}) \leq 0,$$

and therefore also equivalent to

$$\frac{\alpha(\hat{x}_{k:n})}{\beta_k^{\max}(x_k)} \geq \theta_k(\hat{x}_k,x_k|z_{k-1}),$$

considering Eq. (17). Since this inequality must hold for every $x_k \neq \hat{x}_k$, we infer from Eq. (24) that we must have that

$$\alpha(\hat{x}_{k:n}) \geq \alpha_k^{\operatorname{opt}}(\hat{x}_k|z_{k-1}). \tag{27}$$

So we must check this condition for all the candidate sequences $\hat{x}_{k:n}$ in the set (26). We can do this efficiently by using the following backward-forward recursion approach.

Backward-forward recursion. We start by letting k run *backward* from n to 1.

For $k = n$, it is a straightforward matter to determine $\operatorname{opt}(\mathscr{X}_n|z_{n-1},o_n)$ for every $z_{n-1} \in \mathscr{Z}_{n-1}$ using Eq. (25).

For each $k < n$, we now show how we can determine $\operatorname{opt}(\mathscr{X}_k|z_{k-1},o_{k:n})$ by executing the following *forward running* procedure for every $z_{k-1} \in \mathscr{Z}_{k-1}$.

If we combine Eqs. (27) and (18), we see that a necessary condition for $\hat{x}_k$ to be the state at time k in some optimal state sequence in $\operatorname{opt}(\mathscr{X}_k|z_{k-1},o_{k:n})$ is that

$$\alpha_k^{\max}(\hat{x}_k) \geq \alpha_k^{\operatorname{opt}}(\hat{x}_k|z_{k-1}), \tag{28}$$

meaning we can eliminate from our search those sequences for which the first state $\hat{x}_k$ does not satisfy this condition. On the other hand, for any $\hat{x}_k$ that satisfies the condition (28), we know from Eq. (18) that there is at least one state sequence with first state $\hat{x}_k$ that satisfies the condition (27).

So now we consider any $\hat{x}_k$ that satisfies the condition (28), and any $\hat{x}_{k+1}$ that is a first state in some optimal sequence in $\operatorname{opt}(\mathscr{X}_{k+1}|\hat{x}_k,o_{k+1:n})$. Observe that we can determine whether $\hat{x}_{k+1}$ satisfies this condition, because we have determined $\operatorname{opt}(\mathscr{X}_{k+1}|\hat{x}_k,o_{k+1:n})$ in the forward run for $k+1$.

Taking into account the recursion equation (13), we see that the condition (27) is equivalent to

$$\alpha(\hat{x}_{k+1:n}) \geq \alpha^{\operatorname{opt}}(\hat{x}_{k:k+1}|z_{k-1}), \tag{29}$$

where

$$\alpha^{\operatorname{opt}}(\hat{x}_{k:k+1}|z_{k-1}) := \frac{\alpha_k^{\operatorname{opt}}(\hat{x}_k|z_{k-1})}{\overline{S}_k(\{o_k\}|\hat{x}_k)\overline{Q}_{k+1}(\{\hat{x}_{k+1}\}|\hat{x}_k)}.$$

So if we combine Eqs. (29) and (18), we see that a necessary condition for $\hat{x}_{k+1}$ to be a state at time $k+1$ in some optimal sequence starting with $\hat{x}_k$ is that

$$\alpha_{k+1}^{\max}(\hat{x}_{k+1}) \geq \alpha^{\operatorname{opt}}(\hat{x}_{k:k+1}|z_{k-1}), \tag{30}$$

meaning we can eliminate from our search those sequences in $\operatorname{opt}(\mathscr{X}_{k+1}|\hat{x}_k,o_{k+1:n})$ for which the first state $\hat{x}_{k+1}$ does not satisfy this condition. On the other hand, for any $\hat{x}_{k+1}$ that satisfies the condition (30), we know from Eq. (18) [for $k+1$] that there is at least one state sequence in $\operatorname{opt}(\mathscr{X}_{k+1}|\hat{x}_k,o_{k+1:n})$ with first state $\hat{x}_{k+1}$ that satisfies the condition (29).

Next, we consider any $\hat{x}_k$ and $\hat{x}_{k+1}$ that satisfy the condition (30) and any $\hat{x}_{k+2}$ for which $\hat{x}_{k+1}$ and $\hat{x}_{k+2}$ are the first two states in some optimal sequence in $\operatorname{opt}(\mathscr{X}_{k+1}|\hat{x}_k,o_{k+1:n})$. Taking into account the recursion equation (13), we see that the condition (27) is equivalent to

$$\alpha(\hat{x}_{k+2:n}) \geq \alpha^{\operatorname{opt}}(\hat{x}_{k:k+2}|z_{k-1}), \tag{31}$$

where

$$\alpha^{\operatorname{opt}}(\hat{x}_{k:k+2}|z_{k-1}) := \frac{\alpha^{\operatorname{opt}}(\hat{x}_{k:k+1}|z_{k-1})}{\overline{S}_{k+1}(\{o_{k+1}\}|\hat{x}_{k+1})\overline{Q}_{k+2}(\{\hat{x}_{k+2}\}|\hat{x}_{k+1})}.$$

So if we combine Eqs. (31) and (18), we see that a necessary condition for $\hat{x}_{k+2}$ to be a state at time $k+2$ in some optimal sequence starting with $\hat{x}_{k:k+1}$ is that

$$\alpha_{k+2}^{\max}(\hat{x}_{k+2}) \geq \alpha^{\operatorname{opt}}(\hat{x}_{k:k+2}|z_{k-1}), \tag{32}$$

meaning we can eliminate from our search those sequences in $\operatorname{opt}(\mathscr{X}_{k+1}|\hat{x}_k,o_{k+1:n})$ for which the second state $\hat{x}_{k+2}$ does not satisfy this condition. On the other hand, for any $\hat{x}_{k+2}$ that satisfies the condition (32), there is at least one state sequence in $\operatorname{opt}(\mathscr{X}_{k+1}|\hat{x}_k,o_{k+1:n})$ with a second state $\hat{x}_{k+2}$ that satisfies the condition (31).

It should be clear that we can go forward in this way until we reach time n, and that in doing so we construct all the sequences $\hat{x}_{k:n}$ in $\operatorname{opt}(\mathscr{X}_k|z_{k-1},o_{k:n})$.

A brief discussion of the algorithm's complexity. We begin with the preparatory calculations of the quantities in Eqs. (17)–(24). For the thresholds $\theta_k(\hat{x}_k, x_k|z_{k-1})$ in Eq. (17), the computational complexity is clearly cubic in the number of states, and linear in the number of nodes. Calculating the $\alpha_k^{\max}(x_k)$ and $\beta_k^{\max}(x_k)$ in Eqs. (20) and (21) is linear in the number of nodes, and quadratic in the number of states. The complexity of finding the $\alpha_k^{\text{opt}}(\hat{x}_k|z_{k-1})$ in Eq. (24) is linear in the number of nodes, and cubic in the number of states.

On the other hand, the computational complexity of the backward-forward loop is clearly quadratic in the number of nodes, quadratic in the number of states, and roughly speaking linear in the number of maximal sequences.[3]

For precise HMMs, the state sequence estimation problem can be solved very efficiently by the Viterbi algorithm [11, 13], whose complexity is linear in the number of nodes, and quadratic in the number of states. However, this algorithm only emits a single optimal (most probable) state sequence, even in cases where there are multiple (equally probable) optimal solutions: this of course simplifies the problem. If we would content ourselves with giving only a single maximal solution, the ensuing algorithm would have a complexity that is similar to Viterbi's. So, to allow for a fair comparison between Viterbi's algorithm and ours, we would need to alter Viterbi's algorithm in such a way that it no longer resolves ties arbitrarily, and emits all (equally probable) optimal state sequences. This new version will remain linear in the number of nodes, and quadratic in the number of states, but emitting the optimal sequences will be linear in the number of them. For the complexity for the most time-consuming part of our algorithm (the backward-forward loop), the only difference is this: Viterbi's approach is linear and ours quadratic in the number of nodes. Where does this difference come from? In iHMMs we have mutually incomparable solutions, whereas in pHMMs the optimal solutions are indifferent, or equally probable. This makes sure that the algorithm for pHMMs requires no forward loops. We believe that this added complexity is a reasonable price to pay for the robustness that working with imprecise-probabilistic models offers.

Additional comments. All that is needed in order to produce the α- and β-functions are assessments for the lower and upper transition and emission mass functions:

$$\underline{Q}_k(\{z_k\}|z_{k-1}), \overline{Q}_k(\{z_k\}|z_{k-1}), \underline{S}_k(\{o_k\}|z_k), \overline{S}_k(\{o_k\}|z_k)$$

for all $k \in \{1,\ldots,n\}$, $z_{k-1} \in \mathscr{X}_{k-1}$, $z_k \in \mathscr{X}_k$ and $o_k \in \mathscr{O}_k$. The most conservative coherent models $\underline{Q}_k(\cdot|X_{k-1})$ that correspond to such assessments are 2-monotone [4, 8]. Due to their comonotone additivity, this implies that:

$$\underline{Q}_k(\mathbb{I}_{\{x_k\}} - a\mathbb{I}_{\{\hat{x}_k\}}|z_{k-1}) = \underline{Q}_k(\{x_k\}|z_{k-1}) - a\overline{Q}_k(\{\hat{x}_k\}|z_{k-1})$$

for all $a \geq 0$, and therefore Eq. (17) leads to

$$\theta_k(\hat{x}_k, x_k|z_{k-1}) = \frac{\underline{Q}_k(\{x_k\}|z_{k-1})}{\overline{Q}_k(\{\hat{x}_k\}|z_{k-1})}. \tag{33}$$

The right-hand side is the smallest possible value of the threshold $\theta_k(\hat{x}_k, x_k|z_{k-1})$ corresponding to the assessments $\underline{Q}_k(\{x_k\}|z_{k-1})$ and $\overline{Q}_k(\{\hat{x}_k\}|z_{k-1})$, leading to the most conservative inferences, and therefore the largest possible sets of maximal sequences, that correspond to these assessments.

6 Some experiments

While a linear complexity in the number of maximal sequences is probably the best we can hope for, we also see that we will only be able to find all maximal sequences efficiently provided their number is reasonably small. Should it, say, tend to increase exponentially with the length of the chain, then no algorithm, however cleverly designed, could overcome this hurdle. Because this number of maximal sequences is so important, we study its behaviour in more detail. In order to do so, we take a closer look at how this number of maximal sequences depends on the transition probabilities of the model, and how it evolves when we let the imprecision of the local models grow. We shall see that this number displays very interesting behaviour that can be explained, and even predicted to some extent. To allow for easy visualisation, we limit this discussion to binary iHMMs, where both the state and output variables can assume only two possible values, say 0 and 1.

Describing a binary stationary iHMM. We first consider a binary stationary HMM. The (precise) transition probabilities for going from one state to the next are completely determined by numbers in the unit interval: the probability p to go from state 0 to state 0, and the probability q to go from state 1 to state 0. To further pin down the HMM we also need to specify the (marginal) probability m for the first state to be 0, and the two emission probabilities: the probability r of emitting output 0 from state 0 and the probability s of emitting output 0 from state 1.

In this binary case, all imprecise models can be found by contamination: taking convex mixtures of precise models, with mixture coefficient $1-\varepsilon$, and the vacuous model, with mixture coefficient ε, leading to a so-called linear-vacuous model. To simplify the analysis, we let the emission model remain precise, and use the same mixture coefficient ε for the marginal and the transition models. As ε ranges from zero to one, we then evolve from a precise HMM towards an iHMM with vacuous marginal and transition models (and precise emission models).

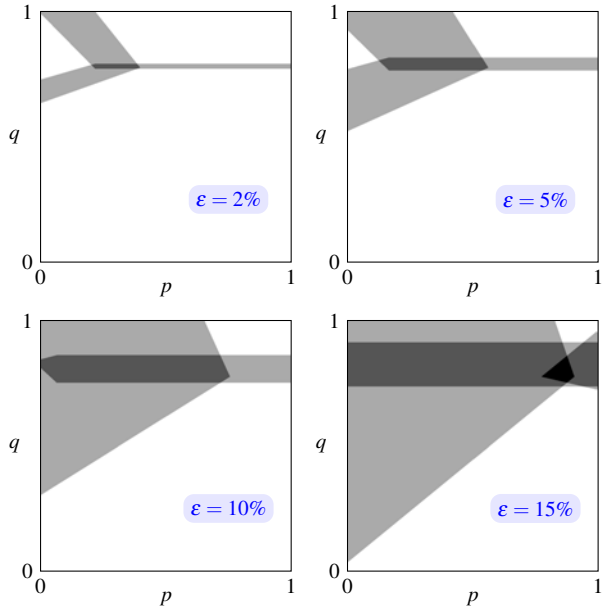
Explaining the basic ideas using a chain of length two. We now examine the behaviour of an iHMM of length two,

[3] Each backward step in the backward-forward loop has a linear complexity in the number of maximal elements at that stage.

with the following (precise) probabilities fixed:[4]

$$m = 0.1,\ r = 0.8 \text{ and } s = 0.3.$$

Fixing an output sequence and a value for ε, we can use our algorithm to calculate the corresponding numbers of maximal state sequences as p and q range over the unit interval. The results can be represented conveniently in the form of a heat plot. The plots below correspond to the output sequence $o_{1:2} = 01$.



The number of maximal state sequences clearly depends on the transition probabilities p and q. In the rather large parts of 'probability space' that are coloured white, we get a single maximal sequence—as we would for HMMs—, but there are contiguous regions where we see a higher number appear. In the present example (binary chain of length two), the highest possible number of maximal sequences is of course four. In the dark grey area, there are three maximal sequences, and two in the light grey regions. The plots show what happens when we let ε increase: the grey areas expand and the number of maximal sequences increases. For $\varepsilon = 15\%$, we even find a small area (coloured black) where all four possible state sequences are maximal: locally, due to the relatively high imprecision of our local models, we cannot give any useful robust estimates of the state sequence producing the output sequence $o_{1:2} = 01$.

For small ε, the areas with more than one maximal state sequence are quite small and seem to resemble strips that narrow down to lines as ε tends to zero. This suggests that we should be able to explain at least qualitatively where these areas come from by looking at compatible precise models: the regions where an iHMM produces different maximal (mutually incomparable) sequences, are widened versions of loci of indifference for precise HMMs.

By a *locus of indifference*, we mean the set of (p,q) that correspond to two given state sequences $x_{1:2}$ and $\hat{x}_{1:2}$ having equal posterior probability:

$$p(x_{1:2}|o_{1:2}) = p(\hat{x}_{1:2}|o_{1:2}),$$

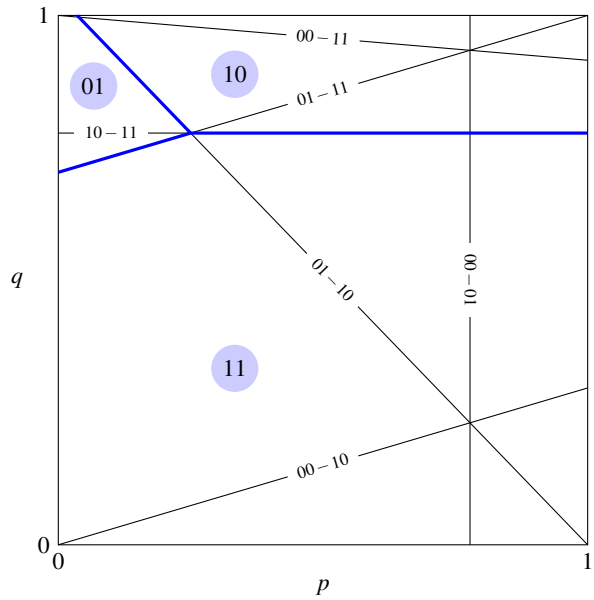
or, provided that $p(o_{1:2}) > 0$,

$$p(x_{1:2}, o_{1:2}) = p(\hat{x}_{1:2}, o_{1:2}).$$

In our example where $o_{1:2} = 01$, we find the following expressions for each of the four possible state sequences:

$$\begin{aligned} p(00,01) &= mr(1-r)p \\ p(01,01) &= mr(1-s)(1-p) \\ p(10,01) &= (1-m)s(1-r)q \\ p(11,01) &= (1-m)s(1-s)(1-q) \end{aligned}$$

By equating any two of these expressions, we express that the corresponding two state sequences have an equal posterior probability. Since the resulting equations are a function of p and q only, each of these six possible combinations defines a locus of indifference. All of them are depicted as lines in the following figure:

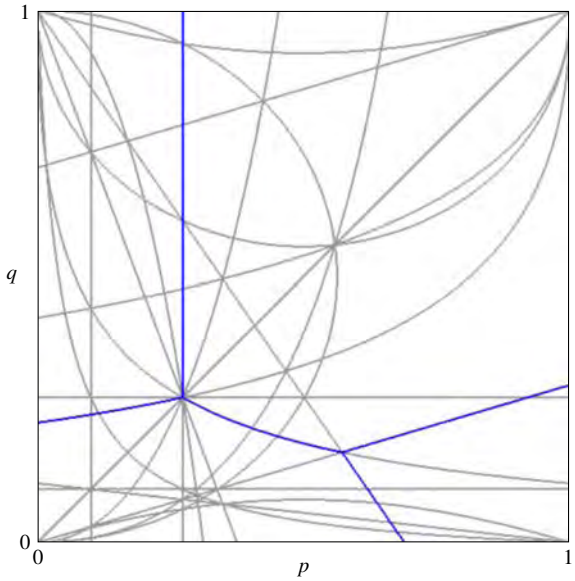


Parts of these loci, depicted in blue (darker and bolder in monochrome versions of this paper) demarcate the three regions where the state sequences 01, 10 and 11 are optimal (have the highest posterior probability).

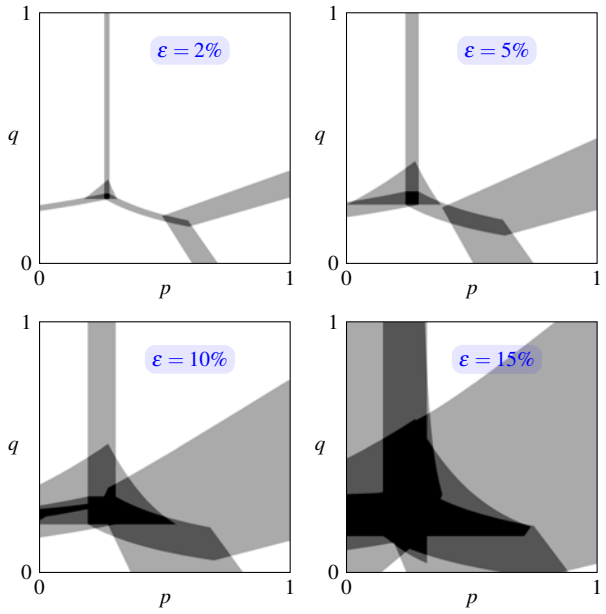
What happens when the transition models become imprecise? Roughly speaking, nearby values of the original p and q enter the picture, effectively turning the loci (lines) of indifference into bands of incomparability: the emergence of regions with two and more maximal sequences can be seen to originate from the loci of indifference; compare the figure for these loci with the heat plots given above.

[4] This choice is of course arbitrary. Different values would yield comparable results.

Extending the argument to a chain of length three. For a chain of length three, we can determine the loci of indifference for precise models in a completely analogous manner. If we use the same marginal model and emission model as in the previous example, the resulting lines of indifference for the output sequence 000 look as follows:



If we compare this with the visualisation below of the number of maximal elements for the same sequence, the resemblance is again quite striking.



7 Showing off the algorithm's power

In order to demonstrate that our algorithm is indeed quite efficient, we let it determine the maximal sequences for a random output sequence of length 100.

We consider the same binary stationary HMM as we presented above, but with the following precise marginal and emission probabilities:

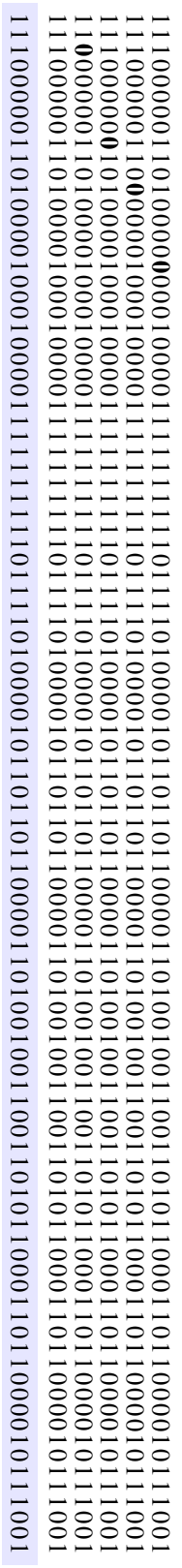
$$m = 0.1,\ r = 0.98 \text{ and } s = 0.01.$$

In practical applications, the probability for an output variable to have the same value as the corresponding hidden state variable is usually quite high, which explains why we have chosen r and s to be close to 1 and to 0, respectively. In contrast with the previous experiments, we do not let the transition probabilities vary, but fix them to the following values:

$$p = 0.6 \text{ and } q = 0.5.$$

The iHMM we use to determine the maximal sequences is then generated by mixing these precise local models with a vacuous one, using the same mixture coefficient ε for the marginal, transition and emission models. On the right, we display the five maximal sequences corresponding to the highlighted output sequence, and $\varepsilon = 2\%$. Since the emission probabilities were chosen to be quite accurate, it is no surprise that the output sequence itself is one of the maximal sequences. In addition, we have indicated in bold face the state values that differ from the outputs in the output sequence. We see that the model represents more indecision about the values of the state variables as we move further away from the end of the sequence. This is a result of a phenomenon called *dilation*, which—as has been noted in another paper [5]—tends to occur when inferences in a credal tree proceed from the leaves towards the root.

As for the efficiency of our algorithm: it took about 0.2 seconds to calculate these 5 maximal sequences. The reason why this could be done so fast is that the algorithm is linear in the number of solutions, which in this case is only 5. If we let ε grow to for example 5%, the number of maximal sequences for the same output sequence is 764 and these can be determined in about 32 seconds. This demonstrates that the complexity is indeed linear in the number of solutions and that the



algorithm can efficiently calculate the maximal sequences even for long output sequences.

8 Conclusions

Interpreting the graphical structure of an imprecise hidden Markov model as a credal network under epistemic irrelevance, leads to an efficient algorithm for finding the maximal state sequences for a given output sequence. Preliminary simulations show that, even for transition models with non-negligible imprecision, the number of maximal elements seems to be reasonably low in fairly large regions of parameter space, with high numbers of maximal elements concentrated in fairly small regions. It remains to be seen whether this observation can be corroborated by a theoretical analysis, and whether increasing the imprecision of the emission models changes this picture appreciably.

It is not clear to us, at this point, whether ideas similar to the ones we discussed above could be used to derive similarly efficient algorithms for imprecise hidden Markov models whose graphical structure is interpreted as a credal network under strong independence [3]. This could be interesting and relevant, as the more stringent independence condition leads to joint models that are less imprecise, and therefore produce fewer maximal state sequences (although they will be contained in our solutions).

Acknowledgements

Jasper De Bock is a Master student of Civil Engineering at Ghent University, and has developed the algorithm described here in the context of his Master's thesis, in close cooperation with Gert de Cooman, who acted as his thesis supervisor.

Research by De Cooman has been supported by SBO project 060043 of the IWT-Vlaanderen. This paper has benefitted from discussions with Marco Zaffalon, Alessandro Antonucci, Alessio Benavoli, Cassio de Campos, Erik Quaeghebeur and Filip Hermans. We are grateful to Marco Zaffalon for providing travel funds allowing us to visit IDSIA and discuss practical applications.

References

[1] Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, 1957.

[2] Inés Couso, Serafín Moral, and Peter Walley. A survey of concepts of independence for imprecise probabilities. *Risk Decision and Policy*, 5:165–181, 2000.

[3] Fabio G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.

[4] L. M. de Campos, J. F. Huete, and S. Moral. Probability intervals: a tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2:167–196, 1994.

[5] Gert de Cooman, Filip Hermans, Alessandro Antonucci, and Marco Zaffalon. Epistemic irrelevance in credal nets: the case of imprecise Markov trees. *International Journal of Approximate Reasoning*, 51:1029–1052, 2010.

[6] Gert de Cooman, Enrique Miranda, and Marco Zaffalon. Independent natural extension. In Eyke Hüllermeier, Rudolf Kruse, and Frank Hoffmann, editors, *Computational Intelligence for Knowledge-Based Systems Design (Proceedings of IPMU 2010, 28 June – 2 July 2010, Dortmund, Germany)*, volume 6178 of *Lecture Notes in Computer Science*, pages 737–746. Springer, Heidelberg, 2010.

[7] Gert de Cooman and Matthias C. M. Troffaes. Dynamic programming for deterministic discrete-time systems with uncertain gain. *International Journal of Approximate Reasoning*, 39:257–278, 2005.

[8] Gert de Cooman, Matthias C. M. Troffaes, and Enrique Miranda. *n*-Monotone exact functionals. *Journal of Mathematical Analysis and Applications*, 347:143–156, 2008.

[9] Nathan Huntley and Matthias C. M. Troffaes. Normal form backward induction for decision trees with coherent lower previsions. *Annals of Operations Research*, 2010. Submitted for publication.

[10] Enrique Miranda. A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 48(2):628–658.

[11] Lawrence R. Rabiner. A tutorial on HMM and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[12] Matthias C. M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29.

[13] Andrew J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

[14] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

Independent natural extension for sets of desirable gambles

Gert de Cooman
SYSTeMS, Ghent University, Belgium
gert.decooman@UGent.be

Enrique Miranda
Dept. of Statistics and O.R., University of Oviedo, Spain
mirandaenrique@uniovi.es

Abstract

We investigate how to combine a number of marginal coherent sets of desirable gambles into a joint set using the properties of epistemic irrelevance and independence. We provide formulas for the smallest such joint, called their independent natural extension, and study its main properties. The independent natural extension of maximal sets of gambles allows us to define the strong product of sets of desirable gambles. Finally, we explore an easy way to generalise these results to also apply for the conditional versions of epistemic irrelevance and independence.

Keywords. Epistemic irrelevance, epistemic independence, independent natural extension, strong product, coherent set of desirable gambles.

1 Introduction

One disadvantage of working with coherent lower previsions (or previsions and probabilities for that matter), is that conditioning a lower prevision does not necessarily lead to uniquely coherent results when the conditioning event has lower probability zero; see for instance Ref. [8, Section 6.4]. For precise probabilities, this difficulty can be circumvented by using full conditional measures [5]. In an imprecise-probabilities context, working with the more informative coherent sets of desirable gambles rather than with lower previsions provides a very elegant and intuitively appealing way out of this problem, as Walley already suggested in 1991 [8, Section 3.8.6 and Appendix F], and argued in much more detail in 2000 [9]. The connection between full conditional measures and maximal coherent sets of desirable gambles was explored by Couso and Moral [1]. De Cooman and Quaeghebeur [4] have shown that working with sets of desirable gambles is especially illuminating in the context of modelling exchangeability assessments.

Exchangeability is a structural assessment, and so is independence. Conditioning and independence are, of course, closely related. In a recent paper [3], we investigated the notions of epistemic independence of finite-valued variables using coherent lower previsions. The above-mentioned problems with conditioning, and the fact that the coherence requirements for conditional lower previsions are, to be honest, quite cumbersome to work with, have turned this into a quite complicated exercise. This is the reason why, in the present paper, we investigate if looking at independence using sets of desirable gambles leads to a more elegant theory that avoids some of the complexity pitfalls of working with coherent lower previsions. In doing this, we build on the strong pioneering work on epistemic irrelevance by Moral [7]. While we focus here on the symmetrised notion of epistemic independence, much of what we do can be seen as an application and continuation of his ideas.

In Section 2 we summarise relevant results in the existing theory of sets of desirable gambles. After mentioning useful notational conventions in Section 3, we recall the basic marginalisation, conditioning and extension operations for sets of desirable gambles in Sections 4 and 5. We use these to combine a number of marginal sets of desirable gambles into a joint satisfying epistemic irrelevance (Section 6), and epistemic independence (Section 7). In Section 8, we study the particular case of maximal sets of desirable gambles, and derive the concept of a strong product. Section 9 deals with conditional independence assessments.

2 Coherent sets of desirable gambles and natural extension

Consider a variable X taking values in some non-empty set $\mathscr{X}$, that we shall assume to be finite. We model information about X by means of sets of desirable gambles. A *gamble* is a real-valued function on $\mathscr{X}$, and we denote the set of all gambles on $\mathscr{X}$ by $\mathscr{G}(\mathscr{X})$. It is a linear space under point-wise addition of gambles and point-wise multiplication of gambles with real numbers. For any subset $\mathscr{A}$ of $\mathscr{G}(\mathscr{X})$, we denote by $\operatorname{posi}(\mathscr{A})$ the set of all positive linear combinations of gambles in $\mathscr{A}$:

$$\operatorname{posi}(\mathscr{A}) := \left\{ \sum_{k=1}^{n} \lambda_k f_k \colon f_k \in \mathscr{A}, \lambda_k > 0, n > 0 \right\}.$$

We call $\mathscr{A}$ a *convex cone* if it is closed under positive linear combinations, meaning that $\text{posi}(\mathscr{A}) = \mathscr{A}$.

For any gambles f and g on $\mathscr{X}$, we write '$f \geq g$' if $(\forall x \in \mathscr{X}) f(x) \geq g(x)$, and '$f > g$' if $f \geq g$ and $f \neq g$. A gamble $f > 0$ is called *positive*. A gamble $g \leq 0$ is called *non-positive*. $\mathscr{G}(\mathscr{X})_{\neq 0}$ denotes the set of all non-zero gambles, $\mathscr{G}(\mathscr{X})_{>0}$ the convex cone of all positive gambles, and $\mathscr{G}(\mathscr{X})_{\leq 0}$ the convex cone of all non-positive gambles.

2.1 Coherence and avoiding non-positivity

Definition 1 ([4]). *A set of desirable gambles $\mathscr{D} \subseteq \mathscr{G}(\mathscr{X})$* avoids non-positivity *if $\mathscr{G}(\mathscr{X})_{\leq 0} \cap \text{posi}(\mathscr{D}) = \emptyset$. It is called* coherent *if:*

D1. $0 \notin \mathscr{D}$;
D2. $\mathscr{G}(\mathscr{X})_{>0} \subseteq \mathscr{D}$;
D3. $\mathscr{D} = \text{posi}(\mathscr{D})$.

We denote by $\mathbb{D}(\mathscr{X})$ the set of all coherent sets of desirable gambles on $\mathscr{X}$.

Requirement D3 turns $\mathscr{D}$ into a convex cone. Due to D2, it includes $\mathscr{G}(\mathscr{X})_{>0}$; due to D1–D3, it excludes $\mathscr{G}(\mathscr{X})_{\leq 0}$, and therefore avoids non-positivity.

2.2 Natural extension

If we consider any non-empty family of coherent sets of desirable gambles $\mathscr{D}_i$, $i \in I$, then their intersection $\bigcap_{i \in I} \mathscr{D}_i$ is still coherent. This is the idea behind the following result. If a subject gives us an *assessment*, a set $\mathscr{A} \subseteq \mathscr{G}(\mathscr{X})$ of gambles on $\mathscr{X}$ that he finds desirable, then we can tell exactly when this assessment can be extended to a coherent set, and how to construct the smallest such set.

Theorem 1 (Natural extension [4]). *Consider an assessment $\mathscr{A} \subseteq \mathscr{G}(\mathscr{X})$, and define its* natural extension *as:*[1]

$$\mathscr{E}(\mathscr{A}) := \bigcap \{\mathscr{D} \in \mathbb{D}(\mathscr{X}) \colon \mathscr{A} \subseteq \mathscr{D}\}$$

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) *$\mathscr{E}(\mathscr{A}) \neq \mathscr{G}(\mathscr{X})$;*
(iv) *the set of desirable gambles $\mathscr{E}(\mathscr{A})$ is coherent;*
(v) *$\mathscr{E}(\mathscr{A})$ is the smallest coherent set of desirable gambles that includes $\mathscr{A}$.*

When any (and hence all) of these equivalent statements hold, then $\mathscr{E}(\mathscr{A}) = \text{posi}\big(\mathscr{G}(\mathscr{X})_{>0} \cup \mathscr{A}\big)$.

2.3 Helpful lemmas

In order to prove a number of results in this paper, we need the following lemmas, one of which is convenient version of the separating hyperplane theorem:

Lemma 2. *Consider a finite subset $\mathscr{A}$ of $\mathscr{G}(\mathscr{X})$. Then $0 \notin \text{posi}(\mathscr{G}(\mathscr{X})_{>0} \cup \mathscr{A})$ if and only if there is some probability mass function p such that $\sum_{x \in \mathscr{X}} p(x) f(x) > 0$ for all $f \in \mathscr{A}$ and $p(x) > 0$ for all $x \in \mathscr{X}$.*

Proof. It clearly suffices to prove necessity. Since $0 \notin \text{posi}(\mathscr{G}(\mathscr{X})_{>0} \cup \mathscr{A})$, we infer from a version of the separating hyperplane theorem [8, Appendix E.1] that there is a linear functional Λ on $\mathscr{G}(\mathscr{X})$ such that

$$(\forall x \in \mathscr{X}) \Lambda(\mathbb{I}_{\{x\}}) > 0 \text{ and } (\forall f \in \mathscr{A}) \Lambda(f) > 0.$$

Then $\Lambda(\mathscr{X}) = \sum_{x \in \mathscr{X}} \Lambda(\mathbb{I}_{\{x\}}) > 0$, and if we let $p(x) := \Lambda(\mathbb{I}_{\{x\}})/\Lambda(\mathscr{X}) > 0$ for all $x \in \mathscr{X}$, then p is a probability mass function on $\mathscr{X}$ for which $\Lambda(f)/\Lambda(\mathscr{X}) = \sum_{x \in \mathscr{X}} p(x) f(x) > 0$ for all $f \in \mathscr{A}$. □

Lemma 3. *Consider a convex cone $\mathscr{A}$ of gambles on $\mathscr{X}$ such that $\max f > 0$ for all $f \in \mathscr{A}$. Consider any non-zero gamble g on $\mathscr{X}$. If $g \notin \mathscr{A}$ then $0 \notin \text{posi}(\mathscr{A} \cup \{-g\})$.*

Proof. Consider a non-zero gamble $g \notin \mathscr{A}$, and assume *ex absurdo* that $0 \in \text{posi}(\mathscr{A} \cup \{-g\})$. Then it follows from the assumptions that there are $f \in \mathscr{A}$ and $\mu > 0$ such that $0 = f + \mu(-g)$. Hence $g \in \mathscr{A}$, a contradiction. □

2.4 Maximal sets of desirable gambles

An element $\mathscr{D}$ of $\mathbb{D}(\mathscr{X})$ is called *maximal* if it is not strictly included in any other element of $\mathbb{D}(\mathscr{X})$, or in other words, if adding any gamble f to $\mathscr{D}$ makes sure we can no longer extend the set $\mathscr{D} \cup \{f\}$ to a set that is still coherent:

$$(\forall \mathscr{D}' \in \mathbb{D}(\mathscr{X}))(\mathscr{D} \subseteq \mathscr{D}' \Rightarrow \mathscr{D} = \mathscr{D}')$$

$\mathbb{M}(\mathscr{X})$ denotes the set of all maximal elements of $\mathbb{D}(\mathscr{X})$.

The following proposition provides a characterisation of such maximal elements.

Proposition 4 ([1, 4]). *Let $\mathscr{D} \in \mathbb{D}(\mathscr{X})$, then $\mathscr{D}$ is a maximal coherent set of desirable gambles if and only if*

$$(\forall f \in \mathscr{G}(\mathscr{X})_{\neq 0})(f \notin \mathscr{D} \Rightarrow -f \in \mathscr{D}).$$

For the following important result, it is easy to provide a constructive proof, based on the same ideas as in Ref. [1]. For the more general case of infinite $\mathscr{X}$, a non-constructive proof can be based on Zorn's Lemma [4].

Theorem 5 ([1, 4]). *A subset $\mathscr{A}$ of $\mathscr{G}(\mathscr{X})$ avoids non-positivity if and only if $m(\mathscr{A}) := \{\mathscr{M} \in \mathbb{M}(\mathscr{X}) \colon \mathscr{A} \subseteq \mathscr{M}\}$ is non-empty. Moreover, $\mathscr{E}(\mathscr{A}) = \bigcap m(\mathscr{A})$.*

2.5 Coherent lower previsions

Given a coherent set of desirable gambles $\mathscr{D}$, the functional $\underline{P}$ defined on $\mathscr{G}(\mathscr{X})$ by

$$\underline{P}(f) := \sup\{\mu \colon f - \mu \in \mathscr{D}\} \tag{1}$$

[1] As usual, in this expression, we let $\bigcap \emptyset = \mathscr{G}(\mathscr{X})$.

is a coherent lower prevision [8, Theorem 3.8.1], and therefore corresponds to taking a lower envelope of expectations with respect a set of probability mass functions. Many different coherent sets of desirable gambles induce the same coherent lower prevision $\underline{P}$. The smallest is called the associated *set of strictly desirable* gambles:

$$\mathscr{D}' := \{f \in \mathscr{G}(\mathscr{X}) \colon f > 0 \text{ or } \underline{P}(f) > 0\}. \tag{2}$$

When $\mathscr{D}$ is a maximal coherent set of desirable gambles, the lower prevision $\underline{P}$ defined by Eq. (1) is a *linear prevision*, meaning that it corresponds to an expectation operator with respect to a probability mass function. For more information, see Refs. [1, Section 5], [6, Proposition 6], [8] and [10].

3 Basic notation

From now on we consider a number of variables X_n, $n \in N$, taking values in the respective finite sets $\mathscr{X}_n$. Here N is some finite non-empty index set.

For every subset R of N, we denote by X_R the tuple of variables (with one component for each $r \in R$) that takes values in the Cartesian product $\mathscr{X}_R := \times_{r\in R}\mathscr{X}_r$. The elements of $\mathscr{X}_R$ are generically denoted by x_R or z_R, with corresponding components $x_r := x_R(r)$ or $z_r := z_R(r)$, $r \in R$.

We will assume that the variables X_n are logically independent, which means that for each subset R of N, X_R may assume all values in $\mathscr{X}_R$.

We denote by $\mathscr{G}(\mathscr{X}_R)$ the set of gambles defined on $\mathscr{X}_R$. We will frequently resort to the simplifying device of *identifying* a gamble on $\mathscr{X}_R$ with a gamble on $\mathscr{X}_N$, namely its cylindrical extension. To give an example, if $\mathscr{K} \subseteq \mathscr{G}(\mathscr{X}_N)$, this trick allows us to consider $\mathscr{K} \cap \mathscr{G}(\mathscr{X}_R)$ as the set of those gambles in $\mathscr{K}$ that depend only on the variable X_R. As another example, this device allows us to identify the gambles $\mathbb{I}_{\{x_R\}}$ and $\mathbb{I}_{\{x_R\}\times\mathscr{X}_{N\setminus R}}$, and therefore also the events $\{x_R\}$ and $\{x_R\} \times \mathscr{X}_{N\setminus R}$. More generally, for any event $A \subseteq \mathscr{X}_R$, we can identify the gambles $\mathbb{I}_A$ and $\mathbb{I}_{A\times\mathscr{X}_{N\setminus R}}$, and therefore also the events A and $A \times \mathscr{X}_{N\setminus R}$.

We draw attention to the case $R = \emptyset$. By definition, $\mathscr{X}_\emptyset$ contains only one element $x_\emptyset$: the empty map $\emptyset \to \emptyset$. There is no uncertainty about the value of the variable $X_\emptyset$: it can assume only one value (the empty map), and $\mathbb{I}_{\mathscr{X}_\emptyset} = \mathbb{I}_{\{x_\emptyset\}} = 1$. We can identify $\mathscr{G}(\mathscr{X}_\emptyset)$ with the set of real numbers $\mathbb{R}$. There is only one coherent set of desirable gambles on $\mathscr{X}_\emptyset$: the set $\mathbb{R}_{>0}$ of positive real numbers.

4 Marginalisation and cylindrical extension

Suppose that we have a set $\mathscr{D}_N \subseteq \mathscr{G}(\mathscr{X}_N)$ of desirable gambles modelling a subject's information about the uncertain variable X_N. We are interested in modelling the information about the variable X_O, where O is some subset of N. This can be done using the set of desirable gambles that belong to $\mathscr{D}_N$ but only depend on the variable X_O:

$$\operatorname{marg}_O(\mathscr{D}_N) := \{g \in \mathscr{G}(\mathscr{X}_O) \colon g \in \mathscr{D}_N\} = \mathscr{D}_N \cap \mathscr{G}(\mathscr{X}_O) \tag{3}$$

is called a *marginal* set of desirable gambles [7]. Observe that $\operatorname{marg}_\emptyset(\mathscr{D}_N) = \mathscr{G}(\mathscr{X}_\emptyset)_{>0}$, which can be identified with the set of positive real numbers $\mathbb{R}_{>0}$. Also, with $O_1, O_2 \subseteq N$, it is obvious that

$$O_1 \subseteq O_2 \Rightarrow \operatorname{marg}_{O_1}(\operatorname{marg}_{O_2}(\mathscr{D}_N)) = \operatorname{marg}_{O_1}(\mathscr{D}_N). \tag{4}$$

Coherence is trivially preserved under marginalisation:

Proposition 6. *Let $\mathscr{D}_N$ be a set of desirable gambles on $\mathscr{X}_N$, and consider any subset O of N.*

(i) *If $\mathscr{D}_N$ avoids non-positivity, then so does* $\operatorname{marg}_O(\mathscr{D}_N)$.
(ii) *If $\mathscr{D}_N$ is coherent, then* $\operatorname{marg}_O(\mathscr{D}_N)$ *is a coherent set of desirable gambles on $\mathscr{X}_O$.*

We now look for a kind of inverse operation to marginalisation. Suppose we have a coherent set $\mathscr{D}_O \subseteq \mathscr{G}(\mathscr{X}_O)$ of desirable gambles modelling a subject's information about the uncertain variable X_O, and we want to extend this to a coherent set of desirable gambles on $\mathscr{X}_N$, representing the same information. So we are looking for a coherent set of desirable gambles $\mathscr{D}_N \subseteq \mathscr{G}(\mathscr{X}_N)$ such that $\operatorname{marg}_O(\mathscr{D}_N) = \mathscr{D}_O$ and that is as small as possible: the most conservative coherent set of desirable gambles on $\mathscr{X}_N$ that marginalises to $\mathscr{D}_O$.

Proposition 7. *Let O be a subset of N and let $\mathscr{D}_O \in \mathbb{D}(\mathscr{X}_O)$. Then the most conservative (smallest) coherent set of desirable gambles on $\mathscr{X}_N$ that marginalises to $\mathscr{D}_O$ is given by*

$$\operatorname{ext}_N(\mathscr{D}_O) := \operatorname{posi}(\mathscr{G}(\mathscr{X}_N)_{>0} \cup \mathscr{D}_O). \tag{5}$$

It is called the cylindrical extension *of $\mathscr{D}_O$ to a set of desirable gambles on $\mathscr{X}_N$, and satisfies*

$$\operatorname{marg}_O(\operatorname{ext}_N(\mathscr{D}_O)) = \mathscr{D}_O. \tag{6}$$

This extension is called *weak extension* by Moral [7, Section 2.1].

Proof. It is clear from the coherence requirements and Eq. (3) that any coherent set that marginalises to $\mathscr{D}_O$ must include $\mathscr{G}(\mathscr{X}_N)_{>0}$ and $\mathscr{D}_O$, and therefore also $\operatorname{posi}(\mathscr{G}(\mathscr{X}_N)_{>0} \cup \mathscr{D}_O) = \operatorname{ext}_N(\mathscr{D}_O)$. It therefore suffices to prove that $\operatorname{posi}(\mathscr{G}(\mathscr{X}_N)_{>0} \cup \mathscr{D}_O)$ is coherent, and that it marginalises to $\mathscr{D}_O$.

To prove coherence, it suffices to prove that $\mathscr{D}_O$ avoids non-positivity, by Theorem 1. But this is obvious because $\mathscr{D}_O$ is a coherent set of desirable gambles on $\mathscr{X}_O$.

We are left to prove that $\operatorname{marg}_O(\operatorname{ext}_N(\mathscr{D}_O)) = \mathscr{D}_O$. Since for any $g \in \mathscr{D}_O$ it is obvious that both $g \in \operatorname{ext}_N(\mathscr{D}_O)$ and $g \in \mathscr{G}(\mathscr{X}_O)$, we see immediately that $\mathscr{D}_O \subseteq \operatorname{marg}_O(\operatorname{ext}_N(\mathscr{D}_O))$, so

we concentrate on proving that $\text{marg}_O(\text{ext}_N(\mathscr{D}_O)) \subseteq \mathscr{D}_O$. Consider $f \in \text{marg}_O(\text{ext}_N(\mathscr{D}_O))$, meaning that both $f \in \mathscr{G}(\mathscr{X}_O)$ and $f \in \text{ext}_N(\mathscr{D}_O)$. The latter means that there are $g \in \mathscr{D}_O$, $h \in \mathscr{G}(\mathscr{X}_N)_{>0}$, and non-negative λ and μ such that $\max\{\lambda,\mu\} > 0$ for which $f = \lambda g + \mu h$. Since we need to prove that $f \in \mathscr{D}_O$, we can assume without loss of generality that $\mu > 0$. But then $h = (f - \lambda g)/\mu \in \mathscr{G}(\mathscr{X}_O)$ and therefore also $h \in \mathscr{G}(\mathscr{X}_O)_{>0}$, whence indeed $f \in \mathscr{D}_O$, by coherence of $\mathscr{D}_O$. □

5 Conditioning

Suppose that we have a set $\mathscr{D}_N \subseteq \mathscr{G}(\mathscr{X}_N)$ of desirable gambles modelling a subject's information about the uncertain variable X_N. Consider a subset I of N, and assume we want to update the model $\mathscr{D}_N$ with the information that $X_I = x_I$. This leads to an updated set of desirable gambles:

$$\mathscr{D}_N|x_I := \left\{f \in \mathscr{G}(\mathscr{X}_N)\colon \mathbb{I}_{\{x_I\}}f \in \mathscr{D}_N\right\}. \tag{7}$$

For technical reasons, and mainly in order to streamline the proofs as much as possible, we also allow the admittedly pathological case that $I = \emptyset$. Since $\mathbb{I}_{\{x_\emptyset\}} = 1$, this amounts to not conditioning at all.

Eq. (7) introduces the conditioning operator '|' essentially used by Walley [9] and Moral [7]. We prefer a slightly modified version '⌋' [4]. Since $\mathbb{I}_{\{x_I\}}f = \mathbb{I}_{\{x_I\}}f(x_I,\cdot)$, we can characterise the updated model $\mathscr{D}_N|x_I$ through the set

$$\mathscr{D}_N\rfloor x_I := \left\{g \in \mathscr{G}(\mathscr{X}_{N\setminus I})\colon \mathbb{I}_{\{x_I\}}g \in \mathscr{D}_N\right\} \subseteq \mathscr{G}(\mathscr{X}_{N\setminus I}),$$

in the specific sense that for all $g \in \mathscr{G}(\mathscr{X}_{N\setminus I})$:

$$g \in \mathscr{D}_N\rfloor x_I \Leftrightarrow \mathbb{I}_{\{x_I\}}g \in \mathscr{D}_N \Leftrightarrow \mathbb{I}_{\{x_I\}}g \in \mathscr{D}_N|x_I, \tag{8}$$

and for all $f \in \mathscr{G}(\mathscr{X}_N)$: $f \in \mathscr{D}_N|x_I \Leftrightarrow f(x_I,\cdot) \in \mathscr{D}_N\rfloor x_I$. Coherence is trivially preserved under conditioning:

Proposition 8. *Let $\mathscr{D}_N$ be a coherent set of desirable gambles on $\mathscr{X}_N$, and consider any subset I of N. Then $\mathscr{D}_N\rfloor x_I$ is a coherent set of desirable gambles on $\mathscr{X}_{N\setminus I}$.*

The order of marginalisation and conditioning can be reversed, under some conditions.

Proposition 9. *Let $\mathscr{D}_N$ be a coherent set of desirable gambles on $\mathscr{X}_N$, and consider any disjoint subsets I and O of N. Then $\text{marg}_O(\mathscr{D}_N\rfloor x_I) = \text{marg}_{I\cup O}(\mathscr{D}_N)\rfloor x_I$ for all $x_I \in \mathscr{X}_I$.*

Proof. Consider any $h \in \mathscr{G}(\mathscr{X}_N)$ and observe the following chain of equivalences:

$$\begin{aligned}
h \in \text{marg}_O(\mathscr{D}_N\rfloor x_I) &\Leftrightarrow h \in \mathscr{G}(\mathscr{X}_O) \text{ and } h \in \mathscr{D}_N\rfloor x_I\\
&\Leftrightarrow h \in \mathscr{G}(\mathscr{X}_O) \text{ and } \mathbb{I}_{\{x_I\}}h \in \mathscr{D}_N\\
&\Leftrightarrow h \in \mathscr{G}(\mathscr{X}_O) \text{ and } \mathbb{I}_{\{x_I\}}h \in \text{marg}_{I\cup O}(\mathscr{D}_N)\\
&\Leftrightarrow h \in \mathscr{G}(\mathscr{X}_O) \text{ and } h \in \text{marg}_{I\cup O}(\mathscr{D}_N)\rfloor x_I\\
&\Leftrightarrow h \in \text{marg}_{I\cup O}(\mathscr{D}_N)\rfloor x_I.
\end{aligned}$$
□

6 Irrelevant natural extension

We are now ready to look at the simplest type of irrelevance judgement. Consider two disjoint subsets I and O of N. We say that X_I is *epistemically irrelevant* to X_O when learning the value of X_I does not influence or change our subject's beliefs about X_O.

When does a set $\mathscr{D}_N$ of desirable gambles on $\mathscr{X}_N$ capture this type of epistemic irrelevance? Observing that $X_I = x_I$ turns $\mathscr{D}_N$ into the updated set $\mathscr{D}_N\rfloor x_I$ of desirable gambles on $\mathscr{X}_{N\setminus I}$, we should clearly require that:

$$\text{marg}_O(\mathscr{D}_N\rfloor x_I) = \text{marg}_O(\mathscr{D}_N) \text{ for all } x_I \in \mathscr{X}_I. \tag{9}$$

As before, for technical reasons we also allow I and O to be empty. It is clear from the definition above that the 'variable' $X_\emptyset$, about whose constant value we are certain, is epistemically irrelevant to any variable X_O. Similarly, we see that any variable X_I is epistemically irrelevant to the 'variable' $X_\emptyset$. This seems to be in accordance with intuition.

The epistemic irrelevance condition can be formulated trivially in an interesting and slightly different manner.

Proposition 10. *Let $\mathscr{D}_N$ be a coherent set of desirable gambles on $\mathscr{X}_N$, and let I and O be any disjoint subsets of N. Then the following statements are equivalent:*

(i) $\text{marg}_O(\mathscr{D}_N\rfloor x_I) = \text{marg}_O(\mathscr{D}_N)$ *for all* $x_I \in \mathscr{X}_I$;
(ii) *for all* $f \in \mathscr{G}(\mathscr{X}_O)$ *and all* $x_I \in \mathscr{X}_I$: $\mathbb{I}_{\{x_I\}}f \in \mathscr{D}_N \Leftrightarrow f \in \mathscr{D}_N$.

Irrelevance assessments are most useful in constructing sets of desirable gambles from other ones. Suppose we have a coherent set $\mathscr{D}_O$ of desirable gambles on $\mathscr{X}_O$, and an assessment that X_I is epistemically irrelevant to X_O, where I and O are disjoint index sets. Then how can we combine $\mathscr{D}_O$ and this structural irrelevance assessment into a coherent set of desirable gambles on $\mathscr{X}_{I\cup O}$, or more generally, on $\mathscr{X}_N$, where $N \supseteq I \cup O$? To see how this can be done in a way that is as conservative as possible, we introduce:

$$\mathscr{A}^{\text{irr}}_{I\to O} := \text{posi}\left(\left\{\mathbb{I}_{\{x_I\}}g\colon g \in \mathscr{D}_O \text{ and } x_I \in \mathscr{X}_I\right\}\right).$$

It follows from the next lemma that for all $h \in \mathscr{G}(\mathscr{X}_{I\cup O})$:

$$h \in \mathscr{A}^{\text{irr}}_{I\to O} \Leftrightarrow h \neq 0 \text{ and } (\forall x_I \in \mathscr{X}_I)h(x_I,\cdot) \in \mathscr{D}_O \cup \{0\}. \tag{10}$$

Clearly, and this will be quite important in streamlining proofs, $\mathscr{A}^{\text{irr}}_{\emptyset\to O} = \mathscr{D}_O$ and $\mathscr{A}^{\text{irr}}_{I\to\emptyset} = \mathscr{G}(\mathscr{X}_I)_{>0}$. We also give two important properties of these sets:

Lemma 11. *Consider disjoint subsets I and O of N, and a coherent set $\mathscr{D}_O$ of desirable gambles on $\mathscr{X}_O$. Then $\mathscr{A}^{\text{irr}}_{I\to O}$ is a coherent set of desirable gambles on $\mathscr{X}_{I\cup O}$.*

Proof. D1. Assume *ex absurdo* that there are $n > 0$, real $\lambda_k > 0$ and $f_k \in \mathscr{A}^{\text{irr}}_{I\to O}$ such that $\sum_{k=1}^n \lambda_k f_k = 0$. It follows from the assumptions that there are $\ell \in \{1,\dots,n\}$ and $x_I \in \mathscr{X}_I$ such that $f_\ell(x_I,\cdot) \neq 0$. This implies that in the sum $\sum_{k=1}^n \lambda_k f_k(x_I,\cdot) = 0$

not all the gambles $\lambda_k f_k(x_I,\cdot)$ are zero. Since the non-zero ones belong to $\mathscr{D}_O$, this contradicts the coherence of $\mathscr{D}_O$.

D2. Consider any $h \in \mathscr{G}(\mathscr{X}_{I\cup O})_{>0}$. Then clearly $h(x_I,\cdot) \geq 0$ and therefore $h(x_I,\cdot) \in \mathscr{D}_O \cup \{0\}$ for all $x_I \in \mathscr{X}_I$. Since $h \neq 0$, it follows that indeed $h \in \mathscr{A}^{\text{irr}}_{I\to O}$.

D3. Trivial if we recall that $\text{posi}(\text{posi}(\mathscr{D})) = \text{posi}(\mathscr{D})$ for any set of desirable gambles $\mathscr{D}$. □

Lemma 12. *Consider disjoint subsets I and O of N, and a coherent set $\mathscr{D}_O$ of desirable gambles on $\mathscr{X}_O$. Then* $\text{marg}_O(\mathscr{A}^{\text{irr}}_{I\to O}) = \mathscr{D}_O$.

Proof. It is obvious from Eq. (10) that indeed:

$$\begin{aligned}\text{marg}_O(\mathscr{A}^{\text{irr}}_{I\to O}) &= \mathscr{A}^{\text{irr}}_{I\to O} \cap \mathscr{G}(\mathscr{X}_O)\\ &= \left\{h \in \mathscr{G}(\mathscr{X}_O)_{\neq 0}\colon (\forall x_I \in \mathscr{X}_I) h \in \mathscr{D}_O \cup \{0\}\right\}\\ &= \left\{h \in \mathscr{G}(\mathscr{X}_O)_{\neq 0}\colon h \in \mathscr{D}_O \cup \{0\}\right\} = \mathscr{D}_O. \quad \square\end{aligned}$$

Theorem 13. *Consider disjoint subsets I and O of N, and a coherent set $\mathscr{D}_O$ of desirable gambles on $\mathscr{X}_O$. Then the smallest coherent set of desirable gambles on $\mathscr{X}_N$ that marginalises to $\mathscr{D}_O$ and satisfies the epistemic irrelevance condition* (9) *of X_I to X_O is given by* $\text{ext}_N(\mathscr{A}^{\text{irr}}_{I\to O}) = \text{posi}(\mathscr{G}(\mathscr{X}_N)_{>0} \cup \mathscr{A}^{\text{irr}}_{I\to O})$.

Proof. Consider any coherent set $\mathscr{D}_N$ on $\mathscr{X}_N$ that marginalises to $\mathscr{D}_O$ and satisfies the irrelevance condition (9). This implies that $\text{marg}_O(\mathscr{D}_N\rfloor x_I) = \mathscr{D}_O$ for any $x_I \in \mathscr{X}_I$, so $g \in \mathscr{D}_N\rfloor x_I$, and therefore $\mathbb{I}_{\{x_I\}}g \in \mathscr{D}_N$ for any $g \in \mathscr{D}_O$, by Eq. (8). So we infer by coherence that $\mathscr{A}^{\text{irr}}_{I\to O} \subseteq \mathscr{D}_N$, and therefore also that $\text{posi}(\mathscr{G}(\mathscr{X}_N)_{>0} \cup \mathscr{A}^{\text{irr}}_{I\to O}) \subseteq \mathscr{D}_N$. As a consequence, it suffices to prove that (i) $\text{ext}_N(\mathscr{A}^{\text{irr}}_{I\to O})$ is coherent, (ii) marginalises to $\mathscr{D}_O$, and (iii) satisfies the epistemic irrelevance condition (9). This is what we now set out to do.

(i). By Lemma 11, $\mathscr{A}^{\text{irr}}_{I\to O}$ is a coherent set of desirable gambles on $\mathscr{X}_{I\cup O}$, so Proposition 7 implies that $\text{posi}(\mathscr{G}(\mathscr{X}_N)_{>0} \cup \mathscr{A}^{\text{irr}}_{I\to O}) = \text{ext}_N(\mathscr{A}^{\text{irr}}_{I\to O})$ is a coherent set of desirable gambles on $\mathscr{X}_N$.

(ii). Marginalisation leads to:

$$\begin{aligned}\text{marg}_O(\text{ext}_N(\mathscr{A}^{\text{irr}}_{I\to O})) &= \text{marg}_O(\text{marg}_{I\cup O}(\text{ext}_N(\mathscr{A}^{\text{irr}}_{I\to O})))\\ &= \text{marg}_O(\mathscr{A}^{\text{irr}}_{I\to O}) = \mathscr{D}_O,\end{aligned}$$

where the first equality follows from Eq. (4), the second from Eq. (6), and the third from Lemma 12.

(iii). It follows from Proposition 9 and Eq. (6) that

$$\begin{aligned}\text{marg}_O(\text{ext}_N(\mathscr{A}^{\text{irr}}_{I\to O})\rfloor x_I) &= \text{marg}_{I\cup O}(\text{ext}_N(\mathscr{A}^{\text{irr}}_{I\to O}))\rfloor x_I\\ &= \mathscr{A}^{\text{irr}}_{I\to O}\rfloor x_I,\end{aligned}$$

and we have just shown in (ii) that $\text{marg}_O(\text{ext}_N(\mathscr{A}^{\text{irr}}_{I\to O})) = \mathscr{D}_O$, so proving that $\text{marg}_O(\text{ext}_N(\mathscr{A}^{\text{irr}}_{I\to O})\rfloor x_I) = \text{marg}_O(\text{ext}_N(\mathscr{A}^{\text{irr}}_{I\to O}))$ amounts to proving that $\mathscr{A}^{\text{irr}}_{I\to O}\rfloor x_I = \mathscr{D}_O$. It is obvious from the definition of $\mathscr{A}^{\text{irr}}_{I\to O}$ that $\mathscr{D}_O \subseteq \mathscr{A}^{\text{irr}}_{I\to O}\rfloor x_I$, so we concentrate on the converse inclusion. Consider any $h \in \mathscr{A}^{\text{irr}}_{I\to O}\rfloor x_I$; then $\mathbb{I}_{\{x_I\}}h \in \mathscr{A}^{\text{irr}}_{I\to O}$, so we infer from Eq. (10) that in particular $h \in \mathscr{D}_O \cup \{0\}$. But since $\mathscr{A}^{\text{irr}}_{I\to O}$ is coherent by Lemma 11, we see that $h \neq 0$ and therefore indeed $h \in \mathscr{D}_O$. □

Theorem 13 is mentioned briefly, with only a hint at the proof, by Moral [7, Section 2.4]. We believe the result is not so trivial and have therefore decided to include our version of the proof here. Our notion of epistemic irrelevance is called *weak* epistemic irrelevance by Moral. For his version of epistemic irrelevance he requires in addition that $\mathscr{D}_N$ should be equal to the irrelevant natural extension of $\mathscr{D}_O$, and therefore be the *smallest* model that satisfies the (weak) epistemic irrelevance condition (9). While we feel comfortable with his reasons for doing so, we have decided not to follow his lead in this.

7 Independent natural extension

We now turn to independence assessments, which constitute a symmetrisation of irrelevance assessments. We say that the variables $X_n, n \in N$ are *epistemically independent* when learning the values of any number of them does not influence or change our beliefs about the remaining ones: for any two disjoint subsets I and O of N, X_I is epistemically irrelevant to X_O.

When does a set $\mathscr{D}_N$ of desirable gambles on $\mathscr{X}_N$ capture this type of epistemic independence?

Definition 2. *A coherent set $\mathscr{D}_N$ of desirable gambles on $\mathscr{X}_N$ is called* independent *if*

$$\begin{aligned}&\text{marg}_O(\mathscr{D}_N\rfloor x_I) = \text{marg}_O(\mathscr{D}_N)\\ &\qquad \textit{for all disjoint } I,O \subseteq N, \textit{ and all } x_I \in \mathscr{X}_I.\end{aligned}$$

In this definition, we allow I and O to be empty too, but doing so does not lead to any substantive requirement, because the condition $\text{marg}_O(\mathscr{D}_N\rfloor x_I) = \text{marg}_O(\mathscr{D}_N)$ is trivially satisfied when I or O are empty.

Independent sets have an interesting factorisation property (see Ref. [3] for another paper where factorisation is considered in this somewhat unusual form).

Proposition 14 (Factorisation). *Let $\mathscr{D}_N$ be an independent coherent set of desirable gambles on $\mathscr{X}_N$. Then for all disjoint subsets I and O of N and for all $f \in \mathscr{G}(\mathscr{X}_O)$:*

$$f \in \mathscr{D}_N \Leftrightarrow (\forall g \in \mathscr{G}(\mathscr{X}_I)_{>0}) fg \in \mathscr{D}_N. \tag{11}$$

Proof. Fix arbitrary disjoint subsets I and O of N and any $f \in \mathscr{G}(\mathscr{X}_O)$; we show that Eq. (11) holds. The '$\Leftarrow$' part is trivial. For the '$\Rightarrow$' part, assume that $f \in \mathscr{D}_N$ and consider any $g \in \mathscr{G}(\mathscr{X}_I)_{>0}$. We have to show that $fg \in \mathscr{D}_N$. Since $g = \sum_{x_I \in \mathscr{X}_I} \mathbb{I}_{\{x_I\}} g(x_I)$, we see that $fg = \sum_{x_I \in \mathscr{X}_I} g(x_I)\mathbb{I}_{\{x_I\}} f$. Now since $f \in \text{marg}_O(\mathscr{D}_N)$, we infer from the independence of $\mathscr{D}_N$ and the assumption (i) in Proposition 10 that $f \in \mathscr{D}_N\rfloor x_I$ and therefore $\mathbb{I}_{\{x_I\}} f \in \mathscr{D}_N$ for all $x_I \in \mathscr{X}_I$. We conclude that fg is a positive linear combination of elements $\mathbb{I}_{\{x_I\}} f$ of $\mathscr{D}_N$, and therefore belongs to $\mathscr{D}_N$ by coherence. □

Independence assessments are useful in constructing joint sets of desirable gambles from marginal ones. Suppose

we have coherent sets $\mathscr{D}_n$ of desirable gambles on $\mathscr{X}_n$, for each $n \in N$ and an assessment that the variables X_n, $n \in N$ are epistemically independent. Then how can we combine the $\mathscr{D}_n$ and this structural independence assessment into a coherent set of desirable gambles on $\mathscr{X}_N$ in a way that is as conservative as possible? If we call *independent product* of the $\mathscr{D}_n$ any independent $\mathscr{D}_N \in \mathbb{D}(\mathscr{X}_N)$ that marginalises to the $\mathscr{D}_n$ for all $n \in N$, this means we are looking for the smallest such independent product.

Further on, we are going to prove that such a smallest independent product always exists. Before we can do this elegantly, however, we need to do some preparatory work involving particular sets of desirable gambles that can be constructed from the $\mathscr{D}_n$. Consider, as a special case of Eq. (10), for any subset I of N and any $o \in N \setminus I$:

$$\mathscr{A}^{\text{irr}}_{I\to\{o\}} := \text{posi}\left(\left\{\mathbb{I}_{\{x_I\}}g\colon g \in \mathscr{D}_o \text{ and } x_I \in \mathscr{X}_I\right\}\right)$$

It is again easy to see that for all $h \in \mathscr{G}(\mathscr{X}_{I\cup\{o\}})$:

$$h \in \mathscr{A}^{\text{irr}}_{I\to\{o\}} \Leftrightarrow h \neq 0 \text{ and } (\forall x_I \in \mathscr{X}_I)h(x_I,\cdot) \in \mathscr{D}_o \cup \{0\}. \tag{12}$$

We use these sets to construct the following set of desirable gambles on $\mathscr{X}_N$:

$$\otimes_{n\in N}\mathscr{D}_n := \text{posi}\left(\mathscr{G}(\mathscr{X}_N)_{>0} \cup \bigcup_{n\in N} \mathscr{A}^{\text{irr}}_{N\setminus\{n\}\to\{n\}}\right). \tag{13}$$

Observe that, quite trivially, $\mathscr{A}^{\text{irr}}_{\{n\}\setminus\{n\}\to\{n\}} = \mathscr{D}_n$ and therefore $\otimes_{m\in\{n\}}\mathscr{D}_m = \mathscr{D}_n$. We now prove a number of important properties for $\otimes_{n\in N}\mathscr{D}_n$.

Proposition 15 (Coherence). *$\otimes_{n\in N}\mathscr{D}_n$ is a coherent set of desirable gambles on $\mathscr{X}_N$.*

Proof. Let, for ease of notation $\mathscr{A}_N := \bigcup_{n\in N}\mathscr{A}^{\text{irr}}_{N\setminus\{n\}\to\{n\}}$. It follows from Theorem 1 that we have to prove that $\mathscr{A}_N$ avoids non-positivity. So consider any $f \in \text{posi}(\mathscr{A}_N)$, and assume *ex absurdo* that $f \leq 0$. Then there are $\lambda_n \geq 0$ and $f_n \in \mathscr{A}^{\text{irr}}_{N\setminus\{n\}\to\{n\}}$ such that $f = \sum_{n\in N}\lambda_n f_n$ and $\max_{n\in N}\lambda_n > 0$ [recall that the $\mathscr{A}^{\text{irr}}_{N\setminus\{n\}\to\{n\}}$ are convex cones, by Lemma 11]. Fix arbitrary $m \in N$. Let

$$\mathscr{A}^N_m := \left\{f_m(x_{N\setminus\{m\}},\cdot)\colon x_{N\setminus\{m\}} \in \mathscr{X}_{N\setminus\{m\}}, f_m(x_{N\setminus\{m\}},\cdot) \neq 0\right\},$$

then it follows from Eq. (12) that $\mathscr{A}^N_m$ is a finite non-empty subset of $\mathscr{D}_m$, so the coherence of $\mathscr{D}_m$, Theorem 1 and Lemma 2 imply that there is some mass function p_m on $\mathscr{X}_m$ with expectation operator E_m such that $(\forall x_m \in \mathscr{X}_m)p_m(x_m) > 0$ and

$$(\forall x_{N\setminus\{m\}} \in \mathscr{X}_{N\setminus\{m\}})\\ (f_m(x_{N\setminus\{m\}},\cdot) \neq 0 \Rightarrow E_m(f_m(x_{N\setminus\{m\}},\cdot)) > 0).$$

So if we define the gamble $g_{N\setminus\{m\}}$ on $\mathscr{X}_{N\setminus\{m\}}$ by letting $g_{N\setminus\{m\}}(x_{N\setminus\{m\}}) := E_m(f_m(x_{N\setminus\{m\}},\cdot))$ for all $x_{N\setminus\{m\}} \in \mathscr{X}_{N\setminus\{m\}}$, then $g_{N\setminus\{m\}} > 0$.

Since we can do this for all $m \in N$, we can define the mass function p_N on $\mathscr{X}_N$ by letting $p_N(x_N) := \prod_{m\in N} p_m(x_m) > 0$ for all $x_N \in \mathscr{X}_N$. The corresponding expectation operator E_N is of course the product operator of the marginals E_m. But then it follows from the reasoning and assumptions above that $E_N(f) = \sum_{m\in N}\lambda_m E_N(f_m) = \sum_{m\in N}\lambda_m E_N(g_m) > 0$, whereas $f \leq 0$ leads us to conclude that $E_N(f) \leq 0$, a contradiction. □

Lemma 16. *Consider any disjoint subsets I, R of N and any $o \in N \setminus (I \cup R)$. Then $f(x_R,\cdot) \in \mathscr{A}^{\text{irr}}_{I\to\{o\}} \cup \{0\}$ for all $f \in \mathscr{A}^{\text{irr}}_{I\cup R\to\{o\}}$ and all $x_R \in \mathscr{X}_R$.*

Proof. Fix $f \in \mathscr{A}^{\text{irr}}_{I\cup R\to\{o\}}$ and $x_R \in \mathscr{X}_R$ and consider the gamble $g := f(x_R,\cdot)$ on $\mathscr{X}_{I\cup O}$. It follows from the assumptions that for all $x_I \in \mathscr{X}_I$, $g(x_I,\cdot) = f(x_R,x_I,\cdot) \in \mathscr{D}_o \cup \{0\}$, whence indeed $g \in \mathscr{A}^{\text{irr}}_{I\to\{o\}} \cup \{0\}$. □

Proposition 17 (Marginalisation). *Let R be any subset of N, then* $\text{marg}_R(\otimes_{n\in N}\mathscr{D}_n) = \otimes_{r\in R}\mathscr{D}_r$.

Proof. Since we are interpreting gambles on $\mathscr{X}_R$ as special gambles on $\mathscr{X}_N$, it is clear from Eq. (12) that for any $r \in R$, $\mathscr{A}^{\text{irr}}_{R\setminus\{r\}\to\{r\}} \subseteq \mathscr{A}^{\text{irr}}_{N\setminus\{r\}\to\{r\}}$. Eqs. (5) and (13) now tell us that $\text{ext}_N(\otimes_{r\in R}\mathscr{D}_r) \subseteq \otimes_{n\in N}\mathscr{D}_n$. If we invoke Eq. (6), this leads to $\otimes_{r\in R}\mathscr{D}_r = \text{marg}_R(\text{ext}_N(\otimes_{r\in R}\mathscr{D}_r)) \subseteq \text{marg}_R(\otimes_{n\in N}\mathscr{D}_n)$, so we can concentrate on the converse inclusion.

Consider therefore any $f \in \text{marg}_R(\otimes_{n\in N}\mathscr{D}_n) = (\otimes_{n\in N}\mathscr{D}_n) \cap \mathscr{G}(\mathscr{X}_R)$, and assume *ex absurdo* that $f \notin \otimes_{r\in R}\mathscr{D}_r$.

It follows from the coherence of $\otimes_{n\in N}\mathscr{D}_n$ [see Proposition 15] that $f \neq 0$. Since $f \in \otimes_{n\in N}\mathscr{D}_n$, there are $S \subseteq N$, $f_s \in \mathscr{A}^{\text{irr}}_{N\setminus\{s\}\to\{s\}}$, $s \in S$ and $g \in \mathscr{G}(\mathscr{X}_N)$ with $g \geq 0$ such that $f = g + \sum_{s\in S} f_s$. Clearly $S \setminus R \neq \emptyset$, because $S \setminus R = \emptyset$ would imply that, with $x_{N\setminus R}$ any element of $\mathscr{X}_{N\setminus R}$, $f = f(x_{N\setminus R},\cdot) = g(x_{N\setminus R},\cdot) + \sum_{s\in S\cap R} f_s(x_{N\setminus R},\cdot) \in \otimes_{r\in R}\mathscr{D}_r$, since we infer from Lemma 16 that $f_s(x_{N\setminus R},\cdot) \in \mathscr{A}^{\text{irr}}_{R\setminus\{s\}\to\{s\}} \cup \{0\}$ for all $s \in S \cap R$.

It follows from the coherence of $\otimes_{r\in R}\mathscr{D}_r$ [Proposition 15], $f \notin \otimes_{r\in R}\mathscr{D}_r$ and Lemma 3 that $0 \notin \text{posi}(\{-f\} \cup \otimes_{r\in R}\mathscr{D}_r)$. Let, for ease of notation, $\mathscr{A}^N_{S\cap R}$ be the set

$$\left\{f_s(z_{N\setminus R},\cdot)\colon s \in S\cap R, z_{N\setminus R} \in \mathscr{X}_{N\setminus R}, f_s(z_{N\setminus R},\cdot) \neq 0\right\}.$$

Then $\mathscr{A}^N_{S\cap R}$ is clearly a finite subset of $\otimes_{r\in R}\mathscr{D}_r$ [to see this, use a similar argument as above, involving Lemma 16], so we infer from Lemma 2 that there is some mass function p_R on $\mathscr{X}_R$ with associated expectation operator E_R such that

$$\begin{cases}(\forall x_R \in \mathscr{X}_R)p_R(x_R) > 0\\ (\forall s \in S\cap R)(\forall z_{N\setminus R} \in \mathscr{X}_{N\setminus R})E_R(f_s(z_{N\setminus R},\cdot)) \geq 0\\ E_R(f) < 0.\end{cases}$$

Since $f = f(z_{N\setminus R},\cdot)$ for any choice of $z_{N\setminus R}$ in $\mathscr{X}_{N\setminus R}$, we see that $f = g(z_{N\setminus R},\cdot) + \sum_{s\in S\cap R} f_s(z_{N\setminus R},\cdot) + \sum_{s\in S\setminus R} f_s(z_{N\setminus R},\cdot)$, whence:

$$\begin{aligned}0 &> E_R(f) - E_R(g(z_{N\setminus R},\cdot)) - \sum_{s\in S\cap R} E_R(f_s(z_{N\setminus R},\cdot))\\ &= \sum_{s\in S\setminus R} E_R(f_s(z_{N\setminus R},\cdot)) = \sum_{s\in S\setminus R}\sum_{x_R\in\mathscr{X}_R} p_R(x_R)f_s(z_{N\setminus R},x_R).\end{aligned}$$

The gambles $f_s(\cdot,x_R)$ on $\mathscr{X}_{N\setminus R}$, with $x_R \in \mathscr{X}_R$ and $s \in S \setminus R$, can clearly not all be zero. The non-zero ones all belong to $\otimes_{s\in N\setminus R}\mathscr{D}_s$, by Lemma 16, so the coherence of the set of desirable gambles $\otimes_{s\in N\setminus R}\mathscr{D}_s$ [Proposition 15] guarantees that their positive linear combination $h := \sum_{s\in S\setminus R}\sum_{x_R\in\mathscr{X}_R} p_R(x_R)f_s(\cdot,x_R)$ also belongs to $\otimes_{s\in N\setminus R}\mathscr{D}_s$. This contradicts $h < 0$. Hence indeed $f \in \otimes_{r\in R}\mathscr{D}_r$.□

Proposition 18 (Conditioning). *$\otimes_{n\in N}\mathscr{D}_n$ is independent: for all disjoint subsets I and O of N, and all $x_I \in \mathscr{X}_I$,*

$$\mathrm{marg}_O(\otimes_{n\in N}\mathscr{D}_n\rfloor x_I) = \mathrm{marg}_O(\otimes_{n\in N}\mathscr{D}_n) = \otimes_{o\in O}\mathscr{D}_o.$$

This could probably be proved indirectly using the 'semi-graphoid' properties of conditional epistemic irrelevance, proved by Moral [7]; it appears we need reverse weak union, reverse decomposition, and contraction. Here we give a direct proof. Proposition 17 can also be seen as a special case of the present result for $I = \emptyset$.

Proof. Fix arbitrary disjoint subsets I and O of N, and arbitrary $x_I \in \mathscr{X}_I$. The second equality follows from Proposition 17, so we concentrate on proving that $\mathrm{marg}_O(\otimes_{n\in N}\mathscr{D}_n\rfloor x_I) = \otimes_{o\in O}\mathscr{D}_o$.

We first show that $\otimes_{o\in O}\mathscr{D}_o \subseteq \otimes_{n\in N}\mathscr{D}_n\rfloor x_I$. Consider any gamble $f \in \otimes_{o\in O}\mathscr{D}_o$, then we have to show that $\mathbb{I}_{\{x_I\}}f \in \otimes_{n\in N}\mathscr{D}_n$. By assumption, there are non-negative reals λ_o and μ, gambles $f_o \in \mathscr{A}^{\mathrm{irr}}_{O\setminus\{o\}\to\{o\}}$ for all $o \in O$ and $g \in \mathscr{G}(\mathscr{X}_O)_{>0}$ such that $f = \mu g + \sum_{o\in O}\lambda_o f_o$ and $\max\{\mu, \max_{o\in O}\lambda_o\} > 0$. Fix $o \in O$ and let $f'_o := \mathbb{I}_{\{x_I\}}f_o \in \mathscr{G}(\mathscr{X}_N)$. Then it follows from the definition of $\mathscr{A}^{\mathrm{irr}}_{O\setminus\{o\}\to\{o\}}$ that $f'_o(z_{N\setminus\{o\}},\cdot) = \mathbb{I}_{\{x_I\}}(z_I)f_o(z_{O\setminus\{o\}},\cdot) \in \mathscr{D}_o \cup \{0\}$ for all $z_{N\setminus\{o\}} \in \mathscr{X}_{N\setminus\{o\}}$. Since $f'_o \neq 0$, the definition of $\mathscr{A}^{\mathrm{irr}}_{N\setminus\{o\}\to\{o\}}$ tells us that $f'_o \in \mathscr{A}^{\mathrm{irr}}_{N\setminus\{o\}\to\{o\}}$. Similarly, if we let $g' := \mathbb{I}_{\{x_I\}}g \in \mathscr{G}(\mathscr{X}_N)$, then $g' > 0$. So it follows from Eq. (13) that indeed $\mathbb{I}_{\{x_I\}}f = \mu g' + \sum_{o\in O}\lambda_o f'_o \in \otimes_{n\in N}\mathscr{D}_n$.

We now turn to the converse inclusion $\otimes_{n\in N}\mathscr{D}_n\rfloor x_I \subseteq \otimes_{o\in O}\mathscr{D}_o$. Consider any gamble $f \in \mathscr{G}(\mathscr{X}_O)$ such that $\mathbb{I}_{\{x_I\}}f$ belongs to $\otimes_{n\in N}\mathscr{D}_n$ and assume *ex absurdo* that $f \notin \otimes_{o\in O}\mathscr{D}_o$. Let, for the sake of notational simplicity, $C := N\setminus(I\cup O)$.

It follows from the coherence of $\otimes_{n\in N}\mathscr{D}_n$ [Proposition 15] that $f \neq 0$. Since $\mathbb{I}_{\{x_I\}}f \in \otimes_{n\in N}\mathscr{D}_n$, there are $S \subseteq N$, $f_s \in \mathscr{A}^{\mathrm{irr}}_{N\setminus\{s\}\to\{s\}}$, $s \in S$ and $g \in \mathscr{G}(\mathscr{X}_N)$ with $g \geq 0$ such that $\mathbb{I}_{\{x_I\}}f = g + \sum_{s\in S} f_s$. Clearly $S\setminus O \neq \emptyset$, because $S\setminus O = \emptyset$ would imply that, with x_C any element of $\mathscr{X}_C$, $f = g(x_I,x_C,\cdot) + \sum_{s\in S\cap O} f_s(x_I,x_C,\cdot) \in \otimes_{o\in O}\mathscr{D}_o$, because $f_s(x_I,x_C,\cdot) \in \mathscr{A}^{\mathrm{irr}}_{O\setminus\{s\}\to\{s\}}$ for all $s \in S\cap O$ by Lemma 16.

It follows from the coherence of $\otimes_{o\in O}\mathscr{D}_o$ [Proposition 15], $f \notin \otimes_{o\in O}\mathscr{D}_o$ and Lemma 3 that $0 \notin \mathrm{posi}(\{-f\}\cup\otimes_{o\in O}\mathscr{D}_o)$. The set

$$\mathscr{A}^N_{S\cap O} := \{f_s(x_I,z_C,\cdot)\colon s \in S\cap O, z_C \in \mathscr{X}_C, f_s(x_I,z_C,\cdot) \neq 0\}$$

is clearly a finite subset of $\otimes_{o\in O}\mathscr{D}_o$ [use Lemma 16 again], so we infer from Lemma 2 that there is some mass function p_O on $\mathscr{X}_O$ with associated expectation operator E_O such that

$$\begin{cases} (\forall x_O \in \mathscr{X}_O)p_O(x_O) > 0 \\ (\forall s \in S\cap O)(\forall z_C \in \mathscr{X}_C)E_O(f_s(x_I,z_C,\cdot)) \geq 0 \\ E_O(f) < 0. \end{cases}$$

Since $f = g(x_I,z_C,\cdot) + \sum_{s\in S\cap O} f_s(x_I,z_C,\cdot) + \sum_{s\in S\setminus O} f_s(x_I,z_C,\cdot)$ for any choice of $z_C \in \mathscr{X}_C$, we see that:

$$\begin{aligned} 0 &> E_O(f) - E_O(g(x_I,z_C,\cdot)) - \sum_{s\in S\cap O} E_O(f_s(x_I,z_C,\cdot)) \\ &= \sum_{s\in S\setminus O} E_O(f_s(x_I,z_C,\cdot)) = \sum_{s\in S\setminus O}\sum_{x_O\in\mathscr{X}_O} p_O(x_O)f_s(x_I,z_C,x_O). \end{aligned}$$

Similarly, for any $z_C \in \mathscr{X}_C$ and any $z_I \in \mathscr{X}_I\setminus\{x_I\}$ we infer from $0 = g(z_I,z_C,\cdot) + \sum_{s\in S\cap O} f_s(z_I,z_C,\cdot) + \sum_{s\in S\setminus O} f_s(z_I,z_C,\cdot)$ that:

$$\begin{aligned} 0 &\geq -E_O(g(z_I,z_C,\cdot)) - \sum_{s\in S\cap O} E_O(f_s(z_I,z_C,\cdot)) \\ &= \sum_{s\in S\setminus O} E_O(f_s(z_I,z_C,\cdot)) = \sum_{s\in S\setminus O}\sum_{x_O\in\mathscr{X}_O} p_O(x_O)f_s(z_I,z_C,x_O). \end{aligned}$$

Hence $h := \sum_{s\in S\setminus O}\sum_{x_O\in\mathscr{X}_O} p_O(x_O)f_s(\cdot,\cdot,x_O) < 0$. The gambles $f_s(\cdot,\cdot,x_O)$ on $\mathscr{X}_{I\cup C}$, with $x_O \in \mathscr{X}_O$ and $s \in S\setminus O$, can clearly not all be zero. The non-zero ones all belong to $\otimes_{s\in I\cup C}\mathscr{D}_s$, by Lemma 16. But then the coherence of the set of desirable gambles $\otimes_{s\in I\cup C}\mathscr{D}_s$ [Proposition 15] guarantees that their positive linear combination h is an element of $\otimes_{c\in C}\mathscr{D}_c$ for which $h < 0$, a contradiction. Hence indeed $f \in \otimes_{o\in O}\mathscr{D}_o$. □

Theorem 19 (Independent natural extension). *$\otimes_{n\in N}\mathscr{D}_n$ is the smallest coherent set of desirable gambles on $\mathscr{X}_N$ that is an independent product of the coherent sets $\mathscr{D}_n$ of desirable gambles on $\mathscr{X}_n$, $n \in N$.*

We call $\otimes_{n\in N}\mathscr{D}_n$ the *independent natural extension* of the marginals $\mathscr{D}_n$.

Proof. It follows from Propositions 15, 17 and 18 that $\otimes_{n\in N}\mathscr{D}_n$ is an independent product $\mathscr{D}_N$ of the $\mathscr{D}_n$. To prove that it is the smallest one, consider any independent product $\mathscr{D}_N$ of the $\mathscr{D}_n$. Fix $n \in N$. If we consider any $x_{N\setminus\{n\}} \in \mathscr{X}_{N\setminus\{n\}}$, then $\mathrm{marg}_n(\mathscr{D}_N\rfloor x_{N\setminus\{n\}}) = \mathscr{D}_n$, by assumption. If we therefore consider any $g \in \mathscr{D}_n$, this in turn implies that $g \in \mathscr{D}_N\rfloor x_{N\setminus\{n\}}$, and therefore $\mathbb{I}_{\{x_{N\setminus\{n\}}\}}g \in \mathscr{D}_N$, by Eq. (8). So we infer by coherence that $\mathscr{A}^{\mathrm{irr}}_{N\setminus\{n\}\to\{n\}} \subseteq \mathscr{D}_N$, and therefore also that $\otimes_{n\in N}\mathscr{D}_n \subseteq \mathscr{D}_N$.□

Theorem 20 (Associativity). *Let N_1, N_2 be disjoint non-empty index sets, and let $\mathscr{D}_{n_k} \in \mathbb{D}(\mathscr{X}_{n_k})$, $n_k \in N_k$, $k = 1,2$. Then $\otimes_{n\in N_1\cup N_2}\mathscr{D}_n = (\otimes_{n_1\in N_1}\mathscr{D}_{n_1}) \otimes (\otimes_{n_2\in N_2}\mathscr{D}_{n_2})$.*

Proof. Consider, for ease of notation, $\mathscr{D}_{N_1} := \otimes_{n_1\in N_1}\mathscr{D}_{n_1}$ and $\mathscr{D}_{N_2} := \otimes_{n_2\in N_2}\mathscr{D}_{n_2}$. We have to prove that $\mathscr{D}_{N_1} \otimes \mathscr{D}_{N_2} = \otimes_{n\in N_1\cup N_2}\mathscr{D}_n$.

We first prove that $\mathscr{D}_{N_1} \otimes \mathscr{D}_{N_2} \subseteq \otimes_{n\in N_1\cup N_2}\mathscr{D}_n$. Fix any gamble $h \in \mathscr{A}^{\mathrm{irr}}_{\{N_1\}\to\{N_2\}}$ and any $x_{N_1} \in \mathscr{X}_{N_1}$, so $h(x_{N_1},\cdot) \in \mathscr{D}_{N_2} \cup \{0\}$ by Eq. (12). It follows from Eq. (13) that there are gambles $h^{n_2}_{x_{N_1}} \in \mathscr{A}^{\mathrm{irr}}_{N_2\setminus\{n_2\}\to\{n_2\}} \cup \{0\}$ for all $n_2 \in N_2$ such that $h(x_{N_1},\cdot) \geq \sum_{n_2\in N_2} h^{n_2}_{x_{N_1}}$. Define, for any $n_2 \in N_2$, the gamble g_{n_2} on $\mathscr{X}_N$ by letting $g_{n_2}(x_{N\setminus\{n_2\}},\cdot) := h^{n_2}_{x_{N_1}}(x_{N_2\setminus\{n_2\}},\cdot)$ for all $x_N \in \mathscr{X}_N$. Then it follows from Eq. (12) that $g_{n_2}(x_{N\setminus\{n_2\}},\cdot) \in \mathscr{D}_{n_2} \cup \{0\}$ for all $x_N \in \mathscr{X}_N$, and therefore $g_{n_2} \in \mathscr{A}^{\mathrm{irr}}_{N\setminus\{n_2\}\to\{n_2\}} \cup \{0\}$. Moreover,

$$\begin{aligned} h &= \sum_{x_{N_1}\in\mathscr{X}_{N_1}} \mathbb{I}_{\{x_{N_1}\}}h(x_{N_1},\cdot) \geq \sum_{x_{N_1}\in\mathscr{X}_{N_1}} \mathbb{I}_{\{x_{N_1}\}}\sum_{n_2\in N_2} h^{n_2}_{x_{N_1}} \\ &= \sum_{n_2\in N_2}\sum_{x_{N_1}\in\mathscr{X}_{N_1}} \mathbb{I}_{\{x_{N_1}\}}h^{n_2}_{x_{N_1}} = \sum_{n_2\in N_2} g_{n_2}, \end{aligned}$$

Since clearly $h \neq 0$, we infer from Eq. (13) that $h \in \otimes_{n\in N_1\cup N_2}\mathscr{D}_n$. We conclude that $\mathscr{A}^{\mathrm{irr}}_{\{N_1\}\to\{N_2\}} \subseteq \otimes_{n\in N_1\cup N_2}\mathscr{D}_n$. Similarly, we can prove the inclusion $\mathscr{A}^{\mathrm{irr}}_{\{N_2\}\to\{N_1\}} \subseteq \otimes_{n\in N_1\cup N_2}\mathscr{D}_n$, and therefore also $\mathscr{D}_{N_1} \otimes \mathscr{D}_{N_2} \subseteq \otimes_{n\in N_1\cup N_2}\mathscr{D}_n$, again by Eq. (13).

To conclude, we turn to the converse inclusion $\otimes_{n\in N_1\cup N_2}\mathscr{D}_n \subseteq \mathscr{D}_{N_1} \otimes \mathscr{D}_{N_2}$. Consider any gamble $h \in \otimes_{n\in N_1\cup N_2}\mathscr{D}_n$, then by Eq. (13) there are $h_n \in \mathscr{A}^{\mathrm{irr}}_{N_1\cup N_2\setminus\{n\}\to\{n\}} \cup \{0\}$, $n \in N_1\cup N_2$,

such that $h \geq h_1 + h_2$, where we let $h_1 := \sum_{n_1 \in N_1} h_{n_1}$ and $h_2 := \sum_{n_2 \in N_2} h_{n_2}$. Fix any $x_{N_1} \in \mathscr{X}_{N_1}$. For any $n_2 \in N_2$, we infer that $h_{n_2}(x_{N_1}, \cdot) \in \mathscr{A}^{\mathrm{irr}}_{N_2 \setminus \{n_2\} \to \{n_2\}} \cup \{0\}$ from $h_{n_2} \in \mathscr{A}^{\mathrm{irr}}_{N_1 \cup N_2 \setminus \{n_2\} \to \{n_2\}} \cup \{0\}$ by Lemma 16. Hence $h_2(x_{N_1}, \cdot) \in \mathscr{D}_{N_2} \cup \{0\}$ by Eq. (13), and therefore $h_2 \in \mathscr{A}^{\mathrm{irr}}_{\{N_1\} \to \{N_2\}} \cup \{0\}$ by Eq. (12). Similarly, $h_1 \in \mathscr{A}^{\mathrm{irr}}_{\{N_2\} \to \{N_1\}} \cup \{0\}$, and therefore $h \in \mathscr{D}_{N_1} \otimes \mathscr{D}_{N_2}$ by Eq. (13), since clearly $h \neq 0$. □

To conclude this section, we establish a connection between independent natural extension for sets of desirable gambles and the eponymous notion for coherent lower previsions studied in detail in Ref. [3]. Given coherent lower previsions $\underline{P}_n$ on $\mathscr{G}(\mathscr{X}_n)$, $n \in N$, their *independent natural extension* is the coherent lower prevision given by

$$\underline{E}_N(f) := \sup_{\substack{h_n \in \mathscr{G}(\mathscr{X}_N) \\ n \in N}} \min_{z_N \in \mathscr{X}_N} \left[f(z_N) - \sum_{n \in N} [h_n(z_N) - \underline{P}_n(h_n(\cdot, z_{N \setminus \{n\}}))] \right] \tag{14}$$

for all gambles f on $\mathscr{X}_N$. It is the point-wise smallest (most conservative) joint lower prevision that is jointly coherent with the marginals $\underline{P}_n$ given an assessment of epistemic independence of the variables X_n, $n \in N$.

Theorem 21. *Let $\mathscr{D}_n$ be coherent sets of desirable gambles on $\mathscr{X}_n$ for $n \in N$, and let $\otimes_{n \in N} \mathscr{D}_n$ be their independent natural extension. Consider the coherent lower previsions $\underline{P}_n$ on $\mathscr{G}(\mathscr{X}_n)$ given by $\underline{P}_n(f_n) := \sup\{\mu \in \mathbb{R} \colon f_n - \mu \in \mathscr{D}_n\}$ for all $f_n \in \mathscr{G}(\mathscr{X}_n)$. Then the independent natural extension $\underline{E}_N$ of the marginal lower previsions $\underline{P}_n$, $n \in N$ satisfies*

$$\underline{E}_N(f) = \sup\{\mu \in \mathbb{R} \colon f - \mu \in \otimes_{n \in N} \mathscr{D}_n\}$$

for all gambles f on $\mathscr{X}_N$.

Proof. Fix any gamble f in $\mathscr{G}(\mathscr{X}_N)$. First, consider any real number $\mu < \underline{E}_N(f)$, then it follows from Eq. (14) that there are $\delta > 0$ and $h_n \in \mathscr{G}(\mathscr{X}_N)$, $n \in N$, such that $f - \mu \geq \sum_{n \in N} g_n$, where we defined the gambles g_n on $\mathscr{X}_N$ by $g_n(z_N) := h_n(z_N) - \underline{P}_n(h_n(z_{N \setminus \{n\}}, \cdot)) + \delta$ for all $z_N \in \mathscr{X}_N$. It follows from the definition of $\underline{P}_n$ that $g_n(z_{N \setminus \{n\}}, \cdot) = h_n(z_{N \setminus \{n\}}, \cdot) - \underline{P}_n(h_n(z_{N \setminus \{n\}}, \cdot)) + \delta \in \mathscr{D}_n$ for all $z_{N \setminus \{n\}} \in \mathscr{X}_{N \setminus \{n\}}$. Since clearly $g_n \neq 0$, Eq. (12) then tells us that $g_n \in \mathscr{A}^{\mathrm{irr}}_{N \setminus \{n\} \to \{n\}}$, and we infer from Eq. (13) that $\sum_{n \in N} g_n \in \otimes_{n \in N} \mathscr{D}_n$, and therefore also $f - \mu \in \otimes_{n \in N} \mathscr{D}_n$. This guarantees that $\underline{E}_N(f) \leq \sup\{\mu \in \mathbb{R} \colon f - \mu \in \otimes_{n \in N} \mathscr{D}_n\}$.

To prove the converse inequality, consider any real μ such that $f - \mu \in \otimes_{n \in N} \mathscr{D}_n$. We infer using Eq. (13) that there are gambles $h_n \in \mathscr{A}^{\mathrm{irr}}_{N \setminus \{n\} \to \{n\}} \cup \{0\}$, $n \in N$, such that $f - \mu \geq \sum_{n \in N} h_n$. For all $n \in N$ and $z_{N \setminus \{n\}} \in \mathscr{X}_{N \setminus \{n\}}$, it follows from Eq. (12) that $h_n(z_{N \setminus \{n\}}, \cdot) \in \mathscr{D}_n \cup \{0\}$, whence $\underline{P}_n(h_n(z_{N \setminus \{n\}}, \cdot)) \geq 0$. This leads to $\sum_{n \in N} [h_n(z_N) - \underline{P}_n(h_n(z_{N \setminus \{n\}}, \cdot))] \leq \sum_{n \in N} h_n(z_N) \leq f(z_N) - \mu$. We then infer from Eq. (14) that $\underline{E}_N(f) \geq \mu$ and so we find that indeed also $\underline{E}_N(f) \geq \sup\{\mu \in \mathbb{R} \colon f - \mu \in \otimes_{n \in N} \mathscr{D}_n\}$. □

8 Maximal sets of desirable gambles and strong products

The following result was (essentially) proved in Ref. [1].

Proposition 22. *Let $\mathscr{M}_N \in \mathbb{M}(\mathscr{X}_N)$, and consider any disjoint subsets I and O of N. Then $\mathrm{marg}_O(\mathscr{M}_N \rfloor x_I) \in \mathbb{M}(\mathscr{X}_O)$ for all $x_I \in \mathscr{X}_I$.*

Now consider the case where we have coherent marginal sets of desirable gambles $\mathscr{D}_n$ for all $n \in N$. We define their *strong product* $\boxtimes_{n \in N} \mathscr{D}_n$ as the set of desirable gambles on the product space $\mathscr{X}_N$ given by:

$$\boxtimes_{n \in N} \mathscr{D}_n := \bigcap \{\otimes_{n \in N} \mathscr{M}_n \colon \mathscr{M}_n \in m(\mathscr{D}_n), n \in N\}$$

Observe that for maximal sets $\mathscr{M}_n \in \mathbb{M}(\mathscr{X}_n)$, $n \in N$ the strong product and the independent natural extension coincide: $\boxtimes_{n \in N} \mathscr{M}_n = \otimes_{n \in N} \mathscr{M}_n$.

The marginalisation properties of the strong product follow from those of the independent natural extension.

Proposition 23 (Marginalisation). *Consider coherent sets of desirable gambles $\mathscr{D}_n$ for all $n \in N$. Let R be any subset of N, then $\mathrm{marg}_R(\boxtimes_{n \in N} \mathscr{D}_n) = \boxtimes_{r \in R} \mathscr{D}_r$.*

Proof. Consider any $f \in \mathscr{G}(\mathscr{X}_R)$ and observe the following chain of equivalences:

$$\begin{aligned} f \in \boxtimes_{n \in N} \mathscr{D}_n &\Leftrightarrow (\forall \mathscr{M}_n \in m(\mathscr{D}_n), n \in N) f \in \otimes_{n \in N} \mathscr{M}_n \\ &\Leftrightarrow (\forall \mathscr{M}_n \in m(\mathscr{D}_n), n \in N) f \in \otimes_{r \in R} \mathscr{M}_r \\ &\Leftrightarrow (\forall \mathscr{M}_r \in m(\mathscr{D}_r), r \in R) f \in \otimes_{r \in R} \mathscr{M}_r \\ &\Leftrightarrow f \in \boxtimes_{r \in R} \mathscr{D}_r, \end{aligned}$$

where the second equivalence follows from Proposition 17. □

As we have come to expect from our treatment of the independent natural extension, the proof of the following independence property is very similar to that of the marginalisation property.

Proposition 24. *Consider coherent sets of desirable gambles $\mathscr{D}_n$ for all $n \in N$. Then their strong product $\boxtimes_{n \in N} \mathscr{D}_n$ is an independent product of these marginals.*

Proof. Consider any disjoint subsets I and O of N, and any $x_I \in \mathscr{X}_I$, then it suffices to prove that, also using Proposition 23, $\mathrm{marg}_O(\boxtimes_{n \in N} \mathscr{D}_n \rfloor x_I) = \boxtimes_{o \in O} \mathscr{D}_o$. So consider any gamble f on $\mathscr{X}_O$ and observe the following chain of equivalences:

$$\begin{aligned} f \in \boxtimes_{n \in N} \mathscr{D}_n \rfloor x_I &\Leftrightarrow \mathbb{I}_{\{x_I\}} f \in \boxtimes_{n \in N} \mathscr{D}_n \\ &\Leftrightarrow (\forall \mathscr{M}_n \in m(\mathscr{D}_n), n \in N) \mathbb{I}_{\{x_I\}} f \in \otimes_{n \in N} \mathscr{M}_n \\ &\Leftrightarrow (\forall \mathscr{M}_n \in m(\mathscr{D}_n), n \in N) f \in \otimes_{o \in O} \mathscr{M}_o \\ &\Leftrightarrow (\forall \mathscr{M}_o \in m(\mathscr{D}_o), o \in O) f \in \otimes_{o \in O} \mathscr{M}_o \\ &\Leftrightarrow f \in \boxtimes_{o \in O} \mathscr{D}_o, \end{aligned}$$

where the third equivalence follows from Proposition 18. □

It is still an open problem at this point whether, like the natural extension, the strong product is associative.

To conclude this section, we establish a connection between the strong product of sets of desirable gambles and the eponymous notion for coherent lower previsions, studied in Ref. [3]. Given coherent lower previsions $\underline{P}_n$ on $\mathscr{G}(\mathscr{X}_n)$, $n \in N$, their *strong product* is defined by

$$\underline{S}_N(f) := \inf\{\times_{n\in N} P_n(f) \colon (\forall n \in N) P_n \in \mathscr{M}(\underline{P}_n)\}$$

for all gambles f on $\mathscr{X}_N$. If we start from linear previsions P_n on $\mathscr{G}(\mathscr{X}_n)$, their strong product corresponds to their linear product $\times_{n\in N} P_n$, and it coincides also with their independent natural extension E_N. If we begin with coherent lower previsions $\underline{P}_n$ on $\mathscr{G}(\mathscr{X}_n)$, their strong product $\underline{S}_N$ is the lower envelope of the set of strong products determined by the dominating linear previsions.

Theorem 25. *Let $\mathscr{D}_n$ be coherent sets of desirable gambles in $\mathscr{G}(\mathscr{X}_n)$ for all $n \in N$, and let $\boxtimes_{n\in N}\mathscr{D}_n$ be their strong product. Consider the coherent lower previsions $\underline{P}_n$ on $\mathscr{G}(\mathscr{X}_n)$ given by $\underline{P}_n(f) := \sup\{\mu \in \mathbb{R} \colon f - \mu \in \mathscr{D}_n\}$. Then the strong product $\underline{S}_N$ of the marginal lower previsions $\underline{P}_n$, $n \in N$ satisfies $\underline{S}_N(f) = \sup\{\mu \in \mathbb{R} \colon f - \mu \in \boxtimes_{n\in N}\mathscr{D}_n\}$.*

Proof. Assume first of all that $\mathscr{D}_n$ is a maximal set of desirable gambles for all n in N. Then it follows from Theorem 3.8.3 in Ref. [8] that $\underline{P}_n$ is a linear prevision, which we denote by P_n, for all $n \in N$. The strong product of the linear previsions P_n, $n \in N$ coincides with their linear independent product $\times_{n\in N} P_n$, which is also their independent natural extension, by Proposition 10 in Ref. [3]. Since we have proved in Theorem 21 that this is the coherent lower prevision associated with $\otimes_{n\in N}\mathscr{D}_n = \boxtimes_{n\in N}\mathscr{D}_n$, we conclude that the strong product $\boxtimes_{n\in N}\mathscr{D}_n$ is associated with the strong product of the linear previsions P_n.

Next, fix any gamble f on $\mathscr{X}_N$. Consider any real number $\mu < \underline{S}_N(f)$. For any $n \in N$, consider any maximal set $\mathscr{M}_n \in m(\mathscr{D}_n)$, and the associated linear prevision P_n, then clearly $P_n \in \mathscr{M}(\underline{P}_n)$. Hence $\times_{n\in N} P_n(f) \geq \underline{S}_N(f) > \mu$, and we infer from the arguments above that then necessarily $f - \mu \in \otimes_{n\in N}\mathscr{M}_n$. Hence $f - \mu \in \boxtimes_{n\in N}\mathscr{D}_n$. This leads to the conclusion that $\underline{S}_N(f) \leq \sup\{\mu \in \mathbb{R} \colon f - \mu \in \boxtimes_{n\in N}\mathscr{D}_n\}$.

Conversely, consider any real μ such that $f - \mu \in \boxtimes_{n\in N}\mathscr{D}_n$. Consider arbitrary $P_n \in \mathscr{M}(\underline{P}_n)$, $n \in N$, then there are maximal sets $\mathscr{M}_n \in m(\mathscr{D}_n)$ inducing them: indeed, the set of strictly desirable gambles $\mathscr{D}_n'$ that induces P_n, given by Eq. (2), is coherent by Theorem 3.8.1 in Ref. [8]; Theorem 5 implies that there is some maximal set $\mathscr{M}_n \in m(\mathscr{D}_n') \supseteq m(\mathscr{D}_n)$, and now Theorem 3.8.3 in Ref. [8] implies that $\mathscr{D}_n'$ and $\mathscr{M}_n$ induce the same P_n by means of Eq. (1). But then $f - \mu \in \otimes_{n\in N}\mathscr{M}_n$, and therefore $\times_{n\in N} P_n(f) \geq \mu$, using the argumentation above. Hence $\underline{S}_N(f) \geq \mu$, and therefore also $\underline{S}_N(f) \geq \sup\{\mu \in \mathbb{R} \colon f - \mu \in \boxtimes_{n\in N}\mathscr{D}_n\}$. □

Together with Theorem 21 and the fact that the strong product of lower previsions may strictly dominate their independent natural extension [see Example 9.3.4 in Ref. [8]], this shows that the strong product of marginal sets of desirable gambles may strictly include their independent natural extension.

9 Conditional irrelevance and independence

We turn to conditional irrelevance judgements. Next to the variables X_N in $\mathscr{X}_N$, we now also consider another variable Y assuming values in a finite set $\mathscr{Y}$.

Consider two disjoint subsets I and O of N. We say that X_I is *epistemically irrelevant* to X_O when, conditional on Y, learning the value of X_I does not influence or change our beliefs about X_O. In order for a set $\mathscr{D}$ of desirable gambles on $\mathscr{X}_N \times \mathscr{Y}$ to capture this type of conditional epistemic irrelevance, we should require that:

$$\mathrm{marg}_O(\mathscr{D}\rfloor x_I, y) = \mathrm{marg}_O(\mathscr{D}\rfloor y)\ \forall x_I \in \mathscr{X}_I, y \in \mathscr{Y}.$$

As before, for technical reasons we also allow I and O to be empty. It is clear from the definition above that the 'variable' $X_\emptyset$, about whose constant value we are certain, is conditionally epistemically irrelevant to any variable X_O. Similarly, we see that any variable X_I is conditionally epistemically irrelevant to the 'variable' $X_\emptyset$. This seems to be in accordance with intuition.

Also, if $\mathscr{Y}$ is a singleton, then there is no uncertainty about Y and conditioning on Y amounts to not conditioning at all: epistemic irrelevance can be seen as a special case of conditional epistemic irrelevance. We now want to argue that, conversely, there is a very specific and definite way in which conditional epistemic irrelevance statements can be reduced to simple epistemic irrelevance statements. The crucial results that allow us to establish this, are the following conceptually very simple theorem and its corollary.

Theorem 26 (Sequential updating). *Consider any subset R of N, and any coherent set $\mathscr{D}$ of desirable gambles on $\mathscr{X}_N \times \mathscr{Y}$. Then*

$$(\mathscr{D}\rfloor y)\rfloor x_R = (\mathscr{D}\rfloor x_R)\rfloor y = \mathscr{D}\rfloor x_R, y \quad \text{for all } x_R \in \mathscr{X}_R \text{ and } y \in \mathscr{Y}. \tag{15}$$

Proof. Fix any x_R in $\mathscr{X}_R$ and any $y \in \mathscr{Y}$. Clearly, all three sets in Eq. (15) are subsets of $\mathscr{G}(\mathscr{X}_{N\setminus R})$. So take any gamble f on $\mathscr{X}_{N\setminus R}$, and consider the following chains of equivalences:

$$\mathbb{I}_{\{y\}}\mathbb{I}_{\{x_R\}} f \in \mathscr{D} \Leftrightarrow \mathbb{I}_{\{x_R\}} f \in \mathscr{D}\rfloor y \Leftrightarrow f \in (\mathscr{D}\rfloor y)\rfloor x_R$$
$$\mathbb{I}_{\{y\}}\mathbb{I}_{\{x_R\}} f \in \mathscr{D} \Leftrightarrow \mathbb{I}_{\{y\}} f \in \mathscr{D}\rfloor x_R \Leftrightarrow f \in (\mathscr{D}\rfloor x_R)\rfloor y$$
$$\mathbb{I}_{\{y\}}\mathbb{I}_{\{x_R\}} f \in \mathscr{D} \Leftrightarrow f \in \mathscr{D}\rfloor x_R, y.$$ □

Corollary 27 (Reduction). *Consider any disjoint subsets I and O of N, and any coherent set $\mathscr{D}$ of desirable gambles on $\mathscr{X}_N \times \mathscr{Y}$. Then the following statements are equivalent:*

(i) $\mathrm{marg}_O(\mathscr{D}\rfloor x_I, y) = \mathrm{marg}_O(\mathscr{D}\rfloor y)$ *for all $x_I \in \mathscr{X}_I$ and all $y \in \mathscr{Y}$;*

(ii) $\mathrm{marg}_O((\mathscr{D}\rfloor y)\rfloor x_I) = \mathrm{marg}_O(\mathscr{D}\rfloor y)$ *for all $x_I \in \mathscr{X}_I$ and all $y \in \mathscr{Y}$.*

This tells us that *a model $\mathscr{D}$ about (X_N, Y) captures epistemic irrelevance of X_I to X_O, conditional on Y if and only if for each possible value $y \in \mathscr{Y}$ of Y, the model $\mathscr{D}\rfloor y$ about X_N captures epistemic irrelevance of X_I to X_O.*

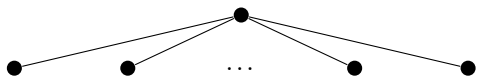
Now suppose we have marginal conditional models $\mathscr{D}_n\rfloor Y$ on $\mathscr{X}_n$, $n \in N$. The notation $\mathscr{D}_n\rfloor Y$ is a concise way of representing the family of conditional models $\mathscr{D}_n\rfloor y$, $y \in \mathscr{Y}$. Then if we combine Corollary 27 and Theorem 19, we see that the smallest conditionally independent product $\mathscr{D}\rfloor Y$ of these marginal models $\mathscr{D}_n\rfloor Y$ is given by $\otimes_{n\in N}(\mathscr{D}_n\rfloor Y)$, meaning that for each $y \in \mathscr{Y}$, $\mathscr{D}\rfloor y = \otimes_{n\in N}(\mathscr{D}_n\rfloor y)$.

10 Conclusions

Sets of desirable gambles are more informative than coherent lower previsions, and they are helpful in avoiding problems involving zero probabilities. They have been overlooked for much of the development of the theory, and it is only in the last five or six years that more effort is being devoted to bringing this simplifying and unifying notion to the fore.

Our results here show that we can model assessments of epistemic independence easily using sets of desirable gambles, and that we can derive from them existing results for lower previsions.

They also indicate that constructing global joint models (i.e. coherent sets of desirable gambles) from local ones is something that can be easily and efficiently done for the following types of simple credal networks:



They may therefore open up the way towards finding efficient algorithms for inference in credal *trees* under epistemic irrelevance using sets of desirable gambles as uncertainty models, building on the ideas proposed in Ref. [2]. We expect that generalising those algorithms towards more general credal networks (polytrees, ...) will be more difficult, and will have to rely heavily on the pioneering work of Moral [7] on graphoid properties for epistemic irrelevance.

Acknowledgements

This work was supported by SBO project 060043 of the IWT-Vlaanderen, and by projects TIN2008-06796-C04-01 and MTM2010-17844.

References

[1] Inés Couso and Serafín Moral. Sets of desirable gambles and credal sets. In Thomas Augustin, Frank P. A. Coolen, Serafin Moral, and Matthias C. M. Troffaes, editors, *ISIPTA '09: proceedings of the Sixth International Symposium on Imprecise Probabilities: Theories and Applications*, pages 99–108, Durham, United Kingdom, 2009. SIPTA.

[2] Gert de Cooman, Filip Hermans, Alessandro Antonucci, and Marco Zaffalon. Epistemic irrelevance in credal nets: the case of imprecise Markov trees. *International Journal of Approximate Reasoning*, 51:1029–1052, 2010.

[3] Gert de Cooman, Enrique Miranda, and Marco Zaffalon. Independent natural extension. *Artificial Intelligence*, 2010. Accepted for publication.

[4] Gert de Cooman and Erik Quaeghebeur. Exchangeability and sets of desirable gambles. *International Journal of Approximate Reasoning*, 2010. In press. Special issue in honour of Henry E. Kyburg, Jr.

[5] Lester E. Dubins. Finitely additive conditional probabilities, conglomerability and disintegrations. *The Annals of Probability*, 3:88–99, 1975.

[6] Enrique Miranda and Marco Zaffalon. Notes on desirability and coherent lower previsions. *Annals of Mathematics and Artificial Intelligence*, 2011. In press.

[7] Serafín Moral. Epistemic irrelevance on sets of desirable gambles. *Annals of Mathematics and Artificial Intelligence*, 45:197–214, 2005.

[8] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[9] Peter Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24:125–148, 2000.

[10] Peter 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.

Modelling Uncertainties in Limit State Functions

Thomas Fetz
Institut für Grundlagen der Bauingenieurwissenschaften
Arbeitsbereich Technische Mathematik
Universität Innsbruck, Austria
Thomas.Fetz@uibk.ac.at

Abstract

In this paper uncertainties in limit state functions g as arising in engineering problems are modelled by adding additional parameters and by introducing parameterized probability density functions which describe the uncertainties of these new additional parameters and of the basic variables of g. This will lead to a function $p_f(a,b)$ for the probability of failure depending on parameters a and b corresponding to the two parameterized density functions. Further the parameters a and b are assumed to be uncertain. Using intervals, sets or random sets to model their uncertainty results in upper probabilities $\overline{p}_f$ of failure. In this context we also discuss different notions of independence such as strong independence, epistemic irrelevance and random set independence and present a simple engineering example.

Keywords. Probability of failure, limit state functions, parameterized probability measures, random sets, random set independence, epistemic irrelevance, strong independence.

1 Introduction

In reliability analysis the probability p_f of failure of a system is obtained by

$$p_f = \int\limits_{\{x:\, g(x)\leq 0\}} f^X(x)\,\mathrm{d}x \tag{1}$$

where $x=(x_1,\ldots,x_n)$ are the basic variables of the system such as material properties and loads and where f^X is a probability density function describing the uncertainty of the variables x. The function g is the limit state function of the system telling us for which x the system fails ($g(x)\leq 0$) or not ($g(x)>0$), see also [14].

In the case of scarce information about the values of the basic variables x and the behaviour of the system it may be neither sufficient to model the uncertainty of the variables x by a single probability density f^X nor to describe the system's reliability by a single deterministic limit state function g. To overcome such difficulties, fuzzy sets [17], random sets [3], credal sets [13] or sets of parameterized probability measures [9] have been used to model the uncertainty of the variables x, cf. also [6, 8, 10, 11]. Uncertainties in the limit state function g have been modelled using additional random variables [5], fuzzy sets, random sets [12] or fuzzy probabilities [1, 15].

The aim of this paper is to develop a function

$$p_f(a,b) = \iint\limits_{\{(x,z):\, h(x,z)\leq 0\}} f_b^Z(z)\,\mathrm{d}z\, f_a^X(x)\,\mathrm{d}x \tag{2}$$

depending on vectors of parameters a and b parameterizing the probability density functions f_a^X and f_b^Z. These density functions describe the uncertainty of the basic variables x and the additional parameters z of an extended limit state function h. These additional variables z are used to parameterize a familiy of limit state functions g_z with $g_z(x)=h(x,z)$.

In a next step we assume that the parameters a and b are uncertain themselves modelling their uncertainty by intervals, sets or random sets. This approach gives us the possibility to describe the uncertainty of x and z by sets of probability measures generated by the density functions f_a^X and f_b^Z and their uncertain parameters a and b. The functions f_a^X and f_b^Z allow us to use more specific probability measures such as Gaussian distributions in contrast to the case where the uncertainty of x and z is directly modelled by sets or random sets. Such coarser models of uncertainty are also encompassed simply by replacing f_a^X and f_b^Z by Dirac measures.

A simple engineering example with one uncertain basic variable x exemplifies different cases and models of uncertainty of a and b and the computation of the upper probability $\overline{p}_f$ of failure by means of $p_f(a,b)$ with respect to different notions of indepence between the limit state functions and the basic variables.

2 Uncertain limit state functions

2.1 Limit state functions

In reliability theory a system and its corresponding continuous *limit state function*

$$g : \mathcal{X} \subseteq \mathbb{R}^n \to \mathcal{Y} \subseteq \mathbb{R} : \; x \to y = g(x) \tag{3}$$

is given with output $y \in \mathcal{Y}$ depending on a vector of *n basic variables* $x = (x_1, \ldots, x_n) \in \mathcal{X} \subseteq \mathbb{R}^n$ where $g(x) \leq 0$ means failure of the system. The *probability p_f of failure* of the system is then defined by

$$p_f = P(g(X) \leq 0) = \int_{\mathcal{X}} \chi(g(x) \leq 0) f^X(x) \, \mathrm{d}x \tag{4}$$

where f^X is the joint probability density function of the basic random variables $X = (X_1, \ldots, X_n)$ and where

$$\chi(\text{expression}) = \begin{cases} 1 & \text{expression true,} \\ 0 & \text{expression false.} \end{cases} \tag{5}$$

The set $R_f = \{x \in \mathcal{X} : \; g(x) \leq 0\}$ is the *failure region* of the system which we describe by the indicator function

$$q : \mathcal{X} \to \{0,1\} : x \to \chi(g(x) \leq 0). \tag{6}$$

2.2 Parameterized limit state functions

We parameterize the limit state function $g : \mathcal{X} \to \mathcal{Y}$ by means of a vector $z = (z_1, \ldots, z_m) \in \mathcal{Z} \subseteq \mathbb{R}^m$ of additional parameters using a function

$$h : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y} : (x,z) \to h(x,z) \tag{7}$$

where again $h(x,z) \leq 0$ means failure. A function

$$g_z : \mathcal{X} \to \mathcal{Y} : x \to g_z(x) = h(x,z) \tag{8}$$

is then one of the available limit state functions specified by a parameter value z. When both the basic variables x and the parameters z are uncertain, the probability p_f of failure is defined by

$$p_f = \int_{\mathcal{X}} \int_{\mathcal{Z}} \chi(h(x,z) \leq 0) f^{X,Z}(x,z) \, \mathrm{d}z \, \mathrm{d}x \tag{9}$$

where $f^{X,Z} : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ is the joint density function of the random variables $X = (X_1, \ldots, X_n)$ and $Z = (Z_1, \ldots, Z_m)$. The uncertainty of the parameters z is the uncertainty in the choice of an appropriate limit state function g_z.

2.3 Independence of the basic variables and the parameters

In the following we always assume that the random variables X and Z are *independent* which has the following meaning:

(a) *If we learn the values of the basic variables x, our knowledge about the parameters z and therefore about the choice of the limit state functions g_z does not change.*

(b) *Learning the values of the parameters z and therefore learning which limit state function g_z to use has no influence on our knowledge about the variables x.*

Then the probability p_f of failure is given by

$$p_f = \int_{\mathcal{X}} \int_{\mathcal{Z}} \chi(h(x,z) \leq 0) f^Z(z) \, \mathrm{d}z \, f^X(x) \, \mathrm{d}x \tag{10}$$

with density functions f^X and f^Z for their corresponding random variables X and Z. The inner integral is a function

$$q : \mathcal{X} \to [0,1] : x \to \int_{\mathcal{Z}} \chi(h(x,z) \leq 0) f^Z(z) \, \mathrm{d}z \tag{11}$$

which is a generalization of q in Eq. (6). For q in Eq. (6) only the function values 1 and 0 are admissible telling us wether an $x \in \mathcal{X}$ is in the failure region R_f or not, but here q describes an uncertain failure region similar to a membership function of a fuzzy set. The value $q(x)$ is the probability that x belongs to the failure region R_f.

2.4 Sets of probability measures and notions of independence

We use now sets $\mathcal{M}_X$ and $\mathcal{M}_Z$ of probability measures to describe the uncertainty of the basic variables x and parameters z of the limit state function h. Since we want to keep the assumption that the basic variables x and the limit state functions g_z are independent we have to compute the upper probability of failure with respect to the different notions of independence for sets of probability measures [2, 6]. We consider here *strong independence*, the weaker and asymmetric *epistemic irrelevance* and later on in Sec. 3.3 *random set independence*.

Strong independence [2, 6, 16]: It is the most restrictive definition of independence simply considering all possible product measures $P_X \otimes P_Z$ for $P_X \in \mathcal{M}_X$ and $P_Z \in \mathcal{M}_Z$. Then the upper probability $\overline{p}_f^{\mathrm{S}}$ of failure in case of strong independence is obtained by

$$\begin{aligned} \overline{p}_f^{\mathrm{S}} &= \sup\left\{(P_X \otimes P_Z)(S_f) : \; P_X \in \mathcal{M}_X, P_Z \in \mathcal{M}_Z\right\} \\ &= \sup_{\substack{P_X \in \mathcal{M}_X \\ P_Z \in \mathcal{M}_Z}} \int_{\mathcal{X}} \int_{\mathcal{Z}} \chi(h(x,z) \leq 0) \, \mathrm{d}P_Z(z) \, \mathrm{d}P_X(x) \\ &= \sup_{\substack{P_X \in \mathcal{M}_X \\ q \in Q}} \int_{\mathcal{X}} q(x) \, \mathrm{d}P_X(x) \end{aligned} \tag{12}$$

where $S_f = \{(x,z) : h(x,z) \le 0\}$ and Q the set

$$Q = \Big\{ q : \mathcal{X} \to [0,1] : \quad q(x) = \int_{\mathcal{Z}} \chi(h(x,z) \le 0)\, \mathrm{d}P_Z(z),\ P_Z \in \mathcal{M}_Z \Big\} \tag{13}$$

of all functions q describing the uncertainty of the failure region R_f as in Eq. (11).

Epistemic irrelevance [2, 4, 16]: We start with the above formula for $\overline{p}_f^{\mathrm{S}}$, but move then $\sup_{P_Z \in \mathcal{M}_Z}$ inside the outer integral:

$$\begin{aligned}
\overline{p}_f^{\mathrm{S}} &= \sup_{\substack{P_X \in \mathcal{M}_X \\ P_Z \in \mathcal{M}_Z}} \int_{\mathcal{X}} \int_{\mathcal{Z}} \chi(h(x,z) \le 0)\, \mathrm{d}P_Z(z)\, \mathrm{d}P_X(x) \\
&\le \sup_{P_X \in \mathcal{M}_X} \int_{\mathcal{X}} \sup_{P_Z \in \mathcal{M}_Z} \int_{\mathcal{Z}} \chi(h(x,z) \le 0)\, \mathrm{d}P_Z(z)\, \mathrm{d}P_X(x) \\
&= \sup_{P_X \in \mathcal{M}_X} \int_{\mathcal{X}} \overline{q}(x)\, \mathrm{d}P_X(x) =: \overline{p}_f^{X \not\to Z}
\end{aligned} \tag{14}$$

with

$$\overline{q}(x) = \sup_{P_Z \in \mathcal{M}_Z} \int_{\mathcal{Z}} \chi(h(x,z) \le 0)\, \mathrm{d}P_Z(z) = \sup_{q \in Q} q(x). \tag{15}$$

The result is a formula for the upper probability $\overline{p}_f^{X \not\to Z}$ in case of epistemic irrelevance, because for each x we can choose its own $P_Z \in \mathcal{M}_Z$ or more exactly a conditional probability measure $P_Z(\cdot\,|x)$ given x. The notation $X \not\to Z$ means that X is epistemically irrelevant to Z, see [4], or in our case that the basic variables are epistemically irrelevant to the parameterized limit state functions g_z. Epistemic irrelevance is an asymmetric notion of independence meaning only what we have stated in (a) in Sec. 2.3, but not necessarily the other way round as in (b). The set $\mathcal{M}_{X \not\to Z}$ of all probability measures according to epistemic irrelevance of X to Z is defined by

$$\mathcal{M}_{X \not\to Z} = \Big\{ P : P(E) = \int_{\mathcal{X}} \int_{\mathcal{Z}} \chi((x,z) \in E)\, \mathrm{d}P_Z(z|x)\, \mathrm{d}P_X(x), \quad P_X \in \mathcal{M}_X,\ P_Z(\cdot\,|x) \in \mathcal{M}_Z \Big\} \tag{16}$$

where E is an event. In Eq. (14) we write $P_Z(z)$ instead of $P_Z(z|x)$ since it is clear that we use different probability measures P_Z and not only one because of the $\sup_{P_Z \in \mathcal{M}_Z}$ in the formula.

When it is possible to assume epistemic irrelevance we have the advantage that we can treat the uncertainty of the basic variables and of the limit state functions completely separately. We can compute $\overline{q}$ in advance and then using $\overline{q}$ for different models of uncertainty of x.

The function $\overline{q}$ is the upper envelope of the set Q defined in Eq. (13). If this upper envelope $\overline{q}$ is an element of Q we have $\overline{p}_f^{\mathrm{S}} = \overline{p}_f^{X \not\to Z}$.

3 The probability of failure $p_f(a,b)$ with uncertain parameters a and b

3.1 The function $p_f(a,b)$

Let us now extend Equation (10) by adding parameters $a = (a_1, \ldots, a_{n_a}) \in \mathbb{R}^{n_a}$ for the probability density function f^X describing the uncertainty of the basic variables x and parameters $b = (b_1, \ldots, b_{n_b}) \in \mathbb{R}^{n_b}$ for the density f^Z of the additional parameters z which leads to a function

$$p_f(a,b) = \int_{\mathcal{X}} \int_{\mathcal{Z}} \chi(h(x,z) \le 0) f_b^Z(z)\, \mathrm{d}z\, f_a^X(x)\, \mathrm{d}x. \tag{17}$$

This function $p_f(a,b)$ provides an interface for controlling the shape of the probability density functions used for modelling the uncertainty of the basic variables x and the parameters z. We also write $p_f(a,b; f_a^X, f_b^Z)$ if it is necessary to emphasize which density functions are used.

In the following the parameters a and b are assumed to be uncertain; intervals, sets or random sets are used to describe their uncertainty. This and the approach with parameterized density functions f_a^X and f_b^Z give us a convenient way to generate the sets $\mathcal{M}_X$ and $\mathcal{M}_Z$ of probability measures and the possibility to model the uncertainty of x and z by means of more specific probability measures than directly using sets or random sets for x and z. An example for such a parameterized density f_a^X or f_b^Z is the density of a Gaussian distribution depending on expectation μ and variance σ^2. Then describing the uncertainty of μ and σ by sets or random sets leads to sets $\mathcal{M}_X$ or $\mathcal{M}_Z$ of probability measures.

3.2 Uncertainty of the parameters a and b modelled by sets A and B

We describe the uncertainty of the parameter $a \in \mathbb{R}^{n_a}$ by a set $A \subseteq \mathbb{R}^{n_a}$ and the uncertainty of $b \in \mathbb{R}^{n_b}$ by a set $B \subseteq \mathbb{R}^{n_b}$ and show how the upper probability of failure is determined in case of strong independence or epistemic irrelevance. But first we have to generate the sets of probability measures $\mathcal{M}_X$ and $\mathcal{M}_Z$.

Generating $\mathcal{M}_X$ and $\mathcal{M}_Z$:

$$\mathcal{M}_X = \Big\{ P : P(E) = \int_{\mathcal{A}} \int_{\mathcal{X}} \chi(x \in E) f_a^X(x)\, \mathrm{d}x\, \mathrm{d}P_A(a), \quad P_A \in \mathcal{M}(A) \Big\} \tag{18}$$

where $\mathcal{M}(A) = \{P : P(A) = 1\}$ is the set of all probability measures living on the set A and where E is an event. The set $\mathcal{M}_Z$ is generated in an analogous way using f_b^Z and $\mathcal{M}(B)$.

Strong independence: Eq. (12) together with Eq. (18) leads to the following formula for the upper probability

$\overline{p}_f^{\mathrm{S}}$ in case of strong independence:

$$\begin{aligned}\overline{p}_f^{\mathrm{S}} &= \sup_{\substack{P_X\in\mathcal{M}_X\\P_Z\in\mathcal{M}_Z}} \int_{\mathcal{X}}\int_{\mathcal{Z}} \chi(h(x,z)\le 0)\,\mathrm{d}P_Z(z)\,\mathrm{d}P_X(x)\\ &= \sup_{\substack{P_A\in\mathcal{M}(A)\\P_B\in\mathcal{M}(B)}} \int_{\mathcal{A}}\int_{\mathcal{X}}\int_{\mathcal{B}}\int_{\mathcal{Z}} \chi(h(x,z)\le 0) f_b^Z(z)\,\mathrm{d}z\cdot\\ &\qquad\qquad \cdot \mathrm{d}P_B(b) f_a^X(x)\,\mathrm{d}x\,\mathrm{d}P_A(a)\\ &= \sup_{\substack{P_A\in\mathcal{M}(A)\\P_B\in\mathcal{M}(B)}} \int_{\mathcal{A}}\int_{\mathcal{B}} p_f(a,b)\,\mathrm{d}P_B(b)\,\mathrm{d}P_A(a) \qquad (19)\\ &= \sup_{\substack{a\in A\\b\in B}} \int_{\mathcal{A}}\int_{\mathcal{B}} p_f(\xi,\eta)\,\mathrm{d}\delta_b(\eta)\,\mathrm{d}\delta_a(\xi) = \sup_{\substack{a\in A\\b\in B}} p_f(a,b).\end{aligned}$$

The Dirac measures δ_a and δ_b are extreme points in the sets $\mathcal{M}(A)$ and $\mathcal{M}(B)$ of probability measures.

Epistemic irrelevance: Eq. (14) together with Eq. (18) leads to the formula for the upper probability $\overline{p}_f^{X\not\to Z}$ in case of epistemic irrelevance:

$$\begin{aligned}\overline{p}_f^{X\not\to Z} &= \sup_{P_A\in\mathcal{M}(A)} \int_{\mathcal{A}}\int_{\mathcal{X}} \sup_{P_B\in\mathcal{M}(B)} \int_{\mathcal{B}}\int_{\mathcal{Z}} \chi(h(x,z)\le 0) f_b^Z(z)\cdot\\ &\qquad\qquad \cdot \mathrm{d}z\,\mathrm{d}P_B(b) f_a^X(x)\,\mathrm{d}x\,\mathrm{d}P_A(a)\\ &= \sup_{a\in A}\int_{\mathcal{X}} \sup_{b\in B}\int_{\mathcal{Z}} \chi(h(x,z)\le 0) f_b^Z(z)\,\mathrm{d}z\, f_a^X(x)\,\mathrm{d}x\\ &= \sup_{a\in A}\int_{\mathcal{X}} \overline{q}(x) f_a^X(x)\,\mathrm{d}x \qquad (20)\end{aligned}$$

with $\overline{q}(x) = \sup_{b\in B}\int_{\mathcal{Z}} \chi(h(x,z)\le 0) f_b^Z(z)\,\mathrm{d}z$, again using that δ_a and δ_b are extreme points in $\mathcal{M}(A)$ and $\mathcal{M}(B)$.

3.3 Uncertainty of a and b modelled by random sets $\mathscr{A}$ and $\mathscr{B}$

A *random set* as introduced by [3] is a family $\mathscr{A}$ of *focal sets* A_i together with *weights* $m_{\mathscr{A}}(A_i)$ which sum up to one. Then the upper probability $\overline{P}(E)$ or *plausibility* $\mathrm{Pl}_{\mathscr{A}}(E)$ of an event E is given in the case of a finite random set with focal sets $A_1,\dots,A_{|\mathscr{A}|}$ by the formula

$$\overline{P}(E) = \mathrm{Pl}_{\mathscr{A}}(E) = \sum_{i:E\cap A_i\neq\varnothing} m_{\mathscr{A}}(A_i) = \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i) \sup_{P\in\mathcal{M}(A_i)} P(E) \qquad (21)$$

where $|\mathscr{A}|$ denotes the number of focal sets where

$$\mathcal{M}(A_i) = \{P:\ P(A_i) = 1\} \qquad (22)$$

is the set of all probability measures on the focal set A_i, cf. [6]. The lower probability $\underline{P}(E)$ or belief $\mathrm{Bel}_{\mathscr{A}}(E)$ is defined by

$$\underline{P}(E) = \mathrm{Bel}_{\mathscr{A}}(E) = \sum_{i:A_i\subseteq E} m_{\mathscr{A}}(A_i) = \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i) \inf_{P\in\mathcal{M}(A_i)} P(E). \qquad (23)$$

First we have to generate the sets $\mathcal{M}_X$ and $\mathcal{M}_Z$ by means of random sets $\mathscr{A}$ and $\mathscr{B}$ modelling the uncertainty of a and b. Then we show how to determine the upper probabilities of failure for strong independence, epistemic irrelevance and random set independence.

Generating the sets $\mathcal{M}_X$ and $\mathcal{M}_Z$:

$$\begin{aligned}\mathcal{M}_X &= \Big\{P:\ P(E) = \int_{\mathcal{A}}\int_{\mathcal{X}} \chi(x\in E) f_a^X(x)\,\mathrm{d}x\,\mathrm{d}P_A(a),\\ &\qquad\qquad P_A\in\mathcal{M}(\mathscr{A})\Big\}\\ &= \Big\{P:\ P(E) = \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i)\cdot \qquad (24)\\ &\qquad \cdot \int_{\mathcal{A}}\int_{\mathcal{X}} \chi(x\in E) f_a^X(x)\,\mathrm{d}x\,\mathrm{d}P_{A_i}(a),\\ &\qquad\qquad P_{A_i}\in\mathcal{M}(A_i), i=1,\dots,n\Big\}\end{aligned}$$

where $\mathcal{M}(\mathscr{A})$ is the set of all probability measures generated by a random set $\mathscr{A}$, cf. [9]. A probability measure in $\mathcal{M}(\mathscr{A})$ is a weighted sum of probability measures $P_{A_i}\in\mathcal{M}(A_i)$ living on the focal sets A_i. The set $\mathcal{M}_Z$ is obtained in a similar way using f_b^Z and the random set $\mathscr{B}$.

Strong independence: Eq. (12) together with Eq. (24) leads to the upper probability

$$\begin{aligned}\overline{p}_f^{\mathrm{S}} &= \sup_{\substack{P_{Ar}\in\mathcal{M}(A_r),r=1,\dots,|\mathscr{A}|\\P_{Bs}\in\mathcal{M}(B_s),s=1,\dots,|\mathscr{B}|}} \sum_{i=1}^{|\mathscr{A}|}\sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{A}}(A_i) m_{\mathscr{B}}(B_j)\cdot \qquad (25)\\ &\qquad\qquad \cdot \int_{\mathcal{A}}\int_{\mathcal{B}} p_f(a,b)\,\mathrm{d}P_{A_i}(a)\,\mathrm{d}P_{B_j}(b)\\ &= \sup_{\substack{a_r\in A_r,r=1,\dots,|\mathscr{A}|\\b_s\in B_s,s=1,\dots,|\mathscr{B}|}} \sum_{i=1}^{|\mathscr{A}|}\sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{A}}(A_i) m_{\mathscr{B}}(B_j) p_f(a_i,b_j)\end{aligned}$$

in case of strong independence replacing the probability measures P_{A_i} and P_{B_j} by Dirac measures δ_{a_i} and δ_{b_j} on their corresponding focal sets A_i and B_j similar to the section before. A general proof that the upper probability can be obtained by means of Dirac measures can be found in [7].

Epistemic irrelevance: Eq. (14) together with Eq. (24) results in the upper probability $\overline{p}_f^{X\not\to Z}$ in case of epistemic irrelevance:

$$\begin{aligned}\overline{p}_f^{X\not\to Z} &= \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i) \sup_{P_A\in\mathcal{M}(A_i)} \int_{\mathcal{A}}\int_{\mathcal{X}} \sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{B}}(B_j)\cdot \qquad (26)\\ &\qquad \cdot \sup_{P_B\in\mathcal{M}(B_j)} \int_{\mathcal{B}}\int_{\mathcal{Z}} \chi(h(x,z)\le 0) f_b^Z(z)\cdot\\ &\qquad\qquad \cdot \mathrm{d}z\,\mathrm{d}P_B(b)\, f_a^X(x)\,\mathrm{d}x\,\mathrm{d}P_A(a)\\ &= \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i) \sup_{a\in A_i} \int_{\mathcal{X}} \sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{B}}(B_j)\cdot\\ &\qquad \cdot \sup_{b\in B_j} \int_{\mathcal{Z}} \chi(h(x,z)\le 0) f_b^Z(z)\,\mathrm{d}z\, f_a^X(x)\,\mathrm{d}x.\end{aligned}$$

The function $\overline{q}$ is given here by

$$\overline{q}(x) = \sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{B}}(B_j) \sup_{b\in B_j} \int_{\mathcal{Z}} \chi(h(x,z)\le 0) f_b^Z(z)\,\mathrm{d}z. \tag{27}$$

Random set independence: Let the uncertainty of a variable a be modelled by a random set $\mathscr{A}$ with focal sets A_i and weights $m_{\mathscr{A}}(A_i)$ and the uncertainty of a variable b by a random set $\mathscr{B}$ with focal sets B_j and weights $m_{\mathscr{B}}(B_j)$. The *joint random set*, in the classical version assuming *random set independence*, is defined as the family $\mathscr{C}$ of all Cartesian products $C_{ij} = A_i \times B_j$ of focal sets A_i and B_j. The weights of these *joint focal sets* C_{ij} are given by the product $m_{\mathscr{C}}(C_{ij}) = m_{\mathscr{A}}(A_i)\, m_{\mathscr{B}}(B_j)$ and the formula for the *joint plausibility measure* Pl by

$$\begin{aligned}\mathrm{Pl}(E) &= \sum_{i=1}^{|\mathscr{A}|}\sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{A}}(A_i) m_{\mathscr{B}}(B_j) \chi(E\cap(A_i\times B_j)\neq\varnothing) \\ &= \sum_{i=1}^{|\mathscr{A}|}\sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{A}}(A_i) m_{\mathscr{B}}(B_j) \sup_{P\in\mathcal{M}(A_i\times B_j)} P(E)\end{aligned} \tag{28}$$

with set $\mathcal{M}(A_i\times B_j) = \{P : P(A_i\times B_j) = 1\}$, cf. [2, 3, 6]. This set is the largest possible set of joint probability measures generated by the marginal sets $\mathcal{M}(A_i)$ and $\mathcal{M}(B_j)$. The key properties of random set independence are that

(i) there are no interactions between focal sets A_i and B_j,
(ii) the focal sets A_i and B_j are chosen in a stochastically independent way;
(iii) the joint plausibility $\mathrm{Pl}(E)$ is obtained by solving optimization problems $\sup_{P\in\mathcal{M}(A_i\times B_j)} P(E)$ on each joint focal set $A_i\times B_j$ *separately* and *independently* of the other joint focal sets.

Our problem here is that density functions are involved in the formulas and that we have to combine not only two random sets $\mathscr{A}$ and $\mathscr{B}$ but also two density functions f_a^X and f_b^Z. So we have to generalize the formula for the joint plausibility measure. One possibility is to replace the set $\mathcal{M}(A_i\times B_j)$ by a set of joint probability measures generated by sets $\mathcal{M}_X^i$ and $\mathcal{M}_Z^j$ defined by

$$\begin{aligned}\mathcal{M}_X^i &= \Big\{P\colon P(E) = \int_{\mathcal{A}}\int_{E} f_a^X(x)\,\mathrm{d}x\,\mathrm{d}P_A(a),\, P_A\in\mathcal{M}(A_i)\Big\}, \\ \mathcal{M}_Z^j &= \Big\{P\colon P(E) = \int_{\mathcal{B}}\int_{E} f_b^Z(z)\,\mathrm{d}z\,\mathrm{d}P_B(b),\, P_B\in\mathcal{M}(B_j)\Big\}\end{aligned} \tag{29}$$

as the sets $\mathcal{M}_X$ and $\mathcal{M}_Z$ in Sec. 3.2. Now the question arises how to combine $\mathcal{M}_X^i$ and $\mathcal{M}_Z^j$. Since $\mathcal{M}(A_i\times B_j)$ is the largest possible set of joint probability measures generated by $\mathcal{M}(A_i)$ and $\mathcal{M}(B_j)$ an analogous approach would be to use here the set of all possible joint probability measures generated by $\mathcal{M}_X^i$ and $\mathcal{M}_Z^j$. But this means to consider also all possible joint density functions with marginals f_a^X and f_b^Z which is not very attractive because of the computational effort and because independence is not taken into account on the level of the density functions.

Another approach is to combine $\mathcal{M}_X^i$ and $\mathcal{M}_Z^j$ according to strong independence or epistemic irrelevance as in Sec. 3.2 which means to replace $\sup_{P\in\mathcal{M}(A_i\times B_j)} P(E)$ in Eq. (28) by the results of Eq. (19) or of Eq. (20):

For strong independence, locally for each pair of sets $\mathcal{M}_X^i$ and $\mathcal{M}_Z^j$, we get the upper probability

$$\overline{p}_f^{\mathrm{RS}} = \sum_{i=1}^{|\mathscr{A}|}\sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{A}}(A_i) m_{\mathscr{B}}(B_j) \sup_{a\in A_i, b\in B_j} p_f(a,b) \tag{30}$$

by means of Eq. (19), cf. also [9]. We denote this upper probability by a superscript "RS" where "R" means random set independence and "S" indicates that the sets $\mathcal{M}_X^i$ and $\mathcal{M}_Z^j$ corresponding to $A_i\times B_j$ are combined according to strong independence. The difference to the "global" version of strong independence in Eq. (25) is that here the "sup" is inside instead of outside the sums. So it is clear that we have the ordering $\overline{p}_f^{\mathrm{S}} \le \overline{p}_f^{\mathrm{RS}}$.

Epistemic irrelevance, locally for each pair of sets $\mathcal{M}_X^i$ and $\mathcal{M}_Z^j$, leads to

$$\overline{p}_f^{\mathrm{R},X\not\to Z} = \sum_{i=1}^{|\mathscr{A}|}\sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{A}}(A_i) m_{\mathscr{B}}(B_j) \sup_{a\in A_i} \int_{\mathcal{X}} \overline{q}_j(x) f_a^X(x)\,\mathrm{d}x \tag{31}$$

$$\overline{q}_j(x) = \sup_{b\in B_j} \int_{\mathcal{Z}} \chi(h(x,z)\le 0) f_b^Z(z)\,\mathrm{d}z, \tag{32}$$

cf. Eq. (20). We use here the superscript "$\mathrm{R}, X\not\to Z$" instead of "RS" to denote the upper probability.

Summarizing the orderings of all upper probabilities we have $\overline{p}_f^{\mathrm{S}} \le \overline{p}_f^{X\not\to Z} \le \overline{p}_f^{\mathrm{R},X\not\to Z}$ and $\overline{p}_f^{\mathrm{S}} \le \overline{p}_f^{\mathrm{RS}} \le \overline{p}_f^{\mathrm{R},X\not\to Z}$.

We note that Dirac measures δ_a and δ_b instead of arbitrary density functions f_a^X and f_b^Z lead to the classical joint plausibility measure: For Dirac measures we always have $\mathcal{M}_X^i = \mathcal{M}(A_i)$, $\mathcal{M}_Z^j = \mathcal{M}(B_j)$. Further the resulting upper probabilities for $\mathcal{M}(A_i\times B_j)$ as in Eq. (28) or for sets of probability measures generated by $\mathcal{M}(A_i)$, $\mathcal{M}(B_j)$ according to strong independence or epistemic irrelevance coincide since Dirac measures $\delta_a\otimes\delta_b$, $a\in A_i$, $b\in B_j$ are extreme points in all these three sets. This means that we have $\overline{p}_f^{\mathrm{R}} := \overline{p}_f^{\mathrm{RS}} = \overline{p}_f^{\mathrm{R},X\not\to Z}$.

4 Alternative approaches and views

Let $Y_{|x} = h(x,Z)$ be the conditional random variable for the uncertain output of the parameterized limit state function h given a value of the basic variables x, Z the random variable corresponding to the parameters z, $f^{Y|x}: \mathcal{Y}\subseteq\mathbb{R}\to\mathbb{R}$ the probability density of $Y_{|x}$ and $F^{Y|x}: \mathcal{Y}\to[0,1]$ the probability distribution function. Then $q:\mathcal{X}\to[0,1]$ describing the uncertainty of the failure region is defined here by

$$q(x) = F^{Y|x}(0) = \int_{-\infty}^{0} f^{Y|x}(y)\,\mathrm{d}y \tag{33}$$

since $y\le 0$ means failure.

On the one hand the functions $f^{Y|x}$, $F^{Y|x}$ and q (or $\overline{F}^{Y|x}$ and $\overline{q}$ if sets of probability measures are used, see below) can be used to visualize the uncertainties in the limit state function. On the other hand the uncertainties in the limit state function can be specified providing these functions. Especially describing the uncertainty in the failure region by means of the function $\overline{q}$ in case of epistemic irrelevance opens the possibility to start also with fuzzy failure regions. Note that there may be a conceptual but not a formal difference between $\overline{q}$ and a membership function of a fuzzy set. To specify the limit state function g in its uncertain format instead of introducing additional parameters was also suggested in [12].

We show now how the two approaches are connected for the case that h is given by $y = h(x,z) = g(x) + z$ which means to add something uncertain to a deterministic limit state function g. Substituting $z = y - g(x)$ in Eq. (17) leads to

$$\begin{aligned} p_f(a,b) &= \int_{\mathcal{X}} \int_{\mathcal{Z}} \chi(g(x)+z \le 0) f_b^Z(z)\,\mathrm{d}z\, f_a^X(x)\,\mathrm{d}x \\ &= \int_{\mathcal{X}} \int_{\mathcal{Y}} \chi(y \le 0) f_b^Z(y-g(x))\,\mathrm{d}y\, f_a^X(x)\,\mathrm{d}x \\ &= \int_{\mathcal{X}} \int_{-\infty}^{0} f^{Y|x}_{(g(x),b)}(y)\,\mathrm{d}y\, f_a^X(x)\,\mathrm{d}x \\ &= \int_{\mathcal{X}} F^{Y|x}_{(g(x),b)}(0) f_a^X(x)\,\mathrm{d}x \end{aligned} \tag{34}$$

with $f^{Y|x}_{(g(x),b)}(y) = f_b^Z(y-g(x))$. The density $f^{Y|x}_{(g(x),b)}$ describes the uncertainty of the output of the limit state function and it is the same density function as f_b^Z, but moved from 0 to $g(x)$. This is indicated by the additional parameter $g(x)$ of the probability density $f^{Y|x}_{(g(x),b)}$ depending now on parameters which are not constant on $\mathcal{X}$. Modelling the uncertainty of parameter b by a set B we get an example for a function

$$\begin{aligned} \overline{q}(x) &= \sup_{b\in B} \int_{\mathcal{Z}} \chi(h(x,z) \le 0) f_b^Z(z)\,\mathrm{d}z \\ &= \sup_{b\in B} F^{Y|x}_{(g(x),b)}(0) =: \overline{F}^{Y|x}(0) \end{aligned} \tag{35}$$

using both approaches. The function $\overline{F}^{Y|x}$ is the upper distribution function of $Y_{|x}$. In an analogous way we obtain the lower bound

$$\begin{aligned} \underline{q}(x) &= \inf_{b\in B} \int_{\mathcal{Z}} \chi(h(x,z) \le 0) f_b^Z(z)\,\mathrm{d}z \\ &= \inf_{b\in B} F^{Y|x}_{(g(x),b)}(0) =: \underline{F}^{Y|x}(0). \end{aligned} \tag{36}$$

It is the lower probability of failure given $x \in \mathcal{X}$.

5 Numerical example

5.1 Problem statement

As a simple numerical example we consider a beam of length 3 m supported on both ends and additionally bedded on a spring, cf. Fig. 1. The values of the beam rigidity $EI = 1\,\mathrm{kNm}^2$ and of the load $f(\xi) = 100\,\mathrm{kN/m}$ are deterministic, but the value of the spring constant x (in our notation for the basic variables) is assumed to be uncertain.

The deterministic limit state function g is given as[1]

$$\begin{aligned} g(x) &= M_{\mathsf{yield}} - \max_{\xi\in[0,3]} |M(\xi,x)| \\ &= M_{\mathsf{yield}} - \frac{qL^2}{4} \max\left(\frac{(1-c(x))^2}{2}, c(x) - \frac{1}{2} \right) \end{aligned} \tag{37}$$

with $c(x) = 5x/(384EI/L^3 + 8x)$, see Fig. 1. $M(\xi,x)$ is the bending moment at $\xi \in [0,3]$ on the beam depending on the spring constant x and $M_{\mathsf{yield}} = 21\,\mathrm{kNm}$ is the elastic limit moment of the beam for both positive and negative moments.

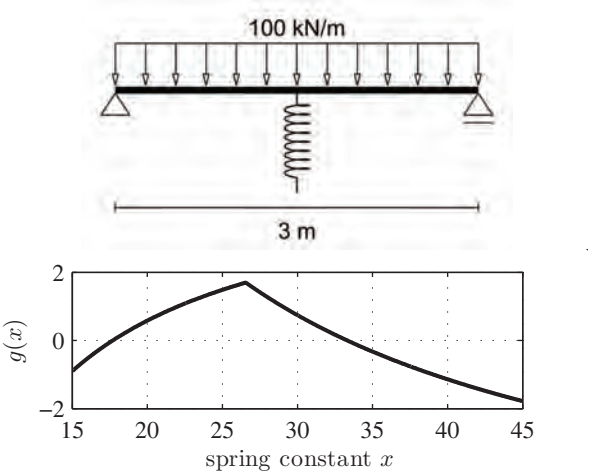


Figure 1: Beam bedded on a spring and deterministic limit state function g.

5.2 Modelling the uncertainty of spring constant x

The uncertainty of the spring constant x ([kN/m]) is modelled either by an interval A, by a random set $\mathscr{A}$ or by a Gaussian distribution. In the following we present what we will use for the basic variable x in the examples in the next section.

Interval A modelling the uncertainty of x:
We will use the interval $A = [\underline{a}, \overline{a}] = [20, 30]\,\mathrm{kN/m}$.

Random set $\mathscr{A}$ modelling the uncertainty of x:
The random set $\mathscr{A}$ is given by the focal sets $A_1 = [17, 30]$, $A_2 = [23, 31]$, $A_3 = [27, 32.5]$ and weights $m_{\mathscr{A}}(A_1) = 0.2$, $m_{\mathscr{A}}(A_2) = 0.3$ and $m_{\mathscr{A}}(A_3) = 0.5$.

[1] Thanks to one of the reviewers for providing an explicit formula.

Probability distribution modelling the uncertainty of the basic variable x:
We assume that x is Gaussian distributed with parameters $\mu = 34$ and $\sigma^2 = 1$.

6 Cases and examples

In this section we present examples and special cases with respect to the different notions of independence.

6.1 Sets of parameterized limit state functions

Let B be a set and

$$G = \{g_z : g_z(x) = h(x,z),\, z \in B\} \tag{38}$$

the family of limit state functions parameterized by $z \in B$. Further let the function $\underline{g}$ be the lower envelope of G defined by $\underline{g}(x) = \inf_{g_z \in G} g_z(x)$ and $\overline{g}$ the upper envelope.

In this case we have to set $f_b^Z := \delta_z$ and $b := z$ in Eqs. (19) and (20) which leads to

$$\begin{aligned}\overline{p}_f^{\mathrm{S}} &= \sup_{\substack{a\in A\\ z\in B}} p_f(x,z; f_a^X, \delta_z) \qquad (39)\\ &= \sup_{\substack{a\in A\\ z\in B}} \int_{\mathcal{X}}\int_{\mathcal{Z}} \chi(h(x,\eta)\le 0)\,\delta_z(\eta)\,\mathrm{d}\eta\, f_a^X(x)\,\mathrm{d}x\\ &= \sup_{\substack{a\in A\\ z\in B}} \int_{\mathcal{X}} \chi(g_z(x)\le 0) f_a^X(x)\,\mathrm{d}x\end{aligned}$$

for strong independence and to $\overline{q}$ and the upper probability for epistemic irrelevance:

$$\begin{aligned}\overline{q}(x) &= \sup_{z\in B}\int_{\mathcal{Z}} \chi(h(x,\eta)\le 0)\,\delta_z(\eta)\,\mathrm{d}\eta \qquad (40)\\ &= \sup_{z\in B}\chi(g_z(x)\le 0) = \chi(\underline{g}(x)\le 0),\\ \overline{p}_f^{X\not\to Z} &= \sup_{a\in A}\int_{\mathcal{X}} \overline{q}(x) f_a^X(x)\,\mathrm{d}x = \sup_{a\in A}\int_{\mathcal{X}} \chi(\underline{g}(x)\le 0) f_a^X(x)\,\mathrm{d}x. \qquad (41)\end{aligned}$$

As an example we use $h(x,z) = g(x+z)$ with $z \in B = [0,2]$ moving g to the left and the limit state function g defined in the previous section. In Fig. 2 the set G, the functions $\overline{q}$, $\underline{q}$ and the upper and lower probability distribution functions $\overline{F}^{Y|x}$ and $\underline{F}^{Y|x}$ at $x = 20$ are depicted, see also Sec. 4.

Uncertainty of x modelled by an interval A:
Here we have to set $f_a^X := \delta_x$ and $a := x$. In this case the results for strong independence and epistemic irrelevance coincide:

$$\begin{aligned}\overline{p}_f^{\mathrm{S}} &= \sup_{\substack{x\in A\\ z\in B}} p_f(x,z;\delta_x,\delta_z) = \sup_{\substack{x\in A\\ z\in B}}\int_{\mathcal{X}} \chi(h(\xi,z)\le 0)\delta_x(\xi)\,\mathrm{d}\xi\\ &= \sup_{\substack{x\in A\\ z\in B}} \chi(g_z(x)\le 0) = \sup_{x\in A}\chi(\underline{g}(x)\le 0), \qquad (42)\end{aligned}$$

$$\begin{aligned}\overline{p}_f^{X\not\to Z} &= \sup_{x\in A}\int_{\mathcal{X}} \overline{q}(\xi)\delta_x(\xi)\,\mathrm{d}\xi = \sup_{x\in A}\overline{q}(x) \qquad (43)\\ &= \sup_{x\in A}\chi(\underline{g}(x)\le 0)\end{aligned}$$

because only one single $x \in A$ is used at the same time in the formulas.

We obtain the upper probabilities for our example by means of

$$\overline{p}_f^{\mathrm{S}} = \overline{p}_f^{X\not\to Z} = \sup_{x\in A}\chi(\underline{g}(x)\le 0) = \chi(\underline{g}(A)\cap(-\infty,0]\neq\varnothing) \tag{44}$$

where $g(A) = \big[\min_{x\in A} g(x), \max_{x\in A} g(x)\big] = [\underline{g(A)}, \overline{g(A)}]$ is the image of A under a function g. For the computation of $\chi(g(A)\cap(-\infty,0]\neq\varnothing)$ it is sufficient to know the lower bound $\underline{g(A)}$ of the image $g(A)$:

$$\chi(g(A)\cap(-\infty,0]\neq\varnothing) = \chi(\underline{g(A)}\le 0). \tag{45}$$

Since in our example all g_z and $\underline{g}$ are a concave functions we have $\underline{g}(A) = \min(\underline{g}(\underline{a}), \underline{g}(\overline{a})) = 0.2763$ for the interval $A = [\underline{a}, \overline{a}] = [20, 30]$ and therefore the upper probability of failure $\overline{p}_f^{\mathrm{S}} = \overline{p}_f^{X\not\to Z} = \chi(0.2763 \le 0) = 0$.

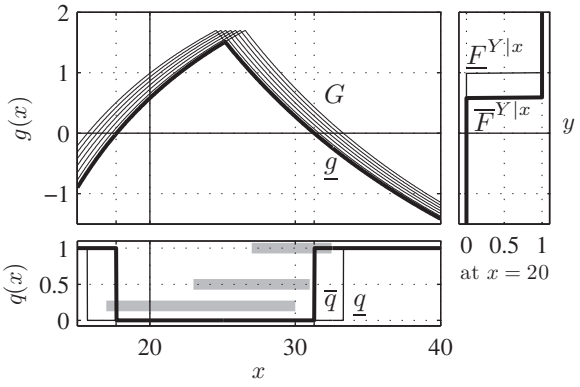


Figure 2: Set G of limit state functions g_z, lower envelope $\underline{g}$; $\underline{q}(x) = \underline{F}^{Y|x}(0), \overline{q}(x) = \overline{F}^{Y|x}(0)$, focal sets of random set $\mathscr{A}$ (gray bars); $\underline{F}^{Y|x}$ and $\overline{F}^{Y|x}$ at $x = 20$.

Uncertainty of x modelled by a random set $\mathscr{A}$:
First we do some preliminary work replacing in Eqs. (25), (26), (30) and (31) the density functions by Dirac measures:

$$\overline{p}_f^{\mathrm{S}} = \sup_{\substack{x_r\in A_r,\, r=1,\dots,|\mathscr{A}|\\ z_s\in B_s,\, s=1,\dots,|\mathscr{B}|}} \sum_{i=1}^{|\mathscr{A}|}\sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{A}}(A_i) m_{\mathscr{B}}(B_j) p_f(x_i, z_j) \tag{46}$$

$$\overline{p}_f^{X\not\to Z} = \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i) \sup_{x\in A_i} \int_{\mathcal{X}} \sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{B}}(B_j) \cdot \tag{47}$$

$$\cdot \sup_{z\in B_j} \int_{\mathcal{Z}} \chi(h(\xi,\eta)\le 0)\delta_z(\eta)\,\mathrm{d}\eta\,\delta_x(\xi)\,\mathrm{d}\xi$$

$$= \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i) \sup_{x\in A_i} \sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{B}}(B_j) \sup_{z\in B_j} \chi(h(x,z)\le 0)$$

$$= \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i) \sup_{x\in A_i} \overline{q}(x)$$

$$\overline{q}(x) = \sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{B}}(B_j) \sup_{z\in B_j} \chi(h(x,z)\le 0), \tag{48}$$

$$\overline{p}_f^{\mathrm{RS}} = \overline{p}_f^{\mathrm{R},X\not\to Z} = \sum_{i=1}^{|\mathscr{A}|}\sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{A}}(A_i) m_{\mathscr{B}}(B_j) \sup_{\substack{x\in A_i\\ z\in B_j}} p_f(x,z) \tag{49}$$

where $p_f(x,z) = \chi(h(x,z)\le 0)$ and where $\overline{p}_f^{\mathrm{RS}} = \overline{p}_f^{\mathrm{R},X\not\to Z}$ coincides for Dirac measures as already mentioned. While these equations are needed in the next section we further use here that we have only one set B with weight 1 which leads to the following simplified versions:

$$\overline{p}_f^{\mathrm{S}} = \sup_{\substack{x_r\in A_r,\, r=1,\ldots,|\mathscr{A}|\\ z\in B}} \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i) p_f(x_i,z), \tag{50}$$

$$\overline{p}_f^{X\not\to Z} = \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i) \sup_{x\in A_i} \sup_{z\in B} \chi(h(x,z)\le 0), \tag{51}$$

$$\overline{p}_f^{\mathrm{RS}} = \overline{p}_f^{\mathrm{R},X\not\to Z} = \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i) \sup_{\substack{x\in A_i\\ z\in B}} p_f(x,z) \tag{52}$$

where $\overline{p}_f^{X\not\to Z} = \overline{p}_f^{\mathrm{R},X\not\to Z}$ because $p_f(x,z) = \chi(h(x,z)\le 0)$.

The difference between $\overline{p}_f^{\mathrm{S}}$ and $\overline{p}_f^{X\not\to Z}$ is that there is a single z used for all x_r together in case of strong independence while for epistemic irrelevance z can be chosen for each x_r separately. The numerical results are obtained by

$$\overline{p}_f^{\mathrm{S}} = \sup_{z\in B} \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i) \sup_{x\in A_i} \chi(g_z(x)\le 0) \tag{53}$$

$$= \sup_{z\in B} \mathrm{Pl}_{\mathscr{A}}(g_z(x)\le 0) = 0.5$$

and

$$\overline{p}_f^{X\not\to Z} = \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i) \sup_{x\in A_i} \chi(\underline{g}(x)\le 0) \tag{54}$$

$$= \mathrm{Pl}_{\mathscr{A}}(\underline{g}(x)\le 0) = 0.2\cdot 1 + 0.3\cdot 0 + 0.5\cdot 1 = 0.7,$$

cf. Eqs. (50), (51).

We have always $\overline{p}_f^{\mathrm{S}} = \overline{p}_f^{X\not\to Z}$ if $\underline{g}\in G$. This holds in the case where $h(x,z) = g(x)+z$, $z\in B = [\underline{b},\overline{b}]$ $\underline{g}(x) = g(x)+\underline{b}$ and $\overline{g}(x) = g(x)+\overline{b}$.

6.2 Random sets of parameterized limit state functions

Modelling the uncertainty of the limit state function:
For modelling the uncertainty of the parameter z we use a random set $\mathscr{B}$ given by the following focal sets B_j and weights $m_{\mathscr{B}}(B_j)$:

$$\begin{aligned} B_1 &= [-0.9, 1.3], \quad & m_{\mathscr{B}}(B_1) &= 0.1,\\ B_2 &= [-0.6, 0.9], \quad & m_{\mathscr{B}}(B_2) &= 0.3,\\ B_3 &= [-0.4, 0.6], \quad & m_{\mathscr{B}}(B_3) &= 0.4,\\ B_4 &= [-0.2, 0.4], \quad & m_{\mathscr{B}}(B_4) &= 0.2. \end{aligned}$$

In the view of Sec. 4 we define a random set $\mathscr{G}$ of limit state functions by the focal sets $G_j = \{g_z : z\in B_j\}$ and weights $m_{\mathscr{G}}(G_j) = m_{\mathscr{B}}(B_j)$. At a point $x\in\mathcal{X}$ we have then a random set $\mathscr{G}(x)$ with focal sets $G_j(x) = \{g(x) : g\in G_j\}$ and the same weights, which describes the output of the uncertain limit state function.

The function $\overline{q}$ is obtained by

$$\overline{q}(x) = \overline{F}^{Y|x}(0) = \sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{B}}(B_j)\chi(G_j(x)\cap(-\infty,0]\neq\varnothing)$$

$$= \mathrm{Pl}_{\mathscr{G}(x)}((-\infty,0]) \tag{55}$$

which is the plausibility measure of $(-\infty,0]$ for the random set $\mathscr{G}(x)$ at x. The lower bound $\underline{q}$ is the belief measure at x (cf. Fig. 3):

$$\underline{q}(x) = \underline{F}^{Y|x}(0) = \sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{B}}(B_j)\chi(G_j(x)\subseteq(-\infty,0])$$

$$= \mathrm{Bel}_{\mathscr{G}(x)}((-\infty,0]). \tag{56}$$

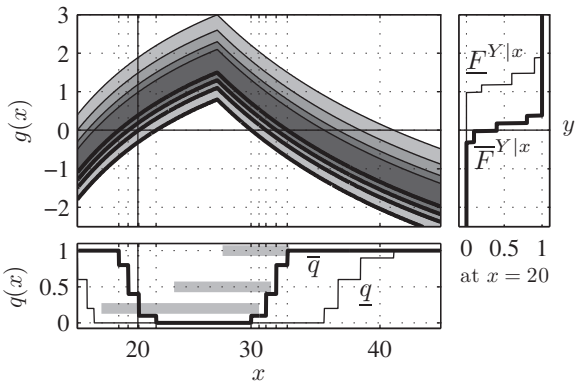


Figure 3: Random set $\mathscr{G}$, lower envelopes g^j; $\underline{q}(x) = \underline{F}^{Y|x}(0)$ and $\overline{q}(x) = \overline{F}^{Y|x}(0)$, focal sets of random set $\mathscr{A}$ (gray bars); $\underline{F}^{Y|x}$ and $\overline{F}^{Y|x}$ at $x = 20$.

Uncertainty of x modelled by a random set $\mathscr{A}$:
We consider now the special case where $h(x,z)$ is given by $g_z(x) = h(x,z) = g(x)+z$ resulting in the uncertain limit state function depicted in Fig. 3.

Then it holds for the lower envelopes $\underline{g}^j$ of the focal sets G_j that $\underline{g}^j \in G_j$. It is clear that we can reduce the focal sets G_j to their lower envelopes which leads to a discrete set of limit state functions eqipped with a probability distribution induced by the weights of the focal sets G_j. But then there is only one single probability distribution and therefore no possibility of choice which leads to $\overline{p}_f^{\mathrm{S}} = \overline{p}_f^{X \not\to Z}$. Further we have $\overline{p}_f^{\mathrm{S}} = \overline{p}_f^{\mathrm{RS}}$ because of the ordering of the four lower envelopes ($\underline{g}^1 \le \underline{g}^2 \le \underline{g}^3 \le \underline{g}^4$, see Fig. 3).

In the following the upper probabilities $\overline{p}_f^{\mathrm{RS}} = \overline{p}_f^{\mathrm{R},X \not\to Z}$ and $\overline{p}_f^{X \not\to Z}$ are computed: Since in our example the results coincide for all notions of independence we have the possibility to choose between two methods for the upper probability of failure where either discontinuous or continuous optimization problems involved: For the upper probability $\overline{p}_f^{X \not\to Z}$ in case of epistemic irrelevance we have to solve $|\mathscr{A}|$ discontinuous ($\overline{q}$ is discontinuous) optimization problems

$$\begin{aligned}\overline{p}_f^{X \not\to Z} &= \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i) \sup_{x \in A_i} \overline{q}(x) \qquad (57)\\ &= \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i) \sup_{x \in A_i} \sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{B}}(B_j) \chi(\underline{g}^j(x) \le 0)\\ &= 0.2 \cdot 1.0 + 0.3 \cdot 0.4 + 0.5 \cdot 1.0 = 0.82\end{aligned}$$

and for the upper probability in case of random set independence $|\mathscr{A}| \cdot |\mathscr{B}|$ continuous one (g is continuous):

$$\begin{aligned}\overline{p}_f^{\mathrm{RS}} = \overline{p}_f^{\mathrm{R},X \not\to Z} &= \sum_{i,j} m_{\mathscr{A}}(A_i) m_{\mathscr{B}}(B_j) \sup_{\substack{x \in A_i \\ z \in B_j}} p_f(x,z) \qquad (58)\\ &= \sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{B}}(B_j) \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i) \chi(\underline{\underline{g}^j(A_i)} \le 0)\\ &= \sum_{j=1}^{|\mathscr{B}|} m_{\mathscr{B}}(B_j) \mathrm{Pl}_{\mathscr{A}}(\underline{g}^j(x) \le 0)\\ &= 0.1 \cdot 1.0 + 0.3 \cdot 1.0 + 0.4 \cdot 0.7 + 0.2 \cdot 0.7 = 0.82.\end{aligned}$$

6.3 Random limit state functions

We have again $h(x,z) = g(x) + z$ and model the uncertainty of the parameter z by a Gaussian distribution (density f_b^Z) with parameters $b = (\mu, \sigma)$.

Let us start with deterministic parameters, say $b = (0, 0.5)$, which leads to $\overline{p}_f^{\mathrm{S}} = \overline{p}_f^{X \not\to Z}$. Using the notation of Sec. 4 we have $Y_{|x} = g(x) + Z$ with random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$ and conditional random variable $Y_{|x} \sim \mathcal{N}(g(x) + \mu, \sigma^2)$ given the basic variable x. The function q is obtained by

$$q(x) = \int_{\mathcal{Z}} \chi(g(x) + z \le 0) f_{(0,0.5)}^Z(z)\, \mathrm{d}z = F_{(g(x),0.5)}^{Y|x}(0) \qquad (59)$$

where $F^{Y|x}$ is the probability distribution function of $Y_{|x}$, cf. Sec. 4 and Fig. 4.

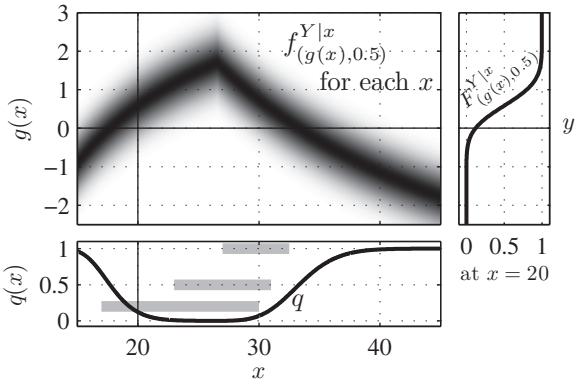


Figure 4: Uncertain limit state function $g(x) + z$ where the uncertainty of z is described by a Gaussian distribution with $\mu = 0$ and $\sigma = 0.5$; $q(x) = F_{(g(x),0.5)}^{Y|x}(0)$, focal sets of random set $\mathscr{A}$; $F_{(g(x),0.5)}^{Y|x}$ at $x = 20$.

Uncertainty of x modelled by an interval $A = [20, 30]$: We obtain

$$\overline{p}_f^{\mathrm{S}} = \overline{p}_f^{X \not\to Z} = \sup_{x \in A} p_f(x, b; \delta_x, f_b^Z) = \sup_{x \in A} q(x) = 0.1222 \qquad (60)$$

using Eqs. (19) and (20).

Uncertainty of x modelled by a random set: We get

$$\begin{aligned}\overline{p}_f^{\mathrm{S}} = \overline{p}_f^{\mathrm{R},X \not\to Z} = \overline{p}_f^{\mathrm{RS}} = \overline{p}_f^{X \not\to Z} &= \sum_{i=1}^{|\mathscr{A}|} m_{\mathscr{A}}(A_i) \sup_{x \in A_i} q(x) \qquad (61)\\ &= 0.2 \cdot 0.6604 + 0.3 \cdot 0.1576 + 0.5 \cdot 0.3682 = 0.3635\end{aligned}$$

for the random set $\mathscr{A}$ given in Sec. 5.2 using very simplified versions of Eqs. (26), (30) and (31).

Uncertainty of x modelled by a single probability distribution: For a Gaussian distribution (density f_a^X) with deterministic parameters $a = (\mu, \sigma) = (34, 1)$ we get the result

$$\begin{aligned}&p_f((34,1),(0,0.5), f_a^X, f_b^Z) = \qquad (62)\\ &= \int_{\mathcal{X}} \int_{\mathcal{Z}} \chi(g(x) + z \le 0) f_{(0,0.5)}^Z(z)\, \mathrm{d}z\, f_{(34,1)}^X(x)\, \mathrm{d}x\\ &= \int_{\mathcal{X}} q(x) f_{(34,1)}^X(x)\, \mathrm{d}x = 0.5976.\end{aligned}$$

Uncertainty of b modelled by a set B:
Let the set B for the parameters b in f_b^Z given by

$$[\underline{\mu}, \overline{\mu}] \times [\underline{\sigma}, \overline{\sigma}] = [-0.3, 0.3] \times [0.2, 0.6].$$

The function $\overline{q}$ and the corresponding lower bound $\underline{q}$ are obtained here by

$$\overline{q}(x) = \overline{F}^{Y|x}(0) = \begin{cases} F_{(g(x)+\underline{\mu},\overline{\sigma})}^{Y|x}(0) & \text{if } g(x) + \underline{\mu} > 0,\\ F_{(g(x)+\underline{\mu},\underline{\sigma})}^{Y|x}(0) & \text{if } g(x) + \underline{\mu} \le 0 \end{cases} \qquad (63)$$

and

$$\underline{q}(x) = \underline{F}^{Y|x}(0) = \begin{cases} F^{Y|x}_{(g(x)+\overline{\mu},\underline{\sigma})}(0) & \text{if } g(x)+\overline{\mu} > 0, \\ F^{Y|x}_{(g(x)+\overline{\mu},\overline{\sigma})}(0) & \text{if } g(x)+\overline{\mu} \leq 0. \end{cases} \tag{64}$$

In Fig. 5 the densities $f^{Y|x}_{(g(x)+\underline{\mu},\overline{\sigma})}$ and $f^{Y|x}_{(g(x)+\underline{\mu},\underline{\sigma})}$ resulting in $\overline{q}$ are depicted as well as the functions $\overline{q}$, $\underline{q}$ and the upper and lower distribution functions $\overline{F}^{Y|x}$ and $\underline{F}^{Y|x}$ at $x = 20$. The numerical results for uncertain x as above (set A, random set $\mathscr{A}$, probability distribution) can be obtained in case of epistemic irrelevance by simply replacing q by $\overline{q}$ in the Eqs. (60), (61) and (62). The results for $\overline{p}_f^{X \not\to Z}$ are 0.3192 for the set A, 0.6817 for the random set $\mathscr{A}$ and 0.9387 for the Gaussian distribution.

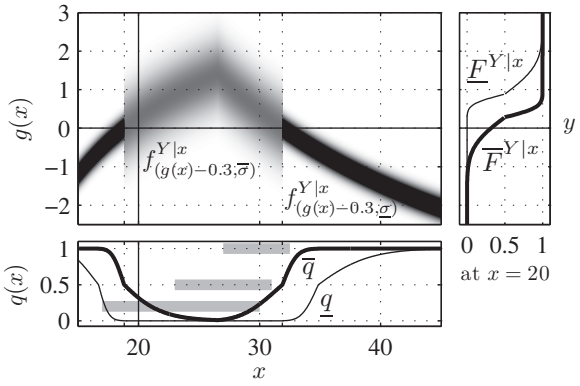


Figure 5: Uncertain limit state function $g(x)+z$ where the uncertainty of z is modelled by a set of Gaussian distributions.

Conclusion

To model uncertainties in limit state functions g we extended g depending on basic variables x to functions h by adding additional parameters z and introduced a function $p_f(a,b)$ for the probability of failure. This function provides an interface for controlling the parameters a and b of the probability density functions f^X_a and f^Z_b used for modelling the uncertainty of the basic variables x and the new additional parameters z. In a next step the two parameters a and b were assumed to be uncertain using sets or random sets to model their uncertainty resulting in sets of proability measures for x and z. In this context we discussed several notions of independence, gave computational formulas for different cases of uncertainty models exemplified by a simple engineering example and addressed visualization methods and alternative approaches as well.

References

[1] M. Beer. Fuzzy probability theory. In Meyers, editor, *Encyclopedia of Complexity and Systems Science*, volume 6, pages 4047–4059. Springer, New York, 2009.

[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 generated by a random closed interval. *Ann. Math. Stat.*, 39:957–966, 1968.

[4] S. Destercke. Independence concepts in evidence theory: some results about epistemic irrelevance and imprecise belief functions. Workshop on the Theory of Belief Functions, Brest 2010.

[5] O. Ditlevsen. Model uncertainty in structural reliability. *Structural Safety*, 1(1):73–86, 1982.

[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. *Mengen von gemeinsamen Wahrscheinlichkeitsmaßen erzeugt von zufälligen Mengen*. PhD thesis, Universität Innsbruck, 2003.

[8] Th. Fetz. Multi-parameter models: rules and computational methods for combining uncertainties. In W. Fellin, H. Lessman, R. Vieider, and M. Oberguggenberger, editors, *Multi-parameter models: rules and computational methods for combining uncertainties*. Springer, Berlin, 2004.

[9] Th. Fetz. Multiparameter models: Probability distributions parameterized by random sets. In *G. de Cooman, J. Vejnarova, M. Zaffalon (Eds.): ISIPTA '07, Proceedings of the Fifth International Symposium on Imprecise Probabilities and Their Applications*, Action M Agency, SIPTA, Prague, 317 - 326, 2007.

[10] 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.

[11] Th. Fetz and F. Tonon. Probability bounds for series systems with variables constrained by sets of probability measures. *Int. J. Reliability and Safety*, 2(4):309–339, 2008.

[12] J. Hall and J. Lawry. Fuzzy label methods for constructing imprecise limit state functions. *Struct. Saf.*, 28:317–341, 2003.

[13] I. Levi. *The Enterprise of Knowledge*. MIT Press, London, 1980.

[14] R. E. Melchers. *Structural Reliability Analysis and Prediction*. Wiley, Chichester, 1999.

[15] B. Möller, W. Graf, and M. Beer. Safety assessment of structures in view of fuzzy randomness. *Comput. Struct.*, 81:1567–1582, 2003.

[16] P. Walley. *Statistical reasoning with imprecise probabilities.* Chapman and Hall, London, 1991.

[17] L. A. Zadeh. Fuzzy sets. *Inf. Contr.*, 8:338–353, 1965.

Coherent conditional probabilities and proper scoring rules

Angelo Gilio
Dipartimento di Scienze di Base e Applicate per l'Ingegneria,
University of Rome "La Sapienza" (Italy)
gilio@dmmm.uniroma1.it

Giuseppe Sanfilippo
Dipartimento di Scienze Statistiche e Matematiche "S. Vianelli",
University of Palermo (Italy)
sanfilippo@unipa.it

Abstract

In this paper we study the relationship between the notion of coherence for conditional probability assessments on a family of conditional events and the notion of admissibility with respect to scoring rules. By extending a recent result given in literature for unconditional events, we prove, for any given strictly proper scoring rule s, the equivalence between the coherence of a conditional probability assessment and its admissibility with respect to s. In this paper we focus our analysis on the case of continuous bounded scoring rules. In this context a key role is also played by Bregman divergence and by a related theoretical aspect. Finally, we briefly illustrate a possible way of defining (generalized) coherence of interval-valued probability assessments by exploiting the notion of admissibility given for precise probability assessments.

Keywords. Conditional probability assessments, coherence, penalty criterion, proper scoring rules, conditional scoring rules, weak dominance, strong dominance, admissibility, Bregman divergence, g-coherence, total coherence, imprecise probability assessments.

1 Introduction

The theory and the applications of proper scoring rules have a long history in statistical literature (see, e.g., [1, 20, 23, 24, 25, 26, 28, 29, 31, 32, 33, 36, 37]). This theory was central to de Finetti's ideas about assessing the relative values of different subjective probability assessments ([9], see also [12]). A review of the general theory, with applications, has been given in [25] and, more recently, in [20]. A scoring rule for the probability of a given event E is a function of both the observation that comes to be observed, E true, or E false, and of the assessed probability $P(E)$. Assume that you were asked to assert $P(E)$, knowing that your assertion were to be scored according to the rule $s(E, P(E))$; moreover, assume that your degree of belief were $P(E) = p$, while you announced instead some other number $P(E) = x$, in the expectation that you would achieve a better score. The rule is said to be proper if you cannot expect a better score by specifying a value x different from p. Proper scoring rules encourage sincerity, because for you the best decision is to announce probabilities which conform to your beliefs.

The connections between the notions of coherence and of admissibility for probability assessments have been investigated in the work of de Finetti ([9, 10, 11]), by means of a penalty criterion based on the Brier quadratic scoring rule ([5]). A generalization of the work of de Finetti to a broad class of scoring rules has been given by Lindley in [26]. In his paper Lindley assumes suitable properties for the score function and admissibility for the numerical values which describe the uncertainty. Then, he shows that such numerical values can be transformed into numerical values which satisfy the basic properties of conditional probabilities.

The relationship between the notions of coherence and of non-dominance, with respect to strictly proper continuous scoring rules, has been investigated in [27]. In the same paper the connection of coherence and strictly proper scoring rules to Bregman divergence has been clarified.

A rich analysis of scoring rules which extends the results obtained in [27] to conditional probability assessments has been given in [33], where different notions of coherence have been discussed. In the same paper, some conditions are given under which the quadratic scoring rule can be replaced by a general strictly proper scoring rule, preserving the equivalence of the notions of coherence introduced through the gambling and the penalty arguments. In [33] are also examined the cases of scoring rules which are discontinuous and/or not strictly proper. In particular, in Example 8 of the same paper, by using a discontinuous strictly proper scoring rule it is shown that an incoherent probability assessment cannot be weakly

dominated by any coherent probability assessment, while it is dominated by other incoherent assessments. Moreover, in Example 9 of [33], by using a discontinuous merely proper scoring rule it is shown that a coherent probability assessment is weakly dominated by another coherent probability assessment.
In our paper we adopt a notion of coherence for conditional events which is different from that ones given in [33] and is based on the *strengthened* coherence principle of de Finetti ([11], vol. 2, Axiom 3, pag. 339). Such a strengthened principle allows to properly manage *conditioning* events with *zero probability* and, as proved in [14, 15] (see also [18]), is equivalent to the notion of coherence for conditional probability assessments studied by other authors; see e.g. [8, 22, 30, 34, 35]. In order to unify the treatment of unconditional and conditional events, the definition of coherence given by de Finetti with the penalty criterion was suitably modified in [15] (see also [16]).
As it can be shown by suitable examples (see [8, 17]), if a function P defined on a family of conditional events satisfies the axiomatic properties of a conditional probability, but the set of conditioning events doesn't have any structure, it may happen that P is not coherent. On the contrary, if P is coherent, then P satisfies all the properties of conditional probabilities. In particular, (strengthened) coherence requires that $0 \leq P(A|B) \leq 1$, for any given conditional event $A|B$. As another example, let us consider the assessment $P(A_1|B) = 0.9$, $P(A_2|B) = 0.7$, with $A_1 \wedge A_2 = \emptyset$ and $B \neq \Omega$ (see [33], p. 204). Such an assessment, which is coherent based on Definition 1 in [33], is not coherent in our approach.
We observe that the notions of coherence given in [33] and strengthened coherence are equivalent in the case of unconditional probabilities. Moreover, in Example 8 and Example 9 illustrated above only unconditional events are considered; hence, the corresponding results also hold in our approach. Then, in our paper we focus the analysis on continuous strictly proper scoring rules.
In this paper, using the strengthened notion of coherence, we extend the result given in [27] to the case of conditional events. We prove that, for any given (continuous) bounded strictly proper scoring rule s, a probability assessment on an arbitrary family of conditional events is coherent if and only if it is admissible with respect to s.
In ([33], p. 204) the authors leave open the question of whether their results still hold if one restricted the notion of coherence to require that the axioms of probability conditional on events with zero probability be satisfied. Our answer to this open question is that the equivalence between coherence and admissibility still holds with our notion of coherence (which restricts the notions of coherence used in [33]).
In our paper, based on the comments of an anonymous referee, we briefly examine how the notion of admissibility for precise probability assessments can be exploited in the case of interval-valued probability assessments.
The paper is organized as follows: In Section 2 we first give some preliminary notions; then, in Subsection 2.1 we recall the notion of coherence with the betting scheme; in Subsection 2.2 we give the notion of coherence with the penalty criterion of de Finetti; in Subsection 2.3 we illustrate, by a suitable alternative theorem, the equivalence of the betting scheme and the penalty criterion. In Section 3 we recall the notion of strictly proper scoring rule for unconditional events; then, we consider scoring rules for conditional events and we give the notions of weak and strong dominance, and of admissibility, for conditional probability assessments with respect to a scoring rule. We also consider a function $s(p, x)$ connected with the prevision of unconditional and conditional scoring rules. In Section 4 we illustrate some well known properties of $s(p, x)$. In Section 5 we recall the notion of Bregman divergence and a related theoretical aspect. Then, we prove for conditional probability assessments the equivalence between coherence and admissibility with respect to any continuous bounded strictly proper scoring rule. In Section 6 we recall the notions of g-coherence, coherence and total coherence for interval-valued probability assessments and we briefly examine how these notions can be defined by means of the admissibility property. Finally, in Section 7 we give some conclusions.

2 Some preliminary notions

Given a real function $P : \mathcal{K} \to \mathbb{R}$, where $\mathcal{K}$ is an arbitrary family of conditional events, let us consider a sub-family $\mathcal{F}_n = \{E_1|H_1, \ldots, E_n|H_n\} \subseteq \mathcal{K}$, and the vector $\mathcal{P}_n = (p_1, \ldots, p_n)$, where $p_i = P(E_i|H_i)\,, \;\; i = 1, \ldots, n$. The vector $\mathcal{P}_n$ represents the restriction of the function P to $\mathcal{F}_n$. We denote by $\mathcal{H}_n$ the disjunction $H_1 \vee \cdots \vee H_n$. Since

$$E_i H_i \vee E_i^c H_i \vee H_i^c = \Omega\,, \;\; i = 1, \ldots, n\,,$$

where Ω is the sure event, by expanding the expression $\bigwedge_{i=1}^{n}(E_i H_i \vee E_i^c H_i \vee H_i^c)$, we can represent Ω as the disjunction of 3^n logical conjunctions, some of which may be impossible. The remaining ones are the constituents generated by the family $\mathcal{F}$. We denote by $C_1, \ldots, C_m$ the constituents contained in $\mathcal{H}_n$ and (if $\mathcal{H}_n \neq \Omega$) by C_0 the further constituent $\mathcal{H}_n^c = H_1^c \cdots H_n^c$, so that

$$\mathcal{H}_n = C_1 \vee \cdots \vee C_m\,,$$
$$\Omega = \mathcal{H}_n^c \vee \mathcal{H}_n = C_0 \vee C_1 \vee \cdots \vee C_m\,, \quad m + 1 \leq 3^n\,.$$

2.1 Coherence with betting scheme

Using the same symbols for the events and their indicators, with the pair $(\mathcal{F}_n, \mathcal{P}_n)$ we associate the random gain

$$\mathcal{G} = \sum_{i=1}^{n} s_i H_i (E_i - p_i) \, ,$$

where $s_1, \ldots, s_n$ are n arbitrary real numbers. Let g_h be the value of $\mathcal{G}$ when C_h is true. Of course $g_0 = 0$ (notice that g_0 will not play any role in the definition of coherence). Denoting by $\mathcal{G}|\mathcal{H}_n$ the restriction of $\mathcal{G}$ to $\mathcal{H}_n$, it is $\mathcal{G}|\mathcal{H}_n \in \{g_1, \ldots, g_m\}$. Then, the function P defined on $\mathcal{K}$ is said *coherent* if and only if, for every integer n, for every finite sub-family $\mathcal{F}_n \subseteq \mathcal{K}$ and for every $s_1, \ldots, s_n$, one has

$$\min \, \mathcal{G}|\mathcal{H}_n \, \leq 0 \leq \max \, \mathcal{G}|\mathcal{H}_n \, . \tag{1}$$

Remark 1. If the function P is coherent, then it is called a *conditional probability* on $\mathcal{K}$. Notice that, if P is coherent, then P satisfies all the well known properties of conditional probabilities (while the converse is not true; see [8], Example 13; or [17], Example 8).

2.2 Coherence with penalty criterion

Another operational definition of probabilities based on the quadratic scoring rule has been proposed by de Finetti ([10, 11]). This definition has been extended to the case of conditional events in [15].
With the pair $(\mathcal{F}_n, \mathcal{P}_n)$ we associate the loss $\mathcal{L} = \sum_{i=1}^{n} H_i (E_i - p_i)^2$; we denote by L_h the value of $\mathcal{L}$ if C_h is true. If You specify the assessment $\mathcal{P}_n$ on $\mathcal{F}_n$ as representing your belief's degrees, You are required to pay a penalty L_h when C_h is true. Then, the function P is said *coherent* if and only if do not exist an integer n, a finite sub-family $\mathcal{F}_n \subseteq \mathcal{K}$, and an assessment ${\mathcal{P}_n}^* = (p_1^*, \ldots, p_n^*)$ on $\mathcal{F}_n$ such that, for the loss $\mathcal{L}^* = \sum_{i=1}^{n} H_i (E_i - p_i^*)^2$, associated with $(\mathcal{F}_n, \mathcal{P}_n^*)$, it results $\mathcal{L}^* \leq \mathcal{L}$ and $\mathcal{L}^* \neq \mathcal{L}$; that is $L_h^* \leq L_h \, , \quad h = 1, \ldots, m$, with $L_h^* < L_h$ in at least one case.
We can develop a geometrical approach to coherence by associating, with each constituent C_h contained in $\mathcal{H}_n$, a point $Q_h = (q_{h1}, \ldots, q_{hn})$, where

$$q_{hj} = \begin{cases} 1, & \text{if } C_h \subseteq E_j H_j, \\ 0, & \text{if } C_h \subseteq E_j^c H_j, \\ p_j, & \text{if } C_h \subseteq H_j^c. \end{cases} \tag{2}$$

Denoting by $\mathcal{I}$ the convex hull of the points $Q_1, \ldots, Q_m$, based on the penalty criterion, the following result can be proved ([15], see also [17])

Theorem 1. The function P is coherent if and only if, for every finite sub-family $\mathcal{F}_n \subseteq \mathcal{K}$, one has $\mathcal{P}_n \in \mathcal{I}$.

2.3 Equivalence between betting scheme and penalty criterion

The betting scheme and the penalty criterion are *equivalent* ([14, 15]). This equivalence can also be proved by the following steps ([18]):
1. The condition $\mathcal{P}_n \in \mathcal{I}$ amounts to solvability of the following system Σ in the unknowns $\lambda_1, \ldots, \lambda_m$

$$(\Sigma) \qquad \begin{cases} \sum_{h=1}^{m} q_{hj} \lambda_h = p_j \, , \;\; j = 1, \ldots, n \, ; \\ \sum_{h=1}^{m} \lambda_h = 1 \, , \;\; \lambda_h \geq 0 \, , \;\; h = 1, \ldots, m. \end{cases}$$

2. Let $\mathbf{x} = (x_1, \ldots, x_m)$, $\mathbf{y} = (y_1, \ldots, y_n)^t$ and $A = (a_{ij})$ be, respectively, a row $m-$vector, a column $n-$vector and a $m \times n-$matrix. The vector $\mathbf{x}$ is said *semipositive* if $x_i \geq 0, \, \forall \, i$, and $x_1 + \cdots + x_m > 0$. Then, we have (cf. [13], Theorem 2.9)

Theorem 2. Exactly one of the following alternatives holds.
(i) the equation $\mathbf{x}A = 0$ has a *semipositive* solution;
(ii) the inequality $A\mathbf{y} > 0$ has a solution.

We observe that, choosing $a_{ij} = q_{ij} - p_j$, $\forall \, i, j$, the solvability of $\mathbf{x}A = 0$ means that $\mathcal{P}_n \in \mathcal{I}$, while the solvability of $A\mathbf{y} > 0$ means that, choosing $s_i = y_i$, $\forall \, i$, one has $\min \, \mathcal{G}|\mathcal{H}_n \, > 0$ (and hence $\mathcal{P}_n$ would be incoherent). Therefore, by applying Theorem 2 with $A = (q_{ij} - p_j)$, we obtain $\max \, \mathcal{G}|\mathcal{H}_n \, \geq 0$ if and only if Σ is solvable, that is, $\max \, \mathcal{G}|\mathcal{H}_n \, \geq 0$ if and only if $\mathcal{P}_n \in \mathcal{I}$.

3 Scoring rules and admissibility for conditional probability assessments

In this section we recall the notion of (strictly) proper scoring rule for unconditional events; then, based on this notion, we consider scoring rules for conditional events, called *conditional scoring rules*. Then, we illustrate the notions of weak and strong dominance, and of admissibility, for a probability assessment with respect to a scoring rule.
A score may represent a reward or a penalty; we think of scores as penalties, so that to improve the score means to reduce it. To introduce strictly proper scoring rules, we use the definition given in [27].

Definition 1. A function $s : \{0, 1\} \times [0, 1] \to [0, +\infty]$ is said to be a strictly proper scoring rule if the following conditions are satisfied:
(a) for every $x, p \in [0, 1]$, with $x \neq p$, it is

$$p \, s(1, x) + (1 - p) \, s(0, x) > p \, s(1, p) + (1 - p) \, s(0, p) \, ; \tag{3}$$

(b) the functions $s(1, x)$ and $s(0, x)$ are continuous.

We observe that, if x is your announced probability for the event E, while p represents your degree of

belief on E, then the quantity $ps(1,x)+(1-p)s(0,x)$ is nothing but your expected score.
For brevity, a *strictly proper scoring rule* will be called *proper scoring rule*.
We indicate by the same symbol the events and their indicators. Then, given any event E, we have

$$s(E,x) = \begin{cases} s(1,x), & E, \\ s(0,x), & E^c. \end{cases}$$

Given a scoring rule s, with any conditional event $E|H$ we associate the conditional scoring rule $s(E|H,x) : \{0,1\} \times [0,1] \to [0,+\infty]$ defined as

$$s(E|H,x) = Hs(E,x) = \begin{cases} s(1,x), & EH, \\ s(0,x), & E^cH, \\ 0, & H^c. \end{cases}$$

We consider, for any given proper scoring rule s defined on the set $\{0,1\} \times [0,1]$, the extension of s to the set $[0,1] \times [0,1]$, defined as

$$s(p,x) = p\, s(1,x) + (1-p)\, s(0,x)\,. \tag{4}$$

We remark that, if x is your announced probability for the conditional event $E|H$, while p numerically represents your degree of belief on $E|H$, then the quantity $s(p,x)$ in (4) represents the *conditional prevision*

$$\mathbb{P}[s(E|H,x)\,|\,H] = \mathbb{P}[Hs(E,x)\,|\,H] = \mathbb{P}[s(E,x)\,|\,H]\,.$$

Moreover,

$$\mathbb{P}[s(E|H,x)] = s(1,x)P(EH) + s(0,x)P(E^cH)\,;$$

of course, $s(p,x) \neq s(1,x)P(EH) + s(0,x)P(E^cH)$.
Given a probability assessment $\mathcal{P}_n = (p_1, p_2, \ldots, p_n)$, with $p_i \in [0,1]$, on a family of conditional event $\mathcal{F}_n = \{E_1|H_1, E_2|H_2, \ldots, E_n|H_n\}$, where $p_i = P(E_i|H_i)$, and a proper scoring rule s, assuming that the scores are additive, we define the random penalty, or loss function, $\mathcal{L}$ associated with the pair $(\mathcal{F}_n, \mathcal{P}_n)$ as

$$\mathcal{L} = \sum_{i=1}^{n} s(E_i|H_i, p_i) = \sum_{i=1}^{n} H_i s(E_i, p_i).$$

For the Brier quadratic scoring rule $s(E,x) = (E-x)^2$ it is $s(E|H,x) = H(E-x)^2$. The loss function associated with this conditional scoring rule was used in [15] (see also [18]), in the framework of the penalty criterion of de Finetti, to give a unified definition of the notion of coherence for conditional and unconditional events.
For the (unbounded and proper) logarithmic scoring rule ([21]) $s(E,x) = -\log(1-|E-x|)$, we have

$$s(E|H,x) = -H\log(1-|E-x|)\,.$$

The associated random penalty is

$$\mathcal{L} = -\sum_{i=1}^{n} [E_i H_i \log p_i + E_i^c H_i \log(1-p_i)]\,,$$

which was proposed in ([25], p. 355) for the case of unconditional events, with $\{E_1, \ldots, E_n\}$ a partition of Ω. The above random penalty was used in [7] to introduce a suitable discrepancy measure with the aim of correcting incoherent conditional probability assessments.
Given the constituents $C_0, C_1, \ldots, C_m$ generated by $\mathcal{F}_n$, we denote by L_k the value of $\mathcal{L}$ associated with C_k, $k = 0, 1, \ldots, m$. Of course, $L_0 = 0$.

Definition 2. Let be given a scoring rule s and a probability assessment $\mathcal{P}_n$ on a family of n conditional events $\mathcal{F}_n$. Given any assessment $\mathcal{P}_n^*$ on $\mathcal{F}_n$, with $\mathcal{P}_n^* \neq \mathcal{P}_n$, we say that $\mathcal{P}_n$ is *weakly dominated* by $\mathcal{P}_n^*$, with respect to s, if denoting by $\mathcal{L}$ (resp., $\mathcal{L}^*$) the penalty associated with the pair $(\mathcal{F}_n, \mathcal{P}_n)$ (resp., $(\mathcal{F}_n, \mathcal{P}_n^*)$), it is $\mathcal{L}^* \leq \mathcal{L}$, that is: $L_k^* \leq L_k$, for every $k = 0, 1, \ldots, m$.

We observe that $\mathcal{P}_n$ is not weakly dominated by $\mathcal{P}_n^*$ if and only if $L_k^* > L_k$ for at least a subscript k.

Definition 3. Let be given a scoring rule s and a probability assessment $\mathcal{P}_n$ on a family of n conditional events $\mathcal{F}_n$. We say that $\mathcal{P}_n$ is *admissible w.r.t. s* if $\mathcal{P}_n$ is not weakly dominated by any $\mathcal{P}_n^* \neq \mathcal{P}_n$.

Remark 2. We observe that, by Definition 3, it follows:
- If the assessment $\mathcal{P}_n$ on $\mathcal{F}_n$ is admissible, then for every subfamily $\mathcal{F}_J \subset \mathcal{F}_n$ the sub-assessment $\mathcal{P}_J$ associated with $\mathcal{F}_J$ is admissible.

In order to manage infinite families of conditional events we give the following

Definition 4. Let be given a scoring rule s and a probability assessment $\mathcal{P}$ on an arbitrary family of conditional events $\mathcal{K}$. We say that $\mathcal{P}$ is admissible with respect to s if, for every finite subfamily $\mathcal{F}_n \subseteq \mathcal{K}$, the restriction of $\mathcal{P}$ on $\mathcal{F}_n$ is admissible w.r.t. s.

By observing that $L_0 = L_0^* = 0$, we give the following

Definition 5. Let be given a scoring rule s and a probability assessment $\mathcal{P}_n$ on a family of n conditional events $\mathcal{F}_n$. Given any assessment $\mathcal{P}_n^*$ on $\mathcal{F}_n$, we say that $\mathcal{P}_n$ is *strongly dominated* by $\mathcal{P}_n^*$, with respect to s, if $L_k^* < L_k$, for every $k = 1, \ldots, m$.

4 Properties of the function $s(p,x)$

For the convenience of the reader and to make our exposition self-contained, in the Proposition below we illustrate some well known properties of the function $s(p,x)$ defined in (4).

Proposition 1. Given a proper scoring rule s, the function $s(p,x)$ satisfies the following properties:

1. $s(\alpha p'+(1-\alpha)p'',x) = \alpha\, s(p',x)+(1-\alpha)\, s(p'',x)$;
2. $s(p,x) \geq s(p,p)$, with $s(p,x) = s(p,p)$ if and only if $x = p$;
3. $s(p,p)$ is strictly concave on $(0,1)$;
4. $s(p,x)$ is partially derivable with respect to x at (p,p), for every $p \in (0,1)$, and it is
$$\frac{\partial s(p,x)}{\partial x}|_{(p,p)} = 0\,;$$
5. for every $p \in (0,1)$, $s(p,p)$ is differentiable, with a continuous decreasing derivative
$$s'(p,p) = a(p) = s(1,p) - s(0,p)\,;$$
6. for every $p \in [0,1]$, $x \in (0,1)$, it is
$$s(p,x) = s(x,x) + s'(x,x)(p-x)\,.$$

Proof. 1. We have $s(p,x) = a(x)p + b(x)$, where
$$a(x) = s(1,x) - s(0,x)\,,\quad b(x) = s(0,x)\,,$$
so that
$$\begin{aligned} &s(\alpha p' + (1-\alpha)p'', x) = \\ &a(x)[\alpha p' + (1-\alpha)p''] + b(x)[\alpha + (1-\alpha)] = \\ &\alpha\, s(p',x) + (1-\alpha)\, s(p'',x). \end{aligned}$$

2. The property immediately follows by observing that the restriction of the function $s(p,x)$ to the set $\{0,1\}\times[0,1]$ is a proper scoring rule.
3. For every $x, y, \alpha \in (0,1)$, by setting $z = \alpha x + (1-\alpha)y$, we have $s(x,x) < s(x,z)$, $s(y,y) < s(y,z)$; then
$$\begin{aligned} &s(z,z) = s(\alpha x + (1-\alpha)y, \alpha x + (1-\alpha)y) = \\ &\alpha\, s(x,z) + (1-\alpha)\, s(y,z) > \alpha\, s(x,x) + (1-\alpha)\, s(y,y) \end{aligned}$$

4. Given any $p \in (0,1)$ and $0 < \varepsilon < 1-p$, by property 2 we have
$$\begin{aligned} &\frac{s(p,p+\varepsilon) - s(p,p)}{\varepsilon} > 0\,, \\ &\frac{s(p+\varepsilon,p+\varepsilon) - s(p+\varepsilon,p)}{\varepsilon} < 0\,. \end{aligned} \tag{5}$$
Moreover
$$\begin{aligned} &\tfrac{s(p+\varepsilon,p+\varepsilon)-s(p+\varepsilon,p)}{\varepsilon} = \\ &= \tfrac{s(p+\varepsilon,p+\varepsilon)-s(p,p+\varepsilon)}{\varepsilon} - \tfrac{s(p+\varepsilon,p)-s(p,p)}{\varepsilon} + \\ &+ \tfrac{s(p,p+\varepsilon)-s(p,p)}{\varepsilon} = \\ &= \tfrac{\varepsilon[s(1,p+\varepsilon)-s(0,p+\varepsilon)]}{\varepsilon} - \tfrac{\varepsilon[s(1,p)-s(0,p)]}{\varepsilon} + \\ &+ \tfrac{s(p,p+\varepsilon)-s(p,p)}{\varepsilon} = \\ &= \tfrac{\varepsilon[s(1,p+\varepsilon)-s(0,p+\varepsilon)]}{\varepsilon} - \tfrac{\varepsilon[s(1,p)-s(0,p)]}{\varepsilon} + \\ &+ \tfrac{s(p,p+\varepsilon)-s(p,p)}{\varepsilon} = \\ &= [s(1,p+\varepsilon)-s(0,p+\varepsilon)] - [s(1,p)-s(0,p)] + \tfrac{s(p,p+\varepsilon)-s(p,p)}{\varepsilon}\,. \end{aligned}$$

Then, by (5), it follows
$$\begin{aligned} &0 < \frac{s(p,p+\varepsilon) - s(p,p)}{\varepsilon} < \\ &< [s(1,p) - s(0,p)] - [s(1,p+\varepsilon) - s(0,p+\varepsilon)]\,, \end{aligned} \tag{6}$$
and by continuity of the function $s(1,x) - s(0,x)$ it follows
$$\lim_{\varepsilon\to 0^+} \frac{s(p,p+\varepsilon) - s(p,p)}{\varepsilon} = 0\,.$$
Analogously, given any $p \in (0,1)$ and $0 < \varepsilon < p$, by property 2 we have
$$\begin{aligned} &\frac{s(p,p-\varepsilon) - s(p,p)}{\varepsilon} > 0\,, \\ &\frac{s(p-\varepsilon,p-\varepsilon) - s(p-\varepsilon,p)}{\varepsilon} < 0\,. \end{aligned} \tag{7}$$
Moreover
$$\begin{aligned} &\tfrac{s(p-\varepsilon,p-\varepsilon)-s(p-\varepsilon,p)}{\varepsilon} = \\ &= \tfrac{s(p-\varepsilon,p-\varepsilon)-s(p,p-\varepsilon)}{\varepsilon} - \tfrac{s(p-\varepsilon,p)-s(p,p)}{\varepsilon} + \\ &+ \tfrac{s(p,p-\varepsilon)-s(p,p)}{\varepsilon} = \\ &= \tfrac{-\varepsilon[s(1,p-\varepsilon)-s(0,p-\varepsilon)]}{\varepsilon} - \tfrac{-\varepsilon[s(1,p)-s(0,p)]}{\varepsilon} + \\ &+ \tfrac{s(p,p-\varepsilon)-s(p,p)}{\varepsilon} = \\ &= -[s(1,p-\varepsilon)-s(0,p-\varepsilon)] + [s(1,p)-s(0,p)] + \tfrac{s(p,p-\varepsilon)-s(p,p)}{\varepsilon}\,. \end{aligned}$$

Then, by (7), it follows
$$\begin{aligned} &0 < \frac{s(p,p-\varepsilon) - s(p,p)}{\varepsilon} < \\ &< [s(1,p-\varepsilon) - s(0,p-\varepsilon)] - [s(1,p) - s(0,p)]\,, \end{aligned}$$
and by continuity of the function $s(1,x) - s(0,x)$ it follows
$$\begin{aligned} &\lim_{\varepsilon\to 0^+} \frac{s(p,p-\varepsilon) - s(p,p)}{-\varepsilon} = \\ &= -\lim_{\varepsilon\to 0^+} \frac{s(p,p-\varepsilon) - s(p,p)}{\varepsilon} = 0\,. \end{aligned}$$

Therefore, for every $p \in (0,1)$, there exists the partial derivative of $s(p,x)$ with respect to x at (p,p) and it is zero.
5. Given any $p \in (0,1)$ and $-p < \varepsilon < 1-p$, $\varepsilon \neq 0$, we have

$$\begin{aligned}&\tfrac{s(p+\varepsilon,p+\varepsilon)-s(p,p)}{\varepsilon} = \\ &= \tfrac{s(p+\varepsilon,p+\varepsilon)-s(p,p+\varepsilon)}{\varepsilon} + \tfrac{s(p,p+\varepsilon)-s(p,p)}{\varepsilon} = \\ &= \tfrac{\varepsilon[s(1,p+\varepsilon)-s(0,p+\varepsilon)]}{\varepsilon} + \tfrac{s(p,p+\varepsilon)-s(p,p)}{\varepsilon} = \\ &= [s(1,p+\varepsilon)-s(0,p+\varepsilon)] + \tfrac{s(p,p+\varepsilon)-s(p,p)}{\varepsilon}\,;\end{aligned}$$

then, by continuity of the function $s(1,x)-s(0,x)$ and by property 4, it follows

$$\begin{aligned}&s'(p,p) = \lim_{\varepsilon\to 0} \frac{s(p+\varepsilon,p+\varepsilon)-s(p,p)}{\varepsilon} = \\ &= a(p) = s(1,p) - s(0,p)\,.\end{aligned}$$

We observe that, in agreement with the strict concavity of $s(p,p)$ and as shown in (6), $a(p)$ is decreasing.
6. For every $p \in [0,1]$, $x \in (0,1)$, we have

$$\begin{aligned}&s(p,x) - s(x,x) = [a(x)p + b(x)] - [a(x)x + b(x)] = \\ &= s'(x,x)(p-x)\,;\end{aligned}$$

hence $s(p,x) = s(x,x) + s'(x,x)(p-x)$. □

5 Coherence and admissibility

In this section we recall the notion of Bregman divergence and a related theoretical aspect. Then, we prove the main result of the paper, by showing the equivalence between the coherence of conditional probability assessments and admissibility with respect to any bounded (strictly) proper scoring rule s.
Given two vectors

$$V_n = (v_1,\ldots,v_n),\ \ \mathcal{P}_n = (p_1,\ldots,p_n) \in [0,1]^n,$$

we set

$$S(V_n,\mathcal{P}_n) = \sum_{i=1}^{n} s(v_i,p_i)\,. \tag{8}$$

By property 3, the function S is strictly concave; moreover, by property 5, S is differentiable in $(0,1)^n$. By property 6, given any $\mathcal{P}_n \in (0,1)^n$ we have

$$\begin{aligned}&S(V_n,\mathcal{P}_n) = \textstyle\sum_{i=1}^{n}[s(p_i,p_i) + s'(p_i,p_i)(v_i-p_i)] = \\ &= S(\mathcal{P}_n,\mathcal{P}_n) + \nabla S(\mathcal{P}_n,\mathcal{P}_n)\cdot(V_n-\mathcal{P}_n)\,;\end{aligned} \tag{9}$$

then, by setting

$$\Phi(\mathcal{P}_n) = -S(\mathcal{P}_n,\mathcal{P}_n)\,,$$

we have

$$S(V_n,\mathcal{P}_n) = -\Phi(\mathcal{P}_n) - \nabla\Phi(\mathcal{P}_n)\cdot(V_n-\mathcal{P}_n)\,. \tag{10}$$

We recall that the function $s(p,p)$ is continuous on $[0,1]$ and strictly concave on $(0,1)$; then $\Phi(\mathcal{P}_n)$ is continuous on $[0,1]^n$ and strictly convex on $(0,1)^n$. Moreover, $s(p,p)$ has a continuous first derivative on $(0,1)$; then, the function $\Phi(\mathcal{P}_n)$ has continuous partial derivatives on $(0,1)^n$. Hence, $\Phi(\mathcal{P}_n)$ is differentiable on $(0,1)^n$ and its gradient $\nabla\Phi(\mathcal{P}_n)$ is a continuous function on $(0,1)^n$. If s is bounded, then $\nabla\Phi(\mathcal{P}_n)$ extends to a bounded continuous function on $[0,1]^n$.
In the definition below we recall the notion of Bregman divergence (see e.g. [6]).

Definition 6. Let $\mathcal{C}$ be a convex subset of $\mathbb{R}^n$ with nonempty interior. Let $\Phi : \mathcal{C} \to \mathbb{R}$ be a strictly convex function, differentiable in the interior of $\mathcal{C}$, whose gradient $\nabla\Phi$ extends to a bounded, continuous function on $\mathcal{C}$. For $V_n, \mathcal{P}_n \in \mathcal{C}$ the **Bregman divergence** $d_\Phi : \mathcal{C}\times\mathcal{C} \to \mathbb{R}$ corresponding to Φ is given by

$$d_\Phi(V_n,\mathcal{P}_n) = \Phi(V_n) - \Phi(\mathcal{P}_n) - \nabla\Phi(\mathcal{P}_n)\cdot(V_n-\mathcal{P}_n)\,.$$

It is $d_\Phi(V_n,\mathcal{P}_n) \geq 0$ and, as Φ is strictly convex, $d_\Phi(V_n,\mathcal{P}_n) = 0$ if and only if $V_n = \mathcal{P}_n$.
We remark that, assuming s bounded, $\mathcal{C} = [0,1]^n$ and $\Phi(\mathcal{X}) = -S(\mathcal{X},\mathcal{X})$, by (10) and Definition 6 it follows

$$d_\Phi(V_n,\mathcal{P}_n) = S(V_n,\mathcal{P}_n) - S(V_n,V_n)\,. \tag{11}$$

We observe that, for $s(E,x) = -\log(1-|E-x|)$, we have

$$S(V_n,\mathcal{P}_n) = -\sum_{i=1}^{n}[v_i\log p_i + (1-v_i)\log(1-p_i)]\,; \tag{12}$$

then, formula (11) becomes

$$d_\Phi(V_n,\mathcal{P}_n) = \sum_{i=1}^{n}\left[v_i\log\left(\frac{v_i}{p_i}\right) + (1-v_i)\log\left(\frac{1-v_i}{1-p_i}\right)\right].$$

This logarithmic Bregman divergence is connected with the discrepancy measure proposed in [7] to correct incoherent conditional probability assessments.
Now, we recall the following result given in [27]; see also [6].

Proposition 2. Let $d_\Phi : \mathcal{C}\times\mathcal{C} \to \mathbb{R}$ be a Bregman divergence and let $\mathcal{I} \subseteq \mathcal{C}$ be a closed convex subset of $\mathbb{R}^n$. For each $\mathcal{P}_n \in \mathcal{C}\setminus\mathcal{I}$, there exists a unique $\mathcal{P}_n^* \in \mathcal{I}$, called the **projection** of $\mathcal{P}_n$ onto $\mathcal{I}$, such that

$$d_\Phi(\mathcal{P}_n^*,\mathcal{P}_n) \ \leq\ d_\Phi(V_n,\mathcal{P}_n)\,,\ \forall\, V_n \in \mathcal{I}\,.$$

Moreover

$$\begin{aligned}&d_\Phi(V_n,\mathcal{P}_n^*) + d_\Phi(\mathcal{P}_n^*,\mathcal{P}_n) \ \leq\ d_\Phi(V_n,\mathcal{P}_n)\,, \\ &\forall\, V_n \in \mathcal{I}\,,\ \mathcal{P}_n \in \mathcal{C}\setminus\mathcal{I}\,.\end{aligned} \tag{13}$$

In the next result we illustrate the relationship between the notion of coherence and the property of non dominance.

Theorem 3. Let be given a probability assessment $\mathcal{P}$ on a family of conditional events $\mathcal{K}$; moreover, let be given any bounded (strictly) proper scoring rule s. The assessment $\mathcal{P}$ is coherent if and only if it is admissible with respect to s.

Proof. ($\Rightarrow$) Assuming $\mathcal{P}$ coherent, let s be any bounded proper scoring rule. Given any subfamily $\mathcal{F}_n = \{E_1|H_1, \ldots, E_n|H_n\}$ of $\mathcal{K}$, let $\mathcal{P}_n = (p_1, \ldots, p_n)$ be the restriction to $\mathcal{F}_n$ of $\mathcal{P}$. Now, given any $\mathcal{P}_n^* = (p_1^*, \ldots, p_n^*) \neq \mathcal{P}_n$, we distinguish two cases:
(a) $p_i^* \neq p_i$, for every $i = 1, \ldots, n$;
(b) $p_i^* = p_i$, for at least one index i.
Case **(a)**. We still denote by $C_0, C_1, \ldots, C_m$, where $C_0 = H_1^c \wedge \cdots \wedge H_n^c$, the constituents generated by $\mathcal{F}_n$ and by $Q_k = (q_{k1}, \ldots, q_{kn})$ the point associated with $C_k, k = 1, \ldots, m$.

We introduce the following binary quantities

$$e_{ki} = \begin{cases} 1, & C_k \subseteq E_i, \\ 0, & C_k \subseteq E_i^c, \end{cases} \quad , \quad h_{ki} = \begin{cases} 1, & C_k \subseteq H_i, \\ 0, & C_k \subseteq H_i^c. \end{cases}$$

Then, by recalling (2), for every $i = 1, \ldots, n$, $k = 1, \ldots, m$ it is

$$q_{ki} = e_{ki}h_{ki} + (1 - h_{ki})p_i \,. \tag{14}$$

With the assessment $\mathcal{P}_n$ it is associated the loss

$$\mathcal{L} = \sum_{i=1}^{n} [E_i H_i s(1, p_i) + E_i^c H_i s(0, p_i)] = \sum_{i=1}^{n} H_i s(E_i, p_i) \,;$$

of course, with any other assessment $\mathcal{P}_n^*$ on $\mathcal{F}_n$ it associated the loss

$$\mathcal{L}^* = \sum_{i=1}^{n} H_i [E_i s(1, p_i^*) + E_i^c s(0, p_i^*)] = \sum_{i=1}^{n} H_i s(E_i, p_i^*) \,.$$

For each constituent $C_k, k = 0, 1, \ldots, m$, the values of $\mathcal{L}$ and $\mathcal{L}^*$ are, respectively

$$L_k = \sum_{i=1}^{n} [e_{ki}h_{ki}s(1, p_i) + (1 - e_{ki})h_{ki}s(0, p_i)] \,,$$

$$L_k^* = \sum_{i=1}^{n} [e_{ki}h_{ki}s(1, p_i^*) + (1 - e_{ki})h_{ki}s(0, p_i^*)] \,.$$

By recalling that $L_0 = L_0^* = 0$, in what follows we will only refer to the values L_k, L_k^*, $k = 1, \ldots, m$.
As $\mathcal{P}_n$ is coherent, there exists a vector $(\lambda_1, \ldots, \lambda_m)$, with $\lambda_k \geq 0$ and $\sum_k \lambda_k = 1$, such that $\mathcal{P}_n = \sum_k \lambda_k Q_k$; that is, by (14)

$$p_i = \textstyle\sum_k \lambda_k q_{ki} = \sum_k \lambda_k e_{ki} h_{ki} + p_i - p_i \sum_k \lambda_k h_{ki} \,,$$

for every $i = 1, \ldots, n$; so that

$$\sum_k \lambda_k e_{ki} h_{ki} = p_i \sum_k \lambda_k h_{ki} \,, \; i = 1, \ldots, n \,,$$

or equivalently

$$\sum_k \lambda_k (1 - e_{ki}) h_{ki} = (1 - p_i) \sum_k \lambda_k h_{ki} \,, \; i = 1, \ldots, n \,.$$

Then,

$\sum_k \lambda_k L_k =$

$= \sum_k \lambda_k \sum_{i=1}^{n} [e_{ki}h_{ki}s(1, p_i) + (1 - e_{ki})h_{ki}s(0, p_i)] =$

$= \sum_i (\sum_k \lambda_k e_{ki} h_{ki})\, s(1, p_i) +$

$+ \sum_i (\sum_k \lambda_k (1 - e_{ki}) h_{ki})\, s(0, p_i) =$

$= \sum_i [p_i \sum_k \lambda_k h_{ki}) s(1, p_i) + (1 - p_i) \sum_k \lambda_k h_{ki}) s(0, p_i)]$

$= \sum_i (\sum_k \lambda_k h_{ki})\, [p_i\, s(1, p_i) + (1 - p_i)\, s(0, p_i)] \,.$

We set $I' = \{i : \sum_k \lambda_k h_{ki} > 0\} \subseteq \{1, 2, \ldots, n\}$. We observe that I' is not empty. In fact, for each $i = 1, \ldots, n$, there exists a constituent C_k such that $C_k \subseteq H_i$ and then $\sum_i h_{ki} \geq 1$. Moreover, as

$$\sum_i \sum_k \lambda_k h_{ki} = \sum_k \lambda_k \sum_i h_{ki} \geq \sum_k \lambda_k = 1 \,,$$

there exists an index i such that $\sum_k \lambda_k h_{ki} > 0$; i.e. $I' \neq \emptyset$.

Then, by recalling that for each $i = 1, \ldots, n$ it is

$$p_i\, s(1, p_i) + (1 - p_i) s(0, p_i) < p_i\, s(1, p_i^*) + (1 - p_i) s(0, p_i^*),$$

we have

$\sum_k \lambda_k L_k =$

$= \sum_{i \in I'} (\sum_k \lambda_k h_{ki}) [p_i\, s(1, p_i) + (1 - p_i)\, s(0, p_i)] <$

$< \sum_{i \in I'} (\sum_k \lambda_k h_{ki}) [p_i\, s(1, p_i^*) + (1 - p_i)\, s(0, p_i^*)] =$

$= \sum_i (\sum_k \lambda_k h_{ki}) [p_i\, s(1, p_i^*) + (1 - p_i)\, s(0, p_i^*)] =$

$= \sum_k \lambda_k L_k^* \,.$

The inequality $\sum_k \lambda_k L_k < \sum_k \lambda_k L_k^*$ implies that there exists an index k such that $L_k < L_k^*$; that is $\mathcal{L}^* > \mathcal{L}$ in at least one case. Hence $\mathcal{P}_n$ is admissible. Since $\mathcal{F}_n$ is arbitrary, it follows that $\mathcal{P}$ is admissible.
Case **(b)**. Let be given any $\mathcal{P}_n^* \neq \mathcal{P}_n$, with $p_i^* = p_i$, for at least one index i. We set $J = \{i : p_i^* \neq p_i\} \subset J_n = \{1, \ldots, n\}$. We denote by $\mathcal{P}_J$ (resp., $\mathcal{P}_{J_n \setminus J}$) the subvector of $\mathcal{P}_n$ associated with J (resp., $J_n \setminus J$).

Analogously, we can consider the subvectors $\mathcal{P}^*_J$ and $\mathcal{P}^*_{J_n\setminus J}$ of $\mathcal{P}^*_n$. Then, we have

$$\mathcal{L} = \mathcal{L}_J + \mathcal{L}_{J_n\setminus J}\,,\quad \mathcal{L}^* = \mathcal{L}^*_J + \mathcal{L}^*_{J_n\setminus J}\,,\quad \mathcal{L}_{J_n\setminus J} = \mathcal{L}^*_{J_n\setminus J}\,.$$

By the same reasoning as in case (a), it holds that $\mathcal{L}^*_J > \mathcal{L}_J$ in at least one case. Then, by observing that $\mathcal{L} - \mathcal{L}^* = \mathcal{L}_J - \mathcal{L}^*_J$, it is $\mathcal{L}^* > \mathcal{L}$ in at least one case; hence $\mathcal{P}_n$ is admissible. Since $\mathcal{F}_n$ is arbitrary, $\mathcal{P}$ is admissible.

($\Leftarrow$). We will prove that, given any bounded proper scoring rule s, if $\mathcal{P}$ is not coherent, then $\mathcal{P}$ is not admissible with respect to s. Assume that $\mathcal{P}$ is not coherent. Then, there exists a subfamily $\mathcal{F}_n = \{E_1|H_1, \ldots, E_n|H_n\} \subseteq \mathcal{K}$ such that, for the restriction $\mathcal{P}_n = (p_1, \ldots, p_n)$ of $\mathcal{P}$ to $\mathcal{F}_n$, denoting by $\mathcal{I}_n \subseteq [0,1]^n$ the associated convex hull, it is $\mathcal{P}_n \notin \mathcal{I}_n$. For each constituent C_k we set $I_k = \{i : C_k \subseteq H_i^c\}$, $J_k = \{i : C_k \subseteq H_i\}$; then, by recalling (11), the value L_k of the penalty $\mathcal{L}$ is given by

$$\begin{aligned}
&L_k = \textstyle\sum_{i=1}^n s(e_{ki}, p_i)h_{ki} = \\
&= \textstyle\sum_{i=1}^n s(q_{ki}, p_i) - \sum_{i\in I_k} s(p_i, p_i) = \\
&= \textstyle\sum_{i=1}^n s(q_{ki}, p_i) - \sum_{i=1}^n s(q_{ki}, q_{ki}) + \\
&+ \textstyle\sum_{i=1}^n s(q_{ki}, q_{ki}) - \sum_{i\in I_k} s(p_i, p_i) = \\
&= \textstyle\sum_{i=1}^n s(q_{ki}, p_i) - \sum_{i=1}^n s(q_{ki}, q_{ki}) + \\
&+ \textstyle\sum_{i\in J_k}^n s(e_{ki}, e_{ki}) = \\
&= S(Q_k, \mathcal{P}_n) - S(Q_k, Q_k) + \alpha_k = \\
&= d_\Phi(Q_k, \mathcal{P}_n) + \alpha_k\,,
\end{aligned} \tag{15}$$

where $\alpha_k = \sum_{i\in J_k}^n s(e_{ki}, e_{ki})$ and $\Phi(\mathcal{X}) = -S(\mathcal{X}, \mathcal{X})$. By applying Proposition 2 with $\mathcal{C} = [0,1]^n$ and $\mathcal{I} = \mathcal{I}_n$, by (13) we have

$$d_\Phi(Q_k, \mathcal{P}^*_n) + d_\Phi(\mathcal{P}^*_n, \mathcal{P}_n) \le d_\Phi(Q_k, \mathcal{P}_n)\,,$$

where $\mathcal{P}^*_n = (p^*_1, \ldots, p^*_n)$ is the projection of $\mathcal{P}_n$ onto $\mathcal{I}_n$. Moreover, as $\mathcal{P}^*_n \neq \mathcal{P}_n$ it is $d_\Phi(\mathcal{P}^*_n, \mathcal{P}_n) > 0$ and hence

$$d_\Phi(Q_k, \mathcal{P}^*_n) < d_\Phi(Q_k, \mathcal{P}_n)\,,\ k = 1, \ldots, m\,.$$

Now, denoting by $Q^*_1 = (q^*_{11}, \ldots, q^*_{1n}), \ldots, Q^*_m = (q^*_{m1}, \ldots, q^*_{mn})$ the points associated with the pair $(\mathcal{F}_n, \mathcal{P}^*_n)$, recalling property 2, for each $k = 1, \ldots, m$ we have

$$\begin{aligned}
&d_\Phi(Q_k, \mathcal{P}^*_n) - d_\Phi(Q^*_k, \mathcal{P}^*_n) = \\
&= S(Q_k, \mathcal{P}^*_n) - S(Q_k, Q_k) - S(Q^*_k, \mathcal{P}^*_n) + S(Q^*_k, Q^*_k) = \\
&= \sum_{i=1}^n [s(q_{ki}, p^*_i) - s(q_{ki}, q_{ki}) - s(q^*_{ki}, p^*_i) + s(q^*_{ki}, q^*_{ki})] = \\
&= \textstyle\sum_{i=1}^n [s(q_{ki}, p^*_i) - s(q^*_{ki}, p^*_i)] + \\
&- \textstyle\sum_{i=1}^n [s(q_{ki}, q_{ki}) - s(q^*_{ki}, q^*_{ki})] = \\
&= \textstyle\sum_{i\in I_k} [s(p_i, p^*_i) - s(p^*_i, p^*_i)] + \\
&- \textstyle\sum_{i\in I_k} [s(p_i, p_i) - s(p^*_i, p^*_i)] = \\
&= \textstyle\sum_{i\in I_k} [s(p_i, p^*_i) - s(p_i, p_i)] \ge 0\,.
\end{aligned}$$

Therefore, for each $k = 1, \ldots, m$, it is

$$d_\Phi(Q^*_k, \mathcal{P}^*_n) \;\le\; d_\Phi(Q_k, \mathcal{P}^*_n) \;<\; d_\Phi(Q_k, \mathcal{P}_n)\,.$$

Then, by (15), for each $k = 1, \ldots, m$ it follows

$$L^*_k = d_\Phi(Q^*_k, P^*_n) + \alpha_k \;<\; d_\Phi(Q_k, P_n) + \alpha_k = L_k\,;$$

that is, $\mathcal{P}_n$ is strongly dominated (and hence weakly dominated) by $\mathcal{P}^*_n$; hence $\mathcal{P}_n$ is not admissible. This implies that $\mathcal{P}$ is not admissible. □

We remark that in the first part of the proof of the previous theorem it has not been necessary to use the Bregman divergence .

We observe that Theorem 3 can be formulated in the following equivalent way.

Theorem 4. Given an arbitrary family of conditional events $\mathcal{K}$, let Π_c the set of coherent conditional probability assessments $\mathcal{P}$ on $\mathcal{K}$. Moreover, denoting by Σ the class of bounded (continuous strictly) proper scoring rules, let be given any $s \in \Sigma$. Then, let Π_s be the set of conditional probability assessments $\mathcal{P}$ on $\mathcal{K}$ which are admissible with respect to s. We have

$$\Pi_s = \Pi_c\,,\ \forall\, s \in \Sigma\,. \tag{16}$$

Remark 3. The equality (16) in the case $s(E, x) = (E - x)^2$ has been proved in [15] (see also [18]).

6 The case of imprecise probability assessments

In this section we illustrate a possible way of studying the relationship between coherence and admissibility with respect to scoring rules in the case of interval-valued conditional probability assessments. An anonymous referee observed that "there is an *impossibility* result due to the authors of [33] (probably still unpublished) showing that there does not exist a real-valued proper IP-scoring rule".

Moreover, the referee claims that in the same paper it is shown that "there is a lexicographic, i.e. non-standard valued, proper scoring rule for eliciting probability intervals".

Here, we just show that the notion of admissibility given for precise assessments can also be exploited in the case of imprecise probabilities.

We recall below the notions of generalized coherence (g-coherence, [2, 3, 4]), coherence and total coherence ([19]) for interval-valued conditional probability assessments.

Definition 7. Let be given an interval-valued probability assessment $\mathcal{A}_n = ([l_i, u_i], i = 1, \ldots, n)$, defined on a family of n conditional events $\mathcal{F}_n = \{E_i|H_i, i = 1, \ldots, n\}$. We say that:

a) $\mathcal{A}_n$ is g-coherent if there exists a coherent precise probability assessment $\mathcal{P}_n = (p_i, i = 1, \ldots, n)$ on $\mathcal{F}_n$, with $p_i = P(E_i|H_i)$, which is consistent with $\mathcal{A}_n$, that is such that $l_i \leq p_i \leq u_i$ for each $i = 1, \ldots, n$;
b) $\mathcal{A}_n$ is coherent if, given any $j \in \{1, \ldots, n\}$ and any $x_j \in [l_j, u_j]$, there exists a coherent precise probability assessment $\mathcal{P}_n = (p_i, i = 1, \ldots, n)$ on $\mathcal{F}_n$, which is consistent with $\mathcal{A}_n$ and is such that $p_j = x_j$;
c) $\mathcal{A}_n$ is totally coherent if every precise probability assessment $\mathcal{P}_n = (p_i, i = 1, \ldots, n)$ on $\mathcal{F}_n$, consistent with $\mathcal{A}_n$, is coherent.

We observe that the notions of g-coherence and coherence above amount to the well known notions of *avoiding uniform loss* and *coherence*, respectively, used in the literature on imprecise probabilities (see, e.g., [34]). Based on Definition 7 we can give the following versions of our main result in the case of interval-valued probability assessments.

Proposition 3. Let be given an interval-valued probability assessment $\mathcal{A}_n = ([l_i, u_i], i = 1, \ldots, n)$, defined on $\mathcal{F}_n = \{E_i|H_i, i = 1, \ldots, n\}$. Moreover, let be given any bounded (continuous and strictly) proper scoring rule s. We have:
a) $\mathcal{A}_n$ is g-coherent if and only if there exists a precise probability assessment $\mathcal{P}_n = (p_i, i = 1, \ldots, n)$ on $\mathcal{F}_n$, consistent with $\mathcal{A}_n$, which is admissible w.r.t. s;
b) $\mathcal{A}_n$ is coherent if, given any $j \in \{1, \ldots, n\}$ and any $x_j \in [l_j, u_j]$, there exists a precise probability assessment $\mathcal{P}_n = (p_i, i = 1, \ldots, n)$ on $\mathcal{F}_n$, with $p_j = x_j$, consistent with $\mathcal{A}_n$, which is admissible w.r.t. s;
c) $\mathcal{A}_n$ is totally coherent if every precise probability assessment $\mathcal{P}_n = (p_i, i = 1, \ldots, n)$ on $\mathcal{F}_n$, consistent with $\mathcal{A}_n$, is admissible w.r.t. s.

7 Conclusions

In this paper we have studied the relationship between the notion of (strengthened) coherence for conditional probability assessments and the property of admissibility with respect to scoring rules. We have extended to the case of conditional events a result given in [27] for unconditional events. We have shown that, given any bounded (continuous and strictly) proper scoring rule s, a conditional probability assessment on an arbitrary family of conditional events is coherent if and only if it is admissible with respect to s. To obtain our main result a key role has also been played by Bregman divergence. Finally, we have shown that the property of admissibility can be exploited to characterize the notions of g-coherence, coherence and total coherence for interval-valued conditional probability assessments.

Acknowledgments

The authors thank the anonymous referees for their very useful comments and suggestions.

References

[1] G. Agró, F. Lad, G. Sanfilippo. Sequentially forecasting economic indices using mixture linear combinations of EP distributions. *Journal of Data Science*, 8(1):101–126, 2010.

[2] V. Biazzo, 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.

[3] V. Biazzo, A. Gilio, G. Sanfilippo. Coherence Checking and Propagation of Lower Probability Bounds, *Soft Computing* 7, 310-320, 2003.

[4] V. Biazzo, A. Gilio, G. Sanfilippo. Generalized coherence and connection property of imprecise conditional previsions. In: *Proceedings of 12th Information Processing and Management of Uncertainty in Knowledge-Based Systems Conference*, IPMU 2008, Malaga, Spain, June 22 - 27, 907–914, 2008.

[5] G. W. Brier. Verification of forecasts expressed in terms of probability, *Monthly Weather Review*, 78:1–3, 1950.

[6] Y. Censor, S.A. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford Univ. Press, Oxford, 1997.

[7] A. Capotorti, G. Regoli, F. Vattari. Correction of incoherent conditional probability assessments. *International Journal of Approximate Reasoning*, 51:718-727, 2010.

[8] G. Coletti, R. Scozzafava. *Probabilistic logic in a coherent setting*. Kluwer, Dordrecht, 2002.

[9] B. de Finetti. Does it make sense to speak of 'good probability appraisers'? In: *The scientist speculates: an anthology of of partly-baked ideas*, I. J. Good (ed.), Heinemann, London, 357–364, 1962.

[10] B. de Finetti. Probabilità composte e teoria delle decisioni. *Rendiconti di Matematica*, 23:128-134, 1964.

[11] B. de Finetti. *Teoria delle probabilità*, voll. 1-2. Einaudi, Torino, 1970, (Engl. transl.: Theory of Probability, voll. 1-2, Wiley, Chichester, 1974, 1975).

[12] B. de Finetti. The Role of 'Dutch Books' and of 'Proper Scoring Rules'. *The British Journal for the Philosophy of Science*, 32(1):55-56, 1981.

[13] D. Gale. *The theory of linear economic models*. McGraw-Hill, New York, 1960.

[14] A. Gilio. Probabilità condizionate $\mathcal{C}_0$-coerenti. *Rendiconti di Matematica*, Serie VII, 9:277–295, 1989.

[15] A. Gilio. Criterio di penalizzazione e condizioni di coerenza nella valutazione soggettiva della probabilità. *Boll. Un. Mat. Ital.*, [7a] 4-B(3): 645–660, 1990.

[16] A. Gilio. C_0-Coherence and Extension of Conditional Probabilities. In: *Bayesian Statistics 4*, J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith (eds.), Oxford University Press, 633-640, 1992.

[17] A. Gilio. Algorithms for precise and imprecise conditional probability assessments. In: *Mathematical Models for Handling Partial Knowledge in Artificial Intelligence*, G. Coletti, D. Dubois, R. Scozzafava (eds.), Plenum Press, New York, 231–254, 1995.

[18] A. Gilio. Algorithms for conditional probability assessments. In: *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*, D. A. Berry, K. M. Chaloner, J. K. Geweke (eds.), John Wiley, 29–39, 1996.

[19] A. Gilio, S. Ingrassia. Totally coherent set-valued probability assessments. *Kybernetika* 34(1): 3–15, 1998.

[20] T. Gneiting, A. Raftery. Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[21] I. J. Good. Rational decisions. *J. Roy. Statist. Soc.*, 14:107-114, 1952.

[22] S. Holzer. On coherence and conditional prevision. *Boll. Un. Mat. Ital.*, 4(6):441–460, 1985.

[23] A. D. Hendrickson, R. J. Buehler. Proper scores for probability forecasters. *Ann. Math. Statist.* 42:1916-1921, 1971.

[24] V. R. R. Jose, R. F. Nau, R. L. Winkler. Scoring Rules, Generalized Entropy, and Utility Maximization, *Operations Research*, 56(5):1146-1157, 2008.

[25] F. Lad. *Operational Subjective Statistical Methods: a mathematical, philosophical, and historical introduction*. John Wiley, New York, 1996.

[26] D. V. Lindley. Scoring rules and the inevitability of probability. *Int. Statist. Rev.*, 50:1-11, 1982.

[27] J. B. Predd, R. Seiringer, E. H. Lieb, D. N. Osherson, H. V. Poor, S. R. Kulkarni. Probabilistic Coherence and Proper Scoring Rules. *IEEE T. Inform. Theory*, 55:4786–4792, 2009.

[28] J. McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42:654–655, 1956.

[29] R. F. Nau, V. R. R. Jose, R. L. Winkler. Scoring Rules, Entropy, and Imprecise Probabilities. In: *Proceedings of the 5th International Symposium on Imprecise Probabilities and their Applications, ISIPTA07*, Prague, Czech Republic, 307–316, 2007.

[30] E. Regazzini. Finitely additive conditional probabilities. *Rend. Sem. Mat. Fis. Milano*, 55:69–89, 1985.

[31] G. Sanfilippo, G. Agró, F. Lad. Assessing fat-tailed sequential forecast distributions for the Dow-Jones index with logarithmic scoring rules. In: *Proc. of 56th Session of the International Statistical Institute*, Lisbon, 22-29, August, 2007.

[32] L. J. Savage. Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.*, 66:783-801, 1971

[33] M. J. Schervish, T. Seidenfeld, J. B. Kadane. Proper Scoring Rules, Dominated Forecasts, and Coherence. *Decision Analysis*, 6(4):202–221, 2009.

[34] P. Walley, R. Pelessoni, P. Vicig. Direct Algorithms for Checking Coherence and Making Inferences from Conditional Probability Assessments, *Journal of Statistical Planning and Inference*, 126(1), 119-151, 2004.

[35] P. M. Williams. 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.

[36] R. L. Winkler. Scoring rules and the evaluation of probability assessors. *J. Amer. Statist. Assoc.*, 64:1073-1078, 1969.

[37] R. L. Winkler. Scoring rules and the evaluation of probabilities. *Test* 5(1):1-60, 1996.

Potential Surprises

Frank Hampel
ETH Zurich, Switzerland
hampel@stat.math.ethz.ch

Abstract

After a brief historical overview over various approaches to the foundations of statistics, the very general, simple and basic concept of (potential) surprises is introduced, which may be subjective or objective and goes beyond previous approaches by I.J. Good and by the author. The surprises are conditional on the background knowledge or belief of the person experiencing it; the updating of the so-called background, and the merging or, if not possible, the contrasting of different backgrounds by two or more persons (otherwise they talk past each other) are very important operations in practice. A number of examples from real life, in complement to two previous, more qualitative papers, are given.

Keywords. Foundations of statistics, historical concepts, (potential) surprises, background knowledge or belief, combining of backgrounds, updating of backgrounds, merging or contrasting of backgrounds, practical application of mathematical models, real life examples.

1 Introduction and overview

Over the centuries, there have been various different approaches towards the fundamental concepts of statistics.

One line of thought focusses on observed and hypothetical frequencies of "random" events, especially – usually under some symmetry assumptions – the expectations for games of chance. (I shall leave aside the various philosophical meanings of "randomness".) After some previous isolated attempts, this led to the work by Pascal ([33], cf. also [18, Ch. 8], and also [3, Part 4, Ch. XVI, p. 387-388]) and Fermat (starting 1654), Huygens, de Moivre, Laplace, and later the frequentist theories by von Mises, Neyman-Pearson (cf., e.g., [31]), and Wald, among others. A basic tool was the law of large numbers [5, Part 4.5] and its later refinements, which in practice allowed to approximately equate the observed percentage of successes in a "long" sequence (whatever that means, cf. [22, Part 5]) of random experiments with their theoretical probability.

Besides a lot of mathematical work building upon the basic assumptions, there is still a chance for new ideas about this line of foundations, as shown by Cattaneo's [8, 9] improvement of Wald's minimax principle.

A very different approach to the foundations of stochastics, which apparently has not found much attention, goes back to Jacob Bernoulli [5, Part 4]. Besides continuing the work of Huygens (and discovering the law of large numbers, his "theorema aurea", as an auxiliary tool for something very different), he tried to develop a quantitative counterpart to the (then very famous) dichotomic "Logique" by Arnauld and Nicole [3]. His aim was to measure the degree of "probability" in the old, qualitative sense of this word (cf. [7, Ch. 2.8, 5.2, 5.3], also briefly [30, Ch. 2, espec. first paragraph, and Ch. 4.3]). Apparently by a misunderstanding of the eloges at the death of Bernoulli in the year 1705, the term "probability" was then used also for games of chance (cf. [7, Ch. 6.4.2]). (It is interesting to observe that both the terms "probability" and "statistics" (cf. [35, pp. 2f and 8f]) originally had a very different meaning.) Bernoulli proposed in a normative way in words (not in formulas) 9 "axioms" or basic (self-evident) properties which the new probability ought to obey and which would exclude, for example, both the Bayesian and the Neyman-Pearson theory and would not even obey in general the rule of additivity (cf. [7, Ch. 5.3.2]). But he still counted cases, as is done in games of chance. Perhaps he hoped to be able to derive "objective" results. Altogether, his approach (left incomplete because of his death) is a bold, fascinating and singular but perhaps shaky edifice; it seems not clear to me whether it can be worked out to a fully functioning system.

Another approach has become highly influential, namely the approach by Bayes [4]. Contrary to common belief, Bayes was basically a frequentist; only in his famous Scholium did he leave some open questions between the lines (which may be even the reason for not publishing his paper during his lifetime) which led to the later "Bayesian" interpretation (cf. [23, Ch. 1.3]). But cf. also the critical remarks by Boole [6].

Variants of the "Bayesian" approach were used by Laplace, Jeffreys (both normative or "logical" in different ways) and de Finetti (subjectivistic). Especially de Finetti [11] was a very sharp, radical, philosophically deep and fascinating thinker, building a pure and clean theory (although he did also applications). But if his theory is taken literally, I find it in its last consequence solipsistic, without any relation to anything like a "real world" – which for him does not exist - or to any fellow scientist. (I only know from L.J. Savage, one of his main pupils, that Savage was sometimes pragmatic in applications; moreover, his (the latter's) pupil D. Ellsberg showed with his paradox that some basic assumptions of Neo-Bayesians do not work in practice.)

All Bayesians (including "logical" or "objective" Bayesians) consider only epistemic probabilities (referring to our knowledge or belief about Nature, not about the (principally unknown) state of Nature itself. (Empirical Bayesians use frequentist methods.) On the other hand, all (traditional) frequentists consider only (usually unknown) aleatory probabilities in Nature, without any reference to what we know or may know (the few cases with known or strongly believed probability models excepted). Perhaps the first one to build a formal bridge between aleatory and epistemic probabilities was R.A. Fisher [15] with his fiducial argument. Unfortunately, he later made a mistake in its interpretation, but this mistake can be corrected, and Fisher's (corrected) fiducial probabilities just turn out to be a very special case of a general theory, using upper and lower probabilities [21, 23, 24, 26, 27].

It seems still unbelievable to many mathematical statisticians that one can derive known epistemic probabilities from unknown aleatory probabilities; but this is correctly done by most intelligent users of statistics who have not given up their own intuition in favor of either the Neyman-Pearson or the Bayesian theory (which, to be sure, are correct as far as they go, but in my eyes do not cover all the needs of good applied statistics, cf. [21, Ch. 4, p. 130], [23, 1.3], [24, Ch. 1.1]). And it has been done long ago, also at the early time of Fisher, cf., e.g., Student [38] or Pearson & Wishart [34]. Even though Fisher seemed only intuitively and not rationally clear about it, his concept of confidence intervals was clearly epistemic (and hence allowing a correct "aposteriori" interpretation), while that of Neyman was clearly aleatory, explaining Fisher's original doubt and later his conviction that despite all superficial formal similarities the two concepts are indeed different, referring to two different probability spaces.

In recent times, there are a number of approaches to statistics in a very broad sense using something like upper and lower probabilities, instead of strictly additive probabilities, as is only too well known at ISIPTA conferences (cf., e.g., [14, 37, 39, 40]; cf. also the ISIPTA conferences). Also some other statisticians, although claiming to be Bayesians, occasionally or inconspicuously use upper and lower probabilities, notably Dempster [12, 13] and Good (cf., e.g., [17]). There are a number of different concepts defined and many results developed. Also one of my lines of work [21, 23, 24] which centrally uses bets (like the Bayesians), but introduces also one-sided bets (thus leading beyond Bayesians) and uses also upper and lower fiducial probabilities, bridging the gap between aleatory and epistemic probabilities, belongs to this body of research.

A main goal of the present paper is to present several new concepts, with the help of practical examples, which ought to be able to describe the inference process on a higher level (cf. also [28, 29]). Although in my opinion inductive reasoning will be done mostly in a qualitative or semi-quantitative framework (using a discrete ordinal scale) as in the previous papers, an attempt is made specifically in this paper to allow also the introduction of a quantitative theory, still leaving a lot of freedom for the precise choices in detail.

There is a concept which looks so simple and at the same time so basic that it seems surprising that it is not more popular in statistical theories: the concept of potential surprises, or of surprises, for short. It can be used as a generic, rather encompassing term; in special situations, it can also be defined as, for example, minus the logarithm of a probability, then giving it a quantitative interpretation. This interpretation is of course inherent in information theory, though it is not normally given a special name. I.J. Good ([16], but not [17]), in the spirit of pure mathematics, has defined a whole mathematical class of surprises. A related definition of surprise is independently given by Hampel [21, Ch. 5]; it turned out that it differs from what Good (orally) considered his most important special case just by an additive constant. But surprises in my present theory can be given any subjective (numerical) interpretation (as is the case with beliefs, subjective probabilities, etc.). This concept, which has been neglected so far, is qualitatively (and

semi-quantitatively) investigated in Hampel [28, 29], with a number of practical examples. Both papers are in close connection with the present paper. It is my belief that (like elsewhere in stochastics) the precise numbers don't matter so much as the more qualitative features. (This is shown in the examples of the two previous papers.)

However, some people may want a general quantitative theory, and for this purpose the present paper is written. Yet, to avoid misunderstandings, this paper is basically philosophical and is derived from practical experience of everyday life (including experience in science). It is not derived from some system of axioms. Personally (and perhaps in an oldfashioned way), I do not start with axioms (not even intuitive looking ones as did Jacob Bernoulli), but rather I think axioms should be the crown at the end of the development of a body of knowledge. Later, there were historical reasons for the Bourbaki style in pure mathematics (trying to derive everything quickly on the highest and most abstract level); but I find this even dangerous, as the connections with the intuitive sources, including the nonmathematical sources, easily tend to get lost. As I try to derive all concepts from practical experience, and as this paper is work in progress (with some open questions, e.g. at the end of Ch. 4.2), I shall not try to present an axiomatic development of surprise. (It may even be argued that the problems Jacob Bernoulli had with his approach - see above - may partly be due to premature axiomatization - even though, or perhaps because, he partly relied on the Logique of Port-Royal, cf. [7, Ch. 5.3.2 and 5.2.2.4.].)

This paper contains several examples for the use of the new concepts; for more examples, see the other two papers [28, 29]. One example could even be continued: while the Arctic Warbler (Phylloscopus borealis) in Poland and South China had exactly the same song [29, Ch. 6.5], to my big surprise the same species in Japan had a very different song. It turned out that in Japan breeds a different subspecies (Ph. b. xanthodryas), and it is presently being investigated whether it ought to be separated as a new species from the nominate form Ph. b. borealis.

As stressed already in the previous two papers, the surprises are conditional on the assumed background belief or available background knowledge (both in their intuitive senses), both formally called background for short; and updating of the background, when new information becomes available, is a very important part of the inference or learning process.

The structure of the background is described more fully in [29] and the corresponding poster. Briefly, it consists of all our knowledge, beliefs, conditional or hypothetical beliefs, etc.; but more importantly, it exists in layers, and normally we use only the uppermost layer, containing our most plausible (or likely, or "normal") world view; only when we get a (complete or almost) contradiction with it by some new information (an infinite or close to infinite surprise), we abandon the uppermost layer and fully switch to the next one [29, Sec. 4]. This is (normally) a qualitative change of the background, not just a belief revision (cf., e.g., the article on Belief revision in Wikipedia (05/02/2011)) which slightly modifies the old belief system by means of some logical operations, or information fusion (information integration) or such an operation. It is not a deductive operation, but an inductive jump (cf. the examples); the old theory is false and not just modified, but replaced by something new, created by inductive thinking from the deeper, more hidden layers of our background.

(It might be argued that by enough logical operations one can change the background also to something qualitatively new; but this is not what I experience in the real world examples that came to my mind. I noted already [29, end of Sec. 1] that in the about 20 pages of [1] I could not find a single real life example, while they abound in my papers. To be sure, there is a place for deductive-logical operations; but I suggest that the creative inductive thinking process which generates genuine new knowledge has been badly neglected in research.)

Another important part of the inference process is the merging, or, if this is not possible, the contrasting of the backgrounds of two or more different persons (cf. Ch. 2.3).

As already briefly mentioned above, surprises in my approach are completely compatible with belief theory, Bayesian theory, and so on. They may be seen as a kind of superstructure over the old theories. As long as the surprises (in whatever reasonable way they are measured) are in an intermediate or low range, nothing essential changes. But if they are equal or close to infinity, then the background has to be changed. –

A word or two on terminology: It seems there are too few words in our language for all the different concepts that have been defined. The editors kindly drew my attention to [36] who used not only the term "surprise", but even "potential surprise" (loc. cit., Part II, espec. Ch. IX). There is much overlap and in part(!) a very similar intuition; moreover, style and basic philosophical attitude are quite similar. But I am mainly interested in inference, and Shackle in decisions, especially in economics; his formal definitions are different from mine (for example, by always demanding also a

surprise of zero in any disjunction); but most importantly, I could not find the change of background in case of (as he calls it) maximum surprise, which is so central in my approach. He also seems to avoid "...or any other..." which for me opens the door to the radical change of background. Some of his argumentation (against traditional probability theory) now may seem outdated, especially at an ISIPTA conference, and he also has run into problems with his attempt at an axiomatization (cf. above); but overall I find his thinking and arguing quite inspiring, although there is only partly an overlap in our approaches. (At least, the use of the same term does still seem bearable, as long as one is conscious of the differences.)

Another author who introduced the term "surprise" is Neumaier [32]. Again, he just tries to modify the old background in view of contradictions, not abandoning it, by finding an optimal compromise (with an "army of computer slaves" in the fictive story of the king on p. 22), minimizing the total surprise. (This may be appropriate if the surprise is so moderate that the background must only be slightly modified, not abandoned entirely.) And again, Neumaier finds much basic intuition in common with Shackle, but many formal differences. –

Our paper starts with basic definitions, properties and examples, which are not only mathematical since the application of the theory needs also close connections with the nonmathematical world (cf. [22]). Then the case of two or more different background assumptions is discussed, the connection with cautious surprises and successful bets is explained, and the problem of two (or more) persons with different backgrounds is brought into view. The updating of the background information is shown with a complex example, and another complex example asks among other points what to do if an event is totally unexpected. A practical example on how to concentrate incomplete knowledge in a fairly effective way concludes the paper.

2 Basic concepts

The following subsections introduce some basic definitions, properties and examples.

2.1 Basic set-up

Consider one person, say, Ted, with his background knowledge and collection of beliefs B, and a class of uncertain (future or unknown) events E which are of interest to Ted.

Let A be an event in E, and define the nonnegative number $s = s(A|B)$ to be the surprise of Ted, given his background B, when A turns out to be true; with $s = 0$ meaning no surprise at all; $s = \infty$ means Ted considers A impossible; and s "close to ∞" means Ted considers A "practically impossible".

More precisely, we have to distinguish
(i) the hypothetical surprise of Ted when he imagines that A shall happen or (unknown to him) has happened (the potential surprise in the strict sense, cf. also [36]);
(ii) the reaction of Ted when (perhaps in the future) he is reliably told that A has happened (or is for sure going to happen); and
(iii) Ted's reaction when he observes A himself.
(There is not much difference between (ii) and (iii) except the additional reliability by observing A oneself; on the other hand, Ted can also err himself.) Situation (i) requires that Ted thinks of the possibility that A may happen.

There may be possible events which Ted does not even think of. In such a case, Ted's surprise under (ii) or (iii) may still be small if he notes he just has forgotten a rather likely possibility; the surprise shall be very large, if on hindsight he considers the event A possible though highly unlikely; the surprise may be even infinite if A is not compatible with the assumed background B. In this case, Ted has to change his background belief B (one of the most fruitful sources of qualitatively new knowledge), or else he has to change his interpretation of the observation A (e.g., by finding an error in the observation).

In general, we shall change the background, going to the next layer (see above), not only when s is infinite, but also when s is "close to infinity". This is in analogy with what is also called Cournot's principle: that in the applications of probability theory, we consider an event with probability "close to one" as "practically certain" or (formerly) "morally certain". The boundary may depend on circumstances; Bernoulli gives as an example 999/1000 [7, p. 230], while Cournot [10, Ch. IV, 48] requires the difference to one to be "infinitesimally small" for an event to be "physically certain". No matter in what way the surprise is defined, I find the change of background as described the most important application of the concept of surprise.

(A logician might ask what is Ted's surprise by A if he has an "empty" background, e.g. if he wakes up from a coma and has lost all memory and all thinking ability. Then all his surprises are zero, because everything is fully possible. As soon as Ted starts thinking again, one has to be very careful in sorting out what he is able to think and learn again.)

2.2 Some basic properties of surprises

As mentioned above, a surprise s is a real number between zero and infinity, depending on the background knowledge or belief B of a person (here called Ted) and on the (perhaps fictive) occurrence of an event A.

It may be fully subjective, or it may be determined by objective circumstances, yielding an intersubjectively determined number (i.e., the same one for every person with the same background B). In either case, it is an epistemic quantity, that is, it refers to the knowledge or belief of a person, and not to some "objective" property of Nature (unless the two happen to coincide).

Example 1: Let F be a probability space with a known probability $P = P(A)$ for every (measurable) event A in F. Then we may define $s(A|B = F) = -\log P(A)$. In this case, s is a very natural "objective" measure for our surprise in case A happens. Some mathematical properties of s follow in this case, for example, its wellknown additivity. In particular, the expected value of s may be termed the entropy of F. And this entropy may be called the minimum possible average surprise of Ted. If Ted entertains another surprise function s', his average surprise, averaged over all possible events A with their probabilities, will be at least as large.

2.3 Two background assumptions

Now consider the situation that Ted entertains two different background belief systems B and C, perhaps being in doubt which one he should adopt. This may be the case in a learning situation, or in a conflict between different beliefs. If he would be not surprised if either told reliably that B is true, or else that C is true, his (minimum) surprise when observing A is $s(A|B \text{ or } C) = \min(s(A|B), s(A|C))$.

A more refined and more realistic situation is that Ted has different ("apriori") surprises $b(B)$ and $c(C)$ if told that B or C, resp., is true. (The functions or numbers b and c measure the surprise if Ted is told that a specific belief system is true. They may be different numbers, therefore the change of notation from s. In the following we require an additivity property of surprises, as in Example 1.) We call the three surprises s, b, and c unrelated if no occurrence of A or B or C (or a subset of these) affects any other surprise. Then Ted's minimum surprise $s(A|(B \text{ with } b) \text{ or } (C \text{ with } c)) = \min(s(A|B) + b(B), s(A|C) + c(C))$. Naturally, this can also be done with more than two beliefs (cf. 2.4).

The observation A in turn influences Ted's ("aposteriori") surprises about B and C, given A: $b(B|A) = b(B) + s(A|B)$, and correspondingly for $c(C|A)$. (The notation is a bit stretched, as A is not a belief system, but the meaning should be clear.) Naturally, this is close to Bayes' theorem, except that we do not introduce and do not need the renormalization.

The main purpose of computing $b(B|A)$ and comparing it with $c(C|A)$ is that if $b(B|A)$ is infinite, B cannot be used anymore as a background belief (except in case of an error in A); but also if $b(B|A)$ is "much" larger than $c(C|A)$, B is "practically impossible". This is in accordance with common sense thinking (except in case of a very strong prejudice in favor of B which, however, would also imply a very small $b(B)$), but it is at variance with the usual procedure in Bayes theory, belief function theory and similar approaches, where even tiny probabilities or beliefs are being renormalized (as long as they are not exactly zero).

If a model assumption or another basic assumption B is clearly shown by the data to be wrong, we have to change the model, rather than computing some fictive numbers which have no relation to reality. This, naturally, holds also for the Neyman-Pearson theory. As C. Daniel, a highly recognized applied statistician, once said: "We are told not to change the horses in the middle of the stream ...", but to continue along his line: If the old horses drowned already, we better use new ones. Cf. also [22].

2.4 More than two background assumptions

This subsection is an obvious generalization of 2.3. But if every B is not a whole belief system, but just a single parameter, Example 2 can be interpreted as a general inference method (related to minimum entropy methods).

Consider now a (finite or infinite) class of background beliefs or assumptions $B_1, B_2, ...$ with (prior) surprises, $b_1 = b(B_1), b_2 = b(B_2), ...$ In practice, we start looking only at the smallest b_i's; however, we have to be able to consider also larger b_i's, once an observation A is made, because now the (near) smallest $s(A|B_i) + b(B_i)$ will be of the greatest interest. And the B_i's with "very large" $s(A|B_i) + b(B_i)$ will be deemed "practically impossible".

Example 2: Let F be a measurable space with a set of parameters B_i and a collection of potential probabilities $P(A|B_i)$ for all measurable events A and the corresponding surprises $s(A|B_i) = -\log P(A|B_i)$. Let $b_i = b(B_i)$ the collection of apriori surprises if the B_i were declared to be true. The b_i may be a constant, or may be determined by a (subjective or objective) Bayesian apriori distribution, a likelihood, a belief function, or some other measure of the apriori

"plausibility" of the B_i. Given an observation A, the aposteriori surprise of $B_i|A$ is $s_i := s(A|B_i) + b(B_i)$. All "small" values of s_i are entirely plausible, and we may just for convenience pick out the minimum or some similar quantity (perhaps depending also on "neighboring" B_i's, doing some local "smoothing"). But all "too large" s_i are ruled out as "practically impossible" (until perhaps – rarely – a very surprising future observation A_2 forces us to either scrutinize A and A_2 more closely or to revolutionize the order of the B_i, digging out hypothetical models not yet considered in practice so far).

For a qualitative and semi-quantitative description of such a set-up, with many practical examples, cf. [28] and [29].

3 Additional aspects

3.1 Cautious surprises and successful bets

This subsection is for the readers who either know the two concepts mentioned, or who may want to study the pertaining literature, and want to see its relation to the present paper. (Obviously, there is no space here to repeat the old theories.) The last three paragraphs contain sketches of related possible future research problems.

In [21] a function m between 0 and 1 (a kind of upper probability describing our lack of surprise about some event A) and two definitions are introduced, namely "cautious surprises" and "successful bets". Now we can put $s = -\log m$, and the property of cautious surprises is nothing but the minimization of the average surprise mentioned above.

When we linearize the logarithm of m, we obtain a linear theory with close relationships to other statistical concepts, especially bets, and the concept of successful bets has been worked out to some extent especially in [23, 24]. A very special case are the (in)famous fiducial probabilities [21, 26, 27].

Somewhat related may be the linearization of approximately linear theories, such as Choquet capacities in a local neighborhood (described, e.g., by the gross-error model or the total-variation model).

Another aspect may be the robustification of the potential surprises, by putting an upper bound on them. The need for this may be only moderate, since $m \log m$ is bounded on the unit interval; but the two factors may not be always so closely related.

The approximate or exact requirements of cautious surprises or successful bets may also help in the robustification of the Bayes theory, as in the "weighted Bayes' theorem" [25, Ch. 5.3], in which basically random weights are treated like fixed weights.

3.2 Ted and Fred: different background information

We now leave Ted considered in isolation and discuss informally some situations where more than one person is involved.

An important practical problem is that two (or more) persons – say, Ted and Fred – may have different background knowledge or beliefs, while some common ground, resulting in common or at least similar surprises, should be achieved, otherwise no general opinion, including no general scientific theory, would be possible.

A first step is to openly discuss the different background opinions of Ted and Fred, until (hopefully) some common agreement can be found.

But a frequent obstacle is that many opinions, or even many reasons for such opinions, are not conscious for either Ted or Fred. They may be subconscious prejudices, which perhaps only by some kind of hard detective work can be elucidated, for example, by auxiliary information given by Ted or Fred, or by their family, educational, sociological or religious background.

Even if the basic reasons for such disagreements can be brought out into the open, it may be that on certain points no agreement is possible. Then Ted and Fred still can "agree to disagree". An example is the technical staff for water, electricity etc. in West and East Berlin during the height of the Cold War, who had to cooperate in the divided city, and they did so productively, agreeing on the political disagreements, but making sure the city would still function.

4 Examples and further aspects

The last chapter provides some more examples of the rich variety of real-life situations which can be described within the framework given. I don't know all the literature, but I am not aware of a theory which, for example, does describe the zigzag in the example of 4.1 in an adequate way. Perhaps it is because most theories are only deductive, while in real life we need (also) inductive thinking. The formal framework may still have to be worked out further and refined, as the example in 4.2 shows. On the other hand, the example in 4.3 should be a relatively easy one also for other theories, as it involves no change of background; moreover, I found it decidedly useful; it may and should already exist somewhere, in some form or other.

The last paragraph offers many opportunities for further work. But at any rate, this paper, together with the two previous ones, provides a broad conceptual framework (if one wishes, even a quantitative one, as shown specifically in this paper) for describing how we can deal with incomplete knowledge and how we can learn in real life.

4.1 Updating of the background information

Updating information clearly is an important operation, which can change potential surprises considerably.

Let us assume Ted is going to visit Fred by train fairly late in the evening. Fred expects to meet Ted at the closest major station, perhaps a few minutes late, but hardly more than half an hour late; any much longer delay would be a big surprise. But then Ted calls that he is stuck somewhere, because of a serious accident on the route, and has no information on how long the delay will be. It is now conceivable that he cannot even reach the last local train. Later, he cites the experience of a fellow traveller that with this type of accident, the delay is usually around 2 hours. This would mean still reaching the last local train. Eventually, after the train moves again, two official delays become available, which both are somewhat below 2 hours, but differ by 20 minutes. The true arrival time is in between.

The consecutive updating of the background information changes the potential surprises, first to much less "knowledge", then to a more realistic expectation (although, as so often in life, not all discrepancies are cleared up).

4.2 Unexpected surprises

As mentioned above, some potential surprises are so unlikely to Ted that he does not even think of them. Sometimes he would consider them more plausible if his background information were updated by some additional information. Let us consider an example with various forms of surprises.

A married couple want to celebrate their wedding anniversary, with the husband secretly organizing it. First, they arrive at a high-level hotel, a fairly big surprise for the wife, but feasible. Then they get their room which turns out to be a ("the") historical room: almost everything like a hundred years ago: a big unexpected surprise for both of them. An excursion by horse-drawn carriage was only a moderate surprise for the wife, since such carriages exist in the area. But an excursion by public boat on the nearby lake was an "impossible" surprise for her, since she knew that such boats didn't exist; she needed the updating of her knowledge that in very recent years public boat connections had been introduced. But then the husband leads his wife, well-dressed and at a fixed time, not to the ordinary hotel elevator, but to the remote staff elevator; they go down and get lost in the subterranean floors; he finds the way again, and they walk amidst the rooms of the staff and end up in a little chapel where a priest performs a small private ceremony for their wedding anniversary. – It seems hard to formalize such surprises and the lack of any knowledge on the way there.

4.3 Informative short knowledge descriptions

Let us close with an example from field ornithology. There are many books on where to find which birds, but some of them I find rather unsatisfactory, either being not sufficiently informative, or not agreeing with my experience (or being even misleading). However, I discovered one book which, to my own surprise, I found very useful [2]: in its bird lists (each for a larger area), the abundance of every species was coded by just 3 symbols: `c` (for common), no symbol, or `r` (for rare). (To be more precise, there are also symbols for the season (summer, winter, migration, or year around) and sometimes for the altitude or other informative features.) Why are just these 3 symbols for abundance so satisfying, according to my experience?

Clearly, `r` means rare: not impossible, but each observation would be a big (pleasant) surprise, unless one knows and visits the restricted areas (if existing) where the species is not so rare. But in general it would be no surprise at all not to find the species, even after a long search. – And `c` means common: the species would be no surprise at all, and with a decently long search in the right habitat (and perhaps time of day, weather, etc.), it would be a big surprise not to find it. – No symbol means neither `c` nor `r`; it would be neither a surprise to find the species, nor a surprise not to find it. The species may be sparsely distributed, or regional, or temporal (e.g., during irregular invasions); a more detailed description of the "probability of encounter" ("Antreff-Wahrscheinlichkeit", cf. [19, 20, 22] would be too complicated on limited space. But two out of three categories are very informative. – I think we can use this set-up much more generally to distinguish the things we are pretty sure to happen, the ones we are pretty sure *not* to happen, and the ones we just don't know.

The method can be easily generalized to situations with more than two alternatives. We can describe profiles of potential surprises – and conditional surprises,

given various backgrounds – in very complicated situations. Surprises imply assumed partial knowledge (that an event is not likely going to happen). Two very special cases are deterministic knowledge (all surprises infinite, except one being zero), and perfect knowledge of a probability space (the sum of the negative antilogarithms of all surprises of disjoint events being one), but obviously there are many more forms of incomplete knowledge.

Acknowledgments: I am grateful to W. Stahel for his help. – Besides the references given by the editors, two referees provided a thorough critical reading and a number of questions and suggestions which gave rise to a considerable enlargement of the original paper.

References

[1] C. F. Alchourron, P. Gardenfors, and D. Makinson. On the logic of theory change: partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530, 1985.

[2] P. Alden and J. Gooders. *Finding Birds Around the World.* Houghton Mifflin, Boston, 1981.

[3] A. Arnauld and P. Nicole. *La logique ou l'art de penser.* Paris, 1662. Reprinted 1965 by Fridrich Fromann Verlag, Stuttgart-Bad Cannstatt.

[4] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. Roy. Soc. London*, A 53:370–418, 1763. Reprinted in *Biometrika* **45** (1958), 293–315, and in Pearson, E. S., Kendall, M. G. (eds.)(1970): *Studies in the History of Statistics and Probability,* 131–153, Griffin, London.

[5] J. Bernoulli. *Ars Conjectandi.* Thurnisiores, Basel, 1713. Reprinted in *Werke von Jakob Bernoulli, volume 1,* Birkhäuser Verlag, Basel, 1975.

[6] G. Boole. *An Investigation of the Laws of Thought, on which are founded the Mathematical Theories of Logic and Probabilities.* Macmillan, London, (1854) 1958. Reprinted by Dover, New York (1958).

[7] D. Brönnimann. Die Entwicklung des Wahrscheinlichkeitsbegriffs von 1654 bis 1718. Diplomarbeit, Seminar für Statistik, Swiss Federal Institute of Technology (ETH) Zurich, 2001. URL: `ftp://stat.ethz.ch/Masters-Theses/Daniel_Broennimann-Wahrschkeit.pdf`.

[8] M. E. G. V. Cattaneo. Likelihood-based statistical decisions. In F. G. Cozman, R. Nau, and T. Seidenfeld, editors, *Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*, 2005.

[9] M. E. G. V. Cattaneo. *Statistical Decisions Based Directly on the Likelihood Function.* PhD thesis, ETH Zurich, 2007.

[10] A. Cournot. *Exposition de la théorie des chances et des probabilités.* Librairie philosophique J. Vrin, Paris, (1843) 1984.

[11] B. de Finetti. La prevision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré*, 7:1–68, 1937. English translation in Kyburg, H. E., and Smokler, H. E. (eds.) (1964) *Studies in Subjective Probability.* Wiley, New York. (2nd, enlarged ed. 1980).

[12] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.*, 38:325–339, 1967.

[13] A. P. Dempster. A generalization of Bayesian inference. *J. Roy. Statist. Soc.*, B 30:205–245, 1968.

[14] D. Dubois and H. Prade. *Theory of Possibility.* Plenum, London, UK., 1988. Original Edition in French (1985) Masson, Paris.

[15] R. A. Fisher. Inverse probability. *Proc. of the Cambridge Philosophical Society*, 26:528–535, 1930. Reprinted in Collected Papers of R. A. Fisher, ed. J. H. Bennett, Volume 2, 428–436, University of Adelaide 1972.

[16] I. J. Good. The probabilistic explication of information, evidence, surprise, causality, explanation, and utility. In V. P. Godambe and D. A. Sprott, editors, *Foundations of Statistical Inference*, pages 108–141. Holt, Rinehart, and Winston, Toronto, 1971. Reprinted partly in Good (1983).

[17] I. J. Good. *Good Thinking; The Foundations of Probability and Its Applications.* University of Minnesota Press, Minneapolis, 1983.

[18] I. Hacking. *The Emergence of Probability. A philosophical study of early ideas about probability, induction and statistical inference.* Cambridge University Press, Cambridge, 1975.

[19] F. Hampel. Artenliste vom Seeburger See 1955-1964 (unter knapper Berücksichtigung des Raumes um Göttingen). mimeographed manuscript, Göttingen, 23 pp., later reprinted, 1965.

[20] F. Hampel. Überwinterung und Verhaltensweisen der Beutelmeise (Remiz pendulinus) am Seeburger See. *J. für Ornithologie*, 107 (Vol. 3/4):359–360, 1966.

[21] F. Hampel. Some thoughts about the foundations of statistics. In S. Morgenthaler, E. Ronchetti, and W. A. Stahel, editors, *New Directions in Statistical Data Analysis and Robustness*, pages 125–137. Birkhäuser Verlag, Basel, 1993.

[22] F. Hampel. Is statistics too difficult? *Canad. J. Statist.*, 26(3):497–513, 1998.

[23] F. Hampel. On the foundations of statistics: A frequentist approach. In M. S. de Miranda and I. Pereira, editors, *Estatística: a diversidade na unidade*, pages 77–97. Edições Salamandra, Lda., Lisboa, Portugal, 1998. URL: `ftp://ftp.stat.math.ethz.ch/Research-Reports/85.pdf`.

[24] F. Hampel. An outline of a unifying statistical theory. In G. de Cooman, T. L. Fine, and T. Seidenfeld, editors, *Proc. of the 2nd Internat. Symp. on Imprecise Probabilities and their Applications, ISIPTA'01, Cornell University, 26-29 June 2001*, pages 205–212. Shaker Publishing Maastricht, 2000, 2001. URL: `ftp://ftp.stat.math.ethz.ch/Research-Reports/95.pdf`.

[25] F. Hampel. Some thoughts about classification. In K. Jajuga, A. Sokołowski, and H.-H. Bock, editors, *Classification, Clustering, and Data Analysis. Recent Advances and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, July 16–19, 2002, Cracow, Poland, pages 5–26. Invited keynote lecture, 8th Conference of the International Federation of Classification Societies, Springer, Berlin, 2002. URL: `ftp://ftp.stat.math.ethz.ch/Research-Reports/102.pdf`.

[26] F. Hampel. The proper fiducial argument. Extended abstract. *Electronic Notes in Discrete Mathematics*, 21:297–300, 2005.

[27] F. Hampel. The proper fiducial argument. *Information Transfer and Combinatorics*, LNCS 4123:512–526, 2006. URL: `ftp://ftp.stat.math.ethz.ch/Research-Reports/114.pdf`.

[28] F. Hampel. Upper and lower probabilities in real life. In *CD-ROM containing the Proc. 56th Session of the ISI, Contrib. Papers, Lisboa, Portugal*, 2007. URL: `ftp://ftp.stat.math.ethz.ch/Research-Reports/145.pdf`.

[29] F. Hampel. How can we get new knowledge? In T. Augustin, F. P. A. Coolen, S. Moral, and M. C. M. Troffaes, editors, *Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications*, Durham, U.K., 2009. Durham University, Department of Mathematical Sciences. URL: `http://www.sipta.org/isipta09/proceedings/037.html`.

[30] F. Hampel. Nonadditive probabilities in statistics. *Journal of Statistical Theory and Practice*, 3(1):11–23, 2009. URL: `ftp://ftp.stat.math.ethz.ch/Research-Reports/146.pdf`.

[31] E. L. Lehmann. *Testing Statistical Hypotheses*. Wiley, New York, 1959. (2nd ed. 1986).

[32] A. Neumaier. Fuzzy modeling in terms of surprise. *Fuzzy Sets and Systems*, 135:21–38, 2003.

[33] B. Pascal. *Oeuvres complètes*, volume 34. Bibliothèques de la Pléiade, Paris, 1954. Edited by Jacques Chevalier, see also many more editions.

[34] E. S. Pearson and J. Wishart, editors. *"Student's" Collected Papers*. Cambridge University Press, 1958.

[35] K. Pearson. *The History of Statistics in the 17th & 18th Centuries; against the changing background of intellectual, scientific and religious thought*. Griffin, London, 1978. Lectures by K. Pearson at University College London 1921-1933, edited by E.S. Pearson.

[36] G. L. S. Shackle. *Decision – Order and Time in Human Affairs*. Cambridge University Press, Cambridge, (1961) 1969.

[37] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J., 1976.

[38] "Student". The probable error of a mean. *Biometrika*, 6:1–25, 1908.

[39] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[40] L. A. Zadeh. Fuzzy sets. *Inform. Control*, 8:338–353, 1965.

Dynamic Programming and Subtree Perfectness for Deterministic Discrete-Time Systems with Uncertain Rewards

Nathan Huntley
Durham University, UK
nathan.huntley@durham.ac.uk

Matthias C. M. Troffaes
Durham University, UK
matthias.troffaes@gmail.com

Abstract

We generalise de Cooman and Troffaes's sufficient condition for dynamic programming to work for deterministic discrete-time systems. To do so, we use the general framework developed by Huntley and Troffaes, for decision trees with arbitrary rewards and arbitrary choice functions. Whence, we allow deterministic discrete-time systems with arbitrary rewards and an arbitrary composition operator on rewards. We show that the principle of optimality reduces to two much simpler conditions on the choice function. We establish necessary and sufficient conditions on choice functions for deterministic discrete-time systems to be solvable by backward induction, that is, for dynamic programming to work. Finally, we also discuss subtree perfectness—which is a stronger form of dynamic consistency—for these systems, and show that, in general, decision criteria from imprecise probability theory violate it, even though dynamic programming may work.

Keywords. Optimal control, dynamic programming, deterministic discrete-time systems, backward induction, subtree perfectness, choice function

1 Introduction

In this paper we formalize and extend the results of de Cooman and Troffaes [4] for deterministic discrete-time systems with uncertain gains. Such systems are typical in *control theory* (see for instance [2, 9]), which more generally covers the behaviour and control of dynamic systems. The particular class of systems we investigate is best illustrated by example: Fig. 1 depicts a system that starts at N_1, and can reach N_4 by multiple paths. The subject, who controls the system, can choose the path the system will take. Travelling down a particular arc gives the subject an associated reward. For instance, choosing the arc from N_1 to N_2 will give the subject X. The subject's task is to find an optimal path for the system to take.

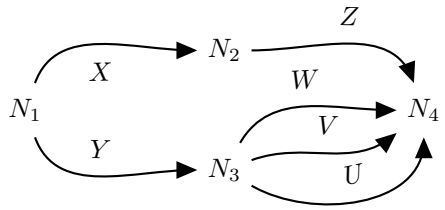


Figure 1: A simple deterministic system.

This is an example of a *deterministic discrete-time system*. If all rewards $U, \ldots, Z$ are certain, so the subject knows exactly what she will receive when choosing a particular route, then this is a system with *certain gains*. Such systems are easily solved: find a path with the highest total reward. We instead consider systems with uncertain gains, so $U, \ldots, Z$ give rewards determined by the as yet unknown state of nature. Such uncertain gains are called *gambles*. The overall reward for a particular path is then determined by the sum of the gambles for all arcs in the path.

This paper deals with *normal form* decision making. In general, normal form decisions involve the subject specifying her decisions in all eventualities, and then acting upon this specification. For deterministic discrete-time systems with certain rewards, a normal form decision is simply a path through the system. In contrast, the *extensive form* involves making decisions only when the relevant decision point is reached, and is expressed differently. We do not investigate the extensive form in this paper, but caution that the two forms do not always lead to the same answer.

With uncertain rewards, there are two possible ways the system can evolve. If the subject receives the reward from a gamble as soon as that arc is chosen, then she can use this information to choose her next arc. For example, an informal strategy for Fig. 1 could be "choose Y, and then choose W if Y has given a large reward, but choose V otherwise". Alternatively, the

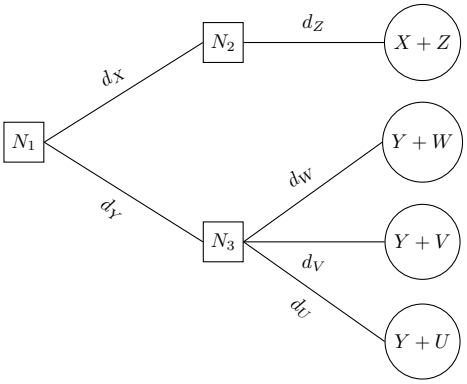


Figure 2: The decision tree for Fig. 1.

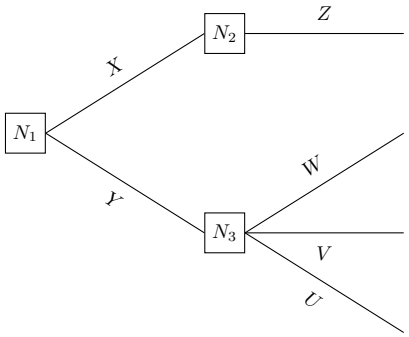


Figure 3: The deterministic system tree for Fig. 1.

subject may only learn about her actual rewards at the end of the process, and so could have no strategy more complicated than, say "choose Y, then W", because she does not learn of the outcome of Y until later. The latter set-up, where the true state of nature is only revealed at the end of the process, is followed by de Cooman and Troffaes, and so we follow it too.

Normal form decisions are thus very simple (indeed, exactly the same as for certain rewards), and no concept of conditioning is required. Also, since in this case everything is completely deterministic until the final decision has been made, it seems natural to use the normal form. Note that the concept of normal form decision making can be criticized [14], however we do not aim to address these issues in this paper.

We aim to apply known results by Huntley and Troffaes [6] on backward induction and subtree perfectness for decision trees to these deterministic discrete-time systems, thereby generalizing the work of de Cooman and Troffaes [4]. Whence, as a first step, we represent these systems as decision trees [8, 7, 3]. An example is given in Fig. 2. In such a tree, square nodes, called decision nodes, represent points at which the subject must choose an arc. The circular nodes, called chance nodes represent points at which the consequence is determined by the state of nature. For completeness, Fig. 3 ought to have arcs leading from the chance nodes to terminal *reward* nodes, representing the rewards given by the gambles for particular states of nature. Since we have not explicitly defined the gambles, this final layer of nodes has been omitted.

This representation can be simplified to a form of decision tree more suited for the special structure of the problem at hand. In this representation, which we call a *deterministic system tree*, there are only two types of nodes: decision nodes and terminal nodes. All branches end with a terminal node, and all terminal nodes appear at the end of branches. Every arc corresponds to a decision, and each arc has an associated gamble. The deterministic system tree for Fig. 1 is shown in Fig. 3. It must be emphasised that, although gambles are acquired upon choosing a decision arc, their value is not discovered until the terminal node is reached. Therefore there is no learning or conditioning involved in this model.

This tree is clearly much more similar to the description of the system. Indeed, normal form decisions for deterministic system trees are again just paths through the tree. How do we find the optimal paths? Following [4], we will use *choice functions on gambles*. Such choice function returns, for every set of gambles, a subset of gambles which are deemed optimal in some sense (which depends on your choice of choice function). For example, maximizing expected utility is one such choice function, but many more exist.

Now, as we saw, each path through the tree has a corresponding gamble. Whence, given a choice function, we can say that a path is optimal whenever its gamble is optimal in the set of gambles induced by all paths. Effectively, we end up with a set of optimal paths.

Two questions arise from this form of solution. The first, addressed by de Cooman and Troffaes, is whether backward induction (more commonly called *dynamic programming* in this field, following Bellman [2]) can be used to reach the normal form solution for a given choice function.

The idea of backward induction is simple. We informally illustrate it on Fig. 3. First, we find which of W, V, and U are optimal. Suppose this is $\{V, W\}$. Then, we determine which of $X+Z$, $Y+W$, and $Y+V$ is optimal. We end up with the backward induction solution, say for instance $\{X + Z, Y + V\}$.

Backward induction thus returns a set of paths, but for many choice functions it can give a different set of paths from the standard normal form solution. In other words, applying the choice function recursively

stage-by-stage may not give the same result as applying the choice function on all gambles at once. De Cooman and Troffaes [4] show that backward induction works if the choice function satisfy Bellman's principle of optimality [2] and another property, insensitivity to the omission of non-optimal elements. This paper contributes a reformulation of these results into a theorem about trees and paths in the same fashion as our method for decision trees [6], a proof of necessity as well as sufficiency, and a decomposition of the principle of optimality into two more basic properties.

The second question is whether the normal form solution is equivalent to the combination of local solutions. For instance, in Fig. 3, if W and V are both optimal at N_3, then both $Y + W$ and $Y + V$ should be optimal at N_1, or neither should—this was violated in our earlier example demonstrating backward induction.

This property has been studied extensively for problems modelled by standard decision trees (see for instance [5, 10, 11, 6]). We call a solution with such a property *subtree perfect* (following Selten's analogous concept of subgame perfectness [15]). We show that subtree perfectness for deterministic system trees corresponds to a stricter version of Bellman's principle of optimality obtained by strengthening set inclusions of all properties involved to equalities.

The paper is structured as follow. Section 2 introduces necessary notation. Section 3 presents the results on dynamic programming. Section 4 presents the results on subtree perfectness. Section 5 provides a brief summary of the consequences of the results for the theory of coherent lower previsions. Section 6 concludes the paper.

2 Definitions and Notation

Let Ω be a *possibility space*, i.e. the set of all possible states of nature. Elements ω of Ω are called outcomes. Let $\mathcal{R}$ be a set of rewards (results the subject can receive; they do not have to be desirable rewards). We assume a binary operator $+$ on $\mathcal{R}$, which we call *addition*.[1] We assume that $\mathcal{R}$ has a left identity element 0, so $0 + r = r$ for all $r \in \mathcal{R}$. We also assume that $r_1 + r_2 = r_1 + r_3$ implies $r_2 = r_3$; this holds for instance if every reward $r \in \mathcal{R}$ has a left inverse $-r \in \mathcal{R}$, so $(-r) + r$ equals the left identity element 0. No other assumptions about $\mathcal{R}$ are required.

A *gamble* is a function $X \colon \Omega \to \mathcal{R}$, with the interpretation that, should ω be the true state of nature, the gamble X gives the subject the reward $X(\omega)$. Addition of gambles is defined in the obvious way: $(X + Y)(\omega) = X(\omega) + Y(\omega)$.

Given a set of gambles $\mathcal{X}$ (in this paper, all sets are assumed to be finite, and non-empty unless otherwise noted) from which our subject must pick one, how should she decide? Ideally, she would like to select a single optimal gamble for every set $\mathcal{X}$, but this may not always be possible, for instance, because she lacks information about ω, or because she has no precise utility over her rewards. She might, at least, be able to specify a (possibly empty) set of gambles in $\mathcal{X}$ she considers unacceptable. Any gamble not so judged remains a plausible candidate, and these could be reported as an optimal set. This procedure is represented by a *choice function*: a function that maps sets of options to non-empty subsets.

Definition 1. *A choice function on gambles,* opt*, is a function that maps each set $\mathcal{X}$ of gambles to a non-empty subset of that set:*

$$\emptyset \neq \mathrm{opt}(\mathcal{X}) \subseteq \mathcal{X}.$$

How do we use the concepts of gambles and choice functions to solve deterministic system trees? First, we introduce the concept of normal form decisions, solutions, and operators.

Definition 2. *A normal form decision of a deterministic system tree T is a path through T.*

Definition 3. *The set of all normal form decisions for a deterministic system tree T is denoted by* $\mathrm{nfd}(T)$.

Definition 4. *A normal form solution of a deterministic system tree T is a non-empty subset of* $\mathrm{nfd}(T)$.

The interpretation of a normal form solution is that the subject may pick any path in this subset and follow it.

Definition 5. *A normal form operator* norm *is a function that maps each deterministic system tree T to a normal form solution of T:*

$$\emptyset \neq \mathrm{norm}(T) \subseteq \mathrm{nfd}(T).$$

Using these definitions, we can define the set of all gambles associated with a deterministic system tree. Recall that any path through a tree has its own gamble, so given the set of all normal form decisions we can find the set of all gambles for that tree.

Definition 6. *The function* gamb *maps deterministic system trees to their set of associated gambles (called* normal form gambles*):*

$$\mathrm{gamb}(T) = \bigcup_{U \in \mathrm{nfd}(T)} \mathrm{gamb}(U).$$

[1] If $\mathcal{R} = \mathbb{R}$, then the operator $+$ does not need to have any resemblance with the usual addition of real numbers, although it is a convenient and popular choice.

This gives us a set of gambles to which to apply the choice function opt. The procedure is as follows: find the set of normal form gambles, apply the choice function to find an optimal subset of normal form gambles, and then list all normal form decisions with gambles in this optimal subset. This defines a normal form operator, norm_{opt}.

Definition 7. *For a choice function* opt*, the* normal form operator induced by opt *is defined for any deterministic system tree T by*

$$\text{norm}_{\text{opt}}(T) = \{U \in \text{nfd}(T) : \text{gamb}(U) \subseteq \text{opt}(\text{gamb}(T))\}.$$

Of course, since U is always a normal form decision, $\text{gamb}(U)$ is always a singleton in this definition. In particular, the following equality holds:

$$\text{gamb}(\text{norm}_{\text{opt}}(T)) = \text{opt}(\text{gamb}(T)).$$

In the above equation, we have used the following notation: for any set of deterministic system trees $\mathcal{T}$,

$$\text{gamb}(\mathcal{T}) = \bigcup_{T \in \mathcal{T}} \text{gamb}(T).$$

To express backward induction in terms of trees, and to help with many proofs, we introduce a notation for representing a deterministic system tree as a combination of smaller deterministic system trees. For any trees $T_1, \ldots, T_n$, we can join them at a decision node, with the arc from this decision node to T_i corresponding to a gamble X_i, and write this as

$$\bigsqcup_{i=1}^{n} X_i T_i.$$

Sometimes we need to work with all the possible ways to join sets of trees $\mathcal{T}_1, \ldots, \mathcal{T}_n$ in a similar way. This is written as

$$\bigsqcup_{i=1}^{n} X_i \mathcal{T}_i = \left\{ \bigsqcup_{i=1}^{n} X_i T_i : T_i \in \mathcal{T}_i \right\}.$$

This allows gamb to be defined recursively:

$$\text{gamb}\left(\bigsqcup_{i=1}^{n} X_i T_i \right) = \bigcup_{i=1}^{n} (X_i + \text{gamb}(T_i)),$$

where we use the notation

$$X + \mathcal{Y} = \{X + Y : Y \in \mathcal{Y}\}.$$

Similarly,

$$\text{gamb}\left(\bigsqcup_{i=1}^{n} X_i \mathcal{T}_i \right) = \bigcup_{i=1}^{n} (X_i + \text{gamb}(\mathcal{T}_i)).$$

Finally, we sometimes need to restrict deterministic system trees to particular subtrees, obtained by removing everything before a certain node.

Definition 8. *A subtree of a deterministic system tree T obtained by removal of all non-descendants of a particular node N, but retaining N, is called the* subtree of T at N *and is denoted by* $\text{st}_N(T)$*.*

This extends to sets of trees in the usual way:

$$\text{st}_N(\mathcal{T}) = \{\text{st}_N(T) : T \in \mathcal{T} \text{ and } N \text{ in } T\}.$$

Usually, the subtrees we need to use are those whose roots are immediate successors of T. Therefore we define $\text{ch}(T)$ to be the set of immediate successors (i.e. children) of the root node of T.

3 Backward Induction Theorem

We introduce a new normal form operator based on backward induction, defined recursively. The operator works by eliminating non-optimal paths in subtrees, then bringing all optimal paths to the next largest subtree, and so on until the root node is reached. To do so elegantly, we extend norm_{opt} to act upon sets of trees:

$$\text{norm}_{\text{opt}}(\mathcal{T}) = \{U \in \text{nfd}(\mathcal{T}) : \text{gamb}(U) \subseteq \text{opt}(\text{gamb}(\mathcal{T}))\}.$$

Definition 9. *The normal form operator* back_{opt} *is defined for any deterministic system tree T that consists only of a terminal node by*

$$\text{back}_{\text{opt}}(T) = T$$

and for any other deterministic system tree $T = \bigsqcup_{i=1}^{n} X_i T_i$ by

$$\text{back}_{\text{opt}}(T) = \text{norm}_{\text{opt}}\left(\bigsqcup_{i=1}^{n} X_i \, \text{back}_{\text{opt}}(T_i) \right).$$

We are interested in determining when back_{opt} and norm_{opt} coincide. This happens if and only if the following two properties hold.

Property 1 (Insensitivity of optimality to the omission of non-optimal elements)**.** *For any sets of gambles $\mathcal{X}$ and $\mathcal{Y}$,*

$$\text{opt}(\mathcal{X}) \subseteq \mathcal{Y} \subseteq \mathcal{X} \Rightarrow \text{opt}(\mathcal{Y}) = \text{opt}(\mathcal{X}).$$

De Cooman and Troffaes [4] explain that this property is crucial for backward induction to work. It also appears in the work of Sen [16], who shows it to be

one "half" of the property of *path independence* (see for instance Plott [12]).

The second property was introduced by Bellman [2] with the following explanation:

> An optimal policy has the property that, whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

Note that, in the context of deterministic system trees, states are simply decision nodes.

Although Bellman states the principle in terms of the first decision only, it implies that the restriction of an optimal policy to any subtree must be optimal. We formalize the principle into the following property.

Property 2 (Principle of Optimality)**.** *A normal form operator* norm *satisfies the principle of optimality if, for any deterministic system tree T, and any node N in at least one element of* $\operatorname{norm}(T)$*,*

$$\operatorname{st}_N(\operatorname{norm}(T)) \subseteq \operatorname{norm}(\operatorname{st}_N(T)).$$

Equivalently, for any normal form decision $U \in \operatorname{norm}(T)$ and any node N in U,

$$\operatorname{st}_N(U) \in \operatorname{norm}(\operatorname{st}_N(T)).$$

For the particular case of $\operatorname{norm}_{\text{opt}}$, the above definition is easily seen to be equivalent to the inclusion formula of de Cooman and Troffaes [4, Definition 13]—but our notation is far more efficient at expressing it.

Interestingly, we can decompose Property 2, the principle of optimality, into two far more basic properties.

Property 3 (Preservation of non-optimality under the addition of elements)**.** *For any sets of gambles $\mathcal{X}$ and $\mathcal{Y}$,*

$$\mathcal{Y} \subseteq \mathcal{X} \Rightarrow \operatorname{opt}(\mathcal{Y}) \supseteq \operatorname{opt}(\mathcal{X}) \cap \mathcal{Y}.$$

This is a type of independence of irrelevant alternatives (see [1, 13]), called property α by Sen [16]. It is the other "half" of path independence (so we show that path independence is necessary for dynamic programming). Property 3 is not explicitly invoked by de Cooman and Troffaes, but it is used in a proof for a particular choice function [4, Proposition 16].

Property 4 (Backward Addition Property)**.** *For any gamble X and any non-empty finite set of gambles $\mathcal{Y}$,*

$$\operatorname{opt}(X + \mathcal{Y}) \subseteq X + \operatorname{opt}(\mathcal{Y}).$$

This property was informally foreseen by de Cooman and Troffaes (see the discussion of "additivity" [4, §3.4]). It is similar to properties relating to backward induction for other decision processes [6, 17].

The proof of equivalence relies on the next lemma.

Lemma 10. *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 $\operatorname{norm}(T)$*,*

$$\operatorname{st}_K(\operatorname{norm}(T)) \subseteq \operatorname{norm}(\operatorname{st}_K(T)),$$

(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 $\operatorname{norm}(\operatorname{st}_K(T))$*,*

$$\operatorname{st}_L(\operatorname{norm}(\operatorname{st}_K(T))) \subseteq \operatorname{norm}(\operatorname{st}_L(\operatorname{st}_K(T))),$$

then, for all nodes N in T such that N is in at least one element of $\operatorname{norm}(T)$*,*

$$\operatorname{st}_N(\operatorname{norm}(T)) \subseteq \operatorname{norm}(\operatorname{st}_N(T)).$$

Proof. If N is the root of T, then the result is immediate. If $N \in \operatorname{ch}(T)$, then the result follows from (i). Otherwise, N must be in $\operatorname{st}_K(T)$ for one $K \in \operatorname{ch}(T)$.

By assumption, there is a $U \in \operatorname{norm}(T)$ that contains N (and of course also K). Therefore, $U \in \operatorname{st}_K(\operatorname{norm}(T))$, and by (i), $\operatorname{st}_K(U) \in \operatorname{norm}(\operatorname{st}_K(T))$, and so N is also in at least one element of $\operatorname{norm}(\operatorname{st}_K(T))$.

We use the fact that, if $\mathcal{U}$ and $\mathcal{V}$ are sets of normal form decisions such that $\mathcal{U} \subseteq \mathcal{V}$, then for any node N, $\operatorname{st}_N(\mathcal{U}) \subseteq \operatorname{st}_N(\mathcal{V})$. Combining everything, by (i),

$$\operatorname{st}_N(\operatorname{st}_K(\operatorname{norm}(T))) \subseteq \operatorname{st}_N(\operatorname{norm}(\operatorname{st}_K(T)))$$

hence, since N is in at least one element of $\operatorname{norm}(\operatorname{st}_K(T))$, by (ii) we have

$$\subseteq \operatorname{norm}(\operatorname{st}_N(\operatorname{st}_K(T))),$$

whence the desired result follows, since $\operatorname{st}_N(\operatorname{st}_K(T)) = \operatorname{st}_N(T)$. □

Theorem 11. $\operatorname{norm}_{\text{opt}}$ *satisfies Property 2 if and only if* opt *satisfies Properties 3 and 4.*

Proof. "only if". Let X be a gamble and $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ be a set of gambles. Consider the upper tree in Fig. 4. If $X + Y_k \in \operatorname{opt}(X + \mathcal{Y})$, then by Property 2 it follows that $Y \in \operatorname{opt}(\mathcal{Y})$, hence Property 4 holds. Next, consider the lower tree. Let $\mathcal{Y} = \{Y_1 \ldots, Y_m\}$, $\mathcal{Z} = \{Z_1, \ldots, Z_n\}$ and suppose

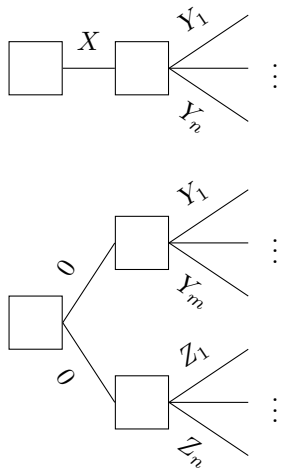


Figure 4: Decision trees for Theorem 11.

$\mathcal{Y} \cap \mathcal{Z} = \emptyset$. Now let $\mathcal{X} = \mathcal{Y} \cup \mathcal{Z}$. By Property 2 we know that if $Y \in \mathcal{Y} \cap \text{opt}(\mathcal{X})$, then $Y \in \text{opt}(\mathcal{Y})$, hence Property 3 holds.

"if". We proceed by structural induction. Let T be a deterministic system tree. The base step, to show the result when T consists of a terminal node only, is trivial. The inductive step is to suppose that Property 2 holds for every $\text{st}_K(T)$ where $K \in \text{ch}(T)$, and then show that Property 2 holds for T. By Lemma 10, we need only show that for every $K \in \text{ch}(T)$ that is in at least one element of $\text{norm}_{\text{opt}}(T)$,

$$\text{st}_K(\text{norm}_{\text{opt}}(T)) \subseteq \text{norm}_{\text{opt}}(\text{st}_K(T)).$$

So, the proof is established if we can show that, for every $U \in \text{norm}_{\text{opt}}(T)$ passing through $K \in \text{ch}(T)$,

$$\text{st}_K(U) \in \text{norm}_{\text{opt}}(\text{st}_K(T)). \tag{1}$$

We now express this in terms of gambles—but first we introduce some notation.

Let $\text{ch}(T) = \{K_1, \ldots, K_n\}$, and $K = K_k$. Let $\text{gamb}(\text{st}_{K_i}(T)) = \mathcal{Y}_i$, and let X_i be the gamble corresponding to the arc to K_i. That is,

$$T = \bigsqcup_{i=1}^{n} X_i \, \text{st}_{K_i}(T).$$

Recall, U contains the node K_k, so $\text{gamb}(U) = X_k + Y_k$ for some $Y_k \in \mathcal{Y}_k$.

Now, because $U \in \text{norm}_{\text{opt}}(T)$, we know that

$$X_k + Y_k \in \text{opt}(\text{gamb}(T)) = \text{opt}\left(\bigcup_{i=1}^{n}(X_i + \mathcal{Y}_i)\right). \tag{2}$$

To establish Eq. (1), we must simply show that $Y_k \in \text{opt}(\mathcal{Y}_k)$.

Indeed. Obviously,

$$X_k + \mathcal{Y}_k \subseteq \bigcup_{i=1}^{n}(X_i + \mathcal{Y}_i).$$

Applying Property 3,

$$\text{opt}(X_k + \mathcal{Y}_k) \supseteq \text{opt}\left(\bigcup_{i=1}^{n}(X_i + \mathcal{Y}_i)\right) \cap (X_k + \mathcal{Y}_k).$$

However, by Eq. (2), $X_k + Y_k$ belongs to the right hand side, whence, it must also belong to the left hand side. Now, apply Property 4, to see that indeed $Y_k \in \text{opt}(\mathcal{Y}_k)$. This completes the inductive step. □

We are now in a position to prove a backward induction theorem. It turns out that we can incorporate another simple concept into this theorem, namely that of *strategic equivalence*. Two trees are strategically equivalent if their set of gambles is the same. We can show easily that back_{opt} and norm_{opt} agreeing is equivalent to back_{opt} preserving strategic equivalence.

Theorem 12. *Let* opt *be any choice function. The following conditions are equivalent.*

(A) For any deterministic system tree T, it holds that $\text{back}_{\text{opt}}(T) = \text{norm}_{\text{opt}}(T)$.

(B) For any strategically equivalent deterministic system trees, T_1 and T_2, it holds that

$$\text{gamb}(\text{back}_{\text{opt}}(T_1)) = \text{gamb}(\text{back}_{\text{opt}}(T_2)).$$

(C) opt *satisfies Properties 1 and 2.*

Lemma 13. *If, for all strategically equivalent deterministic system trees T_1 and T_2, it holds that*

$$\text{gamb}(\text{back}_{\text{opt}}(T_1)) = \text{gamb}(\text{back}_{\text{opt}}(T_2)),$$

then opt *satisfies Property 1.*

Proof. Let $\mathcal{X}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ be sets of gambles such that $\text{opt}(\mathcal{X}) \subseteq \mathcal{Y} \subseteq \mathcal{X}$. Let T_1 be a deterministic system tree with just one decision node and $\text{gamb}(T_1) = \mathcal{X}$. Let T_2 be a deterministic system tree constructed as follows: there is one decision arc with gamble 0 that leads to T_1, and n other decision arcs, each leading immediately to a terminal node, with gambles Y_1 to Y_n. Clearly, $\text{gamb}(T_2) = \mathcal{X}$. We have

$$\text{gamb}(\text{back}_{\text{opt}}(T_2)) = \text{opt}(\text{opt}(\mathcal{X}) \cup \mathcal{Y}) = \text{opt}(\mathcal{Y}).$$

because $\text{opt}(\mathcal{X}) \subseteq \mathcal{Y}$. Since back_{opt} is assumed to preserve strategic equivalence, and T_1 and T_2 are strategically equivalent by construction, it follows that $\text{opt}(\mathcal{Y}) = \text{opt}(\mathcal{X})$, as required. □

Lemma 14. *If, for all strategically equivalent deterministic system trees T_1 and T_2, it holds that*

$$\mathrm{gamb}(\mathrm{back}_{\mathrm{opt}}(T_1)) = \mathrm{gamb}(\mathrm{back}_{\mathrm{opt}}(T_2)),$$

then $\mathrm{norm}_{\mathrm{opt}}$ *satisfies Property 2.*

Proof. We show that opt must satisfy Properties 3 and 4 and invoke Theorem 11. We can again use the two trees from Fig. 4. Let the upper tree be called T_1, and let T_2 be a tree with only one decision node and $\mathrm{gamb}(T_2) = X + \mathcal{Y}$. Then,

$$\begin{aligned}\mathrm{opt}(X+\mathcal{Y}) &= \mathrm{gamb}(\mathrm{back}_{\mathrm{opt}}(T_2)) \\ &= \mathrm{gamb}(\mathrm{back}_{\mathrm{opt}}(T_1)) \\ &= \mathrm{opt}(X + \mathrm{opt}(\mathcal{Y})) \subseteq X + \mathrm{opt}(\mathcal{Y}),\end{aligned}$$

so Property 4 holds.

Let T_1 be the lower tree in Fig. 4, with $\{\mathcal{Y}, \mathcal{Z}\}$ a partition of $\mathcal{X}$. Let T_2 have one decision node and $\mathrm{gamb}(T_2) = \mathcal{X}$. As assumed, $\mathrm{gamb}(\mathrm{back}_{\mathrm{opt}}(T_1)) = \mathrm{opt}(\mathrm{opt}(\mathcal{Y}) \cup \mathrm{opt}(\mathcal{Z})) = \mathrm{opt}(\mathcal{X})$. So,

$$\begin{aligned}\mathrm{opt}(\mathcal{X}) \cap \mathcal{Y} &= \mathrm{opt}(\mathrm{opt}(\mathcal{Y}) \cup \mathrm{opt}(\mathcal{Z})) \cap \mathcal{Y} \\ &\subseteq (\mathrm{opt}(\mathcal{Y}) \cup \mathrm{opt}(\mathcal{Z})) \cap \mathcal{Y} \\ &= \mathrm{opt}(\mathcal{Y}) \cap \mathcal{Y} = \mathrm{opt}(\mathcal{Y}),\end{aligned}$$

so Property 3 holds. □

Lemma 15. *If $\mathcal{T} \subseteq \mathcal{U} \subseteq \mathcal{V}$ are sets of deterministic system trees,* opt *satisfies Property 1, and* $\mathrm{norm}_{\mathrm{opt}}(\mathcal{T}) = \mathrm{norm}_{\mathrm{opt}}(\mathcal{V})$*, then* $\mathrm{norm}_{\mathrm{opt}}(\mathcal{U}) = \mathrm{norm}_{\mathrm{opt}}(\mathcal{V})$.

Proof. By assumption, we have that

$$\begin{aligned}\mathrm{opt}(\mathrm{gamb}(\mathcal{V})) &= \mathrm{opt}(\mathrm{gamb}(\mathcal{T})) \subseteq \mathrm{gamb}(\mathcal{T}) \\ &\subseteq \mathrm{gamb}(\mathcal{U}) \subseteq \mathrm{gamb}(\mathcal{V}).\end{aligned}$$

Hence, by Property 1,

$$\mathrm{opt}(\mathrm{gamb}(\mathcal{T})) = \mathrm{opt}(\mathrm{gamb}(\mathcal{U})) = \mathrm{opt}(\mathrm{gamb}(\mathcal{V})).$$

So,

$$\begin{aligned}\mathrm{norm}_{\mathrm{opt}}(\mathcal{U}) &= \{U \in \mathcal{U} \colon \mathrm{gamb}(U) \subseteq \mathrm{opt}(\mathrm{gamb}(\mathcal{T}))\} \\ &\supseteq \{U \in \mathcal{T} \colon \mathrm{gamb}(U) \subseteq \mathrm{opt}(\mathrm{gamb}(\mathcal{T}))\} \\ &= \mathrm{norm}_{\mathrm{opt}}(\mathcal{T})\end{aligned}$$

because $\mathcal{U} \supseteq \mathcal{T}$, and

$$\begin{aligned}\mathrm{norm}_{\mathrm{opt}}(\mathcal{U}) &= \{U \in \mathcal{U} \colon \mathrm{gamb}(U) \subseteq \mathrm{opt}(\mathrm{gamb}(\mathcal{V}))\} \\ &\subseteq \{U \in \mathcal{V} \colon \mathrm{gamb}(U) \subseteq \mathrm{opt}(\mathrm{gamb}(\mathcal{V}))\} \\ &= \mathrm{norm}_{\mathrm{opt}}(\mathcal{V})\end{aligned}$$

because $\mathcal{U} \subseteq \mathcal{V}$. We conclude that

$$\mathrm{norm}_{\mathrm{opt}}(\mathcal{T}) \subseteq \mathrm{norm}_{\mathrm{opt}}(\mathcal{U}) \subseteq \mathrm{norm}_{\mathrm{opt}}(\mathcal{V}).$$

Now use $\mathrm{norm}_{\mathrm{opt}}(\mathcal{T}) = \mathrm{norm}_{\mathrm{opt}}(\mathcal{V})$. □

Proof of Theorem 12. (A) $\Longrightarrow$ (B). Immediate, since for strategically equivalent trees, $\mathrm{norm}_{\mathrm{opt}}(T_1) = \mathrm{norm}_{\mathrm{opt}}(T_2)$ by definition.

(B) $\Longrightarrow$ (C). See Lemmas 13 and 14.

(C) $\Longrightarrow$ (A). We proceed by structural induction. The base step is trivial. The induction hypothesis is that, for a $T = \bigsqcup_{i=1}^{n} X_i T_i$, we have $\mathrm{norm}_{\mathrm{opt}}(T_i) = \mathrm{back}_{\mathrm{opt}}(T_i)$ for all i. The induction step is to show that this implies $\mathrm{norm}_{\mathrm{opt}}(T) = \mathrm{back}_{\mathrm{opt}}(T)$.

Let K_i be the root node of T_i. For any i such that K_i is in at least one element of $\mathrm{norm}_{\mathrm{opt}}(T)$, we know from Property 2 that $\mathrm{st}_{K_i}(\mathrm{norm}_{\mathrm{opt}}(T)) \subseteq \mathrm{norm}_{\mathrm{opt}}(T_i) = \mathrm{back}_{\mathrm{opt}}(T_i)$. If instead K_i is not in at least one element of $\mathrm{norm}_{\mathrm{opt}}(T)$, then nothing from $\mathrm{back}_{\mathrm{opt}}(T_i)$ is involved in $\mathrm{norm}_{\mathrm{opt}}(T)$. Therefore,

$$\mathrm{norm}_{\mathrm{opt}}(T) \subseteq \bigsqcup_{i=1}^{n} X_i \,\mathrm{back}_{\mathrm{opt}}(T_i) \subseteq \mathrm{nfd}(T).$$

Since $\mathrm{norm}_{\mathrm{opt}}(\mathrm{nfd}(T)) = \mathrm{norm}_{\mathrm{opt}}(T)$ and it follows from Property 1 that $\mathrm{norm}_{\mathrm{opt}}(\mathrm{norm}_{\mathrm{opt}}(T)) = \mathrm{norm}_{\mathrm{opt}}(T)$,[2] we can use Lemma 15 to conclude that

$$\begin{aligned}\mathrm{back}_{\mathrm{opt}}(T) &= \mathrm{norm}_{\mathrm{opt}}\left(\bigsqcup_{i=1}^{n} X_i \,\mathrm{back}_{\mathrm{opt}}(T_i)\right) \\ &= \mathrm{norm}_{\mathrm{opt}}(T).\end{aligned}$$

□

4 Subtree Perfectness

Subtree perfectness means that, when a normal form solution is restricted to a subtree of a deterministic system tree, it is equal to the solution of the subtree.

Definition 16. *A normal form operator* norm *is subtree perfect if, for any deterministic system tree T, and any node N in at least one element of* $\mathrm{norm}(T)$*,*

$$\mathrm{st}_N(\mathrm{norm}(T)) = \mathrm{norm}(\mathrm{st}_N(T)).$$

This is just a stronger form of Property 2, and so it is unsurprising that the necessary and sufficient conditions on opt turn out to be identical apart from having equalities instead of inclusions.

Property 5 (Intersection property)**.** *For any sets of gambles $\mathcal{X}$ and $\mathcal{Y}$ such that $\mathcal{Y} \subseteq \mathcal{X}$ and* $\mathrm{opt}(\mathcal{X}) \cap \mathcal{Y} \neq \emptyset$*,*

$$\mathrm{opt}(\mathcal{Y}) = \mathrm{opt}(\mathcal{X}) \cap \mathcal{Y}.$$

Property 6 (Addition Property)**.** *For any gamble X and any non-empty finite set of gambles $\mathcal{Y}$,*

$$\mathrm{opt}(X + \mathcal{Y}) = X + \mathrm{opt}(\mathcal{Y}).$$

[2] Use $\mathcal{Y} = \mathrm{opt}(\mathcal{X})$ in Property 1.

Note that Property 1 is actually included within Property 5 (which is in fact equivalent to saying that opt defines a total preorder [1]). Another useful reformulation of Property 5 is [16, 6]:

Property 7 (Very strong path independence)**.** *For any sets of gambles $\mathcal{X}_1, \ldots, \mathcal{X}_n$, let $\mathcal{I} = \{i\colon \mathcal{X}_i \cap \operatorname{opt}(\cup_{i=1}^n \mathcal{X}_i) \neq \emptyset\}$. Then,*

$$\operatorname{opt}\left(\bigcup_{i=1}^{n} \mathcal{X}_i\right) = \bigcup_{i\in\mathcal{I}}^{n} \operatorname{opt}(\mathcal{X}_i).$$

Theorem 17. *The normal form operator* $\operatorname{norm}_{\operatorname{opt}}$ *is subtree perfect for deterministic system trees if and only if* opt *satisfies Properties 5 and 6.*

Lemma 18. *Consider a deterministic system tree $T = \bigsqcup_{i=1}^n X_i T_i$, and any choice function* opt*. For each tree T_i, let K_i be its root. Then, K_i is in at least one element of* $\operatorname{norm}_{\operatorname{opt}}(T)$ *if and only if*

$$(X_i + \operatorname{gamb}(T_i)) \cap \operatorname{opt}(\operatorname{gamb}(T)) \neq \emptyset. \tag{3}$$

Proof. Eq. (3) holds if and only if there is a normal form decision $U \in \operatorname{nfd}(T_i)$ such that $X_i + \operatorname{gamb}(U) \subseteq \operatorname{opt}(\operatorname{gamb}(T))$. This is equivalent to there being a U such that $\operatorname{gamb}(\sqcup X_i U) \subseteq \operatorname{opt}(\operatorname{gamb}(T))$. Clearly, $\sqcup X_i U$ is a normal form decision of T, and so by definition of $\operatorname{norm}_{\operatorname{opt}}$, Eq. (3) holds if and only if $\sqcup X_i U$ is in $\operatorname{norm}_{\operatorname{opt}}(T)$, which holds if and only if K_i is in at least one element of $\operatorname{norm}_{\operatorname{opt}}(T)$. □

Lemma 19. *If $T = \bigsqcup_{i=1}^n X_i T_i$, and* opt *is a choice function satisfying Properties 5 and 6, then*

$$\operatorname{gamb}(\operatorname{norm}_{\operatorname{opt}}(T)) = \bigcup_{i\in\mathcal{I}} (X_i + \operatorname{gamb}(\operatorname{norm}_{\operatorname{opt}}(T_i))) \tag{4}$$

implies

$$\operatorname{norm}_{\operatorname{opt}}(T) = \operatorname{nfd}\left(\bigsqcup_{i\in\mathcal{I}} X_i \operatorname{norm}_{\operatorname{opt}}(T_i)\right),$$

where $\mathcal{I} = \{i \in \{1,\ldots,n\}\colon (X_i + \operatorname{gamb}(T_i)) \cap \operatorname{opt}(\operatorname{gamb}(T)) \neq \emptyset\}$.

Proof. We first show that

$$\operatorname{norm}_{\operatorname{opt}}(T) \supseteq \operatorname{nfd}\left(\bigsqcup_{i\in\mathcal{I}} X_i \operatorname{norm}_{\operatorname{opt}}(T_i)\right).$$

Consider a normal form decision $U \in \operatorname{nfd}\left(\bigsqcup_{i\in\mathcal{I}} X_i \operatorname{norm}_{\operatorname{opt}}(T_i)\right)$. To show that $U \in \operatorname{norm}_{\operatorname{opt}}(T)$, we must show that $U \in \operatorname{nfd}(T)$ and $\operatorname{gamb}(U) \subseteq \operatorname{gamb}(\operatorname{norm}_{\operatorname{opt}}(T))$. The former is obvious, and the latter is established by Eq. (4):

$$\begin{aligned}\operatorname{gamb}(U) &\subseteq \bigcup_{i\in\mathcal{I}} (X_i + \operatorname{gamb}(\operatorname{norm}_{\operatorname{opt}}(T_i))) \\ &= \operatorname{gamb}(\operatorname{norm}_{\operatorname{opt}}(T)).\end{aligned}$$

Next we show that

$$\operatorname{norm}_{\operatorname{opt}}(T) \subseteq \operatorname{nfd}\left(\bigsqcup_{i\in\mathcal{I}} X_i \operatorname{norm}_{\operatorname{opt}}(T_i)\right).$$

Let $U \in \operatorname{norm}_{\operatorname{opt}}(T)$. Let V be U with the root node removed, that is, $U = \sqcup X_k V$ for some k. Clearly, $V \in \operatorname{nfd}(T_k)$. It suffices to show that $V \in \operatorname{norm}_{\operatorname{opt}}(T_k)$. Let $\{Y\} = \operatorname{gamb}(V)$ and let $\mathcal{Y} = \operatorname{gamb}(T_k)$. We know that $X_k + Y \in \operatorname{gamb}(T)$, and $Y \in \operatorname{gamb}(T_k)$. Also, $X_k + \mathcal{Y} \subseteq \operatorname{gamb}(T)$. By Property 5 and Lemma 18,

$$\operatorname{opt}(X_k + \mathcal{Y}) = \operatorname{opt}(\operatorname{gamb}(T)) \cap (X_k + \mathcal{Y}).$$

By Property 6,

$$X_k + \operatorname{opt}(\mathcal{Y}) = \operatorname{opt}(X_k + \mathcal{Y}),$$

whence

$$X_k + \operatorname{opt}(\mathcal{Y}) = \operatorname{opt}(\operatorname{gamb}(T)) \cap (X_k + \mathcal{Y}).$$

We know $X_k + Y$ is in the right hand side, so $X_k + Y$ is in the left hand side. Therefore $Y \in \operatorname{opt}(\mathcal{Y})$ and $V \in \operatorname{norm}_{\operatorname{opt}}(T_k)$. □

Lemma 20 (Huntley and Troffaes [6, Lemma 17])**.** *Let* norm *be a normal form operator. Let T be a deterministic system 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 21. *If* $\operatorname{norm}_{\operatorname{opt}}$ *is subtree perfect then* opt *satisfies Property 5.*

Proof. Let $\mathcal{X}$ and $\mathcal{Y}$ be sets of gambles such that $\mathcal{Y} \subseteq \mathcal{X}$. Let T_1 and T_2 be deterministic system trees with exactly one decision node, and $\operatorname{gamb}(T_1) = \mathcal{X}$, $\operatorname{gamb}(T_2) = \mathcal{Y}$. Let $T = T_1 \sqcup T_2$ (so the arcs to T_1 and T_2 have reward 0), and N be the node at the root of T_2. So, $\operatorname{gamb}(T) = \mathcal{X}$. Now, $\operatorname{gamb}(\operatorname{norm}_{\operatorname{opt}}(T)) = \operatorname{opt}(\mathcal{X})$, and $\operatorname{gamb}(\operatorname{st}_N(\operatorname{norm}_{\operatorname{opt}}(T))) = \operatorname{gamb}(\mathcal{Y}) \cap \operatorname{opt}(\mathcal{X})$. By subtree perfectness, Property 5 follows. □

Lemma 22. *If* norm_{opt} *is subtree perfect, then* opt *satisfies Property 6.*

Proof. Let X be a gamble and let $\mathcal{Y}$ be a non-empty finite set of gambles. Let T_1 be a deterministic system tree with exactly one decision node and $\text{gamb}(T_1) = \mathcal{Y}$. Let $T = \sqcup X T_1$, so $\text{gamb}(T) = X + \mathcal{Y}$. Now,

$$\text{gamb}(\text{norm}_{\text{opt}}(T)) = \text{opt}(X + \mathcal{Y})$$

and

$$\text{gamb}(\text{norm}_{\text{opt}}(T_1)) = \text{opt}(\mathcal{Y}).$$

By subtree perfectness and the definition of norm_{opt}, we must have that, first, any gamble $X + Y \in \text{opt}(X + \mathcal{Y})$ must have $Y \in \text{opt}(\mathcal{Y})$ (else there is a $U \in \text{norm}_{\text{opt}}(T)$ that is non-optimal in T_1), and second, any $Y \in \text{opt}(\mathcal{Y})$ must have $X + Y \in \text{opt}(X + \mathcal{Y})$ (else there is a $U \in \text{norm}_{\text{opt}}(T_1)$ with $\sqcup X U$ non-optimal in T). Therefore $\text{opt}(X + \mathcal{Y}) = X + \text{opt}(\mathcal{Y})$. □

Proof of Theorem 17. "only if". Follows from Lemmas 21 and 22.

"if". We proceed by structural induction as usual. The base step is trivial. The induction hypothesis is that, for a $T = \bigsqcup_{i=1}^{n} X_i T_i$, we have subtree perfectness at all T_i. If we can show that

$$\text{gamb}(\text{norm}_{\text{opt}}(T)) = \bigcup_{i \in \mathcal{I}} (X_i + \text{gamb}(\text{norm}_{\text{opt}}(T_i)))$$

for $\mathcal{I} = \{i \in \{1, \ldots, n\} \colon (X_i + \text{gamb}(T_i)) \cap \text{opt}(\text{gamb}(T)) \neq \emptyset\}$, then by Lemma 19 and Lemma 20, subtree perfectness holds for T.

We have

$$\text{gamb}(\text{norm}_{\text{opt}}(T)) = \text{opt}\left(\bigcup_{i=1}^{n} (X_i + \text{gamb}(T_i))\right)$$

whence by Property 7

$$= \bigcup_{i \in \mathcal{I}} \text{opt}(X_i + \text{gamb}(T_i))$$

whence by Property 6

$$= \bigcup_{i \in \mathcal{I}} (X_i + \text{opt}(\text{gamb}(T_i)))$$
$$= \bigcup_{i \in \mathcal{I}} (X_i + \text{gamb}(\text{norm}_{\text{opt}}(T_i)))$$

as required. □

	Property				
	1	3	4	5	6
E-admissibility	✓	✓	✓		✓
Maximality	✓	✓	✓		✓
Γ-maximin	✓	✓		✓	
Interval Dominance	✓	✓			

Table 1: Properties of various choice functions.

5 Imprecise Probability

De Cooman and Troffaes [4, §3.2–3.5] investigate whether dynamic programming works for four common choice functions in *imprecise probability* [18], namely maximality, E-admissibility, Γ-maximin, and interval dominance. The first two satisfy all properties, and the latter two fail Property 4. Γ-maximin and interval dominance fail because of the non-additivity of a coherent lower prevision.

For subtree perfectness, none of the choice functions satisfies all the necessary properties. Property 5 requires a total preorder, and, of the four, only Γ-maximin is. Since Γ-maximin fails Property 4, it automatically fails Property 6. These results mirror those for standard decision trees [6]: only maximality and E-admissibility allow backward induction, and nothing is subtree perfect. A table showing the properties satisfied by each choice function is shown in Table 1.

As mentioned by de Cooman and Troffaes, Γ-maximin could satisfy Property 6 for certain lower previsions. Suppose that Ω is a product of possibility spaces $\Omega_1, \ldots, \Omega_m$, and the gambles on the ith decision arc in any path is a gamble on Ω_i. If the overall lower prevision $\underline{P}$ is a suitable independent product of lower previsions $\underline{P}_i$ on the Ω_i, then additivity will be satisfied. We refer to [4, §3.4] for more details and references.

6 Conclusion

In this paper we have investigated dynamic programming for deterministic discrete-time systems with uncertain gain using normal form operators induced by choice functions. We have brought the work of de Cooman and Troffaes into the decision tree setting of [6]. In doing so, we have extended their Bellman Equation Theorem [4, Theorem 14] by adding necessity to their sufficiency, allowing arbitrary rewards (so a utility function over rewards is no longer assumed), and fairly arbitrary addition operators. Also, we have decomposed Bellman's principle of optimality into two much simpler properties.

Further, we have found simple necessary and sufficient conditions for subtree perfectness, which is a stronger

form of Bellman's principle. The distinction between dynamic programming and subtree perfectness is not often made (see for instance the informal description of Property 2 by Luenberger [9, p. 419]: this is clearly subtree perfectness being described).

A likely reason for this lack of distinction is that, under the assumption of a total preorder (a very popular assumption in decision theory literature) the two concepts become almost identical. We cannot think of a well-known choice function for any uncertainty model that satisfies Properties 4 and 5 but not Property 6. The distinction is much more important with imprecise methods, where a major attraction is the ability to model indecision and incomparability of options. In such cases, subtree perfectness will always fail.

The key observations are that lack of subtree perfectness is not necessarily a barrier to dynamic programming, but nor is success of dynamic programming enough to guarantee that one's normal form solution is completely well-behaved.

Acknowledgements

The first author is supported by the EPSRC. We thank both reviewers for their valuable comments and suggestions.

References

[1] Kenneth J. Arrow. Rational choice functions and orderings. *Economica*, 26(102):121–127, May 1959.

[2] R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, 1957.

[3] Robert T. Clemen and Terence Reilly. *Making Hard Decisions*. Duxbury, 2001.

[4] G. De Cooman and M.C.M. Troffaes. Dynamic programming for deterministic discrete-time systems with uncertain gain. *International Journal of Approximate Reasoning*, 39(2-3):257–278, Jun 2005.

[5] P. Hammond. Consequentialist foundations for expected utility. *Theory and Decision*, 25(1):25–78, Jul 1988.

[6] N. Huntley and M. C. M. Troffaes. Characterizing factuality in normal form sequential decision making. In Thomas Augustin, Frank P. A. Coolen, Serafin Moral, and Matthias C. M. Troffaes, editors, *ISIPTA'09: Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications*, pages 239–248, 2009.

[7] D. V. Lindley. *Making Decisions*. Wiley, London, 2nd edition, 1985.

[8] R.D. Luce and H. Raiffa. *Games and Decisions: introduction and critical survery*. Wiley, 1957.

[9] D. G. Luenberger. *Introduction to Dynamic Systems*. Wiley, 1979.

[10] M.J. Machina. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature*, 27(1622-1688), 1989.

[11] E. F. McClennen. *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge University Press, 1990.

[12] C.R. Plott. Path independence, rationality, and social choice. *Econometrica*, 41(6):1075–1091, Nov 1973.

[13] P. Ray. Independence of irrelevant alternatives. *Econometrica*, 41(5):987–991, Sep 1973.

[14] Teddy Seidenfeld. When normal and extensive form decisions differ. In D. Prawitz, B. Skyrms, and D. Westerstahl, editors, *Logic, Methodology and Philosophy of Science IX, Proceedings of the Ninth International Congress of Logic, Methodology and Philosophy of Science*, volume 134 of *Studies in Logic and the Foundations of Mathematics*, pages 451–463. Elsevier, 1995.

[15] R. Selten. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4(1):25–55, Mar 1975.

[16] A. K. Sen. Social choice theory: A re-examination. *Econometrica*, 45(1):53–89, 1977.

[17] M. C. M. Troffaes, N. Huntley, and R. Shirota Filho. Sequential decision processes under act-state independence with arbitrary choice functions. In E. Huellermeier, R. Kruse, and F. Hoffmann, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 98–107. Springer, 2010.

[18] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

A Note on Local Computations in Dempster-Shafer Theory of Evidence

Radim Jiroušek
Faculty of Management of University of Economics, and
Institute of Information Theory and Automation, Academy of Sciences
Czech Republic
radim@utia.cas.cz

Abstract

When applying any technique of multidimensional models to problems of practice, one always has to cope with two problems: it is necessary to have a possibility to represent the models with a "reasonable" number of parameters and to have sufficiently efficient computational procedures at one's disposal. When considering graphical Markov models in probability theory, both of these conditions are fulfilled; various computational procedures for decomposable models are based on the ideas of local computations, whose theoretical foundations were laid by Lauritzen and Spiegelhalter.

The presented contribution studies a possibility of transferring these ideas from probability theory into Dempster-Shafer theory of evidence. The paper recalls decomposable models, discusses connection of the model structure with the corresponding system of conditional independence relations, and shows that under special additional conditions, one can locally compute specific basic assignments which can be considered to be conditional.

Keywords. Multidimensional models, graphical models, conditional independence, factorisation, computations.

1 Introduction

The great advantage of Dempster-Shafer theory [5, 18] is the fact that it generalises classical probability theory in the way that one can easily describe not only uncertainty but also vagueness (ignorance). Nevertheless, the disadvantage of this approach stems from the fact that belief functions cannot be represented by a point function (like density in probability theory); instead, one has to manipulate with set functions, which leads to exponential increase of algorithmic complexity of all the necessary computational procedures.

With regard to probability theory, substantial decrease of computational complexity was achieved with the help of Graphical Markov Models (GMM), a technique developed in the last quarter of the last century. Here we specifically have in mind a technique based on local computations for which theoretical background was laid by Lauritzen and Spiegelhalter [17]. Its basic idea can be expressed in a few words: a multidimensional distribution represented by a Bayesian network is first converted into a decomposable model, which allows for efficient computation of conditional probabilities.

Studying properly probabilistic GMM one can realise that it is a notion of *conditional independence* (which is closely connected with a notion of *factorisation*) that makes it possible to represent multidimensional probability distributions efficiently. A goal of this paper is to make a brief survey summarising results concerning decomposable models within Dempster-Shafer theory of evidence presented in [10, 11, 12]. In addition to this we will show that, even in Dempster-Shafer theory, one can employ the basic ideas of Lauritzen and Spiegelhalter and compute "conditional" basic assignments locally.

1.1 Notation

In this paper we consider a finite multidimensional space $\mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_n$, and its subspaces (for all $K \subseteq N$)

$$\mathbf{X}_K = \bigtimes_{i \in K} \mathbf{X}_i.$$

For a point $x = (x_1, x_2, \ldots, x_n) \in \mathbf{X}_N$ its projection into subspace $\mathbf{X}_K$ is denoted $x^{\downarrow K} = (x_{i,i \in K})$, and for $A \subseteq \mathbf{X}_N$

$$A^{\downarrow K} = \{y \in \mathbf{X}_K : \exists x \in A, x^{\downarrow K} = y\}.$$

By a *join* of two sets $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_L$ we 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 $A \otimes B = A \times B$, if $K = L$ then $A \otimes B = A \cap B$.

In view of this paper it is important to realise that if $x \in C \subseteq \mathbf{X}_{K \cup L}$, then $x^{\downarrow K} \in C^{\downarrow K}$ and $x^{\downarrow L} \in C^{\downarrow L}$, which means that always $C \subseteq C^{\downarrow K} \otimes C^{\downarrow L}$. However, it does not mean that $C = C^{\downarrow K} \otimes C^{\downarrow L}$. For example, considering two-dimensional frame of discernment $\mathbf{X}_{\{1,2\}}$ with $\mathbf{X}_i = \{a_i, \bar{a}_i\}$ for both $i = 1, 2$, and $C = \{a_1a_2, \bar{a}_1a_2, a_1\bar{a}_2\}$, one gets

$$C^{\downarrow\{1\}} \otimes C^{\downarrow\{2\}} = \{a_1, \bar{a}_1\} \otimes \{a_2, \bar{a}_2\} = \{a_1a_2, \bar{a}_1a_2, a_1\bar{a}_2, \bar{a}_1\bar{a}_2\} \supsetneq C.$$

1.2 Basic assignments

The role played by a probability distribution in probability theory is replaced by that of a set function in Dempster-Shafer theory: belief function, plausibility function or basic (*probability or belief*) assignment. Knowing one of them, one can derive the remaining two. In this paper we will use almost 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

$$\sum_{\emptyset \neq A \subseteq \mathbf{X}_K} m(A) = 1.$$

If $m(A) > 0$, then A is said to be a *focal element* of m. Recall that

$$Bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B),$$

and

$$Pl(A) = \sum_{B \subseteq \mathbf{X}_K : B \cap A \neq \emptyset} m(B).$$

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 $\emptyset \neq B \subseteq \mathbf{X}_L$):

$$m^{\downarrow L}(B) = \sum_{A \subseteq \mathbf{X}_K : A^{\downarrow L} = B} m(A).$$

1.3 Operator of composition

Compositional models were introduced for probability theory in [8] as an alternative to Bayesian networks for efficient representation of multidimensional measures. They were based on recurrent application of an operator of composition. An analogous operator within the framework of Dempster-Shafer theory was introduced in [14]).

Definition 1 ***Operator of Composition.*** *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 $(m_1 \triangleright m_2)(C) = 0$.*

Remark 1 First of all, we want to stress that the operator of composition is something other than the famous Dempster's rule of combination [5], or its non-normalised version, the so called *conjunctive combination rule* [1]

$$(m_1 \textcircled{\cap} m_2)(C) = \sum_{A \subseteq \mathbf{X}_K, B \subseteq \mathbf{X}_L : A \otimes B = C} m_1(A) \cdot m_2(B).$$

For example, the operation of composition is (in contrast with the above-mentioned conjunctive combination rule) neither commutative nor associative. While Dempster's rule of combination was designed to combine different (independent) sources of information (it realises fusion of sources), the operator of composition primarily serves for composing pieces of local information (usually coming from one source) into a global model. The notion of composition is therefore closely connected with the notion of *factorisation*. This fact manifests also in the following difference: while for computation of $(m_1 \triangleright m_2)(C)$ it is enough to know only m_1 and m_2 just for the respective projections of set C, computing $(m_1 \textcircled{\cap} m_2)(C)$ requires knowledge of, roughly speaking, the entire basic assignments m_1 and m_2.

For further intuitive justification of the operator of composition the reader is referred to [14], where a number of its properties were proved. In view of the forthcoming text, those presented in the following assertion are the most important.

Proposition 1 ***Basic Properties.*** *Let m_1 and m_2 be basic assignments defined on $\mathbf{X}_K, \mathbf{X}_L$, respectively. Then:*

1. *$m_1 \triangleright m_2$ is a basic assignment on $\mathbf{X}_{K \cup L}$;*
2. *$(m_1 \triangleright m_2)^{\downarrow K} = m_1$;*
3. *$m_1 \triangleright m_2 = m_2 \triangleright m_1 \iff m_1^{\downarrow K \cap L} = m_2^{\downarrow K \cap L}$.*

The reader probably noticed that Property 2 guarantees idempotency of the operator and gives a hint about how to get a counterexample to its commutativity. From point 1, one immediately gets that for basic assignments $m_1, m_2, \ldots, m_r$ defined on $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}, \ldots, \mathbf{X}_{K_r}$, respectively, the formula $m_1 \triangleright m_2 \triangleright \ldots \triangleright m_r$ defines a (possibly multidimensional) basic assignment defined on $\mathbf{X}_{K_1 \cup \ldots \cup K_r}$.

2 Controlled associativity

As already mentioned above, the operator of composition is not associative. This means that in fact we do not know what the formula $m_1 \triangleright m_2 \triangleright \ldots \triangleright m_r$ means. To avoid the necessity of using too many parentheses, let us make the following convention. In the formulae like $m_1 \triangleright m_2 \triangleright \ldots \triangleright m_r$, when the order of application of the operators of composition is not controlled by parentheses, the operators will be applied from left to right, i.e.,

$$m_1 \triangleright m_2 \triangleright \ldots \triangleright m_r = (\ldots (m_1 \triangleright m_2) \triangleright \ldots \triangleright m_{r-1}) \triangleright m_r.$$

Nevertheless, when designing a process of local computations for compositional models in D-S theory (which is intended to be an analogy to the process proposed by Lauritzen and Spiegelhalter in [17]), one needs a type of associativity expressed in the following assertion.

Proposition 2 ***Controlled associativity.*** *Let m_1, m_2 and m_3 be basic assignments on $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}$ and $\mathbf{X}_{K_3}$, respectively, such that $K_2 \supseteq K_1 \cap K_3$, and*

$$m_1^{\downarrow K_1 \cap K_2}(C^{\downarrow K_1 \cap K_2}) > 0 \Longrightarrow m_2^{\downarrow K_1 \cap K_2}(C^{\downarrow K_1 \cap K_2}) > 0.$$

Then

$$(m_1 \triangleright m_2) \triangleright m_3 = m_1 \triangleright (m_2 \triangleright m_3)\,.$$

Proof. The goal is to prove that for any $C \subseteq \mathbf{X}_{K_1 \cup K_2 \cup K_3}$

$$((m_1 \triangleright m_2) \triangleright m_3)(C) = (m_1 \triangleright (m_2 \triangleright m_3))(C). \quad (1)$$

We have to distinguish five special cases.

A. $C \neq C^{\downarrow K_1} \otimes C^{\downarrow K_2} \otimes C^{\downarrow K_3}$.
This is the simplest situation because, due to associativity of join,

$$(C^{\downarrow K_1} \otimes C^{\downarrow K_2}) \otimes C^{\downarrow K_3} = C^{\downarrow K_1} \otimes (C^{\downarrow K_2} \otimes C^{\downarrow K_3})$$

and therefore in this case both sides of formula (1) equal 0, which follows from Definition 1 (case [c]).

B. $C = C^{\downarrow K_1} \otimes C^{\downarrow K_2} \otimes C^{\downarrow K_3}$
& $m_2^{\downarrow K_1 \cap K_2}(C^{\downarrow K_1 \cap K_2}) > 0, m_3^{\downarrow K_2 \cap K_3}(C^{\downarrow K_2 \cap K_3}) > 0.$

In this case, under the given assumptions,

$$K_3 \cap (K_1 \cup K_2) = K_3 \cap K_2$$

and therefore

$$((m_1 \triangleright m_2) \triangleright m_3)(C) = \frac{m_1(C^{\downarrow K_1}) \cdot m_2(C^{\downarrow K_2})}{m_2^{\downarrow K_2 \cap K_1}(C^{\downarrow K_2 \cap K_1})} \cdot \frac{m_3(C^{\downarrow K_3})}{m_3^{\downarrow K_3 \cap K_2}(C^{\downarrow K_3 \cap K_2})}.$$

Analogously, we can make the following computations (in the last modification we use the fact that in the considered case $K_1 \cap K_2 \cap K_3 = K_1 \cap K_3$):

$$\begin{aligned}
(m_1 \triangleright (m_2 \triangleright m_3))(C) &= \frac{m_1(C^{\downarrow K_1}) \cdot (m_2 \triangleright m_3)(C^{\downarrow K_2 \cup K_3})}{(m_2 \triangleright m_3)^{\downarrow K_1 \cap (K_2 \cup K_3)}(C^{\downarrow K_1 \cap (K_2 \cup K_3)})} \\
&= \frac{m_1(C^{\downarrow K_1})}{(m_2 \triangleright m_3)^{\downarrow K_1 \cap (K_2 \cup K_3)}(C^{\downarrow K_1 \cap (K_2 \cup K_3)})} \\
&\qquad \cdot \frac{m_2(C^{\downarrow K_2}) \cdot m_3(C^{\downarrow K_3})}{m_3^{\downarrow K_2 \cap K_3}(C^{\downarrow K_2 \cap K_3})} \\
&= \frac{m_1(C^{\downarrow K_1}) \cdot m_3^{\downarrow K_1 \cap K_2 \cap K_3}(C^{\downarrow K_1 \cap K_2 \cap K_3})}{m_2^{\downarrow K_1 \cap K_2}(C^{\downarrow K_1 \cap K_2}) \cdot m_3^{\downarrow K_1 \cap K_3}(C^{\downarrow K_1 \cap K_3})} \\
&\qquad \cdot \frac{m_2(C^{\downarrow K_2}) \cdot m_3(C^{\downarrow K_3})}{m_3^{\downarrow K_2 \cap K_3}(C^{\downarrow K_2 \cap K_3})} \\
&= \frac{m_1(C^{\downarrow K_1}) \cdot m_2(C^{\downarrow K_2}) \cdot m_3(C^{\downarrow K_3})}{m_2^{\downarrow K_1 \cap K_2}(C^{\downarrow K_1 \cap K_2}) \cdot m_3^{\downarrow K_2 \cap K_3}(C^{\downarrow K_2 \cap K_3})},
\end{aligned}$$

which proves that the equality (1) holds.

C. $C = C^{\downarrow K_1} \otimes C^{\downarrow K_2} \otimes C^{\downarrow K_3}$
& $m_2^{\downarrow K_1 \cap K_2}(C^{\downarrow K_1 \cap K_2}) > 0, m_3^{\downarrow K_2 \cap K_3}(C^{\downarrow K_2 \cap K_3}) = 0.$
In this case, if $C^{\downarrow K_3 \setminus K_2} \neq \mathbf{X}_{K_3 \setminus K_2}$ then both sides of formula (1) equal 0. This is because, due to Definition 1, both composed assignments $(m_1 \triangleright m_2) \triangleright m_3$ and $m_2 \triangleright m_3$ equal 0 for this C, and therefore also $(m_1 \triangleright (m_2 \triangleright m_3))(C) = 0$.

Therefore, consider $C = C^{\downarrow K_1} \otimes C^{\downarrow K_2} \otimes \mathbf{X}_{K_3 \setminus K_2}$. For this we get from Definition 1

$$((m_1 \triangleright m_2) \triangleright m_3)(C) = (m_1 \triangleright m_2)(C^{\downarrow K_1 \cup K_2}).$$

For the right-hand side of formula (1) we get

$$(m_2 \triangleright m_3)(C^{\downarrow K_2 \cup K_3}) = m_2(C^{\downarrow K_2})$$

and therefore

$$(m_1 \triangleright (m_2 \triangleright m_3))(C) = (m_1 \triangleright m_2)(C^{\downarrow K_1 \cup K_2}).$$

D. $C = C^{\downarrow K_1} \otimes C^{\downarrow K_2} \otimes C^{\downarrow K_3}$
& $m_2^{\downarrow K_1 \cap K_2}(C^{\downarrow K_1 \cap K_2}) = 0, m_3^{\downarrow K_2 \cap K_3}(C^{\downarrow K_2 \cap K_3}) > 0.$

focal elements	$(m_1 \triangleright m_2) \triangleright m_3$
$\{a_1a_2\}$	$\frac{1}{3}$
$\{a_1\bar{a}_2\}$	$\frac{1}{3}$
$\{a_1a_2, a_1\bar{a}_2\}$	$\frac{1}{3}$

Table 1: Composed basic assignment $(m_1 \triangleright m_2) \triangleright m_3$

Since we assume that $m_1^{\downarrow K_1\cap K_2}(C^{\downarrow K_1\cap K_2}) > 0$ implies $m_2^{\downarrow K_1\cap K_2}(C^{\downarrow K_1\cap K_2}) > 0$, we know that for the considered C, $m_1^{\downarrow K_1\cap K_2}(C^{\downarrow K_1\cap K_2}) = 0$, and therefore both sides of formula (1) equal 0 because m_1 is marginal to both $(m_1 \triangleright m_2) \triangleright m_3$ and $m_1 \triangleright (m_2 \triangleright m_3)$.

E. $C = C^{\downarrow K_1} \otimes C^{\downarrow K_2} \otimes C^{\downarrow K_3}$
& $m_2^{\downarrow K_1\cap K_2}(C^{\downarrow K_1\cap K_2}) = 0, m_3^{\downarrow K_2\cap K_3}(C^{\downarrow K_2\cap K_3}) = 0$.
It is obvious from Definition 1 that both sides of formula (1) equal 0 for all C but for $C = C^{\downarrow K_1} \otimes \mathbf{X}_{K_2\setminus K_1} \otimes \mathbf{X}_{K_3\setminus K_1}$. For this special case, however,

$$((m_1 \triangleright m_2) \triangleright m_3)(C) = m_1(C^{\downarrow K_1}),$$
$$(m_1 \triangleright (m_2 \triangleright m_3))(C) = m_1(C^{\downarrow K_1}). \qquad \square$$

Example: Let us illustrate the necessity of the assumption

$$m_1^{\downarrow K_1\cap K_2}(C^{\downarrow K_1\cap K_2}) > 0 \Longrightarrow m_2^{\downarrow K_1\cap K_2}(C^{\downarrow K_1\cap K_2}) > 0$$

required in Lemma 2 by (for the sake of simplicity a rather degenerated) example. Consider three basic assignments m_1, m_2 and m_3. Assume that in this case $K_1 = K_2 = \{1\}$ and $K_3 = \{1, 2\}$, $\mathbf{X}_i = \{a_i, \bar{a}_i\}$ for both $i = 1, 2$. Define $m_1(\{a_1\}) = 1$ and $m_2(\{\bar{a}_1\}) = 1$, which means that both m_1, m_2 have only one focal element, and $m_3(A) = \frac{1}{15}$ for all nonempty subsets of $\mathbf{X}_1 \times \mathbf{X}_2$.

For these basic assignments we immediately get $m_1 = m_1 \triangleright m_2$ (when applying Definition 1, one has to take $C^{\downarrow K_1} \times \mathbf{X}_\emptyset = C^{\downarrow K_1}$), and therefore one gets $m_1 \triangleright m_2 \triangleright m_3$ as indicated in Table 1. Analogously, one gets $m_2 \triangleright m_3$ which is depicted in Table 2. Computing

focal elements	$m_2 \triangleright m_3$
$\{\bar{a}_1a_2\}$	$\frac{1}{3}$
$\{\bar{a}_1\bar{a}_2\}$	$\frac{1}{3}$
$\{\bar{a}_1a_2, a_1\bar{a}_2\}$	$\frac{1}{3}$

Table 2: Composed basic assignment $m_2 \triangleright m_3$

now the basic assignment $m_1 \triangleright (m_2 \triangleright m_3)$, one gets a basic assignment with only one focal element

$$(m_1 \triangleright (m_2 \triangleright m_3))(\{a_1\} \times \mathbf{X}_2) = 1.$$

Thus we have shown that in this case

$$(m_1 \triangleright m_2) \triangleright m_3 \neq m_1 \triangleright (m_2 \triangleright m_3)\,.$$

3 Decomposable models

3.1 Independence and factorisation

What makes the representation and local computations with multidimensional probability distributions feasible is the property of factorisation [17]. Therefore, in [10] we also introduced this notion into Dempster-Shafer theory of evidence.

Definition 2 ***Simple Factorisation.*** *Consider two nonempty sets $K \cup L = N$. We say that basic assignment m* factorises with respect to (K, L) *if there exist two nonnegative set functions*

$$\phi : \mathcal{P}(\mathbf{X}_K) \longrightarrow [0, +\infty), \quad \psi : \mathcal{P}(\mathbf{X}_L) \longrightarrow [0, +\infty),$$

such that for all $A \subseteq \mathbf{X}_{K\cup L}$

$$m(A) = \begin{cases} \phi(A^{\downarrow K}) \cdot \psi(A^{\downarrow L}) & \text{if } A = A^{\downarrow K} \otimes A^{\downarrow L} \\ 0 & \text{otherwise.} \end{cases}$$

Example: Consider $\mathbf{X}_{\{1,2,3\}} = \mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3$ with all three $\mathbf{X}_i = \{a_i, \bar{a}_i\}$ as in the preceding example, and consider basic assignment m factorising with respect to $(\{1, 2\}, \{2, 3\})$. This means that it can be represented with the help of two functions

$$\phi : \mathcal{P}(\mathbf{X}_{\{1,2\}}) \to [0, +\infty), \quad \psi : \mathcal{P}(\mathbf{X}_{\{2,3\}}) \to [0, +\infty).$$

Since both subspaces $\mathbf{X}_{\{1,2\}}$ and $\mathbf{X}_{\{2,3\}}$ have 15 nonempty subsets, each of these functions is defined with the help of maximally 15 numbers, which means that the considered basic assignment can be represented with 30 parameters. Generally, a basic assignment on $\mathbf{X}_{\{1,2,3\}}$ can have up to 255 focal elements, and the number of sets $A \subseteq \mathbf{X}_{\{1,2,3\}}$ for which $A \neq A^{\downarrow\{1,2\}} \otimes A^{\downarrow\{2,3\}}$ is 156.

Remark 2 Notice that the importance of the factorisation does not follow only from the fact that the basic assignment m in the preceding example can be represented by two functions ϕ and ψ, i.e., just with 30 parameters, but especially in the fact that the value $m(A)$ can be computed just from two values: $\phi(A^{\downarrow\{1,2\}})$ and $\psi(A^{\downarrow\{2,3\}})$. Value $m(A)$ does not depend on values of functions ϕ and ψ in other points of their domains of definition.

In probability theory, the notion of factorisation is closely connected with the notion of conditional independence. The same holds in Dempster-Shafer theory under the assumption that one accepts the notion of conditional independence as it appears in the following Definition 3, introduced originally in [13]. Nevertheless, based on the recommendation of the anonymous referee, let us first repeat some intuitive reasoning published in [13] that led us to this definition.

There are at least three ways to introduce a generally accepted concept of unconditional (some authors call it marginal) independence (non-interactivity) for two disjoint groups of variables X_K and X_L. Here we will mention two of them, neither of which requires Dempster's rule of combination. The older one, used for example by Ben Yaghlane et al. [1], Shenoy [19] and Studený [21], is based on the properties of a *commonality function* defined for basic assignment m by the formula

$$Q(A) = \sum_{B \subseteq \mathbf{X}_N : A \subseteq B} m(B).$$

According to this older definition, we say that disjoint groups of variables X_K and X_L are (unconditionally) independent with respect to basic assignment m if

$$Q^{\downarrow K \cup L}(A) = Q^{\downarrow K}(A^{\downarrow K}) \cdot Q^{\downarrow L}(A^{\downarrow L})$$

for any $A \subseteq \mathbf{X}_{K \cup L}$. The other (equivalent) definition says that X_K and X_L are independent if for all $A \subseteq \mathbf{X}_{K \cup L}$ for which $A = A^{\downarrow K} \times A^{\downarrow L}$

$$m^{\downarrow K \cup L}(A) = m^{\downarrow K}(A^{\downarrow K}) \cdot m^{\downarrow L}(A^{\downarrow L}),$$

and $m^{\downarrow K \cup L}(A) = 0$ for all the remaining $A \subseteq \mathbf{X}_{K \cup L}$ for which $A \neq A^{\downarrow K} \times A^{\downarrow L}$. Both of these definitions invite generalisation for the case of overlapping groups of variables, both these generalisations satisfy the so-called semigraphoid properties, and yet these generalisations do not coincide. As it is discussed in [2], Studený showed that the generalisation based on the commonality functions is not consistent with marginalisation (for details the reader is referred to [2]), and this is one of the reasons why we prefer the following definition (another reason is that for the concept of conditional independence from Definition 3, one can prove the Factorisation Lemma - see Proposition 3 below).

Definition 3 ***Conditional Independence.*** *Let m be a basic assignment on $\mathbf{X}_N$ and $K, L, M \subset N$ be disjoint, both $K, L \neq \emptyset$. 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 for any $A \subseteq \mathbf{X}_{K \cup L \cup M}$ such that $A = A^{\downarrow K \cup M} \otimes A^{\downarrow L \cup M}$ the equality*

$$m^{\downarrow K \cup L \cup M}(A) \cdot m^{\downarrow M}(A^{\downarrow M}) = m^{\downarrow K \cup M}(A^{\downarrow K \cup M}) \cdot m^{\downarrow L \cup M}(A^{\downarrow L \cup M})$$

holds true, and $m^{\downarrow K \cup L \cup M}(A) = 0$ for all the remaining $A \subseteq \mathbf{X}_{K \cup L \cup M}$, for which $A \neq A^{\downarrow K \cup M} \otimes A^{\downarrow L \cup M}$.

Remark 3 As already mentioned above, it was shown in [13] that this definition meets all the semigraphoid axioms [21] and that for $M = \emptyset$ it reduces to the generally accepted definition of (unconditional, or marginal) independence (see, e.g., [1]).

Important relationships between this type of conditional independence and factorisation (operator of composition) are presented in the following two assertions proved in [14] and [23], respectively.

Proposition 3 ***Factorisation Lemma.*** *Let $K, L \subseteq N$ be nonempty, $K \cup L = N$. m factorises with respect to (K, L) if and only if*

$$K \setminus L \perp\!\!\!\perp L \setminus K \,|\, K \cap L\ [m].$$

Proposition 4 ***Factorisation of Composition.*** *Let $K, L \subseteq N$ be nonempty, $K \cup L = N$. m factorises with respect to (K, L) if and only if*

$$m = m^{\downarrow K} \triangleright m^{\downarrow L}.$$

3.2 Graphical models

In probability theory, graphical models were defined as probability distributions (measures) factorising with respect to a system of subsets forming cliques of a graph (Daroch, Lauritzen and Speed 1980, Edwards and Havránek 1985). For the sake of this paper we will just define a subclass of graphical models, so-called decomposable models, which factorise with respect to decomposable graphs, i.e., with respect to the graphs whose cliques (maximal complete subsets of nodes) can be ordered to meet the so-called *Running Intersection Property* (RIP): for all $i = 2, \ldots, r$ there exists $j, 1 \leq j < i$, such that

$$K_i \cap (K_1 \cup \ldots \cup K_{i-1}) \subseteq K_j.$$

This offers us a possibility to define decomposable models using Definition 2 recursively.

Definition 4 ***Decomposable Basic Assignments.*** *We say that a basic assignment m is* decomposable *if it* factorises with respect to a

decomposable graph *in the following sense (let* $K_1, K_2, \ldots, K_r$ *be cliques of the considered decomposable graph ordered so that they meet RIP): for all* $i = 2, \ldots, r$ *the marginal* $m^{\downarrow K_1 \cup \ldots \cup K_i}$ *factorises (in the sense of Definition 2) with respect to* $(K_1 \cup \ldots \cup K_{i-1}, K_i)$.

By repeated application of Proposition 4 one can see that a decomposable model can easily be represented by a system of its marginals.

Proposition 5 ***Composition of Decomposable Models.*** *Consider a decomposable graph with cliques* $K_1, \ldots, K_r$. *If this ordering meets RIP then* m *is decomposable with respect to the graph in question if and only if*

$$m = m^{\downarrow K_1} \triangleright m^{\downarrow K_2} \triangleright \ldots \triangleright m^{\downarrow K_{r-1}} \triangleright m^{\downarrow K_r}.$$

This assertion says that a basic assignment is decomposable if it can be composed from a system of its marginals (the structure of the system must correspond to cliques of a decomposable graph). We can also ask the opposite question: having a system of low-dimensional marginal basic assignment $m_1, m_2, \ldots, m_r$ defined on $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}, \ldots, \mathbf{X}_{K_r}$, respectively, what are the properties of the multidimensional basic assignment $m_1 \triangleright m_2 \triangleright \ldots \triangleright m_r$? The answer to this question, which follows from the following assertion proved in [13], is that if $K_1, K_2, \ldots, K_r$ meet RIP then $m_1 \triangleright m_2 \triangleright \ldots \triangleright m_r$ is decomposable.

Proposition 6 *For any sequence* $m_1, m_2, \ldots, m_r$ *of basic assignments defined on* $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}, \ldots, \mathbf{X}_{K_r}$, *respectively, the sequence* $\bar{m}_1, \bar{m}_2, \ldots, \bar{m}_r$ *computed by the following process*

$$\begin{aligned}
\bar{m}_1 &= m_1, \\
\bar{m}_2 &= \bar{m}_1^{\downarrow K_2 \cap K_1} \triangleright m_2, \\
\bar{m}_3 &= (\bar{m}_1 \triangleright \bar{m}_2)^{\downarrow K_3 \cap (K_1 \cup K_2)} \triangleright m_3, \\
&\vdots \\
\bar{m}_r &= (\bar{m}_1 \triangleright \ldots \triangleright \bar{m}_{r-1})^{\downarrow K_r \cap (K_1 \cup \ldots K_{r-1})} \triangleright m_r,
\end{aligned}$$

has the following properties: $m_1 \triangleright \ldots \triangleright m_r = \bar{m}_1 \triangleright \ldots \triangleright \bar{m}_r$; *each* $\bar{m}_i$ *is defined on* $\mathbf{X}_{K_i}$ *and is marginal to* $m_1 \triangleright \ldots \triangleright m_r$.

Remark 4 It is important to realise that if $K_1, K_2, \ldots, K_r$ meet RIP, then each $K_i \cap (K_1 \cup \ldots K_{i-1})$ is a subset of some K_j $(j < i)$ and therefore

$$(\bar{m}_1 \triangleright \ldots \triangleright \bar{m}_{i-1})^{\downarrow K_i \cap (K_1 \cup \ldots K_{i-1})} = \bar{m}_j^{\downarrow K_i \cap K_j}.$$

Therefore, from the computational point of view, the process described in Proposition 6 is simple for systems of low-dimensional assignments corresponding to decomposable graphs, and can be performed locally (see the next section).

Remark 5 Notice that, thanks to Proposition 3, one can deduce that for a decomposable basic assignment m it is possible to read the system of conditional independence relations valid for m exactly in the same way as it is done for decomposable probabilistic measures: If $G = (N, E)$ is a decomposable graph with respect to which decomposable basic assignment m factorises, and if nodes i and j are separated in G by set M then

$$i \perp\!\!\!\perp j \mid M \ [m].$$

However, let us stress once more: this possibility holds only if one accepts Definition 3.

4 Local computations

By local computations we understand a process based on the ideas published in the famous paper by Lauritzen and Spiegelhalter [17]: the considered probabilistic model (Bayesian network) was first converted into a decomposable model which was subsequently used to compute the required conditional probabilities. What is important in the latter part of the process is the fact that when computing the required conditional probability, one performs computations only on the system of marginal distributions defining the decomposable model. During the computational process one does not need to store more data than what is necessary to store for the decomposable model.

In this section we assume that the considered basic assignment is decomposable, i.e.,

$$m = m^{\downarrow K_1} \triangleright m^{\downarrow K_2} \triangleright \ldots \triangleright m^{\downarrow K_r},$$

and $K_1, K_2, \ldots, K_r$ meet RIP. So let us turn our attention to answering a question: What type of computation will correspond to determination of conditional probability?

Consider the simplest possible case. Assume the goal is to compute a one-dimensional marginal basic assignment for variable X_d in a case where we know that the value of variable X_e equals a $(d, e \in K_1 \cup \ldots \cup K_r)$. If we denote by ${}^a_e m$ the basic assignment on $\mathbf{X}_e$ with just one focal element ${}^a_e m(\{a\}) = 1$, then composition ${}^a_e m \triangleright m$ is a basic assignment describing the situation when one knows that $X_e = a$. Therefore, the goal mentioned above is achieved by computation of $({}^a_e m \triangleright m)^{\downarrow \{d\}}$.

Now, we are going to study the possibility of computing

$$({}^{a}_{e}m \triangleright m)^{\downarrow\{d\}} = ({}^{a}_{e}m \triangleright (m^{\downarrow K_1} \triangleright m^{\downarrow K_2} \triangleright \ldots \triangleright m^{\downarrow K_r}))^{\downarrow\{d\}}$$

locally. When evaluating $({}^{a}_{e}m \triangleright m)^{\downarrow\{d\}}$ we take full advantage of the assumption that m is decomposable, but, unfortunately, we also have to assume that $\{a\}$ is a focal element of $(m)^{\downarrow\{e\}}$, i.e., $(m)^{\downarrow\{e\}}(\{a\}) > 0$.

Namely, under these assumptions we can make the following consideration:

Having a decomposable model, we can find a permutation of the considered index sets $K_1, K_2, \ldots, K_r$ such that it meets RIP and the sequence starts with any of the sets containing the index e. Without loss of generality, let it be the sequence $K_1, K_2, \ldots, K_r$ (so, $K_1, K_2, \ldots, K_r$ meet RIP and $e \in K_1$). Then we can apply Proposition 2 because $\{e\} \cap K_r \subseteq K_1 \cup \ldots \cup K_{r-1}$ (recall that we selected the ordering such that $e \in K_1$) and

$$(m)^{\downarrow\{e\}}(\{a\}) > 0,$$

from which we get

$$\begin{aligned} &{}^{a}_{e}m \triangleright m \\ &\quad = {}^{a}_{e}m \triangleright ((m^{\downarrow K_1} \triangleright m^{\downarrow K_2} \triangleright \ldots \triangleright m^{\downarrow K_{r-1}}) \triangleright m^{\downarrow K_r}) \\ &\quad = ({}^{a}_{e}m \triangleright (m^{\downarrow K_1} \triangleright m^{\downarrow K_2} \triangleright \ldots \triangleright m^{\downarrow K_{r-1}})) \triangleright m^{\downarrow K_r}. \end{aligned}$$

However, in the same way we also get

$$\begin{aligned} &{}^{a}_{e}m \triangleright (m^{\downarrow K_1} \triangleright m^{\downarrow K_2} \triangleright \ldots \triangleright m^{\downarrow K_{r-1}}) \\ &\quad = ({}^{a}_{e}m \triangleright (m^{\downarrow K_1} \triangleright m^{\downarrow K2} \triangleright \ldots \triangleright m^{\downarrow K_{r-2}})) \triangleright m^{\downarrow K_{r-1}}, \end{aligned}$$

and after applying Proposition 2 $r-1$ times we get

$${}^{a}_{e}m \triangleright m = {}^{a}_{e}m \triangleright m^{\downarrow K_1} \triangleright m^{\downarrow K_2} \triangleright \ldots \triangleright m^{\downarrow K_{r-1}} \triangleright m^{\downarrow K_r}.$$

So we have shown that if m is a decomposable basic assignment and $(m)^{\downarrow\{e\}}(\{a\}) > 0$, then $({}^{a}_{e}m \triangleright m)^{\downarrow\{d\}}$ can always be computed locally in two steps:

- first order the respective K_i's in the way that they meet RIP and the first K_1 contains index e, and then
- apply Proposition 6 to the decomposable model

  $$({}^{a}_{e}m \triangleright m^{\downarrow K_1}) \triangleright m^{\downarrow K_2} \triangleright \ldots \triangleright m^{\downarrow K_r}$$

  receiving

  $$\begin{aligned} \bar{m}_1 &= {}^{a}_{e}m \triangleright m^{\downarrow K_1}, \\ \bar{m}_2 &= \bar{m}_1^{\downarrow K_2 \cap K_1} \triangleright m^{\downarrow K_2}, \end{aligned}$$

focal elements	$m_1(X_1, X_2)$
$\{a_1a_2, a_1\bar{a}_2\}$	$\frac{1}{4}$
$\{a_1\bar{a}_2, \bar{a}_1\bar{a}_2\}$	$\frac{1}{4}$
$\{a_1a_2, a_1\bar{a}_2, \bar{a}_1a_2\}$	$\frac{1}{2}$
	$m_2(X_2, X_3)$
$\{a_2a_3\}$	$\frac{1}{4}$
$\{\bar{a}_2, a_3\}$	$\frac{1}{4}$
$\{a_2\bar{a}_3, \bar{a}_2\bar{a}_3\}$	$\frac{1}{4}$
$\{a_2\bar{a}_3, \bar{a}_2a_3\}$	$\frac{1}{4}$
	$m_3(X_3, X_4)$
$\{a_3a_4\}$	$\frac{1}{2}$
$\{a_3a_4, \bar{a}_3\bar{a}_4\}$	$\frac{1}{4}$
$\{\bar{a}_3a_4, \bar{a}_3\bar{a}_4\}$	$\frac{1}{4}$

Table 3: Basic assignments m_1, m_2, m_3

$$\begin{aligned} \bar{m}_3 &= (\bar{m}_1 \triangleright \bar{m}_2)^{\downarrow K_3 \cap (K_1 \cup K_2)} \triangleright m^{\downarrow K_3}, \\ &\vdots \\ \bar{m}_r &= (\bar{m}_1 \triangleright \ldots \triangleright \bar{m}_{n-1})^{\downarrow K_n \cap (K_1 \cup \ldots K_{n-1})} \triangleright m^{\downarrow K_r}. \end{aligned}$$

Now we know that

$${}^{a}_{e}m \triangleright m = \bar{m}_1 \triangleright \bar{m}_2 \triangleright \ldots \triangleright \bar{m}_r,$$

each $\bar{m}_i$ is marginal to ${}^{a}_{e}m \triangleright m$, and therefore the required marginal basic assignment $({}^{a}_{e}m \triangleright m)^{\downarrow\{d\}}$ can be obtained by marginalisation of any m_i for which $d \in K_i$. Recall that, due to RIP, all the computations can be performed locally (see also Remark 4).

Example: Consider a 4-dimensional binary space $\mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3 \times \mathbf{X}_4$ with $\mathbf{X}_i = \{a_i, \bar{a}_i\}$, and three two-dimensional basic assignments whose all focal elements are given in Table 3. Let the goal be to compute $(m_1 \triangleright m_2 \triangleright m_3)^{\downarrow\{4\}}$ under the assumption that $X_1 = a_1$, i.e., we want to evaluate

$$({}^{a_1}_{1}m \triangleright (m_1 \triangleright m_2 \triangleright m_3))^{\downarrow\{4\}}.$$

Since X_1 is among the arguments of m_1, and $\{a_1\}$ is a focal element of $(m_1 \triangleright m_2 \triangleright m_3)^{\downarrow\{4\}}$, we can apply the above-introduced procedure (repeated application of Proposition 2) getting that

$$({}^{a_1}_{1}m \triangleright (m_1 \triangleright m_2 \triangleright m_3))^{\downarrow\{4\}} = ({}^{a_1}_{1}m \triangleright m_1 \triangleright m_2 \triangleright m_3)^{\downarrow\{4\}}.$$

So, it remains to apply the process described in Proposition 6. We get that ${}^{a_1}_{1}m \triangleright m_1$ has only one focal

element $(\{a_1a_2, a_1\bar{a}_2\})$, and therefore the same holds also for $({}^{a_1}_{1}m \triangleright m_1)^{\downarrow\{2\}}$: $({}^{a_1}_{1}m \triangleright m_1)^{\downarrow\{2\}}(\mathbf{X}_2) = 1$.

From this we immediately get $({}^{a_1}_{1}m \triangleright m_1)^{\downarrow\{2\}} \triangleright m_2$ with two focal elements

$$(({}^{a_1}_{1}m \triangleright m_1)^{\downarrow\{2\}} \triangleright m_2)(\mathbf{X}_2 \times \{\bar{a}_3\}) = \frac{1}{2}$$
$$(({}^{a_1}_{1}m \triangleright m_1)^{\downarrow\{2\}} \triangleright m_2)(\mathbf{X}_2 \times \mathbf{X}_3) = \frac{1}{2},$$

and therefore also its marginal $(({}^{a_1}_{1}m \triangleright m_1)^{\downarrow\{2\}} \triangleright m_2)^{\downarrow\{3\}}$, which is necessary for the computation of the next (already the last) composition, has two focal elements: $\{\bar{a}_3\}$ and $\mathbf{X}_3$. Evaluating this third composition we get that $(({}^{a_1}_{1}m \triangleright m_1)^{\downarrow\{2\}} \triangleright m_2)^{\downarrow\{3\}} \triangleright m_3$ has again two focal elements $\{a_3a_4, \bar{a}_3\bar{a}_4\}$ and $\{\bar{a}_3a_4, \bar{a}_3\bar{a}_4\}$; for each of them the computed composed basic assignment equals $\frac{1}{2}$. Marginalising the last two-dimensional basic assignment we get the desired result:

$$({}^{a_1}_{1}m \triangleright (m_1 \triangleright m_2 \triangleright m_3))^{\downarrow\{4\}}$$
$$= (((({}^{a_1}_{1}m \triangleright m_1)^{\downarrow\{2\}} \triangleright m_2)^{\downarrow\{3\}} \triangleright m_3)^{\downarrow\{4\}}$$

has only one focal element, namely

$$({}^{a_1}_{1}m \triangleright (m_1 \triangleright m_2 \triangleright m_3))^{\downarrow\{4\}})(\bar{a}_4) = 1.$$

Remark 6 If the goal is to compute a basic assignment for variable X_d under the condition that $X_e = a$ and simultaneously $X_f = b$, then one can first compute the decomposable model ${}^{a}_{e}m \triangleright m = \bar{m}_1 \triangleright \bar{m}_2 \triangleright \ldots \triangleright \bar{m}_r$ by the process described above, and afterwards

$${}^{b}_{f}m \triangleright ({}^{a}_{e}m \triangleright m) = {}^{b}_{f}m \triangleright (\bar{m}_1 \triangleright \bar{m}_2 \triangleright \ldots \triangleright \bar{m}_r)$$

in an analogous way finding a new permutation of $K_1, K_2, \ldots, K_r$ meeting RIP such that the first index set contains f. This time, naturally, we have to assume that $m^{\downarrow\{f\}}(\{b\}) > 0$, too.

5 Conclusions

Inspired by Graphical Markov Models in probability theory, we introduced decomposable models in Dempster-Shafer theory of evidence. For this we used two recently introduced concepts: operator of composition and factorisation.

Based on a *factorisation lemma* it is possible to deduce the fact that the introduced decomposable models possess the same conditional independence structure as their probabilistic counterparts; it can be read from the respective graphs following exactly the same rules as in the probabilistic case. This, however, holds only under the assumption that we accept the definition of conditional independence as presented here in Definition 3. Recall that our papers are not the only ones showing evidence in favour of this definition. As it was already presented in [2], Studený showed that the concept of conditional independence based on application of the conjunctive combination rule is not *consistent with marginalisation*. He found two consistent basic assignments for which there does not exist a common extension manifesting the respective conditional independence (for more details and Studený's example see [2]). Let us stress here once more that Definition 3 does not suffer from this insufficiency.

Nevertheless, it was not the main goal of this paper to support the new concept of conditional independence. Here we dealt with the question of whether the ideas of local computations can also be applied to computations in Dempster-Shafer theory of evidence. At this time we have, unfortunately, obtained only a partial answer. The results presented in the last section show that we are able to theoretically support local computations in the cases when the associativity of the operator of composition holds. We did it under the additional assumption that $m^{\downarrow e}(\{a\}) > 0$, i.e., under the assumption that

$$Bel(X_e = a) = m^{\downarrow e}(\{a\}) > 0.$$

From the point of view of real-world application, we would prefer if the designed computational process were applicable under a weaker condition, for example, in a case where

$$Pl(X_e = a) = \sum_{A \subseteq \mathbf{X}_e : a \in A} m^{\downarrow e}(A) > 0.$$

However, as we showed in Example in Section 2, this condition does not guarantee the associativity of the operator of composition. Therefore, there remains an open problem for the further research: either to show that the proposed (or similar) computational process corresponding to local computations can be performed without the assumption of associativity, or to modify the definition of the operator of composition (here we have in mind modification of case [b] of Definition 1) so that associativity would be valid under weaker conditions.

Acknowledgements

This work was supported by GAČR under the grants no. ICC/08/E010, and 201/09/1891, and by MŠMT ČR under grants 1M0572 and 2C06019.

References

[1] B. Ben Yaghlane, Ph. Smets, and K. Mellouli, "Belief Function Independence: I. The Marginal Case," *Int. J. of Approximate Reasoning*, vol. 29, no. 1, pp. 47–70, 2002.

[2] 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.

[3] I. Couso, S. Moral and P. Walley, "Examples of independence for imprecise probabilities," in *Proceedings of ISIPTA'99,* G. de Cooman, F. G. Cozman, S. Moral, P. Walley, Eds., Ghent, 1999, pp. 121–130.

[4] J. N. Daroch, S. Lauritzen and T. P. Speed, "Markov Fields and Log Linear Interaction Models for Contingency Tables," *Ann. Stat.* vo. 8, pp. 522-539, 1980.

[5] A. Dempster, "Upper and lower probabilities induced by a multi-valued mapping," *Annals of Math. Statistics* vol. 38, pp. 325–339, 1967.

[6] D. E. Edwards and T. Havránek, "A Fast Procedure for Model Search in Multidimensional Contingency Tables," *Biometrika*, vol. 72, no. 2, pp. 339-351, 1985.

[7] F. V. Jensen, *Bayesian Networks and Decision Graphs.* IEEE Computer Society Press, New York, 2001.

[8] 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.

[9] R. Jiroušek, "On a conditional irrelevance relation for belief functions based on the operator of composition," in *Dynamics of Knowledge and Belief, Proceedings of the Workshop at the 30th Annual German Conf. on Artificial Intelligence,* Ch. Beierle, G. Kern-Isberner, Eds., Fern Universität in Hagen, Osnabrück, 2007, pp. 28-41.

[10] R. Jiroušek, "Factorization and Decomposable Models in Dempster-Shafer Theory of Evidence," in *Proceedings of the Workshop on Theory of Belief Functions,* Brest, 2010.

[11] R. Jiroušek, "Is It Possible to Define Graphical Models in Dempster-Shafer Theory of Evidence?" in: *Proceedings of the 13th Int. Workshop on Non-Monotonic Reasoning,* Toronto, 2010.

[12] R. Jiroušek, "An Attempt to Define Graphical Models in Dempster-Shafer Theory of Evidence," in: *Proceedings of the 5th International Conference on Soft Methods in Probability and Statistics,* 2010, pp. 361–368.

[13] R. Jiroušek and J. Vejnarová, "Compositional models and conditional independence in evidence theory," *Int. J. Approx. Reasoning*, **52** (2011), 3, pp. 316–334.

[14] R. Jiroušek, J. Vejnarová and M. Daniel, "Compositional models of belief functions," in *Proc. of the 5th Symp. on Imprecise Probabilities and Their Applications,* G. de Cooman, J. Vejnarová, M. Zaffalon, Eds., Praha, 2007, pp. 243–252.

[15] G. J. Klir, *Uncertainty and Information. Foundations of Generalized Information Theory.* Wiley, Hoboken, 2006.

[16] S. L. Lauritzen, *Graphical models.* Oxford University Press, 1996.

[17] Lauritzen S. L. and Spiegelhalter D. J., Local computation with probabilities on graphical structures and their application to expert systems, *Journal of the Royal Statistical Society series B* **50** (1988), pp. 157–224.

[18] G. Shafer, *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, New Jersey, 1976.

[19] P. P. Shenoy, "Conditional independence in valuation-based systems," *Int. J. of Approximate Reasoning*, vol. 10, no. 3, pp. 203–234, 1994.

[20] P. P. Shenoy, "Binary join trees for computing marginals in the Shenoy-Shafer architecture," *Int. J. of Approximate Reasoning*, vol. 17, no. (2-3), pp. 239–263, 1997.

[21] M. Studený, "Formal properties of conditional independence in different calculi of AI," in *Proceedings of European Conference on Symbolic and quantitative Approaches to Reasoning and Uncertainty ECSQARU'93*, K. Clarke, R. Kruse, S. Moral, Eds., location and date, Springer-Verlag, 1993, pp. 341–351.

[22] M. Studený, "On stochastic conditional independence: the problems of characterization and description," *Annals of Mathematics and Artificial Intelligence,* vol. 35, p. 323-341, 2002.

[23] J. Vejnarová, "On conditional independence in evidence theory," in *Proc. of the 6th Symposium on Imprecise Probability: Theories and Applications,* Durham, UK, 2009, pp. 431–440.

Overcoming some limitations of imprecise reliability models

Igor Kozine
Technical University of Denmark
Kongens Lyngby
igko@man.dtu.dk

Victor Krymsky
State Academy of Economics and Service
Ufa; Russia
vikrymsky@mail.ru

Abstract

The application of imprecise reliability models is often hindered by the rapid growth in imprecision that occurs when many components constitute a system and by the fact that time to failure is bounded from above. The latter results in the necessity to explicitly introduce an upper bound on time to failure which is in reality a rather arbitrary value. The practical meaning of the models of this kind is brought to question. We suggest an approach that overcomes the issue of having to impose an upper bound on time to failure and makes the calculated lower and upper reliability measures more precise. The main assumption consists in that failure rate is bounded. Lagrange method is used to solve the non-linear program. Finally, an example is provided.

Keywords. Imprecise reliability, variational calculus, bounded failure rate.

1 Introduction

The appropriate incorporation of uncertainty into reliability and risk analyses is a topic of importance and widespread interest. Perhaps the most widely recognised distinction in uncertainty types is between aleatory and epistemic uncertainty and the presence of these two in the analyses of complex systems is a challenge systems analysts face. To address it, a number of mathematical structures able to capture the both types have been developed. The reader can find good overviews of the methods of uncertainty representation in different sources, for example, in [1] – [4]. Some of the mathematical structures are based on the two simple notions: interval-valued probabilities and imprecisely specified probability distributions. These structures are interval probability, probability bound analysis, Dempster-Shafer theory, robust Bayes methods, and the theory of imprecise probabilities that can be considered as the most general approach. The theory of imprecise probabilities, as it was introduced in [1] and [5], has served as the theoretical basis for generalising a large number of reliability models to imprecise probabilities. For a brief overview see [6]. More specifically, the reliability models of non-reparable systems of general structures (series, parallel and complex connection) generalised to imprecise probabilities are presented in [7], generalised discrete Markov chains used to model repairable systems are described in [8] and [9], stress-strength models for structural reliability are reported in [10] and [11]. The theory of imprecise probabilities has been applied to other important issues for reliability and risk analyses like aggregation of imprecise data having' different degrees of confidence to different pieces of evidence, expert judgement elicitation procedures, and decision making based on imprecise probabilities.

In spite of the seemingly rich arsenal of applied models built on imprecise statistical reasoning, they are nevertheless hesitantly used in practice and remain firmly in the academic realm. Do they lack adequate promotion by their practitioners, or are there other primary obstacles that prevent them from being widely applied? In [12] the authors' belief was that the main obstacle to the practical application of this knowledge is a tangible imprecision in lower and upper probability bounds constructed from a set of imprecise probabilistic pieces of evidence or/and the rapid growth in imprecision that occurs when intervals are propagated through mathematical models. The main cause in mathematical terms of the tangible imprecision was arguably identified as lying in the main mechanism of constructing coherent imprecise probability measures, which was originally called by Walley natural extension [1], and which in fact is a linear program. The crux of this linear program is that the solutions obtained are defined on the family of degenerate probability distributions[1], which are included on equal footing in the set of all admissible probability distributions over which the solution is sought. As proven in [13], solving this optimisation problem on the set of all admissible probability distributions gives the same solution as that obtained on only the set of degenerate distributions. This would simply be

[1] The probability distribution of a continuous random variable is referred to as degenerate if the probability masses are concentrated in a finite number of points belonging to the continuous set of possible states.

mathematical subtlety – that is, of little interest to practitioners – if it did not give us a clue to deriving more precise previsions of interest for continuous random variables. For some variables it is often not realistic to assume that the probability masses are concentrated in a few points as opposed to being continuously distributed over the set of possible outcomes. In reliability applications probability masses of time to failure cannot (except for very special cases) concentrate in a very few points of the positive real line. Ignoring this fact is one of the causes (we hold it to be the root cause) of high imprecision in reliability as well as in other applications. Or at least this is where some improvements are possible.

Several attempts have been undertaken to introduce some extra judgements to the set of constraints of the natural extension to limit the set of admissible probability distributions on which a solution is sought. That is, the desire is to remove from the admissible set the distributions that are obviously do not provide a reasonable model of the underlying random values like time to failure.

An attempt to mitigate the influence of degenerate probability distributions on the solutions was undertaken in [14]. No significant effect was obtained through the introduction of judgements on the skewness and unimodality of the distributions as, in this case, the peaks of degenerate distributions simply become repositioned and probability masses become redistributed among the peaks. The nature of the distributions defining the solutions remains unchanged.

Another approach was suggested in [15]. It consists in employing the calculus of variations to solve the optimisation problems instead of attempting to solve them with linear programming techniques. As it was demonstrated in [15] and then in [12] and [16] this way enables us to utilise a broader spectrum of statistical judgements, which results in tighter bounds on probability measures. The introduction of direct constraints on probability distributions like an upper bound on a probability density function (pdf) or/and on the absolute value of its derivative turned to be especially efficient. This type of constraints is not possible to utilise if the conventional natural extension in the form of a linear program is used as a tool for construction of imprecise probability measures. Direct constraints on pdfs make the problem nonlinear that can be solved with variational calculus. The direct constraints result in good improvements in precision so that we can see room for even better improvements.

Despite the obvious improvements in the precision of the constructed measures there is yet one more obstacle on the way of applying the theory of imprecise probabilities to reliability calculations. This obstacle stems from the underlying constraint imposed on the values of random variables. The random variables are bounded and this feature has a pernicious consequence on imprecise reliability models. This consequence consists in having an upper bound on time to failure explicitly present in the reliability models. (The lower bound is present too but since it is equal to zero, seemingly it is not part of the models.) Why the consequence is so harmful? This is because the upper bound on time to failure of any systems cannot be known. That is to say, the imposed necessity to choose this bound makes the reliability measures rather arbitrary values, as the upper bound is not known. The only non-arbitrary and true assertion about the sample space of time to failure is that it stretches from zero to infinity. All conventional reliability models reside in this presupposition.

In this paper we continue to use the calculus of variations for constructing imprecise probability measures and we introduce constraints on failure rate. It has a double effect: better precision in the results and avoidance of the necessity to have the upper bound on time to failure.

2 Exhibiting imprecise reliability models with the troublesome parameter

Let us look at several reliability models generalised to imprecise probabilities. The notations used are the following: $\underline{a}_i$ and $\overline{a}_i$ are a lower and upper m-th moments of time to failure of an i-th component for $i \le n$, $\underline{A}$ and $\overline{A}$ are a lower and upper m-th moments of a system compounded of n components, and T is an upper bound of time to failure that is assumed the same for all components.

For a system with independent components connected in series from the reliability point of view the following results are valid [7]:

$$\underline{A} = \frac{1}{(T^{n-1})^m} \prod_{i=1}^{n} \underline{a}_i , \qquad \overline{A} = \min_{i=1,\dots,n} \overline{a}_i$$

If the components are connected in parallel, then [7]

$$\underline{A} = \max_{i=1,\dots,n} \underline{a}_i , \qquad \overline{A} = T - T \prod_{i=1}^{n} \left(1 - \frac{\overline{a}_i}{T}\right)$$

Consider a couple of more examples. Let K is an upper bound of the pdf of time to failure of a component and this is the only reliability data available. Then we have the following results for the mean time to failure $M(t)$ [12]:

$$\underline{M}(t) = \frac{1}{2K} , \qquad \overline{M}(t) = T - \frac{1}{2K}$$

If in addition to K a bound on the absolute value of the pdf's derivative L is known, then [16]

$$\underline{M}(t) = \frac{1}{2K} + \frac{K}{2L}, \qquad \overline{M}(t) = T - \frac{1}{2K} - \frac{K}{2L}$$

As seen from the above expressions, one of the bounds of the expected values is explicitly dependent on the upper bound of time to failure T. Assuming that $T \to \infty$ gives us a very imprecise result that in many cases is practically useless. The two interrelated issues - high imprecision and dependence on the upper bound of time to failure – have motivated us to attempt to find a better solution.

The following section suggests a new problem statement that - as it will be demonstrated further in this paper - results in improved solutions.

3 Problem statement

Let us formulate first a rather general problem of computing bounds $\underline{M}$ and $\overline{M}$ on the expected value of an arbitrary function $g(x)$ given the upper, $\overline{f}_i = \overline{M}(f_i(t))$, and lower, $\underline{f_i} = \underline{M}(f_i(t))$, bounds of the expected values of other arbitrary functions $f_i(t)$, $i \le n$. As a particular case, the expected values can be known precisely meaning that the bounds are equal to each other. If $f_i(t) = t$, the expected value is the first moment. If $f_i(t) = t^2$, the expected value is the second moment, etc. In case $f_i(t) = I_{[t_1,t_2]}(t)$, where $I_{[t_1,t_2]}(t)$ is an indicator function equal to 1 when $t \in [t_1, t_2]$, and equal to 0 otherwise, the expected value is the probability $Pr(t \in [t_1, t_2])$.

The problem is stated as follows:

$$\left.\begin{aligned} \underline{M}(g) &= \inf_{\{\rho(x)\}} \int_0^T g(x)\rho(x)dx \\ \overline{M}(g) &= \sup_{\{\rho(x)\}} \int_0^T g(x)\rho(x)dx \end{aligned}\right\} \quad (1)$$

subject to

$$\left.\begin{aligned} \underline{f_i} \le \int_0^T f_i(x)\rho(x)dx \le \overline{f_i}, \;\; i = 1,2,\ldots,n \\ \rho(x) \ge 0, \text{ and } \int_0^T \rho(x)dx = 1 \end{aligned}\right\} \quad (2)$$

where $\rho(x)$ is the pdf of a random variable x defined on $[0,T]$. Here the inf and sup are taken over the set $\{\rho(x)\}$ of all pdfs matching constraints (2). That is, each constraint in (2) is associated with a subset of $\{\rho(x)\}$, and the intersection of those subsets, if not empty, defines the solutions of the optimization problems (1)-(2). If some of the subsets of $\{\rho(x)\}$ become disjoint, the solution does not exist. It should be noted that problems (1)-(2) are linear and the dual optimization problems can be written for them. The primal optimisation problems (1)-(2) and their duals have served as the key tools to derive a number of imprecise reliability models (see, for example, [7], [8] and [14]). The results were explicitly dependent on the upper bound, T, imposed on the random variable time to failure, as it was demonstrated in the previous section.

This is namely problems (1)-(2) the solutions to which are defined on the family of degenerate probability distributions [13]. This finding was a point of departure for introducing constraints that rule out the degenerate distribution from the set of admissible ones. Being guided by this finding, tighter bounds for probability measures have been derived for several problem statements [12], [15], [16]. In this paper we seek to solve the more ambitious problem: obtaining tighter bounds for a constructed probability measure of interest and getting rid of the need to impose an upper bound, T, on time to failure.

Now we introduce some new constraints and reformulate problems (1)-(2). In the following we will think of the random variable t as time to failure.

The cumulative distribution function of time to failure takes the form

$$F(t) = \int_0^t \rho(x)dx$$

and the reliability function is $P(t) = 1 - F(t)$. According to its definition (see, for example, [17]) the failure rate is

$$\lambda(t) = \frac{\rho(t)}{P(t)},$$

from which $P(t) = exp\left(-\int_0^t \lambda(x)dx\right)$.

Denote $\int_0^t \lambda(t)dt = y(t)$, then $\lambda(t) = \frac{dy(t)}{dt} = y'(t)$

Based on the above formulas and introduced notation the expression for the pdf, $\rho(t)$, appears as follows

$$\rho(t) = P(t)\lambda(t) = y'(t)exp(-y(t)).$$

Assuming that the failure rate is bounded from below and above by $\underline{\lambda}$ and $\overline{\lambda}$, that is $\underline{\lambda} \le \lambda(t) = y'(t) \le \overline{\lambda}$ and considering the lower, $\underline{f}_i$, and upper, $\overline{f}_i$, bound on the expected value of random variable $f_i(t)$ known, the following optimisation problem can be formulated

$$\left.\begin{aligned} \underline{M}(g) &= \inf_{\{y(t)\}} \int_0^T g(t)y'(t)\exp\big(-y(t)\big)dt \\ \overline{M}(g) &= \sup_{\{y(t)\}} \int_0^T g(t)y'(t)\exp\big(-y(t)\big)dt \end{aligned}\right\} \quad (3)$$

subject to

$$\underline{f}_i \le \int_0^T f_i(t)y'(t)\exp\big(-y(t)\big)dt \le \overline{f}_i, \quad i \le n, \quad (4)$$

$$\int_0^T y'(t)\exp\big(-y(t)\big)dt = 1 \quad (5)$$

$$\underline{\lambda} \le y'(t) \le \overline{\lambda} \quad (6)$$

Problems (3)-(6) are nonlinear and in order to solve them we suggest employing the calculus of variations as it was done in [12], [15], and [16].

4 Solving the problem with the calculus of variations

Problems similar to (3)-(6) have to be modified slightly to make them amenable to the calculus of variations. The constraint $\underline{\lambda} \le y'(t) \le \overline{\lambda}$ can be rewritten as follows:

$$\begin{aligned} y'(t) - u^2(t) &= \underline{\lambda}, \\ y'(t) + v^2(t) &= \overline{\lambda}. \end{aligned} \quad (7)$$

Here $u(t)$, $v(t)$ are unknown real-valued functions.

The solution of problems (3) subject to constraints (4), (5) and (7) is based on the following theorem

Theorem. *If for any interval $\alpha \le t \le \beta$, $0 \le \alpha < \beta \le T$ and for any $h_0, h_1, \ldots, h_n \in \mathbf{R}$ it holds that*

$$g(t) \ne h_0 + \sum_{i=1}^{n} h_i f_i(t),$$

then the failure rate $\lambda(t) = y'(t)$, on which inf and sup are attained in problems (3) subject to constraints (4), (5) and (7), is a step-wise function which is equal either to $\underline{\lambda}$ or to $\overline{\lambda}$.

The proof of this theorem is given in the Appendix and the meaning of it is that $\lambda(t)$ cannot take any other values between $\underline{\lambda}$ and $\overline{\lambda}$ but only either $\underline{\lambda}$ or $\overline{\lambda}$. This statement has a direct influence on the pdf, $\rho(t)$, on which inf and sup are attained in problems (3). That is, the pdf consists of the pieces $\underline{\rho}(t) = p(t_0, \ldots, t_i) \cdot \underline{\lambda} \cdot exp\big(-\underline{\lambda}(t - t_i)\big)$, $t \ge t_i$ and $\overline{\rho}(t) = p(t_0, \ldots, t_{i-1}) \cdot \overline{\lambda} \cdot exp\big(-\overline{\lambda}(t - t_{i-1})\big)$, $t \ge t_{i-1}$ that switch at some instances $t_1, t_2, \ldots, t_i$. The term $p(t_0, \ldots, t_i)$ is interpreted as the probability of being free of failure until time instant t_i. The correspondence between $\underline{\lambda}$, $\overline{\lambda}$ and optimizing ρ(t) is shown in Fig. 1.

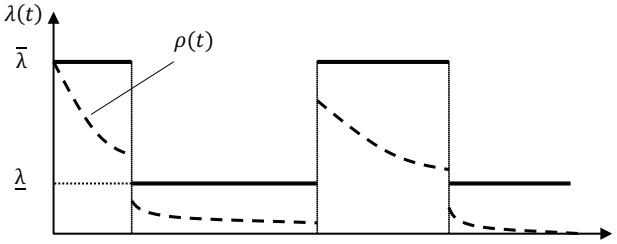


Figure 1. Optimizing pdf, $\boldsymbol{\rho(t)}$, and connected to it $\underline{\boldsymbol{\lambda}}$ and $\overline{\boldsymbol{\lambda}}$

Noticeable, the distribution of probability masses over time tends to zero when time tends to infinity. This in fact means that the very strong limiting requirement of imprecise probability theory that the random variable must be bounded is no longer valid and the "troublesome parameter" will not enter the expressions for reliability measures. It will be demonstrated in an example below.

As now the optimizing pdf is known (except for t_i) we can return to optimization problems (1)-(2) where $\rho(t)$ explicitly appears in the formulas. That what is not known now is the instances t_i when $\lambda(t)$, and consequently, $\rho(t)$ switch from one to the other value.

Assume that the optimal failure rate $\lambda(t)$ commutes $2m$ times between $\underline{\lambda}$ and $\overline{\lambda}$. That is,

$[0,t_1)$, $[t_2,t_3)$, $[t_4,t_5)$,...,$[t_{2j},t_{2j+1})$,...

are intervals of time in which $\lambda(t) = \overline{\lambda}$. Similarly,

$[t_1,t_2)$, $[t_3,t_4)$, $[t_5,t_6)$,...,$[t_{2j+1},t_{2j+2})$,... are the intervals on which $\lambda(t) = \underline{\lambda}$, $j \le m$.

Note that if $m = 0$, we have 2 intervals: one with the failure rate equal to $\overline{\lambda}$ and other with the failure rate equal to $\underline{\lambda}$. There may be some cases for which the optimizing failure rate for the whole time interval [0, T] is constant and equal either to $\overline{\lambda}$ or $\underline{\lambda}$.

The formula $\int_0^T f_i(x)\rho(x)dx$ for the expected value appearing in the constraints (2) as well as $\int_0^T g(t)\rho(t)dt$ appearing in (1) can now be rewritten

$$\Phi_i = \int_0^T f_i(x)\rho(x)dx =$$
$$\overline{\lambda}\left(\sum_{j=0}^{m} p(t_0,...,t_{2j}) \int_{t_{2j}}^{t_{2j+1}} f_i(t)\exp\left(-\overline{\lambda}(t-t_{2j})\right)dt\right) +$$
$$\underline{\lambda}\left(\sum_{j=1}^{m} p(t_0,...,t_{2j+1}) \int_{t_{2j+1}}^{t_{2j+2}} f_i(t)\exp\left(-\underline{\lambda}(t-t_{2j+1})\right)dt\right).$$

$$G = \int_0^T g(t)\rho(t)dt =$$
$$\overline{\lambda}\left(\sum_{j=0}^{m} p(t_0,...,t_{2j}) \int_{t_{2j}}^{t_{2j+1}} g(t)\exp\left(-\overline{\lambda}(t-t_{2j})\right)dt\right) +$$
$$\underline{\lambda}\left(\sum_{j=1}^{m} p(t_0,...,t_{2j+1}) \int_{t_{2j+1}}^{t_{2j+2}} g(t)\exp\left(-\underline{\lambda}(t-t_{2j+1})\right)dt\right).$$

$$R = \int_0^T \rho(t)dt = \overline{\lambda}\left(\sum_{j=0}^{m} p(t_0,...,t_{2j}) \int_{t_{2j}}^{t_{2j+1}} \exp\left(-\overline{\lambda}(t-t_{2j})\right)dt\right) +$$
$$\underline{\lambda}\left(\sum_{j=1}^{m} p(t_0,...,t_{2j+1}) \int_{t_{2j+1}}^{t_{2j+2}} \exp\left(-\underline{\lambda}(t-t_{2j+1})\right)dt\right) =$$
$$\sum_{j=0}^{m} p(t_0,...,t_{2j})\cdot\left(1-\exp(-\overline{\lambda}(t_{2j+1}-t_{2j}))\right) +$$
$$\sum_{j=0}^{m} p(t_0,...,t_{2j+1})\cdot\left(1-\exp(-\underline{\lambda}(t_{2j+2}-t_{2j+1}))\right).$$

Finally, the reformulated problem statement is as follows:

$$\min_{t_1,t_2,...,t_{2m+2}} G(t_1,t_2,...,t_{2m+2}) \text{ and}$$

$$\max_{t_1,t_2,...,t_{2m+2}} G(t_1,t_2,...,t_{2m+2})$$

subject to constraints

$$\underline{a}_i \le \Phi_i(t_1,t_2,...,t_{2m+2}) \le \overline{a}_i,\ i \le n, \text{ and}$$

$$R(t_1,t_2,...,t_{2m+2}) = 1.$$

This is rather an easy optimisation problem with algebraic constraints. Once one knows the number of intervals *m*, this optimization problem can be solved by using standard numerical techniques such as gradient methods, simplex-based search methods, genetic algorithms, etc. In simple cases, the solution can be obtained in an analytical form as it takes place in the example below.

The number of intervals in which the failure rate remains constant is a priori unknown. In the following we suggest an algorithm, similar to that introduced in [12] and [16], which solves this problem. We start with the verification if only one of the two $\overline{\lambda}$ or $\underline{\lambda}$ for the whole time period [0, *T*] satisfies the constraints. If the result is positive we can compute the value of the objective function. Then we set $m = 0$, solve the optimization problem and compare the obtained value of the objective function with the previous result. If it is different, we may continue and increase *m* by 1, and so on. The process will be stopped if the expression for the density function $\rho(t)$ does not change (or changes negligibly) and the improvement of the objective function also is not observed.

5 Example

Assume we are interested in knowing bounds $\underline{\tau}$ and $\overline{\tau}$ on the mean time to failure $\tau = \int_0^\infty t\rho(t)dt$ of a system and the following data (constraints) are known:

$$\Pr(q) = 1 - \int_0^\infty I_{[0,q]}(t)\rho(t)dt = p \text{ and } \underline{\lambda} \le \lambda(t) \le \overline{\lambda}.$$

That is, we know precisely the probability $\Pr(q)$, which we interpret as system's reliability at time *q*, and the lower $\underline{\lambda}$ and upper bound $\overline{\lambda}$ on the failure rate. $I_{[0,q]}(t)$ is the indicator function equal to 1 if $t \in [0,q]$ or equal to 0 otherwise. The consistency relation between the reliability and failure rate is expressed by the two inequalities $\exp(-\overline{\lambda}q) \le p \le \exp(-\underline{\lambda}q)$. If $\exp(-\overline{\lambda}q) = p$ or $p = \exp(-\underline{\lambda}q)$, the solution to the problem is simple, as there is only one pdf satisfying the either equality. The problem of this kind was described in [17]. This problem becomes more complicated if the strong inequalities hold $\exp(-\overline{\lambda}q) < p < \exp(-\underline{\lambda}q)$. For

this case, there are intervals on which the failure rate switches. Hence we start with $m = 0$. However, immediately it becomes clear that for $m = 0$ the expression for $\rho(t)$ contains only one unknown parameter t_1 while there are two constraints

$$\Pr(q) = 1 - \int_0^{\infty} I_{[0,q]}(t)\rho(t)dt = p\,,\ \int_0^{\infty} \rho(t)dt = 1.$$

This is why we have to increase *m* by 1

Determining $\underline{\tau}$. The graph of the pdf, $\rho(t)$, for which τ attains its minimum takes the form as shown in Fig. 2:

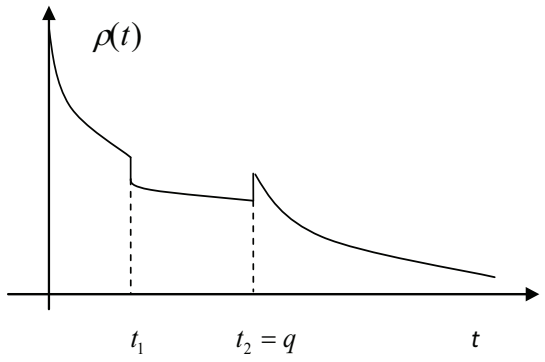


Figure 2. The behaviour of the pdf for which τ attains its minimum

Hence $\Pr(q) = \exp(-\overline{\lambda}t_1)\cdot\exp\left(-\underline{\lambda}(q-t_1)\right) = p$.

From this equation we obtain

$$t_1 = -\frac{1}{\overline{\lambda}-\underline{\lambda}}\ln\left(p\cdot\exp(\underline{\lambda}q)\right) = -\frac{1}{\overline{\lambda}-\underline{\lambda}}\left(\ln p + \underline{\lambda}q\right)$$

Now we compute the value of t_2. First, assume that $t_2 \neq q$ (e.g. $t_2 > q$). Then the following equation must hold:

$$1 - p + \underline{\lambda}p\int_q^{t_2}\left(\exp\left(-\underline{\lambda}(t-q)\right)\right)dt +$$

$$\overline{\lambda}p(1-\exp\left(-\underline{\lambda}(t_2-q)\right)\int_{t_2}^{\infty}\left(\exp\left(-\overline{\lambda}(t-t_2)\right)\right)dt = 1.$$

It is true if $t_2 = q$. Finally,

$$\underline{\tau} = \int_0^{\infty} P(t)dt = \int_0^{t_1}\exp(-\overline{\lambda}t)dt +$$

$$\exp(-\overline{\lambda}t_1)\cdot\int_{t_1}^{q}\exp\left(-\underline{\lambda}(t-t_1)\right)dt + p\cdot\int_q^{\infty}\exp\left(-\overline{\lambda}(t-q)\right)dt =$$

$$\frac{1}{\overline{\lambda}}\left(1+p-\exp(-\overline{\lambda}t_1)\right) + \frac{1}{\underline{\lambda}}\left(1-\exp(-\underline{\lambda}(q-t_1))\right)\cdot\exp(-\overline{\lambda}t_1).$$

Increasing *m* by 1 does not lead to any improvement. Thus the obtained formula value is optimal one.

Determining $\overline{\tau}$. The graph of the pdf, $\rho(t)$, for which τ attains its maximum takes the form as shown in Fig. 3.

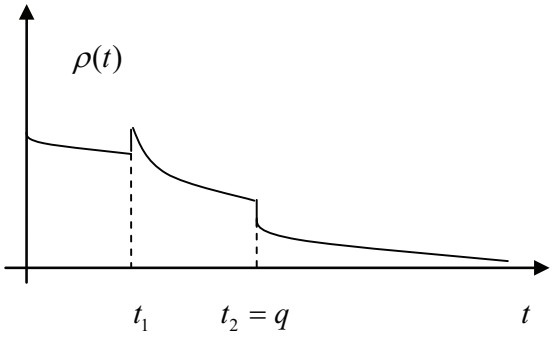


Figure 3. The behaviour of the pdf for which τ attains its maximum

For this case we can perform computations similar to the above and arrive at the result

$$\overline{\tau} = \frac{1}{\underline{\lambda}}\left(1+p-\exp(-\underline{\lambda}t_1)\right) + \frac{1}{\overline{\lambda}}\left(1-\exp(-\overline{\lambda}(q-t_1))\right)\cdot\exp(-\underline{\lambda}t_1).$$

6 Concluding notes

In spite of the existence of a number of risk/reliability and other applied models built on imprecise statistical reasoning, only a few of them have ever been used in practice – and then only hesitantly –, the rest remaining firmly in the academic realm. Perhaps the complexity of imprecise statistical reasoning as a whole is such as to severely limit the accessibility of this kind of models to potential practitioners. We nevertheless believe that the main obstacles to the practical application of this knowledge are different. One which is thoroughly familiar to the group of experts who practise interval computations and which we have repeatedly mentioned [12], [16]: it is namely the rapid growth in imprecision that occurs when intervals are propagated through mathematical models and when the number of components in a system is large. The other one stems

from the requirement of imprecise probability theory that the random value is to be bounded. This requirement appears very restrictive for reliability applications, as some reliability models explicitly contain an upper bound on time to failure which is in reality an arbitrary value.

Our main finding was that bounding the failure rate allows deriving reliability measures devoid of an upper bound on time to failure. That is, the sample space of time to failure is now as it must be from zero to infinity. This is the basic assumption on which all conventional reliability models rest and deviations from that can hardly be practical. Making judgements on the lower and upper bounds of failure rates is meaningful and can often be substantiated by observed events taking place in the system of interest or analogous ones.

Acknowledgements

This work was supported by the OECD Halden reactor project.

References

[1] Walley, P., 1991, *Statistical Reasoning with Imprecise Probabilities* (New York: Chapman and Hall).

[2] Alternative Representation of Epistemic Uncertainty, Edited by J.C. Helton and W.L. Oberkampf. *Reliability Engineering & System Safety*. Volume 86, Issues 1-3, pages 1-369

[3] Helton, J.C. et al. Representation of analysis results involving aleatory and epistemic uncertainty. *International Journal of General Systems*. Vol: 39, No: 6, 2010, pp. 605- 646

[4] Ferson, S., RAMAS Risk Calc: Risk Assessment with Uncertain Numbers. Lewis Publishers, 2002

[5] Kuznetsov, V., 1991, *Interval Statistical Models*. (Moscow: Radio and Sviaz, in Russian).

[6] Utkin, L.V. and Coolen, F., 2007. Imprecise Reliability: an Introductory Overview. In: *Computational Intelligence in Reliability Engineering, Vol. 2: New Metaheuristics, Neural and Fuzzy Techniques in Reliability*, Chapter 10, G. Levitin (Ed.), pp. 261-306 (Springer, 2007).

[7] Kozine, I. and Utkin, L.V., 2005, Computing System Reliability Given Interval-Valued Characteristics of the Components. *Reliable Computing,* **11**, pp. 19–34.

[8] Kozine, I. and Utkin, L.V., 2002, Interval-Valued Finite Markov Chains. *Reliable Computing*, **8**, pp. 97–113.

[9] Skujl, D. Discrete time Markov chains with interval probabilities. *Int. Journal of Approximate Reasoning*, V. 50, Issue 8, 2009

[10] Utkin, L. and Kozine, I., 2002, Stress-Strength Reliability Models under Incomplete Information. *International Journal of General Systems*, **31**, pp. 549-568.

[11] Utkin, L. and Kozine, I., On new cautious structural reliability models in the framework of imprecise probabilities. Structural Safety 32 (2010) 411-416

[12] Kozine, I. and Krymsky V. Computing interval-valued statistical characteristics: what is the stumbling block for reliability applications? Int. Journal of General Systems. 38:5,547 — 565, 2008

[13] Utkin, L. and Kozine, I., 2001, Different Faces of the Natural Extension. In *Proc. of the Second Intern. Symp. on Imprecise Probabilities and Their Applications - ISIPTA '01*, G. De Cooman, T. Fine and T. Seidenfeld (Eds.), pp.316-323 (Ithaca, NY, USA: Shaker, 2001).

[14] Utkin, L., 2002, Imprecise Calculation with the Qualitative Information about Probability Distributions. In *Proc. of the Conf. on Soft Methods in Probability and Statistics*, P. Grzegorzewski, O. Hryniewicz and M.A. Gil (Eds.), pp. 164-169 (Heidelberg, New York: Phisica-Verlag, 2002),.

[15] Kozine, I., Krymsky, V., 2007. Enhancement of natural extension. *Proceedings. 5. International symposium on imprecise probability*: *Theories and applications (ISIPTA '07), Prague*, 16-19 Jul 2007. Cooman, G. de; Vejnarova, J.; Zaffalon, M. (eds.), Action M Agency, Prague, 253-262

[16] Kozine, I., Krymsky, V., Bounded Densities and Their Derivatives: Extension to Other Domains. Journal of Statistical Theory and Practice, V. 3, No. 1, pp. 25-38, 2009

[17] Barlow, R.E. and Proshan, F., 1975, *Statistical Theory of Reliability and Life Testing: Probability Models* (New York: Holt, Rinehart and Winston).

[18] Gelfand, N.M. and Fomin, S.V., 2000, *Calculus of Variations* (New York: Dover Pubns).

Appendix

Theorem. *If for any interval* $\alpha \le t \le \beta$, $0 \le \alpha < \beta \le T$ *and for any* $h_0, h_1, ..., h_n \in \mathbf{R}$ *it holds that*

$$g(t) \ne h_0 + \sum_{i=1}^{n} h_i f_i(t), \tag{8}$$

then the failure rate $\lambda(t)$*, on which inf and sup are attained in problems (3) subject to constraints (4), (5) and (7), is a step-wise function which is equal either to* $\underline{\lambda}$ *or to* $\overline{\lambda}$.

Proof. According to the method of Lagrange [18] the primal form of optimization problem (3) subject to constraints (4), (5) and (7) is to be replaced by the equivalent unconstrained optimization problem. To do so the following function is introduced

$$J^*(t) = g(t)y'(t)\exp\big(-y(t)\big) +$$
$$\sum_{i=1}^{n} \mu_i f_i(t) y'(t)\exp\big(-y(t)\big) + \mu_0 y'(t)\exp\big(-y(t)\big) +$$
$$\mu^*(t)\Big(y'(t) - u^2(t)\Big) + \mu^{**}(t)\Big(y'(t) + v^2(t)\Big)$$

Where $\mu, i \le n$ and $\mu^*(t)$, $\mu^{**}(t)$ are unknown Lagrange multipliers.

Then the Euler-Lagrange equations (the necessary condition of optimality) take the form:

$$\frac{\partial J^*}{\partial y} - \frac{d}{dt}\left(\frac{\partial J^*}{\partial y'}\right) = 0; \qquad \frac{\partial J^*}{\partial u} = 0; \qquad \frac{\partial J^*}{\partial v} = 0.$$

In our case these equations become:

$$-g(t)y'(t)\exp\big(-y(t)\big) - \sum_{i=1}^{n} \mu_i f_i(t) y'(t)\exp\big(-y(t)\big)$$
$$-\mu_0 y'(t)\exp\big(-y(t)\big) + g(t)y'(t)\exp\big(-y(t)\big) +$$
$$\sum_{i=1}^{n} \mu_i f_i(t) y'(t)\exp\big(-y(t)\big) + \mu_0 y'(t)\exp\big(-y(t)\big) -$$
$$g'(t)\exp\big(-y(t)\big) - \sum_{i=1}^{n} \mu_i f_i'(t)\exp\big(-y(t)\big) +$$
$$d\Big(\mu^*(t)\Big)/\,dt + d\Big(\mu^{**}(t)\Big)/\,dt = 0$$
$$\mu^*(t)u(t) = 0 \text{ and } \mu^{**}(t)v(t) = 0$$

Here

$$g'(t) = dg(t)/dt;\ f_i'(t) = df_i(t)/dt, i \le n;$$
$$y''(t) = d^2 y(t)/dt^2.$$

It can be concluded that if $u(t) \ne 0$ and $v(t) \ne 0$ simultaneously then $\mu^*(t) = \mu^{**}(t) = 0$. Hence $d\Big(\mu^*(t)\Big)/\,dt = 0$ and $d\Big(\mu^{**}(t)\Big)/\,dt = 0$ resulting in

$$g'(t) + \sum_{i=1}^{n} \mu_i f_i'(t) = 0,$$

or after integration

$$g(t) + \sum_{i=1}^{n} \mu_i f_i(t) + c = 0, \tag{9}$$

in which c is arbitrary constant. (9) contradicts to (8). To resolve this conflict, one of the functions $u(t)$, $v(t)$ must be equal to zero inside the interval $\alpha \le t \le \beta$. On the other hand, they cannot be both equal to zero because the equalities $\lambda(t) = \underline{\lambda}$ and $\lambda(t) = \overline{\lambda}$ cannot hold simultaneously.

Finally, we conclude that the failure rate alternates between $\underline{\lambda}$ and $\overline{\lambda}$ within the time period $[0,T]$.

A study on updating belief functions for parameter uncertainty representation in Nuclear Probabilistic Risk Assessment

Tu Duong Le Duy,
Dominique Vasseur, Mathieu Couplet
Electricity of France R&D
Risk Management Department MRI
Clamart cedex, France
tu-duong.le-duy@edf.fr;
dominique.vasseur@edf.fr; mathieu.couplet@edf.fr

Laurence Dieulle,
Christophe Bérenguer
University of Technology of Troyes
Institut Charles Delaunay/LM2S, UMR STMR
Troyes Cedex, France
laurence.dieulle@utt.fr;
christophe.berenguer@utt.fr

Abstract

Probabilistic Risk Assessments (PRA) are used to achieve a safe design and operation of Nuclear Power Plants. The impact of uncertainties which may affect PRA results must thus be taken into account in the decision making process. These uncertainties due to the lack of data have been recently seen as mainly epistemic ones and it has been recommended to characterize them by the belief functions of Dempster-Shafer Theory rather than a presumed single probability distribution. The current construction of these functions is based on the data provided by PRA data handbooks using traditional statistical tools like Maximum Likelihood Estimation (MLE). However, this approach is only appropriate when data coming from the operating feedback observations are sufficiently large as required in the MLE approach. Furthermore, when wishing to incorporate other sources of information, such as expert's opinions, the pooling data of MLE has limits to account for these kinds of information. Therefore, in order to overcome this problem, two alternative perspectives based on the Dempster's rule of combination and the Generalized Bayesian Theorem for constructing and updating the belief functions in a more effective way will be presented in this paper. These two approaches will be studied for the use in the context of PRA. The comparison of these two approaches with the current method is carried out through a practical example. Some conclusions about the application of these approaches will be drawn.

Keywords. Parameter uncertainty, belief functions, generalized Bayesian theorem, nuclear risk assessment.

1 Introduction

Probabilistic Risk Assessment (PRA) [10] is a methodology which provides a quantitative assessment of the risk of accidents at Nuclear Power Plants (NPP). It involves the development of models that delineate the response of systems and of operators to initiating events that could lead to core damage or a release of radioactivity to the environment. The evaluation of the frequency of such an accident relies on the assessment of the failure probability of systems by means of event/fault trees. In PRA, parametric statistical models are used to characterize the random occurrence of accidents at nuclear power plants [2][10]. Some usual parametric models like Poisson model, exponential model…are used for this purpose. The parameters associated to these models in PRA are reliability parameters such as the failure rates of individual components or the probability of failure on demand and so on. The values of these parameters are generally unknown and estimated with statistical tools. These estimated values are therefore subjected to uncertainty due to insufficient feedback data which can impact the decision making process. As a consequence, the results in the nuclear PRA context for decision making need to take into account these uncertainties.

In the traditional PRA practice of uncertainty analysis, the epistemic parameter uncertainty is generally represented by a presumed probability distribution, such as the log-normal distribution which is viewed as the subjective interpretation of probability (i.e. degree of belief) for the possible values of the parameter. Nevertheless, the choice of this distribution which is made for some practical reasons has been shown to be questionable because it could have major impacts on the final results of decision making [20]. Recently, a general framework of parameter uncertainty quantification within the Dempster-Shafer Theory (DST) framework has been proposed in the nuclear PRA context [20][21]. In this framework, parameter uncertainty is no longer characterized by an assumed probability distribution but by belief and plausibility functions which represent the current state of knowledge about the possible values of the parameter. The approach proposed in [20] for the construction of these belief functions is based on the statistical data provided by EDF PRA data handbooks using traditional statistical tools such as Maximum Likelihood Estimation (MLE). Therefore, when new data become available, statistical tools are first used to

provide estimated values from the pooled data (e.g. nominal values and confidence intervals) from which belief functions for uncertainty representation are constructed. However, this approach is only appropriate to the case where data come from the operating feedback observations and when the number of observations is sufficiently large. Furthermore, if additional sources of information are to be incorporated, such as expert's opinions, the pooling data of MLE has limits to account for these kinds of information. The expert's opinions are often used in the context of PRA model for the events whose the frequency of occurrence is very small i.e. rarely or never observed. Therefore, in order to incorporate the experts' opinions with the available operating feedback data, two alternative perspectives for constructing belief functions in a more effective way are studied in this paper. The updated belief functions are built by combining the belief functions given each data. In doing so, the incorporation of other sources of information, such as expert's opinions will be done in a natural manner. The two proposed approaches also allow us to deal with the prior ignorance in a more appropriate manner than the classical way. In the first approach, we still use the MLE but in a different way. For each independent serie of observations, the belief functions are firstly built from the confidence intervals provided by MLE, and then the updated belief functions are obtained by using the Dempster's rule of combination (ROC) to aggregate all the belief functions. In the same manner but within the perspective of Bayesian theorem, the second approach relies on the General Bayesian Theorem (GBT) to provide belief functions given each data. The GBT introduced by Smets in [13] performs the same task as the classical Bayesian theorem but within the context of belief functions instead of probability functions. This theorem and the pignistic transformation are the essential tools of the so-called Transferable Belief Model (TBM) which is a subjective interpretation of the DST [14]. The main objective of this paper is to study the use of these approaches for updating belief functions in the context of nuclear PRA data.

The section 2 of this article presents shortly basic notions of the Dempster–Shafer theory of belief functions. In section 3, the updating of belief functions with the ROC and the GBT is presented. The section 4 studies the application of these two approaches in the context of PRA. The comparison of these two approaches with the currently used method is carried out through a practical example in the section 5. In the section 6, some conclusions and perspective are finally given.

2 The Dempster-Shafer Theory of Belief Functions

The Dempster-Shafer Theory of evidence [6], also known as the theory of belief functions, is a generalization of the Bayesian theory of subjective probability in that it allows less restrictive assumptions about the likelihood than in the case of probabilistic characterization of uncertainty. In literature, this theory has been used in risk assessment for industrial applications [11][17][18] and recently studied in the context of PRA for treating the uncertainty [20][21]. In this framework, the epistemic uncertainty associated to the input parameters of PRA is no longer characterized by a single probability distribution but by so-called belief and plausibility functions. In doing so, we can avoid the problem of choosing an appropriate probability distribution for uncertainty representation in a context of lack of data. The definition of these functions is shortly outlined now.

Let $\Theta=\{\theta_1, \theta_2,\ldots,\theta_N\}$ be a finite set of possible values for parameter θ called the frame of discernment. Unlike the probability distribution which is completely defined by the weight of each singleton θ_i, the belief functions are defined on the set of subsets of Θ, called 2^Θ. In the DST, the basic measure is represented by a so-called basic belief assignment

$$m: 2^\Theta \to [0,1] \qquad \sum_{A\subseteq\Theta} m(A) = 1 \tag{1}$$

Where $m(\Theta)=1$ and $m(\emptyset)=0$. The basic belief assignment (BBA) m(A) represents the degree of belief that the actual solution is exactly committed to A and due to lack of knowledge cannot be attributed any more specific event. The state of complete ignorance is represented by the so-called vacuous BBA defined by $m(\Theta)=1$ that is no information is available for the more likely values among Θ. A Bayesian BBA is a BBA whose focal sets are singletons. A BBA is said to be consonant if its focal sets are nested.

The belief function *Bel*, the plausibility function *Pl* and the commonality function *q* are defined for all $B \subseteq \Theta$ as follows

$$Bel(B) = \sum_{A\subseteq B} m(A) \tag{2}$$

$$Pl(B) = \sum_{A\cap B\neq\emptyset} m(A) \tag{3}$$

$$q(B) = \sum_{A\supseteq B} m(A) \tag{4}$$

The belief $Bel(B)$ obtained by the summation of BBAs for all elements A which are fully included in proposition B expresses the "total" degree of belief. The degree of plausibility $Pl(B)$ is calculated by adding BBAs of elements A whose the intersection with proposition B is not an empty set. The commonality function q is used for mathematical purposes only. In the perspective of Walley [16], these belief and plausibility functions consist of lower and upper bounding probability functions of the true but unknown probability distribution.

When a decision needs to be made, we use a so-called pignistic[(1)] transformation which induces a pignistic probability function from the belief functions. This is the

[(1)] *Pignistic* means '*bet*' in Latin

result of applying the TBM model introduced by Smets [14] which is a subjective interpretation of the DST. The TBM is a two-level mental model in which the beliefs are represented and quantified at the credal level by belief functions, whereas decision making is based on the probability distributions and takes place at the pignistic level. The use of the TBM model for decision making in the context of PRA has been studied in [21].

In the next sections, the Dempster-Shafer Theory is studied for the use of updating the belief functions when new evidence is available.

3 Approaches for updating belief within the Theory of Belief functions.

Combination of different sources of evidence is one of the important fields when dealing with uncertainty. The Dempster-Shafer Theory of belief functions offers many approaches for aggregating belief functions in a natural way. Two approaches often studied and used in some real applications are outlined hereafter. These two approaches allow the belief functions to be updated by taking account of the prior sources of information (e.g. experts' opinions or previous data) in addition with new available data.

In the following we consider a random variable X on the state space Ψ and characterized by its probability distribution P_θ, with the parameter θ taking its values in Θ.

3.1 Dempster's Rule of Combination (ROC)

Suppose that the uncertainty associated to the parameter of the model is characterized by belief functions. These functions need to be updated when new data on the space Ψ become available. If data observations are independently collected, the belief functions of the parameter given each data can be all combined together using the Dempster's rule of combination (ROC). Let BBA m_1 and BBA m_2 represent respectively the belief functions given the first data and the second data over the frame Θ, according to the ROC, then the combined BBA is calculated as follows

$$m_{12}(A) = (m_1 \oplus m_2)(A)$$
$$= \frac{1}{1-K} \sum_{A_1 \cap A_2 = A} m_1(A_1).m_2(A_2) \text{ for } \forall A \subseteq \Theta \quad (5)$$

Where $K = \sum_{A_1 \cap A_2 = \emptyset} m_1(A_1).m_2(A_2)$ is a measure of the amount of conflict between the two BBAs.

Therefore, by considering m_1 as the prior BBA and m_2 as the BBA given new available data, the posterior belief functions can be obtained using the above ROC. In some contexts, the prior information can be simply vacuous belief functions i.e. $m(\Theta)=1$ which express the total ignorance.

As wee can see in equation (5), since the operator $\oplus$ used in this rule is both associative and commutative, thus the order of these functions to combine is not relevant. Note that when the belief functions are Bayesian functions, Shafer [7] proved that the Bayes' rule of conditioning is a special case of the Dempster's rule of combination.

3.2 Generalized Bayesian Theorem in TBM

The previous approach for aggregating the belief functions of the uncertain parameter θ involved a fairly standard application of DST. However, a generalization of the Bayes' rule within the TBM may be used to update the belief functions in a manner more closely aligned with updating of probability distributions via the classical Bayes' rule. This approach is now outlined.

3.2.1 Generalized Bayesian Theorem

As we know, in probability theory, the Bayesian theorem allows the computation of the posterior probability function of θ given observed realizations of X from the likelihood of X given θ and some prior probability distribution of θ. The same idea has been extended in the TBM context [13] where conditional belief functions of θ given observations of X is built from the conditional belief function of X given each $\theta_i \in \Theta$ and a vacuous prior belief of θ. Thus, if we know the conditional plausibilites $pl^{\Psi}(x|\theta_i)$ of X given each $\theta_i \in \Theta$ and according to the GBT, the conditional belief functions for all $A \subseteq \Theta$ given an observation $x \in \Psi$ are computed as follows:

$$m^{\Theta}(A|x) =$$
$$= C.\prod_{\theta_i \in A} pl^{\Psi}(x|\theta_i) \prod_{\theta_i \in \bar{A}} (1 - pl^{\Psi}(x|\theta_i)) \quad (6)$$

$$Bel^{\Theta}(A|x) =$$
$$= C\left(\prod_{\theta_i \in \bar{A}} (1 - pl^{\Psi}(x|\theta_i) - \prod_{\theta_i \in \Theta} (1 - pl^{\Psi}(x|\theta_i)) \right) \quad (7)$$

and

$$pl^{\Theta}(A|x) = C\left(1 - \prod_{\theta_i \in A} (1 - pl^{\Psi}(x|\theta_i)) \right) \quad (8)$$

Where $C^{-1} = 1 - \prod_{\theta_i \in \Theta} (1 - pl^{\Psi}(x|\theta_i))$ is the normalized factor which is introduced when the assumption of closed-world is made i.e. the BBA $m(\emptyset) = 0$ is assumed. The interesting point in the GBT is that the needed prior belief on Θ is a vacuous belief function which is the perfect representation of total

ignorance. We can thus avoid one of the delicate problems of classical Bayesian approach related to choosing an appropriate a priori. In the context of updating belief functions, the posterior beliefs can be obtained using the Dempster's rule of combination applied to the above conditional belief function given new data and the prior belief function built from the previous data.

In the case of having n independent series of observations with event counts x_1, x_2, … x_n resulting from the same probabilistic model (e.g. Poisson model), in order to aggregate belief functions given these observations, we can construct n conditional belief functions of θ given each event count x_i and then combine these belief functions by the ROC. The same result can be obtained in a different way by considering the joint conditional plausibility function $pl^{\Psi}(x_1,...,x_n \mid \theta_i)$ directly obtained from the joint observations $(x_1,x_2,\ldots,x_n)$ using the notion of "conditional cognitive independence" as proposed in [5][13]. As a result, the plausibility function $pl^{\Psi}(x_1,...,x_n \mid \theta_i)$ of observing the joint observation $(x_1, x_2 \ldots x_n)$ given each $\theta_i \in \Theta$ is the product of the individual plausibility functions of all observations i.e.:

$$pl^{\Psi}(x_1,...x_n \mid \theta_i) = \prod_{k=1}^{n} pl^{\Psi}(x_k \mid \theta_i) \qquad (9)$$

Then, the equations above (6,7,8) can be applied to calculate the conditional belief functions on Θ given the joint observation. This above property is essential and in fact the core of the axiomatic derivations of the GBT [12]. Let us now discuss about the performance of two ways for calculating the conditional belief functions given the data in GBT. From a computational point of view, the way of constructing conditional belief functions of θ given joint observations $(x_1, x_2 \ldots x_n)$ is more efficient than calculating the conditional belief functions of θ given each x_i and combing them by ROC. This is because the former way is simply involved in the "product" operations (9) while the later concern with the orthogonal sums of ROC which require practically much more computational time. However, if we have some other sources of information such as expert's judgments or any source which is distinct from the observations resulted from the same random process of probabilistic model, the Dempster's rule would be more appropriate to use to construct the overall belief functions. This situation is often encountered in the context of PRA model.

As can be seen so far, the updating of the belief functions of the uncertain parameter θ of the probabilistic model $\{P_\theta : \theta \in \Theta\}$ using the GBT just requires to calculate the conditional plausibility functions $pl^{\Psi}(x \mid \theta_i)$ given each $\theta_i \in \Theta$. In the following paragraph we will discuss about the calculation of this conditional plausibility function.

3.2.2 About the calculation of the conditional plausibility functions $pl^{\Psi}(x \mid \theta_i)$

As we know, the probabilistic distribution of a random variable X describes the degree of chance (estimated by the long run frequency) of its independent realizations x_1, $x_2,\ldots,x_n$. If the probability distribution of the random variable X is known then the Hacking's frequency principle [8] claims that the degree of belief of an event is equal to its probability i.e. $Bel=P_\theta$. However, in the TBM model, the degrees of chance are not equated with the degrees of belief. Thus, if asked about the belief held by an agent regarding the future realization of X, as argued in [1], this degree of belief should be distinguished from the degree of chance which is only handled at pignistic level in the TBM model. Hence, according to [1], "*we replace the Hacking's principle by the weaker requirement that pignistic probability of an event is considered as its long run frequency when the latter is known"*. In other words, the belief functions on credal level quantifying the belief regarding the next realization of a random variable should be such that its pignistic probability distribution is the probabilistic model $\{P_\theta : \theta \in \Theta\}$. In order to be consistent with the underlying assumptions of the TBM used in our context, we will adopt in this paper this point of view to derive the beliefs with regard to the future observations of a random variable.

If the pignistic probability distribution equated with a probabilistic distribution is known while the corresponding belief and plausibility functions are unknown, then we can recover these functions using the least commitment principle proposed in [3]. Since the pignistic transformation is not bijective, an infinite number of BBA, called a set of isopignistic belief functions, can induce the same *BetP*. In the absence of additional information, the least commitment principle suggests to choose, in the set of all isopignistic BBA, the one that maximizes the commonality function q, named q-least committed (q-LC). Dubois, Prade and Smets [3] demonstrated that the (q-LC) BBA associated with a given pignistic probability distribution *BetP* is unique and consonant (i.e. a possibility distribution). Therefore, according to the results of [3], the conditional plausibility $pl^{\Psi}(x)$ of observing x over the discrete space Ψ given each $\theta_i \in \Theta$ is calculated from *BetP* as follows:

$$pl^{\Psi}(x) = \sum_{y \in \Psi} \min(p(x), p(y)) \qquad (10)$$

Where $p(x)=BetP(x)$ which is a unimodal discrete probability distribution. In the case where Ψ is continuous, the conditional plausibility of a probability density is defined in the same way by substituting the finite sums by integrals.

After calculating the posterior belief functions, similarly to the classical way for updating a probability distribution with the Baye's theorem, it is possible to

estimate the parameter by constructing the pignistic probability induced by the posterior belief functions.

In this section, we studied two approaches for updating the belief functions when new knowledge is available. The first approach is simply based on a standard application of the Dempster-Shafer theory while the second is based on the generalization of the Bayes' rule within the TBM. Both approaches do not require prior belief functions to be set. In literature, theses two approaches have been criticized by [16] and recently discussed in [4]. In practice, the use of Dempster's rule and GBT has been studied for updating the belief functions in some applications [5][11]. In the next sections, we will consider these two approaches in the context of nuclear PRA data.

4 Application of belief updating approaches to Nuclear PRA context

The use of belief functions for modeling the uncertainty associated to reliability parameters in the PRA context has been studied in [20][21]. In these works, the focal elements are constructed from the data as the closed intervals (focal intervals) and then the belief functions are derived. From a computational point of view, this construction is helpful to propagate the uncertainty through a given model function by simulation code. In this section, we will study the use of the approaches presented previously for updating belief functions when new data are available. But let us start by recalling the method currently used in this purpose and based on the MLE [20][21].

4.1 Belief updating from pooled data with Maximum Likelihood Estimation

The MLE is often used to estimate the value of parameters of probabilistic models given observations as the current practice of EDF's Nuclear PRA. Basically, this method relies on the principle of long run frequency to estimate the value of parameters given the number of observations over a time period. For example, the failure rate (often noted as λ) of a component with exponential lifetime is estimated by:

$$\hat{\lambda} = \frac{x}{t} \tag{11}$$

Where x is the number of observed failure events over the time period t. Associated with the estimator, the confidence interval is provided to represent the range of possible values of parameter in which the true value is contained "in most cases" (i.e. for a fraction 100(1-α) of the samples). In the practice of PRA, a 90% confidence interval is often used. When new observations become available, they are combined with previous ones using the pooled data technique to give an updated estimator and a new confidence interval. The new estimator is calculated as:

$$\hat{\lambda} = \frac{\sum_i x_i}{\sum_i t_i} \tag{12}$$

Where $\sum_i x_i$ is the total number of observations and $\sum_i t_i$ is the total exposure time. The confidence interval is also recalculated given this new information. In the traditional uncertainty analysis of PRA, on the basis of this information, a presumed probability distribution such as a log-normal distribution is used in the sense that the subjective probability will reflect our beliefs regarding the values of parameter. However, this point of view has been questioned due to the potential impact of the choice of probability distribution on the results of decision making. An approach using the belief functions of DST is proposed to overcome the issue as studied in [19]. The construction of these functions is based totally on the information given in the form of a nominal value (i.e. an estimated value) and a confidence interval. Obviously, the updating of belief functions when new information is available is not carried out by mean of an aggregation of degrees of belief. Such an approach may have difficulty to incorporate with other sources of information such as those given by expert's opinion. This problem can be addressed using the ROC presented in section 3. This approach allows integrating the prior information given by experts' opinions or past experiences in a natural way. We will see hereafter how this approach is used in the context of PRA data.

4.2 Belief updating with Dempster's rule of combination

When the information about the values of uncertain parameters comes from experts' judgments, the belief functions of DST are appropriate to represent the degrees of beliefs regarding the uncertainty. As independent expert's judgments are given, the combination of these sources of information can be done using the ROC. The same manner can be applied to the case where operating feedback data become available and new belief functions are calculated by taking account of this data as well as the information given by expert's judgments. In this case, the belief functions given the operating data are obtained from the MLE approach and then aggregated with those assessed from expert's judgments. Obviously, one may also apply the ROC for statistical independent data within the MLE context by constructing the belief functions obtained from the confidence intervals of MLE given each data and then aggregating all these functions to obtain the updated belief functions. However some precautions should be taken when using the ROC since the belief functions are constructed on the basis of the confidence intervals of MLE which are randomly derived from a random probabilistic process. This can lead to some cases where the BBA is equal to zero because these confidence intervals may not overlap each other, i.e. they are disjoint intervals each other. This problem can be only addressed if we admit that all the confidence

intervals contain the true value of parameter although this is only true in "most of the cases" (e.g. 90% of chance). This is an unavoidable drawback of the approaches based on the intervals of confidence of MLE to construct belief functions. In [19] some other approaches for the combination of sources of evidence such as mixing or enveloping approaches can be applied for addressing this issue. However, these methods are not appropriate in this context because they tend to widen the uncertainty while we aim to construct the belief functions concentrated around the true value, as new information is available. The GBT inspired from the classical Bayes' rule could be more suitable to construct belief functions given statistical independent observations since this approach does not rely on the use of random confidence intervals of MLE.

4.3 Belief updating with Generalized Bayesian Theorem.

The classical Bayes' theorem has been studied for the parameter estimation and the updating of uncertainty probability distributions in the context of nuclear PRA as in [2][10]. The major issue of this approach resides in choosing an appropriate prior probabilistic distribution since results of an uncertainty analysis could be impacted by this choice. The GBT approach within the theory of belief functions presented in section 3.2 could be the solution to this problem and allows us moreover to characterize the epistemic uncertainty in a more appropriate manner. In general, the probabilistic models in PRA are often supposed to be known in order to characterize the random occurrence of accidents that may occur at nuclear power plants. Therefore, when the belief functions are used to represent epistemic uncertainty associated to its parameters, the updating of these belief functions using GBT can be carried out by considering these probabilistic models as pignistic probability distributions as discussed in section 3. The conditional plausibility $pl^{\Psi}(x|\theta_i)$ on the space of data Ψ given each $\theta_i \in \Theta$ is calculated using the least commitment principle. The probabilistic model in PRA can be divided into two principal types: discrete model and continuous models. However, since the information provided in PRA databook is often given in the form of number of observations, it is usually enough to consider the conditional plausibility $pl^{\Psi}(x|\theta_i)$ on the discrete space of data Ψ. Let us study for instance a Poisson model with an event rate $\lambda^{(2)}$ over an operational time t, the probability of having x accidental events over Ψ given the value of event rate $\lambda_i \in \Theta$ is given as follows:

$$p(x \mid \lambda_i) = e^{-\lambda_i t} \frac{(\lambda_i t)^x}{x!} \qquad (13)$$

Therefore, when the evidence in form of x failures is available, the conditional plausibility $pl^{\Psi}(x \mid \lambda_i)$ is simply calculated by:

$$pl^{\Psi}(x \mid \lambda_i) = \sum_{y \in \Psi} \min(p(x \mid \lambda_i), p(y \mid \lambda_i)) \qquad (14)$$

For example let us consider a frame of data $\Psi=\{x_1, x_2, x_3, x_4\}$, the Poisson model given a specified value of λ_i has the probability distribution such that $p(x_1)=0.3$, $p(x_2)=0.4$, $p(x_3)=0.2$ and $p(x_4)=0.1$. The conditional plausibility $pl^{\Psi}(x_3 \mid \lambda_i)$ of having x_3 failures according to equation (14) is

$$pl^{\Psi}(x_3|\lambda_i)=\min(0.2,0.3)+ \min(0.2,0.4)+ \min(0.2,0.2)+ + \min(0.2,0.1)=0.7.$$

Having calculated the conditional plausibility over space Ψ, the conditional belief functions for all subset $A \subseteq \Theta$ given any observation $x \in \Psi$ can be obtained using equations (7,8). Nevertheless, as we can see, these belief functions from theses equations are computed for the subsets of the discrete frame Θ while it is proposed to construct them on the basis of focal elements which are closed intervals (i.e. focal intervals) for uncertainty propagation in later [21]. Therefore, in order to allow us to update the belief functions using the GBT in our context, it is necessary to transform the belief functions defined on discrete frame to those defined on the real line. We propose for this purpose to build the "empirical" cumulative belief functions and then get the focal intervals from the discretization process. Some additional tasks need therefore to be performed. First of all, to get the discrete frame Θ of a continuous variable θ, we partition the frame Θ such that we have an increasing ordered set of $\theta_1, \theta_2, \ldots, \theta_N$. Then we apply the GBT approach using above equations (7,8) or (6) to calculate the conditional belief and plausibility functions for sets $\{\theta_1\}$, $\{\theta_1, \theta_2\}$, $\{\theta_1, \theta_2, \theta_3\}\ldots\{\theta_1, \theta_2, \theta_3, \ldots, \theta_N\}$. Since these are nested sets, we have always for the belief function (the same for the plausibility function) that $Bel(\{\theta_1\}) \leq Bel(\{\theta_1,\theta_2\}) \leq \ldots \leq Bel(\{\Theta\})=1$. Therefore, similarly to the discrete probability theory if we consider elements $\theta_1, \theta_2, \ldots, \theta_N$ as order statistics and previous belief (plausibility) values as cumulative probabilities then we can build the "empirical" cumulative belief functions on the frame Θ of the continuous variable θ by using a step function. Thus, let note B sets $\{\theta_1\}$, $\{\theta_1, \theta_2\}$, $\{\theta_1, \theta_2, \theta_3\}\ldots\{\theta_1, \theta_2, \theta_3, \ldots, \theta_N\}$, theses functions are expressed as follows

$$Bel^{\Theta}((-\infty,\theta]|x) = \sum_{B \subseteq \Theta} Bel^{\Theta}(B|x).1_{B \subseteq (-\infty,\theta] \subseteq \Theta} \qquad (15)$$

and

$$Pl^{\Theta}((-\infty,\theta]|x) = \sum_{B \subseteq \Theta} Pl^{\Theta}(B|x).1_{B \cap (-\infty,\theta] \neq \emptyset} \qquad (16)$$

where $1_A(x)$ equals one if x is in A and zero in the opposite.

(2) *we use the notation λ instead of θ for this example to keep the same notation used in PRA practical example of the section 5.*

These two functions could be considered as the bounds of a p-box because they are both non-decreasing functions from the real values into the interval [0,1] and the function $Bel^{\Theta}((-\infty,\theta]\,|\,x)$ is less than or equal to $Pl^{\Theta}((-\infty,\theta]\,|\,x)$ for every value of θ. By adopting this view, the Dempster-Shafer focal intervals can be approximately obtained using the discretization methods as described in [19]. The principle of discretization is illustrated in the Figure 1.

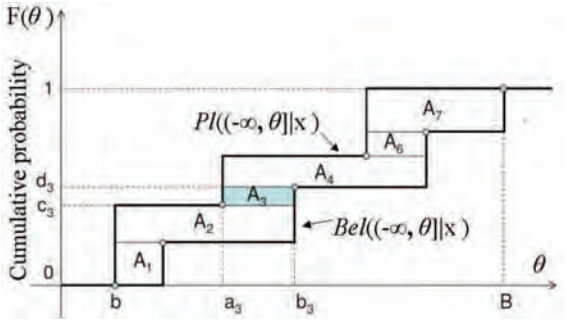


Figure 1 Principle of the construction of focal intervals from a p-box.

The lower and upper bounding functions are assumed to be right and left continuous, respectively. Each rectangle A_i in this figure corresponds to a focal interval $[a_i, b_i]$ with mass $([a_i,b_i])=d_i-c_i$ where d_i and c_i are probability values. These focal intervals then can be used to propagate the parameter uncertainty as done in the framework proposed in [21]. In summary, in order to use the GBT for updating the belief functions of an uncertain parameter θ of a PRA probabilistic model, we go through the following steps:

Step 1: Define the discrete frame Θ of possible values of uncertain parameter θ and then sort them in an increasing order for example, $\theta_1, \theta_2,\ldots,\theta_N$. In practice, the uncertain parameter θ is often given by a bounded confidence interval; the frame Θ can be obtained by discretizing this interval into N possible discrete values.

Step 2: When a new observation x_0 becomes available, compute the plausibilities $pl(x_0|\,\theta_i)$ of observing x_0 given each θ_i using the formula of *least commitment principle* (14).

Step 3: Use the *Generalized Bayesian Theorem*, to calculate the cumulative beliefs for $\theta_1, \theta_2,\ldots, \theta_N$ and then construct "empirical" cumulative belief functions of equations (15,16) for each semi-closed interval $(-\infty, \theta_i]$ on the frame Θ given the observation x_0.

Step 4: Use the discretization methods to obtain focal intervals from "empirical" cumulative belief functions.

Step 5: If the belief functions of some other independent observations are available and/or the prior belief functions come from other sources (e.g. expert's judgment), the final posterior belief functions can be obtained using Dempster's rule of combination.

Step 6: When it is required to provide a point estimate value of parameter θ as in the PRA context, compute the mean (or median or mode) of the pignistic probability distribution induced from posterior belief functions.

In this section, we considered the application of updating belief functions for parameter uncertainty representation in the context of PRA. As we can see, since the mechanism of constructing the belief functions given new information of each method is quite different, thus the results obtained from each one could be different from one to another. Since our main goal is to build posterior belief and plausibility functions such that they should be concentrated around the true value of the parameter, the width between the belief function and the plausibility function should be reduced as new information are available. In order to measure this width of the belief functions obtained from each approach, the measure uncertainty as proposed in [12] can be applied. This measure is defined as follows

$$AW = \sum_{[a_i,b_i]_i \subseteq \Theta} m([a_i,b_i]).(b_i - a_i) \tag{17}$$

This is called a non-specificity measure which quantifies the amount of uncertainty represented by belief functions. As we can see, it measures the aggregated width of all intervals which is the area between the belief and the plausibility functions. The smaller non-specificity measure *AW,* the more specific is the resulting of belief functions. In the following section, this measure will be employed to compare results of updating belief approaches through a practical example.

5 Practical example

In order to illustrate the above approaches through a practical example, we propose to take the example that has been used in [2]. The following example is addressed for the study of an initiating event of PRA but the principle can be applied for other types of failure events.

Problem: Considering a Poisson model with the true but unknown value of an initating event rate λ= 1.2 events per year (13.69E-5/h) over the time period of observation t= 6 years. Thus, the event count follows a Poisson distribution with mean λt = 7.2. In PRA, due to lack of data, the event rate λ is subjected to epistemic parameter uncertainty.

Suppose that we had already prior information about the event rate λ given by a point estimate and an error factor ($EF^{(3)}$), say, λ_{mean}= 5E-5/h and EF=5 which can be

(3) *EF is often used in PRA context to indicate the range of possible values of an uncertain parameter.*

interpreted as the 90% confidence interval such as $\lambda \in [1E-5,\ 25E-5]$ see [20]. This prior information can be viewed as obtained from either expert's judgment or from previous experience. The prior belief functions based on this information are constructed by the approach studied in [2] by considering the point estimate λ_{mean} as the mean value of the uncertain variable λ. These belief functions are displayed in the Figure 2.

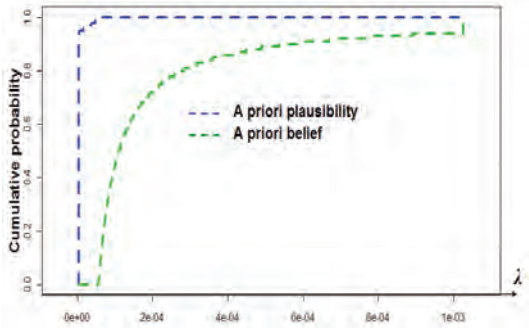


Figure 2 Prior belief and plausibility functions

Now let us suppose that we have new data that are observed from nuclear plants. This can be done by considering the above Poisson process be repeated in a number of times, say, 40 event counts are generated. These may be interpreted as counts from 40 identical plants, each observed for 6 years, or from 40 possible six-year periods at the same plant. Figure 3 shows that the first randomly generated event count was 10, the next was 5, the next was again 10, and so on. Some of the event counts were less than the long-term mean of 7.2, and some were greater. The maximum likelihood estimates of event rate λ are plotted as dots in Figure 3. The corresponding 90% confidence intervals for λ are also plotted. In the Figure 3, the vertical dashed line shows the true value of λ, 1.2.

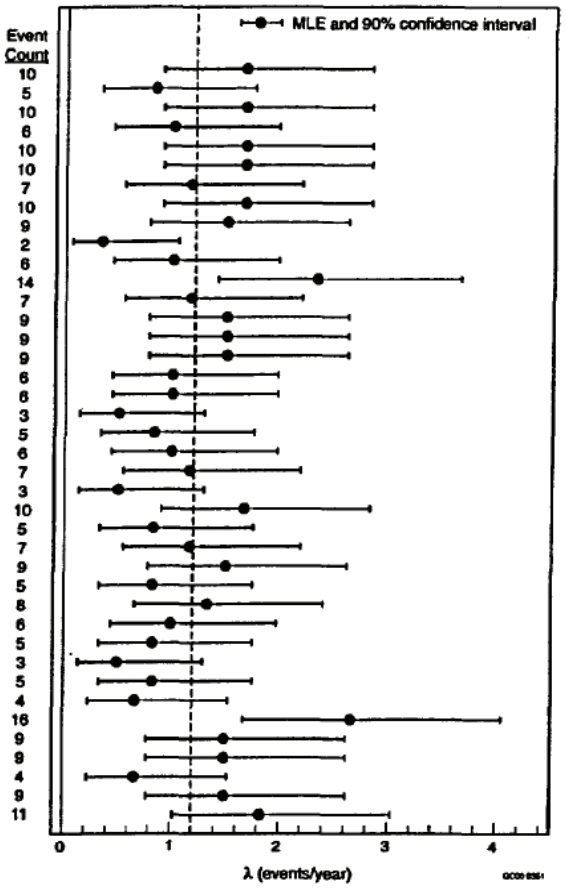


Figure 3 Confidence intervals from random data, all generated from the Poisson process [2].

Given new data, we will next construct the belief functions using the studied approaches. We will consider two cases: one observation and multiple independent observations.

Case 1: One observation

In this case, in order to show the advantage of the GBT with regard to the classical Bayes' theorem, we will distinguish the two following cases.

a. No prior information is available (prior ignorance)

Suppose that we have only the information about the initiating event from the first period of observation of the Poisson process which gives 10 event counts i.e. x=10. In this case, the point estimate value and the 90% confidence interval given by the MLE method are 19.02E-5/h (1.66/year) and [10.32E-5, 32.27E-5] respectively. The belief and plausibility functions can be constructed from this information as showed in the Figure 4.

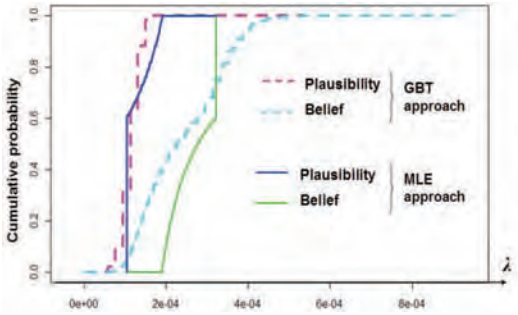


Figure 4 Belief and plausibility functions constructed from MLE approach and GBT without available prior information.

Since we have only one single data (i.e. the first period of observation) while no prior information is available, the Dempster's rule of combination of evidence is not necessary. The Figure 4 displays also the belief functions obtained from the GBT approach. Unlike the classical Baye's theorem where a prior probability distribution is required, no such requirement is needed in the GBT approach. In the absence of prior information, a vacuous belief function i.e. $m(\Theta)=1$ which represents perfectly

the total ignorance is sufficient. This allows us to avoid any assumption about the choice of an appropriate prior probability distribution as in the classical Bayes' theorem. As we can see from the Figure 4, compared to the belief functions of the MLE approach, the results of GBT in this case are more specific because the non-specificity measure *AW* (12.63E-5) is smaller that of the MLE (14.7E-5). The mean pignistic value of the GBT is 17.7E-5/h compared to true value of event rate (13.69E-5/h). Note that, the construction of belief functions is based on the information of 90% confidence interval which is viewed as the upper and lower bounds of the parameter. As discussed in [20][21], in some context this consideration may be helpful to eliminate the values outside the interval which are viewed as unrealistic. However, in other contexts, this can lead to loss of information. The results of the GBT approach are not impacted by this consideration.

b. A priori information is available

When prior information is available, the belief functions of this information can be combined with the belief functions given the new observations. Suppose that we have the prior information of *event rate* as from the Figure 2 i.e. λ_{mean}= 5E-5 and 90% confidence interval [1E-5, 25E-5]. This information is often given by experts' opinions, however, if desired, it can be also viewed as obtained from a previous observation. In this case, the point estimate λ_{mean}= 5E-5 can be considered as if the *event count*s over the time period of 6 year was 3. As a consequence, when the first data of the Poisson process comes with 10 event counts of the first period of observation, the event rate estimated from the pooled data by MLE approach is λ= (10+3)/(6+6)=1.083 events per year (12.36E-5/h) and the 90% confidence interval is [7.31E-5, 19.66E-5]. The belief functions constructed from the pooled data of MLE are showed in the Figure 5. On the other hand, instead of constructing the belief functions from the pooled data, we can use the ROC to build the belief functions given each data. In this case, it is merely sufficient to apply the ROC to prior belief functions (Figure 2) and the belief functions given the first new data (Figure 4). The same way applied to the conditional belief functions with GBT approach. The results of these approaches are displayed in the Figure 5.

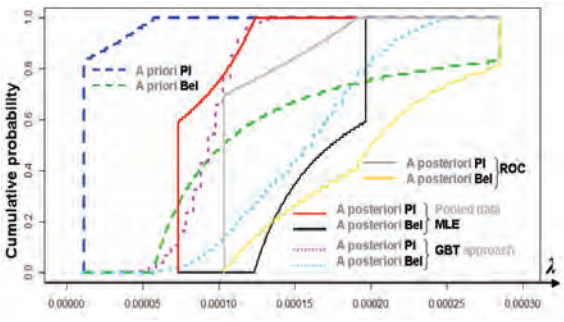


Figure 5 Posterior belief and plausibility functions of approaches vs. prior belief and plausibility functions.

As can be seen, the area between the belief and plausibility functions of GBT approach is smallest since its non-specificity measure (*AW*=6.37E-5) is smaller than that of pooled data MLE approach (8.33E-5) and that of ROC approach (8.35E-5). The mean pignistic value of GBT approach is 12.48E-5/h compared to this value given ROC approach (15.8E-5/h). These values are not far from the true value (13.69E-5/h).

In the first case study, we considered that we had only one data from the first period of observation. In the next case, we suppose having multiple independent observations.

Case 2: Multiple independent observations

In this case, suppose that we have 10 independent series of observations which are collected either from 10 identical power plants during the same time period or from 10 possible six-year periods at the same plant. Thus we have a series of event counts (10,5,10,6,10, 10,7,10,9,2). In the Figure 3, we use the first ten event counts among 40 event counts generated from the random Poisson process.

As in case 1b, given these observations in conjunction with the prior information, the pooled data MLE approach gives the estimated value 14.18E-5/h and 90% confidence intervals [11.70E-5, 17.04E-5]. The belief functions constructed from the pooled data are showed in the Figure 6. In this figure, the results of the ROC approach are also displayed. As can be seen, the resulting belief and plausibility functions do not contain the true

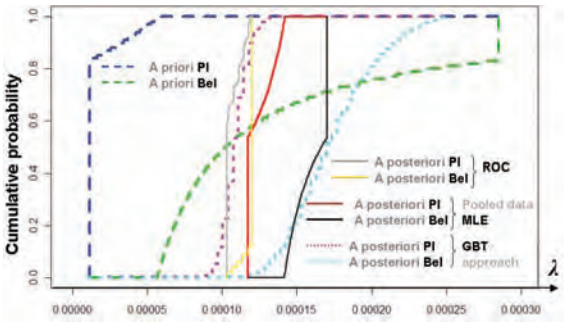


Figure 6 Posterior belief and plausibility functions of approaches in case of multiple independent observations.

value of the event rate because the highest value of these functions on horizontal axe (the maximum value) is smaller than 13.69E-5/h. This is explained by the fact that the assumption that all 90% confidence intervals must contain the true value of parameter is not verified (see the 10^{th} event count).The belief and plausibility functions coming from the GBT are slightly less specific than those coming from the MLE approach in this case but the results allow taking into account possible values located outside the 90% confidence interval of MLE.

6 Conclusions and perspective

In this paper, we studied different approaches for updating belief functions representing parameter uncertainty given new available information in the context of PRA. Although the method of constructing belief functions from pooled data of MLE is intuitive and consistent with the current practice of EDF PRA data, it has some drawbacks regarding the incorporation with other sources of information such experts onions. The method of using the ROC to aggregate belief functions given data within the MLE context is not recommended since its results are too sensitive to random sampling process. The GBT approach appears to be the most appropriate approach to use in PRA context. This approach can address the major issue in the classical Baye's rule about the assumption of prior probability distribution and moreover allows us to overcome the existing drawbacks associated to the MLE approach.

The use of DST for uncertainty representation has been only recently studied in PRA context. A number of challenges of this framework come up for its application within the industrial risk analysis. These approaches studied in this paper for constructing and updating the belief functions need to be reviewed in PRA community and studied within industrial contexts to be integrated in the formal regulatory process.

References

[1] A. Aregui and T. Denoeux. Constructing Consonant Belief Functions from Sample Data using Confidence Sets of Pignistic Probabilities. *International journal of approximate reasoning*, 49: 575-594, 2008.

[2] C. Atwood, J. La Chance, H. Martz, D. Anderson, M. Englehardt, D. Whitehead and T. Wheeler. Handbook of Parameter Estimation for Probabilistic Risk Assessment. *NUREG/CR-6823*, 2002.

[3] D. Dubois, H. Prade and P. Smets. A definition of possibility. *International journal of Approximate reasonning*, 48: 352-364, 2008.

[4] D. Dubois and T. Denoeux. Statistical inference with belief functions and possibility Measures: A discussion of basic assumptions. *In C. Borgelt et al. (Eds), Combining soft computing and statistical methods in data analysis* (*Proceedings of SMPS 2010*, Oviedo, Spain, September 28- October 1, 2010), *Advances in Intelligent and Soft Computing*, 217-225, Springer, 2010.

[5] F. Delmotte and P. Smets. Target identification based on the Transferable Belief Model interpretation of Dempster-Shafer model. *IEEE Transactions on Systems*, Man and Cybernetics A, 34: 457–471, 2004.

[6] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, 1976.

[7] G. Shafer and G. Logan**.** Implementing Dempster's Rule for Hierarchical Evidence. *Artificial Intelligence*, 33:271-298, 1987.

[8] L. Hacking. *Logic of statistical inference*. Cambridge University Press, Cambridge, 1965.

[9] L.S Shapley, A value for n-person games, Contributions to the Theory of Games, *Princeton Univ. Press, Princeton, N.*J. Volume 2, pp. 307-317, 1953.

[10] M. Drouin, G. Parry, J. Lehner, G. Martinez-Guridi, J. LaChance and T. Wheeler. Guidance on the Treatment of Uncertainties Associated with PRAs in Risk-informed Decision making. *NUREG-1855-V.1*, 2009.

[11] M. Owenham, S. Challa and M. Morelande. Fusion of disparate identity estimates for shared situation awareness in a network-centric environment. *Information fusion, 7:*395-417, 2006.

[12] P. Limbourg and E. Rocquigny. Uncertainty analysis using evidence theory confronting level-1 and level-2 approaches with data availability and computational constraints. *Reliability Engineering & System Safety*, 95:550-564, 2010.

[13] P. Smets. Belief Functions: the Disjunctive Rule of Combination and the Generalized Bayesian Theorem. *International Journal of Approximate Reasoning*, 9:1-35, 1993.

[14] P. Smets and R. Kenne. The Transferable Belief Model. *Artificial Intelligence, 66:*191-234, 1994.

[15] P. Smets. Belief Functions on Real Numbers. *International Journal of Approximate Reasoning*, 40:181-223. 2005.

[16] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall London, 1991.

[17] R. G. Almond. *Graphical belief models*. Chapman & Hall, London, 1995.

[18] S. Démotier. W. Schön and T. Denoeux. Risk Assessment Based on Weak Information using Belief Functions: A Case Study in Water Treatment. *IEEE Transactions on Systems*, *Man and Cybernetics C*, 36:382-396, 2006.

[19] S. Ferson, L. Ginzburg, V. Kreinovich, D. Myers and K. Sentz. Construction of Probability Boxes and Dempster-Shafer structures. *Sandia National Laboratories*, *Technical report*, 2003.

[20] T.D. Le-Duy, D.Vasseur, L. Dieulle, C. Bérenguer and M. Couplet. Representation of parameter uncertainty with evidence theory in Probabilistic Risk Assessment. *Proceeding of the Workshop on the Theory of Belief Functions*, *Brest, France,* 2010.

[21] T.D. Le-Duy, D.Vasseur, L. Dieulle, C. Bérenguer and M. Couplet. Uncertainty Analysis by Dempster-Shafer theory in Probabilistic Risk Assessment. *Proceeding of the ESREL 2010 Conference*, *Rhodes, Greece,* 2010.

Robust Equilibria under Linear Tracing Procedure

Hailin Liu
Department of Philosophy
Carnegie Mellon University, Pittsburgh, USA
hailinl@andrew.cmu.edu

Abstract

In Harsanyi and Selten's equilibrium selection theory, the linear tracing procedure has been used to model the hypothetical reasoning process of expectation formation. This paper reconsiders the linear tracing procedure from the perspective of the relationship between priors and Nash equilibria. A prior belongs to the source set of a Nash equilibrium if the linear tracing procedure based on this prior leads to that equilibrium. We show that for any Nash equilibrium, its source set is always nonempty and closed, but not generally convex. This paper also constructs an approach of iterative application of the linear tracing procedure to the auxiliary games that are used to model the hypothetical reasoning under the procedure. We present a notion of robustness of Nash equilibria based on this idea, by replacing uncertainty modelled by a single probability measure with uncertainty modelled by sets of probability measures. This approach attempts to capture the fact that players may not be sufficiently confident in the available information in order to single out one probability distribution that represents their initial beliefs about the other players' possible strategy choices.

Keywords. Equilibrium refinement, linear tracing procedure, stability, robustness, sets of probabilities.

1 Introduction

There are a variety of nontrivial games, with important applications in economics, which generate (sometimes infinitely) many different Nash equilibria. In game theory, a strategy profile is a Nash equilibrium if each player's choice is an optimal response to other players' choices. The fundamental assumption behind this definition is that one player's optimal choice maximizes her own expected utility given the other players' choices. The fact that there are typically multiple Nash equilibria seems to suggest that the equilibrium solution concept is too weak a criterion for predicting the players' behavior. Therefore, a great deal of effort has been devoted to refining the concept of Nash equilibrium by providing more stringent strategy-selection criteria. Examples of suggested equilibrium refinement concepts are Harsanyi and Selten's risk dominance ([2]), Kohlberg and Mertens's stability ([5]), Kreps and Wilson's sequentiality ([6]), and Selten's perfectness ([8]).

Harsanyi and Selten's idea of *risk dominance* captures the idea that, without knowing which equilibrium would be played, the players undergo an introspective process of expectation formation, which may eventually single out one equilibrium as less risky than another. This process is fully modelled by the so-called *linear tracing procedure*, which is thus the mathematical foundation of risk dominance. One of the basic assumptions of this model is that the uncertainty among all players' likely strategy choices is represented by a common prior strategy. However, it could be the case that the uncertainty among the Nash equilibria in question cannot be completely resolved as the assumed reasoning process proceeds. Nevertheless, the linear tracing procedure itself is a mathematical mechanism for modelling the players' hypothetical deliberation process about uncertainty. We shall later return to the linear tracing procedure and describe it in detail.

Moreover, we extend the framework of the linear tracing procedure to accommodate sets of probabilities as a representation of uncertainty. We then examine the possibility of iteratively applying the linear tracing procedure to a sequence of auxiliary games. This may be regarded as a minor generalization of the traditional game-theoretic framework, by only dropping the so-called "dogma of precision" ([9]), namely, that uncertainty should always be represented by a single probability measure. This enables us to assess the robustness of Nash equilibria in the traditional game-theoretic context, where uncertainty is represented in a more realistic manner.

To explain the basic ideas, consider the two-person coordination game described by Figure 1. In this game,

player 1 has two pure strategies denoted by s_{11} and s_{12}, while player 2 also has two pure strategies denoted by s_{21} and s_{22}.

	s_{21}	s_{22}
s_{11}	$1, 1$	$0, 0$
s_{12}	$0, 0$	$3, 3$

Figure 1: Coordination Game

The game has two Nash equilibria in pure strategies, namely $E_1 = (s_{11}, s_{21})$ and $E_2 = (s_{12}, s_{22})$. It also has one Nash equilibrium in mixed strategies, $E_3 = ((\frac{3}{4}, \frac{1}{4}), (\frac{3}{4}, \frac{1}{4}))$, where the first and second pairs of numbers denote the probabilities assigned to player 1's and player 2's two pure strategies respectively. For convenience, the strategy space of the game can be described by the square $ABCD$ in Figure 2. Any point X of this square will represent a mixed strategy profile $\delta = ((q_{11}, q_{12}), (q_{21}, q_{22}))$. In particular, the horizontal distances XY and XZ will represent the probabilities q_{11} and $q_{12} = 1 - q_{11}$ respectively, and the vertical distances XV and XU will represent the probabilities q_{21} and $q_{22} = 1 - q_{21}$ respectively. Accordingly, the three Nash equilibria of the game can be represented by the corner points A and C, and the point E, as shown in Figure 2.

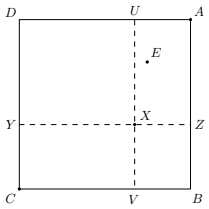


Figure 2: The Strategy Space

A natural question concerning this game is which of the three equilibria would be played. Harsanyi and Selten attempt to answer this question by employing the linear tracing procedure to examine the risk associated with different Nash equilibria when belief-uncertainty is represented by a single probability distribution. Here, we want to investigate the viability of Nash equilibria under the recursive application of the linear tracing procedure when uncertainty is modelled by *sets of probabilities*. We thereby hope to shed light on how traditional solution concepts can be informed by the notion of imprecise probabilities.

The remainder of the paper is structured as follows. Section 2 provides a formal description of the linear tracing procedure and some characterization results concerning source sets. In Section 3 we describe an approach which involves iterative application of the linear tracing procedure to a self-generated sequence of hypothetical games, where uncertainty is represented by sets of probabilities. On the basis of such a recursively applied linear tracing procedure, we then formalize and investigate a notion of *stability,* which measures the tenability of a prior strategy with respect to a certain Nash equilibrium under this procedure. The rest of this section extends the analysis of the linear tracing procedure to allow for representing uncertainty by sets of probabilities, and proposes a notion of *robustness* of Nash equilibria. Section 4 consists of concluding remarks and suggestions for future work along these lines.

2 Linear Tracing Procedure and Source Sets

The linear tracing procedure is a mathematical tool first introduced by Harsanyi ([3]), which underpins the equilibrium refinement concept proposed by Harsanyi and Selten ([2]). Informally speaking, it models how the players gradually update their strategy plans in light of what they know about the opponents' strategic reactions to their own expectations. The procedure can be regarded as a rational deliberation process of expectation formation, after which each player comes to choose a particular Nash equilibrium and to expect the others to make the same choice. The linear tracing process begins with a common probability distribution over all players' strategies, which represents their initial expectations about the other players' likely strategy choices. This way of setting up the initial belief state is often called the *Harsanyi doctrine* or, alternatively, the common prior assumption. Under such an assumption, it would seem that all players should adopt the best responses against the assumed common prior. And this typically gives rise to a different strategy combination that generally does not constitute a Nash equilibrium. Throughout the linear tracing procedure, all players gradually change their own tentative strategy plans, as well as their expectations about the other players' possible strategies, until they arrive at a certain Nash equilibrium. It has been shown ([2]) that the linear tracing procedure determines a unique Nash equilibrium for almost every game. In this section we shall explore the linear tracing procedure from a different perspective, focusing on characterizing the set of priors associated with a certain Nash equilibrium under the linear tracing procedure.

Let us begin with some basic notations and concepts. Let $G = \langle I, \{S_i\}, \{u_i\}\rangle_{i \in I}$ be a finite non-cooperative strategic form game, where I denotes a finite set of players, and S_i denotes the finite set of pure strategies of player i, and $u_i : S \to \mathbb{R}$ denotes a continuous payoff function of this player (where $S = \prod_{i \in I} S_i$).

As usual, we can extend the strategy space to include mixed strategies. In general, we let Δ_i represent the set of mixed strategies of player i, and similarly $\Delta = \prod_{i \in I} \Delta_i$. Likewise, the payoff function of a given player i can be extended in the standard way to the set of all mixed strategy combinations Δ, and we usually write $u_i(\delta)$ for the expected payoff of player i when $\delta \in \Delta$ is played. Let δ_{-i} denote the strategy combination $(\delta_1, \ldots, \delta_{i-1}, \delta_{i+1}, \ldots, \delta_n)$. For any $\delta_{-i} \in \Delta_{-i}$, the set of player i's best responses given δ_{-i} is defined as $B_i(\delta_{-i}) = \{\delta_i \in \Delta_i : u_i(\delta_i, \delta_{-i}) \geq u_i(\delta_i', \delta_{-i})$ for all $\delta_i' \in \Delta_i\}$. A strategy profile $\delta^* \in \Delta$ is a Nash equilibrium of G if and only if each player's strategy is a best response to the other players' strategies, i.e., $\delta_i^* \in B_i(\delta^*_{-i})$ for every player $i \in I$. Henceforth, the set of all Nash equilibria of the game G will be denoted by $NE(G)$. Also, we shall assume that some finite non-cooperative game G is already given.

The linear tracing procedure is a mapping φ from the strategy space Δ into the equilibrium set $NE(G)$. It transforms each strategy profile into a certain Nash equilibrium of the game G. In order to define the linear tracing procedure, consider a one-parameter family of auxiliary games $\Gamma^{t,p}$ with $t \in [0,1]$ and $p \in \Delta$. Each game $\Gamma^{t,p}$ is of the same structure as the original game G, except for the payoff functions. In $\Gamma^{t,p}$, for each $\delta \in \Delta$, each player i's payoff function $u_i^{t,p}$ satisfies

$$u_i^{t,p}(\delta_i, \delta_{-i}) = t u_i(\delta_i, \delta_{-i}) + (1-t) u_i(\delta_i, p_{-i})$$

where u_i is player i's payoff function in the original game G. Obviously, $u_i^{1,p}(\delta_i, \delta_{-i}) = u_i(\delta_i, \delta_{-i})$, which implies that $\Gamma^1 = G$. While, for $t = 0$ the game $\Gamma^{0,p}$ decomposes into n independent and separate one-person maximization problems, one for each player. Now consider the equilibrium correspondence $\psi : t \to NE(\Gamma^{t,p})$ for $t \in [0,1]$ and $p \in \Delta$:

$$\psi = \{(t, \delta) \mid t \in [0,1], \delta \in NE(\Gamma^{t,p})\}$$

Let $\varphi = \varphi(G, p)$ be the graph of the correspondence ψ. Thus, any point x of graph φ has the mathematical form $x = (t, \delta)$, where t is a specific t value whereas δ is an equilibrium point of the corresponding auxiliary game Γ^t. Note that the graph is not always a function. It can be shown that the graph φ contains at least one distinguished path L, the so-called *feasible path*, which connects a starting point $x_0 = (0, \delta_0)$ with an end point $x_1 = (1, \delta^*)$. Hence, for a given game G and for a given prior strategy $p \in \Delta$, we call Θ a *linear tracing procedure* if it consists in selecting an outcome q^* by tracing a feasible path L from its starting point $x_0 = (0, \delta_0)$ to its end point $x_1 = (1, \delta^*)$. And δ^* will be called the *outcome* of the linear tracing procedure Θ. Figure 3 shows the graph of a linear tracing procedure for the game in Figure 1. For this linear tracing procedure, $B'''B''C''C'C$ is the unique feasible path and the equilibrium E_3 (point C) is the outcome.

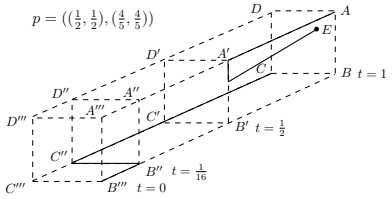


Figure 3: The linear tracing procedure based on p

The linear tracing procedure will always lead to at least one equilibrium, and select one unique equilibrium as the solution for "almost all" games[1]. The linear tracing procedure is called *feasible* if the graph $\varphi = \varphi(G, p)$ contains at least one feasible path L, and is called *well-defined* if X contains exactly one feasible path L. It can be shown that, for any possible pair (G, p), the linear tracing procedure is always feasible but is not always well-defined. In light of its fundamental importance, we state this result as follows.

Proposition 1. *([2]) For any possible choice of game G and prior vector p, the linear tracing procedure is always feasible. However, for some choices of G and p, the linear tracing procedure is not well-defined.*[2]

It is worth noting that the above proposition tells us nothing about the set of priors associated with a certain Nash equilibrium. There are several interesting questions that are worthy of further investigation. For instance, is the set non-empty, closed or convex? Before considering these problems, we first define the concept of *source sets* as follows.

Definition 2. For a given game G and a strategy $\delta^* \in NE(G)$, the *source set* for δ^*, denoted by $\Phi(\delta^*)$, is defined as the set of all prior strategies, based on which the linear tracing procedure yields the Nash equilibrium δ^* as outcome.

Our next proposition shows that for each Nash equilibrium δ^*, its source set always includes itself as an

[1] See Harsanyi and Selten ([2]) for a more detailed explanation of the term "almost all".

[2] The proof provided by Harsanyi and Selten is heavily dependent on the result showing that the logarithmic tracing procedure (also introduced by them) is always well defined. It is worth pointing out that a mathematical proof of feasibility of the linear tracing procedure can be easily derived from a theorem given in [7]. Using techniques from the field of algebraic geometry, Schanuel et al. first show that the logarithmic tracing procedure always connects the prior strategy to exactly one equilibrium point. Based on this result, one can argue that the feasibility of the linear tracing procedure is exactly a limit case of the feasibility of the logarithmic tracing procedure. More recently, Herings ([4]) directly shows the feasibility of the linear tracing procedure without appealing to the logarithmic tracing procedure. The two simple proofs provided by Herings are based on theorems related to the fixed-point theorems of Brouwer and Kakutani.

element, and is thus non-empty.

Proposition 3. *Let G be a finite non-cooperative game. For each Nash equilibrium δ^* of game G, δ^* belongs to its own source set, i.e., $\delta^* \in \Phi(\delta^*)$.*[3]

Next, we might ask whether the source sets are closed under the topology of pointwise convergence. To characterize the closure property of source sets, we must first introduce the concept of *pointwise convergence* on the strategy space of a game G, as well as that of a *limit point* of a set. Recall that we are considering only games with a finite number of pure strategies. Thus, the topology that we are considering is relatively easy to work with. We now define pointwise convergence as follows.

Definition 4. Let Δ be the strategy space of a finite game $G = \langle I, \{S_i\}, \{\pi_i\}\rangle$. A sequence $\{\delta^r\}$ *converges pointwise* to $\delta \in \Delta$, denoted by $\{\delta^r\} \to \delta$, if for each player $i \in I$, all $s_i \in S_i$, and all $\epsilon > 0$, there exist some k such that $|\delta_i^j(s_i) - \delta_i(s_i)| < \epsilon$ for each $j \geq k$. And δ is called the *limit point* of the sequence $\{\delta^r\}$.

Let us compare pointwise convergence and uniform convergence defined in the following sense. We say that a sequence $\{\delta^r\}$ *converges uniformly* over players' strategies to δ if for all $\epsilon > 0$, there exists some k such that for each player i, all $s_i \in S_i$, and all $j \leq k$, it holds that $|\delta_i^j(s_i) - \delta_i(s_i)| < \epsilon$. Clearly, uniform convergence is a stronger concept, and always implies pointwise convergence, but not vice versa. In our framework, however, pointwise convergence implies uniform convergence, since the set of players is finite, as well as each player's set of pure strategies.

In this paper, a point $p \in \Delta$ is called a *limit point* of the source set $\Phi(\delta^*)$ if there exists some sequence $\{p^r\}$ such that each element of $\{p^r\}$ belongs to $\Phi(\delta^*)$ and $\{p^r\} \to p$. We shall employ the notion of limit points to obtain a characterization of closed sets. Loosely speaking, a set A is *closed* in a space X if it contains all its limit points. Our main result is that the source sets of Nash equilibria are always closed. More formally:

Proposition 5. *Let G be a finite non-cooperative game and δ^* be a Nash equilibrium of G. If $p \in \Delta$ is a limit point of the source set $\Phi(\delta^*)$, then $p \in \Phi(\delta^*)$.*

We now give some definitions and lemmas that will be used in the proof of the foregoing proposition.

Definition 6. Let $G^r = \langle N^r, (S_i^r), (\pi_i^r)\rangle$ be a finite non-cooperative game with $r = 1, 2, \ldots$. A sequence of games $\{G^r\}$ converges to a game G if all the games in question have the same set of players $N^r = N$ and the identical set of pure strategies $S_i^r = S_i$, and the payoff function π_i^r converges uniformly to π_i, that is, for all $\epsilon > 0$, there exists some k such that for each player $i \in I$, all $s \in S$ and for all $j \geq k$, it holds that $|\pi_i^j(s) - \pi_i(s)| < \epsilon$.

Obviously, it follows from the definition that the sequence $\{G^r\}$ converges to G if all games under consideration share the same set of players $N^r = N$ and strategy space $\Delta^r = \Delta$ and, moreover, the payoff function u_i^r converges uniformly to u_i. We say that a game G is the limit game of a sequence of games $\{G^r\}$ if the sequence converges to G. The following lemma can be regarded as a special version of the well-known result ([1]) in game theory, which relates the Nash equilibria of a convergent sequence of games to the Nash equilibria of the limit game.

Proposition 7. *Let $\{G^r\}$ be a sequence of finite non-cooperative games converging to G. If the strategy profiles δ^r are Nash equilibria of G^r respectively with $\{\delta^r\} \to \delta$, then δ is a Nash equilibrium of game G.*

Now consider a sequence of prior strategies $\{p^r\}$, which converges to a prior strategy p. It is easy to verify that for each $t \in [0,1]$ the sequence of games $\{\Gamma_{p^r}^t\}$ converges to the game Γ_p^t. In order to see this, let us recall that in game $\Gamma_{p^r}^t$ the payoff function $u_{i,p^r}^t : \Delta \to \mathbb{R}$ is given by $u_{i,p^r}^t(\delta_i, \delta_{-i}) = tu_i(\delta_i, \delta_{-i}) + (1-t)u_i(\delta_i, p_{-i}^r)$, where u_i denotes the payoff function of the original game G. Since the payoff function u_i is assumed to be continuous, it directly follows from $\{p^r\} \to p$ that $\{u_{i,p^r}^t\}$ converges to $u_{i,p}^t$. Moreover, if a sequence $\{t^m\}$ converges to t where $t^m, t \in [0,1]$, then it still holds that the sequence of games $\{\Gamma_{p^r}^{t^m}\}$ converges to the game Γ_p^t, since the sequence $\{u_i^{t^m,r}\}$ of payoff functions converges to u_i^t. Thus, it follows from the above lemma that the limit of Nash equilibria relative to the sequence of games is the Nash equilibrium of the limit game in both cases. These noteworthy facts turn out to play a significant role in the proof of the closure property of source sets. We now state the foregoing results as follows.

Corollary 8. *Suppose that $\{p^r\} \to p$ and $t \in [0,1]$ where $p^r, p \in \Delta$. If δ^r are Nash equilibria relative to game $\Gamma_{p^r}^t$ with $\{\delta^r\} \to \delta$, then δ is a Nash equilibrium of game Γ_p^t.*

Corollary 9. *Suppose that $\{p^r\} \to p$ and $\{t^m\} \to t$ where $t^m, t \in [0,1]$ and $p^r, p \in \Delta$. If $\delta^{m,r}$ are Nash equilibria relative to game $\Gamma_{p^r}^{t^m}$ with $\{\delta^{m,r}\} \to \delta$, then δ is a Nash equilibrium of game Γ_p^t.*

Let us turn to the convexity of source sets. In some games, the source sets are convex, in the sense that any mixture combination between two strategies from a source set also belongs to the source set. It is not the

[3] Proofs not given in the main text can be found in the Appendix.

case, however, that convexity holds in general. This point can be easily illustrated by considering the coordination game in Figure 1. As mentioned before, this game has three Nash equilibria, i.e., E_1, E_2, and E_3. It can be verified by simple computation that the source sets of these three Nash equilibria can be described as shown in Figure 4. In particular, the source set of E_1 consists of all points within the area $AHEF$, the source set of E_2 consists of all points within the area $BCDFEH$, and the source set of E_3 consists of all points lying on the border segment HEF. Clearly, the source sets of the equilibria E_1 and E_3 are not convex.

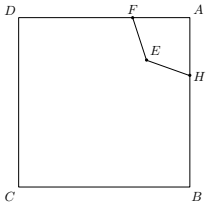


Figure 4: The source Sets

3 Robustness to Sets of Probabilities

As we mentioned above, the purpose of the linear tracing procedure is to provide a rational and effective mechanism for selecting a Nash equilibrium as the solution of non-cooperative game. Now let us recall how the linear tracing procedure works. The linear tracing procedure is not merely an examination of the game itself. Instead, we invoke a sequence of hypothetical games to investigate how the equilibria of the original game behave in auxiliary games. It is worthwhile to note that these auxiliary games are also non-cooperative, and typically resemble the original game in other respects. In other words, the auxiliary games themselves are also amenable to the linear tracing procedure. Therefore, it seems reasonable to apply the linear tracing procedure recursively to solve these auxiliary games. In this section, we will investigate such recursive applications of the linear tracing procedure.

Consider the finite non-cooperative strategic form game $G = \langle I, \{S_i\}, \{u_i\}\rangle_{i\in I}$ and the linear tracing procedure for G as described in Section 2. Let $p \in \Delta$ be one prior strategy and define a one-parameter family of auxiliary games Γ_p^t with $t \in [0,1]$. Generally speaking, any such game Γ_p^t will be a game of the same structure as the original G except for the payoff functions. More precisely, Γ_p^t can be specified as $\Gamma_p^t = \langle I, \{S_i\}, \{u_i^t\}\rangle_{i\in I}$, where, for each $\delta \in \Delta$, the payoff function u_i^t is defined by

$$u_i^t(\delta_i, \delta_{-i}) = tu_i(\delta_i, \delta_{-i}) + (1-t)u_i(\delta_i, p_{-i}).$$

Now let us consider an application of the linear tracing procedure to the game Γ_p^t for some $t \in [0,1]$. That is, for some $t \in [0,1]$ assume that Γ_p^t is the original game, denoted by G^t. Define a one-parameter family of auxiliary games $\Lambda^{t'}$ with $t' \in [0,1]$ as follows. Given a prior strategy $p' \in \Delta$, game $\Lambda^{t'}$ can be defined as $\Lambda_{p'}^{t'} = \langle I, \{S_i\}, \{\mu_i^{t'}\}\rangle_{i\in I}$, where, for each $\delta \in \Delta$, the payoff function $\mu_i^{t'}$ satisfies

$$\mu_i^{t'}(\delta_i, \delta_{-i}) = t'u_i^t(\delta_i, \delta_{-i}) + (1-t')u_i^t(\delta_i, p'_{-i}).$$

It was shown in Proposition 2 that the source set of any equilibrium point for the original game G is not empty. In order to examine the robustness of an equilibrium, we focus on how its source set changes when applying the linear tracing procedure recursively to the auxiliary games.

Before entering into further analysis of the source set, we first consider one interesting case: what happens if, throughout the recursive application of the linear tracing procedure, we always use the same prior as a starting point? Suppose that δ^* is an equilibrium of game G, and p is an element of the source set of δ^*, that is, $p \in \Phi(\delta^*)$. Now consider the games Γ_p^t, which can be represented as $\Gamma_p^t = \langle I, \{S_i\}, \{u_i^t\}\rangle_{i\in I}$, where, for each $\delta \in \Delta$, the payoff function u_i^t satisfies

$$u_i^t(\delta_i, \delta_{-i}) = tu_i(\delta_i, \delta_{-i}) + (1-t)u_i(\delta_i, p_{-i}).$$

Then apply the linear tracing procedure to game Γ_p^t with the same prior p. As mentioned above, we have to consider a new one-parameter class of auxiliary games $\Lambda_p^{t'} = \langle I, \{S_i\}, \{\mu_i^{t'}\}\rangle_{i\in I}$ with $t' \in [0,1]$, where, for each $\delta \in \Delta$, the payoff function $\mu_i^{t'}$ satisfies

$$\mu_i^{t'}(\delta_i, \delta_{-i}) = t'u_i^t(\delta_i, \delta_{-i}) + (1-t')u_i^t(\delta_i, p_{-i}).$$

Obviously, $\Lambda_p^0 = \Gamma_p^0$, since the payoff functions are identical, that is, $\mu_i^0 = u_i^0$. For the same reason, we have $\Lambda_p^1 = \Gamma_p^t$. Thus, the class of auxiliary games $\Lambda_p^{t'}$ is a subset of the family of auxiliary games Γ_p^t with respect to the game G. In other words, when considering the linear tracing procedure applied to the game Γ_p^t, we are in fact investigating a smaller subset of the family of auxiliary games generated by the linear tracing procedure applied to the *original* game. Hence, we can show that whenever δ^* is an equilibrium point of game Γ_p^t, the linear tracing procedure starting from p always feasibly leads to δ^*. On the basis of this observation, the following result is immediate:

Theorem 10. *Let $G = \langle I, \{S_i\}, \{u_i\}\rangle_{i\in I}$ be a finite non-cooperative strategic form game, and let δ^* be one equilibrium point of G. If $p \in \Phi(\delta^*)$ and δ^* is a Nash equilibrium of game Γ_p^t, then p is an element of the source set of δ^* with respect to game Γ_p^t.*

Now we can ask: given a certain equilibrium, under what constraint would a prior strategy belong to its source set throughout the recursive application of the linear tracing procedure? It turns out that, whenever the equilibrium δ^* under consideration maximizes the expected payoff for each player with respect to the prior strategy p, then p is always an element of the source set of δ^* pertaining to game Γ_p^t for any $t \in [0,1]$. More precisely, we have:

Theorem 11. *Let $G = \langle I, \{S_i\}, \{u_i\}\rangle_{i\in I}$ be a finite non-cooperative strategic form game, and let δ^* be an equilibrium point of G. For any $t \in [0,1]$, if δ^* maximizes all players' expected payoffs with respect to the prior strategy p, then $p \in \Phi^t(\delta^*)$ with respect to Γ_p^t.*

Clearly, the prior strategy determines how far into the recursive application of the procedure the prior p remains an element of a source set of the same equilibrium. This suggests that, when recursively applying the linear tracing procedure, the duration in which the prior strategy p belongs to the same source set can be considered a measure of the stability of p. According to the foregoing theorem, when a certain Nash equilibrium δ^* maximizes all the players' expected payoffs with respect to a prior p, then p is the most stable element of the source set of δ^*. This is because the linear tracing procedure that begins with p always points to the same equilibrium δ^*. Thus, we can say that such a prior strategy p is *maximally stable* with respect to δ^*. We now define the measure of stability.

Definition 12. Given a finite non-cooperative strategic form game G and one equilibrium δ^*, the *stability* of a prior strategy $p \in \Delta$ with respect to δ^* is a real-valued function γ on $\Phi(\delta^*)$, which is defined as $\gamma(p, \delta^*) = 1 - t^*$, where t^* is the smallest value of t such that $p \in \Phi^t(\delta^*)$. We say that p is *maximally stable* with respect to δ^* when $\gamma(p, \delta^*) = 1$. The set consisting of all such prior strategies is called the maximally stable source set of δ^*.

To illustrate the notion of stability with respect to Nash equilibria, consider the coordination game mentioned in section 1. The general description is shown in Figure 5. In particular, the source set of E_1 (the area $AHEF$) can be divided into two parts: the area $AGEI$ contains all the maximally stable priors with respect to the equilibrium E_1, and the remaining area consists of the priors with $\gamma(p, E_1) < 1$. Similarly, the source set of E_2 is composed of the maximally stable source set $ENCM$ and the rest of the non-maximally stable prior strategies with $\gamma(p, E_2) < 1$. In contrast, the maximally stable source set of the mixed strategy equilibrium E_3 consists of only *one* prior strategy, namely itself. All other prior strategies in its source set are not maximally stable with respect to E_3.

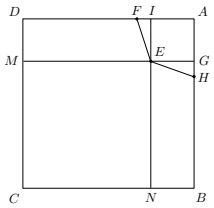


Figure 5: The Stability of Prior strategies

For each element p of the source set of a certain equilibrium, there is a measure of the stability of p, indicating how long the prior p remains associated with the same source set in the recursive application of the procedure. Recall from the previous section, that the source set of each equilibrium is non-empty and closed. Note that all the intermediate games invoked by the linear tracing procedure are closely related to the original game. Thus, the stability of the prior strategies under the recursive application of the linear tracing procedure indicates the robustness of the Nash equilibria with respect to the original game. On the basis of t-value as a measure for the stability of the priors, it is reasonable to employ the stability measure to compare the robustness of the Nash equilibria of a given game.

Before we define the measure of robustness, let us informally motivate the very idea of introducing sets of probabilities into the game-theoretic framework. As mentioned above, there are many games that have multiple Nash equilibria. This fact has given rise to a wide discussion of the equilibrium refinement problem in game theory. We believe that the linear tracing procedure is an appropriate mathematical mechanism for comparing Nash equilibria, since it accords with a common intuition regarding relative degrees of risk associated with different Nash equilibria. As mentioned above, the linear tracing procedure invokes a family of auxiliary games closely resembling the original game in question. Thus, by applying it recursively to the auxiliary games, we provide further information about the original game. In fact, it indicates the stability of one prior strategy with respect to a certain Nash equilibrium.

On the other hand, the linear tracing procedure assumes that all players employ the same probability distribution to represent their initial beliefs about the other players' likely strategy choices. In their analysis, Harsanyi and Selten choose a single probability distribution to express the uncertainty among players regarding which strategy the others would adopt. In decision theory, however, there are numerous suggested methods to represent decision makers' uncertainty besides using a single probability. Some salient approaches involve modelling uncertainty using sets of probabilities, upper and lower probabilities, upper

and lower previsions, and belief functions ([9]).

Here we want to employ a non-trivial, convex set of probability measures $\mathcal{P}$ to represent all players' ignorance about the other players' likely behaviors. More precisely, we want to extend Harsanyi and Selten's framework framework by employing sets of prior strategies, rather than one single prior, to represent players' initial beliefs. Note that each of the prior strategies under consideration leads to a certain equilibrium under the linear tracing procedure, which simply means that it belongs to the source set of that equilibrium. Moreover, when we recursively apply the linear tracing procedure to the auxiliary games, we can determine the stability measure associated with each of the prior strategies with respect to a certain equilibrium. Based on these measures of stability, we can now define the robustness of equilibria with respect to a set of prior strategies as follows.

Definition 13. Let G be a finite non-cooperative strategic form game, and let the players' initial beliefs about the other players' possible behaviors be represented by a set of prior strategies $\mathcal{P}$. The robustness of an equilibrium δ^* with respect to $\mathcal{P}$ is defined as $R(\delta^*, \mathcal{P}) = \min_{p \in \mathcal{P}} \gamma(p, \delta^*)$, i.e., the minimum stability index associated with the priors with respect to $\mathcal{P}$.

This notion can be regarded as a further refinement of Nash equilibria based on the stability measures of the priors under the iterative application of the linear tracing procedure. Basically, one equilibrium is more robust than another if the least stability index associated with the elements of its source set is higher than the one associated with the other equilibrium under the recursive application of the linear tracing procedure. Given certain sets of prior strategies, we employ the maximin principle to assess the robustness of equilibria, where uncertainty is represented by sets of probabilities. That is, we select the equilibrium that maximizes the possible minimum stability of the prior strategies in its source set.

In order to illustrate the idea, let us consider an ϵ-*contaminated class* of probabilities given by $M = \{(1-\epsilon)P + \epsilon Q, Q \in \mathscr{P}\}$, where P is a particular prior distribution and ϵ is a fixed number in $[0, 1]$. $\mathscr{P}$ is the class of probability distributions that represents the possible deviations of the prior P. And the fixed ϵ represents the degree of contamination that players want to introduce into P.

Example 14. (ϵ-contamination under equilibria coordination) Consider the game described above. Suppose that all players believe that they will play the game in a coordination way. That is, the players collectively choose some mixed strategies involving the equilibria E_1, E_2, and E_3. Let $\mathscr{P} = \{Q : Q = p_1E_1 + p_2E_2 + p_3E_3, \text{ where } p_1 + p_2 + p_3 = 1\}$. Figure 6 (the dark segment on the diagonal AC) illustrates the corresponding ϵ-contaminated class $\mathcal{P} = \{(1-\epsilon)P + \epsilon Q, Q \in \mathscr{P}\}$ when $P(E_1) = \frac{7}{10}$, $P(E_2) = \frac{1}{5}$, $P(E_3) = \frac{1}{10}$ and $\epsilon = 0.2$, which represents the players' initial beliefs. Observe that each prior in $\mathcal{P}$ is maximally stable with respect to either E_1, E_2, or E_3. Thus, $R(E_1, M) = R(E_2, M) = R(E_3, M) = 1$. This suggests that in this game when all players believe they will coordinate on an equilibrium, the notion of maximin robustness proposed here does not distinguish among these equilibria.

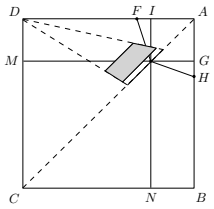


Figure 6: ϵ-contaminated class

Example 15. (Coordination failure) Reconsider the same game again. Suppose that all players initially believe that they will fail to coordinate with small probability. More precisely, the players believe that they will mostly choose a strategy from the ϵ-contaminated class $\mathcal{P}$, or otherwise adopt the strategy $D = (s_{12}, s_{21})$ with some probability in $[0.05, 0.2]$. In this case, the players' initial beliefs can be represented by $\mathcal{P}' = \{(1-\alpha)\mathcal{P} + \alpha D, 0.05 \leq \alpha \leq 0.2\}$, which is illustrated by the grey area in Figure 6. Simple calculation gives us the following result: for all $p \in \mathcal{P}$, $\frac{57}{64} = 0.8\bar{9} \leq \gamma(p, E_1) \leq 1$, $\frac{375}{482} = 0.7\bar{8} \leq \gamma(p, E_2) \leq 1$, and $\gamma(p, E_3) = 0$. Thus, $R(E_1, \mathcal{P}') = 0.8\bar{9}$, $R(E_2, \mathcal{P}') = 0.7\bar{8}$, and $R(E_3, \mathcal{P}') = 0$. Thus, according to the notion of maximin robustness we defined, E_1 is the most robust equilibrium with respect to $\mathcal{P}'$.

4 Conclusion

Why should one Nash equilibrium be more likely to be played than any other in a game? There is a large literature providing different criteria for selecting a particular equilibrium among many. Harsanyi and Selten propose a notion of risk dominance based on the linear tracing procedure. Here we do not attempt to address the issue of whether risk dominance is an appropriate criterion for equilibrium comparison. Instead, we extend the manner in which the linear tracing procedure is applied, as well as to replace a single probability distribution with sets of priors to represent players' uncertainty about strategy choices.

We first showed that, for any Nash equilibrium, its source set is always nonempty and closed, but not

necessarily convex. We then considered a recursive application of the linear tracing procedure onto sequences of hypothetical games generated by the procedure itself. Based upon this, we formalized a notion of stability of priors, as well as a sufficient condition for characterizing the set of maximally stable priors with respect to certain Nash equilibria. Finally, we introduced a notion of maximin robustness of equilibria by considering the recursively applied linear tracing procedure when uncertainty is represented by a set of probabilities rather a single probability measure. We employed the maximin criterion to measure the robustness index associated with the Nash equilibria related to certain sets of prior strategies under the recursive procedure. The approach considered here is thus meant to demonstrate how one might accommodate the idea of imprecise probabilities within the traditional game-theoretic framework.

In the future, we intend to continue our examination of robustness in more general games, for instance symmetric games. This would provide further characterization about how sets of probabilities can be incoporated within game-theoretic framework. We shall also compare this approach to other existing theories of equilibrium refinement to investigate the relationships among them. Moreover, we shall consider the possibility of developing a new solution concept based on sets of probabilities that possesses more appealing features.

Acknowledgements

I am particularly grateful to my supervisor, Professor Teddy Seidenfeld for valuable discussions, supervision and encouragement. And I would like to thank Assistant Professor Kevin Zollman and two anonymous referees for detailed and useful comments.

Appendix

Proof of Proposition 3: Let G be a finite non-cooperative game. Assume that $\delta^* \in \Delta$ is a Nash equilibrium of G. We must show that $\delta^* \in \Phi(\delta^*)$. In fact, we need to show that for the prior strategy δ^* the linear tracing procedure will feasibly select δ^* as the outcome.

Consider the games Γ^t invoked by the linear tracing procedure based on δ^*. We will show that δ^* is a Nash equilibrium for any game Γ^t for $t \in [0,1]$, which of course suffices to establish the result that $\delta^* \in \Phi(\delta^*)$.

According to the definition of Nash equilibrium, we have that $\delta_i^* \in B_i(\delta_{-i}^*)$ for every player i. More precisely, it means that for any player i

$$\delta_i^* \in B_i(\delta_{-i}^*) = \{\delta_i \in \Delta_i \mid u_i(\delta_i, \delta_{-i}^*) \geq u_i(\delta_i', \delta_{-i}^*), \forall \delta' \in \Delta_i\}.$$

Consider first the separable game Γ^0. Since the strategy δ^* is the prior strategy, the payoff functions for each player i are given by $u_i^0(\delta_i, \delta_{-i}) = u_i(\delta_i, \delta_{-i}^*)$ for any $\delta_i \in \Delta_i$. Thus, for any $\delta \in \Delta$ the best response correspondence B_i^0 can be represented by $B_i^0(\delta_{-i}) = \{\delta_i \in \Delta_i \mid u_i(\delta_i, \delta_{-i}^*) \geq u_i(\delta_i', \delta_{-i}^*),$ for all $\delta_i' \in \Delta_i\}$. Clearly, it implies $B_i^0(\delta_{-i}) = B_i(\delta_{-i}^*)$ for any $\delta \in \Delta$. We then have that $\delta_i^* \in B_i^0(\delta_{-i}^*)$ for every player i, and thus $\delta^* \in NE_{\Gamma^0}$.

Now, let us consider the generic games Γ^t. First, the payoff functions for each player i are given by $u_i^t(\delta_i, \delta_{-i}) = tu_i(\delta_i, \delta_{-i}) + (1-t)u_i(\delta_i, \delta_{-i}^*)$. Thus, the best response correspondence B_i^t can be represented as follows.

$$\begin{aligned} B_i^t(\delta_{-i}) = \{&\delta_i \in \Delta_i \mid tu_i(\delta_i, \delta_{-i}) + (1-t)u_i(\delta_i, \delta_{-i}^*) \\ &\geq tu_i(\delta_i', \delta_{-i}) + (1-t)u_i(\delta_i', \delta_{-i}^*), \forall \delta_i' \in \Delta_i\}. \end{aligned}$$

In particular, $B_i^t(\delta_{-i}^*) = \{\delta_i \in \Delta_i \mid tu_i(\delta_i, \delta_{-i}^*) + (1-t)u_i(\delta_i, \delta_{-i}^*) \geq tu_i(\delta_i', \delta_{-i}^*) + (1-t)u_i(\delta_i', \delta_{-i}^*), \forall \delta_i' \in \Delta_i\}$. It follows that

$$B_i^t(\delta_{-i}^*) = \{\delta_i \in \Delta_i \mid u_i(\delta_i, \delta_{-i}^*) \geq u_i(\delta_i', \delta_{-i}^*), \forall \delta_i' \in \Delta_i\},$$

which is independent of the value of t. This means that $B_i^t(\delta_{-i}^*) = B_i(\delta_{-i}^*)$ for each player i.

Hence we have that $\delta_i^* \in B_i^t(\delta_{-i}^*)$ for every player i, and thus $\delta^* \in NE(\Gamma^t)$ for $t \in [0,1]$. We can therefore conclude that $\delta^* \in \Phi(\delta^*)$. ■

In order to prove our main result (Proposition 5), we need to show some properties concerning Nash equilibria of convergent sequences of games. Since the results are required in proving the main result, we first present their proofs.

Proof of Proposition 7: Suppose that $\{G^r\}$ converges to game G. Assume for contradiction that δ is not a Nash equilibrium of the limit game G. Then there exists some player i with some $t_i \in \Delta_i$ such that

$$u_i(\delta) < u_i(t_i, \delta_{-i}).$$

First, note that $\{G^r\}$ converges to G which thus implies that u_i^r converges uniformly to u_i. Thus we can find a continuous approximation of u_i, denoted by u_i^j, such that

$$u_i^j(\delta) < u_i^j(t_i, \delta_{-i}).$$

Moreover, we know that the sequence $\{\delta^r\}$ converges pointwise to δ, and thus converges uniformly to δ. Hence, when j is large enough, we have that

$$u_i^j(\delta^j) < u_i^j(t_i, \delta_{-i}^j),$$

which contradicts the assumption that δ^j is a Nash equilibrium of game G^j. Therefore, δ must be a Nash equilibrium of the limit game G. ■

Proof of Corollary 8: Suppose that $p^r \to p$ and $t \in [0,1]$. First, we show that the sequence of games $\{\Gamma_{p^r}^t\}$ converges to game Γ_p^t. As described in the linear tracing procedure, the sets of players and the strategy spaces, denoted by I and Δ respectively, are all the same as the original game G. Note that the payoff function of game $\Gamma_{p^r}^t$ is given by

$$u_{i,p^r}^t(\delta_i, \delta_{-i}) = tu_i(\delta_i, \delta_{-i}) + (1-t)u_i(\delta_i, p_{-i}^r)$$

where u_i is the player i's payoff function in the original game G. Note that the first term on the right side is independent of p^r. And since it is assumed that u_i is continuous, it thus follows from $\{p^r\} \to p$ that $\{u^t_{i,p^r}\}$ converge to $u^t_{i,p}$. Hence, the sequence $\{\Gamma^t_{p^r}\}$ converges to Γ^t_p. Moreover, it is assumed that δ^r are Nash equilibria relative to game $\Gamma^t_{p^r}$ with $\{\delta^r\} \to \delta$. Therefore by Proposition 7, δ is a Nash equilibrium of the limit game Γ^t_p. ■

Proof of Corollary 9: Suppose that $\{p^r\} \to p$ and $\{t^m\} \to t$ where $t^m, t \in [0,1]$ and $p^r, p \in \Delta$. In order to apply Proposition 7, we have to show that the sequence of games $\{\Gamma^{t^m}_{p^r}\}$ converges to game Γ^t_p. Similarly, we have that the sets of players and the strategy spaces are all the same as the original game G, denoted by I and Δ respectively. Now consider the payoff function of game $\Gamma^{t^m}_{p^r}$ which is defined as

$$u^{t^m}_{i,p^r}(\delta_i, \delta_{-i}) = t^m u_i(\delta_i, \delta_{-i}) + (1 - t^m) u_i(\delta_i, p^r_{-i})$$

where u_i is the player i's payoff function in the original game G. Note that u_i is assumed to be continuous. And since $\{t^m\} \to t$ and $\{p^r\} \to p$, it implies that $\{u^{t^m}_{i,p^r}\}$ converge to $u^t_{i,p}$. Thus, according to the defintion of convergent sequence of games, we have that the sequence of games $\{\Gamma^{t^m}_{p^r}\}$ converges to game Γ^t_p. And since it is assumed that $\delta^{m,r}$ are Nash equilibria relative to game $\Gamma^{t^m}_{p^r}$ with $\{\delta^{m,r}\} \to \delta$, it follows from Proposition 7 that δ is a Nash equilibrium of the limit game Γ^t_p. ■

With the aid of the foregoing results, we can now present the proof of our main result in section 2.

Proof of Proposition 5: Let $p \in \Delta$ be a limit point of the source set $\Phi(\delta^*)$. This means that there exists some sequence of priors $\{p^r\}$ such that $p^j \in \Phi(\delta^*)$ for each $p^j \in \{p^r\}$ and $\{p^r\}$ pointwise converges to p, i.e., $\{p^r\} \to p$. According to the definition, $p^j \in \Phi(\delta^*)$ means that there exists a feasible path, denoted by L_{p^j}, connecting the starting point $\delta^0_{p^j}$ and the end point δ^*, where $\delta^0_{p^j}$ is a Nash equilibrium of the game $\Gamma^0_{p^j}$ corresponding to the separable game that used p^j as the prior strategy. Here for $t \in [0,1]$ and $p^j \in \{p^r\}$, we let $\Gamma^t_{p^j}$ denote the game generated by using p^j as the prior strategy, and let $\delta^t_{p^j}$ denote the Nash equilibrium point(s) of game $\Gamma^t_{p^j}$ appearing on the feasible path L_{p^j}.

We must show that there exists a feasible path L_p for p which connects some equilibrium point(s) of game Γ^0_p to δ^*. Clearly, the set of t-values T is totally bounded, and thus can be covered by finitely many sets, each of which is centered at a point of T with diameter at most ϵ, for any $\epsilon > 0$. Now let $\epsilon > 0$. The set T can then be written as the union of finitely many sets with diameters $< \epsilon$. Let us denote these sets by $T_1, \ldots, T_m$. To show the existence of such a feasible path L_p, let us consider whether infinitely many feasible paths of $\{L_{p^j}\}$ have continuous segments of equilibrium points for the corresponding games at these sets $T_1, \ldots, T_m$.

Case 1: There is no such set where infinitely many feasible paths of $\{L_{p^j}\}$ have continuous segments of equilibrium points. This implies that either all the feasible paths are straight lines or only finite many feasible paths have continuous segments of equilibrium points somewhere.

First, consider the former case. Since all the feasible paths L_{p^j} are straight lines passing from some points to the same point δ^*, these feasible paths thus can be fully characterized by the corresponding slopes of the lines. Note that we are considering only finite games. It thus follows that the strategy space can be viewed as a subset of a finite-dimensional Euclidean space R^n, and the slopes of the feasible paths must be bounded to a certain region. Recall that by the Bolzano–Weierstrass theorem, each bounded sequence in R^n has a convergent subsequence. Thus, there exists a convergent subsequence of the slopes of the straight feasible paths, which means that there exists some subsequence of the straight feasible paths converging to a straight line determined by the limit slope, denoted by L_p. And we know that each feasible path corresponds to a prior strategy in the sequence $\{p^r\}$, and, therefore, that convergent subsequence of the straight feasible paths also correspond to a convergent subsequence of $\{p^r\}$, denoted by $\{p^{r'}\}$. Of course, the subsequence $\{p^{r'}\}$ must converge to the same limit as $\{p^{r'}\}$, that is $\{p^{r'}\} \to p$.

Now consider the subsequence $\{p^{r'}\}$ converging to p. As pointed out above, for each $t \in [0,1]$ the sequence of games $\{\Gamma^t_{p^{r'}}\}$ converges to Γ^t_p. Since that subsequence of the straight feasible paths converges to a straight line L_p, the sequence of Nash equilibria $\{\delta^t_{p^{r'}}\}$ thus converges to δ^t_p for each $t \in [0,1]$. It thus follows from Corollary 9 that δ^t_p must be a Nash equilibrium of game Γ^t_p, which shows that each point of the straight line L_p is one Nash equilibrium of the corresponding game Γ^t_p. Hence, the straight line L_p is a feasible path for p which connects some starting point belonging to Γ^0_p to the end point δ^*.

Now, consider the latter case, where only finite many feasible paths have continuous segments of equilibrium points somewhere. We can always ignore such feasible paths and only consider the other infinitely many feasible paths that are straight lines. Since each feasible path is associated with a prior p_j, there thus exists some subsequence $\{p^{r'}\}$ corresponding to these infinitely many feasible paths. Hence, the above argument can be applied to this subsequence $\{p^{r'}\}$. Therefore, in this case we have that there exists one feasible path L_p for p as well.

Case 2: There exists one and only one set, say T_k, where infinitely many feasible paths of $\{L_{p^j}\}$ have continuous segments of equilibrium points. Assume that the set T_k is centered at t_k. Similarly, we have that there exists some subsequence $\{p^{r'}\}$ corresponding to these infinitely many feasible paths, and $\{p^{r'}\} \to p$. We are going to show that there exists a feasible path L_p for p.

Note that all or infinitely many feasible paths of $\{L_{p^j}\}$ do not have any continuous segments in the interval $(t_k, 1]$. This means that there are infinitely many feasible paths that are straight lines in $(t_k, 1]$. A similar argument as that of case 1 shows that these infinitely many feasible paths converge to L_p in $(t_k, 1]$.

Now consider the set T_k. As described above, T_k is a set centered at t_k with diameter$< \epsilon$ where ϵ is arbitrarily small. We have that $\{p^{r'}\} \to p$, and the corresponding feasible paths of $\{p^{r'}\}$ have continuous segments of equilibrium points at T_k. Then, according to the Bolzano–Weierstrass theorem, there exists a subsequence $\{p^{r''}\}$ of $\{p^{r'}\}$ such that $\{p^{r''}\}$ uniformly converges to p with $\{t^m\} \to t_k$. Thus the corresponding continuous segments of equilibrium points uniformly converge to one continuous segment of equilibrium points for the game $\Gamma_p^{t_k}$. This implies that there exists one continuous segment of equilibrium points of the game $\Gamma_p^{t_k}$. Next, we show that the coming-in and coming-out points are exactly the two endpoints of this continuous segment. The reason is that the coming-in and coming-out points should be the limits of the coming-in and coming-out points of the infinitely many paths corresponding to $\{p^{r''}\}$, which must coincide with the the limits of the endpoints of these infinitely many paths. So far we have established that there exists a continuous path from t_k to 1, which has a continous segment at t_k.

Note again that there are infinitely many feasible paths of $\{L_{p^j}\}$ that are straight line in $[0, t_k)$. By a similar argument as in case 1, these infinitely many feasible paths converge to L_p in $[0, t_k)$. Taking these together, we can therefore conclude that there exists a feasible path L_p for p.

We can employ the above argument to examine all the sets $T_1, \ldots T_k$. Since these sets are finite, we know that there exists a feasible path L_p for p, which implies that $p \in \Phi(\delta^*)$. ■

Proof of Theorem 11: Assume that δ^* is an equilibrium of the game G, and δ^* maximizes all players' expected payoff with respect to p. In order to check whether $p \in \Phi^t(\delta^*)$, let us regard Γ_p^t as the original game, which can be represented as $\Gamma_p^t = \langle I, \{S_i\}, \{u_i^t\}\rangle_{i\in I}$, where, for each $\delta \in \Delta$, the payoff function u_i^t is defined as

$$u_i^t(\delta_i, \delta_{-i}) = tu_i(\delta_i, \delta_{-i}) + (1-t)u_i(\delta_i, p_{-i}).$$

We then consider a new one-parameter class of auxiliary games $\Lambda_p^{t'} = \langle I, \{S_i\}, \{\mu_i^{t'}\}\rangle_{i\in I}$ with $t' \in [0, 1]$, where, for each $\delta \in \Delta$, the payoff function $\mu_i^{t'}$ is given by

$$\mu_i^{t'}(\delta_i, \delta_{-i}) = t'u_i^t(\delta_i, \delta_{-i}) + (1-t')u_i^t(\delta_i, p_{-i}).$$

Obviously, $\Lambda_p^0 = \Gamma_p^0$, since the payoff functions are identical, that is, $\mu_i^0 = u_i^0$; and $\Lambda_p^1 = \Gamma_p^t$ for the same reason. In view of this, the class of auxiliary games $\Lambda_p^{t'}$ is a subset of the family of auxiliary games Γ_p^t with respect to the game G. In other words, when considering the linear tracing procedure with respect to game Γ_p^t, we are merely examining a small subset of the family of auxiliary games previously considered.

As was assumed, δ^* is an equilibrium point of G, that is, for each player i,

$$u_i(\delta_i^*, \delta_{-i}^*) \geq u_i(\delta_i, \delta_{-i}^*), \text{ for all } \delta_i \in \Delta_i.$$

Moreover, we assume that δ^* maximizes the expected payoff with respect to p, which means that $u_i(\delta_i^*, p_{-i}) \geq u_i(\delta_i, p_{-i})$ for each player i and each $\delta_i \in \Delta_i$. From these two conditions, it is easy to verify that $u_i^t(\delta_i^*, \delta_{-i}^*) \geq u_i^t(\delta_i, \delta_{-i}^*)$ for each player i and each $\delta_i \in \Delta_i$, which means that δ^* is an equilibrium of game Γ_p^t. Note that $u_i^t(\delta_i, p_{-i}) = u_i(\delta_i, p_{-i})$ for all $\delta_i \in \Delta_i$. Thus, we have that $u_i^t(\delta_i^*, p_{-i}) \geq u_i^t(\delta_i, p_{-i})$ for all $\delta_i \in \Delta_i$. Together, these two coniditions, which specfy the best response conditions for games Γ_p^t and Γ_p^0, guarantee the existence of a feasible path for the equilibrium δ^*. This point can be easily illustrated by the following inequality: for each player i and each $\delta_i \in \Delta_i$

$$\begin{aligned} \mu_i^{t'}(\delta_i^*, \delta_{-i}^*) &= t'u_i^t(\delta_i^*, \delta_{-i}^*) + (1-t')u_i^t(\delta_i^*, p_{-i}) \\ &\geq t'u_i^t(\delta_i, \delta_{-i}^*) + (1-t')u_i^t(\delta_i, p_{-i}) \\ &= \mu_i^{t'}(\delta_i, \delta_{-i}^*) \end{aligned}$$

Since this inequality holds for each player i and each $t' \in [0, 1]$, it implies that there exists a feasible path continuously connecting game Λ_p^0 to Γ_p^t. We can therefore conclude that $p \in \Phi^t(\delta^*)$ for each $t \in [0, 1]$. ■

References

[1] D. Fudenberg and J. Tirole, *Game Theory*, The MIT Press, 1991.

[2] J. C. Harsanyi and R. Selten, *A General Theory of Equilibrium Selection in Games*, The MIT Press, 1988.

[3] J. C. Harsanyi, The Tracing Porcedure: a Bayesian approach to defining a solution for n-person noncooperative games, *International Journal of Game Theory* **4**, pp. 61-94, 1975.

[4] P. J. J. Herings, Two Simple Proofs of the Feasibility of the Linear Tracing Procedure, *Economic Theory* **15**, pp. 485-490, 2000.

[5] E. Kohlberg and J. F. Mertens, On the Strategic Stability of Equilibria, *Econometrica* **54**(5), pp. 1003-1037, 1986.

[6] D. M. Kreps and R. Wilson, Sequential Equilibria, *Econometrica* **50**, pp. 863-894, 1982.

[7] S. H. Schanuel, L. K. Simon, and W. R. Zame, The Algebraic Geometry of Games and the Tracing Procedure. In: Selten, R. (ed.) *Game Equilibrium Models II: methods, morals and markets*, pp. 9-43. Berlin Heidelberg New York: Springer 1991.

[8] R. Selten, A reexamination of the perfectness concept for equilibrium points in extensive games, *International Journal of Game Theory* **4**, pp. 25-55, 1975.

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

Bounds for Self-consistent CDF Estimators for Univariate and Multivariate Censored Data

Xuecheng Liu
Research Unit on Children's Psychosocial Maladjustment,
University of Montreal,
Canada
xuecheng.liu@umontreal.ca

Alain C. Vandal
Faculty of Health and Environmental Sciences,
AUT University;
Centre for Clinical Research and effective practice,
New Zealand
alain.vandal@aut.ac.nz

Abstract

In this paper, lower bounds and upper bounds are given for the mass assigned to a set of maximal cliques in self-consistent estimates of CDF NPMLEs for multivariate (including univariate) interval censored data under the assumption that the censoring mechanism is ignorable for the purpose of likelihood inference. The bounds are applied to give upper bounds of the diameter and size of the polytope of CDF NPMLEs for multivariate censored data.

Keywords. Interval censoring, maximal clique, clique matrix, self-consistent estimator, bounds, NPMLE, mixture nonuniqueness

1 Introduction

Survival analysis is the statistical analysis of event times, assumed nonnegative. It must account for, and is largely characterized by, censoring. Censoring is a type of coarsening of the data whereby an event time is only known up to an interval. While *right-censored data* consist of exactly observed times and intervals unbounded on the right, collections of positive values and of bounded and (right-) unbounded intervals on the nonnegative half-line are known as *interval censored data*. Right-censoring will occur in studies where follow-up is limited by design at a deterministic or random time. Interval censored data will typically arise in medical longitudinal studies, where patients can be assessed for a condition continuously, or at regular or irregular intervals.

The first task to undertake given interval censored data is often to estimate the underlying cumulative distribution function (CDF) F or equivalently the survival function $S = 1 - F$. In many instances, a nonparametric approach will be preferred to the constraining assumption of a parametric form for the CDF. In such situations the nonparametric maximum likelihood estimator (NPMLE) of the CDF will be the estimator of choice in the univariate case (Peto [18], Turnbull [22]), even when smoothing estimators are sought (Braun, Duchesne & Stafford [3]).

Event times can sometimes be stochastically associated, for instance through clustering. It is then useful to treat them as multivariate. Multivariate interval censored data are geometrically represented as the Cartesian product of the marginal event times or intervals that enter in a given observation.

Computing the CDF NPMLE can be a complex endeavor. Generally this computation can be carried out in two phases: in the first the effective support of the NPMLE is determined (Gentleman & Vandal [9], Bogaerts & Lesaffre [2], Maathuis [16]). This effective support consists in the *real representations (RR)* of the *maximal cliques* of the data, concepts to be defined in Section 2; for now it suffices to describe an RR as a generalized, possibly degenerate, hypercube in $\mathbb{R}_0^{+d}$, ($\mathbb{R}_0^+ = [0, +\infty)$, $d = 1, 2, ...$), with edges parallel to the axes. In the second phase a nonparametric likelihood with the CDF as argument is maximized; the maximizer assigns a probability mass to each RR (Wang [26]).

The probability vector obtained thus completely characterizes the CDF NPMLE. It is worth noting that this probability vector is always unique with univariate data (Vandal [23]). Arbitrary mass placement within an RR does not however affect the nonparametric likelihood, a situation to which we refer as *R-nonuniqueness* (Gentleman & Vandal [10]). With multivariate data, the probability vector itself may not be unique, a somewhat more serious situation we label *M-nonuniqueness*.

In this paper, we are interested in obtaining lower and upper bounds of the total CDF NPMLE mass assigned to an RR or a set of RRs of maximal cliques *without* conducting NMPL estimation. This is done by considering bounds on a class of more general estimators, namely self-consistent estimators (SCE), to

which NPMLEs belong.

These bounds can be obtained much more quickly than the probability vector that maximizes the likelihood (whether unique or not). There are good reasons for providing such bounds. First, even when one NPMLE vector is available, M-nonuniqueness will prevent us from deducing bounds for the probability mass on a *collection* of RRs. Second, reliable lower and upper bounds may enable us to select good starting probability vectors for NPML estimation: currently all algorithms used in for NPML estimation with general interval censored data are iterative. Third, there are self-contained applications of the bounds; in Section 5, we use them to provide upper bounds for the diameter and size of the polytope of NPMLEs.

The present paper focuses on nonparametric (and non-smoothed) maximum likelihood estimation. In that respect it differs from works such as those of Ferson et al. [7], whose statistical focus lies in parametric analysis with some forays in smoothing estimators. It also differs from the works such as that of Manski [17], that focus on the consequences of unobservability. This paper can be thought of as an inferential addition to the "catalogue" of techniques for symbolic data analysis, described in Billard & Diday [1].

We will assume in the sequel that the true CDF and the CDF NPMLE have support in $\mathbb{R}_0^{+d}$. We will also assume that the censoring mechanism is ignorable in the sense of Heitjian & Rubin [12], which implies in particular that likelihood-based inference relying on the data can be performed without reference to the censoring mechanism. A sufficient condition for ignorability of the censoring mechanism is for the underlying inspection process to be independent of the event times.

The rest of the paper is divided into 4 sections. In Section 2, we provide some necessary concepts and notation. In Section 3, we provide SCE bounds for any given collection of maximal clique RRs. In Section 4, we consider two special cases: one concerns the bounds on the SCE mass of a single maximal clique; the other the bounds on the SCEs given univariate censored data. In Section 5, we apply SCE bounds to give upper bounds of the diameter and size of the polytope of CDF NPMLEs for multivariate censored data.

2 Preliminaries and Notation

We provide some concepts and notation used in subsequent sections.

Let $R_1, ..., R_n$ be the n observations of an interval censored data set in $\mathbb{R}_0^{+d}$. Throughout this paper, we *always* assume that the censoring mechanism is ignorable in the sense of Heitjian & Rubin [12], which implies in particular that likelihood-based inference relying on the data can be performed without reference to the censoring mechanism. A sufficient condition for ignorability of the censoring mechanism is for the underlying inspection process to be independent of the event times. For any CDF F, the likelihood of F given the data is

$$L(F) = \prod_{i=1}^{n} P_F(R_i). \tag{1}$$

2.1 Intersection Graph, Maximal Clique, Clique Matrix, Real Representation

We can form the *intersection graph* of the data set in the following way: each observation corresponds to a vertex and two vertices are connected if and only if their corresponding observations intersect. A *clique* is a subset of vertices such that every pair are connected. A clique is called *maximal* if it is not a proper subset of another clique. The clique structure can be represented by the *clique matrix*, which is a 0/1 matrix, each row corresponding to a maximal clique and each column corresponding to an observation. An entry in the clique matrix is 1 if and only if the corresponding observation (i.e., vertex) belongs to the corresponding maximal clique. The clique matrix is unique up to permutations of rows and columns. In addition, each maximal clique has a real representation (RR), namely, the intersection of all its observations. The following is an illustrative example.

Example 2.1 *Let R_i, $i = 1, ..., 7$, be bivariate censored data as shown on Figure 1. Their intersection graph* [1] *is displayed in Figure 2. There are 4 maximal cliques M_1, M_2, M_3 and M_4:*

$$M_1 = \{R_1, R_2, R_4\}, \qquad M_2 = \{R_3, R_4, R_7\},$$
$$M_3 = \{R_4, R_5, R_6\} \quad \text{and} \quad M_4 = \{R_4, R_6, R_7\}.$$

Their corresponding maximal intersections (i.e., real representations of maximal cliques) are shaded in Figure 1. The clique matrix of these data is given in Table 1.

2.2 NPML and Self-consistent Estimators of the CDF

The importance of maximal cliques lies in two facts: the possible support of NPMLE is limited to the RRs

[1] Note that each R_i intersects itself and hence corresponds to a loop in the intersection graph. The loops are ignored in Figure 1.

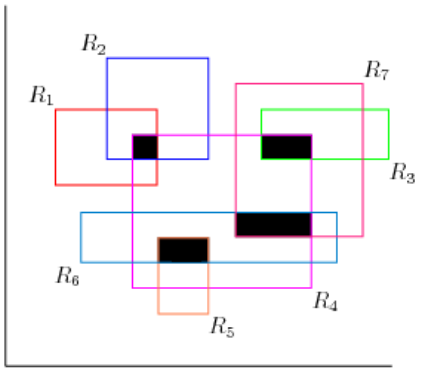


Figure 1: An example of bivariate interval censored data set

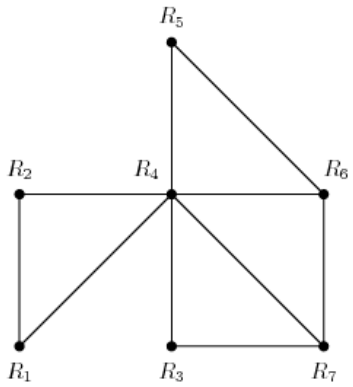


Figure 2: The intersection graph for the data in Figure 1

of maximal cliques; and the clique matrix is sufficient for the probability vector corresponding to the NPMLE. For a detailed discussion of the first fact, see Peto [18] and Turnbull [22]. For the second, refer to Gentleman & Vandal [10]. In the multivariate case, maximal cliques are most efficiently identified using the HeightMap algorithm of Maathuis [16] and the marked iso-graph algorithm (Liu [15]).

Suppose we have m maximal cliques $M_1, ..., M_m$, which are assigned masses $p_1, ..., p_m$ respectively. The likelihood (1) can then be redefined as a function of $\mathbf{p}$:

$$L(\mathbf{p}) = \prod_{j=1}^{n} \sum_{i=1}^{m} a_{ij} p_i, \tag{2}$$

where a_{ij}s valued in $\{0, 1\}$ are the entries of the clique matrix $\mathbf{A}_{m\times n}$. (That is, $a_{ij} = 1$ if and only if the observation R_j is in the maximal clique M_i.) The NPMLE corresponds to a probability vector $\mathbf{p} = [p_1, ..., p_m]'$. An NPMLE of the CDF will be constant except for increases of sizes p_i concentrated on the on the real representations of the maximal cliques.

	R_1	R_2	R_3	R_4	R_5	R_6	R_7
M_1	1	1	0	1	0	0	0
M_2	0	0	1	1	0	0	1
M_3	0	0	0	1	1	1	0
M_4	0	0	0	1	0	1	1

Table 1: Clique matrix of the data in Example 2.1

The precise placement of the mass within the real representations does not affect the likelihood, a situation to which we refer as R-nonuniqueness.

An important feature of a CDF NPMLE under censored data is that it must satisfy the self-consistency condition (Turnbull [22]). There are several equivalent definitions of self-consistency of estimators in the literature. We use the following, which precisely identifies fixed points of the EM algorithm:

Definition 2.2 *Let $\mathbf{A}_{m\times n}$ be the clique matrix for the multivariate censored data. A probability vector $\tilde{\mathbf{p}}$ is a self-consistent estimate if and only if*

$$n\tilde{\mathbf{p}} = \mathbf{D}_{\tilde{\mathbf{p}}}\mathbf{A}(\mathbf{A}'\tilde{\mathbf{p}})^{-\mathbf{I}}, \tag{3}$$

where $\mathbf{I}$ is the identity matrix of order m, and

- $\mathbf{D}_{\mathbf{x}}$ denotes the diagonal matrix with diagonal $\mathbf{x}$;
- For any column vector $\mathbf{a}_{m\times 1} := [a_1, ..., a_m]'$, $a_i \neq 0$, $\mathbf{a}^{-\mathbf{I}}$ is the column vector whose $i-$th element is $1/a_i, i = 1, 2, ..., m$. [2]

The product-limit estimator for univariate right-censored data, first proposed by Kaplan & Meier [13], was later shown by Efron [6] to be self-consistent. Turnbull [21, 22] then used self-consistency as the basis for an estimation algorithm, later shown in Dempster, Laird & Rubin [5] to be a particular application of the EM algorithm. It is now a well recognized fact (Groeneboom & Wellner [11], Gentleman & Geyer [8], Wellner & Zhan [27]) that in general there exist several distinct values of $\tilde{\mathbf{p}}$ which are self-consistent but do not maximize the likelihood. In order to be the NPMLE, a self-consistent estimate must also satisfy the Kuhn-Tucker conditions listed in Gentleman & Geyer [8].

2.3 Further Notation

Let $\mathcal{C}$ be a set of maximal cliques of a multivariate censored data (MCD) set with n observations. Throughout this chapter, we use $\tilde{\mathbf{p}}$ to denote a self-consistent

[2] The notation $\mathbf{a}^{-\mathbf{I}}$ is a special case of Hadamard exponentiation. For more detailed information, see Gentleman & Vandal [9].

estimate. For such an estimate, define $\tilde{\mathbf{p}}_{\mathcal{C}}$ to be the total mass assigned to $\mathcal{C}$.

Let $n^+(\mathcal{C})$ and $n^-(\mathcal{C})$ be the numbers of observations in $\underset{C\in\mathcal{C}}{\cup} C$ and only in $\underset{C\in\mathcal{C}}{\cup} C$ respectively. Equivalently, we may interpret $n^+(\mathcal{C})$ as the number of observations covering some maximal clique RRs in $\mathcal{C}$ and $n^-(\mathcal{C})$ as the number of the observations covering only some maximal clique RRs in $\mathcal{C}$. Formally,

$$n^+(\mathcal{C}) := \Big|\bigcup_{C\in\mathcal{C}} C\Big|$$

and

$$n^-(\mathcal{C}) := \Big|\bigcup_{C_1\in\mathcal{C}} C_1 \setminus \bigcup_{C_2\notin\mathcal{C}} C_2\Big|.$$

We have

$$n^-(\mathcal{C}) = n - n^+(\mathcal{C}^c)$$

where $\mathcal{C}^c$ is the complement of $\mathcal{C}$ with respect to the set of all maximal cliques.

3 Bounds on Self-consistent CDF Estimates for MCD: General Case

3.1 Main Result

The main result of this section is the following theorem.

Theorem 3.1 *Let $\mathcal{C}$ be a set of maximal cliques of an MCD set with n observations, there holds*

$$\frac{n^-(\mathcal{C})}{n} \leqslant \tilde{\mathbf{p}}_{\mathcal{C}} \leqslant \frac{n^+(\mathcal{C})}{n}. \tag{4}$$

Proof. First, we prove the right-hand side of (4), that is

$$\tilde{\mathbf{p}}_{\mathcal{C}} \leqslant \frac{n^+(\mathcal{C})}{n}. \tag{5}$$

Without any loss of generality, we assume that in the clique matrix $\mathbf{A}$, the first $|\mathcal{C}|$ rows correspond to maximal cliques in $\mathcal{C}$ and first $n^+(\mathcal{C})$ columns correspond to observations in $\underset{C\in\mathcal{C}}{\cup} C$. Therefore, $\mathbf{A}$ is of the form

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{O} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

where the size of $\mathbf{A}_{11}$ is $|\mathcal{C}|$ by $|n^+(\mathcal{C})|$ and $\mathbf{O}$ denotes a matrix whose entries are all 0.

Rewrite $\tilde{\mathbf{p}}$ as $\tilde{\mathbf{p}} = \begin{bmatrix} \tilde{\mathbf{p}}_1 \\ \tilde{\mathbf{p}}_2 \end{bmatrix}$, where $\tilde{\mathbf{p}}_1 \in \mathbb{R}_+^{|\mathcal{C}|}$ and $\tilde{\mathbf{p}}_2 \in \mathbb{R}_+^{m-|\mathcal{C}|}$. Then $\tilde{\mathbf{p}}_{\mathcal{C}} = \sum_{i=1}^{|\mathcal{C}|} \tilde{p}_i$. Also let $\mathbf{I}$, $\mathbf{I}_1$ and $\mathbf{I}_2$ be the identity matrices of orders m, $|\mathcal{C}|$ and $m - |\mathcal{C}|$ respectively. The self-consistency condition on $\tilde{\mathbf{p}}$ becomes

$$\begin{aligned}
\begin{bmatrix} \tilde{\mathbf{p}}_1 \\ \tilde{\mathbf{p}}_2 \end{bmatrix} &= \frac{1}{n}\begin{bmatrix} \mathbf{D}_{\tilde{\mathbf{p}}_1} & \mathbf{O} \\ \mathbf{O} & \mathbf{D}_{\tilde{\mathbf{p}}_2} \end{bmatrix}\begin{bmatrix} \mathbf{A}_{11} & \mathbf{O} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \\
&\quad \left(\begin{bmatrix} \mathbf{A}'_{11} & \mathbf{A}'_{21} \\ \mathbf{O} & \mathbf{A}'_{22} \end{bmatrix}\begin{bmatrix} \tilde{\mathbf{p}}_1 \\ \tilde{\mathbf{p}}_2 \end{bmatrix}\right)^{-\mathbf{I}} \\
&= \frac{1}{n}\begin{bmatrix} \mathbf{D}_{\tilde{\mathbf{p}}_1} & \mathbf{O} \\ \mathbf{O} & \mathbf{D}_{\tilde{\mathbf{p}}_2} \end{bmatrix}\begin{bmatrix} \mathbf{A}_{11} & \mathbf{O} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \\
&\quad \begin{bmatrix} \mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2 \\ \mathbf{A}'_{22}\tilde{\mathbf{p}}_2 \end{bmatrix}^{-\mathbf{I}} \\
&= \frac{1}{n}\begin{bmatrix} \mathbf{D}_{\tilde{\mathbf{p}}_1} & \mathbf{O} \\ \mathbf{O} & \mathbf{D}_{\tilde{\mathbf{p}}_2} \end{bmatrix}\begin{bmatrix} \mathbf{A}_{11} & \mathbf{O} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \\
&\quad \begin{bmatrix} (\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_1} \\ (\mathbf{A}'_{22}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_2} \end{bmatrix},
\end{aligned}$$

which implies that

$$\tilde{\mathbf{p}}_1 = \frac{1}{n}\mathbf{D}_{\tilde{\mathbf{p}}_1}\mathbf{A}_{11}(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_1}.$$

Hence, by letting $\mathbf{e}$ be the vector $[1]_{|\mathcal{C}|\times 1}$,

$$\begin{aligned}
\sum_{i=1}^{|\mathcal{C}|}\tilde{p}_i &= \mathbf{e}'\tilde{\mathbf{p}}_1 \\
&= \frac{1}{n}\mathbf{e}'\mathbf{D}_{\tilde{\mathbf{p}}_1}\mathbf{A}_{11}(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_1} \\
&= \frac{1}{n}\tilde{\mathbf{p}}'_1\mathbf{A}_{11}(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_1} \\
&= \frac{1}{n}(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1)'(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_1}.
\end{aligned}$$

Since $\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 \geqslant \mathbf{0}$ and $\mathbf{A}'_{21}\tilde{\mathbf{p}}_2 \geqslant \mathbf{0}$, we have

$$\begin{aligned}
\sum_{i=1}^{|\mathcal{C}|}\tilde{p}_i &\leqslant \frac{1}{n}(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2)'(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_1} \\
&= \frac{1}{n}\times n^+(\mathcal{C})
\end{aligned}$$

and (5) is proved.

Now we show the left part of (4). Denoting by $\mathcal{C}^c$ the complement of the set $\mathcal{C}$ of maximal cliques, we obtain from (5)

$$\tilde{\mathbf{p}}_{\mathcal{C}^c} \leqslant \frac{n^+(\mathcal{C}^c)}{n}$$

and therefore

$$\tilde{\mathbf{p}}_{\mathcal{C}} = 1 - \tilde{\mathbf{p}}_{\mathcal{C}^c} \geqslant 1 - \frac{n^+(\mathcal{C}^c)}{n} = \frac{n - n^+(\mathcal{C}^c)}{n} = \frac{n^-(\mathcal{C})}{n}.$$

The proof is complete. □

Note that, in the proof of Theorem 3.1, since $(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 + \mathbf{A}'_{21}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_1} > \mathbf{0}$ (without which the notation $(\mathbf{A}'_{11}\tilde{\mathbf{p}}_1 +$

$\mathbf{A}'_{21}\tilde{\mathbf{p}}_2)^{-\mathbf{I}_1}$ does not make sense), the equality in (5) is valid (that is, $\tilde{\mathbf{p}}_{\mathcal{C}}$ reaches its upper bound in (5)) if and only if $\mathbf{A}'_{21}\tilde{\mathbf{p}}_2 = 0$. The latter condition is equivalent to the following statement: if the r^{th} entry in $\tilde{\mathbf{p}}_2$ is positive, then the r^{th} row of $\mathbf{A}_{21}$ is a zero row-vector.

Specifically, when $\mathbf{A}_{21}$ is the null matrix, $\tilde{\mathbf{p}}_{\mathcal{C}}$ reaches its upper bound in (5). In this case, the observations $\mathcal{R}$ corresponding to $\mathbf{A}$ can be divided into two groups: the observations which are only in $\mathcal{C}$ and observations only in $\mathcal{C}^c$. A similar argument is applicable to the left-hand side of (4). We can therefore conclude that (4) cannot be improved for any data set.

The lower and upper bounds described by Theorem 3.1 correspond to belief and plausibility measures in Dempster-Shafer Theory ([4, DST]). These measures are obtained from a basic assignment induced by equiprobability on the original data; this basic assignment is normalized to the set of maximal cliques rather than the power set of the data. To our knowledge this is the first time a relationship is established (via self-consistency) between Dempster-Shafer theory and non-smoothing/non-penalized nonparametric likelihood estimation.

3.2 Two Examples

Example 3.2 *Consider the data depicted in Figure 3: Applying (4) to the data, we obtain the bounds shown in Table 3. The "True region" heading indicates the bounds on the total mass of $\hat{\mathbf{p}}_{\mathcal{C}}$ implied by the M-nonuniqueness of the NPMLE. Indeed, the two end-points of the true region are the lower and upper probabilities defined by the NPMLE probability vectors.*

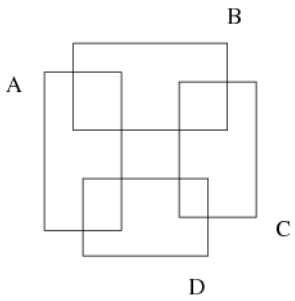


Figure 3: An example bivariate censored data set

Example 3.3 *Consider Pruitt's data (Pruitt [19]) depicted in Figure 4. Applying (4) to the data, bounds of $\tilde{\mathbf{p}}_{\mathcal{C}}$ for some given subsets $\mathcal{C}$ of maximal cliques are given in Table 5.*

	A	B	C	D
M_1	1	1	0	0
M_2	0	1	1	0
M_3	0	0	1	1
M_4	1	0	0	1

Table 2: The clique matrix for the data on Figure 3

$\mathcal{C}$	Lower bound	Upper bound	True region
$\{\tilde{p}_1\}, \{\tilde{p}_2\}, \{\tilde{p}_3\}, \{\tilde{p}_4\}$	0	2/4	[0, 0.5]
$\{\tilde{p}_1{+}\tilde{p}_2\}, \{\tilde{p}_2{+}\tilde{p}_3\},$ $\{\tilde{p}_3{+}\tilde{p}_4\}, \{\tilde{p}_4{+}\tilde{p}_1\}$	1/4	3/4	[0.5, 0.5]
$\{\tilde{p}_1{+}\tilde{p}_3\}, \{\tilde{p}_2{+}\tilde{p}_4\}$	0/4	4/4	[0,1]
$\{\tilde{p}_i{+}\tilde{p}_j{+}\tilde{p}_k,$ $1 \leqslant i < j < k \leqslant 4\}$	2/4	4/4	[0.5, 1]

Table 3: Mass bounds on $\mathbf{p}_{\mathcal{C}}$ for the data set in Example 3.2

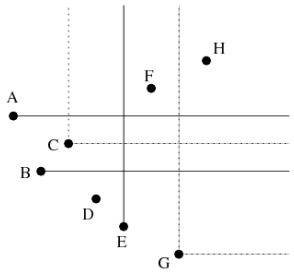


Figure 4: Pruitt's data set

	A	B	C	D	E	F	G	H
M_1	0	0	0	1	0	0	0	0
M_2	0	1	0	0	1	0	0	0
M_3	1	0	1	0	1	0	0	0
M_4	0	0	1	0	0	1	0	0
M_5	0	1	0	0	0	0	1	0
M_6	1	0	1	0	0	0	1	0
M_7	0	0	1	0	0	0	1	1

Table 4: The clique matrix for Pruitt's data set

$\mathcal{C}$	Lower bound	Upper bound	True region
$\{\tilde{p}_1\}$	1/8	1/8	[0.125, 0.125]
$\{\tilde{p}_2\}$	0/8	2/8	[0.095,0.191]
$\{\tilde{p}_5 + \tilde{p}_6\}$	0/8	4/8	[0.096, 0.096]
$\{\tilde{p}_2{+}\tilde{p}_3{+}\tilde{p}_5{+}\tilde{p}_6\}$	3/8	5/8	[0.457, 0.457]

Table 5: Mass bounds on $\mathbf{p}_{\mathcal{C}}$ for some $\mathcal{C}$'s for the data set in Example 3.3

3.3 Discussion

For uncensored data, the lower and upper bounds in (4) are always equal. Hence, (4) is an extension from uncensored to censored data.

M-nonuniqueness of NPMLEs for MCD can potentially create large differences between the lower and upper bounds in (4). From Examples 3.2 and 3.3, we notice that some intervals are wide and that we get no information at all in some cases. For instance, in Example 3.2, the lower and upper bounds for $\tilde{p}_1+\tilde{p}_3$ are 0 and 1 respectively. Note, however, that tighter bounds on $\tilde{p}_1+\tilde{p}_3$ are not available, since, the M-nonuniqueness interval of $\tilde{p}_1+\tilde{p}_3$ is $[0,1]$.

4 The Bounds in some Special Cases

4.1 Bounds on the SCE Mass of a Single Maximal Clique

In this section, we focus on the bounds for the mass assigned to one maximal clique by an SCE.

Theorem 4.1 *Let M_i be any maximal clique of an MCD set with n observations, there holds*

$$\frac{n^-(\{M_i\})}{n-n^+(\{M_i\})+n^-(\{M_i\})} \leqslant \tilde{p}_i \leqslant \frac{n^+(\{M_i\})}{n}. \quad (6)$$

Note that, the lower bound in (6) improves the lower bound in (4) in Section 3, and the upper bounds in (6) and (4) are the same.

Proof of Theorem 4.1 . We only need to show the left-hand side of (6), that is

$$\tilde{p}_i \geqslant \frac{n^-(\{M_i\})}{n-n^+(\{M_i\})+n^-(\{M_i\})}.$$

Denote by

$$\mathcal{J}_i := \{j; R_j \in M_i\}$$

the index set of $M_i \in \mathcal{M}$. So,

$$|\mathcal{J}_i| = n^+(\{M_i\}).$$

Also, denote

$$\tilde{\eta} := \mathbf{A}'\tilde{\mathbf{p}}.$$

Clearly, for every $i=1,...,m$ and all $j \in \mathcal{J}_i$,

$$\tilde{p}_i \leqslant \eta_j \leqslant 1. \quad (7)$$

Put

$$\mathcal{S}_i = \{j \in \mathcal{J}_i; R_j \text{ is only contained in } \mathcal{M}_i\}. \quad (8)$$

Then $|\mathcal{S}_i| = n^-(\{M_i\})$ and hence,

$$\begin{aligned} n &= \sum_{j\in\mathcal{J}_i}\frac{1}{\eta_j} \\ &= \frac{n^-(\{M_i\})}{\tilde{p}_i} + \sum_{j\in\mathcal{J}_i\setminus\mathcal{S}_i}\frac{1}{\tilde{\eta}_j} \quad (9) \\ &\geqslant \frac{n^-(\{M_i\})}{\tilde{p}_i} + \sum_{j\in\mathcal{J}_i\setminus\mathcal{S}_i} 1 \qquad \text{[from (7)]} \\ &= \frac{n^-(\{M_i\})}{\tilde{p}_i} + n^+(\{M_i\}) - n^-(\{M_i\}) \quad (10) \end{aligned}$$

whence the result follows.

Note that

$$n^+(\{M_i\}) = n^-(\{M_i\})$$

if and only if

$$n^-(\{M_i\})/(n-n^+(\{M_i\})+n^+(\{M_i\}) = n^+(\{M_i\})/n.$$

□

4.2 Bounds on Self-consistent Estimates for Univariate Censored Data

In this section, we give the form of (4) and (6) for univariate censored data based on the characteristic matrix notation introduced by Vandal [23]. For univariate censored data, we further improve the lower bound for one maximal clique in (6).

4.2.1 Characteristic Matrix for Univariate Data

The clique matrix of a univariate censored data set is equivalent to its characteristic matrix, defined as follows.

Definition 4.2 *Let $\mathbf{A} = [a_{ij}]_{m\times n}$ be the clique matrix for a univariate censored data set $\{R_1,...,R_n\}$ with maximal cliques $M_1,...,M_m$ and corresponding RRs $H_1,...,H_m$, ordered in the natural way. For each pair $i,j \in \{1,...,m\}$ with $i \leqslant j$, define $\chi_{i,j}$ to be the number of columns in $\mathbf{A}$ such that the sub-column of 1's starts at i and ends at j.*[3] *The following upper-right triangle matrix*

$$\boldsymbol{\chi} := \begin{bmatrix} \chi_{1,1} & \chi_{1,2} & \cdots & \chi_{1,m-1} & \chi_{1,m} \\ & \chi_{2,2} & \cdots & \chi_{2,m-1} & \chi_{2,m} \\ & & \ddots & \vdots & \vdots \\ & & & \chi_{m-1,m-1} & \chi_{m-1,m} \\ & & & & \chi_{m,m} \end{bmatrix}$$

is called the characteristic matrix of the data. [4]

[3] Recall that the clique matrix of univariate censored data has the consecutive-1's property.

[4] The lower-left triangle in characteristic matrix is left undefined.

Example 4.3 *The following is the clique matrix of a univariate censored data set:*

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

The equivalent characteristic matrix is

$$\boldsymbol{\chi} = \begin{bmatrix} 1 & 2 & 1 & 0 \\ & 0 & 3 & 0 \\ & & 0 & 2 \\ & & & 2 \end{bmatrix}.$$

4.2.2 Bounds on SCEs for Univariate Censored Data

The inequalities (4) for univariate censored data can be expressed using the entries of the characteristic matrix. Let $\mathbf{A}_{m\times n}$ be the clique matrix for a univariate data set with rows ordered according to the natural order of the maximal cliques[5]. Let $\tilde{\mathbf{p}} = [\tilde{p}_i]_{m\times 1}$ be a self-consistent estimate based on $\mathbf{A}$. Also, in this section, we always assume that $\chi_{1,m} = 0$ in $\mathbf{A}$'s characteristic matrix $\boldsymbol{\chi}$, since $\chi_{1,m}$ corresponds universal observations and have no bearing on the CDF estimation. For any given $j \in \{1, ...m\}$, let $\mathcal{C} := \{M_1, ..., M_k\}$. We then have

$$n^-(\mathcal{C}) = \sum_{\substack{s\leqslant k \\ r \text{ free}}} \chi_{rs} = \sum_{s=1}^{k}\sum_{r=1}^{s} \chi_{rs},$$

and

$$n^+(\mathcal{C}) = \sum_{\substack{r\leqslant k \\ s \text{ free}}} \chi_{rs} = \sum_{r=1}^{k}\sum_{s=r}^{m} \chi_{rs}.$$

From (4), the bounds on $\sum_{i=1}^{k} \tilde{p}_i$ can be given as

Theorem 4.4

$$\frac{1}{n}\sum_{s=1}^{k}\sum_{r=1}^{s}\chi_{r,s} \leqslant \sum_{i=1}^{k}\tilde{p}_i \leqslant \frac{1}{n}\sum_{r=1}^{k}\sum_{s=r}^{m}\chi_{r,s} \qquad (11)$$

Corollary 4.5 *When $j > 1$,*

$$\sum_{i=j}^{k}\tilde{p}_i \geqslant \frac{1}{n}\left(\sum_{s=1}^{k}\sum_{r=1}^{s}\chi_{r,s} - \sum_{r=1}^{j-1}\sum_{s=r}^{m}\chi_{r,s}\right), \quad (12)$$

$$\sum_{i=j}^{k}\tilde{p}_i \leqslant \frac{1}{n}\left(\sum_{r=1}^{k}\sum_{s=r}^{m}\chi_{r,s} - \sum_{s=1}^{j-1}\sum_{r=1}^{s}\chi_{r,s}\right). \quad (13)$$

[5] That is, $H < H'$ if and only if $x < x'$ for all $x \in H$ and $x' \in H'$.

Proof. Apply (11) to $\sum_{i=j}^{k}\tilde{p}_i$ and $\sum_{i=1}^{j-1}\tilde{p}_i$ and subtract. □

When we focus on the bounds of mass for one maximal clique, it is not difficult to show that for $i = 1, ..., m$,

$$\begin{aligned} n^-(\{M_i\}) &= \chi_{i,i}, \\ n^+(\{M_i\}) &= \sum_{j_1\leqslant i\leqslant j_2} \chi_{j_1,j_2} =: n_i. \end{aligned}$$

from Theorem 4.1, we have

Theorem 4.6

$$\frac{\chi_{i,i}}{n - n_i + \chi_{i,i}} \leqslant \tilde{p}_i \leqslant \frac{n_i}{n}. \qquad (14)$$

If $0 < \chi_{i,i} < n_i < n$ for some $i = 1, ..., m$, then we can further improve the lower bound on $\tilde{p}_i$ in (14). First, for $i = 1, ..., m$, introduce the following notation:

$$\begin{aligned} l_i &:= \min\{r;\ \chi_{r,s} > 0 \text{ and } r \leqslant i \leqslant s\}, \\ u_i &:= \max\{s;\ \chi_{r,s} > 0 \text{ and } r \leqslant i \leqslant s\}, \\ d_i &:= \sum_{r=1}^{u_i}\sum_{s=r}^{m}\chi_{r,s} - \sum_{s=1}^{l_i-1}\sum_{r=1}^{s}\chi_{r,s} \end{aligned}$$

(We adhere to the usual convention that a summation over an empty index set is 0.) Since d_i is always smaller than n, the lower bound on $\tilde{p}_i$ can be improved in the following theorem.

Theorem 4.7

$$\tilde{p}_i \geqslant \frac{\chi_{i,i}d_i}{n(d_i - n_i + \chi_{i,i})}. \qquad (15)$$

Proof. The proof is similar to that of Theorem 4.1, except that for every $i = 1, ..., m$ and all $j \in \mathcal{J}_i$, $\tilde{\eta}_j$ now satisfies that,

$$\begin{aligned} \tilde{\eta}_j &\leqslant \sum_{c=l_i}^{u_i}\tilde{p}_c \\ &\leqslant \frac{d_i}{n}. \quad \text{[from (13)]} \end{aligned}$$

Therefore,

$$\begin{aligned} n &= \sum_{j\in\mathcal{J}_i}\frac{1}{\eta_j} \\ &= \frac{\chi_{i,i}}{\tilde{p}_i} + \sum_{j\in\mathcal{J}_i\setminus\mathcal{S}_i}\frac{1}{\tilde{\eta}_j} \\ &\geqslant \frac{\chi_{i,i}}{\tilde{p}_i} + \frac{n_i - \chi_{i,i}}{\frac{d_i}{n}}, \end{aligned}$$

and (15) follows. □

Corollary 4.8

$$\tilde{p}_1 \geqslant \frac{\chi_{1,1}(n-\chi_{m,m})}{n(n-\chi_{m,m}-n_1+\chi_{1,1})}$$
$$\tilde{p}_m \geqslant \frac{\chi_{m,m}(n-\chi_{1,1})}{n(n-\chi_{m,m}-n_i+\chi_{m,m})}$$

and for $i = 2, ..., m-1$,

$$\tilde{p}_i \geqslant \frac{\chi_{i,i}(n-\min(\chi_{1,1},\chi_{m,m}))}{n(n-\min(\chi_{1,1},\chi_{m,m})-n_i+\chi_{i,i})}.$$

Proof. Proof is obtained from the facts that

$$d_1 \leqslant n-\chi_{m,m}, \quad d_m \leqslant n-\chi_{1,1},$$

and for every $i = 2, ..., m-1$,

$$d_i \leqslant n-\min(\chi_{1,1},\chi_{m,m}).$$

□

Example 4.9 *Consider a univariate data set* $\{R_1, ..., R_{12}\} = \{1, 2, [3, 5], [4, 7], [6, 10], [8, 12], 9, [11, \infty), 13, [14, \infty), 15, [16, \infty)\}$ *which are represented in Figure 5. (The vertical positions hold no special meaning.) The RRs of the data are represented at the lowest vertical position and labeled* $H_1, ..., H_9$. *The*

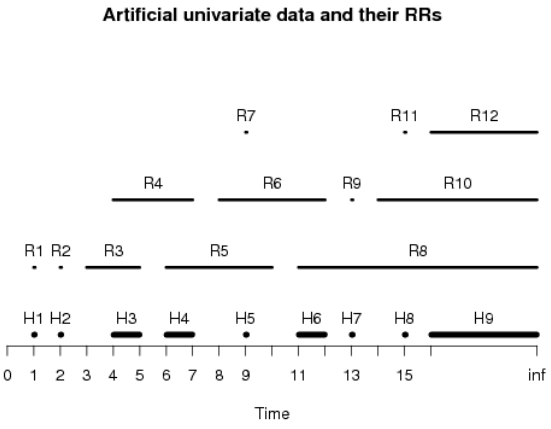


Figure 5: An artificial univariate data set

following is the clique matrix of this univariate censored data set:

$$\mathbf{A} = \begin{bmatrix} 1&0&0&0&0&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0&0&0&0&0 \\ 0&0&1&1&0&0&0&0&0&0&0&0 \\ 0&0&0&1&1&0&0&0&0&0&0&0 \\ 0&0&0&0&1&1&1&0&0&0&0&0 \\ 0&0&0&0&0&1&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&1&1&0&0&0 \\ 0&0&0&0&0&0&0&1&0&1&1&0 \\ 0&0&0&0&0&0&1&1&0&1&0&1 \end{bmatrix}.$$

The equivalent characteristic matrix is

$$\boldsymbol{\chi} = \begin{bmatrix} 1&0&0&0&0&0&0&0&0 \\ &1&0&0&0&0&0&0&0 \\ &&1&1&0&0&0&0&0 \\ &&&0&1&0&0&0&0 \\ &&&&1&1&0&0&0 \\ &&&&&0&0&0&1 \\ &&&&&&1&0&1 \\ &&&&&&&1&0 \\ &&&&&&&&1 \end{bmatrix}.$$

The (unique) NPMLE probability vector is $\hat{\mathbf{p}} = [0.083, 0.083, 0.167, 0, 0.25, 0, 0.104, 0.156, 0.156]'$. *The CDF NPMLE is displayed in Figure 6.*

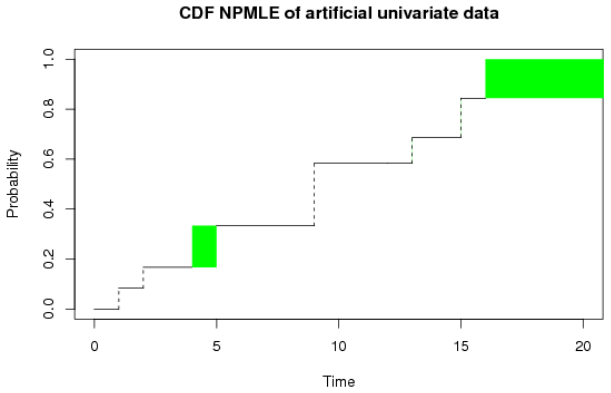


Figure 6: Example CDF NPMLE. Shaded boxes indicate areas of R-nonuniqueness, i.e. nonuniqueness related to arbitrariness of mass placement.

Applying Theorem 4.4, we can compare the NPMLE and the SCE lower and upper bounds as shown in Table 6.

$\mathcal{C}$	Lower bound	NPMLE	Upper bound
$\tilde{p}_1$	0.083	0.083	0.083
$\tilde{p}_1+\tilde{p}_2$	0.167	0.167	0.167
$\tilde{p}_1+\tilde{p}_2+\tilde{p}_3$	0.250	0.333	0.333
$\tilde{p}_1+\cdots+\tilde{p}_4$	0.333	0.333	0.417
$\tilde{p}_1+\cdots+\tilde{p}_5$	0.500	0.583	0.583
$\tilde{p}_1+\cdots+\tilde{p}_6$	0.583	0.583	0.667
$\tilde{p}_1+\cdots+\tilde{p}_7$	0.667	0.688	0.750
$\tilde{p}_1+\cdots+\tilde{p}_8$	0.750	0.844	0.917
$\tilde{p}_1+\cdots+\tilde{p}_9$	1.000	1.000	1.000

Table 6: NPMLE, lower and upper bounds comparison for a univariate data set

5 Application to M-Nonuniqueness

For MCD, the NPMLEs may display the aforementioned mixture or *M-nonuniqueness*, which occurs when different probability vectors have the same likelihood, that is, mass may be exchanged between maximal cliques without changing the likelihood. Gentleman & Vandal [9] proved that M-nonuniqueness cannot occur with univariate censored data but that it may arise with multivariate censored data. Liu [14, 15] and Vandal, Gentleman & Liu [24] discuss conditions for uniqueness of the NPMLEs and apply methods from convex optimization theory to show that the set of all NPMLEs is a polytope.

Suppose that the size of the clique matrix $\mathbf{A}$ for a MCD set is $m \times n$. When we have one CDF NPMLE $\hat{\mathbf{p}}$ for $\mathbf{A}$, the NPMLE's polytope $\mathcal{P}$ can be described by the following so called H-representation (Liu [14, 15], Vandal, Gentleman & Liu [24]):

$$\mathcal{P} = \{\mathbf{p} = [p_1, ...p_m]'; \mathbf{A}'\mathbf{p} = \mathbf{A}'\hat{\mathbf{p}}, \mathbf{p} \geqslant \mathbf{0} \text{ and } \mathbf{e}'\mathbf{p} = 1\}.$$

We consider three descriptions of the NPMLE's polytope $\mathcal{P}$ designed to quantify the extent of M-nonuniqueness. The first description is the so-called V-representation of $\mathcal{P}$, that is, the list of all its vertices. The second is the *diameter* of $\mathcal{P}$, corresponding to the longest distance between two of its vertices. The third is the *size* of $\mathcal{P}$, defined as the longest projection on one of the m axes corresponding to the vector entries. The diameter and size of $\mathcal{P}$ have been considered in the study of CDF NPMLE M-nonuniqueness and asymptotic properties. For more detail, see Liu [14, 15] and Vandal, Gentleman & Liu [24].

From (6), upper bounds for the diameter and the size of $\mathcal{P}$ for a censored data set with clique matrix $\mathbf{A}_{m\times n}$ can be obtained respectively as

$$\text{dia}(\mathcal{P}) \leqslant \left(\sum_{i=1}^{m}(U_i - L_i)^2\right)^{1/2}, \quad (16)$$

$$\text{size}(\mathcal{P}) \leqslant \max_{i=1,...,m}(U_i - L_i), \quad (17)$$

where for $i = 1, 2, ..., n$,

$$L_i := \frac{n^-(\{M_i\})}{n - n^+(\{M_i\}) + n^-(\{M_i\})}$$

and

$$U_i := \frac{n^+(\{M_i\})}{n}$$

are lower and upper bounds of $\tilde{p}_i$ given in (6). As an application of (16) and (17), consider a cyclical data set with order $2k$, $k = 2, 3, ...$, circulant clique matrix defined as follows:

$$\begin{bmatrix} 1 & & & & & & 1 \\ 1 & 1 & & & & & \\ & 1 & 1 & & & & \\ & & \ddots & \ddots & & & \\ & & & 1 & 1 & & \\ & & & & 1 & 1 & \end{bmatrix} \quad (18)$$

where all unspecified entries in the matrix are 0.

Then the diameter and size of the corresponding NPMLE polytope are at most $\sqrt{2/k}$ and $\frac{1}{k}$ respectively. [6]

The last theorem provides a sufficient condition for asymptotic mixture uniqueness.

Theorem 5.1 *When*

$$\frac{1}{n}\max_{i=1,...m}(n^+(M_i) - n^-(M_i)) \xrightarrow{\text{a.s}} 0,$$

the M-nonuniqueness of the CDF NPMLE will disappear asymptotically, in the sense that,

$$\text{size}(\mathcal{P}) \xrightarrow{\text{a.s}} 0.$$

Proof. Since for every $i = 1, ..., m$,

$$n^+(\{M_i\}) \geqslant n^-(\{M_i\}),$$

then from (17),

$$\text{size}(\mathcal{P}) \leqslant \max_{i=1,...,m}\Big(\frac{n^+(\{M_i\}) - n^-(\{M_i\})}{n}\Big).$$

The conclusion follows. □

Acknowledgements. The authors wish to thank two anonymous reviewers for their useful comments and suggestions.

References

[1] Billard, L. & Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic data analysis. *J Amer Statist Assoc* **98**, 470–487.

[2] Bogaerts, K. & Lesaffre, E. (2004). A new, fast algorithm to find the regions of possible support for bivariate interval-censored data. *J Comp Graph Statist* **13**, 330–340.

[6] We can in fact check that the diameter and size are $\sqrt{2/k}$ and $\frac{1}{k}$ exactly.

[3] Braun, J., Duchesne, T. & Stafford J.E. (2005). Local likelihood density estimation for interval censored data. *Can J Statist* **33**, 39–60.

[4] Dempster, A.P. (1967). Upper and Lower Probability inferences based on a sample from a finite univariate population. *Biometrika* **54**, 325-339.

[5] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximal likelihood estimation from incomplete data via the EM algorithm (with discussion). *J Roy Statist Soc B* **39**, 1–38.

[6] Efron, B. (1967). The two-sample problem with censored data. *Proc. Fifth Berkeley Symp Math Statist Probab* **4**, 831–853.

[7] Ferson, S., Kreinovich, V., Hajagos, J., Oberkampf, W. & Ginzburg, L. (2007). Experimental uncertainty estimation and statistics for data having interval uncertainty. Sandia National Laboratories Technical Report SAND2007-0939.

[8] Gentleman, R. & Geyer, C.J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika* **81**, 618–623.

[9] Gentleman, R. & Vandal, A.C. (2001). Computational algorithms for censored data using intersection graphs. *J Comp Graph Statist* **10**, 403–421.

[10] Gentleman, R. & Vandal, A.C. (2002). Graph-theoretical aspects of bivariate censored data. *Can J Statist* **30**, 557–571.

[11] Groeneboom, P. & Wellner, J.A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser, Basel.

[12] Heitjian, D. F. & Rubin, D.B. (1991). Ignorability and coarse data. *Ann Statist* **19**, 2244–2253.

[13] Kaplan E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J Amer Statist Assoc* **53**, 457–481.

[14] Liu, X. (2002). *Nonparametric Maximum Likelihood Estimation of the Cumulative Distribution Function with Multivariate Interval Censored Data: Computation, Identifiability and Bounds*. M.Sc. Thesis, Department of Mathematics and Statistics, McGill University, Montréal.

[15] Liu, X. (2005). *Nonparametric Estimation with Censored Data: a iscrete Approach*. Ph.D. Thesis, Department of Mathematics and Statistics, McGill University, Montréal.

[16] Maathuis, M.H. (2005). Reduction algorithm for the NPMLE for the distribution function of bivariate interval censored data. *J Comp Graph Statist* **14**, 252–262.

[17] Manski, C.F. (2003). *Partial Identification of Probability Distribution*. Springer-Verlag:Berlin.

[18] Peto, R. (1973). Experimental survival curves for interval censored data. *Appl Statist* **22**, 86–91.

[19] Pruitt, R.C. (1993). Small sample comparison of six bivariate survival curve estimators. *J Statist Comp Simul* **45**, 147–167.

[20] Rakowski, U.K. (2007). Fundamentals of the Dempster-Shafer Theory and its application to system safety and reliability modelling. *Reliability: Theory & Applications* (special issue) **3-4**, 173–185.

[21] Turnbull, B.W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J Amer Statist Assoc* **69**, 169–173.

[22] Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J Roy Statist Soc B* **38**, 290–295.

[23] Vandal, A.C. (1999). *Order theory and nonparametric analysis for interval censored data*. Ph.D. Thesis, Department of Statistics, University of Auckland.

[24] Vandal, A.C., Gentleman, R. & Liu, X. (2005a). Some comments on the uniqueness of the CDF NPMLE for censored data. Technical report, Department of Mathematics and Statistics, McGill University.

[25] Vandal, A.C., Gentleman, R. & Liu, X. (2005b). Constrained estimation and likelihood intervals for censored data, *Can J Statist* **33**, 71–83.

[26] Wang, Y. (2008). Dimension-reduced nonparametric maximum likelihood computation for interval-censored data, *Comp Statist Data Anal* **52**, 2388–2402.

[27] Wellner, J.A. & Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric likelihood estimator from censored data. *J Amer Statist Assoc* **92**, 945–959.

A Fully Polynomial Time Approximation Scheme for Updating Credal Networks of Bounded Treewidth and Number of Variable States

Denis D. Mauá
IDSIA, Switzerland
denis@idsia.ch

Cassio P. de Campos
IDSIA, Switzerland
cassio@idsia.ch

Marco Zaffalon
IDSIA, Switzerland
zaffalon@idsia.ch

Abstract

Credal networks lift the precise probability assumption of Bayesian networks, enabling a richer representation of uncertainty in the form of closed convex sets of probability measures. The increase in expressiveness comes at the expense of higher computational costs. In this paper we present a new algorithm which is an extension of the well-known variable elimination algorithm for computing posterior inferences in extensively specified credal networks. The algorithm efficiency is empirically shown to outperform a state-of-the-art algorithm. We then provide the first fully polynomial time approximation scheme for inference in credal networks with bounded treewidth and number of states per variable.

Keywords. Probabilistic graphical models, credal networks, approximation scheme, valuation algebra.

1 Introduction

Credal networks [11] are generalizations of Bayesian networks that allow for a richer representation of uncertainty in the form of set-valued probabilities—in contrast to the sharp numeric values required by their Bayesian counterpart. They are models of imprecise probability as advocated by Walley [18]. In a nutshell, credal networks rely on a directed acyclic graph (DAG) to encode a compact and computationally efficient representation of a closed convex set of joint probability mass functions over a set of variables, much in the same way that Bayesian networks do for single joint probability mass functions. Namely, credal networks respect the local Markov condition that each variable (uniquely represented by a node in the DAG) is (strongly) independent of its non-descendant non-parents conditional on its parents. Strong independence is justified by a sensitivity analysis interpretation, where we assume that there exists a single probability mass function representing our knowledge which we cannot know precisely for lack of resources; epistemic irrelevance, on the other hand, is arguably more consistent with a behavioral interpretation of inherent imprecision [18]. In the following, we assume credal networks to operate under strong independence.

In order to enable efficient computation, additional constraints need to be imposed to the set-valued specifications of the local probabilities. The two most common choices are *extensively specified* sets, in which local models are given as sets of probability potentials, and *separately specified* sets, in which local models are specified as collections containing one set of probability mass functions for each configuration of the parents. Separately specified networks can be mapped to extensively specified and vice-versa [2].

There is also another subtlety when computing with such local models, which concerns the way they are represented in a computer. The sets of local (conditional) probability mass functions can be encoded either as sets of points (e.g., the sets of vertices of a convex polytope), or as sets of (linear) inequalities. Although these two encodings can represent any finitely-generated closed convex set, moving from an inequality-based encoding to a vertex-based encoding can dramatically increase the length of the representation of the local models. For example, a simple 8-dimensional polytope specified by 729 inequalities has between 5 thousand and 12 billion vertices [4].

Inference with credal networks has been theoretically and empirically shown to be a difficult problem. For example, computing exact marginals in credal networks is known to be NP-hard even for polytree-shaped networks, a particular case that can be computed in polynomial time in Bayesian networks [7]. Despite the hardness of the problem, several algorithms are known to perform reasonably well under certain conditions. Most notably, the 2U algorithm [12], which computes exact posterior bounds in polytree-shaped credal networks with binary variables, continues to be the only known polynomial time algorithm available, and its generalizations to arbitrary networks (e.g., the GL2U [3]), which perform approximate inference, are among the fastest algorithms. A notable example, against which we compare our results in this paper, is the algorithm of de Campos and Cozman [8], which

algorithm	complexity	topology	inference	representation
2U [12]	polynomial	polytree	exact	inequality
GL2U [3]	polynomial	all	approximate	vertex
A/R+ [16]	exponential	polytree	approximate	inequality
IP [8]	exponential	all	exact/approx.	inequality
ML [6]	exponential	all	exact/approx.	inequality
HC [9]	exponential	all	exact/approx.	vertex

Table 1: Comparison of some existing algorithms for inference in credal networks.

finds exact posterior bounds in general networks by converting the problem into a mixed integer program, which can be solved exactly for small networks, or relaxed to provide approximate results in large networks. Other approaches mix branch-and-bound methods for exact inference and local searches for approximate results [6, 9, 16]. Table 1 contrasts some of the available algorithms. To date, no algorithm is known to provide approximations within given bounds in polynomial time. Recently, de Cooman et al. [10] developed a polynomial time algorithm for tree-shaped credal networks, but it operates under epistemic irrelevance.

In this paper, we present a new algorithm for computing exact posterior bounds in extensively specified credal networks encoded by vertices, as well as a fully polynomial time approximation scheme (FPTAS) for networks with bounded treewidth and number of states per variable. We begin by stating the basic elements of our formalism (Section 2), followed by a formal definition of inference in extensively specified credal networks (Section 3). Then we present a modified variable elimination algorithm for exact inference, which has worst-case complexity exponential in both the treewidth of the graph and the size of local sets (Section 4). We address this issue by devising an FPTAS (Section 5). Experiments showing the performance of the algorithms are presented and discussed in Section 6. Finally, Section 7 contains our concluding thoughts.

Due to the limited space, we only present proofs for the most important results.

2 An Algebra of Ordered Potentials

In this section, we introduce the main ingredients of the message passing algorithms that we present later as well as the basic results needed to guarantee the correctness and efficiency of computations.

From an algebraic viewpoint, the primitive entities of our formalism are the so-called *labeled valuations* (ϕ, x), which encode information about a (local) domain through a *valuation* ϕ and a set of *variables* x. Here we adopt the equivalent notation ϕ_x to denote the pair (ϕ, x). More concretely, valuations can take as straightforward forms as bounded real-valued functions (Section 2.2), or represent more complicated objects such as sets of pairs of probability potentials (Section 2.3).

The set of all variables we consider relevant to a problem, denoted by $\mathcal{U}$, is the largest set of variables that can be considered for a (labeled) valuation in our setting, which we assume to be bounded. We write variables with capital letters (e.g., $X_1, \ldots, X_n \in \mathcal{U}$) and sets of variables in lower case (e.g., $x = \{X_1, \ldots, X_n\}$). Any variable X is assumed to be associated with a finite set of values Ω_X called its *frame*. The elements of Ω_X are called states. If x is a set of variables, the domain Ω_x is given by the Cartesian product of the frames of variables in x, $\Omega_x \triangleq \times_{X \in x} \Omega_X$. Any element of Ω_x is called a configuration. If $\mathbf{x}$ is a configuration in Ω_x, the notation $\mathbf{x}^{\downarrow y}$ denotes the projection of $\mathbf{x}$ onto $y \subseteq x$, with $\mathbf{x}^{\downarrow \emptyset} \triangleq \lambda$, where λ denotes the null element that does not appear in any frame.

The set of all valuations (ϕ, x) over a subset $x \subseteq \mathcal{U}$ is denoted by Φ_x. The set of all valuations is denoted by $\Phi \triangleq \bigcup_{x \subseteq \mathcal{U}} \Phi_x$. The algebra comes with two basic operations of *combination* and *marginalization*. Intuitively, combination represents aggregation of two pieces of information. If ϕ_x and ϕ_y are two arbitrary valuations, then $\phi_x \times \phi_y$ is a valuation $\phi_{x \cup y}$ with domain $\Omega_{x \cup y}$. Marginalization, on the other hand, acts by coarsening information. If ϕ_x is a valuation then the marginal $\phi_x^{\downarrow y}$ is a valuation with domain Ω_y. Sometimes, it is convenient to define the elimination operation, which is in a one-to-one correspondence to marginalization. Formally, if ϕ_x is a valuation then $\phi_x^{-y} \triangleq \phi_x^{\downarrow x \setminus y}$ is the result of the elimination of variables in y. When clear from the context, we write Y to denote a singleton $y = \{Y\}$, for example $\phi_x^{-Y} = \phi_x^{\downarrow x \setminus \{Y\}}$. A system $(\Phi, \mathcal{U}, \times, \downarrow)$ closed under combination and marginalization is said to be a *valuation algebra* if it satisfies the following three axioms [15, 17].

(A1) Combination is commutative and associative.

(A2) For $y \subseteq x \subseteq z$, $\left(\phi_z^{\downarrow x}\right)^{\downarrow y} = \phi_z^{\downarrow y}$.

(A3) If $x \subseteq z \subseteq x \cup y$ then $(\phi_x \times \phi_y)^{\downarrow z} = \phi_x \times \phi_y^{\downarrow z \cap y}$.

The purpose of a valuation algebra is the computation of marginals of the form $(\times_i \phi_{u_i})^{\downarrow y}$, where the joint valu-

ation $\times_i \phi_{u_i}$ is computationally too expensive to be obtained explicitly. The complexity of the operations of combination and marginalization is given by the size of the valuations involved, which is in general a function of the cardinality of the domain. Hence, as a rule-of-thumb, the larger the domain of a valuation the more expensive are the operations involving it. The axioms of valuation algebras provide the necessary framework for breaking down the computation of costly marginals into a sequence of computations of marginals over smaller domains. The pseudo-code in Algorithm 1 exhibits the variable elimination procedure (also known as fusion algorithm), which more efficiently computes marginals of factorized valuations.

Algorithm 1: Variable Elimination

input : A finite set of valuations Ψ, a set of target variables $y \subset \mathcal{U} \triangleq \bigcup_{\phi_u \in \Psi} u$, and an ordering $o = (X_1, \ldots, X_n)$ of the variables in $\mathcal{U} \setminus y$

output: The marginal $(\bigtimes_{\phi \in \Psi} \phi)^{\downarrow y}$

for $i \leftarrow 1$ **to** n **do**
  Set $\mathcal{B}_i \leftarrow \{\phi_u \in \Psi : X_i \in u\}$;
  Compute $\Psi^i \triangleq (\bigtimes_{\phi \in \mathcal{B}_i} \phi)^{-X_i}$;
  Set $\Psi \leftarrow (\Psi \setminus \mathcal{B}_i) \cup \{\Psi^i\}$;
end
return $\Gamma \triangleq \bigtimes_{\phi \in \Psi} \phi$;

Instead of computing a valuation $\bigtimes_{\phi \in \Psi} \phi$ over a large domain $\Omega_{\mathcal{U}}$ and then marginalizing to y, the algorithm computes marginals $(\bigtimes_{\phi \in \mathcal{B}_i} \phi)^{-X_i}$ over possibly much smaller domains. The overall complexity of the algorithm is given by the size of the largest valuation Ψ^i generated at the loop step. If such a size is bounded then (A1)–(A3) are sufficient to show that the algorithm efficiently outputs the desired marginal [15].

Some optimization tasks like the credal network inferences we aim at here admit a partial ordering over the valuations. Let $\leq$ denote a partial order over Φ (i.e., a reflexive, antisymmetric and transitive relation). An *ordered valuation algebra* [13] is a system $(\Phi, \mathcal{U}, \times, \downarrow, \leq)$, where $(\Phi, \mathcal{U}, \times, \downarrow)$ is a valuation algebra and $\leq$ is monotonic with respect to $\times$ and $\downarrow$:

(A4) If $\phi_x \leq \psi_x$ and $\phi_y \leq \psi_y$ then $(\phi_y \times \phi_x) \leq (\psi_y \times \psi_x)$ and $\phi_x^{\downarrow y} \leq \psi_x^{\downarrow y}$.

Given a finite set of ordered valuations $\Psi \subseteq \Phi$, we say that $\phi \in \Psi$ is *maximal* if for all $\psi \in \Psi$ such that $\phi \leq \psi$ it holds that $\psi \leq \phi$. The operation $\max(\Psi)$ returns the set of maximal valuations of a set Ψ. Given any relation R on Ψ, a subset $\Psi' \subseteq \Psi$ is called an *R-covering* of Ψ if for every $\phi \in \Psi$ there is $\psi \in \Psi'$ such that $\phi R \psi$. For example, the set $\max(\Psi)$ is a $\leq$-covering for Ψ.

2.1 Set-Valuations

The algorithms we develop use the more complex entities of sets of valuations, called *set-valuations*. Theses entities can nevertheless be casted in the algebra of valuations, and manipulated by the variable elimination algorithm to produce sets of marginal valuations.

Let 2^{Φ_x} denote the power set of Φ_x, that is, the set of all subsets of it. Thus, 2^{Φ} denotes the set of all subsets of valuations in Φ. If $\Psi_x \in 2^{\Phi_x}$ and $\Psi_y \in 2^{\Phi_y}$, we define their set-combination $\otimes$ as the set-valuation resulting from element-wise combination of their elements, $\Psi_x \otimes \Psi_y \triangleq \{\phi_x \times \phi_y : \phi_x \in \Psi_x, \phi_y \in \Phi_y\}$. Likewise, we define the set-marginalization operation $\Downarrow$ on 2^{Φ} as the element-wise marginalization of the valuations in a set, $\Psi_x^{\Downarrow y} \triangleq \{\phi_x^{\downarrow y} : \phi_x \in \Psi_x\}$.

Proposition 1. *The system* $(2^{\Phi}, \mathcal{U}, \otimes, \Downarrow)$ *of set-valuations with set-combination and set-marginalization is a valuation algebra.*

The exact variable elimination algorithm we develop in Section 4 obtains its (relative) efficiency by propagating only maximal valuations. Let $\max(2^{\Phi}) \triangleq \{\max(\Psi) : \Psi \in 2^{\Phi}\}$ denote the set of all sets of maximal valuations in 2^{Φ}. We define the max-combination $\oplus$ and max-marginalization $\Downarrow\Downarrow$ as $\Psi_x \oplus \Psi_y \triangleq \max(\Psi_x \otimes \Psi_y)$ and $\Psi_x^{\Downarrow\Downarrow y} \triangleq \max(\Phi_x^{\Downarrow y})$.

Proposition 2. *The system* $(\max(2^{\Phi}), \mathcal{U}, \oplus, \Downarrow\Downarrow)$ *of maximal set valuations with max-combination and max-marginalization is also a valuation algebra.*

If $(\Phi_1, \mathcal{U}, \times_1, \downarrow_1)$ and $(\Phi_2, \mathcal{U}, \times_2, \downarrow_2)$ are two valuation algebras, we say that a mapping $h : \Phi_1 \to \Phi_2$ is a *homomorphism* if for any $\phi_x, \phi_y \in \Phi_1$ we have that $h(\phi_x) \times_2 h(\phi_y) = h(\phi_x \times_1 \phi_y)$ and $h(\phi_x)^{\downarrow_2 y} = h(\phi_x^{\downarrow_1 y})$. Thus, if we are interested in computing $h(\phi_1^{\downarrow_1 y})$ for some valuation $\phi_1 \in \Phi_1$ that we know that factorizes as $\phi_1 = \psi_1 \times_1 \cdots \times_1 \psi_m$, we can equivalently obtain $(h(\psi_1) \times_2 \cdots \times_2 h(\psi_m))^{\downarrow_2 y}$, which might be computationally more convenient. The following result relates the algebras of set-valuations and maximal set-valuations.

Proposition 3. $\max$ *is a homomorphism from* $(2^{\Phi}, \mathcal{U}, \otimes, \Downarrow)$ *to* $(\max(2^{\Phi}), \mathcal{U}, \oplus, \Downarrow\Downarrow)$.

Since the set of maximal elements of a set is in the worst case as large as the set itself, but often much smaller, the homomorphism $\max$ allows us to conveniently obtain a set of maximal marginals $\max([\bigotimes_i \Psi_{x_i}]^{\Downarrow y})$ by computing the equivalent $[\bigoplus_i \max(\Psi_{x_i})]^{\Downarrow\Downarrow y}$. Recall that $\otimes$ is defined as element-wise combination of valuations in the cartesian product, and assume that the set-valuations Ψ_{x_i} can not be factorized as combinations of other set-valuations. Hence, the set $\bigotimes_i \Psi_{x_i}$ is exponentially large in the size of each Ψ_{x_i} and often intractable. On the other hand, the combination of maximal set-valuations $\bigoplus_i \max(\Psi_{x_i})$ can mitigate the exponential explosion if the number of

maximal points is kept bounded after each pairwise combination. For instance, if each of the local maximal sets $\max(\Psi_{x_i})$ is half as large as its original set Ψ_{x_i}, then computing $\max([\bigotimes_i \max(\Psi_{x_i})]^{\Downarrow y})$ involves $O(2^n)$ less computations than $\max([\bigotimes_i \Psi_{x_i}]^{\Downarrow y})$. The speed up strongly depends on the number of non-maximal elements that are discarded after each max-combination.

In the rest of this section we introduce the concrete valuation algebras our framework relies on.

2.2 Probability Potentials

Probability potentials are perhaps the most common example of valuation algebras. They generalize (conditional) probability mass functions. If $x \subseteq \mathcal{U}$ is a nonempty set of variables, we define a *potential* p_x as a mapping from Ω_x to the set of nonnegative reals. A potential $p_\emptyset$ over the empty set is defined as a nonnegative real number. The size of a potential p_x is the cardinality of its domain. The following operations are defined over potentials. Combination of potentials is done by element-wise multiplication: for $\mathbf{z} \in \Omega_{x \cup y}$,

$$(p_x \times p_y)(\mathbf{z}) \triangleq p_x(\mathbf{z}^{\downarrow x}) p_y(\mathbf{z}^{\downarrow y})\,. \tag{1}$$

Marginalization is defined as the sum of compatible elements. For $\mathbf{y} \in \Omega_y$,

$$p_x^{\downarrow y}(\mathbf{y}) \triangleq \sum_{\mathbf{x} \in \Omega_x : \mathbf{x}^{\downarrow y} = \mathbf{y}} p_x(\mathbf{x})\,. \tag{2}$$

Note that if $y = \emptyset$, the marginal $p_x^{\downarrow y}$ is a (nonnegative real) number.

Partial ordering is given by weak Pareto dominance. Given two potentials p_x and q_x over Ω_x, we define $p_x \geq q_x$ if $p_x(\mathbf{x}) \geq q_x(\mathbf{x})$ for all $\mathbf{x} \in \Omega_x$. Note that if p_x and q_x have equal sum (i.e., $\sum_{\mathbf{x} \in \Omega_x} p_x(\mathbf{x}) = \sum_{\mathbf{x} \in \Omega_x} q_x(\mathbf{x})$) then $p_x \not\geq q_x$ and $q_x \not\geq p_x$ (unless $p_x = q_x$). This is the case, for example, of potentials representing (conditional) probability mass functions. Therefore, the identity $\mathcal{P}_x = \max(\mathcal{P}_x)$ holds for any set $\mathcal{P}_x$ of (conditional) probability mass functions. Let $\mathcal{P}$ denote the set of all probability potentials.

Proposition 4. *The system $(\mathcal{P}, \mathcal{U}, \times, \downarrow, \leq)$ is an ordered valuation algebra.*

Given a real number $\alpha > 1$, we define an equivalence relation $\equiv_\alpha$ over potentials such that any two potentials p_x and q_x are α-equivalent (i.e., $p_x \equiv_\alpha q_x$) if for all $\mathbf{x} \in \Omega_x$ either $p_x(\mathbf{x}) = q_x(\mathbf{x}) = 0$ or $p_x(\mathbf{x})$ and $q_x(\mathbf{x})$ are both positive and $\lfloor \log_\alpha p_x(\mathbf{x}) \rfloor = \lfloor \log_\alpha q_x(\mathbf{x}) \rfloor$.

2.3 Pairs of Potentials

The algorithms we develop in Sections 4 and 5 rely on a more abstract structure over pairs of potentials. Let $\phi_x = (p_x^\ell, p_x^r)$ denote a pair of probability potentials over x. The potentials p_x^ℓ and p_x^r are referred to as the left and right potentials of ϕ_x, respectively. For any two pairs of potentials ϕ_x and ψ_x, we define $\phi_x = (p_x^\ell, p_x^r) \geq (q_x^\ell, q_x^r) = \psi_x$ if $p_x^\ell \leq q_x^\ell$ and $p_x^r \geq q_x^r$. The partial order defined in this way reflects the nature of computations with credal networks. We seek for a solution that partly dominates (according to right potentials) all other potentials and partly is dominated by them (according to left potentials). It is in part this dichotomy in the objective that makes posterior inferences in credal networks much harder than their Bayesian counterpart.

If $\phi_x = (p_x^\ell, p_x^r)$ and $\phi_y = (p_y^\ell, p_y^r)$ are two pairs of potentials, we define their combination as the pair of left and right combinations of potentials, that is, $\phi_x \times \phi_y \triangleq (p_x^\ell \times p_y^\ell, p_x^r \times p_y^r)$. Similarly, the marginalization of a pair $\phi_x = (p_x^\ell, p_x^r)$ is performed on both potentials, $\phi_x^{\downarrow y} \triangleq ((p_x^\ell)^{\downarrow y}, (p_x^r)^{\downarrow y})$. Let Φ be the set of all pairs of potentials.

Proposition 5. *The system $(\Phi, \mathcal{U}, \times, \downarrow, \leq)$ is an ordered valuation algebra.*

Let 2^Φ and $\max(2^\Phi)$ denote, respectively, the set of all sets of pairs of potentials and the set of all sets of maximal pairs of potentials. It follows from Propositions 1 and 2 that the systems $(2^\Phi, \mathcal{U}, \otimes, \Downarrow)$ and $(\max(2^\Phi), \mathcal{U}, \oplus, \downdownarrows)$ are valuation algebras. Moreover, $\max$ is a homomorphism from 2^Φ to $\max(2^\Phi)$. Thus, given a collection of finite sets of pairs $\Psi_{x_1}, \ldots, \Psi_{x_n}$, we can obtain the set $\max(\Psi_y) \triangleq \max((\bigotimes \Psi_{x_i})^{\Downarrow y})$ of maximal marginal valuations potentially more efficiently by performing computations in the algebra of sets of maximal pairs, that is, by computing $\max((\bigoplus_i \max(\Psi_{x_i}))^{\downdownarrows y})$. Bentley et al. [5] showed that sets with n uniformly distributed pairs of potentials over a domain Ω_y have, on average, $O((\log n)^{2|\Omega_y|-1})$ maximal elements. Unfortunately, the uniformity assumption does not hold in the computations we perform, and we expect the average number of maximal elements to be higher than this. To our knowledge, it remains to be obtained any bounds or expectations on the size of maximal sets obtained from propagated valuations such as those generated by variable elimination. Note that, as with sets of probability potentials, if Ψ contains only valuations whose left or right potentials specify a probability mass function, then $\Psi = \max(\Psi)$.

We can have an upper bound on the cardinality of sets by relaxing the partial order to allow approximate Pareto dominance. Given a real number $\alpha > 1$, we define a relation $\leq_\alpha$ such that $\phi \leq_\alpha \psi$ denotes that by mistakenly assuming $\phi \leq \psi$ we introduce an error no greater than α in each coordinate. More formally, we define $\phi \leq_\alpha \psi$ if $(\alpha^{-1}, \alpha) \times \psi \geq \phi$. Note that $\leq_\alpha$ is neither transitive nor antisymmetric, and that we may have $\phi \leq_\alpha \psi$ for $\phi \not\leq \psi$.

The α-equivalence relation over potentials can easily be

extended to pairs. Two pairs (p_x^ℓ, p_x^r) and (p_y^ℓ, p_y^r) are α-equivalent if $p_x^\ell \equiv_\alpha p_y^\ell$ and $p_y^r \equiv_\alpha p_y^r$. It is not difficult to see that $\phi \equiv_\alpha \psi$ implies both $\phi \leq_\alpha \psi$ and $\psi \leq_\alpha \phi$.

A $\leq_\alpha$-covering for a set of pairs of potentials Ψ_x provides an approximated version of Ψ_x, one in which for each $\phi_x \in \Psi_x$ we are guaranteed to have a pair ψ_x in the covering such that the left and right potentials of ψ_x and ϕ_x differ in each coordinate by a factor no greater than α. We can easily obtain a $\leq_\alpha$-covering of Ψ_x of bounded cardinality from its quotient set Ψ_x/α, that is, by discarding one of any two α-equivalent pairs in Ψ_x. The approximation algorithm we develop in Section 5 strongly relies on the following results.

Lemma 6. *If $k_1, \ldots, k_m$ are positive integers and $\Psi_{x_1}, \Psi'_{x_1}, \ldots, \Psi_{x_m}, \Psi'_{x_m}$ are set valuations such that for $i = 1, \ldots, m$ Ψ'_{x_i} is a $\leq_{\alpha^{k_i}}$-covering for Ψ_i, then $\Psi'_{x_1} \otimes \cdots \otimes \Psi'_{x_m}$ is a $\leq_\beta$-covering for $\Psi_{x_1} \otimes \cdots \otimes \Psi_{x_m}$, where $\beta = \alpha^{\sum_{i=1}^m k_i}$.*

Proof. We work by induction on $j = 1, \ldots, m$. For $j = 1$, it follows directly that Ψ'_1 is a $\leq_{\alpha^{k_1}}$-covering for Ψ_1. Assume the result holds for $1 \leq j < m-1$, and consider any pair $\phi = \phi' \times \phi''$ in $\Psi_{x_1} \otimes \cdots \otimes \Psi_{x_{j+1}}$, where $\phi' \in \Psi_{x_1} \otimes \cdots \otimes \Psi_{x_j}$ and $\phi'' \in \Psi_{x_{j+1}}$. There is $\psi = \psi' \times \psi''$ in $\Psi'_{x_1} \otimes \cdots \otimes \Psi'_{x_{j+1}}$, where $\psi' \in \Psi'_{x_1} \otimes \cdots \otimes \Psi'_{x_j}$ and $\psi'' \in \Psi_{x_{j+1}}$, such that $(\alpha^{-\sum_{i=1}^j k_i}, \alpha^{\sum_{i=1}^j k_i}) \times \psi' \geq \phi'$ (by assumption) and $(\alpha^{-k_{j+1}}, \alpha^{k_{j+1}}) \times \psi'' \geq \phi''$. It follows from (A4) that $(\alpha^{-\sum_{i=1}^{j+1} k_i}, \alpha^{\sum_{i=1}^{j+1} k_i}) \times \psi \geq \phi$.□

Let $\Psi_{x_1}, \ldots, \Psi_{x_m}$ denote sets of pairs of potentials which take values on the interval $[0, 1]$, and let b be the number of bits required to encode these sets.

Proposition 7. *The number of elements in $(\Psi_{x_1} \otimes \cdots \otimes \Psi_{x_m})^{\Downarrow y}/\alpha$ is $O((bm\alpha/(\alpha-1))^{2|\Omega_y|})$.*

The latter result is in fact an adaptation of Papadimitriou and Yannakakis' result on the boundedness of ϵ-approximate Pareto curves in multi-objective optimization problems [1, Theorem 1].

3 Credal Networks

In this section we review the basic concepts and computational challenges of extensively specified credal networks. Let $\mathcal{G} = (\mathcal{U}, \mathcal{E})$ be a DAG, and X a node in $\mathcal{U}$. We write $\mathrm{pa}(X) \triangleq \{Y \in \mathcal{U} : (Y, X) \in \mathcal{E}\}$ to denote the parents of X, $\mathrm{ch}(X) \triangleq \{Y \in \mathcal{U} : (X, Y) \in \mathcal{E}\}$ to denote the children of X in $\mathcal{U}$, and $\mathrm{fa}(X) \triangleq \{X\} \cup \mathrm{pa}(X)$ to denote the family of X. We call Y a descendant of X if there is a directed path from X to Y in $\mathcal{G}$.

An *extensive credal set* K_x is a set of probability potentials p_x over domain Ω_x. Given an extensive credal set K_x, we write $\mathrm{H}(K_x)$ to denote its convex hull (i.e., the set obtained by all convex combinations of elements in K_x), and $\mathrm{ext}[\mathrm{H}(K_x)]$ to denote its extreme points (i.e., the elements of $\mathrm{H}(K_x)$ that cannot be written as a convex combination of other elements). The convex hull of a set and the set of its extreme points are themselves extensive credal sets.

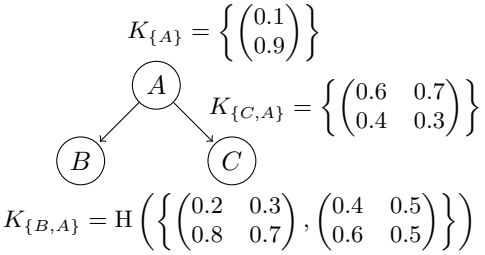


Figure 1: Example of extensively specified credal network.

An *extensively specified credal network* is a pair $(\mathcal{G}, \mathbb{K})$, where $\mathbb{K}$ is a collection of finitely-generated closed convex extensive credal sets $K_{\mathrm{fa}(X)}$, one for each $X \in \mathcal{U}$, such that each potential $p_{\mathrm{fa}(X)} \in K_{\mathrm{fa}(X)}$ satisfies $\sum_{\mathbf{x}^{\downarrow \mathrm{pa}(X)} = \pi} p_{\mathrm{fa}(X)}(\mathbf{x}) = 1$ for all $\pi \in \Omega_{\mathrm{pa}(X)}$ (i.e., they represent conditional probability mass functions $p(X|\,\mathrm{pa}(X))$). Figure 1 depicts a simple extensively specified credal network over 3 binary-valued variables.

The *strong extension* of a credal network is given by the credal set generated by the convex closure of the product of all extensive credal sets in $\mathbb{K}$,

$$K_{\mathcal{U}}^{\text{strong}} \triangleq \mathrm{H}\left(\bigotimes_{X \in \mathcal{U}} K_{\mathrm{fa}(X)} \right). \tag{3}$$

Since the product of local extremes $K_{\mathcal{U}}^{\text{ext}} \triangleq \bigotimes_{X \in \mathcal{U}} \mathrm{ext}[K_{\mathrm{fa}(X)}]$ is a subset of the strong extension (by definition), we have that $\mathrm{ext}[K_{\mathcal{U}}^{\text{strong}}] = \mathrm{ext}[\mathrm{H}(K_{\mathcal{U}}^{\text{ext}})] \subseteq K_{\mathcal{U}}^{\text{ext}}$. Notice that $K_{\mathcal{U}}^{\text{ext}}$ contains a finite number of elements.

Let $q, e \subset \mathcal{U}$ denote disjoint sets of query and evidence variables, respectively, and $(\mathbf{q}, \mathbf{e})$ an element of $\Omega_{q \cup e}$. Inference with credal networks consists in computing lower and upper posterior probabilities (we assume $p^{\downarrow e}(\mathbf{e}) > 0$ for all $p \in K_{\mathcal{U}}^{\text{strong}}$):

$$\underline{p}(\mathbf{q}|\mathbf{e}) \triangleq \min_{p \in K_{\mathcal{U}}^{\text{strong}}} \frac{p^{\downarrow q \cup e}(\mathbf{q}, \mathbf{e})}{p^{\downarrow e}(\mathbf{e})}, \tag{4}$$

$$\overline{p}(\mathbf{q}|\mathbf{e}) \triangleq \max_{p \in K_{\mathcal{U}}^{\text{strong}}} \frac{p^{\downarrow q \cup e}(\mathbf{q}, \mathbf{e})}{p^{\downarrow e}(\mathbf{e})}. \tag{5}$$

Our goal in the rest of this section is to show that the continuous optimizations of Equations (4) and (5) can be mapped into problems of computing maximal sets of marginals of the combinations of finite sets of pairs of potentials. We begin with a well-known result that the solutions to the convex optimizations in Equation (5) are attained at extreme points of the strong extension [18]. Since

any non-extreme point of $K^{\text{ext}}_{\mathcal{U}}$ is also a non-extreme point of the strong extension, we have that

$$\overline{p}(\mathbf{q}|\mathbf{e}) = \max_{p \in K^{\text{ext}}_{\mathcal{U}}} \frac{p^{\downarrow q\cup e}(\mathbf{q},\mathbf{e})}{p^{\downarrow e}(\mathbf{e})} \tag{6}$$

$$= \max_{p_{\mathcal{U}} \in K^{\text{ext}}_{\mathcal{U}}} \frac{p^{\downarrow q\cup e}(\mathbf{q},\mathbf{e})}{p^{\downarrow q\cup e}(\mathbf{q},\mathbf{e}) + p^{\downarrow q\cup e}(\neg\mathbf{q},\mathbf{e})}\,, \tag{7}$$

where $p^{\downarrow q\cup e}(\neg\mathbf{q},\mathbf{e}) \triangleq \sum_{\mathbf{q}'\in\Omega_q:\mathbf{q}'\neq\mathbf{q}} p^{\downarrow q\cup e}(\mathbf{q}',\mathbf{e})$. We can derive analogous equations for the lower bound. The passage from Equation (6) to (7) follows from the definition of marginalization. Notice that Equation (7) states a combinatorial problem over products of local extreme points. If $\underline{p}(\mathbf{q}|\mathbf{e}) > 0$, we can divide the numerator and the denominator of Equation (7) by $p^{\downarrow q\cup e}(\mathbf{q},\mathbf{e}) > 0$ and obtain

$$\overline{p}(\mathbf{q}|\mathbf{e}) = \max_{p \in K^{\text{ext}}_{\mathcal{U}}} \left(1 + \frac{p^{\downarrow q\cup e}(\neg\mathbf{q},\mathbf{e})}{p^{\downarrow q\cup e}(\mathbf{q},\mathbf{e})}\right)^{-1}. \tag{8}$$

For any potential $p \in K^{\text{ext}}_{\mathcal{U}}$, let $p_{\mathbf{q}|\mathbf{e}}$ denote the posterior probability obtained by p, that is, $p_{\mathbf{q}|\mathbf{e}} \triangleq [1 + p^{\downarrow q\cup e}(\neg\mathbf{q},\mathbf{e})/p^{\downarrow q\cup e}(\mathbf{q},\mathbf{e})]^{-1}$. Now consider two potentials p and r such that $p^{\downarrow q\cup e}(\neg\mathbf{q},\mathbf{e}) \leq r^{\downarrow q\cup e}(\neg\mathbf{q},\mathbf{e})$ and $p^{\downarrow q\cup e}(\mathbf{q},\mathbf{e}) \geq r^{\downarrow q\cup e}(\mathbf{q},\mathbf{e})$. Clearly, $p_{\mathbf{q}|\mathbf{e}} \geq r_{\mathbf{q}|\mathbf{e}}$, and r is not a solution of the maximization problem (conversely, p is not a solution of the minimization problem). This allows us to define a partial ordering among solutions $p \in K^{\text{ext}}_{\mathcal{U}}$.

Let $\Phi_{\mathbf{q}|\mathbf{e}}$ denote the set of pairs of potentials $(p^{\downarrow q\cup e}(\neg\mathbf{q},\mathbf{e}), p^{\downarrow q\cup e}(\mathbf{q},\mathbf{e}))$, where $p \in K^{\text{ext}}_{\mathcal{U}}$. Then Equation (8) can be rewritten as

$$\overline{p}(\mathbf{q}|\mathbf{e}) = \max_{(p^\ell,p^r)\in\max(\Phi_{\mathbf{q}|\mathbf{e}})} \left(1 + p^\ell/p^r\right)^{-1}. \tag{9}$$

Basically, what Equation (9) states is that we can narrow down the optimization space to the set of potentials whose corresponding pairs in $\Phi_{\mathbf{q}|\mathbf{e}}$ are not smaller than any other pair in the set (conversely, we take the set of minimal elements in the minimization case). Although this set could be as large as $K^{\text{ext}}_{\mathcal{U}}$, our experiments show that most often it is significantly smaller. Thus, if $\max(\Phi_{\mathbf{q}|\mathbf{e}})$ is sufficiently small, we can find the solution by a simple enumerative scheme, and the optimization problem is then converted into the problem of computing the maximal elements of $\Phi_{\mathbf{q}|\mathbf{e}}$, which can be done by the variable elimination procedure in Algorithm 1, as the following section shows.

4 Exact Inference

In this section we describe an algorithm for exact computation of upper posterior probabilities in credal networks. An algorithm for obtaining lower probabilities can be obtained in a very similar way.

For any variable X and a subset $\mathcal{X} \subset \Omega_X$, we define the identity potential $I_{\mathcal{X}}$ as a potential over X that returns 1 for $\mathbf{x} \in \mathcal{X}$ and 0 otherwise. If $\mathcal{X} = \{\mathbf{x}\}$ is a singleton, we write $I_{\mathbf{x}}$. For any $\mathbf{x} \in \Omega_X$, we define the set $\neg\mathbf{x} \triangleq \Omega_X \setminus \{\mathbf{x}\}$.

Consider a credal network $(\mathcal{G}, \mathbb{K})$, an elimination ordering $o = (X_1, \ldots, X_n)$ of the variables in $\mathcal{U}$, sets of query and evidence variables q and e, and a query-evidence pair $(\mathbf{q},\mathbf{e}) \in \Omega_{q\cup e}$. The variable elimination algorithm (Algorithm 1) can be used to compute exact upper posterior probabilities using the valuation algebra of sets of maximal pairs of potentials in the following way. Let Ψ be the set that contains (i) for each $X \in \mathcal{U}$ a set-valuation $\Psi_X \triangleq \{(p_{\text{fa}(X)}, p_{\text{fa}(X)}) : p_{\text{fa}(X)} \in \text{ext}[K_{\text{fa}(X)}]\}$ in Φ; (ii) a set-valuation $\Psi_q \triangleq \{(I_{\neg\mathbf{q}}, I_{\mathbf{q}})\}$ in Ψ; and (iii) for each $E \in e$ a set-valuation $\Psi_E \triangleq \{(I_{\mathbf{e}^{\downarrow E}}, I_{\mathbf{e}^{\downarrow E}})\}$ in Ψ. Let Γ be the output of the variable elimination algorithm with max-combination, and max-marginalization and inputs Ψ, $y = \emptyset$ and ordering o, and let $p_{\mathbf{q}|\mathbf{e}} \triangleq \max_{(p^\ell,p^r)\in\Gamma} \left(1 + p^\ell/p^r\right)^{-1}$. Finally, let $\overline{p}(\mathbf{q}|\mathbf{e})$ be the solution of the maximization problem in Equation (5). The following result states the correctness of the upper posterior probability obtained the procedure.

Theorem 8. $p_{\mathbf{q}|\mathbf{e}} = \overline{p}(\mathbf{q}|\mathbf{e})$.

Proof. The sets $\Psi_X, \Psi_q, \Psi_E \in \Psi$ as well as the sets Ψ^i generated by the variable elimination algorithm are valuations in the valuation algebra of sets of maximal pairs of potentials. It follows from (A1)–(A3) that

$$\Gamma = \left(\Psi_q \bigoplus_{E\in e} \Psi_E \bigoplus_{X\in\mathcal{G}} \Psi_X\right)^{\downarrow\downarrow\emptyset} \tag{10}$$

$$= \max\left(\left[\Psi_q \bigotimes_{E\in e} \Psi_E \bigotimes_{X\in\mathcal{G}} \Psi_X\right]^{\downarrow\emptyset}\right), \tag{11}$$

where the last equivalence is obtained by repeatedly applying Proposition 3. Recall that combination of pairs is defined as the pair formed by the combination of left potentials and the combination of right potentials. Therefore, Γ is a set of maximal pairs of potentials (p^ℓ, p^r), where by definition of Ψ_q, Ψ_E, and Ψ_X,

$$p^\ell = \left(I_{\neg\mathbf{q}} \bigotimes_{E\in e} I_{\mathbf{e}^{\downarrow E}} \bigotimes_{X\in\mathcal{G}} p_{\text{fa}(X)}\right)^{\downarrow\emptyset} \tag{12}$$

$$= p^{\downarrow q\cup e}_{\mathcal{U}}(\neg\mathbf{q},\mathbf{e})\,, \tag{13}$$

$$p^r = \left(I_{\mathbf{q}} \bigotimes_{E\in e} I_{\mathbf{e}^{\downarrow E}} \bigotimes_{X\in\mathcal{G}} p_{\text{fa}(X)}\right)^{\downarrow\emptyset} \tag{14}$$

$$= p^{\downarrow q\cup e}_{\mathcal{U}}(\mathbf{q},\mathbf{e})\,. \tag{15}$$

Moreover, p^ℓ and p^r are compatible, that is, for any potential $p_{\text{fa}(X)}$ in p^ℓ taken from a local extensive credal set $K_{\text{fa}(X)}$, the same potential appears in p^r and no other potential from $K_{\text{fa}(X)}$. Hence, $\Gamma = \max(\Phi_{\mathbf{q}|\mathbf{e}})$. The result

is obtained by comparing the definition of $p_{\mathbf{q}|\mathbf{e}}$ and Equation (9). □

The complexity of the algorithm is upper bounded by the cost of the combination of sets of pairs in computing Ψ^i during the variable elimination part. Each of these computations takes time polynomial in the size of the largest set, which might be exponential in the size of the input sets. For instance, the size of the largest potential is a function of the topology of $\mathcal{G}$ and the given elimination ordering o. The number of elements of a set, on the other hand, depends on the number of non-maximal elements that are discarded at each combination or marginalization operation. In the worst-case scenario where no element is ever discarded, the algorithm runs in exponential time even if the network treewidth and the cardinality of the frames of the input sets are bounded (which is not surprising given that the problem is NP-hard under such assumptions).

An algorithm for lower posterior probabilities can be obtained by substituting sets of maximal valuations and maximizations by sets of minimal valuations and minimizations, respectively. The correctness and complexity analyses are analogous to the maximization case.

5 FPTAS

The computational bottleneck of the variable elimination procedure presented in Section 4 is the existence of large sets at some point in the propagation step (apart from the inherent difficulty of manipulating potentials over large domains). We can remedy the large set problem by trading off accuracy and running time. In this section, we devise a multiplicative approximation scheme that runs in time polynomial in the number of potentials of the input extensive credal sets, but it is still exponential in the size of the largest pair ψ^{X_i} generated during the propagation step, which depends only on the sizes of the frames of the variables and the network treewidth. For the rest of this section, we assume the size of variable frames and the network treewidth to be bounded by a constant. Additionally, we require the input potentials to be represented by rational numbers, so that the length of the input is well-defined. The approximation scheme we obtain is an FPTAS, that is, a family of algorithms parameterized by $\epsilon > 0$ that returns in time polynomial to $1/\epsilon$ and to the input size a feasible solution that is no worse than the optimal solution by a factor of ϵ. If x^* is the optimal solution (of a maximization problem), the approximation algorithm returns a solution x such that $x^*/(1+\epsilon) \leq x \leq x^*$.

Given a real number α greater than one, we define the α-combination of two set-valuations Ψ_x and Ψ_y as the quotient set of the their set-combination, that is, $\Psi_x \boxtimes_\alpha \Psi_y \triangleq (\Psi_x \otimes \Psi_y)/\alpha$. The operation $\boxtimes_\alpha$ is not associative, that is, there are set-valuations Ψ_x, Ψ_y and Ψ_z such that $(\Psi_x \boxtimes_\alpha \Psi_y) \boxtimes_\alpha \Psi_z$ differs from $\Psi_x \boxtimes_\alpha (\Psi_y \boxtimes_\alpha \Psi_z)$. Nevertheless, the order in which sets are α-combined does not alter the combined approximation factor, as the following result states.

Lemma 9. *If $\Psi_1, \ldots, \Psi_m$ are set-valuations, then $\Psi_1 \boxtimes_\alpha \cdots \boxtimes_\alpha \Psi_m$ (where the operations are applied in any order) is a $\leq_\beta$-covering for $\Psi_1 \otimes \cdots \otimes \Psi_m$, where $\beta = \alpha^{m-1}$.*

Proof. We work by induction on $k = 2, \ldots, m$. For $k = 2$, it follows directly from the definition of α-combination that $\Psi_1 \boxtimes_\alpha \Psi_2$ is an $\leq_\alpha$-combination for $\Psi_1 \otimes \Psi_2$. Assume for $k \in \{2, \ldots, m-1\}$ that $\Psi_1 \boxtimes_\alpha \cdots \boxtimes_\alpha \Psi_{k-1}$ is a $\leq_\beta$-covering for $\Psi_1 \otimes \cdots \otimes \Psi_{k-1}$, where $\beta = \alpha^{k-2}$. Consider any pair $\phi = \phi' \times \phi''$ in $\Psi_1 \otimes \cdots \otimes \Psi_k$, where $\phi' \in \Psi_1 \otimes \cdots \otimes \Psi_{k-1}$ and $\phi'' \in \Psi_k$. There is $\psi = \psi' \times \psi''$ in $\Psi_1 \boxtimes_\alpha \cdots \boxtimes_\alpha \Psi_{k-1} \otimes \Psi_k$, where $\psi' \in \Psi_1 \boxtimes_\alpha \cdots \boxtimes_\alpha \Psi_{k-1}$ and $\psi'' \in \Psi_k$, such that $\psi' \geq_\beta \phi'$ (by assumption) and $\psi'' = \phi''$. Then it follows from (A4) that $\psi \geq_\beta \phi$, or equivalently that $(\beta^{-1}, \beta) \times \psi \geq \phi$. But since $\Psi_1 \boxtimes_\alpha \cdots \boxtimes_\alpha \Psi_k$ is a $\leq_\alpha$-covering for $\Psi_1 \boxtimes_\alpha \cdots \boxtimes_\alpha \Psi_{k-1} \otimes \Psi_k$, there is $\psi''' \in \Psi_1 \boxtimes_\alpha \cdots \boxtimes_\alpha \Psi_k$ such that $\psi''' \geq_\alpha \psi$, or equivalently that $(\alpha^{-1}, \alpha) \times \psi''' \geq \psi$. By combining both sides with (β^{-1}, β) and applying (A4) we get to

$$(\beta^{-1}, \beta) \times (\alpha^{-1}, \alpha) \times \psi''' \geq (\beta^{-1}, \beta) \times \psi \geq \phi\,,$$

and hence $(\alpha^{-(k-1)}, \alpha^{k-1}) \times \psi''' \geq \phi$, and $\psi''' \geq_{\alpha^{k-1}} \phi$. The lemma follows from the induction. □

Thus, by properly choosing the value of α we can obtain a covering that approximates a combination of set-valuations with errors as small as we want. In addition, Proposition 7 guarantees that the sets obtained after each α-combination have cardinality polynomial in the input length and in the maximum error, and so the covering.

We can then modify the exact variable elimination algorithm devised in Section 4 to provide an FPTAS by substituting max-combination and max-marginalization by α-combination with $\alpha = 1 + \epsilon/4n$ and set-marginalization. Let Ψ^i and Ψ^i_α denote, respectively, the sets obtained in the ith iteration of the loop step of variable elimination using set-combination and α-combinations (and both with set-marginalization). In other words, Ψ^i is the set obtained by a brute-force elimination algorithm, whereas Ψ^i_α denote the sets obtained by the approximation algorithm. Similarly, we let Γ and Γ_α denote the outputs of variable elimination with set-combination and α-combination, respectively.

Let s_1 denote the number of set-valuations that are combined to compute Ψ^1_α (and also Ψ^1) minus one, that is, $s_1 \triangleq |\mathcal{B}_1| - 1$. Then, for $i = 2, \ldots, n$, we define s_i recursively as $s_i \triangleq |\mathcal{B}_i| - 1 + \sum_{j:\Psi^j_\alpha \in \mathcal{B}_i} s_j$. Intuitively, s_i denote the number of valuations from the input that are required either directly or indirectly to compute Ψ^i_α (and also Ψ^i) minus one. Hence, if Ψ is the set obtained after the loop step, we have that $|\Gamma_\alpha| + \sum_{i:\Psi^i_\alpha \in \Psi} s_i = n$, since there are n set-valuations given as input and each is used exactly once in the computation of some Ψ^i_α (or Ψ^i).

The following lemma relates the set-valuations propagated by variable elimination with α-combination to the corresponding sets obtained by set-combination.

Lemma 10. *For $i = 1, \ldots, n$, the set-valuation Ψ^i_α is a $\leq_{\alpha^{s_i}}$-covering for Ψ^i.*

Proof. For $i = 1$ the result follows directly from Lemma 9. Without loss of generality, let $\Psi^i = [\Psi_1 \otimes \cdots \otimes \Psi_k \otimes \cdots \otimes \Psi_{|\mathcal{B}_i|}]^{-X_i}$, where $\Psi_1, \ldots, \Psi_k$ denote set-valuations given as input and $\Psi_{k+1}, \ldots, \Psi_{|\mathcal{B}_i|}$ denote sets Ψ^j ($j < i$) generated in the propagation step. Similarly, let $\Psi^i_\alpha = [\Psi_1 \boxtimes_\alpha \cdots \boxtimes_\alpha \Psi_k \boxtimes_\alpha \Psi'_{k+1} \boxtimes_\alpha \cdots \boxtimes_\alpha \Psi'_{|\mathcal{B}_i|}]^{-X_i}$, where, for $k+1 < \ell < |\mathcal{B}_i|$, $\Psi_\ell = \Psi^j$ implies $\Psi'_\ell = \Psi^j_\alpha$. Assume by induction that the result holds for $1, \ldots, i-1$. Hence, if $\Psi'_\ell = \Psi^j_\alpha$ then Ψ'_ℓ is a $\leq_{\alpha^{s_j}}$-covering for Ψ_ℓ. Now, consider any pair $\phi = [\phi' \times \phi'']^{-X_i} \in \Psi^i$, where $\phi' \in \Psi_1 \otimes \cdots \otimes \Psi_k$ and $\phi'' \in \Psi_{k+1} \otimes \cdots \otimes \Psi_{|\mathcal{B}_i|}$. From Lemma 9, we have that there is $\psi' \in \Psi_1 \boxtimes_\alpha \cdots \boxtimes_\alpha \Psi_k$ such that $(\alpha^{-k+1}, \alpha^{k-1}) \times \psi' \geq \phi'$. Likewise, since $\Psi'_{k+1} \boxtimes_\alpha \cdots \boxtimes_\alpha \Psi'_{|\mathcal{B}_i|}$ is a $\leq_{\alpha^{|\mathcal{B}_i|-(k+1)}}$-covering for $\Psi'_{k+1} \otimes \cdots \otimes \Psi'_{|\mathcal{B}_i|}$ (by Lemma 9) and $\Psi'_{k+1} \otimes \cdots \otimes \Psi'_{|\mathcal{B}_i|}$ is a $\leq_{\alpha^{\sum_\ell s_\ell}}$-covering for $\Psi_{k+1} \otimes \cdots \otimes \Psi_{|\mathcal{B}_i|}$ (by Lemma 6 and the induction hypothesis), there is $\psi'' \in \Psi'_{k+1} \boxtimes_\alpha \cdots \boxtimes_\alpha \Psi'_{|\mathcal{B}_i|}$ such that $(\alpha^{-s_i+k}, \alpha^{s_i-k}) \times \psi'' \geq \phi''$. Since $\equiv_\alpha$ implies $\leq_\alpha$, there is $\psi \in (\Psi_1 \boxtimes_\alpha \cdots \boxtimes_\alpha \Psi_k) \boxtimes_\alpha (\Psi_{k+1} \boxtimes_\alpha \cdots \boxtimes_\alpha \Psi_{|\mathcal{B}_i|})$ such that $(\alpha^{-1}, \alpha) \times \psi \geq \psi' \times \psi''$. Thus, it follows from (A4) that $[(\alpha^{-s_i}, \alpha^{s_i}) \times \psi]^{-X_i} \geq \phi$. But from (A3) we have that $[(\alpha^{-s_i}, \alpha^{s_i}) \times \psi]^{-X_i} = (\alpha^{-s_i}, \alpha^{s_i}) \times \psi^{-X_i}$, where $\psi^{-X_i} \in \Psi^i_\alpha$. Since this is true for any $\phi \in \Psi^i$, the result holds for i. The lemma follows from the induction. □

Consider a credal network $(\mathcal{G}, \mathbb{K})$, an elimination ordering $o = (X_1, \ldots, X_n)$ of the variables in $\mathcal{U}$, sets of query and evidence variables q and e, and a query-evidence pair $(\mathbf{q}, \mathbf{e}) \in \Omega_{q \cup e}$. Let Ψ be a collection of sets of pairs as defined in Section 4, and consider the variable elimination algorithm with inputs Ψ, $y = \emptyset$ and o, and α-combination and set-marginalization. Finally, return $p_{\mathbf{q}|\mathbf{e}} \triangleq \max_{(p^\ell, p^r) \in \Gamma_\alpha} (1 + p^\ell/p^r)^{-1}$ as the approximate solution output.

Theorem 11. *The procedure described is an FPTAS for computing upper posterior probabilities for networks of bounded treewidth and number of states per variable.*

Proof. First, we analyze the time complexity of the algorithm. We are thus interested in the maximum cardinality of a set Ψ^i_α, and in the cardinality of the domain of a valuation generated in the loop step. The boundedness assumptions imply that the cardinality of the domain of any propagated valuation is smaller than a constant. Hence, the polynomial time complexity depends on $|\Psi^i_\alpha|$ being bounded. For $i = 1, \ldots, n$, any valuation $\phi^i \in \Psi^i_\alpha$ is produced by first combining valuations that are either in some previously generated set Ψ^j_α ($j < i$) or in a set given as input, and then eliminating X_i from it. Thus, by recursively applying (A1)–(A3) to factorize each valuation from a Ψ^j_α into a combination of valuations and moving the eliminations out, we have that $\phi^i = [\phi_1 \times \cdots \times \phi_{s_i+1}]^{-\{X_1, \ldots, X_i\}}$, where each ϕ_j is in a set-valuation given as input. Hence, each Ψ^i_α can be factorized as $[\Psi_1 \otimes \cdots \otimes \Psi_{s_i}]^{-\{X_1, \ldots, X_i\}}$, where each Ψ_i is a subset of a set-valuation given as input. It follows then from Proposition 7 that Ψ^i_α has $O([bs_i\alpha/(\alpha-1)]^{2\omega})$, where ω is a constant greater than the cardinality of the domain of any ϕ^i. Since $\alpha = 1 + \epsilon/4n$, $O([bs_i\alpha/(\alpha-1)]^{2\omega}) \leq O((4n^2b/\epsilon)^{2\omega})$, where b is the length of the input in bits. Therefore the algorithm runs in time polynomial in the input, in the given approximation factor ϵ, and in the number of variables n.

Let $\overline{p}(\mathbf{q}|\mathbf{e}) \triangleq \max_{(p^\ell_*, p^r_*) \in \Gamma} (1 + p^\ell_*/p^r_*)^{-1}$ denote the optimum value. We now show that the approximation algorithm yields a solution such that $p_{\mathbf{q}|\mathbf{e}} \geq \overline{p}(\mathbf{q}|\mathbf{e})/(1+\epsilon)$ for any given positive ϵ. Let $\Psi'_1, \ldots, \Psi'_m$ denote the sets Ψ^i_α in Ψ after the loop step of the approximation algorithm, where $m = |\Gamma_\alpha|$, and let $\Psi_1, \ldots, \Psi_m$ be the sets Ψ^i in Ψ after the loop step of the brute-force version. Then, $\Gamma_\alpha = \Psi'_1 \boxtimes_\alpha \cdots \boxtimes_\alpha \Psi'_m$ and $\Gamma = \Psi_1 \otimes \cdots \otimes \Psi_m$. It follows from Lemma 9 that Γ_α is a $\leq_{\alpha^{m-1}}$-covering for $\Psi'_1 \otimes \cdots \otimes \Psi'_m$, which in turn is a $\leq_{\alpha^{n-m}}$-covering for Γ, by Lemma 10. Hence, for any $\phi \in \Gamma$ there is $\psi \in \Gamma_\alpha$ such that $(\alpha^{-(n-1)}, \alpha^{n-1}) \times \psi \geq \phi$ an thus $(\alpha^{-n}, \alpha^n) \times \psi \geq \phi$. In particular, there is $\psi = (p^\ell, p^r) \in \Gamma_\alpha$ such that $\psi \geq_{\alpha^n} (p^\ell_*, p^r_*) = \phi^*$. Therefore, $p^\ell \leq \alpha^n p^\ell_*$, $\alpha^n p^r \geq p^r_*$, and

$$\begin{aligned}(1 + p^\ell/p^r)^{-1} &\geq (1 + \alpha^{2n} p^\ell_*/p^r_*)^{-1} \\ &\geq \alpha^{-2n} (1 + p^\ell_*/p^r_*)^{-1}\,.\end{aligned}$$

Since $\alpha = (1 + \epsilon/4n)$, we have that

$$\begin{aligned}(1 + p^\ell/p^r)^{-1} &\geq (1 + \epsilon/4n)^{-2n} (1 + p^\ell_*/p^r_*)^{-1} \\ &\geq (1+\epsilon)^{-1} (1 + p^\ell_*/p^r_*)^{-1} \\ &= (1+\epsilon)^{-1} \overline{p}(\mathbf{q}|\mathbf{e})\,,\end{aligned}$$

where the second passage is due to the inequality $(1 + x/z)^z \leq 1 + 2x$, valid for any $x \in [0, 1]$ and any positive integer z. Hence, $p_{\mathbf{q}|\mathbf{e}} \geq \overline{p}(\mathbf{q}|\mathbf{e})/(1+\epsilon)$. □

Finally, we note that the approximation algorithm can be made more efficient by discarding non-maximal pairs from sets Ψ^i_α like in the exact algorithm in Section 4. This is done in our implementation of the algorithm whose performance we evaluate in the next section.

6 Experiments

We evaluate the performance of the exact and the approximation algorithms on a collection of extensively specified credal networks randomly generated using the BNGen package [14]. The graph topology of these networks is divided in three types, namely (from the simplest to the

Type	Exact Method				Approx. ($\epsilon = 0.1$)				Integer Programming			
	%	Time(sec)			%	Time(sec)			%	Time(sec)		
	solved	Median	Avg.	SD	solved	Median	Avg.	SD	solved	Median	Avg.	SD
M10-2-16	20	824	5617	9923	21	955	6978	11157	6	40464	35079	10451
M10-2-2	100	0.04	0.04	0.03	100	0.04	0.04	0.03	100	2	6	8
M10-2-4	100	4	1096	3906	100	3	276	1025	73	11445	13487	9206
M10-4-2	100	0.19	0.38	0.46	100	0.2	0.41	0.49	75	1320	5699	8922
M10-4-4	100	248	2030	4407	100	238	1992	4335	3	8459	8459	1108
M20-2-2	95	113	1835	4304	96	95	1592	3747	46	8039	12601	10654
M20-4-2	81	1154	5864	9584	81	1266	6009	9594	0	–	–	–
M30-2-2	26	8560	12170	11710	30	4032	13775	13734	3	9484	9484	0
P10-4-16	10	15428	16877	14159	10	16719	16470	13080	0	–	–	–
P10-4-2	100	0.04	0.04	0.03	100	0.04	0.05	0.03	96	248	2451	7752
P10-4-4	100	4	1977	5075	100	4	2095	5476	6	15101	15101	1564
P20-4-2	100	39	2055	5097	100	32	1691	4483	0	–	–	–
P20-4-4	6	20669	20669	20588	6	13484	13484	13400	36	5393	8931	6016
P30-4-2	6	8207	8207	1385	6	5171	5171	1306	0	–	–	–
T10-4-16	13	1559	1381	687	16	1855	9778	16704	0	–	–	–
T10-4-2	100	0.04	0.04	0.02	100	0.04	0.05	0.02	100	12	14	7
T10-4-4	100	6	784	3554	100	6	674	3129	0	–	–	–
T20-4-2	96	89	2415	6164	96	73	2597	7009	13	29022	29587	4839

Table 2: Performance of proposed methods and the integer programming idea. Columns show percentage of solved cases, median, mean and standard deviation (SD) for each group. Numbers greater than one are truncated.

most complicated): trees (graphs with maximum in-degree one), polytrees (graphs where the underlying undirected graph has no cycles), and multi-connected (DAGs without restrictions). All networks have treewidth no greater than four, 10 to 30 nodes, 2 to 4 states per variable, and 2 to 16 potentials in each local extensive credal set. In order to have statistically significant measures, we group networks of similar structure which we identify by the notation *Sn-k-c*, where S is one of T (for trees), P (for polytrees), or M (for multi-connected), n is the number of nodes in the graph, k is the number of states per node, and c is the cardinality of the credal sets. The number of networks in each group is either 30 or 60 (see second column of Table 3). For each network, we set some evidence to every leaf node and arbitrarily choose a node with no parents as query. This creates problems where a brute-force approach would have to execute c^n Bayesian network inferences. The elimination ordering is obtained by a greedy algorithm that attempts to minimize the size of propagated set-valuations. To make the removal of non-maximal valuations effective, we ensure the set-valuation Ψ_q is in $\mathcal{B}_1$, even if it is not required (i.e., if $X_1 \notin q$). Since the query has no parents, this can only increase the treewidth by one.

Table 2 reports the performance of the exact and the approximation algorithms along with the integer programming method of de Campos and Cozman [8]. The latter is a state-of-the-art solver for inference in credal networks that performs a symbolic inference in the credal network to obtain a set of linear constraints over continuous and binary optimization variables, which is then processed by a mixed integer programming solver. For each inference method and network group, Table 2 contains the percentage of cases that were correctly solved using at most 12 hours of CPU time and 2GB of RAM, and the median, average and standard deviation of the time spent. Regarding the mixed integer programming, we considered an instance solved only if the lower and upper bounds given returned by the solver matched. As the networks become more complicated, the percentage of solved cases reduces and the time to solve each case increases. The superiority against the integer programming is clear, though we suspect the integer programming might be suffering from numerical issues that are preventing it to achieve better results. Regarding the approximation, we see no significant reduction in time nor increase in the number of solved cases with respect to the exact method. Some facts contribute to that: (i) the limit of 12 hours of computation might be too short to get a consistent difference in the performance of the methods; (ii) the approximation has an additional computational cost in removing α-equivalent pairs, which is asymptotically irrelevant but significant otherwise; (iii) the number of discarded potentials in each step depends on the elimination order, the dimension of the potentials, and the randomness of input values. Table 3 shows average and standard deviation of the maximum number of elements in a set generated by the exact and approximation algorithms in the loop step. Recall that the complexity is related to the number of elements (as well as the cardinality) of the set-valuations generated. For instance, there would eventually be c^n propagated potentials if no $\leq$ relation (conversely, $\leq_\alpha$ relation) was observed.

Type	# of nets	Exact Avg.	Exact SD	Approximation Avg.	Approximation SD
M10-2-16	60	36046	28928	34579	28563
M10-2-2	60	154	141	130	109
M10-2-4	60	24642	64632	7254	9439
M10-4-2	60	225	128	224	127
M10-4-4	60	46147	65056	42664	55941
M20-2-2	60	37515	61606	28977	46774
M20-4-2	60	67573	73868	66185	73362
M30-2-2	30	93213	55519	81624	57996
P10-4-16	30	104468	75687	92784	64183
P10-4-2	30	115	100	114	100
P10-4-4	30	37155	78008	31361	64117
P20-4-2	30	24856	44469	20337	37219
P20-4-4	30	76083	68966	58358	51241
P30-4-2	30	92744	5476	65708	16654
T10-4-16	30	11840	9570	11834	9572
T10-4-2	30	135	108	132	107
T10-4-4	30	17178	49396	13706	41225
T20-4-2	30	57055	104187	49044	96469

Table 3: Average and standard deviation (SD) of the maximum number of pairs of a set for the cases where both methods solved the inference. Numbers are truncated.

7 Conclusion

We derived a new algorithm for exact posterior inference in extensively specified credal networks under strong independence. The algorithm is empirically shown to outperform an state-of-the-art method, being able to solve medium-sized networks in feasible time. We then showed that for networks of bounded treewidth and number of states per variable, a FPTAS for the problem exists. In our experiments, approximation and exact algorithms performed similar, likely due to the large constants hidden by the boundedness assumptions in the asymptotic complexity analysis.

Acknowledgements

This work was partially supported by the Swiss NSF grants n. 200020_134759 / 1, 200020-121785 / 1, and by the Hasler foundation grant n. 10030.

References

[1] C. Papadimitriou and M. Yannakakis. On the approximability of the trade-offs and optimal access of web sources. In *Proc. of the Annual Symp. on Foundations of Computer Science*, 2000.

[2] A. Antonucci and M. Zaffalon. Decision-Theoretic Specification of Credal Networks: A Unified Language for Uncertain Modeling with Sets of Bayesian Networks. *Intl. J. Approx. Reasoning* 49(2), 2008.

[3] A. Antonucci, Y. Sun, C. P. de Campos and M. Zaffalon. Generalized loopy 2U: a new algorithm for approximate inference in credal networks. *Intl. J. Approx. Reasoning* 55(5), 2010.

[4] D. Avis. Living with lrs. In *Discrete and Computational Geometry*, Lecture Notes in Computer Science, Springer, 2000.

[5] J. L. Bentley, H. T. Kung, M. Schkolnick and C. D. Thompson. On the average number of maxima in a set of vectors and applications. *J. of the ACM* 25, 1978.

[6] C. P. de Campos and F. G. Cozman. Inference in credal networks using multilinear programming. In *Proc. of the Starting AI Researchers' Symp.*, 2004.

[7] C. P. de Campos and F. G. Cozman. The inferential complexity of Bayesian and credal networks. In *Intl. Joint Conf. on Artif. Intelligence*, 2005.

[8] C. P. de Campos and F. G. Cozman. Inference in credal networks through integer programming. In *Proc. of the Intl. Symp. on Imprecise Probability: Theories and Applications*, 2007.

[9] A. Cano, M. Gomez, S. Moral and J. Abellan. Hill-climbing and branch-and-bound algorithms for exact and approximate inference in credal networks. In *Intl. J. Approx. Reasoning* 44(3), 2007.

[10] G. de Cooman, F. Hermans, A. Antonucci, and M. Zaffalon. Epistemic irrelevance in credal nets: The case of imprecise Markov trees. In *Intl. J. Approx. Reasoning* 51(9), 2010.

[11] F. G. Cozman. Credal networks. *Artif. Intelligence* 120(2), 2000.

[12] E. Fagiuoli and M. Zaffalon. 2U: an exact interval propagation algorithm for polytrees with binary variables. *Artif. Intelligence* 106(1), 1998.

[13] R. Haenni. Ordered valuation algebras: a generic framework for approximating inference. *Intl. J. Approx. Reasoning* 37(1), 2004.

[14] J. S. Ide , F. G. Cozman and F. T. Ramos. Generating Random Bayesian Networks with Constraints on Induced Width. In *Proc. of the European Conf. on Artif. Intelligence*, 2004.

[15] J. Kohlas. Information Algebras: Generic Structures for Inference. Springer-Verlag, 2003.

[16] J. C. F. da Rocha and F. G. Cozman. Inference in credal networks: branch-and-bound methods and the A/R+ algorithm. *Intl. J. Approx. Reasoning* 39(3), 2005.

[17] P. Shenoy and G. Shafer Axioms for Probability and Belief-Function Propagation. In *Proc. of the Conf. on Uncertainty in Artif. Intelligence*, 1988.

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

Conglomerable Natural Extension

Enrique Miranda
University of Oviedo, Spain
mirandaenrique@uniovi.es

Marco Zaffalon
IDSIA, Lugano, Switzerland
zaffalon@idsia.ch

Gert de Cooman
SYSTeMS, Ghent University, Belgium
gert.decooman@UGent.be

Abstract

We study the weakest conglomerable model that is implied by desirability or probability assessments: the *conglomerable natural extension*. We show that taking the natural extension of the assessments while imposing conglomerability—the procedure adopted in Walley's theory—does not yield, in general, the conglomerable natural extension (but it does so in the case of the marginal extension). Iterating this process produces a sequence of models that approach the conglomerable natural extension, although it is not known, at this point, whether it is attained in the limit. We give sufficient conditions for this to happen in some special cases, and study the differences between working with coherent sets of desirable gambles and coherent lower previsions. Our results indicate that it might be necessary to re-think the foundations of Walley's theory of coherent conditional lower previsions for infinite partitions of conditioning events.

Keywords. Conglomerability, natural extension, desirable gambles, coherent lower previsions.

1 Introduction

You are offered a *gamble* f (that is, a bounded real-valued function representing an uncertain reward) on a possibility space Ω. You assess that, whatever event B you consider in a certain partition $\mathcal{B}$ of Ω, you would desire f conditional on B. Does this imply that you should *unconditionally* desire f?

Common axioms of desirability, such as those in Refs. [11, Section 3.7] or [12], imply that this should indeed be the case, at least when $\mathcal{B}$ is finite. When $\mathcal{B}$ is infinite, some authors have proposed to impose the above requirement through an axiom of so-called *conglomerability*. In fact, conglomerability is a key founding axiom for Walley's theory of *coherent lower previsions* in the conditional case with infinite partitions of conditioning events.

Conglomerability was introduced by de Finetti [2, 3] as a property that a finitely—but not countably—additive probability need not satisfy. In fact, de Finetti was also the first to reject the idea that conglomerability should be required as an axiom of rationality. The concept was studied later by Dubins [5], who established a connection with *disintegrability*. The property of conglomerability was also studied by Seidenfeld, Schervisch and Kadane (e.g., in Refs. [9, 10]). They show in particular [9] that when a probability is defined on all events and takes infinitely many values, countable additivity is equivalent to *full* conglomerability, that is, for conglomerability to hold with respect to all the possible partitions of Ω. See Ref. [4] for an interesting connection with imprecise probability models.

Requiring conglomerability, even only with respect to a single partition $\mathcal{B}$, comes at the expense of some undesirable mathematical properties: for example, a conglomerable coherent lower prevision may not be the lower envelope of conglomerable linear previsions. Perhaps also because of this, conglomerability was rejected in some extensions of de Finetti's work, such Williams's [12] (see also Ref. [8]). In our view, what appears to be mostly controversial is in particular the idea of full conglomerability, as opposed to conglomerability only for the partitions that are actually used for updating beliefs.[1] This is for instance the approach taken by De Cooman and Hermans in Ref. [1] when they require modes to be 'cut conglomerable'.

Here, we do not take any philosophical position about whether models should be conglomerable. Our aim is to perform a technical study of the impact of conglomerability on the possible extensions of an initial set of assessments. We focus in particular on what we call the *conglomerable natural extension*: loosely speaking, this is the weakest (least committal) conglomerable model that is implied by the initial assessments. A related concept is the *natural extension*, which is defined

[1] This is also called *partial conglomerability*. Here, when we talk about conglomerability, we mean partial conglomerability.

in a similar way except for not requiring the extension to be conglomerable.

We start in Section 2 by introducing some basic notions: desirability, coherent lower previsions and the connections between them. We introduce conglomerability in a few different forms: for desirable gambles, in the traditional form and in a weaker variant; and for coherent lower previsions, in the traditional way and in a strengthened form. We uncover the relationships between these notions, which allows us to convert problems formulated for one model into the other.

In Section 3, we focus on desirability. We show that, if it exists, then the conglomerable natural extension $\mathcal{F}$ of a set $\mathcal{R}$ of desirable gambles with respect to a partition $\mathcal{B}$ is the intersection of all conglomerable sets including $\mathcal{R}$. Moreover, we relate $\mathcal{F}$ to the natural extension: we start from $\mathcal{R}$, take its natural extension, and close it with respect to conglomerability, obtaining $\mathcal{E}_1$; we reiterate this process, yielding the sequence $\mathcal{E}_2$, $\ldots$, $\mathcal{E}_n$, $\ldots$. We show that $\mathcal{E}_n \subseteq \mathcal{F}$ for all n, and that the sequence stabilises if and only if one if its elements coincides with $\mathcal{F}$. We provide some sufficient conditions for this, as well as a few examples to illustrate the situation. One of them, in particular, shows that the gambles in $\mathcal{R}$ that do not satisfy conglomerability may be only in the border of the set, and yet the closure with respect to conglomerability may extend the set beyond this border.

In Section 4, we study the conglomerable natural extension $\underline{F}$ of a coherent lower prevision $\underline{P}$ with respect to a partition $\mathcal{B}$. Here, too, we consider a sequence: we start from $\underline{P}$, compute its conditional natural extension $\underline{P}(\cdot|\mathcal{B})$, and then the natural extension of the two of them together, $\underline{E}_1$; we iterate the process, yielding the sequence $\underline{E}_2$, $\ldots$, $\underline{E}_n$, $\ldots$. We show that $\underline{E}_n \leq \underline{F}$ for all n, and again that the sequence stabilises if and only if one of its elements coincides with $\underline{F}$. We then provide what is arguably the most important result of this paper: we show in Example 5 that $\underline{E}_1$ may not equal $\underline{F}$. This is interesting because, when it comes to natural extension (as well as coherence), Walley's theory is implicitly based on stopping at the first element of the sequence: $\underline{E}_1$. We show that this is not enough to fully capture the implications of conglomerability, and give sufficient conditions for $\underline{E}_1 = \underline{F}$.

In Section 5, we relate the results obtained for desirable gambles and coherent lower previsions: we start from the set $\mathcal{R}$ and induce from this a coherent lower prevision $\underline{P}$. We then create the sequences of sets $\mathcal{E}_n$, on the one hand, and the sequences of coherent lower previsions $\underline{E}_n$, on the other. We investigate the relationship between the elements of these sequences. This allows us, in Example 7, to exploit Example 5 to show that $\mathcal{E}_1$ may not coincide with $\mathcal{F}$: this means that taking the one-step conglomerable closure falls short of the mark for desirable gambles as well. We give sufficient conditions for $\mathcal{E}_1 = \mathcal{F}$, as well as for the two sequences to be made up of equivalent models.

To conclude, we consider in Section 6 the problem of dealing with more than one partition. We show that under the assumptions of the *Marginal Extension Theorem* (see Refs. [11, Theorem 6.7.2] and [6]), it does hold that $\mathcal{E}_1 = \mathcal{F}$.

Due to lack of space, we must assume the reader has a working knowledge of the basics of the theory of coherent lower previsions [11]. We refrain from giving proofs of most technical results for the same reason.

2 Introductory Notions

Consider a possibility space Ω. In this paper Ω will frequently be $\mathbb{N}$, the set of natural numbers without zero, but our results will be applicable to more general spaces. A *gamble* is a map $f\colon \Omega \to \mathbb{R}$. The set of all gambles defined on Ω is denoted by $\mathcal{L}(\Omega)$, or simply $\mathcal{L}$ when there is no ambiguity about the possibility space we are working with. In particular, we use '$f \lneq 0$' to mean '$f \leq 0$ and $f \neq 0$' (and we then say that the gamble f is *negative*), and we write $f \gneq 0$ if $-f \lneq 0$.

Given a set of gambles $\mathcal{R}$, we consider the following axioms of desirability:[2]

D1. $f \gneq 0 \Rightarrow f \in \mathcal{R}$;

D2. $0 \notin \mathcal{R}$;

D3. $f \in \mathcal{R}, \lambda > 0 \Rightarrow \lambda f \in \mathcal{R}$;

D4. $f, g \in \mathcal{R} \Rightarrow f + g \in \mathcal{R}$.

Let us define

$$\operatorname{posi}(\mathcal{R}) := \left\{ \sum_{k=1}^{n} \lambda_k f_k \colon f_k \in \mathcal{R}, \lambda_k > 0, n \geq 1 \right\}.$$

We call $\mathcal{R}$ a *convex cone* if it is closed under positive linear combinations, meaning that $\operatorname{posi}(\mathcal{R}) = \mathcal{R}$. This is equivalent to $\mathcal{R}$ satisfying conditions D3 and D4.

Given a partition $\mathcal{B}$ of Ω, $\mathcal{R}$ is called $\mathcal{B}$-*conglomerable* when it satisfies the following axiom:

D5. if $f \neq 0$ and $(\forall B \in \mathcal{B}' \subseteq \mathcal{B}) Bf \in \mathcal{R}$ then $\sum_{B \in \mathcal{B}'} Bf \in \mathcal{R}$.

[2] This axiomatic definition is related to *strict* and *almost*-desirability, see Ref. [11, Section 3.7]. The differences between these concepts lie mostly in the topological properties of the set of desirable gambles and in whether the zero gamble is considered to be desirable.

Axiom D5 is a consequence of D4 when $\mathcal{B}$ is finite. It can be easily checked that D5 is equivalent to:

D5'. if $f \neq 0$ and $(\forall B \in \mathcal{B})Bf \in \mathcal{R} \cup \{0\}$ then $\sum_{B\in\mathcal{B}} Bf \in \mathcal{R}$.

A *lower prevision* is a real-valued functional defined on some set of gambles $\mathcal{K} \subseteq \mathcal{L}$. When $\mathcal{K}$ is a linear space, $\underline{P}$ is called *coherent* when it satisfies the following conditions:

C1. $\underline{P}(f) \geq \inf f$ for all $f \in \mathcal{K}$;

C2. $\underline{P}(\lambda f) = \lambda \underline{P}(f)$ for all $f \in \mathcal{K}$ and $\lambda > 0$;

C3. $\underline{P}(f + g) \geq \underline{P}(f) + \underline{P}(g)$ for all $f, g \in \mathcal{K}$.

When $\mathcal{K} = \mathcal{L}$ and $\underline{P}$ satisfies C3 with equality, it is called a *linear prevision*. The set of linear previsions that dominate a coherent lower prevision $\underline{P}$ on its domain is denoted by $\mathcal{M}(\underline{P})$.

Given a partition $\mathcal{B}$ of Ω, a *conditional lower prevision* $\underline{P}(\cdot|\mathcal{B})$ on $\mathcal{L}$ is a functional such that for every $B \in \mathcal{B}$, $\underline{P}(\cdot|B)$ is a lower prevision on $\mathcal{L}$. It is called *separately coherent* when $\underline{P}(\cdot|B)$ is coherent and $\underline{P}(B|B) = 1$ for every $B \in \mathcal{B}$. For a lower prevision $\underline{P}$ and a conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$, we use the notation

$$G_{\underline{P}}(f) := f - \underline{P}(f),\ G_{\underline{P}}(f|B) := B(f - \underline{P}(f|B)$$
$$G_{\underline{P}}(f|\mathcal{B}) := f - \underline{P}(f|\mathcal{B}) = \sum_{B\in\mathcal{B}} G_{\underline{P}}(f|B).$$

When both $\underline{P}$ and $\underline{P}(\cdot|\mathcal{B})$ are defined on $\mathcal{L}$, they are called *coherent* if and only if $\underline{P}(G_{\underline{P}}(f|\mathcal{B})) \geq 0$ and

$$\underline{P}(G_{\underline{P}}(f|B)) = 0 \qquad \text{(GBR)}$$

for every gamble f and every $B \in \mathcal{B}$. This last condition is called the *Generalised Bayes Rule*.

Definition 1. Let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}$, and $\mathcal{B}$ a partition of Ω. $\underline{P}$ is called $\mathcal{B}$-*conglomerable* when the following condition holds:

WC. $\underline{P}(\sum_{n\in N} B_n f) \geq 0$ for any $f \in \mathcal{L}$ and any countable number of distinct sets B_n in $\mathcal{B}$ such that $\underline{P}(B_n) > 0$ and $\underline{P}(B_n f) \geq 0$ for all $n \in N$.

Again, WC holds trivially when N is finite, and in particular when the partition $\mathcal{B}$ is finite, because of the super-additivity of coherent lower previsions.

Let us shed more light on the relation between the coherence and conglomerability concepts for lower previsions and sets of desirable gambles. On the one hand, given a coherent lower prevision $\underline{P}$, we define its associated set of *strictly desirable gambles* by

$$\underline{\mathcal{R}} := \{f \in \mathcal{L} \colon f \gtrdot 0 \text{ or } \underline{P}(f) > 0\}, \qquad (1)$$

and its set of *almost-desirable gambles* by

$$\overline{\mathcal{R}} := \{f \in \mathcal{L} \colon \underline{P}(f) \geq 0\}. \qquad (2)$$

$\underline{\mathcal{R}}$ satisfies the axioms D1–D4 considered above, and $\overline{\mathcal{R}}$ is a convex cone that includes all non-negative gambles. Moreover, it follows from the equations above that $\underline{\mathcal{R}} \subseteq \overline{\mathcal{R}}$, and that $\underline{\mathcal{R}}$ contains all positive gambles and is closed under dominance.

Conversely, given a set $\mathcal{R}$ of gambles satisfying D1–D4, we can define the corresponding lower prevision by

$$\underline{P}(f) := \sup\{\mu \colon f - \mu \in \mathcal{R}\}. \qquad (3)$$

It follows from Theorem 6 in Ref. [7] that $\underline{P}$ is a coherent lower prevision. Moreover, if we consider the sets $\underline{\mathcal{R}}$ and $\overline{\mathcal{R}}$ given by Eqs. (1) and (2), it follows from Theorem 3.8.1 in Ref. [11] that

$$\sup\{\mu \colon f - \mu \in \underline{\mathcal{R}}\} = \underline{P}(f) = \sup\left\{\mu \colon f - \mu \in \overline{\mathcal{R}}\right\}.$$

As a consequence, any set $\mathcal{R}$ such that $\underline{\mathcal{R}} \subseteq \mathcal{R} \subseteq \overline{\mathcal{R}}$ induces the same lower prevision $\underline{P}$ through Equation (3) [11, Theorem 3.8.1].

The set $\overline{\mathcal{R}}$ is the closure of $\underline{\mathcal{R}}$ (and as a consequence also of any $\underline{\mathcal{R}} \subseteq \mathcal{R} \subseteq \overline{\mathcal{R}}$) in the topology of uniform convergence [7, Proposition 4]. In addition,

$$\underline{\mathcal{R}} := \{f \in \mathcal{R} \colon f \gtrdot 0 \text{ or } f - \varepsilon \in \mathcal{R} \text{ for some } \varepsilon > 0\},$$

for all $\underline{\mathcal{R}} \subseteq \mathcal{R} \subseteq \overline{\mathcal{R}}$.

We now establish a conglomerability condition for sets of desirable gambles that is equivalent to WC.

Theorem 1. *Let $\mathcal{R}$ be a set of desirable gambles that satisfies* D1–D4*, and $\underline{P}$ be the coherent lower prevision it induces through Equation* (3)*. Then $\underline{P}$ satisfies* WC *if and only if $\mathcal{R}$ satisfies the following condition:*

WD5. *if $(\forall B \in \mathcal{B})Bf \in \underline{\mathcal{R}} \cup \{0\}$ then $f \in \overline{\mathcal{R}}$.*

Since $\underline{\mathcal{R}} \subseteq \mathcal{R} \subseteq \overline{\mathcal{R}}$, D5 implies WD5. On the other hand, when we consider a coherent set of almost-desirable gambles $\overline{\mathcal{R}}$ (see Ref. [11, Section 3.7.3] for the definition), condition D5 is equivalent to:

D5". if $(\forall B \in \mathcal{B})Bf \in \overline{\mathcal{R}}$ then $f \in \overline{\mathcal{R}}$.

By definition, condition D5" is a consequence of D5. To see that they are equivalent when we work with a coherent set of almost-desirable gambles, note that the zero gambles belongs to $\overline{\mathcal{R}}$, and as a consequence if $Bf \in \overline{\mathcal{R}}$ for all $B \in \mathcal{B}' \subseteq \mathcal{B}$, then also $B_1 \sum_{B\in\mathcal{B}'} Bf$ belongs to $\overline{\mathcal{R}}$ for all $B_1 \in \mathcal{B}$; using D5" we then deduce that $\sum_{B\in\mathcal{B}'} Bf$ belongs to $\overline{\mathcal{R}}$.

We next show that D5 can also be related to a notion of conglomerability for coherent lower previsions:

Definition 2. Let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}$, and $\mathcal{B}$ a partition of Ω. $\underline{P}$ is called *strongly $\mathcal{B}$-conglomerable* when the following condition holds:

SC. if $f \in \mathcal{L}$ and $(\forall B \in \mathcal{B}' \subseteq \mathcal{B})\underline{P}(Bf) \geq 0$, then $\underline{P}(\sum_{B\in\mathcal{B}'} Bf) \geq 0$.

Theorem 2. *Let $\underline{P}$ be a coherent lower prevision, and let $\overline{\mathcal{R}}$ be its associated set of almost-desirable gambles. Then $\underline{P}$ is strongly $\mathcal{B}$-conglomerable if and only if $\overline{\mathcal{R}}$ satisfies* D5. *Conversely, a coherent set of almost-desirable gambles satisfies* D5 *if and only if the coherent lower prevision $\underline{P}$ it induces satisfies* SC.

We deduce from Theorems 1 and 2 that if a coherent lower prevision is strongly $\mathcal{B}$-conglomerable, then it is also $\mathcal{B}$-conglomerable.

3 Conglomerability for Sets of Desirable Gambles

Let us consider a set of gambles $\mathcal{R}$, and look for the smallest superset $\mathcal{F}$ (if it exists) that satisfies D1–D5 with respect to a fixed partition $\mathcal{B}$. This set is called the *$\mathcal{B}$-conglomerable natural extension* of $\mathcal{R}$. A first characterisation of this set is given in the following:

Proposition 1. *If there is some set of gambles including $\mathcal{R}$ and satisfying* D1–D4 *and* D5 *(resp.* WD5*), then $\mathcal{F}$ is the intersection of all such sets.*

From now on, we assume that $\mathcal{R}$ satisfies conditions D1–D4; D2 is necessary for the existence of a $\mathcal{B}$-conglomerable natural extension, and D1, D3 and D4 can be satisfied by replacing $\mathcal{R}$ with the convex cone $\text{posi}(\mathcal{R} \cup \{f\colon f \gtrdot 0\})$.

The existence of a superset of $\mathcal{R}$ that satisfies D1–D5 does not guarantee that there is a half-space that includes $\mathcal{R}$ and satisfies these axioms. The example that establishes this is a reformulation of [11, Example 6.6.9]:

Example 1. Let Ω be the set of integers without zero, and consider the partition $\mathcal{B} := \{B_n\colon n \in \mathbb{N}\}$ given by $B_n := \{-n, n\}$.

Let P_1 be a linear prevision on $\mathcal{L}$ satisfying $P_1(\{n\}) = \frac{1}{2^{n+1}}$ and $P_1(\{-n\}) = 0$ for all $n \in \mathbb{N}$, and $P_1(\mathbb{N}) = \frac{1}{2}$. Also consider a linear prevision P_2 satisfying $P_2(\{-n\}) = \frac{1}{3^n}$, $P_2(\{n\}) = 0$ for all $n \in \mathbb{N}$, and $P_2(\mathbb{N}) = \frac{1}{2}$. Let $\underline{P} := \min\{P_1, P_2\}$.

Consider $\underline{\mathcal{R}} := \{f\colon f \gtrdot 0 \text{ or } \underline{P}(f) > 0\}$, the set of strictly desirable gambles associated with $\underline{P}$. Then $\underline{\mathcal{R}}$ satisfies D1–D4. To see that it also satisfies D5, note that if for a gamble $0 \neq f$, $B_n f \in \underline{\mathcal{R}} \cup \{0\}$ for all $n \in \mathbb{N}$, then either $\underline{P}(B_n f) > 0$ or $B_n f \geq 0$, and in the latter case $\underline{P}(B_n f) \geq 0$. But since $\underline{P}(B_n f) > 0$ implies that both $P_1(B_n f) > 0$ and $P_2(B_n f) > 0$, and since this in turn means that both $f(-n)$ and $f(n)$ are non-negative, we also deduce that $\underline{P}(B_n f) > 0$ implies that $B_n f \gtrdot 0$. As a consequence, if $B_n f \in \underline{\mathcal{R}} \cup \{0\}$ for all $B_n \in \mathcal{B}$, then $f \geq 0$, and since $f \neq 0$ we deduce that $f \in \underline{\mathcal{R}}$.

Let us now show that there is no half-space including $\underline{\mathcal{R}}$ and satisfying WD5 (and as a consequence neither D5). Assume *ex absurdo* that $\mathcal{D}$ is such a space. Let P be the associated linear prevision, given by $P(f) := \sup\{\mu\colon f - \mu \in \mathcal{D}\}$. Since $\underline{\mathcal{R}} \subseteq \mathcal{D}$, we deduce that P dominates $\underline{P}$. But Walley has shown in Ref. [11, Example 6.6.9] that no dominating linear prevision satisfies WC, and using Theorem 1, we deduce that $\mathcal{D}$ does not satisfy WD5, and as a consequence it does not satisfy D5 either. ♦

Our next goal is to derive a more practical expression for $\mathcal{F}$. In order to do this, let us define the following sequence of sets of desirable gambles, starting with:

$$\mathcal{R}^* := \{f \neq 0\colon (\forall B \in \mathcal{B})Bf \in \mathcal{R} \cup \{0\}\}$$
$$\mathcal{E}_1 := \text{posi}(\mathcal{R} \cup \mathcal{R}^*)$$

and for all $n \geq 2$:

$$\mathcal{E}^*_{n-1} := \{f \neq 0\colon (\forall B \in \mathcal{B})Bf \in \mathcal{E}_{n-1} \cup \{0\}\}$$
$$\mathcal{E}_n := \text{posi}(\mathcal{E}_{n-1} \cup \mathcal{E}^*_{n-1}). \tag{4}$$

We will also use $\mathcal{E}_0 := \mathcal{R}$ and $\mathcal{E}^*_0 := \mathcal{R}^*$.

Lemma 1. *Let $\mathcal{F}' \supseteq \mathcal{R}$ and suppose that $\mathcal{F}'$ satisfies* D1–D5. *Then $\mathcal{F}' \supseteq \mathcal{E}_n$ for all $n \in \mathbb{N}$.*

It follows that the $\mathcal{B}$-conglomerable natural extension of $\mathcal{R}$, if it exists, must include $\bigcup_n \mathcal{E}_n$. As a consequence, in that case we can also express the sets $\mathcal{E}$ as

$$\mathcal{E}_1 = \{f + g\colon f \in \mathcal{R} \cup \{0\}, g \in \mathcal{R}^* \cup \{0\}\} \setminus \{0\},$$
$$\mathcal{E}_n = \{f + g\colon f \in \mathcal{E}_{n-1} \cup \{0\}, g \in \mathcal{E}^*_{n-1} \cup \{0\}\} \setminus \{0\}.$$

We next investigate which desirability axioms are satisfied by the sets $\mathcal{E}_n$ and $\mathcal{E}^*_n$.

Proposition 2. *Assume that there is some superset $\mathcal{F}$ of $\mathcal{R}$ satisfying* D1–D5. *Then:*

1. *$\mathcal{E}_n$ satisfies* D1–D4 *for all $n \in \mathbb{N}$.*
2. *$\mathcal{E}^*_n$ satisfies* D1–D5 *for all $n \in \mathbb{N}$.*

We can now characterise under which conditions $\mathcal{E}_n$ coincides with the $\mathcal{B}$-conglomerable natural extension, in terms of the desirability axioms:

Proposition 3. *The following conditions are equivalent for any natural number $n \geq 0$:*

1. *$\mathcal{E}^*_n \subseteq \mathcal{E}_n$.*
2. *$\mathcal{E}_n$ satisfies* D5.
3. *$\mathcal{F} = \mathcal{E}_n$.*

This simple result has interesting consequences: on the one hand, if $\mathcal{E}_n$ is not the $\mathcal{B}$-conglomerable natural extension of $\mathcal{R}$, then there must be some gamble f in $\mathcal{E}_n^* \setminus \mathcal{E}_n$, and as a consequence $\mathcal{E}_n$ is a proper subset of $\mathcal{E}_{n+1}$. In other words, the sequence $\mathcal{E}_n$ does not stabilise unless we get to the $\mathcal{B}$-conglomerable natural extension. On the other hand, if $\mathcal{E}_n^* = \mathcal{E}_{n+1}^*$ for some n then $\mathcal{E}_{n+1}^*$ is included in $\mathcal{E}_{n+1}$, and Proposition 3 implies that $\mathcal{E}_{n+1}$ is the $\mathcal{B}$-conglomerable natural extension of $\mathcal{R}$. Hence, we can use both sequences to determine at which step we get to $\mathcal{F}$: $\mathcal{E}_n = \mathcal{F}$ if $\mathcal{E}_{n-1}^* = \mathcal{E}_n^*$, and also if and only if $\mathcal{E}_n = \mathcal{E}_{n+1}$.

Next we provide a sufficient condition for $\mathcal{E}_1$ to coincide with $\mathcal{F}$:

Proposition 4. *Let $\mathcal{R}$ be a set of desirable gambles satisfying* D1–D4*, and assume that its $\mathcal{B}$-conglomerable natural extension $\mathcal{F}$ exists.*

1. $(\forall f \in \mathcal{R})(\forall B \in \mathcal{B}) Bf \in \mathcal{R} \cup \{0\} \Leftrightarrow \mathcal{R}^* = \mathcal{F} \Leftrightarrow \mathcal{R} \subseteq \mathcal{R}^*$.
2. *If there is some superset $\mathcal{Q}$ of $\mathcal{R}$ satisfying* D1–D5 *and such that $\mathcal{Q}^* = \mathcal{R}^*$, then $\mathcal{E}_1 = \mathcal{F}$.*

As a consequence, when $\mathcal{R}$ is included in $\mathcal{R}^*$ the sequence $\mathcal{E}_n$ stabilises in the first step: $\mathcal{E}_1 = \mathcal{F}$.

Let us give an example showing that the inclusion $\mathcal{R} \subseteq \mathcal{R}^*$ does not imply that $\mathcal{R} = \mathcal{R}^*$, or, equivalently, that we may have $\mathcal{R} \subsetneq \mathcal{E}_1 = \mathcal{F}$:

Example 2. Consider $\Omega = \mathbb{N}$, $B_n := \{2n-1, 2n\}$ and $\mathcal{B} := \{B_n : n \in \mathbb{N}\}$. Let $\mathcal{R}$ be the set of gambles f for which there is some $n_f \in \mathbb{N}$ such that

$$f(n_f \to) \gtrsim 0 \text{ and}$$
$$f(2n) + f(2n-1) \geq 0 \text{ and } f(2n) \geq 0 \text{ for all } n \in \mathbb{N},$$

where $(n_f \to) := \{n_f, n_f+1, \dots\}$. Then $\mathcal{R}$ satisfies D1–D4:

D1. Any $f \gtrsim 0$ belongs to $\mathcal{R}$ by definition: take $n_f = 1$.

D2. $0 \notin \mathcal{R}$ by definition.

D3. Let $f \in \mathcal{R}$ and $\lambda > 0$. Then there is some $n_f \in \mathbb{N}$ such that $f(n_f \to) \gtrsim 0$, $f(2n) + f(2n-1) \geq 0$ and $f(2n) \geq 0$ for all $n \in \mathbb{N}$, whence $(\lambda f)(n_f \to) = \lambda(f(n_f \to)) \gtrsim 0$, $(\lambda f)(2n) + (\lambda f)(2n-1) = \lambda(f(2n) + f(2n-1)) \geq 0$ and $\lambda(f(2n)) \geq 0$ for all $n \in \mathbb{N}$. Since moreover $\lambda f \neq 0$ because $f \neq 0$ and $\lambda > 0$, we conclude that $\lambda f \in \mathcal{R}$.

D4. Let $f, g \in \mathcal{R}$. Then there are $n_f, n_g \in \mathbb{N}$ such that $f(n_f \to) \geq 0$ and $g(n_g \to) \geq 0$, whence given $n^* := \max\{n_f, n_g\}$, we infer that $(f+g)(n^* \to) \gtrsim 0$. On the other hand, $(f+g)(2n) + (f+g)(2n-1) = f(2n)+g(2n)+f(2n-1)+g(2n-1) \geq 0$ and $(f+g)(2n) \geq 0$ for all $n \in \mathbb{N}$, whence also $f + g \in \mathcal{R}$.

To see that $\mathcal{R} \subseteq \mathcal{R}^*$, observe that given a gamble $f \in \mathcal{R}$ and $B_n \in \mathcal{B}$, $B_n(f(2m)+f(2m-1)) \geq 0$ and $B_n(f(2m)) \geq 0$ for all $m \in \mathbb{N}$. Moreover, if $B_n f = 0$ then automatically $B_n f \in \mathcal{R} \cup \{0\}$; and if $B_n f \neq 0$ then either $f(2n) > 0$, in which case $B_n f \in \mathcal{R}$ by letting $n_{B_n f} = 2n$, or $f(2n) = 0$, in which case $f(2n-1) > 0$ and $B_n f \in \mathcal{R}$ by letting $n_{B_n f} = 2n-1$.

However, $\mathcal{R}$ does not satisfy D5, and as a consequence it does not coincide with $\mathcal{R}^*$: the gamble g given by $g(2n) = 1, g(2n-1) = -1$ for all n does not belong to $\mathcal{R}$ because there is no natural number n_g for which $g(n_g \to) \gtrsim 0$. On the other hand, for every natural number n, $B_n g$ does belong to $\mathcal{R}$: consider $n_{B_n g} = 2n$. Therefore $g \in \mathcal{R}^*$. ♦

This example also allows us to show that conditions D5 and WD5 are not equivalent:

Example 3. Consider the set $\mathcal{R}$ from Example 2. We have already shown there that $\mathcal{R}$ does not satisfy D5. To see that it satisfies WD5, observe that given a gamble f and $B_n \in \mathcal{B}$, $B_n f$ belongs to $\underline{\mathcal{R}} \cup \{0\}$ if and only if $B_n f \geq 0$, because there is no $\delta > 0$ such that $B_n f - \delta \in \mathcal{R}$. As a consequence, $(\forall B_n \in \mathcal{B}) B_n f \in \underline{\mathcal{R}} \cup \{0\}$ implies that $0 \leq f \in \overline{\mathcal{R}}$. ♦

The same example shows us something else: even if the gambles that violate D5 are only on the border of $\mathcal{R}$, taking the closure of $\mathcal{R}$ with respect to D5 will require us in general to enlarge the set beyond its border.

Example 4. Consider set $\mathcal{R}$ and gamble g from Example 2. Taking into account the observations in Example 3, there is no $\delta > 0$ such that $B_n g - \delta \in \mathcal{R}$, because this gamble is not positive, and on the other hand, we know that $B_n g \in \mathcal{R}$. This means that $B_n g \in \mathcal{R} \setminus \underline{\mathcal{R}} \subseteq \overline{\mathcal{R}} \setminus \underline{\mathcal{R}}$ for all $B_n \in \mathcal{B}$. Now consider any $\delta \in (-1, 0)$, and observe that $g - \delta \notin \mathcal{R}$: in fact, $g(2n-1) - \delta < 0$ for all $n \geq 1$, so there is no $n_g \in \mathbb{N}$ such that $(g-\delta)(n_g \to) \geq 0$. On the other hand, $g + 1 \gtrsim 0$ and hence belongs to $\mathcal{R}$. This means that

$$\sup\{\mu : g - \mu \in \overline{\mathcal{R}}\} = \sup\{\mu : g - \mu \in \mathcal{R}\} = -1,$$

and therefore $g \notin \overline{\mathcal{R}}$. ♦

It is an open problem whether the sequence $\mathcal{E}_n$ always stabilises in a finite number of steps, and, if it does not, whether the sequence limit $\bigcup_{n \in \mathbb{N}} \mathcal{E}_n$ always coincides with the $\mathcal{B}$-conglomerable natural extension $\mathcal{F}$ of $\mathcal{R}$.

4 Conglomerability for Coherent Lower Previsions

We now turn to the relationship between the natural extension studied in Ref. [11, Chapter 8] and the conglomerable natural extension, which we define next. Throughout this section, $\mathcal{B}$ is a partition of Ω.

Definition 3. Let $\underline{P}$ be a coherent lower prevision on $\mathcal{K}$. Its *$\mathcal{B}$-conglomerable natural extension* is the smallest coherent lower prevision $\underline{F}$ on $\mathcal{L}$ that dominates $\underline{P}$ and is $\mathcal{B}$-conglomerable.

There may be no dominating $\mathcal{B}$-conglomerable coherent lower prevision, and then the $\mathcal{B}$-conglomerable

natural extension will not exist. On the other hand, if there is some dominating $\mathcal{B}$-conglomerable coherent lower prevision, then there is a $\mathcal{B}$-conglomerable natural extension, because $\mathcal{B}$-conglomerability is preserved by taking lower envelopes.

We may assume without loss of generality that the domain $\mathcal{K}$ of $\underline{P}$ is the set $\mathcal{L}$ of all gambles: otherwise, it suffices to consider the natural extension $\underline{E}$ of $\underline{P}$ to $\mathcal{L}$. To see that the $\mathcal{B}$-conglomerable natural extensions of $\underline{P}$ and $\underline{E}$ coincide, denote these by $\underline{F}_1$ and $\underline{F}_2$, respectively. Trivially $\underline{F}_2 \geq \underline{F}_1$. Conversely, $\underline{F}_1$ is by definition a $\mathcal{B}$-conglomerable coherent lower prevision that dominates $\underline{P}$ on $\mathcal{K}$, and therefore also dominates its natural extension $\underline{E}$. Hence $\underline{F}_1 \geq \underline{F}_2$ as well.

Given a coherent lower prevision $\underline{P}$, Walley defines its *conditional natural extension* as

$$\underline{P}(f|B) := \begin{cases} \sup\{\mu\colon \underline{P}(B(f-\mu)) \geq 0\} & \text{if } \underline{P}(B) > 0 \\ \inf_{\omega\in B} f(\omega) & \text{otherwise} \end{cases} \tag{5}$$

for every $f \in \mathcal{L}$ and $B \in \mathcal{B}$. In fact, when $\underline{P}(B) > 0$ then $\underline{P}(f|B)$ is to the unique value of μ such that $\underline{P}(B(f-\mu)) = 0$, i.e., for which (GBR) is satisfied.

From Theorem 6.8.2 in Ref. [11], $\underline{P}$ is $\mathcal{B}$-conglomerable if and only if it is coherent with the conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$ derived from $\underline{P}$ by natural extension. In Ref. [11, Section 6.6], Walley gives a number of examples of coherent lower previsions that are not $\mathcal{B}$-conglomerable. We give a sufficient condition for conglomerability:

Proposition 5. *If the conditional natural extension $\underline{P}(\cdot|\mathcal{B})$ of $\underline{P}$ is given by $\underline{P}(f|B) = \inf_{\omega\in B} f(\omega)$ for all $B \in \mathcal{B}$ and $f \in \mathcal{L}$, then $\underline{P}$ is $\mathcal{B}$-conglomerable, and so is any $\underline{Q} \leq \underline{P}$.*

When $\underline{P}$ is not $\mathcal{B}$-conglomerable, we can consider the natural extensions $\underline{E}$, $\underline{E}(\cdot|\mathcal{B})$ of $\underline{P}$, $\underline{P}(\cdot|\mathcal{B})$, determined by Theorem 8.1.5 in Ref. [11]:

$$\underline{E}(f) := \sup_{g,h\in\mathcal{L}} \sup\left\{\mu\colon f - \mu \geq G_{\underline{P}}(g) + G_{\underline{P}}(h|\mathcal{B})\right\},$$

and it can be checked that $\underline{E}(\cdot|\mathcal{B})$ coincides with the conditional natural extension of $\underline{E}$: it can be obtained using Eq. (5).

Proposition 6. *The natural extension $\underline{E}$ of $\underline{P}$ and $\underline{P}(\cdot|\mathcal{B})$ is dominated by the $\mathcal{B}$-conglomerable natural extension $\underline{F}$ of $\underline{P}$. They coincide if and only if $\underline{E}$ and $\underline{E}(\cdot|\mathcal{B})$ are coherent. Moreover, if we let $\underline{Q} := \underline{P}(\underline{P}(\cdot|\mathcal{B}))$, we have*

$$\begin{aligned}\mathcal{M}(\underline{E}) &= \left\{P \in \mathcal{M}(\underline{P})\colon (\forall f \in \mathcal{L})P(G_{\underline{P}}(f|\mathcal{B})) \geq 0\right\} \\ &= \mathcal{M}(\underline{P}) \cap \mathcal{M}(\underline{Q}).\end{aligned}$$

As a consequence, if $\underline{Q} \geq \underline{P}$, then $\underline{Q}$ coincides with $\underline{E}$ and it is the $\mathcal{B}$-conglomerable natural extension of $\underline{P}$.

Next we show that $\underline{E}$ does not necessarily coincide with the conglomerable natural extension:

Example 5. Consider $\Omega := \mathbb{N} \cup -\mathbb{N}$, $B_n := \{-n, n\}$ and let $\mathcal{B}$ be the partition of Ω given by $\mathcal{B} := \{B_n\colon n \in \mathbb{N}\}$. Let P be a finitely additive probability on $\mathcal{P}(\mathbb{N})$ that satisfies $P(\{n\}) = 0$ for every n (it follows from Ref. [9] that P is not conglomerable), and consider the linear previsions P_1, $\ldots$, P_4, where P_1 is the expectation functional associated with the σ-additive probability measure with

$$P_1(\{n\}) = P_1(\{-n\}) = \frac{1}{2^{n+1}} \text{ for all } n \in \mathbb{N}$$

and P_2, P_3 and P_4 are given, by

$$\begin{aligned} P_2(h) &= \frac{1}{2}\sum_{n=1}^{\infty} h(n)\frac{1}{2^n} + \frac{1}{2}P(h_2) \\ P_3(h) &= \frac{3}{4}P(h_1) + \frac{1}{4}P(h_2) \\ P_4(h) &= \frac{1}{2}P_1(h) + \frac{1}{2}P_3(h), \end{aligned}$$

where for every $h \in \mathcal{L}$ the gambles h_1, h_2 are defined on $\mathbb{N}$ by $h_1(n) := h(n)$ and $h_2(n) := h(-n)$ for every $n \in \mathbb{N}$.

First, we consider the coherent lower prevision $\underline{P} := \min\{P_1, P_2, P_4\}$. Since

$$\underline{P}(B_n) = \min\left\{\frac{1}{2^n}, \frac{1}{2^{n+1}}, \frac{1}{2^{n+1}}\right\} > 0$$

for all $n \in \mathbb{N}$, we see that for every gamble f:

$$\underline{P}(f|B_n) = \min\left\{f(n), \frac{f(n)+f(-n)}{2}\right\}. \tag{6}$$

To see that $\underline{P}$ is not $\mathcal{B}$-conglomerable, consider the gamble f given by

$$f(n) := 1 - \frac{1}{n} \text{ and } f(-n) := -f(n) \text{ for all } n \in \mathbb{N}.$$

It follows from Eq. (6) that $\underline{P}(f|B_n) = 0$ for every n, whence $G_{\underline{P}}(f|\mathcal{B}) = f$. On the other hand,

$$\underline{P}(G_{\underline{P}}(f|\mathcal{B})) \leq P_2(f) = \sum_{n=1}^{\infty}\frac{1}{2^{n+1}}(1-\frac{1}{n}) - \frac{1}{2} < 0,$$

taking into account that $P_2(-\mathbb{N}f) := \frac{1}{2}P(f_2) = -\frac{1}{2}$.

Next we show that $P_4(G_{\underline{P}}(h|\mathcal{B})) \geq 0$ for every gamble h. Note first of all that

$$G_{\underline{P}}(h|\mathcal{B})(n) = \begin{cases} 0 & \text{if } h(n) \leq h(-n) \\ \dfrac{h(n)-h(-n)}{2} & \text{otherwise} \end{cases}$$

$$G_{\underline{P}}(h|\mathcal{B})(-n) = \begin{cases} h(-n)-h(n) & \text{if } h(n) \leq h(-n) \\ \dfrac{h(-n)-h(n)}{2} & \text{otherwise.} \end{cases}$$

As a consequence, $G_{\underline{P}}(h|B_n) \geq 0$ when $h(n) \leq h(-n)$, and this means that $P_4(G_{\underline{P}}(h|\mathcal{B})) \geq P_4(G_{\underline{P}}(h|\mathcal{B})C)$, where $C := \bigcup\{B_n\colon h(n) \geq h(-n)\}$. On the other hand, $G_{\underline{P}}(h|\mathcal{B})(n) = -G_{\underline{P}}(h|\mathcal{B})(-n) \geq 0$ for every $n \in C$, so

$$P_4(G_{\underline{P}}(h|\mathcal{B})C) = 0 + \frac{1}{2}P_3(G_{\underline{P}}(h|\mathcal{B})C)$$

and

$$P_3(G_{\underline{P}}(h|\mathcal{B})C) = \frac{3}{4}P(h') - \frac{1}{4}P(h') \geq 0,$$

where h' is the non-negative gamble on $\mathcal{L}(\mathbb{N})$ given by $h'(n) \coloneqq G_{\underline{P}}(h|\mathcal{B})(n)C(n)$, and where the second term on the right-hand side follows from the definition of P_3.

To determine the natural extension $\underline{E}$ of $\underline{P}$ and $\underline{P}(\cdot|\mathcal{B})$, we apply Proposition 6. First of all, for every linear prevision $Q \in \mathcal{M}(\underline{P})$, there are α_1, α_2 and $\alpha_4 \in [0,1]$ such that $\alpha_1 + \alpha_2 + \alpha_4 = 1$ and $Q = \alpha_1 P_1 + \alpha_2 P_2 + \alpha_4 P_4$. We are going to check which of these combinations satisfies $Q(G_{\underline{P}}(f|\mathcal{B})) \geq 0$ for every gamble f. On the one hand, if $\alpha_2 = 0$ then Q belongs to $\mathcal{M}(\underline{E})$, since we have just proven that P_4 dominates $\underline{E}$ and P_1 is conglomerable. Assume now that $\alpha_2 > 0$, and consider an arbitrary gamble f. As before, since $G_{\underline{P}}(f|\mathcal{B}) \geq G_{\underline{P}}(f|\mathcal{B})C$, where $C \coloneqq \bigcup\{B_n\colon f(n) \geq f(-n)\}$, we can concentrate on gambles f such that $f(n) \geq f(-n)$ for every $n \in \mathbb{N}$. In that case, if we denote $h \coloneqq G_{\underline{P}}(f|\mathcal{B})$, it holds that $h_1 \geq 0$ and $h_2 = -h_1$. As a consequence,

$$\begin{aligned} Q(h) &= \alpha_1 P_1(h) + \alpha_2 P_2(h) + \alpha_4 P_4(h) \\ &= \alpha_2 P_1(h\mathbb{N}) + P(h_1)(\frac{1}{4}\alpha_4 - \frac{1}{2}\alpha_2). \end{aligned}$$

When $\alpha_4 \geq 2\alpha_2 > 0$, we deduce from the non-negativity of $h\mathbb{N}$ (and as a consequence of h_1) that $Q(h) \geq 0$ and therefore $Q \in \mathcal{M}(\underline{E})$. When $\alpha_4 < 2\alpha_2$, there is some natural number n^* such that

$$\frac{1}{2^{n^*}} < \frac{\frac{1}{2}\alpha_2 - \frac{1}{4}\alpha_4}{\alpha_2}.$$

We consider the gamble f given by $f(n) \coloneqq 0$ for $n \leq n^*$, $f(n) \coloneqq 1$ for $n > n^*$ and $f(-n) \coloneqq -f(n)$ for all $n \in \mathbb{N}$. Then $h = G_{\underline{P}}(f|\mathcal{B}) = f$, and using the equation above we obtain $P_1(h\mathbb{N}) = \frac{1}{2^{(n^*+1)}}$ and $P(h_1) = 1$. As a consequence, $Q(h) = \alpha_2 P_1(h\mathbb{N}) + P(h_1)(\frac{1}{4}\alpha_4 - \frac{1}{2}\alpha_2) < 0$, since by construction $P_1(h\mathbb{N}) < \frac{\frac{1}{2}\alpha_2 - \frac{1}{4}\alpha_4}{\alpha_2}$.

We deduce from all this that $\underline{E}$ is the lower envelope of the set $\{P_1, P_4, \frac{1}{3}P_2 + \frac{2}{3}P_4\}$, and as a consequence it induces the conditional lower prevision $\underline{E}(\cdot|\mathcal{B})$ determined by

$$\underline{E}(f|B_n) = \min\left\{\frac{f(n)+f(-n)}{2}, \frac{2f(n)+f(-n)}{3}\right\}. \quad (7)$$

To see that $\underline{E}$ is not $\mathcal{B}$-conglomerable, consider any gamble g such that $g(n) \leq g(-n)$ for all $n \in \mathbb{N}$, then Eq. (7) yields

$$\underline{E}(g|B_n) = \frac{2g(n)+g(-n)}{3},$$

and consequently

$$\begin{aligned} G_{\underline{E}}(g|B_n)(n) &= \frac{g(n)-g(-n)}{3}, \\ G_{\underline{E}}(g|B_n)(-n) &= \frac{2g(-n)-2g(n)}{3}. \end{aligned}$$

Thus, given $h \coloneqq G_{\underline{E}}(g|\mathcal{B})$ we obtain $h_2 = -2h_1 \geq 0$, whence

$$\begin{aligned} P_4(h) &= \frac{1}{2}P_1(h) + \frac{3}{8}P(h_1) + \frac{1}{8}P(h_2) \\ &= \frac{1}{2}(P_1(h\mathbb{N}) + P_1(h-\mathbb{N})) + \frac{1}{8}P(h_1) \\ &= -\frac{1}{2}P_1(h\mathbb{N}) + \frac{1}{8}P(h_1). \end{aligned}$$

Now, if we make for instance $P(h_1) < 4P_1(h\mathbb{N})$, as is the case for $g(n) \coloneqq g(-n) \coloneqq 0$ for $n = 1, 2$ and $g(n) \coloneqq -1$ and $g(-n) \coloneqq 1$ for $n > 2$, then we get $P_4(G_{\underline{E}}(g|\mathcal{B})) < 0$, whence $\underline{E}(G_{\underline{E}}(g|\mathcal{B})) < 0$. Hence, $\underline{E}$ is not $\mathcal{B}$-conglomerable, and therefore it does not coincide with the conglomerable natural extension $\underline{F}$, which exists because $P_1 \geq P$ is $\mathcal{B}$-conglomerable. ♦

On the other hand, we can give a number of sufficient conditions for $\underline{E}$ to be $\mathcal{B}$-conglomerable.

Proposition 7. *If the conditional natural extension derived from $\underline{P}$ is linear and the $\mathcal{B}$-conglomerable natural extension $\underline{F}$ exists, then it coincides with the natural extension $\underline{E}$ of $\underline{P}$ and $P(\cdot|\mathcal{B})$. More generally, if there is a coherent lower prevision $\underline{Q} \geq \underline{P}$ that is coherent with the conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$ derived from $\underline{P}$ using natural extension, then the natural extension $\underline{E}$ of $\underline{P}$ and $\underline{P}(\cdot|\mathcal{B})$ coincides with the $\mathcal{B}$-conglomerable natural extension $\underline{F}$.*

Hence, if $\underline{P}$ is not $\mathcal{B}$-conglomerable, we can consider the natural extension $\underline{E}$ of $\underline{P}$ and $\underline{P}(\cdot|\mathcal{B})$. If then $\underline{E}$ is not $\mathcal{B}$-conglomerable, we can consider the natural extension $\underline{E}_1$ of $\underline{E}$ and $\underline{E}(\cdot|\mathcal{B})$, and so on. Our next result shows that the resulting sequence $\underline{E}_n$ of coherent lower previsions does not stabilise unless we get to a $\mathcal{B}$-conglomerable coherent lower prevision.

Proposition 8. *If $\underline{P}$ is not $\mathcal{B}$-conglomerable, then it does not coincide with the natural extension $\underline{E}$ of $\underline{P}$ and $\underline{P}(\cdot|\mathcal{B})$. On the other hand, if $\underline{E}(\cdot|\mathcal{B}) = \underline{P}(\cdot|\mathcal{B})$ then $\underline{E}$ is $\mathcal{B}$-conglomerable.*

The sequence $\underline{E}_n$ is increasing and therefore converges to a coherent lower prevision $\underline{E}_\infty$, which by construction is dominated by the $\mathcal{B}$-conglomerable natural extension $\underline{F}$ of $\underline{P}$: it suffices to use induction on n and to take into account that at each step n, $\underline{E}_{n+1}$ is a lower bound of any coherent extension of $\underline{E}_n$ and $\underline{E}_n(\cdot|\mathcal{B})$, and is therefore bounded by the $\mathcal{B}$-conglomerable natural extension $\underline{F}$. It is an open problem whether the two coherent lower previsions $\underline{E}_\infty$ and $\underline{F}$ coincide, and also to find an example where $\underline{E}_n$ does not coincide with $\underline{F}_\infty$ for any n, i.e., where we cannot get to the $\mathcal{B}$-conglomerable natural extension in a finite number of steps.

5 Connecting the Two Approaches

The correspondence between sets of desirable gambles and coherent lower previsions we have summarised

in Section 2 does not extend towards the notion of $\mathcal{B}$-conglomerable natural extension we have discussed in Sections 3 and 4. The reason is that in our definition of the $\mathcal{B}$-conglomerable natural extension of a set of gambles we are using condition D5, while the $\mathcal{B}$-conglomerable natural extension for coherent lower previsions is based on condition WC, which is equivalent to WD5, and therefore weaker than D5 in general. We now exhibit all this in more detail.

Let $\mathcal{R}$ be a set of desirable gambles satisfying D1–D4, and let $\underline{P}$ be its associated coherent lower prevision, given by Eq. (3). If $\mathcal{R}$ does not satisfy D5, then we can consider the increasing sequence of sets of desirable gambles $\mathcal{E}_n$, defined by means of Eq. (4). From each of these sets of desirable gambles we can induce a coherent lower prevision $\underline{P}_n$, again by means of Eq. (3). At the same time, we can consider the sequence $\underline{E}_n$ of coherent lower previsions derived from $\underline{P}$ in the manner discussed in Section 4: $\underline{E}_1$ is the natural extension of $\underline{P}$ and $\underline{P}(\cdot|\mathcal{B})$, where $\underline{P}(\cdot|\mathcal{B})$ is the conditional natural extension of $\underline{P}$; $\underline{E}_2$ is the natural extension of $\underline{E}_1$ and $\underline{E}_1(\cdot|\mathcal{B})$; and so on.

Proposition 9. *$\underline{E}_n(f) \leq \underline{P}_n(f)$ for all $f \in \mathcal{L}$.*

However, $\underline{E}_n$ and $\underline{P}_n$ do not coincide in general:

Example 6. Consider the set of desirable gambles $\mathcal{R}$ from Example 2, and let $\underline{P}$ be its associated coherent lower prevision. We have shown in Example 3 that $\mathcal{R}$ satisfies WD5, so Theorem 1 implies that $\underline{P}$ is $\mathcal{B}$-conglomerable, and in particular $\underline{E}_1(f) = \underline{P}(f)$ for every f. On the other hand, we have seen in Example 2 that $\mathcal{R}$ does not satisfy D5, and in particular that the gamble $g = \text{even} - \text{odd}$ belongs to $\mathcal{R}^* \setminus \mathcal{R}$. Moreover, we have seen in Example 4 that $\sup\{\mu\colon g - \mu \in \mathcal{R}\} = -1$. From all this, we infer that

$$\underline{P}_1(g) \geq 0 > -1 = \sup\{\mu\colon g - \mu \in \mathcal{R}\} = P(f) = \underline{E}_1(f).$$

This shows that the inequality in Proposition 9 may be strict. ♦

The reason for this lies in the next result:

Proposition 10. *$\underline{P}_n$ is the natural extension of $\underline{P}_{n-1}$ and $\underline{P}'_{n-1}(\cdot|\mathcal{B})$, where $\underline{P}'_{n-1}(\cdot|\mathcal{B})$ is derived from the set $\mathcal{E}_{n-1}$ by*

$$\underline{P}'_{n-1}(f|B) := \sup\{\mu\colon B(f-\mu) \in \mathcal{E}_{n-1}\} \qquad (8)$$

for all $f \in \mathcal{L}$ and $B \in \mathcal{B}$.

$\underline{P}'_{n-1}(\cdot|\mathcal{B})$ satisfies (GBR) with respect to $\underline{P}_{n-1}$: given a gamble f and a set $B \in \mathcal{B}$, then for all $\varepsilon > 0$,

$$\underline{P}_{n-1}(G_{\underline{P}'_{n-1}}(f|B) + \varepsilon) \geq \underline{P}_{n-1}(B(f - \underline{P}'_{n-1}(f|B) + \varepsilon)) \geq 0,$$

whence $\underline{P}_{n-1}(G_{\underline{P}'_{n-1}}(f|B)) \geq -\varepsilon$ for every $\varepsilon > 0$ and therefore $\underline{P}_{n-1}(G_{\underline{P}'_{n-1}}(f|B)) \geq 0$. Conversely, if there is some $\varepsilon > 0$ such that $\underline{P}_{n-1}(G_{\underline{P}'_{n-1}}(f|B)) \geq \varepsilon$, then the gamble $G_{\underline{P}'_{n-1}}(f|B)) - \frac{\varepsilon}{2}$ must belong to $\mathcal{E}_{n-1}$, and therefore also the gamble $B(f - \underline{P}'_{n-1}(f|B) - \frac{\varepsilon}{2})$, which is greater. But this means that we can increase the value $\underline{P}'_{n-1}(f|B)$ by $\frac{\varepsilon}{2} > 0$, a contradiction with Eq. (8). As a consequence, $\underline{P}'_{n-1}(\cdot|\mathcal{B})$ can strictly dominate the conditional natural extension $\underline{P}_{n-1}(\cdot|\mathcal{B})$ of $\underline{P}_{n-1}$ only when some of the conditioning events have lower probability zero.

From Proposition 9, we can infer the following:

Proposition 11. *Let $\mathcal{R}$ be a coherent set of strictly desirable gambles, and let $\underline{P}$ be its associated coherent lower prevision. Then $\underline{P}_1 = \underline{E}_1$. As a consequence, if $\mathcal{E}_1$ is the $\mathcal{B}$-conglomerable natural extension of $\mathcal{R}$, then $\underline{E}_1$ is the $\mathcal{B}$-conglomerable natural extension of $\underline{P}$.*

Note however that the number of steps necessary to compute the $\mathcal{B}$-conglomerable natural extension can be different in the two cases, as Example 6 shows.

As a consequence of Proposition 11, if $\underline{E}_1$ is not $\mathcal{B}$-conglomerable, then $\mathcal{E}_1$ does not satisfy D5, provided we start from a set of strictly desirable gambles. Using this, we give an example where the sequence of sets $\mathcal{E}_n$ does not stabilise at the first step:

Example 7. Consider the coherent lower prevision $\underline{P}$ from Example 5 and let $\mathcal{R}$ be its associated set of strictly desirable gambles. We have shown in Example 5 that the natural extension $\underline{E}$ of $\underline{P}$ and $\underline{P}(\cdot|\mathcal{B})$ is not $\mathcal{B}$-conglomerable, and therefore it does not coincide with the $\mathcal{B}$-conglomerable natural extension of $\underline{P}$. Applying Proposition 11, we deduce that $\mathcal{E}_1$ cannot be the $\mathcal{B}$-conglomerable natural extension of $\mathcal{R}$, and therefore the sequence $\mathcal{E}_n$ does not stabilise at the first step. ♦

Next we give another sufficient condition for the two sequences of coherent lower previsions to coincide:

Proposition 12. *If $\underline{P}(B) > 0$ for all $B \in \mathcal{B}$, then $\underline{P}_n(f) = \underline{E}_n(f)$ for all $f \in \mathcal{L}$.*

The intuition behind this result is that when the conditioning events have all positive lower probability, then the corresponding conditional lower prevision is uniquely determined by (GBR), and then it necessarily coincides with the natural extension of the unconditional. It implies the following:

Corollary 1. *If $\underline{P}(B) > 0$ for all $B \in \mathcal{B}$ and $\mathcal{E}_n$ is the $\mathcal{B}$-conglomerable natural extension of $\mathcal{R}$, then $\underline{E}_n$ is the $\mathcal{B}$-conglomerable natural extension of $\underline{P}$.*

The condition $\underline{P}(B) > 0$ for every $B \in \mathcal{B}$ does not imply that the sequence stabilises at the first step, as Example 5 shows. On the other hand, the sequences $\underline{E}_n$ and $\mathcal{E}_n$ need not stabilise at the same time: there are examples where $\mathcal{R}$ satisfies WD5, so the associated coherent lower prevision $\underline{P}$ is $\mathcal{B}$-conglomerable, but it does not satisfy D5, so $\mathcal{R}$ is strictly included in $\mathcal{E}_1$.

6 The Case of More Partitions

Next we consider a finite number of sets $\mathcal{R}_1, \ldots, \mathcal{R}_m$, where $\mathcal{R}_i$ satisfies D1–D5 with respect to a partition $\mathcal{B}_i$, and we look for the smallest superset $\mathcal{F}$, if it exists, that satisfies D1–D5 with respect to all partitions in $\mathbb{B} := \{\mathcal{B}_1, \ldots, \mathcal{B}_m\}$.

We first show that conglomerability with respect to the partitions $\mathcal{B}_1, \ldots, \mathcal{B}_m$ is equivalent to conglomerability with respect to all partitions that can be derived from them. Let us define $\mathbb{B}'$ as the (finite) set of partitions $\mathcal{B}$ such that

$$(\forall B \in \mathcal{B})(\exists j \in \{1, \ldots, m\})B \in \mathcal{B}_j.$$

Proposition 13. *Let $\mathcal{R}$ be a set of gambles satisfying* D1–D4*. If it satisfies* D5 *(resp.* WD5*) with respect to all partitions in $\mathbb{B}$, then it also satisfies* D5 *(resp.* WD5*) with respect to all partitions in $\mathbb{B}'$.*

Taking into account Proposition 1, we can show:

Proposition 14. *If there is a set of gambles that includes $\bigcup_{i=1}^m \mathcal{R}_i$ and satisfies* D1–D4 *and* D5 *(resp.,* WD5*) with respect to all partitions in $\mathbb{B}$, then the smallest such set is given by the intersection of all sets that do so.*

On the other hand, if we consider the notion of conglomerability for coherent lower previsions, this time with respect to a finite number of partitions, we can make a link with the property of weak coherence studied in Ref. [11, Section 7.1]:

Proposition 15. *Let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}$. The following statements are equivalent:*

1. *$\underline{P}$ is $\mathcal{B}$-conglomerable for all $\mathcal{B} \in \mathbb{B}$.*
2. *$\underline{P}$ is $\mathcal{B}$-conglomerable for all $\mathcal{B} \in \mathbb{B}'$.*
3. *There are conditional lower previsions $\underline{P}_1(\cdot|\mathcal{B}_1)$, $\ldots$, $\underline{P}_m(\cdot|\mathcal{B}_m)$ that are weakly coherent with $\underline{P}$.*

6.1 The Marginal Extension Theorem

We next prove that when the partitions are nested, the sequence stabilises after one step. This is a version in terms of sets of desirable gambles of the Marginal Extension Theorem 6.7.2 established in Ref. [11] and generalised to any finite number of partitions in Ref. [6]. In a different context, using different notations, this result was also proved (in a different manner) by De Cooman and Hermans [1, Theorem 3]. To proceed, we need to introduce a number of definitions:

Definition 4. Let $\mathcal{B}$ be a partition of Ω. A gamble f on Ω is called *$\mathcal{B}$-measurable* when it is constant on the elements of $\mathcal{B}$. The set of all $\mathcal{B}$-measurable gambles is denoted by $\mathcal{G}(\mathcal{B})$.

Definition 5. Let $\mathcal{Q}$ be a linear subspace of gambles containing all constant gambles, and let $\mathcal{R} \subseteq \mathcal{Q}$. We say that $\mathcal{R}$ is coherent *relative to* $\mathcal{Q}$ if it satisfies D2–D4 and

D1*. if $f \in \mathcal{Q}$ and $f \gtrsim 0$ then $f \in \mathcal{R}$.

When $\mathcal{Q} = \mathcal{L}$, this reduces to the usual coherence notion characterised by axioms D1–D4.

We begin by establishing our result for the case of one partition only.

Proposition 16. *Let $\mathcal{R}_0$ be a set of desirable gambles coherent relative to $\mathcal{G}(\mathcal{B})$. For each $B \in \mathcal{B}$, let $\mathcal{R}|B$ be a coherent set of desirable gambles on $\mathcal{L}(B)$. The $\mathcal{B}$-conglomerable natural extension of $\mathcal{R}_0$ and $\mathcal{R}|B$, $B \in \mathcal{B}$, is the set $\mathcal{F}$ given by*

$$\left\{ f + \sum_{B \in \mathcal{B}} B g_B \colon f \in \mathcal{R}_0 \cup \{0\}, g_B \in \mathcal{R}|B \cup \{0\} \right\} \setminus \{0\}.$$

Proof. Let us show that $\mathcal{F}$ satisfies D1–D5:

D1. Consider $h \gtrsim 0$. Write it as $h = \sum_{B \in \mathcal{B}\colon Bh \neq 0} Bh = \sum_{B \in \mathcal{B}\colon Bh \neq 0} B g_B$, where gamble $g_B \in \mathcal{L}(B)$ is defined by $g_B(\omega) := h(\omega)$ for all $\omega \in B$. Since $g_B \gtrsim 0$, it belongs to the coherent set $\mathcal{R}|B$. Hence h belongs to $\mathcal{F}$.

D2. We know that $0 \notin \mathcal{F}$ by definition.

D3. Consider $h \in \mathcal{F}$ and $\lambda > 0$. We know that $\lambda h = \lambda f + \sum_{B \in \mathcal{B}} B \lambda g_B$. Since $\mathcal{G}(\mathcal{B})$ is a linear space containing all constant gambles, and $\mathcal{R}_0$ is coherent relative to it, it follows that $\lambda f \in \mathcal{R}_0 \cup \{0\}$; moreover, $\lambda g_B \in \mathcal{R}|B \cup \{0\}$, because $\mathcal{R}|B$ is a coherent set. It follows that $\lambda h \in \mathcal{F}$.

D4. Consider $h, h' \in \mathcal{F}$. Then $h + h' = f + f' + \sum_{B \in \mathcal{B}} B(g_B + g'_B)$, where $f, f' \in \mathcal{R}_0 \cup \{0\}$ and $g_B, g'_B \in \mathcal{R}|B \cup \{0\}$. For analogous reasons as in the previous step, it holds that $f + f' \in \mathcal{R}_0 \cup \{0\}$ and $g_B + g'_B \in \mathcal{R}|B \cup \{0\}$. From this, we obtain that $h + h' \in \mathcal{F}$, provided that $h + h' \neq 0$. To see that this is indeed the case, assume that $h + h' = 0$; then either $0 = f + f'$ or $f + f' \neq 0$. In the first case, the coherence of $\mathcal{R}_0$ implies that $f = f' = 0$, and similarly since $g_B + g'_B = 0$ for every B we should have that $g_B = g'_B = 0$ for all B. But then $h = h' = 0$, a contradiction. In the second case, $0 \neq f + f' = -\sum_{B \in \mathcal{B}} B(g_B + g'_B)$. Taking into account that $f + f'$ is $\mathcal{B}$-measurable, there must be some $B \in \mathcal{B}$ such that $B(f + f') \gtrsim 0$: otherwise $f + f' \leq 0$ and $\mathcal{R}_0$ would incur partial loss. But on such a B we obtain that $g_B + g'_B \lesssim 0$, so $\mathcal{R}|B$ would incur partial loss, a contradiction.

D5. Consider $0 \neq h \in \mathcal{L}$ such that $Bh \in \mathcal{F} \cup \{0\}$ for all $B \in \mathcal{B}$. We focus on the case $Bh \neq 0$, where it holds that $Bh = f + \sum_{B \in \mathcal{B}} B g_B$. If $f = 0$, then $Bh = B g_B$. If $f \neq 0$, then consider $B' \in \mathcal{B}$ such that $B' \neq B$. Bh is zero on B', and hence $B' f + B' g_{B'} = 0$. Now, recalling that f is $\mathcal{B}$-measurable, it is only possible that $f < 0$ on B': otherwise, $\mathcal{R}|B'$ would incur partial loss. Since we can repeat this reasoning for all $B' \neq B$, we deduce that $f > 0$ on B, as otherwise $\mathcal{R}_0$ would incur partial loss. In

other words, f is a positive constant, say k_B, on B. Then $g_B + k_B \in \mathcal{R}|B$, so that if we re-define $g_B := g_B + k_B$, we obtain that $Bh = Bg_B$. Thus, $h = \sum_{B\in\mathcal{B}:\ Bh\neq 0} Bh = \sum_{B\in\mathcal{B}:\ Bh\neq 0} Bg_B \in \mathcal{F}$.

Since $\mathcal{F}$ is included in any superset of $\mathcal{R}_0 \cup \mathcal{R}|\mathcal{B}$ satisfying D1–D5, this completes the proof. □

The result also holds for a finite number of partitions.

Proposition 17. *Let $\mathcal{B}_1, \dots, \mathcal{B}_n$ be partitions of Ω such that $\mathcal{B}_{i+1}$ is finer than $\mathcal{B}_i$ for $i = 1, \dots, n-1$. Let $\mathcal{R}_0$ be a set of desirable gambles coherent relative to $\mathcal{G}(\mathcal{B}_1)$. For each $i = 1, \dots, n-1$ and each $B_i \in \mathcal{B}_i$, let $\mathcal{B}_{i+1}|B_i := \{B_{i+1} \in \mathcal{B}_{i+1} : B_{i+1} \subseteq B_i\}$ and $\mathcal{R}_i|B_i$ be a coherent set of desirable gambles on $\mathcal{L}(B_i)$ relative to $\mathcal{G}(\mathcal{B}_{i+1}|B_i)$. Finally, for each $B_n \in \mathcal{B}_n$, let $\mathcal{R}_n|B_n$ be a coherent set of desirable gambles on $\mathcal{L}(B_n)$. The conglomerable natural extension $\mathcal{F}_n$ of $\mathcal{R}_0$ and $\mathcal{R}_i|B_i$, $B_i \in \mathcal{B}_i$, is given by*

$$\Bigg\{ f_0 + \sum_{i=1}^{n} \sum_{B_i\in\mathcal{B}_i} B_i g_{B_i} : f_0 \in \mathcal{R}_0 \cup \{0\}, g_{B_i} \in \mathcal{R}_i|B_i \cup \{0\} \Bigg\} \setminus \{0\}.$$

7 Conclusions

We have studied the extension of desirability and probabilistic assessments under the requirement of conglomerability. Our main finding is that taking the natural extension while imposing conglomerability (which is the procedure adopted in Walley's theory), does not yield the conglomerable natural extension in general (but it does so in the case of the Marginal Extension Theorem); and that although iterating that process yields models ever closer to it, it is an open problem whether or not the conglomerable natural extension is achieved in the limit, or whether the limit is achieved in a finite number of steps. Future work could consist in (i) addressing these problems, and extending everything to the case of multiple partitions; (ii) defining a new coherence notion that follows from the conglomerable natural extension; (iii) investigating the relationship between such an extension and envelope theorems; and (iv) more generally, investigating whether the conglomerable natural extension always allows the most informative conclusions to be drawn.

Acknowledgements

This work was supported by projects TIN2008-06796-C04-01, MTM2010-17844, by the Swiss NSF grants n. 200020_134759 / 1, 200020-121785 / 1, by the Hasler foundation grant n. 10030, and by the SBO project 060043 of the IWT-Vlaanderen.

References

[1] G. De Cooman and F. Hermans. Coherent immediate prediction: bridging two theories of imprecise probability. *Artificial Intelligence*, 172(11):1400–1427, 2008.

[2] B. de Finetti. Sulla proprietà conglomerativa della probabilità subordinate. *Rendiconti del Reale Instituto Lombardo*, 63:414–418, 1930.

[3] B. de Finetti. *Probability, Induction and Statistics*. Wiley, London, 1972.

[4] S. Doria. Coherent upper and lower previsions defined by hausdorff outer and inner measures. In A. Rauh and E. Auer, editors, *Modeling, Design and Simulation Systems with Uncertainties*, pages 175–195. Springer, 2011.

[5] L. E. Dubins. Finitely additive conditional probabilities, conglomerability and disintegrations. *The Annals of Probability*, 3:88–99, 1975.

[6] E. Miranda and G. de Cooman. Marginal extension in the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 46(1):188–225, 2007.

[7] E. Miranda and M. Zaffalon. Notes on desirability and conditional lower previsions. *Annals of Mathematics and Artificial Intelligence*, 2011. In press.

[8] R. Pelessoni and P. Vicig. Williams coherence and beyond. *International Journal of Approximate Reasoning*, 50(4):612–626, 2009.

[9] M. Schervisch, T. Seidenfeld, and J. Kadane. The extent of non-conglomerability of finitely additive probabilities. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, 66:205–226, 1984.

[10] T. Seidenfeld, M. Schervisch, and J. Kadane. Nonconglomerability for finite-valued finitely additive probability. *Sankhya*, 60(3):476–491, 1998.

[11] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[12] P. M. Williams. Notes on conditional previsions. Technical report, School of Mathematical and Physical Science, University of Sussex, UK, 1975. Reprinted in [13].

[13] P. M. Williams. Notes on conditional previsions. *International Journal of Approximate Reasoning*, 44:366–383, 2007. Revised journal version of [12].

Imprecise Probabilities in Non-cooperative Games

Robert Nau
Fuqua School of Business
Duke University
Durham, NC 27708 USA
robert.nau@duke.edu

Abstract

Game-theoretic solution concepts such as Nash equilibrium are commonly used to model strategic behavior in terms of precise probability distributions over outcomes. However, there are many potential sources of imprecision in beliefs about the outcome of a game: incomplete knowledge of payoff functions, non-uniqueness of equilibria, heterogeneity of prior probabilities, unobservable background risk, and distortions of revealed beliefs due to risk aversion, among others. This paper presents a unified approach for dealing with these issues, in which the typical solution of a game is a convex set of probability distributions that, unlike Nash equilibria, may be correlated between players. In the most general case, where players are risk averse, the probabilities do not represent beliefs alone. Rather they must be interpreted as products of subjective probabilities and relative marginal utilities for money.

Keywords: coherence, previsions, lower and upper probabilities, correlated equilibrium, risk neutral probabilities, risk neutral equilibrium

1 Introduction

Game theory occupies the increasingly large middle ground of rational choice theory: the problem of "2, 3, 4… bodies" in which agents must reason about the strategic behavior of other rational agents as well as reflect on their own preferences and compete in markets. The modeling of interactive decisions of this kind requires some special tools and assumptions. First, the rules of the game are (in the most general case) parameterized in units of utility rather than money or goods in order to allow for differences in tastes and attitudes toward risk. Second, the utility functions of different players are assumed to be common knowledge, enabling them to model each other's decisions as well as their own, and to all know that they can all do this, and so on. Third, common knowledge of rationality and common knowledge of the rules of the game are assumed to lead to an equilibrium, usually a Nash equilibrium or one of its refinements or extensions, in which the decision of each player is individually rational given the decisions simultaneously made by the other players, and randomization (if any) is performed independently. And fourth, when there is uncertainty about any of the game parameters, the beliefs of the players are assumed to be consistent with a common prior distribution, which generates an infinite hierarchy of mutually consistent reciprocal beliefs. In applications these assumptions are usually applied at maximum strength in order to tightly (often uniquely) constrain the solution, yet all of them are open to question. This paper will pursue some of these questions and show how they lead to solutions that are characterized by exactly the same rationality conditions as individual decisions and competitive markets. Their common priors and equilibria are generally expressed in terms of imprecise probabilities that need not satisfy an independence condition and do not always represent the players' true subjective beliefs.

The approach to modeling games that will be used in this paper follows that of Nau and McCardle (1990) and Nau (1992), which is just a multi-player extension of de Finetti's operational approach to defining subjective probabilities, which in turn is a microcosm of a financial market. It lends itself naturally to modeling imprecise probabilities; in fact, its behavioral primitives are assertions of lower and upper bounds on probabilities and expectations

2 Imprecise subjective probabilities

Virtually all of the fundamental theorems of rational choice theory—subjective probability, expected utility, subjective expected utility, asset pricing, welfare economics, cardinal utilitarianism, *and* non-cooperative games—are duality theorems that can be proved by using a separating hyperplane argument. In the versions of these theorems that involve finite sets of states and/or consequences, it is a variant of Farkas' lemma, the basis of the duality theorem of linear programming:

LEMMA 1: For any matrix $\boldsymbol{G}$, either there exists a non-negative vector $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha} \bullet \boldsymbol{G} < \mathbf{0}$ or else there exists a non-negative vector $\boldsymbol{\pi}$ such that $\boldsymbol{G\pi} \geq \mathbf{0}$, $\boldsymbol{\pi} \neq \mathbf{0}$.

LEMMA 2: For any matrix $\boldsymbol{G}$, either there exists a non-negative vector $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha} \bullet \boldsymbol{G} \leq \mathbf{0}$ and $[\boldsymbol{\alpha} \bullet \boldsymbol{G}]_k < 0$ or else there exists a non-negative vector $\boldsymbol{\pi}$, with $\pi_k > 0$, such that $\boldsymbol{G\pi} \geq \mathbf{0}$.

De Finetti's (1974) "fundamental theorem of probability," as it applies to imprecise probabilities and expectations, can be proved as follows, using the language of financial markets. Consider an agent ("she") who is uncertain about which element of a finite set $\boldsymbol{S}$ of states of the world will occur. Let N denote the number of states and let $\boldsymbol{x}$ denote an *asset*, which is an N-vector of payoffs assigned to states. The agent's *lower prevision for* $\boldsymbol{x}$ is the price $\underline{P}(\boldsymbol{x})$ that she is publicly willing to pay per unit of $\boldsymbol{x}$ in arbitrary (but small) quantities chosen by someone else. This means that for any small positive number α chosen by an observer ("he"), the agent will accept a bet whose payoff vector for her is $\alpha(\boldsymbol{x} - \underline{P}(\boldsymbol{x}))$, with the opposite payoffs to the observer.[1] For example, if N=3, $\boldsymbol{x}$ = (3, 1, –2), and $\underline{P}(\boldsymbol{x})$ = 1.4, the agent will accept a bet whose payoff vector for her is $(1.6\alpha, -0.4\alpha, -3.4\alpha)$ for any small positive α chosen by the observer. A lower prevision for an asset may be considered as a *lower expectation*, i.e., a lower bound on its subjective expected value for the agent. In the special case where $\boldsymbol{x}$ is a binary vector that is the indicator of an event, its prevision is a *lower probability* for the event.

Lower previsions can also be assessed conditionally. If $\boldsymbol{x}$ is the payoff vector of an asset and $\boldsymbol{e}$ is the indicator vector of an event, the agent's *conditional lower prevision for* $\boldsymbol{x}$ *given* $\boldsymbol{e}$ is the price $\underline{P}(\boldsymbol{x}|\boldsymbol{e})$ that she is publicly willing to pay per unit of $\boldsymbol{x}$ in arbitrary (but small) multiples chosen by an observer, subject to the condition that the bet will be called off if $\boldsymbol{e}$ does not occur. This means that the agent will agree to accept a bet whose payoff vector for her is $\alpha(\boldsymbol{x} - \underline{P}(\boldsymbol{x}|\boldsymbol{e}))\boldsymbol{e}$, for any small positive α. To continue the previous example, if $\boldsymbol{e}$ = (1, 1, 0), i.e., the indicator for the event in which either state 1 or state 2 occurs, and $\underline{P}(\boldsymbol{x}|\boldsymbol{e})$ = 2.1, the agent will accept a bet whose payoff vector for her is $(0.9\alpha, -1.1\alpha, 0)$. In the special case where $\underline{P}(\boldsymbol{x}|\boldsymbol{e}) = 0$, the agent is willing to pay zero for $\boldsymbol{x}$ conditional on $\boldsymbol{e}$, i.e., she will accept a small bet whose payoff vector is proportional to $\boldsymbol{x}$ conditional on the occurrence of $\boldsymbol{e}$. This is equivalent to an unconditional bet with payoffs proportional to $\boldsymbol{xe}$.

It remains to show that rational lower previsions satisfy the laws that ought to be satisfied by lower bounds on probabilities and expectations. Suppose that the agent assigns a conditional lower prevision $\underline{P}(\boldsymbol{x}_m|\boldsymbol{e}_m)$ to asset $\boldsymbol{x}_m$ given the occurrence of event $\boldsymbol{e}_m$, m = 1,…, M, subject to the further requirement that bets on different events are additive, which is the way a bookmaker or financial market normally operates. For example, if the agent simultaneously assigns lower previsions $\underline{P}(\boldsymbol{x}_1|\boldsymbol{e}_1)$ and $\underline{P}(\boldsymbol{x}_2|\boldsymbol{e}_2)$ to asset $\boldsymbol{x}_1$ conditional on event $\boldsymbol{e}_1$ and asset $\boldsymbol{x}_2$ conditional on event $\boldsymbol{e}_2$, this means that for any positive real numbers α_1 and α_2 chosen by the observer, she will accept a bet whose payoff for her in state n is $\alpha_1(x_{1n} - \underline{P}(\boldsymbol{x}_1|\boldsymbol{e}_1))e_{1n} + \alpha_2(x_{2n} - \underline{P}(\boldsymbol{x}_2|\boldsymbol{e}_2))e_{2n}$, where x_{mn} and e_{mn} denote the values of $\boldsymbol{x}_m$ and $\boldsymbol{e}_m$ in state n for m = 1, 2.

The agent is rational *ex ante* if her previsions do not expose her to arbitrage, i.e., if the opponent is not able to make a riskless profit through a clever combination of bets. She is rational *ex post* in state k if they do not allow the opponent to earn a riskless profit if state k occurs. These rationality conditions are called "coherence" and "ex post coherence," respectively. More precisely:

DEFINITION: The conditional lower previsions $\{\underline{P}(\boldsymbol{x}_1|\boldsymbol{e}_1), \ldots, \underline{P}(\boldsymbol{x}_M|\boldsymbol{e}_M)\}$ are *coherent* if there do not exist non-negative numbers $\{\alpha_1, \ldots, \alpha_M\}$ such that $\sum_{m=1}^{M} \alpha_m(x_{mn} - \underline{P}(\boldsymbol{x}_m|\boldsymbol{e}_m))e_{mn} < 0 \ \forall\, n$, i.e., the payoff to the agent is strictly negative in all states. They are *ex post coherent in state* k if and only if there do not exist non-negative numbers $\{\alpha_1, \ldots, \alpha_M\}$ such that $\sum_{m=1}^{M} \alpha_m(x_{mn} - \underline{P}(\boldsymbol{x}_m|\boldsymbol{e}_m)e_{mn} \leq 0 \ \forall\, n$ with strict inequality when $n = k$, i.e., the agent's payoff is surely non-positive and strictly negative in state k.

Coherence entails ex post coherence in at least one state.

THEOREM 1 (de Finetti and others): The conditional lower previsions $\{\underline{P}(\boldsymbol{x}_1|\boldsymbol{e}_1), \ldots, \underline{P}(\boldsymbol{x}_M|\boldsymbol{e}_M)\}$ are coherent [ex post coherent in state k] if and only if there exists a non-empty convex set Π of probability distributions on states of the world [satisfying $\pi_k > 0$] such that, for all m and all $\boldsymbol{\pi} \in \Pi$, $P_{\boldsymbol{\pi}}(\boldsymbol{x}_m|\boldsymbol{e}_m) \geq \underline{P}(\boldsymbol{x}_m|\boldsymbol{e}_m)$ or else $P_{\boldsymbol{\pi}}(\boldsymbol{e}_m) = 0$.

Proof: Let $\boldsymbol{G}$ denote the matrix whose m^{th} row is the vector $(\boldsymbol{x}_m - \underline{P}(\boldsymbol{x}_m|\boldsymbol{e}_m))\boldsymbol{e}_m$ of payoffs to the agent for the

[1] Notational conventions: Lower-case boldface letters such as $\boldsymbol{x}$ and $\boldsymbol{e}$ are used interchangeably for payoff vectors of assets and indicator vectors of events as well as for their proper names (e.g., "event $\boldsymbol{e}$" is the event whose indicator vector is $\boldsymbol{e}$). In the expression $\alpha(\boldsymbol{x} - \underline{P}(\boldsymbol{x}))$, $\boldsymbol{x}$ is a vector and α and $\underline{P}(\boldsymbol{x})$ are scalars, and the multiplication and subtraction are performed pointwise, yielding a vector whose n^{th} element is $\alpha(x_n - \underline{P}(\boldsymbol{x}))$. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors of the same length, then $\boldsymbol{xy}$ denotes their pointwise product (another vector of the same length), and $\boldsymbol{x} \bullet \boldsymbol{y}$ denotes their inner product (a scalar). If $\boldsymbol{G}$ is a matrix and $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors of appropriate length, then $\boldsymbol{x} \bullet \boldsymbol{G}$ and $\boldsymbol{G}\,\boldsymbol{y}$ denote matrix multiplication of $\boldsymbol{G}$ by $\boldsymbol{x}$ on the left or by $\boldsymbol{y}$ on the right, yielding vectors. If $\boldsymbol{\pi}$ is a probability distribution on states and $\boldsymbol{x}$ is a payoff vector and $\boldsymbol{e}$ is an indicator vector for an event, then $P_{\boldsymbol{\pi}}(\boldsymbol{x})$ is the corresponding expected value of $\boldsymbol{x}$ and $P_{\boldsymbol{\pi}}(\boldsymbol{e})$ is the probability of $\boldsymbol{e}$, i.,e. $P_{\boldsymbol{\pi}}(\boldsymbol{x}) = \boldsymbol{\pi} \bullet \boldsymbol{x}$ and $P_{\boldsymbol{\pi}}(\boldsymbol{e}) = \boldsymbol{\pi} \bullet \boldsymbol{e}$. $P_{\boldsymbol{\pi}}(\boldsymbol{x}|\boldsymbol{e})$ denotes the conditional expectation of $\boldsymbol{x}$ given the occurrence of $\boldsymbol{e}$ that is determined by $\boldsymbol{\pi}$, i.e, $P_{\boldsymbol{\pi}}(\boldsymbol{x}|\boldsymbol{e}) = P_{\boldsymbol{\pi}}(\boldsymbol{xe})/P_{\boldsymbol{\pi}}(\boldsymbol{e})$ provided that $P_{\boldsymbol{\pi}}(\boldsymbol{e}) > 0$.

conditional bet determined by the assignment of prevision $\underline{P}(\boldsymbol{x}_m|\boldsymbol{e}_m)$ to asset $\boldsymbol{x}_m$ conditional on event $\boldsymbol{e}_m$. The conditional lower previsions $\{\underline{P}(\boldsymbol{x}_1|\boldsymbol{e}_1),\ldots, \underline{P}(\boldsymbol{x}_M|\boldsymbol{e}_M)\}$ are coherent if and only if there does not a exist non-negative vector $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha} \bullet \boldsymbol{G} < \boldsymbol{0}$. By Lemma 1, this is true if and only there exists a non-negative vector $\boldsymbol{\pi}$ such that $\boldsymbol{G\pi} \geq \boldsymbol{0}$, $\boldsymbol{\pi} \neq \boldsymbol{0}$, which can be normalized so that its elements sum to 1, a probability distribution. The condition $\boldsymbol{G\pi} \geq \boldsymbol{0}$ means $P_{\boldsymbol{\pi}}((\boldsymbol{x}_m - \underline{P}(\boldsymbol{x}_m|\boldsymbol{e}_m))\boldsymbol{e}_m) \geq 0$, or equivalently $P_{\boldsymbol{\pi}}(\boldsymbol{x}_m\boldsymbol{e}_m) \geq \underline{P}(\boldsymbol{x}_m|\boldsymbol{e}_m))P_{\boldsymbol{\pi}}(\boldsymbol{e}_m)$, for all m. This is trivially true if $P_{\boldsymbol{\pi}}(\boldsymbol{e}_m) = 0$, because both sides are zero. If $P_{\boldsymbol{\pi}}(\boldsymbol{e}_m) > 0$, it rearranges to $P_{\boldsymbol{\pi}}(\boldsymbol{x}_m\boldsymbol{e}_m)/P_{\boldsymbol{\pi}}(\boldsymbol{e}_m) \geq \underline{P}(\boldsymbol{x}_m|\boldsymbol{e}_m)$, which by definition means $P_{\boldsymbol{\pi}}(\boldsymbol{x}_m|\boldsymbol{e}_m) \geq \underline{P}(\boldsymbol{x}_m|\boldsymbol{e}_m)$. The corresponding result for ex post coherence in state k follows by applying Lemma 2 in place of Lemma 1. ■

Coherent lower previsions therefore have the properties of lower probabilities and expectations determined by a convex set of probability distributions, which can be interpreted to represent the possibly-imprecise beliefs of the agent, if she has linear utility for money.

An under-appreciated property of de Finetti's operational definition of subjective probabilities and expectations is that it does not merely define them: it makes them common knowledge in the everyday specular sense of the term. The prices are visible to both actors in the scene, and the actors both know it, and both know that they both know it, and so on, and the meaning of the numbers is commonly understood by virtue of the opportunities that they create for reciprocal financial transactions. This is a property of posted prices in general. They do not only simplify the decision-making of consumers and investors: they are also credible and commonly known numerical measurements of the comparative beliefs and values of those who post them.

It might be argued that game-theoretic techniques should be used to address the question of why and how the agent should offer distinct lower and upper previsions (bid and ask prices) in her interaction with the observer, or whether she should offer to bet at all. There might be asymmetric information or incentives for secrecy or deception or speculation that would motivate the agent to set her bid prices for assets at levels other than her true lower bounds on their expected payoffs, whatever "true" might mean. This would merely beg the question of how the rules of the higher-order game would come to be commonly known in numerical terms. If an infinite regress is to be avoided, then at some level of description the amount of private information about her beliefs and values that an agent is willing to publicly reveal is a behavioral primitive. In the sequel, the game-theoretic argument will be turned on its head: the fundamental theorem of non-cooperative games is merely an extension of the fundamental theorem of probability to multiple actors in the same scene.

3 Previsions conditioned on one's own moves

In the assessment of previsions via offers to bet, there is no requirement that states of the world should be events that are beyond the agent's control. However, an observer might be reluctant to take the other side of any bet whose payoff depends on an event that they both know the agent *does* control, and by the same token, the agent might be reluctant to offer to bet on events that she knows to be controlled by others. An important special case is one in which the state space can be partitioned as $\boldsymbol{S} = \boldsymbol{S}_1 \times \boldsymbol{S}_2$, where $\boldsymbol{S}_1$ is a set of events that the agent controls (her own moves) while $\boldsymbol{S}_2$ is a set of events outside her control (moves of nature or other agents). If $\boldsymbol{e}$ is an event that is measurable with respect to $\boldsymbol{S}_1$ (the indicator for a move or subset of moves of the agent), and $\boldsymbol{x}$ is the payoff vector of an asset that is measurable with respect to $\boldsymbol{S}_2$ (a bet whose payoff depends only on moves of others), it might be reasonable for the agent to assert a lower prevision for $\boldsymbol{x}$ conditional on $\boldsymbol{e}$. If she asserts that $\underline{P}(\boldsymbol{x}|\boldsymbol{e}) = 0$, it means that she will accept a small bet whose payoff vector is proportional to $\boldsymbol{x}$ under the same conditions in which she would choose the move $\boldsymbol{e}$, or equivalently, she will accept a small bet whose payoff vector is proportional to $\boldsymbol{xe}$. Such a bet reveals some information about the agent's payoff function in the game she is playing against nature or her adversaries, without necessarily revealing the move she intends to make. Namely, her payoffs in the game are such that her best move is $\boldsymbol{e}$ only under conditions where her prevision for $\boldsymbol{x}$ is non-negative. This method for revealing limited information about one's payoff function yields enough detail about the rules of a non-cooperative game to determine its equilibria, as will be shown next.

4 Imprecise equilibria of games

Let $\mathcal{G}$ denote a non-cooperative game among I players, each having a finite set of strategies. Let $\boldsymbol{S} = \boldsymbol{S}_1 \times\ldots\times \boldsymbol{S}_I$ denote the set of outcomes, where $\boldsymbol{S}_i$ is the set of index numbers for strategies of player i. Let $\boldsymbol{s} = (s_1, \ldots, s_I)$ denote a particular outcome, in which s_i is the strategy chosen by player i. Let $\boldsymbol{x}_i$ denote the payoff function (an $|\boldsymbol{S}|$-dimensional vector) for player i, whose value in outcome $\boldsymbol{s}$ is $x_{\boldsymbol{i}}(\boldsymbol{s})$. Assume that payoffs are measured in units of a common money and that the players are risk neutral. (The risk neutrality assumption will be relaxed later.) The "true" game $\mathcal{G}$ is therefore defined by the sets of strategies $\{\boldsymbol{S}_i\}$ and payoff vectors $\{\boldsymbol{x}_i\}$.

Let $\boldsymbol{e}_{ij}$ denote the event in which player i plays her j^{th} strategy, and for every $j \in \boldsymbol{S}_i$, let $\boldsymbol{x}_{ij}$ denote a vector of payoffs that is obtained from $\boldsymbol{x}_i$ as follows: $\boldsymbol{x}_{ij}(\boldsymbol{s}) = \boldsymbol{x}_i(s_1, \ldots, j, \ldots, s_N)$, where the j occurs in the i^{th} position. In other words, $\boldsymbol{x}_{ij}(\boldsymbol{s})$ is the profile of payoffs that player i receives by playing her j^{th} strategy while all other players play according to $\boldsymbol{s}$. Note that there is some duplication

of information in the structure of $\boldsymbol{x}_{ij}(\boldsymbol{s})$: it contains multiple copies of the payoff profile that player *i* obtains by playing *j*, because the element of $\boldsymbol{x}_{ij}(\boldsymbol{s})$ in coordinate $(s_1, \ldots, s_i, \ldots, s_N)$ is the same for all values of s_i.

Suppose that the payoff functions $\{\boldsymbol{x}_i\}$ are not commonly known *a priori* and must therefore be revealed through some credible language of communication. The language that will be used here is the same one that was sketched in the previous section. To see how it works in the game, observe that in the event that player *i* chooses her j^{th} strategy, she must weakly prefer the profile of payoffs she gets by playing strategy *j* to the profile of payoffs she would have gotten by playing any other strategy *k*. In the terms introduced above, she evidently prefers $\boldsymbol{x}_{ij}$ over $\boldsymbol{x}_{ik}$ in the event that $\boldsymbol{e}_{ij}$ occurs, which means that she would trade $\boldsymbol{x}_{ik}$ for $\boldsymbol{x}_{ij}$ conditional on $\boldsymbol{e}_{ij}$. Such a trade is equivalent to an unconditional bet with a payoff vector of $(\boldsymbol{x}_{ij} - \boldsymbol{x}_{ik})\boldsymbol{e}_{ij}$. If the agent wants to let this information about her payoff function become common knowledge, she can publicly offer to accept a small bet whose payoff vector is proportional to $(\boldsymbol{x}_{ij} - \boldsymbol{x}_{ik})\boldsymbol{e}_{ij}$ at the discretion of an observer. Or, to turn the story around, if by magic her payoff function $\boldsymbol{x}_i$ is already common knowledge, then it is also common knowledge that she will accept such a bet.[2] Note that she is not betting directly on her own strategy. Rather, her own strategy is used as a conditioning event for bets on what other players will do. Bets that are conditioned on the player's own strategy, which may be uncertain to the observer and the other players, do not necessarily reveal her actual state of information or her intended move.

Suppose that all the players offer to accept small conditional bets that are determined by their true payoff functions in the manner described above. Let $\boldsymbol{G}$ denote the matrix whose columns are indexed by outcomes of the game, whose rows are indexed by *ijk*, and whose ijk^{th} row is $(\boldsymbol{x}_{ij} - \boldsymbol{x}_{ik})\boldsymbol{e}_{ij}$, the payoff vector of the bet that is acceptable to player *i* in the event that she chooses strategy *j* in preference to strategy *k*. Then, under the assumption that such bets may be non-negatively linearly combined, an observer of the game may choose a non-negative vector of multipliers $\boldsymbol{\alpha}$ to construct an acceptable bet that yields a total payoff vector of $\boldsymbol{\alpha} \bullet \boldsymbol{G}$ to the players, with the opposite total payoffs to himself.

$\boldsymbol{G}$ will be henceforth called the "revealed rules of the game matrix" because, as will be shown, it contains all the commonly-knowable information about the rules that is actually used in determining the equilibria of non-cooperative games. However, $\boldsymbol{G}$ does not contain all the information about the true game $\mathcal{G}$ that is economically important to the players. In particular, it does not reveal the benefits that a given player might obtain from changes in the strategies of the other players, holding her own strategy fixed. The latter information is subtracted out when the calculation $(\boldsymbol{x}_{ij} - \boldsymbol{x}_{ik})\boldsymbol{e}_{ij}$ is performed. All that remains is information about how a given player would benefit by changing her own strategy, holding the strategies of the *other* players fixed. This is the essence of "non-cooperative" game-playing. The players do not consider the implications of their own play for the payoffs of other players, nor do they expect the other players to show that consideration to them.

Under the assumptions given above, we can define what it means for the game to be played rationally by applying the concept of ex post coherence jointly to all the players. Consider an observer who knows nothing about the game except the bets that the players have offered, which is the minimal information about the game's rules that is common knowledge. Suppose that he does not want to speculate on the game's outcome, but he would like to make a riskless profit if possible. From the observer's perspective, if several bets are placed on the same table at the same time, it doesn't matter if they are offered by one individual or by many who are all looking each other in the eye. If the observer manages to pick their pockets, the players have behaved irrationally as a group.

DEFINITION: The strategy $\boldsymbol{s}$ is *jointly coherent* if there does *not* exist a non-negative $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha} \bullet \boldsymbol{G} \le \boldsymbol{0}$ and $[\boldsymbol{\alpha} \bullet \boldsymbol{G}](\boldsymbol{s}) < 0$, i.e., if, under the revealed rules of the game, there is no system of system of bets under which the observer cannot lose and will win a positive amount from the players if they play $\boldsymbol{s}$.

Fortunately for the players, there is always at least one jointly coherent strategy: they are not doomed to exploitation if they honestly reveal some information about their payoff functions.[3] The interesting question is whether there are strategies that are *not* jointly coherent, and if so, how are they characterized.

In general, the players might choose either pure or randomized strategies, and randomized strategies might be either independent or correlated. Correlated randomization of strategies could be carried out with the help of a mediator but does not necessarily require it: flipping a coin or playing paper-scissors-rock are familiar

[2] Strictly speaking, the *choice* of strategy *j* in the presence of *k* can only be interpreted to mean a *preference* for *j* over *k* if the agent has complete preferences, requiring precise beliefs. Here, offers to bet are assumed to occur at a point in time when the agents may not yet have formed precise beliefs about what their opponents will do, but they expect that they will have done so by the time they are called upon to move. In the meantime they are making assertions about constraints that precise beliefs would have to satisfy in order for them to prefer one strategy over another, thereby partially revealing their payoff functions.

[3] A proof of this result is given in Nau and McCardle (1990). A proof of the dual condition, which (by Theorem 2) is the existence of a correlated equilibrium, is given by Hart and Schmeidler (1989). These proofs are more elementary than the proof of existence of a Nash equilibrium insofar as they do not invoke a fixed-point theorem. In Nau and McCardle's proof, the result follows from the existence of a stationary distribution of a Markov chain.

correlation devices that do not require a mediator, and a taking-turns convention in repeated play could be viewed as a correlation device from the perspective of an observer who doesn't who whose turn it is. Let $\boldsymbol{\pi}$ denote a (possibly-degenerate) probability distribution over the outcomes of the game, and suppose, hypothetically, that the players *do* employ a mediator who is instructed to randomly draw a joint strategy $\boldsymbol{s}$ according to the distribution $\boldsymbol{\pi}$ and then privately recommend to each player that she should play her own part of it. Thus, player i hears only her own recommended strategy, s_i, not those of the other players. Under these conditions, $\boldsymbol{\pi}$ is a common prior distribution over recommended joint strategies in the game, and each player can use Bayesian updating to compute a posterior distribution for the recommendations that were received by the other players, given her own recommendation. If each player's recommended strategy is optimal for her *a posteriori* when the others play their own recommended strategies, then $\boldsymbol{\pi}$ is a correlated equilibrium of the game (Aumann 1974, 1987). More precisely:

DEFINITION: $\boldsymbol{\pi}$ is a *correlated equilibrium* of $\mathcal{G}$ if and only if $\boldsymbol{G\pi} \geq \boldsymbol{0}$, which means that for every player i and every recommended strategy j and alternative strategy k of that player, either $P_{\boldsymbol{\pi}}(\boldsymbol{e}_{ij}) = 0$ (the probability of strategy j being recommended to player i is zero) or else $P_{\boldsymbol{\pi}}(\boldsymbol{x}_{ij}(\boldsymbol{s}) - \boldsymbol{x}_{ik}(\boldsymbol{s})|\boldsymbol{e}_{ij}) \geq 0$ (the conditional expected payoff of strategy j is greater than or equal to the conditional expected payoff of strategy k when j is recommended).

Because the set of all correlated equilibria of $\mathcal{G}$ is determined by a system of linear inequalities, it is a convex polytope—a tractable geometrical object—which will henceforth be denoted by $\Pi_{\mathcal{G}}$. A *Nash equilibrium* is a special case of a correlated equilibrium in which $\boldsymbol{\pi}$ is independent between players, allowing each player to perform her own randomization (if necessary) without a mediator. The set of Nash equilibria is not necessarily convex or connected or bounded by points with rational coordinates, and it can be rather difficult to compute, particularly in games with more than 2 players.

In these terms we can prove a "fundamental theory of non-cooperative games" which is the strategic generalization of the fundamental theorem of probability. Actually, the theorem and its proof are merely a restatement of the fundamental theorem of probability and *its* proof for the special case in which conditional previsions are jointly announced by two or more individuals and the assets and conditioning events to which they refer have a special structure that is determined by a non-cooperative game they are playing.

THEOREM 2 (Nau and McCardle 1990): In a game among risk neutral players, a strategy is jointly coherent if and only if there exists a correlated equilibrium in which it has positive probability.

Proof: By Lemma 2, either there exists a non-negative vector $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha} \bullet \boldsymbol{G} \leq \boldsymbol{0}$ and $[\boldsymbol{\alpha} \bullet \boldsymbol{G}](\boldsymbol{s}) < 0$ or else there exists a non-negative vector $\boldsymbol{\pi}$, with $\pi(\boldsymbol{s}) > 0$, such that $\boldsymbol{G\pi} \geq \boldsymbol{0}$. ■

Hence, the players are rational ex post if and only if they behave as if they had implemented a correlated equilibrium, i.e., if they play a strategy that could have occurred with positive probability in such an equilibrium.[4] But even more can be said: lower and upper bounds can be placed on the players' jointly-held previsions for outcomes of the game and any side bets that might be placed on it, namely the bounds that are determined by the convex polytope $\Pi_{\mathcal{G}}$ of correlated equilibria. On this basis it is appropriate to consider $\Pi_{\mathcal{G}}$ to be the rational "solution" of the game when it is played non-cooperatively in the absence of any constraints other than coherence, and in general it is a solution in terms of imprecise probabilities.[5]

A canonical example of a game in which a non-Nash correlated equilibrium is an attractive strategy is the coordination game known as "battle-of-the-sexes," one version of which has the following payoff matrix:

	Left	Right
Top	2, 1	0, 0
Bottom	0, 0	1, 2

The players would prefer to coordinate on either TL or BR as the solution, but Row has a slight preference for TL and Column has a slight preference for BR. The corresponding rules-of-the-game matrix, $\boldsymbol{G}$, is

	TL	TR	BL	BR
1TB	2	-1	0	0
1BT	0	0	-2	1
2LR	1	0	-2	0
2RL	0	-1	0	2

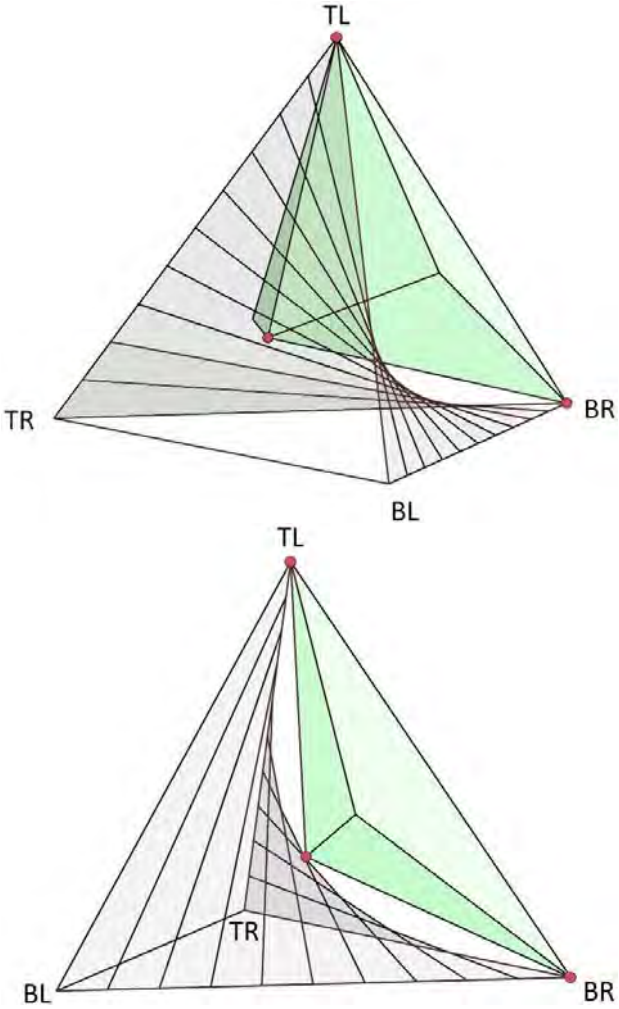
The row label 1TB means G_{1TB}, the payoff vector of the bet for player 1 choosing Top in preference to Bottom, etc. The correlated equilibrium polytope is a hexahedron with 5 vertices, of which 3 are Nash equilibria:

[4] In games of incomplete information, joint coherence leads to a correlated generalization of Bayesian equilibrium (Nau 1992).

[5] This approach can be generalized to the situation in which players do not exactly know their own payoffs. If each payoff in the game matrix is known by its recipient only to lie within some interval, then the ijk^{th} row of $\boldsymbol{G}$ becomes $(\boldsymbol{x}_{ij}^{max} - \boldsymbol{x}_{ik}^{min})\boldsymbol{e}_{ij}$, where $\boldsymbol{x}_{ij}^{max}$ and $\boldsymbol{x}_{ik}^{min}$ are pointwise maxima and minima of the possible payoffs of strategies j and k for player i. This means that in the event that player i chooses strategy j over strategy k, the minimal requirement that her conditional beliefs must satisfy is that her best possible lower prevision for the payoff of j should be at least as great as her worst possible lower prevision for the payoff of k. In general, this sort of payoff-imprecision weakens the constraints and therefore enlarges the set of correlated equilibria.

	TL	TR	BL	BR	Nash?
Vertex 1	1	0	0	0	Yes
Vertex 2	0	0	0	1	Yes
Vertex 3	2/9	4/9	1/9	2/9	Yes
Vertex 4	2/5	0	1/5	2/5	No
Vertex 5	1/4	1/2	0	1/4	No

Two views of the geometry of the correlated equilibrium polytope are shown below The simplex of all probability distributions on outcomes of the game is a tetrahedron, the set of distributions that are independent between players is a saddle, the correlated equilibrium polytope is a hexahedron, and their 3 points of intersection are the Nash equilibria. Nash equilibria always lie on the surface of the correlated equilibrium polytope, but in larger games they need not be vertices of it (Nau et al. 2004).



The mixed-strategy Nash equilibrium is on the inefficient frontier, as is often true of completely mixed strategies in games with multiple equilibria. An obvious and appealing solution of this game that is neither a Nash equilibrium nor an extremal correlated equilibrium is to flip a coin to choose between TL and BR, which is the midpoint of the edge connecting their two vertices.

The players can further restrict the set of rational solutions of the game through the acceptance of additional bets that reflect joint beliefs more precise than the whole set of correlated equilibria. For example, in the battle-of-sexes game, the row player could say "in the event that I play Top [Bottom], I will assign probability 1 (for betting purposes) to the event that my opponent will play Left [Right]," and the column player could similarly say that in the event that she plays Left [Right], she will assign probability 1 to the event that her opponent plays Top [Bottom]. This would indicate that, perhaps through cheap talk or some mechanism such as coin-flipping, the players have coordinated their moves, thereby reducing the set of joint probability distributions to the edge of the simplex that connects TL and BR.

5 Risk aversion & risk neutral probabilities

The results of the previous sections require the players to be risk neutral, i.e., to have state-independent linear utility for money. The more general case of risk averse players will be considered next, and it will be shown that risk aversion leads them to hedge their bets, making the revealed set of equilibria larger than it would have been otherwise. Furthermore, when players are risk averse, side bets may provide opportunities for Pareto-improving modifications of the rules of the game, which leads to some blurring of the distinction between strategic and competitive equilibria. In extreme cases, players may be able to hedge their positions so as to decouple their payoff functions and exit from the game altogether. To set the stage, some general remarks on the modeling of risk aversion are appropriate.

If an agent is risk averse rather than risk neutral, and if she has substantial prior stakes in events ("background risk"), then Theorem 1 still holds, but its parameters have a different interpretation. Suppose that the agent has subjective expected utility preferences and her risk attitude is represented by a strictly concave von Neumann-Morgenstern utility function $U(x)$, with its derivative denoted by $U'(x)$, and suppose that her background risk is represented by a payoff vector $\boldsymbol{z}$ whose elements differ across states by amounts that are large enough to cause substantial variations in the marginal utility of money. Then her acceptance of an additional small bet $\boldsymbol{x}$ will not be based on its expected value but rather on its expected marginal utility in the context of $\boldsymbol{z}$. If the agent's beliefs are represented by a precise probability distribution $\boldsymbol{p}$, then her status quo expected utility is $E_{\boldsymbol{p}}[U(\boldsymbol{z})]$. A bet $\boldsymbol{x}$ will be acceptable to her if it maintains or increases her expected utility, i.e., if $E_{\boldsymbol{p}}[U(\boldsymbol{z}+\boldsymbol{x})] - E_{\boldsymbol{p}}[U(\boldsymbol{z})] \geq 0$.

If the elements of $\boldsymbol{x}$ are small enough in magnitude so that only first-order effects are important, then $\boldsymbol{x}$ is acceptable if $E_{\boldsymbol{p}}[U'(\boldsymbol{z})\boldsymbol{x}] \geq 0$, or equivalently if $E_{\pi}[\boldsymbol{x}] \geq 0$, where $\boldsymbol{\pi}$ is a probability distribution obtained by multiplying the true probability distribution $\boldsymbol{p}$ pointwise by the marginal utility vector $U'(\boldsymbol{z})$ and then re-normalizing, i.e., $\pi(\boldsymbol{s}) \propto p(\boldsymbol{s})U'(z(\boldsymbol{s}))$. This is the *risk neutral probability distribution of the agent* at $\boldsymbol{z}$, because she evaluates small bets in a seemingly risk neutral way using $\boldsymbol{\pi}$ rather than her true subjective probability distribution $\boldsymbol{p}$. The risk neutral distribution of the agent is not uniquely determined by beliefs: it also depends on her background risk and her attitude toward it.[6]

In a financial market, the necessary and sufficient condition for asset prices to create no arbitrage opportunities is that there should exist a probability distribution under which every asset's expected payoff (discounted at the risk-free rate of interest if time is a factor), lies between its bid and ask prices. This result is known as the "fundamental theorem of asset pricing," and it is merely de Finetti's fundamental theorem of probability applied to asset prices offered by the whole market rather than by a single individual. The probability distribution that prices the assets is called the *risk neutral probability distribution of the market*, because it prices them in a seemingly risk neutral way, and it can be determined from prices of options or Arrow securities.[7] Because of friction and incompleteness, the market's risk neutral distribution is usually not unique. Rather, there is a convex set of risk neutral distributions determined by bid and ask prices for assets.

In equilibrium, the marginal prices that agents are willing to pay for financial assets must agree with market prices, which means that the risk neutral probability distributions of all the agents must agree with the risk neutral probability distribution of the market. More precisely, the set of risk neutral distributions that is determined by bid and ask prices in the market is the intersection of all the sets of risk neutral distributions that are determined by bid and ask prices of individual agents, which is non-empty if and only if there are no arbitrage opportunities. Thus, rational behavior in markets requires the agents to "agree" on risk neutral probabilities in the sense that their sets of personal risk neutral probabilities must overlap to some extent. In the special case where the agents have complete preferences and the market is also complete and frictionless, the risk neutral probabilities of the agents and the market are uniquely determined and must be identical.

[6] The role of risk neutral probabilities in modeling a single agent's aversion to risk—and also ambiguity—is discussed in more detail by Nau (2001, 2003, 2011).

[7] The literature on arbitrrage pricing and risk neutral probabilities in finance traces back to the seminal work of Black and Scholes, Merton, Cox, Ross, Rubinstein, and many others in the 1970's, although the connection with de Finetti's use of the no-arbitrage principle in subjective probability, dating to the 1930's, was not noticed until later.

6 Risk neutral equilibria

When agents are risk averse with significant prior stakes in events, their lower and upper previsions determined by offers to accept small bets must be interpreted as lower and upper expectations with respect to convex sets of risk neutral probabilities, rather than true subjective probabilities, as discussed above. The same consideration applies to the analysis of games. A game's own payoffs are a source of background risk with respect to bets on its outcome, and if the players are sufficiently risk averse, this will give rise to distortions when the rules of the game are revealed through betting. The result will be that a rational solution of the game is characterized by a convex set of equilibria whose parameters are risk neutral probabilities.

Suppose that each player has strictly risk averse subjective-expected-utility preferences with respect to profiles of monetary payoffs in the game, and let U_i denote the strictly-concave von Neumann-Morgenstern utility function of player i. Then the payoff profiles $\{x_i(\boldsymbol{s})\}$ translate into utility profiles $\{U_i(x_i(\boldsymbol{s}))\}$. Let $\mathcal{G}^*$ denote the "true" game that is determined by the utility profiles. If U_i' denotes the first derivative of U_i, strict concavity requires that $U_i'(x) < U_i'(y)$ whenever $x > y$. Let $\boldsymbol{u}_i$ denote the utility payoff vector for player i, whose value in outcome $\boldsymbol{s}$ is $U_i(x_i(\boldsymbol{s}))$, and let $\boldsymbol{u}_i'$ denote the corresponding marginal utility vector whose value in outcome $\boldsymbol{s}$ is $U_i'(x_i(\boldsymbol{s}))$. Also, let $\boldsymbol{u}_{ij}$ denote the vector constructed from $\boldsymbol{u}_i$ in the same way that $\boldsymbol{x}_{ij}$ was constructed from $\boldsymbol{x}_i$, namely $u_{ij}(\boldsymbol{s}) = U_i(x_{ij}(\boldsymbol{s}))$. In other words, $u_{ij}(\boldsymbol{s})$ is the utility that player i would receive by playing her j^{th} strategy when all others play according to $\boldsymbol{s}$. Let $\boldsymbol{u}_{ij}'$ denote the corresponding profile of marginal utilities for money, i.e., $u_{ij}'(\boldsymbol{s}) = U_i'(x_{ij}(\boldsymbol{s}))$. As in the case of $\boldsymbol{x}_{ij}$, there is some duplication of information insofar as $u_{ij}(\boldsymbol{s})$ and $u_{ij}'(\boldsymbol{s})$ do not depend on the value of s_i.

By an argument analogous to the one used in the risk neutral case, player i will choose strategy j in preference to strategy k only if her beliefs are such that she would be willing to exchange the utility profile $\boldsymbol{u}_{ik}$, for the utility profile $\boldsymbol{u}_{ij}$, hence a small monetary bet yielding a profile of changes in *marginal* utility that is proportional to $\boldsymbol{u}_{ij} - \boldsymbol{u}_{ik}$ should be acceptable if the event $\boldsymbol{e}_{ij}$ is observed to occur. When strategy j is chosen, the agent's profile of marginal utilities for money is $\boldsymbol{u}_{ij}'$, and a monetary bet that yields a profile of marginal utilities proportional to $\boldsymbol{u}_{ij} - \boldsymbol{u}_{ik}$ can be obtained by dividing the utilities by the corresponding marginal utilities. Thus, agent i should be willing to accept a small bet whose monetary payoffs are proportional to $(\boldsymbol{u}_{ij} - \boldsymbol{u}_{ik})/\boldsymbol{u}_{ij}'$ conditional on the occurrence of $\boldsymbol{e}_{ij}$. Such a bet has an unconditional payoff vector of $((\boldsymbol{u}_{ij} - \boldsymbol{u}_{ik})/\boldsymbol{u}_{ij}')\boldsymbol{e}_{ij}$ in units of money.

Let $\boldsymbol{G}^*$ now denote the matrix whose rows are indexed by *ijk* and whose columns are indexed by $\boldsymbol{s}$ and whose ijk^{th} row is the vector $((\boldsymbol{u}_{ij} - \boldsymbol{u}_{ik})/\boldsymbol{u}_{ij}')\boldsymbol{e}_{ij}$. This is the revealed-rules matrix for the game $\mathcal{G}^*$, representing the information about the game that can be made common knowledge through unilateral offers to accept small bets when the players are risk averse. An observer may choose a small non-negative vector $\boldsymbol{\alpha}$ of multipliers for these bets, and the players as a group will receive the vector of payoffs $\boldsymbol{\alpha} \bullet \boldsymbol{G}^*$, with the opposite payoffs for the observer. The same rationality criterion that was applied in the risk neutral case also applies here in the risk averse case: an outcome $\boldsymbol{s}$ is jointly coherent if and only if there is no non-negative $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha} \bullet \boldsymbol{G}^* \le \boldsymbol{0}$ and $[\boldsymbol{\alpha} \bullet \boldsymbol{G}^*](\boldsymbol{s}) < 0$.[8] The definition of correlated equilibrium and the fundamental theorem of games can now be generalized accordingly. The proof is the same.

DEFINITION: $\boldsymbol{\pi}$ is a *risk neutral equilibrium* of $\mathcal{G}^*$ if and only if $\boldsymbol{G}^*\boldsymbol{\pi} \ge \boldsymbol{0}$, which means that for every player *i* and every strategy *j* and alternative strategy *k* of that player, either $P_{\boldsymbol{\pi}}(\boldsymbol{e}_{ij}) = 0$ or else $P_{\boldsymbol{\pi}}((\boldsymbol{u}_{ij} - \boldsymbol{u}_{ik})/\boldsymbol{u}_{ij}' | \boldsymbol{e}_{ij}) \ge 0$.

THEOREM 3: In a game among risk averse players, a strategy is jointly coherent if and only if there is a risk neutral equilibrium in which it has positive probability.

To provide a story to go with this solution concept, suppose that the players employ a mediator who will use a possibly-correlated randomization device to recommend strategies to them privately, but in this more general case they do not necessarily agree on the true prior probabilities of the outputs of the device. For example, the device may take some of its input data from financial markets or from political or sporting or weather events. Suppose that through side bets with each other or through participation in a public betting market for the input events, they have arrived at a common prior *risk neutral* probability distribution $\boldsymbol{\pi}$ for the outputs of the device. Finally, suppose they will not have the opportunity to directly observe any of the input or output data prior to making their moves *except* for the private recommendations they receive from the mediator, who *will* have observed the data. Under these conditions, for all *i*, *j*, and *k*, the constraint $P_{\boldsymbol{\pi}}((\boldsymbol{u}_{ij} - \boldsymbol{u}_{ik})/\boldsymbol{u}_{ij}' | \boldsymbol{e}_{ij}) \ge 0$ implies $\boldsymbol{p}_{ij} \bullet (\boldsymbol{u}_{ij} - \boldsymbol{u}_{ik}) \ge 0$, i.e., according to player *i*'s own private beliefs, strategy *j* yields an expected utility greater than or equal to that of the alternative strategy *k* when *j* is recommended to her, so it is optimal for each player to follow the mediator if all others do, and this is common knowledge. Thus, a game among risk averse players is played coherently if and only if it is played "as if" with the help of a mediator who uses an incentive-compatible device with respect to whose outputs the players have a common prior risk neutral distribution, although their unobserved true distributions may differ.

A risk neutral equilibrium is a special case of a *subjective correlated equilibrium* (Aumann 1974, 1987), one that can be implemented with the use of a randomizing device about whose properties the players may hold differing beliefs. Such a device would be welcome in playing a zero-sum game—all players might believe their expected payoffs to be positive! Aumann (1987) remarks that such a result depends on "a conceptual inconsistency between the players." By permitting such inconsistencies, subjective correlated equilibrium places only weak restrictions on solutions of many games. A risk neutral equilibrium adds the nontrivial restriction that the players' risk neutral prior probabilities should be mutually consistent, as in an equilibrium of a financial market. When players are risk averse, their true probabilities may be unobservable, and inconsistencies among them are neither surprising nor problematic.

As in the risk neutral case, there is more to be said about the rational solution of the game than to identify the outcomes that are jointly coherent. It is also possible to place bounds on risk neutral probabilities of events or risk neutral expectations of financial assets that depend on the outcome of the game, namely whatever bounds are determined by the system of inequalities $\boldsymbol{G}^*\boldsymbol{\pi} \ge \boldsymbol{0}$ that defines the convex polytope of risk neutral equilibria. These bounds are bid-ask spreads for assets that the players are jointly offering to the observer through their bets that reveal information about the rules of the game.

A simple example of the concept of risk neutral equilibrium is provided by the zero-sum game of "matching pennies," whose payoff matrix is:

	Left	Right
Top	1, –1	–1, 1
Bottom	–1, 1	1, –1

When played by risk neutral players, the revealed-rules matrix $\boldsymbol{G}$, scaled to a maximum value of 1, is:

	TL	TR	BL	BR
1TB	1	-1	0	0
1BT	0	0	-1	1
2LR	1	0	-1	0
2RL	0	-1	0	1

[8] When the utility functions of the players are strictly concave rather than linear, the bet with payoff vector $((\boldsymbol{u}_{ij} - \boldsymbol{u}_{ik})/\boldsymbol{u}_{ij}')\boldsymbol{e}_{ij}$ is technically only "marginally" acceptable to player *i*, so a bet with an aggregate payoff vector of $\boldsymbol{\alpha} \bullet \boldsymbol{G}^*$ may not be quite acceptable to the players for finite $\boldsymbol{\alpha}$. In such a case the observer may need to make a small side payment to the players to get them to agree to the deal, which makes the observer's position not entirely riskless. However, if $\boldsymbol{\alpha} \bullet \boldsymbol{G}^* \le \boldsymbol{0}$ and $[\boldsymbol{\alpha} \bullet \boldsymbol{G}^*](\boldsymbol{s}) < 0$, then by choosing $\boldsymbol{\alpha}$ sufficiently small, the magnitude of the required side payment can be made arbitrarily small in relative terms in comparison to the aggregate loss the players will suffer if they play $\boldsymbol{s}$, which will be considered here as sufficient grounds for not playing $\boldsymbol{s}$. This could be made precise by using the concept of ε-acceptable bets introduced in Nau (1995), but it will not be pursued here in the interest of brevity.

This game has a unique correlated/Nash equilibrium in which the players use independent 50-50 randomization, so the graph of the set of equilibria consists of the single point (¼, ¼, ¼, ¼) in the center of the saddle.

Now suppose that both players are risk averse and, in particular, assume that they both have exponential utility functions, $U(x) = 1 - \exp(-\rho x)$, where the risk aversion parameter is $\rho = \mathrm{LN}(\sqrt{2})$. In units of utility, the payoff matrix of the matching-pennies game is then:

	Left	Right
Top	*a*, *b*	*b*, *a*
Bottom	*b*, *a*	*a*, *b*

where $a = 1 - \sqrt{½} \approx 0.293$ and $b = 1 - \sqrt{2} \approx -0.414$. The corresponding marginal utilities of money under the outcomes *a* and *b* are 0.245 and 0.49, respectively, which conveniently differ by a factor of exactly 2.

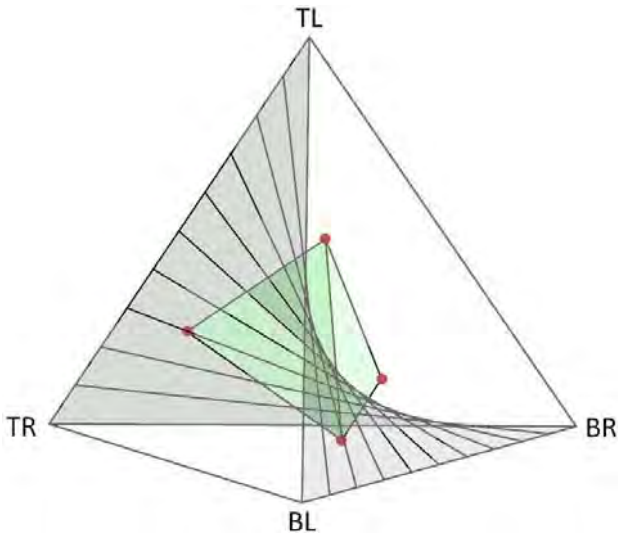
This game is constant-sum and strategically equivalent to the original one, having the same unique correlated/Nash equilibrium. However, the rules matrix of the corresponding revealed game, $\boldsymbol{G}^*$, is *not* equivalent because of the distortions of nonlinear utility for money. It looks like this when scaled to a maximum value of 1:

	TL	TR	BL	BR
1TB	1	–1/2	0	0
1BT	0	0	–1/2	1
2LR	–1/2	0	1	0
2RL	0	1	0	–1/2

The polytope of risk neutral equilibria determined by the inequalities $\boldsymbol{G}^*\boldsymbol{\pi} \geq \boldsymbol{0}$ is a tetrahedron with these vertices:

	TL	TR	BL	BR	EV>0?
Vertex 1	2/15	4/15	1/15	8/15	1BT
Vertex 2	8/15	1/15	4/15	2/15	1TB
Vertex 3	4/15	8/15	2/15	1/15	2RL
Vertex 4	1/15	2/15	8/15	4/15	2LR

None of them lies on the saddle of distributions that are independent between {T,L} and {B,R}, so none is a Nash equilibrium of a game with these strategy sets.[9] Each of these probability distributions satisfies 3 out of the 4 incentive constraints with equality, i.e., assigns an expected value of zero to 3 out of the 4 rows of $\boldsymbol{G}^*$. (The label of the row whose expected value is positive is shown in the rightmost column.). A graph of their configuration is shown below. The polytope of risk neutral equilibria is suspended in the middle of the probability simplex, and the saddle of independent distributions cuts through its interior, a situation that would be impossible for a set of correlated equilibria.



The uniform distribution that is the unique equilibrium of the game when the true utility functions of the players are common knowledge lies in the interior of the polytope of risk neutral equilibria. When players are risk averse, the small side bets they are willing to accept do not fully reveal the between-strategy differences in utility profiles that they face in the game, so the set of risk neutral equilibria is larger than the set of correlated equilibria. This is true in general, as summarized by:

THEOREM 4: The set of correlated equilibria of a game with monetary payoffs played by risk neutral players is a subset of the set of risk neutral equilibria of the same game played by risk averse players.

Proof: If player *i* is risk neutral, she will accept a bet with payoff vector $(\boldsymbol{x}_{ij} - \boldsymbol{x}_{ik})\boldsymbol{e}_{ij}$, while if she is risk averse, she will accept a bet with payoff vector $((\boldsymbol{u}_{ij} - \boldsymbol{u}_{ik})/\boldsymbol{u}_{ij}')\boldsymbol{e}_{ij}$, where $u_{ij}(\boldsymbol{s}) = U_i(x_{ij}(\boldsymbol{s}))$, and $u_{ij}'(\boldsymbol{s}) = U_i'(x_{ij}(\boldsymbol{s}))$. The term $\boldsymbol{e}_{ij}$ will be ignored henceforth because it zeroes-out the same elements of both vectors. By the subgradient inequality, $U(z) < U(y) - U'(y)(y - z)$, because the value of a strictly concave function *U* at *z* must lie below the tangent to its graph at any other point *y*. Letting $y = x_{ij}(\boldsymbol{s})$ and $z = x_{ik}(\boldsymbol{s})$ yields $u_{ik}(\boldsymbol{s}) \leq u_{ij}(\boldsymbol{s}) - u_{ij}'(\boldsymbol{s})\,(x_{ij}(\boldsymbol{s}) - x_{ik}(\boldsymbol{s}))$, which rearranges to $(u_{ij}(\boldsymbol{s}) - u_{ik}(\boldsymbol{s}))/u_{ij}'(\boldsymbol{s}) \geq x_{ij}(\boldsymbol{s}) - x_{ik}(\boldsymbol{s})$, with strict inequality if $x_{ij}(\boldsymbol{s}) \neq x_{ik}(\boldsymbol{s})$. Hence, the bet that player *i* is willing to accept when she chooses strategy *j* in preference to *k* if she is risk neutral is weakly dominated by the bet she will accept in the same game if she is risk averse. This means $\boldsymbol{G}^* \geq \boldsymbol{G}$ pointwise, from which it follows that $\boldsymbol{G}\boldsymbol{\pi} \geq \boldsymbol{0}$ implies $\boldsymbol{G}^*\boldsymbol{\pi} \geq \boldsymbol{0}$, so if $\boldsymbol{\pi}$ is

[9] These distributions are the unique Nash equilibria of the game:

	L*	R*
T*	2, -1	-1, 1
B*	-2, 4	1, -4

under different mappings of {TL, TR, BL, BR} to {T*L*, T*R*, B*L*, B*R*}. They lie on the two other saddles that can be drawn within the original simplex: the one that omits the edges BL-BR and TL-TR and the one that omits the edges TL-BL and TR-BR

a correlated equilibrium of the game played by risk neutral players, then it is a risk neutral equilibrium of the same game when it is played by risk averse players. ■

Hence, risk aversion introduces even more imprecision into the probabilistic solutions of non-cooperative games when their rules must be revealed through credible bets.

7 Rewriting the rules of the game

It was pointed out earlier, in the discussion of the battle-of-sexes game, that players could accept additional bets with an observer, beyond those that determine the rules of the game, in order to reveal more precise information about their joint beliefs. However, if they are risk neutral and have in fact implemented a Nash or correlated equilibrium, which induces a common prior distribution over outcomes of the game, they cannot both be made strictly better off through bets with each other. When players are risk averse, this is not necessarily true, and the matching-pennies game provides a good example. When played by risk averse players, it is a negative-sum game in units of utility, and for both players the unique Nash equilibrium (coin-flipping) has an expected utility that is below their status quo utility. Risk averse players would rather not play this game at all. Furthermore, player 1's marginal utility of money is greater in outcomes TR and BL (her losing outcomes) than in the other two, and vice versa for player 2. The Nash equilibrium is therefore not a *competitive* equilibrium of a financial market in which it is possible for the players to make additional bets that reveal their *solution* of the game in addition to the bets that reveal the *rules* of the game (the latter being the rows of G^*). In the context of the Nash equilibrium, it is desirable to both players to make a bet in which player 1 wins \$$x$ if TR or BL occurs and player 2 wins \$$x$ if TL or BR occurs, for any positive $x \leq 1$. Such a bet changes the rules of the game to a finite extent, but coin-flipping remains a Nash equilibrium. By choosing $x = 1$ they can even zero-out their payoffs, dissolving the game altogether. If they do not bet with each other in this fashion, but instead bet separately with an observer, there is an arbitrage opportunity for the observer that arises from the fact that, at the outset, the players' risk neutral probabilities do not agree if their true probability distributions are uniform.

8 Conclusions

The concept of coherent lower and upper previsions extends in a natural way to non-cooperative game theory, where it can be applied to the process of revealing the rules of the game as well as expressing the beliefs of the players. A rational solution of the game, from the perspective of an observer, is typically a convex set of correlated equilibria rather than a Nash equilibrium. The presence of aversion to risk changes the units of analysis from "true" subjective probabilities to "risk neutral" probabilities, as in asset pricing theory, and it typically renders the solutions even more imprecise. When risk averse players make bets with each other that reflect their beliefs about the solution of the game as well as the rules from which they started, they may be able to rewrite those rules in a mutually beneficial way, merging the concepts of strategic and competitive equilibrium

These results address some of the issues raised by Kadane and Larkey (1982) concerning the relation between game theory and subjective probability theory. The theory of game-playing presented here is a direct extension of subjective probability theory à la de Finetti, and it exploits the underappreciated common-knowledge property of de Finetti's use of bets to measure beliefs. Common knowledge of a game's rules constrains rational beliefs but in general it does not uniquely determine them, leaving room for subjective differences, particularly when players are risk averse and/or have incomplete knowledge of their own payoff functions.

References

[1] Aumann, R.J. (1974) Subjectivity and Correlation in Randomized Games. *Econometrica* **30**, 445-462

[2] Aumann, R.J. (1987) Correlated Equilibrium as an Expression of Bayesian Rationality. *Econometrica* **55**, 1-18

[3] de Finetti, B. (1974), *Theory of Probability, Vol. 1*. Wiley, New York.

[4] Hart, S. and D. Schmeidler (1989) Existence of Correlated Equilibria. *Mathematics of Operations Research* **14**, 18-25

[5] Kadane, J. and P. Larkey (1982) Subjective Probability and the Theory of Games, *Management Science* **28**, 113-120

[6] Nau, R.F. and K. F. McCardle (1990) Coherent Behavior in Non-cooperative Games. *J. Economic Theory* **50**, 424-444

[7] Nau, R.F. (1992) Joint Coherence in Games of Incomplete Information. *Management Science* **38**, 374-387

[8] Nau, R.F. (1995) Coherent Decision Analysis with Inseparable Probabilities and Utilities. *Journal of Risk and Uncertainty* **10**, 71-91

[9] Nau. R. F. (2001) De Finetti Was Right: Probability Does Not Exist. *Theory and Decision* **51**, 89-124

[10] Nau, R.F. (2003) A Generalization Of Pratt-Arrow Measure To Non-Expected-Utility Preferences And Inseparable Probability And Utility. *Management Science* **49**, 1089-1104

[11] Nau, R. F., S. Gomez Canovas and P. Hansen (2004) On the Geometry of Nash Equilibria and Correlated Equilibria. *International Journal of Game Theory* **32**, 443-453

[12] Nau, R.F. (2011) Risk, Ambiguity, and State-Preference Theory. *Economic Theory*, forthcoming

Characterizing joint distributions of random sets with an application to set-valued stochastic processes

Bernhard Schmelzer
Unit of Engineering Mathematics,
University of Innsbruck, Austria
bernhard.schmelzer@uibk.ac.at

Abstract

By the Choquet theorem, distributions of random closed sets can be characterized by a certain class of set functions called capacity functionals. In this paper a generalization to the multivariate case is presented, that is, it is proved that the joint distribution of finitely many random sets can be characterized by a set function fulfilling certain properties. Furthermore, we use this result to formulate an existence theorem for set-valued stochastic processes.

Keywords. Random set, Choquet theorem, capacity functional, joint distribution, Daniell-Kolmogorov theorem.

1 Introduction

Random sets, or set-valued maps, can be used to model uncertainty. They can be interpreted as imprecise observations of random variables ([10]) which assign to each element of the underlying probability space a set instead of a single value. These sets (called focal sets) are supposed to contain the true value of the variable.

We will consider random closed sets, that is, random maps whose values are closed subsets of a topological space $\mathbb{E}$, since they have favorable properties. The family of all closed subsets of $\mathbb{E}$ will be denoted by $\mathcal{F}$ which can in turn be topologized by the so-called Fell topology ([1]). Random closed sets can then be seen as random elements with values in $\mathcal{F}$ and classical probability theory can be applied. As already mentioned, they can also be interpreted as imprecise observations of random variables ([10]). In this case, one is more interested in events from the Borel-σ-algebra $\mathcal{B}(\mathbb{E})$, than from $\mathcal{B}(\mathcal{F})$ and non-additive set functions (so-called lower and upper probabilities, see [4]) are introduced to measure if the focal elements hit or miss a certain set from $\mathcal{B}(\mathbb{E})$. The link between these two interpretations is given by the so-called Choquet theorem (also referred to as the Choquet-Matheron-Kendall theorem, see [13, 15, 17]), which states a one-to-one correspondence between probability distributions on $\mathcal{B}(\mathcal{F})$ and a certain class of non-additive set functions, called capacity functionals, on $\mathcal{B}(\mathbb{E})$.

The goal of this paper is to present characterizations of the joint distribution of finitely many random sets. More precisely, given n random sets we will link their joint distribution defined on the product-σ-algebra $\mathcal{B}(\mathcal{F})^{\otimes n}$ to set functions defined on the compacts of the co-product $\mathbb{E} \times \{1, \dots, n\}$ or a certain class of subsets of $\mathbb{E}^n$.

The plan of the paper is as follows. In Section 2 we review the most important facts on random sets and their distributions including the classical Choquet theorem. The main part of the paper is Section 3 where joint distributions of random sets are considered and characterized by multivariate capacities. In Section 4 the latter is used to formulate a Daniell-Kolmogorov existence theorem ([5, 7]) for set-valued stochastic processes. Furthermore, we consider Brownian motion as an example.

2 Random closed sets and Choquet theorem

In this section we review the most important facts about random closed sets. As already mentioned in the introduction we consider maps whose values are closed subsets of some topological space $\mathbb{E}$. Throughout the paper, $\mathcal{G}$, $\mathcal{F}$, $\mathcal{K}$ will denote the families of open, closed, compact subsets of $\mathbb{E}$, respectively. Furthermore, we will use the following notation

$$\mathcal{F}_A = \{F \in \mathcal{F} : F \cap A \neq \emptyset\}$$
$$\mathcal{F}^A = \{F \in \mathcal{F} : F \cap A = \emptyset\}$$
$$\mathcal{F}^A_{A_1,\dots,A_k} = \mathcal{F}^A \cap \mathcal{F}_{A_1} \cap \dots \cap \mathcal{F}_{A_k}$$

for arbitrary subsets $A, A_1, \ldots, A_k$ of $\mathbb{E}$. The family $\mathcal{F}$ is endowed with the Fell topology ([1]). Recall that the latter has as a sub-base $\{\mathcal{F}_G\}_{G\in\mathcal{G}} \cup \{\mathcal{F}^K\}_{K\in\mathcal{K}}$, that is, sets of the form $\mathcal{F}^K_{G_1,\ldots,G_k}$ $(K \in \mathcal{K},\ G_i \in \mathcal{G})$ constitute a base. We shall always assume that $\mathbb{E}$ is a locally compact Hausdorff second countable (LCHS) space. In this case, $\mathcal{F}$ together with the Fell topology becomes a compact Hausdorff second countable space ([1]). In addition, we introduce on $\mathcal{F}$ the so-called Effros-σ-algebra $\mathcal{B}(\mathcal{F})$ which is generated by the sets $\{\mathcal{F}_G\}_{G\in\mathcal{G}}$. By virtue of the LCHS property of $\mathbb{E}$, the Effros-σ-algebra is also generated by $\{\mathcal{F}_K\}_{K\in\mathcal{K}}$ and is the Borel-σ-algebra with respect to the Fell topology. For details and further information about topologies on $\mathcal{F}$ the reader is referred to the monograph [1].

A map $X : \Omega \to \mathcal{F}$ on a probability space (Ω, Σ, P) will be called Effros-measurable if

$$X^-(G) = \{\omega : X(\omega) \cap G \neq \emptyset\} = X^{-1}(\mathcal{F}_G) \in \Sigma$$

for all $G \in \mathcal{G}$ whereas X will be called random (closed) set if it is strongly measurable ([18]), i.e., $X^-(B) \in \Sigma$ for all $B \in \mathcal{B}(\mathbb{E})$. Note that in general the two conditions are not equivalent unless (Ω, Σ, P) is complete (see [2, 8]). The distribution of an Effros-measurable map X is then the image measure P_X of P on $\mathcal{B}(\mathcal{F})$. For the generating sets $\mathcal{F}_K$ $(K \in \mathcal{K})$ of $\mathcal{B}(\mathcal{F})$ the probabilities $P_X(\mathcal{F}_K) = P(X^-(K))$ can be expressed by a set function $\varphi : \mathcal{K} \to [0,1], K \mapsto P_X(\mathcal{F}_K)$. This set function corresponds to the upper probability of a random set introduced by Dempster and Shafer ([4]) and has (among others) the following properties:

(CF1) $0 \le \varphi \le 1$ and $\varphi(\emptyset) = 0$,

(CF2) For $K, K_1, \ldots, K_n \in \mathcal{K}$, $n \ge 0$, the probabilities $P_X(\mathcal{F}^K_{K_1,\ldots,K_n})$ can be written in terms of φ as

$$P_X(\mathcal{F}^K_{K_1,\ldots,K_n}) = \Delta_n\varphi(K; K_1, \ldots, K_n)$$

where $\Delta_0\varphi(K) = 1 - \varphi(K)$ and for $n \ge 1$

$$\begin{aligned}\Delta_n\varphi(K; K_1, \ldots, K_n) &= \Delta_{n-1}\varphi(K; K_1, \ldots, K_{n-1}) \\ &\quad - \Delta_{n-1}\varphi(K \cup K_n; K_1, \ldots, K_{n-1}).\end{aligned}$$

Thus, $\Delta_n\varphi \ge 0$ for $n \ge 0$.

(CF3) φ is continuous from above, that is, for a decreasing sequence $\{K_n\}_{n\in\mathbb{N}}$ with limit $K = \bigcap_{n\in\mathbb{N}} K_n$ it holds that $\varphi(K_n) \searrow \varphi(K)$.

Note that a set function fulfilling Condition (CF2) is called completely alternating. Furthermore, for $n \ge 1$ the successive differences can be expressed as follows:

$$\Delta_n\varphi(K; K_1, \ldots, K_n) = -\sum_{I\subseteq\{1,\ldots,n\}} (-1)^{|I|}\,\varphi\big(K \cup \bigcup_{i\in I} K_i\big) \quad (1)$$

where the union over $\emptyset$ is set to $\emptyset$. A set function on $\mathcal{K}$ fulfilling these three properties is called capacity functional. The following theorem known as the Choquet theorem (see [13, 15, 17]) says that there is a one-to-one correspondence between capacity functionals and probability measures on $\mathcal{B}(\mathcal{F})$.

Theorem 1. Let $\mathbb{E}$ be an LCHS space and let $\varphi : \mathcal{K} \to [0,1]$ be a capacity functional. Then there exists a unique probability measure Π on $\mathcal{B}(\mathcal{F})$ such that $\varphi(K) = \Pi(\mathcal{F}_K)$ for all $K \in \mathcal{K}$.

For later reference we give a sketch of the proof ([13]): First, note that a capacity functional φ can be extended to the power set $\mathcal{P}$ of $\mathbb{E}$ by setting

$$\begin{aligned}\varphi_*(G) &= \sup\{\varphi(K) : K \subseteq G, K \in \mathcal{K}\} \text{ if } G \in \mathcal{G}, \\ \varphi^*(A) &= \inf\{\varphi_*(G) : G \supseteq A, G \in \mathcal{G}\} \text{ if } A \in \mathcal{P}.\end{aligned} \quad (2)$$

The extension φ^* is a completely alternating Choquet-$\mathcal{K}$-capacity, that is, φ^* is continuous from above on $\mathcal{K}$ and continuous from below on $\mathcal{P}$ ([3, 14]). Furthermore, the extension is consistent, i.e., on $\mathcal{K}$ the extension yields the same results as if φ is directly applied. To obtain the desired probability measure on $\mathcal{B}(\mathcal{F})$ the set function φ^* is considered on $\mathcal{V} = \{G \cup K : G \in \mathcal{G}, K \in \mathcal{K}\}$ and a set function Π is defined on $\mathcal{H} = \{\mathcal{F}^V_{V_1,\ldots,V_k} : V, V_j \in \mathcal{V}, k \ge 0, 1 \le j \le k\}$ by $\Pi(\mathcal{F}^V_{V_1,\ldots,V_k}) = \Delta_k\varphi^*(V; V_1, \ldots, V_k)$. Π is proved to be (finitely) additive and extended to a measure on $\mathcal{B}(\mathcal{F})$ (which is generated by $\mathcal{H}$) by using [16, Prop. I.6.2] and continuity properties of φ^*. Moreover, one can show (cf. [6], Appendix, 2, Satz 2) that for all $B \in \mathcal{B}(\mathbb{E})$ it holds that $\mathcal{F}_B \in \mathcal{B}(\mathcal{F})^0$ and $\varphi^*(B) = \Pi^0(\mathcal{F}_B)$ where $(\mathcal{F}, \mathcal{B}(\mathcal{F})^0, \Pi^0)$ denotes the completed probability space with respect to Π.

3 The multivariate case

Let $n \ge 2$ and $\mathbb{E}_i$ be LCHS spaces with $\mathcal{G}_i$, $\mathcal{F}_i$, $\mathcal{K}_i$ denoting the families of open, closed, compact subsets of $\mathbb{E}_i$, respectively, $1 \le i \le n$. As already outlined in the introduction the goal is to characterize probability measures on the Borel sets of

$$\mathcal{F}^n = \mathcal{F}_1 \times \cdots \times \mathcal{F}_n = \{(F_1, \ldots, F_n) : F_i \in \mathcal{F}_i\}$$

by set functions. $\mathcal{F}^n$ will be endowed with the product Fell topology which is generated by the cylindrical sets

$$\mathcal{F}^{K_1}_{G_{11},\ldots,G_{1k_1}} \times \cdots \times \mathcal{F}^{K_n}_{G_{n1},\ldots,G_{nk_n}}$$

where $G_{ij_i} \in \mathcal{G}_i$, $K_i \in \mathcal{K}_i$. From the one-dimensional case one can infer that the product-Effros-σ-algebra $\mathcal{B}(\mathcal{F}^n) = \mathcal{B}(\mathcal{F}_i)^{\otimes n} = \mathcal{B}(\mathcal{F}_1) \otimes \cdots \otimes \mathcal{B}(\mathcal{F}_n)$ is generated by the sets

$$\mathcal{F}_{K_1} \times \cdots \times \mathcal{F}_{K_n}$$

where $K_i \in \mathcal{K}_i$. For n Effros-measurable maps (random sets) $X_i : \Omega \to \mathcal{F}_i$ on a probability space (Ω, Σ, P) their joint distribution is then given by

$$\begin{aligned} &P_{X_1,\dots,X_n}(\mathcal{F}_{K_1} \times \cdots \times \mathcal{F}_{K_n}) \\ &= P(\{\omega : (X_1(\omega), \dots, X_n(\omega)) \in \mathcal{F}_{K_1} \times \cdots \times \mathcal{F}_{K_n}\}) \\ &= P(\{\omega : X_1(\omega) \cap K_1 \neq \emptyset, \dots, X_n(\omega) \cap K_n \neq \emptyset\}) \\ &= P\Big(\bigcap_{i=1}^n X_i^-(K_i)\Big). \end{aligned} \quad (3)$$

The latter can be expressed by using $K_1 \times \cdots \times K_n$ which is a subset of $\mathbb{E}^n = \mathbb{E}_1 \times \cdots \times \mathbb{E}_n$:

$$\begin{aligned} &P_{X_1,\dots,X_n}(\mathcal{F}_{K_1} \times \cdots \times \mathcal{F}_{K_n}) \\ &= P(\{\omega : X_1(\omega) \times \cdots \times X_n(\omega) \cap K_1 \times \cdots \times K_n \neq \emptyset\}) \end{aligned} \quad (4)$$

Motivated by this, we use the following notation for arbitrary $V, V_1, \dots, V_k \subseteq \mathbb{E}^n$

$$\begin{aligned} {}^n\mathcal{F}_V &= \{(F_1, \dots, F_n) \in \mathcal{F}^n : F_1 \times \cdots \times F_n \cap V \neq \emptyset\} \\ {}^n\mathcal{F}^V &= \{(F_1, \dots, F_n) \in \mathcal{F}^n : F_1 \times \cdots \times F_n \cap V = \emptyset\} \\ {}^n\mathcal{F}^V_{V_1,\dots,V_k} &= {}^n\mathcal{F}^V \cap {}^n\mathcal{F}_{V_1} \cap \cdots \cap {}^n\mathcal{F}_{V_k} \end{aligned}$$

which implies $\mathcal{F}_{K_1} \times \cdots \times \mathcal{F}_{K_n} = {}^n\mathcal{F}_{K_1 \times \cdots \times K_n}$. The event $(X_1, \dots, X_n)^{-1}({}^n\mathcal{F}_V)$ corresponds to the event that the set-valued map

$$X : \omega \mapsto X_1(\omega) \times \cdots \times X_n(\omega) \quad (5)$$

hits V. Note that the values of X are closed subsets of $\mathbb{E}^n$, more precisely closed cylindrical sets, and not elements of $\mathcal{F}^n$. One can prove ([19]) that X is Effros-measurable by using selections and the so-called Fundamental measurability theorem for multifunctions ([2, 8]). Consequently, the map

$$K \mapsto P(X^-(K))$$

is a capacity functional on the compact subsets of $\mathbb{E}^n$ denoted by $\mathcal{K}(\mathbb{E}^n)$. One could thus think of characterizing joint distributions of n random sets by capacity functionals on $\mathcal{K}(\mathbb{E}^n)$. But applying the Choquet theorem leads to a probability measure on the Borel sets of $\mathcal{F}(\mathbb{E}^n)$ denoting the family of closed subsets of $\mathbb{E}^n$. The latter is clearly different from $\mathcal{F}^n$ which can only be identified with the cylindrical closed subsets of $\mathbb{E}^n$, that is, $\{F_1 \times \cdots \times F_n : F_i \in \mathcal{F}_i\}$ which is a proper subset of $\mathcal{F}(\mathbb{E}^n)$.

Hence, there is the need for a different concept. In the following, we will consider the co-product of the spaces $\mathbb{E}_i$, that is,

$$\mathbb{E}^n_{\amalg} = \bigcup_{i=1}^n \mathbb{E}_i \times \{i\}$$

which is a union of n mutually disjoint sets. We endow $\mathbb{E}^n_{\amalg}$ with the sum topology, that is, we take

$$\mathcal{G}^n_{\amalg} = \bigcap_{i=1}^n \{G \subseteq \mathbb{E}^n_{\amalg} : \iota_i^{-1}(G) \in \mathcal{G}_i\}$$

as the family of open sets. The latter is the smallest topology on $\mathbb{E}^n_{\amalg}$ such that the canonical injections $\iota_i : \mathbb{E}_i \to \mathbb{E}^n_{\amalg}, x \mapsto (x, i)$ are continuous. Moreover, $\mathcal{G}^n_{\amalg} = \{\bigcup_{i=1}^n G_i \times \{i\} : G_i \in \mathcal{G}_i\}$ and the analogous relations hold for the families of closed, compact and Borel subsets of $\mathbb{E}^n_{\amalg}$, respectively. It easy is to see that all topological properties of the $\mathbb{E}_i$ carry over to the co-product and so $\mathbb{E}^n_{\amalg}$ is an LCHS space, too.

The question is how the co-product can be used to characterize probability distributions on $\mathcal{B}(\mathcal{F}^n)$. Obviously, each subset A of $\mathbb{E}^n_{\amalg}$ can be written in the form $A = \amalg A_i = \bigcup_{i=1}^n A_i \times \{i\}$ where the A_i are the sections of A, i.e. $A_i = \{x \in \mathbb{E}_i : (x, i) \in A\}$, and consequently A can be identified with the tuple $(A_1, \dots, A_n)$. Hence, we have a one-to-one correspondence between subsets of the co-product $\mathbb{E}^n_{\amalg}$ and tuples of subsets of the $\mathbb{E}_i$. But this means that we have a one-to-one correspondence between $\mathcal{F}^n_{\amalg}$ and $\mathcal{F}^n$ and similarly between $\mathcal{K}^n_{\amalg}$ and $\mathcal{K}^n = \mathcal{K}_1 \times \cdots \times \mathcal{K}_n$.

Consequently, each set function φ on $\mathcal{K}^n_{\amalg}$ is related to a set function ψ on $\mathcal{K}^n$ by

$$\varphi(\amalg K_i) = \psi(K_1, \dots, K_n). \quad (6)$$

The following lemma shows that φ is a capacity functional if and only if ψ is completely alternating and continuous from above in each component. From now on a set function on $\mathcal{K}^n$ fulfilling Conditions (MCF1) - (MCF3) of the following lemma shall be called multivariate capacity functional.

Lemma 1. Let $\varphi : \mathcal{K}^n_{\amalg} \to [0, 1]$ and $\psi : \mathcal{K}^n \to [0, 1]$ satisfying Equation (6) for all $(K_1, \dots, K_n) \in \mathcal{K}^n$. Then φ is a capacity functional if and only if ψ fulfills the following conditions:

(MCF1) $\psi(\emptyset, \dots, \emptyset) = 0$

(MCF2) For all $k \geq 0$, $1 \leq j \leq k$, $K = (K_1, \dots, K_n)$, $K^j = (K^j_1, \dots, K^j_n) \in \mathcal{K}^n$ it holds that

$$\Delta_k \psi(K; K^1, \dots, K^k) \geq 0$$

where $\Delta_0\psi(K) = 1 - \psi(K_1, \dots, K_n)$,

$$\begin{aligned}\Delta_k\psi(K; K^1, \dots, K^k) &= \Delta_{k-1}\psi(K; K^1, \dots, K^{k-1}) \\ &\quad - \Delta_{k-1}\psi(K \cup K^k; K^1, \dots, K^{k-1})\end{aligned}$$

and $K \cup K^k = (K_1 \cup K_1^k, \dots, K_n \cup K_n^k)$.

(MCF3) For all decreasing sequences $\{K_i^k\}_{k\in\mathbb{N}} \subseteq \mathcal{K}_i$, $1 \le i \le n$, it holds that $\psi(K_1^k, \dots, K_n^k) \searrow \psi(K_1, \dots, K_n)$ for $k \to \infty$ where $K_i = \bigcap_{k\in\mathbb{N}} K_i^k$.

Proof. The equivalence follows from the relation $\varphi(\bigcup_{i=1}^n K_i \times \{i\}) = \psi(K_1, \dots, K_n)$. Indeed, we get $\psi(\emptyset, \dots, \emptyset) = \varphi(\bigcup_{i=1}^n \emptyset \times \{i\}) = \varphi(\emptyset)$. Furthermore, by Formula (1) we have

$$\begin{aligned}&\Delta_k\psi(K; K^1, \dots, K^k) \\ &\quad = -\sum_{J\subseteq\{1,\dots,k\}} (-1)^{|J|}\psi\Big(K \cup \bigcup_{j\in J} K^j\Big) \\ &= -\sum_{J\subseteq\{1,\dots,k\}} (-1)^{|J|}\psi\Big(K_1 \cup \bigcup_{j\in J} K_1^j, \dots, K_n \cup \bigcup_{j\in J} K_n^j\Big) \\ &\quad = -\sum_{J\subseteq\{1,\dots,k\}} (-1)^{|J|}\varphi\Big(\bigcup_{i=1}^n (K_i \cup \bigcup_{j\in J} K_i^j) \times \{i\}\Big) \\ &\quad = -\sum_{J\subseteq\{1,\dots,k\}} (-1)^{|J|}\varphi\Big((\amalg K_i) \cup \bigcup_{j\in J} (\amalg K_i^j)\Big) \\ &\quad = \Delta_k\varphi\Big(\amalg K_i; \amalg K_i^1, \dots, \amalg K_i^k\Big).\end{aligned}$$

The equivalence of (MCF3) and (CF3) follows from the fact that $K_i^k \searrow K_i$ for all $1 \le i \le n$ if and only if $\amalg K_i^k = \bigcup_{i=1}^n K_i^k \times \{i\} \searrow \bigcup_{i=1}^n K_i \times \{i\} = \amalg K_i$. □

Given a multivariate set function $\psi : \mathcal{K}^n \to [0,1]$ fulfilling conditions (MCF1) - (MCF3) of the foregoing lemma, the Choquet theorem (Theorem 1) can be applied to the capacity functional $\varphi : \mathcal{K}_\amalg^n \to [0,1]$ defined by $\amalg K_i \mapsto \psi(K_1, \dots, K_n)$. This yields a probability measure $Q : \mathcal{B}(\mathcal{F}_\amalg^n) \to [0,1]$ such that for all $\amalg K_i \in \mathcal{K}_\amalg^n$ it holds that

$$\varphi(\amalg K_i) = Q(\{\amalg F_j \in \mathcal{F}_\amalg^n : \amalg F_j \cap \amalg K_i \neq \emptyset\}). \quad (7)$$

The right-hand side of Equation (7) can further be written in the following form:

$$\begin{aligned}&Q(\{\amalg F_j \in \mathcal{F}_\amalg^n : \amalg F_j \cap \amalg K_i \neq \emptyset\}) \\ &= Q\Big(\Big\{\amalg F_j \in \mathcal{F}_\amalg^n : (\bigcup_{j=1}^n F_j \times \{j\}) \cap (\bigcup_{i=1}^n K_i \times \{i\}) \neq \emptyset\Big\}\Big) \\ &= Q\Big(\bigcup_{i=1}^n \Big\{\amalg F_j \in \mathcal{F}_\amalg^n : (\bigcup_{j=1}^n F_j \times \{j\}) \cap (K_i \times \{i\}) \neq \emptyset\Big\}\Big) \\ &= Q\Big(\bigcup_{i=1}^n \{\amalg F_j \in \mathcal{F}_\amalg^n : F_i \cap K_i \neq \emptyset\}\Big) \\ &= Q\Big(\bigcup_{i=1}^n \{\amalg F_j \in \mathcal{F}_\amalg^n : (F_1, \dots, F_n) \in \widehat{\mathcal{F}}_{K_i}\}\Big) \\ &= Q\Big(\Big\{\amalg F_j \in \mathcal{F}_\amalg^n : (F_1, \dots, F_n) \in \bigcup_{i=1}^n \widehat{\mathcal{F}}_{K_i}\Big\}\Big) \quad (8)\end{aligned}$$

where $\widehat{\mathcal{F}}_{K_i} = \{(F_1, \dots, F_n) \in \mathcal{F}^n : F_i \cap K_i \neq \emptyset\}$.

As already mentioned we have a one-to-one correspondence between $\mathcal{F}_\amalg^n$ and $\mathcal{F}^n$. This can be used to define a probability measure Π on $\mathcal{B}(\mathcal{F}^n)$ from the probability measure Q on $\mathcal{B}(\mathcal{F}_\amalg^n)$ as the following lemma shows.

Lemma 2. It holds that

$$\mathcal{B}(\mathcal{F}_\amalg^n) = \big\{\{\amalg F_i \in \mathcal{F}_\amalg^n : (F_1, \dots, F_n) \in B\} : B \in \mathcal{B}(\mathcal{F}^n)\big\}.$$

Furthermore, if $Q : \mathcal{B}(\mathcal{F}_\amalg^n) \to [0,1]$ is a probability measure then $\Pi : \mathcal{B}(\mathcal{F}^n) \to [0,1]$ defined by

$$\Pi(B) = Q(\{\amalg F_i \in \mathcal{F}_\amalg^n : (F_1, \dots, F_n) \in B\}) \quad (9)$$

is a probability measure, too.

Proof. Let

$$\mathcal{A}_1 = \big\{\{\amalg F_i \in \mathcal{F}_\amalg^n : (F_1, \dots, F_n) \in B\} : B \in \mathcal{B}(\mathcal{F}^n)\big\}.$$

The σ-algebra $\mathcal{B}(\mathcal{F}_\amalg^n)$ is generated by sets of the form $\{\amalg F_i \in \mathcal{F}_\amalg^n : \amalg F_i \cap \amalg K_i \neq \emptyset\}$, $\amalg K_i \in \mathcal{K}_\amalg^n$. As in Equation (8) we obtain

$$\begin{aligned}&\{\amalg F_j \in \mathcal{F}_\amalg^n : \amalg F_j \cap \amalg K_i \neq \emptyset\} \\ &\quad = \Big\{\amalg F_j \in \mathcal{F}_\amalg^n : (F_1, \dots, F_n) \in \bigcup_{i=1}^n \widehat{\mathcal{F}}_{K_i}\Big\}\end{aligned}$$

which lies in $\mathcal{A}_1$ since $\bigcup_{i=1}^n \widehat{\mathcal{F}}_{K_i} \in \mathcal{B}(\mathcal{F}^n)$. It is easy to see that $\mathcal{A}_1$ is a σ-algebra and thus $\mathcal{B}(\mathcal{F}_\amalg^n) \subseteq \mathcal{A}_1$. On the other hand, $\mathcal{B}(\mathcal{F}^n)$ is generated by sets of the form $\mathcal{F}_{K_1} \times \cdots \times \mathcal{F}_{K_n}$, $K_i \in \mathcal{K}_i$. We obtain

$$\begin{aligned}&\{\amalg F_j \in \mathcal{F}_\amalg^n : (F_1, \dots, F_n) \in \mathcal{F}_{K_1} \times \cdots \times \mathcal{F}_{K_n}\} \\ &\quad = \bigcap_{i=1}^n \{\amalg F_j \in \mathcal{F}_\amalg^n : (F_1, \dots, F_n) \in \widehat{\mathcal{F}}_{K_i}\} \\ &\quad = \bigcap_{i=1}^n \{\amalg F_j \in \mathcal{F}_\amalg^n : \amalg F_j \cap (K_i \times \{i\}) \neq \emptyset\}\end{aligned}$$

which lies in $\mathcal{B}(\mathcal{F}_\amalg^n)$ since $K_i \times \{i\} \in \mathcal{K}_\amalg^n$. Furthermore, it is easy to see that

$$\mathcal{A}_2 = \{B \in \mathcal{B}(\mathcal{F}^n) : \{\amalg F_i : (F_1, \dots, F_n) \in B\} \in \mathcal{B}(\mathcal{F}_\amalg^n)\}$$

is a σ-algebra. Thus $\mathcal{B}(\mathcal{F}^n) = \mathcal{A}_2$ which further implies $\mathcal{A}_1 \subseteq \mathcal{B}(\mathcal{F}_\amalg^n)$. It can be easily checked that Π is a probability measure. □

From Equations (6), (7), (8) and (9) we obtain the following relation between the multivariate capacity functional ψ and the probability measure Π:

$$\begin{aligned}&\psi(K_1,\dots,K_n)=\varphi(\amalg K_i)\\&=Q(\{\amalg F_j\in\mathcal{F}_{\amalg}^n:\amalg F_j\cap\amalg K_i\neq\emptyset\}=\Pi\Big(\bigcup_{i=1}^n\widehat{\mathcal{F}}_{K_i}\Big)\end{aligned}$$

We are now ready to formulate the following proposition which can be viewed as a multivariate version of the Choquet theorem.

Proposition 1. Let $\psi:\mathcal{K}^n\to[0,1]$ be a multivariate capacity functional (that is a set function fulfilling Conditions (MCF1) - (MCF3) of Lemma 1). Then there exists a unique probability measure $\Pi:\mathcal{B}(\mathcal{F}^n)\to[0,1]$ such that

$$\psi(K_1,\dots,K_n)=\Pi\Big(\bigcup_{i=1}^n\widehat{\mathcal{F}}_{K_i}\Big)$$

for all $(K_1,\dots,K_n)\in\mathcal{K}^n$.

This means that the probability of events of the form $\bigcup_{i=1}^n\widehat{\mathcal{F}}_{K_i}$ can be directly computed by ψ. Probabilities of other events like $\mathcal{F}_{K_1}\times\cdots\times\mathcal{F}_{K_n}$ can be computed by using the exclusion-inclusion principle and the complete alternation property:

$$\begin{aligned}\Pi(\mathcal{F}_{K_1}\times\cdots\times\mathcal{F}_{K_n})&=\Pi\Big(\bigcap_{i=1}^n\widehat{\mathcal{F}}_{K_i}\Big)\\&=-\sum_{I\subseteq\{1,\dots,n\}}(-1)^{|I|}\Pi\Big(\bigcup_{i\in I}\widehat{\mathcal{F}}_{K_i}\Big)\\&=-\sum_{I\subseteq\{1,\dots,n\}}(-1)^{|I|}\varphi\Big(\bigcup_{i\in I}K_i\times\{i\}\Big)\\&=\Delta_n\varphi(\emptyset;K_1\times\{1\},\dots,K_n\times\{n\})\\&=\Delta_n\psi(\emptyset;\check{K}_1,\dots,\check{K}_n)\end{aligned}\quad(10)$$

where $\check{K}_i=(\emptyset,\dots,\emptyset,K_i,\emptyset,\dots,\emptyset)\in\mathcal{K}^n$. We can state an additional result concerning the probability of $\mathcal{F}'^n=\mathcal{F}_1'\times\cdots\times\mathcal{F}_n'$, that is, the set of tuples of non-empty closed subsets.

Corollary 1. In the situation of Proposition 1, if ψ fulfills in addition for all $1\le i\le n$

$$\sup\{\psi(\check{K}_i):K_i\in\mathcal{K}_i\}=1$$

then $\Pi(\mathcal{F}'^n)=1$, that is, a tuple of closed sets almost surely consists of non-empty sets.

Proof. Let $\{L_i^k\}_{k\in\mathbb{N}}\in\mathcal{K}_i$ be increasing sequences such that $L_i^k\nearrow\mathbb{E}_i$ for all $1\le i\le n$, let $\{M_i^k\}_{k\in\mathbb{N}}\subseteq\mathcal{K}_i$ be increasing sequences such that $\psi(\check{M}_i^k)\nearrow1$ for all $1\le i\le n$ and let $K_i^k=L_i^k\cup M_i^k$ for all $1\le i\le n$ and $k\in\mathbb{N}$. Consequently,

$$\begin{aligned}\mathcal{F}'^n&=\bigcup_{k\in\mathbb{N}}\mathcal{F}_{K_1^k}\times\cdots\times\mathcal{F}_{K_n^k}\\&=\bigcup_{k\in\mathbb{N}}\bigcap_{i=1}^n\widehat{\mathcal{F}}_{K_i^k}=\bigcap_{i=1}^n\bigcup_{k\in\mathbb{N}}\widehat{\mathcal{F}}_{K_i^k}.\end{aligned}$$

By the exclusion-inclusion principle we obtain

$$\begin{aligned}\Pi(\mathcal{F}'^n)&=\Pi\Big(\bigcap_{i=1}^n\bigcup_{k\in\mathbb{N}}\widehat{\mathcal{F}}_{K_i^k}\Big)\\&=-\sum_{\emptyset\neq I\subseteq\{1,\dots,n\}}(-1)^{|I|}\Pi\Big(\bigcup_{i\in I}\bigcup_{k\in\mathbb{N}}\widehat{\mathcal{F}}_{K_i^k}\Big)\\&=-\sum_{\emptyset\neq I\subseteq\{1,\dots,n\}}(-1)^{|I|}\Pi\Big(\bigcup_{k\in\mathbb{N}}\bigcup_{i\in I}\widehat{\mathcal{F}}_{K_i^k}\Big)\\&=-\sum_{\emptyset\neq I\subseteq\{1,\dots,n\}}(-1)^{|I|}\sup_{k\in\mathbb{N}}\Pi\Big(\bigcup_{i\in I}\widehat{\mathcal{F}}_{K_i^k}\Big).\end{aligned}$$

For all $I\neq\emptyset$, $i\in I$ and $k\in\mathbb{N}$ we have

$$\psi(\check{K}_i^k)=\Pi(\widehat{\mathcal{F}}_{K_i^k})\le\Pi\Big(\bigcup_{i\in I}\widehat{\mathcal{F}}_{K_i^k}\Big)\le1$$

and thus

$$\sup_{k\in\mathbb{N}}\psi(\check{K}_i^k)=\sup_{k\in\mathbb{N}}\Pi\Big(\bigcup_{i\in I}\widehat{\mathcal{F}}_{K_i^k}\Big)=1.$$

Hence,

$$\Pi(\mathcal{F}'^n)=-\sum_{\emptyset\neq I\subseteq\{1,\dots,n\}}(-1)^{|I|}=1.$$

□

Note that if we consider n almost surely non-empty random sets $X_1,\dots,X_n$ on a probability space (Ω,Σ,P) then the multivariate capacity functional of the product random set defined by Equation (5) is given by

$$\psi(K_1,\dots,K_n)=P\Big(\bigcup_{i=1}^nX_i^-(K_i)\Big).$$

Hence, if the $\mathbb{E}_i$ are σ-compact spaces (which is the case if the $\mathbb{E}_i$ are LCHS spaces) and $\{K_i^k\}_{k\in\mathbb{N}}\subseteq\mathcal{K}_i$ are increasing sequences converging to $\mathbb{E}_i$, respectively, we obtain

$$\lim_{k\to\infty}\psi(\check{K}_i^k)=\lim_{k\to\infty}P(X_i^-(K_i^k))=P(X_i\neq\emptyset)=1$$

and thus the condition of Corollary 1 is fulfilled.

We will now relate multivariate capacity functionals to set functions on special classes of subsets of the product space $\mathbb{E}^n$. Up to now we have used the fact that a tuple $(A_1, \ldots, A_n)$ of subsets of the $\mathbb{E}_i$ can be identified with the set $\amalg A_i = \bigcup_{i=1}^n A_i \times \{i\}$ which is a subset of the co-product $\mathbb{E}_{\amalg}^n$. On the other hand, a tuple $(A_1, \ldots, A_n)$ can be identified with $\bigcup_{i=1}^n \hat{A}_i$ where

$$\hat{A}_i = \{(x_1, \ldots, x_n) \in \mathbb{E}^n : x_i \in A_i\}.$$

Consequently, we have a one-to-one correspondence between $\mathcal{K}^n$ and

$$\hat{\mathcal{K}}_{\cup}^n = \Big\{ \bigcup_{i=1}^n \hat{K}_i : K_i \in \mathcal{K}_i \Big\}$$

and each set function ψ on $\mathcal{K}^n$ is related to a set function ϕ on $\hat{\mathcal{K}}_{\cup}^n$ by

$$\psi(K_1, \ldots, K_n) = \phi\Big(\bigcup_{i=1}^n \hat{K}_i \Big). \tag{11}$$

Similar to Lemma 1 one has the following lemma.

Lemma 3. Let $\psi : \mathcal{K}^n \to [0,1]$ and $\phi : \hat{\mathcal{K}}_{\cup}^n \to [0,1]$ satisfying Equation (11) for all $(K_1, \ldots, K_n) \in \mathcal{K}^n$. Then ϕ is a capacity functional (that is, ϕ fulfills Conditions (CF1), (CF2) and (CF3) for sets from $\hat{\mathcal{K}}_{\cup}^n$) if and only if ψ is a multivariate capacity functional (that is, ψ fulfills Conditions (MCF1) - (MCF3) of Lemma 1).

Proof. The equivalence follows from the relation $\phi(\bigcup_{i=1}^n \hat{K}_i) = \psi(K_1, \ldots, K_n)$. Indeed, we have $\bigcup_{i=1}^n \hat{K}_i = \emptyset$ if and only if $K_i = \emptyset$ for all i and thus $\phi(\emptyset) = \psi(\emptyset, \ldots, \emptyset)$. Furthermore, by Formula (1) we have for all $K = (K_1, \ldots, K_n) \in \mathcal{K}^n$, $K^j = (K_1^j, \ldots, K_n^j) \in \mathcal{K}^n$

$$\begin{aligned}
&\Delta_k \psi(K; K^1, \ldots, K^k) \\
&\quad = -\sum_{J \subseteq \{1,\ldots,k\}} (-1)^{|J|} \psi\Big(K \cup \bigcup_{j \in J} K^j \Big) \\
&= -\sum_{J \subseteq \{1,\ldots,k\}} (-1)^{|J|} \psi\Big(K_1 \cup \bigcup_{j \in J} K_1^j, \ldots, K_n \cup \bigcup_{j \in J} K_n^j \Big) \\
&\quad = -\sum_{J \subseteq \{1,\ldots,k\}} (-1)^{|J|} \phi\Big(\bigcup_{i=1}^n \big(K_i \cup \bigcup_{j \in J} K_i^j \big)\widehat{\ } \Big) \\
&\quad = -\sum_{J \subseteq \{1,\ldots,k\}} (-1)^{|J|} \phi\Big(\bigcup_{i=1}^n \hat{K}_i \cup \bigcup_{j \in J} \bigcup_{i=1}^n \hat{K}_i^j \Big) \\
&\quad = \Delta_k \phi\Big(\bigcup_{i=1}^n \hat{K}_i; \bigcup_{i=1}^n \hat{K}_i^1, \ldots, \bigcup_{i=1}^n \hat{K}_i^k \Big).
\end{aligned}$$

The equivalence of (MCF3) and (CF3) follows from the fact that $K_i^k \searrow K_i$ for all $1 \le i \le n$ if and only if $\bigcup_{i=1}^n \hat{K}_i^k \searrow \bigcup_{i=1}^n \hat{K}_i$. □

Together with Proposition 1 this implies the following proposition which gives a characterization of the joint distribution of n random sets by a set function on $\hat{\mathcal{K}}_{\cup}^n$.

Proposition 2. Let $\phi : \hat{\mathcal{K}}_{\cup}^n \to [0,1]$ be a capacity functional, that is, ϕ fulfills Conditions (CF1), (CF2) and (CF3) for sets from $\hat{\mathcal{K}}_{\cup}^n$. Then there exists a unique probability measure $\Pi : \mathcal{B}(\mathcal{F}^n) \to [0,1]$ such that

$$\phi\Big(\bigcup_{i=1}^n \hat{K}_i \Big) = \Pi\Big(\bigcup_{i=1}^n \widehat{\mathcal{F}}_{K_i} \Big)$$

for all $\bigcup_{i=1}^n \hat{K}_i \in \hat{\mathcal{K}}_{\cup}^n$. If, in addition, for all $1 \le i \le n$ it holds that $\sup\{\phi(\hat{K}_i) : K_i \in \mathcal{K}_i\} = 1$ then $\Pi(\mathcal{F}'^n) = 1$ and for all $L \in \hat{\mathcal{K}}_{\cup}^n$ it holds that

$$\phi(L) = \Pi({}^n\mathcal{F}_L).$$

Proof. The main assertion directly follows from applying Proposition 1 to $\psi : \mathcal{K}^n \to [0,1]$ defined by $\psi(K_1, \ldots, K_n) = \phi(\bigcup_{i=1}^n \hat{K}_i)$ which is a multivariate capacity functional by Lemma 3. The additional statement follows from the fact that $\psi(\hat{K}_i) = \phi(\hat{K}_i)$. By virtue of Corollary 1 this implies $\Pi(\mathcal{F}'^n) = 1$ which further leads to

$$\Pi\Big(\bigcup_{i=1}^n \widehat{\mathcal{F}}_{K_i} \Big) = \Pi\Big(\mathcal{F}'^n \cap \bigcup_{i=1}^n \widehat{\mathcal{F}}_{K_i} \Big)$$

for all $K_i \in \mathcal{K}_i$. Furthermore, we obtain

$$\begin{aligned}
&\mathcal{F}'^n \cap \bigcup_{i=1}^n \widehat{\mathcal{F}}_{K_i} \\
&\qquad = \bigcup_{i=1}^n \{(F_1, \ldots, F_n) \in \mathcal{F}'^n : F_i \cap K_i \ne \emptyset\} \\
&= \bigcup_{i=1}^n \{(F_1, \ldots, F_n) \in \mathcal{F}^n : F_1 \times \cdots \times F_n \cap \hat{K}_i \ne \emptyset\} \\
&\qquad = {}^n\mathcal{F}_{\bigcup_{i=1}^n \hat{K}_i}.
\end{aligned}$$

Hence, $\phi(L) = \Pi({}^n\mathcal{F}_L)$ for all $L \in \hat{\mathcal{K}}_{\cup}^n$. □

One can think of extending the various set functions to wider classes of sets. In case of a capacity functional $\varphi : \mathcal{K}_{\amalg}^n \to [0,1]$ the extensions from Equation (2) can be used to obtain a completely alternating Choquet-$\mathcal{K}_{\amalg}^n$-capacity $\varphi^* : \mathcal{P}_{\amalg}^n \to [0,1]$ on the power set of $\mathbb{E}_{\amalg}^n$. In case of a multivariate capacity functional $\psi : \mathcal{K}^n \to [0,1]$ or a capacity functional $\phi : \hat{\mathcal{K}}_{\cup}^n \to [0,1]$ one can define a corresponding capacity functional φ on $\mathcal{K}_{\amalg}^n$ by the relation $\varphi(\amalg K_i) = \psi(K_1, \ldots, K_n)$ or $\varphi(\amalg K_i) = \phi(\bigcup_{i=1}^n \hat{K}_i)$ and use φ^* to obtain ψ^* or ϕ^*. On the other hand, the extension procedure given by Equation (2) can be

directly applied to ψ or ϕ which yields the same ψ^* or ϕ^* since $\amalg A_i \subseteq \amalg B_i$ if and only if $A_i \subseteq B_i$ for all i if and only if $\bigcup_{i=1}^n \hat{A}_i \subseteq \bigcup_{i=1}^n \hat{B}_i$.

We have seen how a capacity functional ϕ defined on $\hat{\mathcal{K}}_\cup^n$ can be extended to a Choquet-$\hat{\mathcal{K}}_\cup^n$-capacity on

$$\hat{\mathcal{P}}_\cup^n = \Big\{ \bigcup_{i=1}^n \hat{A}_i : A_i \in \mathcal{P}_i \Big\}.$$

We point out that a further extension to all subsets of $\mathbb{E}^n = \mathbb{E}_1 \times \cdots \times \mathbb{E}_n$ would not make much sense since this extension would not be unique. Indeed, consider the following two (deterministic) sets $X_1 = [0,1]^2$ and $X_2 = \{(x,y) \in [0,1]^2 : x + y \geq 1\}$. They can be seen as random compact sets in $\mathbb{R}^2$ on a one point probability space. The corresponding capacity functionals ϕ_1 and ϕ_2 are given by

$$\phi_i(A) = \begin{cases} 1 & \text{if } X_i \cap A \neq \emptyset \\ 0 & \text{if } X_i \cap A = \emptyset \end{cases}$$

for each $A \subseteq \mathbb{R}^2$. Obviously, ϕ_1 and ϕ_2 coincide on $\hat{\mathcal{K}}_\cup^2$ but they have different values on other sets, for example, $\phi_1(A) = 1$ and $\phi_2(A) = 0$ for $A = [0, 1/3]^2$.

4 Application to set-valued processes

Let T denote a time set, let $(\mathbb{M}, \mathcal{M})$ be a measurable space and let (Ω, Σ, P) be a probability space. Then a map $x : T \times \Omega \to \mathbb{M}$ is a stochastic process if for each $t \in T$ the partial map $x_t : \Omega \to \mathbb{M}$ is measurable, that is, $x_t^{-1}(B) \in \Sigma$ for all $B \in \mathcal{M}$. Denoting by $\mathcal{T}$ the set of all finite subsets of T, the process x induces a family $\{\mu_{\underline{t}}\}_{\underline{t} \in \mathcal{T}}$ of probability measures where

$$\begin{aligned} \mu_{\underline{t}} : \mathcal{M}^{\otimes n} &\to [0,1], \\ B &\mapsto P(\{\omega \in \Omega : (x_{t_1}(\omega), \ldots, x_{t_n}(\omega)) \in B\}), \end{aligned}$$

$\underline{t} = (t_1, \ldots, t_n)$, $\mathcal{M}^{\otimes n} = \mathcal{M} \otimes \cdots \otimes \mathcal{M}$. The latter is called the family of finite-dimensional distributions of x and obviously fulfills the following two conditions:

(i) For all $\underline{t} = (t_1, \ldots, t_n) \in \mathcal{T}$, $B_1, \ldots, B_n \in \mathcal{M}$ and each permutation σ of $\{1, \ldots, n\}$ it holds that

$$\mu_{\underline{t}}(B_1 \times \cdots \times B_n) = \mu_{\sigma(\underline{t})}(B_{\sigma(1)} \times \cdots \times B_{\sigma(n)})$$

where $\sigma(\underline{t}) = (t_{\sigma(1)}, \ldots, t_{\sigma(n)})$.

(ii) For all $\underline{t} = (t_1, \ldots, t_n) \in \mathcal{T}$, $t_{n+1} \in T$, $B \in \mathcal{M}^{\otimes n}$ it holds that

$$\mu_{t_1,\ldots,t_{n+1}}(B \times \mathbb{M}) = \mu_{\underline{t}}(B).$$

A family of finite-dimensional distributions is said to be consistent if these two conditions are fulfilled. Under the assumption that $\mathbb{M}$ is a complete separable metric space endowed with its Borel sets $\mathcal{B}(\mathbb{M})$, the well-known Daniell-Kolmogorov theorem [5, 7] says that for any consistent family of finite-dimensional distributions there exists a stochastic process whose finite dimensional distributions coincide with that family. More precisely, consider the set of maps from T to $\mathbb{M}$ denoted by $\mathbb{M}^T$ which is endowed with the σ-algebra $\mathcal{B}(\mathbb{M}^T)$ generated by sets of the form $\{\omega \in \mathbb{M}^T : (\omega(t_1), \ldots, \omega(t_n)) \in B\}$, $B \in \mathcal{B}(\mathbb{M}^n)$, $t_i \in T$, $n \geq 1$. Then there exists a probability measure μ on $\mathcal{B}(\mathbb{M}^T)$ such that for all $\underline{t} = (t_1, \ldots, t_n) \in \mathcal{T}$ $(n \geq 1)$ and $B \in \mathcal{B}(\mathbb{M}^n)$ it holds that

$$\mu_{\underline{t}}(B) = \mu(\{\omega \in \mathbb{M}^T : (\omega(t_1), \ldots, \omega(t_n)) \in B\}).$$

The desired process is then given by $(t, \omega) \to \omega(t)$.

By set-valued stochastic processes we mean stochastic processes where $\mathbb{M} = \mathcal{F}$, that is, maps of the form

$$X : T \times \Omega \to \mathcal{F}$$

where $X_t : \Omega \to \mathcal{F}$ is Effros-measurable for all $t \in T$. With the aid of Proposition 1 we can now formulate an existence theorem for set-valued processes by using multivariate capacity functionals.

Proposition 3. Let $\{\psi_{\underline{t}} : \underline{t} \in \mathcal{T}\}$ be a family of multivariate capacity functionals (i.e. set functions fulfilling Conditions (MCF1) - (MCF3) of Lemma 1). Assume that the following consistency conditions are fulfilled:

(i) For all $n \geq 1$, $\underline{t} = (t_1, \ldots, t_n) \in \mathcal{T}$, $K_1, \ldots, K_n \in \mathcal{K}$ and each permutation σ of $\{1, \ldots, n\}$ it holds that

$$\psi_{\underline{t}}(K_1, \ldots, K_n) = \psi_{\sigma(\underline{t})}(K_{\sigma(1)}, \ldots, K_{\sigma(n)})$$

where $\sigma(\underline{t}) = (t_{\sigma(1)}, \ldots, t_{\sigma(n)})$.

(ii) For all $n \geq 1$, $\underline{t} = (t_1, \ldots, t_n) \in \mathcal{T}$, $t_{n+1} \in T$, $K_1, \ldots, K_n \in \mathcal{K}$ it holds that

$$\psi_{t_1,\ldots,t_{n+1}}(K_1, \ldots, K_n, \emptyset) = \psi_{\underline{t}}(K_1, \ldots, K_n).$$

Then the family $\{\Pi_{\underline{t}} : \underline{t} \in \mathcal{T}\}$ obtained from Proposition 1 is a consistent family of probability measures and there exists a probability measure Π on $\mathcal{B}(\mathcal{F}^T)$ such that for all $\underline{t} = (t_1, \ldots, t_n) \in \mathcal{T}$ $(n \geq 1)$ and $(K_1, \ldots, K_n) \in \mathcal{K}^n$ it holds that

$$\begin{aligned} &\psi_{\underline{t}}(K_1, \ldots, K_n) \\ &= \Pi_{\underline{t}}\Big(\Big\{\omega \in \mathcal{F}^T : (\omega(t_1), \ldots, \omega(t_n)) \in \bigcup_{i=1}^n \widehat{\mathcal{F}}_{K_i}\Big\}\Big). \end{aligned} \tag{12}$$

In addition, the condition $\sup\{\psi_t(K) : K \in \mathcal{K}\} = 1$ implies $\Pi_t(\{\omega \in \mathcal{F}^T : \omega(t) \neq \emptyset\}) = 1$ for all $t \in T$.

Proof. Since $\mathbb{E}$ is an LCHS space, $\mathcal{F}$ is a compact Hausdorff second countable space. Thus, $\mathcal{F}$ is also a Polish space, that is, separable and completely metrizable. Hence, if we show that $\{\Pi_{\underline{t}} : \underline{t} \in \mathcal{T}\}$ is a consistent family of probability measures the classical Daniell-Kolmogorov theorem can be applied directly and Equation (12) is obtained from Proposition 1:

$$\begin{aligned}\psi_{\underline{t}}(K_1,\dots,K_n) &= \Pi_{\underline{t}}\Big(\bigcup_{i=1}^{n}\widehat{\mathcal{F}}_{K_i}\Big)\\ &= \Pi\Big(\Big\{\omega\in\mathcal{F}^T : (\omega(t_1),\dots,\omega(t_n))\in\bigcup_{i=1}^{n}\widehat{\mathcal{F}}_{K_i}\Big\}\Big)\end{aligned}$$

It is enough to prove that the consistency conditions for $\{\Pi_{\underline{t}} : \underline{t} \in \mathcal{T}\}$ are fulfilled for cylindrical sets of the form

$$\mathcal{F}_{K_{11},\dots,K_{1k_1}} \times \cdots \times \mathcal{F}_{K_{n1},\dots,K_{nk_n}},$$

$K_{ij_i} \in \mathcal{K}$, since they constitute a generating class of $\mathcal{B}(\mathcal{F}^n) = \mathcal{B}(\mathcal{F})^{\otimes n}$ which is closed under finite intersections. Similarly as in Equation (10) we obtain the following formula

$$\begin{aligned}&\Pi_{\underline{t}}(\mathcal{F}_{K_{11},\dots,K_{1k_1}} \times \cdots \times \mathcal{F}_{K_{n1},\dots,K_{nk_n}})\\ &\quad= \Pi_{\underline{t}}\Big(\bigcap_{i=1}^{n}\widehat{\mathcal{F}}_{K_{i1},\dots,K_{ik_i}}\Big) = \Pi_{\underline{t}}\Big(\bigcap_{i=1}^{n}\bigcap_{j_i=1}^{k_i}\widehat{\mathcal{F}}_{K_{ij_i}}\Big)\\ &\quad= -\sum_{J\in\mathcal{J}}(-1)^{|J|}\Pi_{\underline{t}}\Big(\bigcup_{i=1}^{n}\bigcup_{j_i\in J_i}\widehat{\mathcal{F}}_{K_{ij_i}}\Big)\\ &\quad= -\sum_{J\in\mathcal{J}}(-1)^{|J|}\Pi_{\underline{t}}\Big(\bigcup_{i=1}^{n}\widehat{\mathcal{F}}_{\bigcup_{j_i\in J_i}K_{ij_i}}\Big)\\ &= -\sum_{J\in\mathcal{J}}(-1)^{|J|}\psi_{\underline{t}}\Big(\bigcup_{j_1\in J_1}K_{1j_1},\dots,\bigcup_{j_n\in J_n}K_{nj_n}\Big)\end{aligned}$$

where $\mathcal{J} = \{(J_1,\dots,J_n) : J_i \subseteq \{1,\dots,k_i\}\}$ and $|J| = \sum_{i=1}^{n}|J_i|$. Together with (i) this implies

$$\begin{aligned}&\Pi_{\underline{t}}(\mathcal{F}_{K_{11},\dots,K_{1k_1}} \times \cdots \times \mathcal{F}_{K_{n1},\dots,K_{nk_n}})\\ &= \Pi_{\sigma(\underline{t})}(\mathcal{F}_{K_{\sigma(1)1},\dots,K_{\sigma(1)k_{\sigma(1)}}}\times\cdots\times\mathcal{F}_{K_{\sigma(n)1},\dots,K_{\sigma(n)k_{\sigma(n)}}})\end{aligned}$$

In a similar manner as before we obtain

$$\begin{aligned}&\Pi_{t_1,\dots,t_{n+1}}(\mathcal{F}_{K_{11},\dots,K_{1k_1}} \times \cdots \times \mathcal{F}_{K_{n1},\dots,K_{nk_n}} \times \mathcal{F})\\ &= -\sum_{J\in\mathcal{J}}(-1)^{|J|}\psi_{t_1,\dots,t_{n+1}}\Big(\bigcup_{j_1\in J_1}K_{1j_1},\dots,\bigcup_{j_n\in J_n}K_{nj_n},\emptyset\Big).\end{aligned}$$

and thus (ii) implies

$$\begin{aligned}&\Pi_{t_1,\dots,t_{n+1}}(\mathcal{F}_{K_{11},\dots,K_{1k_1}} \times \cdots \times \mathcal{F}_{K_{n1},\dots,K_{nk_n}} \times \mathcal{F})\\ &\qquad= \Pi_{\underline{t}}(\mathcal{F}_{K_{11},\dots,K_{1k_1}} \times \cdots \times \mathcal{F}_{K_{n1},\dots,K_{nk_n}}).\end{aligned}$$

The additional statement that $\sup\{\psi_t(K) : K \in \mathcal{K}\} = 1$ implies $\Pi_t(\{\omega \in \mathcal{F}^T : \omega(t) \neq \emptyset\}) = 1$ directly follows from Corollary 1. □

It should be mentioned that in [9] a Daniell-Kolmogorov theorem for supremum preserving (also called maxitive) upper probabilities has been proved.

With the aid of the foregoing proposition we can now try to construct something like a set-valued Brownian motion. Brownian motion is a real-valued stochastic process in continuous time which is defined via a consistent family of Gaussian distributions. More precisely, it is a process with continuous sample functions starting at time 0 with value 0, and it has independent, Gaussian distributed increments with mean 0. We denote by $\{\beta_{\underline{t}}\}_{\underline{t}\in\mathcal{T}}$ ($T = [0,\infty)$) its family of finite dimensional distributions which is clearly consistent. According to Equation (11) and Lemma 3 we get a family of multivariate capacity functionals $\{\psi_{\underline{t}}\}_{\underline{t}\in\mathcal{T}}$ which can be easily seen to be consistent. In addition, we have for all $\underline{t} = (t_1,\dots,t_n) \in \mathcal{T}$ and for all $1 \leq i \leq n$ that

$$\begin{aligned}\sup\{\psi_{\underline{t}}(\check{K}_i) : K_i \in \mathcal{K}\} &= \sup\{\beta_{\underline{t}}(\hat{K}_i) : K_i \in \mathcal{K}\}\\ &= \sup\{\beta_{t_i}(K_i) : K_i \in \mathcal{K}\} = 1.\end{aligned}$$

By applying Propositions 2 and 3 we get a probability measure Π on $\mathcal{B}(\mathcal{F}^{[0,\infty)})$ such that for each $\underline{t} = (t_1,\dots,t_n) \in \mathcal{T}$ it holds that $\Pi_{\underline{t}}(\mathcal{F}'^n) = 1$ and for each $(K_1,\dots,K_n) \in \mathcal{K}^n$ we get

$$\begin{aligned}&\Pi\Big(\Big\{\omega\in\mathcal{F}^{[0,\infty)} : (\omega(t_1),\dots,\omega(t_n))\in {}^{n}\mathcal{F}_{\bigcup_{i=1}^{n}\hat{K}_i}\Big\}\Big)\\ &= \Pi_{\underline{t}}\Big(\bigcup_{i=1}^{n}\widehat{\mathcal{F}}_{K_i}\Big) = \psi_{\underline{t}}(K_1,\dots,K_n) = \beta_{\underline{t}}\Big(\bigcup_{i=1}^{n}\hat{K}_i\Big).\end{aligned}$$

By defining

$$B : [0,\infty)\times\mathcal{F}^{[0,\infty)} \to \mathcal{F},\ (t,\omega)\mapsto B_t(\omega) = \omega(t)$$

we get a set-valued process with finite dimensional distributions $\{\Pi_{\underline{t}}\}_{\underline{t}}$ and finite dimensional capacity functionals $\{\psi_{\underline{t}}\}_{\underline{t}}$. For time $t \in [0,\infty)$ and $G \in \mathcal{G}$ we get

$$\begin{aligned}\Pi(\{\omega : B_t(\omega)\cap G \neq \emptyset\}) &= \Pi_t(\mathcal{F}_G) = \Pi_t\Big(\bigcup_{n\in\mathbb{N}}\mathcal{F}_{K_n}\Big)\\ &= \lim_{n\to\infty}\Pi_t(\mathcal{F}_{K_n}) = \lim_{n\to\infty}\beta_t(K_n) = \beta_t(G)\end{aligned}$$

where $\{K_n\}_{n\in\mathbb{N}} \subseteq \mathcal{K}$ is an increasing sequence such that $\bigcup_{i=1}^{n}K_n = G$. On the other hand, if we approximate G^{c} by an increasing sequence $\{K_n\}_{n\in\mathbb{N}} \subseteq \mathcal{K}$ we

obtain

$$\begin{aligned}\Pi(\{\omega : B_t(\omega) \subseteq G\}) &= \Pi_t(\mathcal{F}^{G^c}) = 1 - \Pi_t(\mathcal{F}_{G^c}) \\ &= 1 - \lim_{n\to\infty} \Pi_t(\mathcal{F}_{K_n}) = 1 - \lim_{n\to\infty} \beta_t(K_n) \\ &= 1 - \beta_t(G^c) = \beta_t(G).\end{aligned}$$

Consequently, the lower and the upper probability of B_t coincide and thus, B_t is almost surely a singleton. This means that although B has values in $\mathcal{F}$ it is actually not a set-valued process but a version of classical Brownian motion.

Note that there are other approaches to define a set-valued Brownian motion via support functions (see [11, 12]), but at least in the real-valued case they also lead to set-valued processes that almost surely consist of singletons.

5 Summary and conclusion

The goal of this paper was to give a characterization of probability measures on the Borel subsets of $\mathcal{F}^n$ ($n \geq 2$) by set functions. The first approach was to use a set function φ defined on the compact subsets of the co-product $\mathbb{E}^n_{\amalg}$ and to apply the (classical) Choquet theorem leading to a probability measure Q on the Borel-σ-algebra of the closed subsets of $\mathbb{E}^n_{\amalg}$. It has been shown that instead of φ one can equivalently use a set function ψ defined on the cartesian product $\mathcal{K}^n = \mathcal{K}_1 \times \cdots \times \mathcal{K}_n$ (Lemma 1). Moreover, it has been demonstrated how to obtain a probability measure Π on $\mathcal{B}(\mathcal{F}^n)$ from Q (Lemma 2). This resulted in a characterization of probability measures on $\mathcal{B}(\mathcal{F}^n)$ by set functions on $\mathcal{K}^n$ called multivariate capacity functionals (Proposition 1). In addition, Proposition 2 stated a characterization using set functions on $\hat{\mathcal{K}}^n_{\cup}$ which is a special class of subsets of the product space $\mathbb{E}^n = \mathbb{E}_1 \times \cdots \times \mathbb{E}_n$. Figure 1 gives an overview of the proposed characterizations.

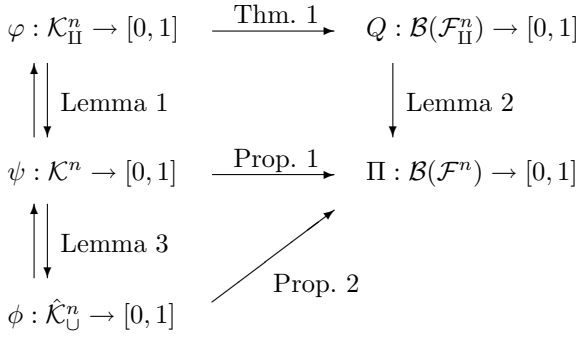


Figure 1: Overview over characterizations of probability measures by set functions.

In Section 4, we have stated a Daniell-Kolmogorov theorem for set-valued stochastic processes, that is, we have demonstrated that for a consistent family of multivariate capacity functionals there exists a set-valued process whose finite dimensional upper probabilities coincide with these multivariate capacity functionals.

Acknowledgements

I would like to thank the reviewers for their useful comments and suggestions. Especially, one of the remarks led to a substantial improvement of the paper.

References

[1] G. Beer. Topologies on Closed and Closed Convex Sets. Kluwer Academic Publishers, Dordrecht, 1993.

[2] C. Castaing, M. Valadier. Convex analysis and measurable multifunctions. Lecture notes in mathematics 580, Springer, 1977.

[3] G. Choquet. Theory of capacities. Annales de l'Institut Fourier, 5:131–295, 1953.

[4] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics 38:325–339, 1967.

[5] J. L. Doob. Stochastic Processes. Wiley, 1990.

[6] E. B. Dynkin. Die Grundlagen der Theorie der Markoffschen Prozesse. Springer, 1961.

[7] I. I. Gikhman, A. V. Skorokhod. Introduction to the theory of random processes. Saunders Company, 1969.

[8] C. J. Himmelberg. Measurable relations. Fundamenta Mathematicae, 87:53–72, 1975.

[9] H. Janssen, G. de Cooman, E. E. Kerre. A Daniell-Kolmogorov theorem for supremum preserving upper probabilities. Fuzzy Sets and Systems 102:429–444, 1999.

[10] R. Kruse, K. D. Meyer. Statistics with vague data. D. Reidel Publishing Company, Dordrecht, 1987.

[11] Shoumei Li, Li Guan. Fuzzy set-valued Gaussian processes and Brownian motions. Information Sciences 177:3251–3259, 2007.

[12] C. Y. Liu, Xu. Han, M. Feng, S. K. Li. Itô integral for bounded closed convex set valued Wiener stochastic processes. In Y. Liu, G. Chen, M. Ying, editors, Fuzzy Logic, Soft Computing and Computational Intelligence, Proceedings of the Eleventh International Fuzzy Systems Association World Congress, Beijing, China, July 28–31, 2005, pages 184–187.

[13] G. Matheron. Random Sets and Integral Geometry. Wiley, 1975.

[14] P. A. Meyer. Probability and Potentials. Waltman, London, 1966.

[15] I. Molchanov. Theory of random sets. Springer, 2005.

[16] J. Neveu. Mathematical Foundations of the Calculus of Probability. Holden-Day, Inc., 1965.

[17] H. T. Nguyen. An Introduction to Random Sets. Chapman & Hall, 2006.

[18] H. T. Nguyen. On random sets and belief functions. J. Math. Anal. Appl., 65:531-542, 1978.

[19] B. Schmelzer. On solutions of stochastic differential equations with parameters modelled by random sets. PhD thesis, 2011.

Forecasting with Imprecise Probabilities

Teddy Seidenfeld
teddy@stat.cmu.edu

Mark J. Schervish
mark@cmu.edu

Joseph B. Kadane
kadane@stat.cmu.edu

Carnegie Mellon University

Abstract

We review de Finetti's two coherence criteria for determinate probabilities: *coherence*$_1$ defined in terms of previsions for a set of random variables that are undominated by the status quo – previsions immune to a sure-loss – and *coherence*$_2$ defined in terms of forecasts for events undominated in Brier score by a rival forecast. We propose a criterion of IP-coherence$_2$ based on a generalization of Brier score for IP-forecasts that uses 1-sided, lower and upper, probability forecasts. However, whereas Brier score is a strictly proper scoring rule for eliciting determinate probabilities, we show that there is no *real-valued* strictly proper IP-score. Nonetheless, with respect to either of two decision rules – Γ-*Maximin* or (Levi's) **E**-a*dmissibility*-+-Γ-*Maximin* – we give a *lexicographic* strictly proper IP-scoring rule that is based on Brier score.

Keywords. Brier score, coherence, dominance, **E**-*admissibility*, Γ-*Maximin,* proper scoring rules.

1. Introduction

Starting in about 1960, de Finetti emphasized two coherence criteria – coherence$_1$ for previsions and coherence$_2$ for forecasts assessed by Brier score. He established [2, 4] that these two criteria are equivalent for purposes of distinguishing between sets of previsions or sets of forecasts that are undominated versus those that are dominated. *Coherence* is the common requirement that a decision maker avoids dominated alternatives. That is, a set of previsions are coherent$_1$ i.e., they are undominated by the alternative of the status-quo – there is no "Book" – if and only if those same quantities, when used as forecasts evaluated by Brier score, are coherent$_2$, i.e., they are undominated by any rival set of forecasts. In his later presentations de Finetti favored coherence$_2$ over coherence$_1$ because, in addition to providing an equivalent criterion for coherence, also proper scores provide a method for incentive compatible elicitation, unlike the situation with coherence$_1$ and the *prevision game*, as we call it. In section 2, we make precise and explain these claims.

De Finetti's theory of coherent previsions, coherence$_1$, serves as the basis for numerous *IP* generalizations – see [7, 18, 19] for examples. However, we know of no parallel development of IP theory based on proper scoring rules. It is our purpose in this essay to report basic findings about scoring-rule based IP theory. In section 3 we explain one approach to an IP version of coherence$_2$. In section 4 we present an impossibility result for a *real-valued* proper IP scoring rule. By contrast, we illustrate a strictly proper, lexicographic (vector-valued) IP version of Brier score. In section 5 we conclude with remarks about the approach begun here.

2. De Finetti's two criteria for coherence

2.1 Coherence$_1$ and coherence$_2$. The *prevision game*, is formulated for a class of bounded variables, $\mathcal{X} = \{X_i: i \in \mathbf{I}\}$ each of which is measurable with respect to a space $\{\Omega, \mathcal{B}\}$, where **I** serves an index set.
One player, the *bookie*, posts a *fair*, or *2-sided* prevision $P(X_i)$ for each $X_i \in \mathcal{X}$. The bookie's opponent, the *gambler*, may choose *finitely many* non-zero real numbers $\{\alpha_i\}$ where, when the state $\omega \in \Omega$ obtains, the bookie's payoff is $\Sigma_i\, \alpha_i(X_i(\omega) - P(X_i))$, and the gambler's payoff is the negative, $-\Sigma_i\, \alpha_i(X_i(\omega) - P(X_i))$. That is, the bookie is obliged either to buy (if $\alpha > 0$), or to sell (if $\alpha < 0$) $|\alpha|$-many units of X at the price, $P(X)$. Hence, the previsions are described as being *2-sided* or *fair* buy/sell prices.

The bookie's previsions are *incoherent$_1$* if the gambler has a strategy that insures a uniformly negative payoff for the bookie, i.e., if there exist a finite set $\{\alpha_i\}$ and $\varepsilon > 0$ such that, for each $\omega \in \Omega$, $\Sigma_i\, \alpha_i(X_i(\omega) - P(X_i)) < -\varepsilon$. Otherwise, the bookie's previsions are *coherent$_1$*.

De Finetti's *Fundamental Theorem of Previsions*:

> The bookie's previsions $\{P(X): X \in \mathcal{X}\}$ are coherent$_1$ if and only if there is a finitely additive probability P whose expected value for X, $\mathbf{E}_P[X]$, is the *bookie*'s prevision:

- Coherence$_1$ *if and only if* $\mathbf{E}_P[X] = P(X)$.

This result extends to include *coherence*$_1$ for conditional expectations given non-null events, using the device of called-off previsions. Let F be an event with $F(\omega)$ its indicator function. The bookie's called-off prevision,

$P_F[X]$, for X given event F has payoff in state ω to the bookie: $F(\omega)\alpha(X(\omega) - P_F(X))$, which equals 0 – the transaction is called-off – in case event F fails. Assuming that the conditioning event is not null, i.e., $P(F) \neq 0$, then

- *Coherence*$_1$ for called-off previsions requires: $\mathbf{E}_P[X \mid F] = P_F[X]$.

When the conditioning event F is null, coherence$_1$ places no substantive constraints on the called-off prevision $P_F[X]$. That is $E_P[F(\omega)\alpha(X(\omega) - P_F(X))] = 0$ regardless the real-value of $P_F[X]$. This defect in de Finetti's formulation has been discussed many times in the literature, and with a variety of different proposals to remedy the situation. For three different corrections to this defect in coherence$_1$ see [8, 10, and 20]. However, the problem with conditioning on null events does not arise for the questions addressed in this essay. So we use de Finetti's version of coherence$_1$.

De Finetti [3] noted that *strategic* aspects of betting may affect elicitation of a bookie's *fair* previsions. For example, when the bookie (believes he/she) knows the gambler's betting odds, then announcing a previsions is subject to strategic play in the game and may fail to reveal the bookie's fair prevision.

Example 1: Suppose the bookie's *fair* (2-sided) prevision for an event G is .50. But suppose the bookie is confident the gambler's fair prevision for G is .75. So the bookie announces $P(G) = .70$, anticipating that the gambler will find it profitable to buy units of G at the inflated price. *Elicitation* using the prevision game fails to identify the bookie's fair price for G. ◊

Aside: There are other issues concerning elicitation in the prevision game. Among these is the challenge of state-dependent utilities [13], which we mention in section 5.

To mitigate strategic aspects of the prevision game, de Finetti turned to a different coherence criterion: probabilistic forecasting subject to Brier score. Hereafter we focus on forecasting events, represented by their indicator functions. $E(\omega) = 1$ if $\omega \in E$ and $E(\omega) = 0$ if $\omega \notin E$.

The bookie's previsions serve as probabilistic forecasts subject to Brier score: squared-error loss. The penalty for the forecast $P(E)$ when $\omega \in \Omega$ is given by two functions $\{g_1, g_0\}$ depending upon the state:

$g_1(P(E), \omega) = (1 - P(E))^2$ if event $\omega \in E$ obtains;
$g_0(P(E), \omega) = (0 - P(E))^2$ if event $\omega \in E^c$ obtains,

which is summarized by the squared-error penalty score

$$(E(\omega) - P(E))^2$$

For the conditional (called-off) forecast $P_F(E)$, on condition that event F obtains, the score is

$$F(\omega)(E(\omega) - P(E))^2.$$

And just as in the prevision game, the score for a finite set of forecasts is the sum of the separate scores.

Definition: A forecast set $\{P(X): X \in \mathcal{X}\}$ is *coherent*$_2$ if, for each finite subset of $\mathcal{X}$, there is no rival forecast set $\{P'(X): X \in \mathcal{X}\}$ whose scores uniformly dominates in Ω.

The two senses of coherence are equivalent, as de Finetti established.

Proposition 1: A set of previsions is coherent$_1$ in the prevision-game *if and only if* those same previsions are a coherent$_2$ set of forecasts under Brier score.

Proof: Here is a geometric version of de Finetti's projection argument for establishing that coherence$_1$ = coherence$_2$ with unconditional previsions/forecasts. We use these ideas in Section 3 to extend coherence$_2$ to an IP setting.

Let $\mathcal{X} = \{X_1, X_2\}$ where X_1 is the indicator for an event A and X_2 is the indicator for the complementary event A^c. In Figure 1, below, a pair of forecasts, $\{Q(A), Q(A^c)\}$ with $0 \leq Q(A), Q(A^c) \leq 1$, is depicted by the point $(Q(A), Q(A^c))$ in the unit square. Note: If either forecast is outside the unit interval, then it is outside the range for the variable being forecasted. And then it is trivial to dominate that forecast with a rival forecast chosen to be closer to the nearest endpoint of the range of the variable in question.

The coherent$_1$ forecasts lie along the reverse diagonal, the simplex on two states, where $Q(A) + Q(A^c) = 1$. No such point is dominated by any other coherent$_1$ forecast, since moving along this line segment increases the distance, and hence increases the squared error relative to one endpoint or the other.

Example 2: Consider, the incoherent$_1$ previsions: $P(A) = .6$ and $P(A^c) = .7$. A *Book* is achieved against these previsions with the gambler's strategy $\alpha_1 = \alpha_2 = 1$. Then the net payoff to the bookie is -0.3 regardless which state ω obtains. In order to see that these are also incoherent$_2$ forecasts, review Figure 1. ◊

If the forecast previsions are not coherent$_1$, they lie outside the probability simplex. Project these incoherent$_1$ forecasts into the simplex. As in Example$_2$, (.60, .70) projects onto the coherent$_1$ previsions depicted by the point (.45, .55). By elementary properties of Euclidean projection, the resulting coherent$_1$ forecasts are closer to each endpoint of the simplex. Thus, the projected forecasts have a dominating Brier score regardless which state obtains. This establishes that the initial forecasts are incoherent$_2$. Since no coherent$_1$ forecast set can be so dominated, we have coherence$_1$ of the previsions if and only coherence$_2$ of the corresponding forecasts.

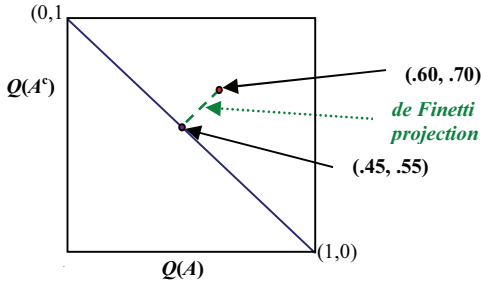


Figure 1

Just as coherence$_1$ fails to regulate called-off previsions given a null event, coherence$_2$ does not regulate called-off forecasts given a null event. See [5] for a parallel revision to coherence$_2$.

2.2 Incentive Compatible Scoring

Brier score is just one of an infinite class of (*strictly*) *proper* scoring rules: A coherent forecaster (uniquely) minimizes expected score by announcing previsions. Thus, forecasting with a (strictly) proper scoring rule avoids the problem of strategic behavior present in the prevision game: there is no opponent. Even allowing different proper scoring rules for different forecasts, by taking the combined score for a finite set of forecasts as the sum of the individual scores, the result is again (strictly) proper. Savage [11] and Schervish [12] characterize the (g_0, g_1) pairs for proper scoring rules. In [14] we establish that all (proper) scoring rules produce the same distinction between coherent$_1$ and incoherent$_1$ forecasts as with Brier score, both for unconditional forecasts and for conditional forecasts given a non-null event.

Proposition 2 [14]:
2.1 When the scoring rule is proper, finite, and continuous, each incoherent$_1$ forecast set is dominated by some coherent$_1$ forecast set.
2.2 When the scoring rule is proper, finite, but *not* continuous, each incoherent$_1$ forecast set is dominated, but not necessarily by a coherent$_1$ forecast set.

Note: Result 2.1 can be established by a generalization of de Finetti's geometric argument, where the projection depends upon the scoring rule. See [9]. The demonstration in [14] uses game-theoretic reasoning.

3. Coherence$_2$ with a Brier *IP* scoring rule.

Recall C.A.B.Smith's [17] modification of de Finetti's prevision game that provides a criterion of IP-coherence$_1$ for (closed, convex) IP sets. Rather than requiring a 2-sided, *fair* price, permit the bookie to fix a pair of 1-sided previsions for each $X \in \mathcal{X}$:

- The bookie announces one rate $\underline{P}(X)$ as a buying price for use when $\alpha > 0$, and a possibly different selling price $\overline{P}(X)$ for use when $\alpha < 0$.

The result is a generalized *Book* argument. See [19, chapter 2] for some history and basic results.

Proposition 3:
(*3.1*) A bookie's 1-sided previsions *avoid sure loss* if and only if there is a maximal, non-empty (closed, convex) set of finitely additive probabilities $\mathcal{P}$ where

$$\underline{P}(X) \leq \textit{infemum}_{P \in \mathcal{P}} \mathbf{E}_P[X]$$

And $$\overline{P}(X) \geq \textit{supremum}_{P \in \mathcal{P}} \mathbf{E}_P[X].$$

When these inequalities are equalities, the 1-sided previsions are said to be *IP-coherent*$_1$.

(*3.2*) By requiring lower and upper previsions for sufficiently many variables (from the linear span of $\mathcal{X}$), the 1-sided previsions avoid sure loss if and only if they are also IP-coherent$_1$. See Theorem 1.ii of [15].

We offer a parallel version for defining IP-coherence$_2$ based on Brier score for 1-sided forecasts, as follows:

Use a *lower forecast* to assess a penalty score when the event forecasted <u>*fails*</u>;

Use an *upper forecast* to assess a penalty score when the event forecasted <u>*obtains*</u>.

Let $\{E_i: i = 1, \ldots, m\}$ be m events defined over a finite partition $\Omega = \{\omega_j: j = 1, \ldots, n\}$. The forecaster gives *lower* and *upper* probability forecasts $\{p_i, q_i\}$ for each event E_i.

<u>*Scoring forecasts with a Brier-styled IP scoring rule*</u>:
Fix a state $\omega \in \Omega$.
If $\omega \in E_i$ the score for the forecast of E_i is

$$(1-q_i)^2 = g_1(q_i, \omega)$$

If $\omega \notin E_i$ the score for the forecast of E_i is

$$p_i^2 = g_0(p_i, \omega)$$

That is, use the most favorable forecast value from the pair $\{p_i, q_i\}$ for determining the score. Just as with the other coherence criteria discussed here, the score for a set of forecasts is the sum of the individual forecast scores.

<u>*Dominance*</u>: A forecast set $\mathcal{G}$ <u>*(strictly) dominates*</u> another $\mathcal{F}$ if, for each $\omega \in \Omega$, the score for $\mathcal{G}$ is (strictly) less than the score for $\mathcal{F}$.

But, since the vacuous $\{0 = p_i, q_i = 1\}$ forecast dominates each rival $\{0 < p_i', q_i' < 1\}$, we require an additional restriction on the class of competing forecasts in order to avoid triviality of the resulting theory of IP-coherence. Aside: This is analogous to a problem that is usually ignored within traditional IP theory. With 1-sided previsions, it remains coherent to be strategic: announce a lower buying (and/or a higher selling) price than one is prepared to accept. That is, knowing who is the *Gambler*

in the 1-sided Prevision Game, the *Bookie* may play strategically and mimic having a less determinate IP-coherent$_1$ set of previsions in order to secure strictly favorable gambles.

We propose that *IP-coherence$_2$* takes into account both a *rival model class M* of coherent$_1$ forecasts and the *relative imprecision* in a forecast set. Stated informally, a set of 1-sided forecasts $\mathcal{F}$ are incoherent$_2$ when:
(i) there exists a dominating set of forecast $\mathcal{G}$ that are
(ii) at least as precise/determinate as $\mathcal{F}$ and
(iii) where $\mathcal{G}$ belongs to the IP-coherent$_1$ model class *M*.
We illustrate this idea by filling in the details of the two concepts: the *rival model class M* and *relative informativeness* between forecast sets.

Example 3*: M* is the ε-*contamination* class. Let *P* be a particular probability distribution over $\Omega = \{\omega_1, \ldots, \omega_n\}$. Fix $0 \leq \varepsilon \leq 1$. Let $\boldsymbol{Q}$ be the simplex of all probability distributions on Ω. The ε-contamination model with focus *P*, $\mathcal{P}_\varepsilon$, is the set of probability distributions on Ω defined by $\mathcal{P}_\varepsilon = \{(1-\varepsilon)P + \varepsilon Q: Q \in \boldsymbol{Q}\}$. For our purposes, it is useful to know that this class is characterized by specifying (IP-coherent$_1$) lower probabilities *for atomic events*, and using the largest closed convex set of distributions satisfying those bounds.◊

In what follows we illustrate one index of *relative indeterminacy* associated with our Brier-styled *IP*-scoring rule.

IP-forecasts over a finite partition for Brier-styled, ε-contamination coherence$_2$:

Let $\mathcal{F} = \{ \{p_i, q_i\}: i = 1, \ldots, n\}$ be forecasts for each state $\omega_i \in \Omega = \{\omega_1, \ldots, \omega_n\}$.

Define $\mathcal{F}$'s *score set* $\mathcal{S}$ by an ordered *n*-tuple of *n*-dimensional points:
$\mathcal{S} = \{(q_1, p_2, \ldots, p_n), (p_1, q_2, \ldots, p_n), \ldots, (p_1, p_2, \ldots, q_n)\}$.
Thus, $\mathcal{S}$ contains at most *n*-many distinct points. Each point in $\mathcal{S}$ has *n*-many coordinates.

Observe that the *IP-Brier-style* score for $\mathcal{F}$ evaluated at state ω_j is the square of the Euclidean distance from the j^{th} point of $\mathcal{S}$ to the j^{th} corner of the probability simplex on Ω. Clearly, the *IP*-score for a forecast set can be improved merely by moving a lower forecast closer to 0, or by moving an upper forecast closer to 1. So, consider dominating forecast sets only when the dominating forecast has a score set that is *less indeterminate* than the score set for the dominated forecast. Here is a candidate for *relative indeterminacy* which, when combined with our Brier-style IP-score, allows a characterization of *ε-contamination* IP-coherence$_2$.

Definition: Forecast set $\mathcal{F}_2$ is *at least as indeterminate as* forecast set $\mathcal{F}_1$ (or $\mathcal{F}_1$ is *at least as determinate as* $\mathcal{F}_2$) if the convex hull of score set $\mathcal{S}_1$, $\boldsymbol{H}(\mathcal{S}_1)$, is isomorphic under rigid movements (where both shape and sized are held fixed) to a subset of the convex hull of score set $\mathcal{S}_2$, $\boldsymbol{H}(\mathcal{S}_2)$.

Note that this relation of *relative imprecision*, or *relative indeterminacy*, is merely a partial order. We opt for such a concept so that relative indeterminacy may be extended to a variety of different real-valued indices of imprecision, e.g., by using generalized volume of the score set to quantify indeterminacy.

We use these notions to define IP-coherence$_2$ generally, and then continue with our illustration of IP-coherence$_2$ with respect to the ε-contamination model.
Definition: Given an *IP*-scoring rule, *a* set $\mathcal{F}$ of IP-forecasts is *IP-incoherent$_2$ with respect to the IP-model M* provided that there is a dominating set of rival forecasts $\mathcal{G}$ from the model *M* where the set $\mathcal{G}$ is at least as determinate than the set $\mathcal{F}$. Say that $\mathcal{F}$ is IP-coherent$_2$ with respect to *M* if it is not IP-incoherent$_2$ with respect to *M*. For convenience we will write these as *M*-coherent$_2$ and *M*-incoherent$_2$

Observe that IP-incoherence$_2$ reduces to de Finetti's incoherence$_2$ when all forecasts in $\mathcal{F}$ are determinate, i.e., when $p_i = q_i$ for each forecasted event E_i ($i \in \mathbf{I}$), and when *M* is the class of determinate, coherent$_1$ forecasts. To see this, assume that $|\Omega| = k$. Then the score set $\mathcal{S}$ is the ordered set with *k*-many repetitions of the same $|\mathbf{I}|$-dimensional point. Since the lower and upper $\mathcal{F}$ forecasts for an event are identical, the *k*-many points in $\mathcal{S}$ do not vary with ω. So a dominating rival forecast set $\mathcal{G} = \{p'_i, q'_i\}$ must also assign the same lower and upper values to each event E_i (that is, for each $i \in \mathbf{I}$, $p'_i = q'_i\}$, in order for $\mathcal{G}$ to be at least as determinate as $\mathcal{F}$. By *Proposition* 2.1, then if $\mathcal{G}$ dominates $\mathcal{F}$ the rival forecast set $\{p_i'\}$ establish that $\mathcal{F}$ is incoherent$_2$ and incoherent$_1$.

Next, we provide two basic results for IP-coherence$_2$ with respect to the ε-contamination model.

Proposition 4: Let $0 \leq p_i \leq q_i \leq 1$, with *n*-many forecasts $\mathcal{F}$ solely *for atoms* in a finite algebra $\Omega = \{\omega_1, \ldots, \omega_n\}$.

(*4.1*) The score set $\mathcal{S}$ for $\mathcal{F}$ lies entirely within the probability simplex on Ω *if and only if* the lower and upper forecasts $\mathcal{F}$ match an ε-*contamination* model. And then $\mathcal{F}$ cannot be dominated by rival forecasts from a more determinate ε-contamination model.

(*4.2*) If all the elements of a score set $\mathcal{S}$, associated with forecast set $\mathcal{F}$, lie outside the probability simplex on Ω, there is a dominating ε-contamination forecast model $\mathcal{F}^*$ with greater determinacy than $\mathcal{F}$. $\mathcal{F}$ is IP-incoherent$_2$ against rivals from the ε-contamination model.

Proof:
(*4.1*) is established by elementary calculations. If and only if each point of the score set $\mathcal{S}$ belongs to the probability simplex then, when state ω_j obtains, corresponding to the j^{th} point of $\mathcal{S}$, $1 = q_j + \sum_{i \neq j} p_i$, and this equality obtains for each $j = 1, \ldots, n$. Then there exists an $\varepsilon \geq 0$ such that for each $i = 1, \ldots, n$, $q_i = p_i + \varepsilon$, which defines an ε-contamination model. In the opposite direction, if forecasts for the atoms are based on an ε-contamination model, for $i = 1, \ldots, n$, $q_i = p_i + \varepsilon$, and then $1 = q_j + \sum_{i \neq j} p_i$ so that all of the score set $\mathcal{S}$ lies in the probability simplex.

Last, if $\mathcal{S}$ belongs to the probability simplex and a rival ε-contamination model $\mathcal{F}'$ (with corresponding score set $\mathcal{S}'$) dominates, then $\boldsymbol{H}(\mathcal{S})$ is a proper subset of $\boldsymbol{H}(\mathcal{S}')$ because for each $j = 1, \ldots, n$, the j^{th} point of $\mathcal{S}'$ is closer to the j^{th} extreme point of the probability simplex than is the j^{th} point of $\mathcal{S}$. So, $\mathcal{F}'$ is *less* determinate than $\mathcal{F}$. Thus $\mathcal{F}$ is IP-coherent$_2$ with respect to the ε-contamination model.

(*4.2*) follows by the *Brouwer Fixed-Point* Theorem. Begin with a forecast set $\mathcal{F} = \mathcal{F}_0$, whose score set $\mathcal{S}_0$ has each of its n-many ordered points outside the simplex of coherent$_1$ forecasts. Recursively create rival forecast sets as follow. Apply the (de Finetti) projection to each of these n-many ordered points of $\mathcal{S}_0$ taking them into the probability simplex of coherent$_1$ forecasts. This creates (at most) n-points $\boldsymbol{T}_1 = \{t_1, \ldots, t_n\}$ where each $t \in \boldsymbol{T}_1$ is a probability distribution $P(\bullet)$ over Ω. Form the new forecast set $\mathcal{F}_1 = \{\{p_{1i}, q_{1i}\}: i = 1, \ldots, n\}$ where $p_{1i} = min_{t \in \boldsymbol{T}^1}\{P(\omega_i)\}$ and $q_{1i} = max_{t \in \boldsymbol{T}^1}\{P(\omega_i)\}$. This determines a new score set $\mathcal{S}_1$. Since none of the points in $\mathcal{S}_0$ belongs to the probability simplex, by the same reasoning used in de Finetti's analysis for *Proposition* 1, $\mathcal{F}_1$ dominates $\mathcal{F}_0$.

Just in case $\mathcal{S}_1$ lies in the simplex, when result (4.1) applies, the recursive procedure halts. Otherwise forecast set $\mathcal{F}_2$ is created from a projection of score set $\mathcal{S}_1$ into the probability simplex, etc. (See Appendix 2 for an illustration.)

Since Euclidean projections are continuous functions and the probability simplex is compact, the recursive process with forecast sets $\mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2, \ldots$. has a fixed point $\mathcal{F}^*$ in the class of ε-contamination models. By a simple adaptation of de Finetti's argument for *Proposition* 1, the forecast set $\mathcal{F}_{i+1}$ (weakly) dominates the forecast set $\mathcal{F}_i$ unless $\mathcal{F}_i$ is a fixed point of the process.
Note: It may be that $\mathcal{F}_{i+1}$ merely weakly dominates $\mathcal{F}_i$ for $i \geq 1$, since some but not all the points in $\mathcal{S}_1$ may lie in the probability simplex. However, since all the points of $\mathcal{S}_0$ lie outside the probability simplex, $\mathcal{F}_1$ dominates $\mathcal{F}_0$.

Last, the projection of a closed, convex set, e.g., the projection of $\boldsymbol{H}(\mathcal{S})$ into the probability simplex, is isomorphic to a subset of $\boldsymbol{H}(\mathcal{S})$. Thus, assuming that the each of the points of $\mathcal{S}_0$ is outside the probability simplex on Ω, the fixed point $\mathcal{F}^*$ of the process $\mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2, \ldots$, which belongs to the ε-contamination model class, strictly dominates $\mathcal{F}_0$, and is at least as determinate as $\mathcal{F}_0$. Hence, $\mathcal{F}_0$ is IP-incoherent$_2$ with respect to the ε-contamination class.

Example$_4$: Here is an illustration of *Proposition* 4, IP-coherence$_2$ with respect to the ε-contamination model, using 5 different forecast sets. Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$. Forecasts are for the three atoms only. The five forecast sets $\mathcal{F}^j$ ($j = 1, \ldots, 5$) are given in the form $\{\{p_i, q_i\}$ for ω_i: $i = 1, 2, 3\}$. The respective score sets have three points with coordinates $\{(q_1, p_2, p_3), (p_1, q_2, p_3), (p_1, p_2, q_3)\}$, as described above. Figure 2 diagrams the convex hull of each score set and shows the shaded 2-dimensional, triangular simplex of probability functions on Ω.

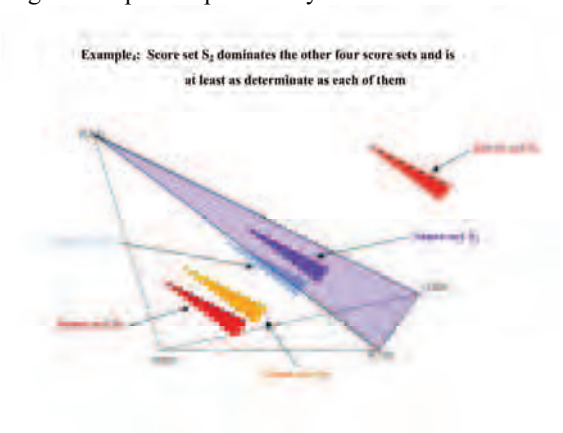


Figure 2 (for *Example* 4)
The convex hull of the five score sets are color coded. The simplex of probability distributions is shaded. Each score set projects onto $\mathcal{S}^2$, the score set for forecast set $\mathcal{F}^2$, corresponding to an ε-contamination model.

$\mathcal{F}^1 = \{ \{.55, .80\}, \{.55, .80\}, \{.55, .80\}\}$
$\mathcal{S}^1 = \{(.80, .55, .55), (.55, .80, .55), (.55, .55, .80)\}$

$\mathcal{F}^2 = \{ \{.25, .50\}, \{.25, .50\}, \{.25, .50\}\}$
$\mathcal{S}^2 = \{(.50, .25, .25), (.25, .50, .25), (.25, .25, .50)\}$

$\mathcal{F}^3 = \{ \{.20, .45\}, \{.20, .45\}, \{.20, .45\}\}$
$\mathcal{S}^3 = \{(.45, .20, .20), (.20, .45, .20), (.20, .20, .45)\}$

$\mathcal{F}^4 = \{ \{.10, .35\}, \{.10, .35\}, \{.10, .35\}\}$
$\mathcal{S}^4 = \{(.35, .10, .10), (.10, .35, .10), (.10, .10, .35)\}$

$\mathcal{F}^5 = \{ \{.05, .30\}, \{.05, .30\}, \{.05, .30\}\}$
$\mathcal{S}^5 = \{(.30, .05, .05), (.05, .30, .05), (.05, .05, .30)\}$

The two forecast sets $\mathcal{F}^1$ and $\mathcal{F}^5$ are IP-incoherent$_1$ in accord with *Proposition* 3. Their 1-sided previsions lead to sure losses as, respectively, their lower (upper) forecasts are too great (too small). There is no determinate probability distribution agreeing with either set's lower and upper forecasts.

Forecast set $\mathcal{F}^2$ corresponds to an ε-contamination model with focus the uniform probability P = (1/3, 1/3, 1/3) and ε = 1/6. The convex hull of the score set $\mathcal{S}^2$ lies in the probability simplex, as per *Proposition* (4.1). It is IP-coherent$_1$ and IP-coherent$_2$ with respect to the ε-contamination model class.

Forecast set $\mathcal{F}^3$ is IP-coherent$_1$ as it has lower and upper forecasts agreeing with a closed convex set of probabilities. Those values agree with an *ALUP* model, but not with an ε-contamination model. That is, $\mathcal{F}^3$ is IP-coherent$_2$ with respect to an IP-model class defined by specifying **a**tomic **l**ower and **u**pper **p**robabilities [ALUP], but not so with respect to the ε-contamination class, which is an IP-model class determined solely by atomic lower probabilities. (See Appendix 1 for details.)

Forecast set $\mathcal{F}^4$ has lower and upper forecasts that do not match those from a closed convex set of probabilities. Its intervals are too wide. However, the uniform probability agrees with these forecasts, i.e., the probability values (1/3, 1/3, 1/3) fall inside the forecast intervals from $\mathcal{F}^4$. Thus, in accord with *Proposition* 3, the forecasts from $\mathcal{F}^4$ do not suffer a sure-loss in the 1-sided prevision game; however, $\mathcal{F}^4$ is IP-incoherent$_1$ and IP-incoherent$_2$ with respect to the ε-contamination model class.

As indicated by Figure 2, each of the other four convex hulls projects to $\boldsymbol{H}(\mathcal{S}^2)$. That is, the process described in the proof of Proposition (4.2) has $\mathcal{F}^2$ as its fixed point for each of the five forecast sets, and the process terminates after (at most) one projection.◊
See Appendix 2 for an illustration of Proposition (4.2) where the fixed point is merely a limit of the process.

4. Incentive compatible IP-elicitation

Recall that de Finetti favored coherence$_2$ over coherence$_1$ because, in addition to serving as an equivalent criterion of coherence, Brier score provides a strictly proper score. It provides incentive compatible elicitation for determinate probabilities. For a forecaster whose degrees of belief about events are represented by a single probability function $P(\bullet)$ and who maximizes expected utility, she/he has a unique strategy for announcing forecasts (and called-off forecasts) that minimize expected Brier score. Announce the probability $P(E)$ for the forecast of event E. If H is not-null, then announce the conditional probability $P(E\,|H)$ for the called-off forecast of event E, on condition that H obtains. Recall that when H is null, coherence$_2$ places no restrictions on the called-off forecasts given H. There is no difference to the expected score contributed by any conditional forecast of E, called-off if H fails, regardless whether that forecast is or is not coherent$_2$. See [5] for an improved version of coherence$_2$.

What can be done to extend Brier score to an incentive compatible IP-scoring rule? The question is ill-formed without a decision rule that extends maximizing expected utility to IP contexts. We consider only decision rules that reduce to the rule of maximizing expected utility when those IP sets collapse onto the special case of a singleton set, where upper and lower probabilities are identical and a single probability distribution represents uncertainty. Also, we require that decision rules respect the following weak form admissibility. Let $\boldsymbol{S}(\mathcal{F}, \omega)$ be a real-valued IP-scoring rule for forecast set $\mathcal{F}$ in state ω. Recall that scores are given in the form of a loss so that smaller is better.

> *Admissibility Principle*: If for each $\omega \in \Omega$ $\boldsymbol{S}(\mathcal{F}, \omega) \leq \boldsymbol{S}(\mathcal{F}', \omega)$, then $\mathcal{F}$ is admissible in a pairwise choice between rival forecasts $\mathcal{F}$ and $\mathcal{F}'$. Moreover, if for each ω this inequality is strict then $\mathcal{F}'$ is inadmissible whenever $\mathcal{F}$ is an option.

In this section we report two results about eliciting upper and lower probabilities for events when the forecaster's opinion is represented by a closed, convex sets of probabilities on a finite state space.

Proposition 5: There is no *real-valued* (strictly) proper IP continuous scoring rule.

By contrast, however,
Proposition 6: Under either the Γ-*Maximin* decision rule, or using one of Levi's [8] lexicographic decision rules – **E**-*admissibility* followed by Γ-*Maximin* security – there is a strictly proper *lexicographic* IP-Brier scoring rule.

The IP-decision rules we investigate in Proposition 6 are summarized as follows, with details given in Section 4.2:
Γ-*Maximin*: The admissible options in **D** are those that maximize their lower expected value.
E-*admissibility*: An option $X \in \mathbf{D}$ is **E**-*admissible* if for some $P \in \mathcal{P}$ and each $Y \in \mathbf{D}$, $\mathbf{E}_P[X] \geq \mathbf{E}_P[Y]$.
E-*admissibility-followed-by*-Γ-*Maximin*: Apply Γ-*Maximin* to the set of **E**-*admissible* options in **D**.

Next, we establish and explain these findings.

4.1 Proof of *Proposition* 5 The impossibility reported in this result is made evident by considering the demands on a real-valued strictly proper IP-scoring rule $\boldsymbol{S}(\mathcal{F}', \omega)$, for forecasting one event, E.

Let the interval $[p, q]$, $0 \le p \le q \le 1$, represent the forecaster's uncertainty for E. In general, the IP-scoring rule may be written

$g_1([p, q], \omega)$ if $\omega \in E$ obtains,
and $g_0([p, q], \omega)$ if $\omega \in E^{\mathbf{c}}$ obtains.

When $p = q$, in order to be strictly proper and real-valued, the scoring rule must satisfy Theorem 4.2 of Schervish [12]. Specifically, with $0 \le x \le 1$, the loss for the point forecast $\boldsymbol{S}([x, x], \omega)$, x satisfies

$$g_1(x) = g_1(1) + \int_x^1 (1-q)\lambda(dq) \quad \text{if } \omega \in E \text{ obtains;}$$
$$g_0(x) = g_0(0) + \int_0^x q\lambda(dq) \quad \text{if } \omega \in E^{\mathbf{c}} \text{ obtains,}$$

where $g_1(1)$ and $g_0(0)$ are finite, and $\lambda(dq)$ is a measure on $[0, 1)$ that gives positive measure to every non-degenerate interval. Continuity of the scoring rule results from a continuous measure λ with no point masses. For example, Brier score results by letting λ have the constant density 2 on the unit interval.

When $p < q$, the impossibility of a strictly proper IP-scoring rule is a consequence of the fact that, since λ is positive on non-degenerate sub-intervals of the unit interval [0,1] and continuous, there will be rival interval forecasts $[p, q]$ and $[p', q']$ with

$g_1([p, q]) - g_1([p', q']) \ge 0$,
and $g_0([p, q]) - g_0([p', q']) \ge 0$.

Then the interval forecast $[p', q']$ is admissible against the rival interval forecast $[p, q]$. When the interval $[p, q]$ is the forecaster's IP-uncertainty for event E, she/he will not have reason to announce that as her/his forecast rather than the rival forecast $[p', q']$ and the IP-scoring rule is not strictly proper. If for each ω the inequality is strict, then the IP-scoring rule is not proper.

Example 5. We illustrate *Proposition* 5 using the ideas about IP-coherence$_2$ presented in section 4. Consider Brier score adapted to a forecast interval $[p, q]$. That is, let $\boldsymbol{b}([p,q], \omega) = g_1([p, q], \omega) = (1-q)^2$ if $\omega \in E$,
and $\boldsymbol{b}([p,q], \omega) = g_0([p, q], \omega) = p^2$ if $\omega \in E^{\mathbf{c}}$.
Introduce a *real-valued* index of indeterminacy for a forecast set $\mathcal{F}$, $\boldsymbol{I}(\mathcal{F})$, where $\boldsymbol{I}$ agrees with the partial order of relative imprecision used to define IP-coherence$_2$. For instance, let $\boldsymbol{I}([p, q]) = q-p$. For real values x, y, let $\boldsymbol{H}(x,y)$ be a real-valued function increasing in each of its arguments, e.g., $\boldsymbol{H}(x,y) = x + y$. Define an IP-Brier score for forecast set $\mathcal{F}$ by $\boldsymbol{B}(\mathcal{F},\omega) = \boldsymbol{H}(\boldsymbol{b}(\mathcal{F},\omega), \boldsymbol{I}(\mathcal{F}))$. Then by *Proposition* 5, $\boldsymbol{B}$ is an improper-IP scoring rule. To complete the example, consider event E and compare the two interval forecasts [.25, .75] and [.50, .50]. Then

$\boldsymbol{B}([.25, .75], \omega) = 1/16 + 1/2 = 9/16$
and $\boldsymbol{B}([.50, .50], \omega) = 1/4 + 0 = 1/4$.

Hence, the interval forecast [.25, .75] is inadmissible under this IP-Brier scoring rule $\boldsymbol{B}$.◊

4.2 Proof of *Proposition* 6 First we review the two decision rules mentioned in the result. Let $\mathcal{P}$ be a closed, convex set of probabilities P on the space $\{\Omega, \mathcal{B}\}$. Let χ be the class of bounded random variables, X, each measurable with respect to this space. For each X, write $\underline{X}$ for the *infemum* over $\mathcal{P}$ of the expected value of X,

$$\underline{X} = \inf_{P \in \mathcal{P}} \mathbf{E}_P[X],$$

which identifies the *lower expected value* for X with respect to $\mathcal{P}$. Identify a decision problem, **D**, with a closed subset of χ. That is, the options in a decision problem form a closed set of bounded variables.

The two IP-decision rules we investigate in Proposition 6 are defined as follows:

Γ-*Maximin*: The admissible options in **D** are those that maximize their lower expected value.

Note: By making both $\mathcal{P}$ and **D** closed sets, this *max-min* operation is well defined.

E-*admissibility*: An option $X \in \mathbf{D}$ is **E**-*admissible* if for some $P \in \mathcal{P}$ and each $Y \in \mathbf{D}$, $\mathbf{E}_P[X] \ge \mathbf{E}_P[Y]$.

E-*admissibility-followed-by*-Γ-*Maximin*: Apply Γ-*Maximin* to the set of **E**-*admissible* options in **D**.

In general, these decision rules have very different axiomatic characterizations. Γ-*Maximin* is represented by a real-valued ordering of χ using $\underline{X}$-values to index each option. But that ordering violates the independence axiom for preferences. **E**-*admissibility* is not represented by an ordering. In fact, it does not even reduce to pairwise comparisons. (See [16] for related discussion.) Nonetheless, next we construct a lexicographic IP-Brier score that is strictly proper under either of the two decision rules mentioned in *Proposition* 6.

Proposition 5 precludes a proper IP-scoring rule that elicits both endpoint of the interval forecast $[p,q]$ for event E. However, we may elicit either endpoint alone. Define the *lower-Brier scoring rule*, $\underline{\mathbf{b}}([x,y], \omega) = \underline{\mathbf{b}}(x,\omega)$ as:

$\underline{g}_1(x) = (1-x)^2$ if $\omega \in E$
$\underline{g}_0(x) = 1 + x^2$ if $\omega \in E^{\mathbf{c}}$.

and the *upper-Brier scoring rule*, $\overline{\mathbf{b}}([x,y], \omega) = \overline{\mathbf{b}}(y,\omega)$ as:

$\overline{g}_1(y) = (1-y)^2 + 1$ if $\omega \in E$
$\overline{g}_0(x) = x^2$ if $\omega \in E^{\mathbf{c}}$.

Each of these is a strictly proper scoring rule for eliciting determinate forecasts. This follows immediately from Schervish's representation (above,) where $\underline{g}_1(1) = \overline{g}_0(0) = 0$, $\underline{g}_1(0) = \overline{g}_1(1) = 1$, and $\lambda = 2$ is the uniform (Brier) score density for both rules.

Lemma 1: Under the Γ-*Maximin* decision rule, respectively, the lower- (upper-) Brier score is strictly proper for the lower (upper) endpoint of the IP-forecast $[p,q]$ of event E.

Proof of Lemma 1: We give the argument for the lower-Brier score. The reasoning for the upper-Brier score is

similar. Let $p = min_{P\in\mathcal{P}}\, P[E]$ and $q = max_{P\in\mathcal{P}}\, P[E]$, so that $\forall P \in \mathcal{P}\ \ p \le P(E) \le q$, and these bounds are tight. The lower-Brier score of the forecast $[r, s]$ for E depends solely on r. The P-Expected score for forecast [r,s] is:

$$\mathbf{E}_P[\underline{\mathbf{b}}[r,s]] = P(E)(1-r)^2 + (1-P(E))(1+r^2)$$
$$= (1-r)^2 + 2r(1-P(E)).$$

By simple dominance, $0 \le r \le 1$. For a given forecast r, this expected penalty score is greatest at $P(E) = p$, when the expected score is $(1-r)^2 + 2r(1-p)$. But since lower-Brier score is strictly proper, this worst value is best, i.e., the worst of these expected scores is smallest uniquely for a forecast with $r = p$. Lemma 1

Lemma 2: Under the **E**-*admissibility-followed-by*-Γ-*Maximin* decision rule, respectively, the lower- (upper-) Brier score is strictly proper for the lower (upper) endpoint of the IP-forecast $[p,q]$ of event E.
Proof of Lemma 2: Again, we give the argument only for the lower-Brier score. Since lower-Brier score is a strictly proper scoring rule for determinate forecasts, the **E**-*admissible* forecasts are those of the form $[r, s]$ where $p \le r \le q$. Then, by *Lemma* 1, the Γ-*Maximin* solution from this set is uniquely solved at $r = p$. Lemma 2

By *Proposition* 5, unfortunately, the real-valued composite score obtained by adding together these two scores, $\overline{\underline{\mathbf{b}}}([r,s]) = \underline{\mathbf{b}}([r,s]) + \overline{\mathbf{b}}([r,s]$, is not IP-proper, which we illustrate with the following example.

Example 6: We illustrate the impropriety of the real-valued IP-score, $\overline{\underline{\mathbf{b}}}([r,s])$, in accord with *Proposition* 5. Consider an extreme case where the forecaster is maximally uncertain of event E, so that the vacuous probability interval [0, 1] represents her/his uncertainty. The forecast [.5, .5] has constant $\overline{\underline{\mathbf{b}}}$-score, i.e.,

$$\overline{\underline{\mathbf{b}}}([.5, .5], \omega) = 1 + ¼ + ¼ = 1.5,$$

independent of ω.
The straightforward forecast [0,1] has the constant score

$$\overline{\underline{\mathbf{b}}}([0, 1], \omega) = 1+1 = 2,$$

independent of ω. So forecast [.5, .5] strictly dominates forecast [0,1] under the $\overline{\underline{\mathbf{b}}}$-scoring rule.◊

Therefore, we use a 2-tier *lexicographical* composite scoring to combine these two rules in a manner that create a strictly proper IP-Brier score.
Definition: The two-tier, lexicographic IP-Brier score for the interval forecast $[p, q]$ of event E, which we write as $\mathbf{b}_{LU}([r,s])$, is the 2-tier lexicographic loss function

$$\mathbf{b}_{LU}([r,s], \omega) = \langle \underline{\mathbf{b}}([r,s], \omega), \overline{\mathbf{b}}([r,s], \omega) \rangle.$$

That is, lexicographically, first apply the loss function $\underline{\mathbf{b}}([r,s])$, and among those forecasts have equal $\underline{\mathbf{b}}$-value, then apply the $\overline{\mathbf{b}}([r,s])$ loss function. By the preceding two lemmas, under the two decision rules named in Proposition 6, only the interval $[p,q]$ is $\mathbf{b}_{LU}$-optimal for forecasting event E when the forecaster's uncertainty for that event is the IP-interval $[p,q]$.

Aside: It is evident that the order of the components is irrelevant in this 2-tiered, lexicographic IP-Brier score.

To elicit an IP-forecast set $\mathcal{F} = \{ \{p_i, q_i\}: i = 1, \ldots, n\}$ for the events $\{E_1, E_2, \ldots, E_n\}$ use, e.g., the $2n$ tiered lexicographic IP-Brier score

$$\langle \underline{\mathbf{b}}_1([r_1,s_1]), \overline{\mathbf{b}}_1([r_1,s_1]), \ldots, \underline{\mathbf{b}}_n([r_n,s_n]), \overline{\mathbf{b}}_n([r_n,s_n]) \rangle.$$

Then the following is immediate from *Proposition* 6.

Corollary. The $2n$-tiered, lexicographic IP-Brier score is strictly proper under either the Γ-*Maximin* or **E**-*admissibility*-followed-by-Γ-*Maximin* decision rules. As above, the order of the $2n$-terms is irrelevant.

5. Summary

When coherence$_1$ of 2-sided previsions is not enough, and elicitation also matters, then Brier score offers an incentive compatible scoring rule with an equivalent coherence criterion: coherence$_2$ – avoid dominated forecasts. This is de Finetti's analysis, *Proposition* 1.

We extend Brier scoring to IP-coherence$_2$ of interval-valued forecasts, analogous to the familiar use of 1-sided (*lower* and *upper*) previsions for defining IP-coherence$_1$. Subject to an IP-scoring rule for forecasting events, the coherent forecaster gives lower and upper probabilistic forecasts for a particular set of events that characterize elements of an IP-model class M – e.g., the ε-*contamination* class is characterized by IP-forecasts for the atoms of the measure space – *Proposition* 4. Coherence$_2$ of the set of IP-forecasts requires that these lower and upper forecasts are not dominated by any *more determinate IP* model within the model class M, subject to the same *IP* scoring rule.

However, a distinguishing feature between coherence$_1$ and coherence$_2$, namely that Brier score is incentive compatible for elicitation of 2-sided (real-valued) forecasts for events, does not extend to 1-sided forecasts. That is, according to *Proposition* 5, there is no strictly proper, real-valued IP-scoring rule for events. However, by relaxing the conditions on scoring rules to permit lexicographic utility, subject to either of two IP-decision rules, there do exist strictly proper IP-scoring rules for eliciting closed, interval-valued probability forecasts.

There are numerous open questions relating to the preliminary work reported in this paper. We list three topics on which we are currently at work.

1) A different challenge to elicitation, even when probability is determinate, is the problem posed by state-dependent utilities. This arises in the choice of the

numeraire that is to be used, either with outcomes of previsions for coherence$_1$, or in scoring forecasts for coherence$_2$. (See [13] for discussion of the problem in the setting of coherence$_1$.)

Does forecasting afford any advantage over betting in this context and is there a difference also with IP-elicitation?

2) As noted in Section 2, neither coherence$_1$ nor coherence$_2$ constrains, respectively, a called-off prevision for an event or a called-off forecast for an event, given a null event. However, lexicographic expected utility [8] is one approach among several others available [5, 10, 20] for improving the treatment of 2-sided conditional probability with called-off previsions given a null event. (See [1] for a review of some of the open issues.) *Proposition* 6 relies on a lexicographic scoring rule to establish propriety with respect to interval valued forecasts.

Can we use lexicographic scoring rules also to elicit called-off forecasts given a null event?

3) De Finetti's theory of coherence is designed to accommodate all finitely additive probabilities. That is, countable additivity is not a requirement of coherence$_1$ or coherence$_2$. This is achieved by insisting that incoherence, i.e., a failure of simple dominance, is achieved using only finitely many previsions or only finitely many forecasts at one time. In other words, a coherent set of previsions or forecasts may be dominated when more than finitely many are combined at once, even though they cannot be dominated when only finitely many are combined. It is interesting, we find, that even with determinate probabilities, coherence$_1$ and coherence$_2$ are not equivalent in this regard. There are settings where countably many coherent$_2$ forecasts may be combined and remain undominated by all rival forecasts, though these same previsions may result in a sure-loss when countably many are combined into a single option [17].

In order to accommodate all finitely additive probabilities, when does IP-coherence$_2$ depend upon the restriction that violations of dominance matter only when finitely many forecasts are scored at the same time?

Acknowledgements

Earlier versions of these results were presented at the University of Warwick's *Subjective Bayes Workshop*, the *Purdue Wimer Memorial Lectures* workshop, and CMU's *Games and Decisions* discussion group, and we thank the participants at these meetings for their helpful comments. In particular we appreciate suggestions from Timos Athanasiou, Luca Rigotti, and Kevin Zollman.

Appendix 1

The Atomic Lower-Upper Probability [*ALUP*] class. This IP-class consists of closed, convex sets of probabilities defined by lower and upper probabilities for atomic events. That is an ALUP model is the largest (closed) convex set of distributions that satisfy such bounds, where the bounds are achieved by the lower and upper probability values given for the atoms of the space. See [6] for discussion about this IP-class of models.

IP-coherence$_2$, where rival forecasts are taken from the ALUP class, arises when the forecaster is called upon to give lower-and-upper forecasts for each atom, ω, and *for the complement to each atom*, ω^c, in the space. That is, in order to duplicate Proposition 4 for the ALUP class the forecaster is called upon to give $2n$-many forecasts when $\Omega = \{\omega_1, \ldots, \omega_n\}$. Example 7 illustrates this.

Example 7 (a continuation of Example 4): An illustration of ALUP-coherence$_2$. We provide 3 forecast sets for the atoms, and the their complements in a space defined by $\Omega = \{\omega_1, \omega_2, \omega_3\}$. That is, each forecast set includes IP-forecasts for 6 events. Forecast sets $\mathcal{F}^j$ ($j = 2, 3, 4$) are given as 6 pairs: $\{p_i, q_i\}$ for ω_i, ω_i^c $i = 1, 2, 3$. Each of the corresponding 3 score sets is comprised by 3 points, corresponding to the 3 states in Ω. Each point in a score set has 6 coordinates, corresponding to the scores for forecasts of $(\omega_1, \omega_1^c, \omega_2, \omega_2^c, \omega_3, \omega_3^c)$.

$\mathcal{F}^2 =$

ω_1	ω_1^c	ω_2	ω_2^c	ω_3	ω_3^c
{.25, .50}	{.50, .75}	{.25, .50}	{.50, .75}	{.25, .50}	{.50, .75}}

$\mathcal{S}^2$ = (.50, .50, .25, .75, .25, .75) for ω_1
(.25, .75, .50, .50, .25, .75) for ω_2
(.25, .75, .25, .75, .50, .50) for ω_2

$\mathcal{F}^3 =$

ω_1	ω_1^c	ω_2	ω_2^c	ω_3	ω_3^c
{ .20, .45}	{.55, .80}	{.20, .45}	{.55, .80}	{.20, .45}	{.55, .80} }

$\mathcal{S}^3$ = (.45, .55, .20, .80, .20, .80) for ω_1
(.20, .80, .45, .55, .20, .80) for ω_2
(.20, ,80, .20, .80 .45, .55)} for ω_3

$\mathcal{F}^4 =$

ω_1	ω_1^c	ω_2	ω_2^c	ω_3	ω_3^c
{ .10, .35}	{.65, .90}	{.10, .35}	{.65, .90}	{.10, .35}	{.65, .90} }

$\mathcal{S}^4$ = (.35, .65, .10, .90, .10, .90) for ω_1
(.10, .90 .35, .65, .10, .90) for ω_2
(.10, .90, .10, .90, .35, .65)} for ω_3

Forecast sets $\mathcal{F}^2$ and $\mathcal{F}^3$ are ALUP-*coherent*. There do not exist more precise forecast sets from the ALUP-model that dominate either of these sets of forecasts. Their score sets lie in the probability simplex for these 6 events.

Forecast set $\mathcal{F}^4$ is ALUP-incoherent. A de Finetti projection of $\mathcal{S}^4$ produces a more determinate rival ALUP forecast with dominating IP Brier score. In fact, the projection produces a more informative ε-*contamination* model that dominates. The respective IP-Brier scores for $\mathcal{F}^4$ and for $\mathcal{F}^2$ are independent of

ω: For $\mathcal{F}^1$ the score is a constant penalty of 0.885. For $\mathcal{F}^2$ it is a constant penalty of 0.750.

Appendix 2

Example 8 – This construction provides a more complicated illustration of *Proposition* 4 where the fixed point $\mathcal{F}^*$ of the process is a limit of the recursive procedure given in the proof of (4.2). Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$. Forecast sets $\mathcal{F}_j$ are of the form $\{\{p_i, q_i\}$: for events ω_i: $i = 1, 2, 3\}$.

$\mathcal{F} = \mathcal{F}_0 = \{ \{.25, .60\}, \{.20, .50\}, \{.10, .40\} \}$

$\mathcal{S} = \mathcal{S}_0 = \{(.60, .20, .10), (.25, .50, .10), (.25, .20, .40)\}$

(Step 1) Project score set $\mathcal{S}_0$ to form set

$\boldsymbol{T}_1 = \{ (.6\overline{3}\,, .2\overline{3}\,,, .1\overline{3}\,,), (.30, .55, .15), (.30, .25, .45)\}$

Form the new forecast and score sets $\mathcal{F}_1$, $\mathcal{S}_1$ based on the probabilities in set $\boldsymbol{T}_1$

$\mathcal{F}_1 = \{ \{.30, .6\overline{3}\,\} \{.2\overline{3}\,, .55\} \{.1\overline{3}\,, .45\} \}$

$\mathcal{S}_1 = \{(.6\overline{3}\;\; .2\overline{3}\;\; .1\overline{3}\,) (.30, .55, .1\overline{3}\,) (.30, .2\overline{3}\,, .45)\}$

(Step 2) Project set $\mathcal{S}_1$ to form set

$\boldsymbol{T}_2 = \{ (.6\overline{3}\,, .2\overline{3}\,, .1\overline{3}\,) (.30\overline{5}\,, .5\overline{5}\,, .1\overline{5}\,) (.30\overline{5}\,, .2\overline{5}\,, .4\overline{5}\,)\}$

Form the new forecast and score sets $\mathcal{F}_2$, $\mathcal{S}_2$ based on the probabilities in set $\boldsymbol{T}_2$

$\mathcal{F}_2 = \{ \{.30\overline{5}\,, .63\overline{3}\,\} \{.23\overline{3}\,, .55\overline{5}\,\} \{.13\overline{3}\,, .45\overline{5}\,\} \}$

$\mathcal{S}_2 = \{(.6\overline{3}\,, .2\overline{3}\,, .1\overline{3}\,) (.30\overline{5}\,, .5\overline{5}\,, .1\overline{3}\,) (.30\overline{5}\,, .2\overline{3}\,, .4\overline{5}\,)\}$

(Step 3) Project $\mathcal{S}_2$ to form set

$\boldsymbol{T}_3 = \{ (.6\overline{3}\,, .2\overline{3}\,, .1\overline{3}\,) (.30\overline{7\,4\,0}\,, .55\overline{7\,4\,0}\,, .13\overline{7\,4\,0}\,)$
$(.30\overline{7\,4\,0}\,, .23\overline{7\,4\,0}\,, .45\overline{7\,4\,0}\,)\}$

Form the new forecast and score sets $\mathcal{F}_3$, $\mathcal{S}_3$ based on the probabilities in set $\boldsymbol{T}_3$

$\mathcal{F}_3 = \{ \{.30\overline{7\,4\,0}\,, .6\overline{3}\,\} \{.2\overline{3}\,, .55\overline{7\,4\,0}\,\} \{.1\overline{3}\,, .45\overline{7\,4\,0}\,\} \}$

$\mathcal{S}_3 = \{(.6\overline{3}\,, .2\overline{3}\,, .1\overline{3}\,) (.30\overline{7\,4\,0}\,, .55\overline{7\,4\,0}\,, .1\overline{3}\,)$
$(.30\overline{7\,4\,0}\,, .2\overline{3}\,, .45\overline{7\,4\,0}\,)\}$

(Step 4) Project $\mathcal{S}_4$ to form set

$\boldsymbol{T}_4 \approx \{ (.6\overline{3}\,, .2\overline{3}\,, .1\overline{3}\,) (.308, .558, .134) (.308, .234, .458)\}$

Form the new forecast and score sets $\mathcal{F}_4$, $\mathcal{S}_4$ based on the probabilities in set $\boldsymbol{T}_4$

$\mathcal{F}_4 = \{ \{.308, .6\overline{3}\,\} \{.2\overline{3}\,, .558\} \{.1\overline{3}\,, .458\} \}$

$\mathcal{S}_4 = \{(.6\overline{3}\,, .2\overline{3}\,, .1\overline{3}\,) (.308, .558, .1\overline{3}\,) (.308, .2\overline{3}\,, .458)\}$

Iterate the process which converges to forecast set

$\mathcal{F}^* = \{ \{.308\overline{6}\,, .6\overline{3}\,\} \{.2\overline{3}\,, .558\} \{.1\overline{3}\,, .458\} \}$

and score set

$\mathcal{S}^* = \{(.6\overline{3}\,, .2\overline{3}\,, .1\overline{3}\,) (.308\overline{6}\,, .558, .1\overline{3}\,)$
$(.308\overline{6}\,, .2\overline{3}\,, .458)\}$

$\mathcal{F}^*$ is an ε-contamination model whose IP-Brier score dominates $\mathcal{F}$'s score. $\mathcal{F}^*$ has greater *informativeness* (greater *determinacy*) than forecast $\mathcal{F}$ as the hull $\boldsymbol{H}(\mathcal{S}^*)$ is isomorphic to a proper subset of the hull $\boldsymbol{H}(\mathcal{S})$.

References

[1] Cozman, F. and T.Seidenfeld (2009) Independence for Full Conditional Measures, Graphoids and Bayesian Networks. In *Foundations of the Formal Sciences VI*. B.Lowe, E.Pacuit, and J-W Romeijn (eds.). College Publications: London

[2] de Finetti, B. (1974) *Theory of Probability* (vol. 1). John Wiley: New York.

[3] de Finetti, B. (1981) The role of *dutch books* and *proper scoring rules*. *Brit. J. Phil. Sci.* 32: 55-56.

[4] de Finetti, B (2008) *Philosophical Lectures on Probability* (A.Mura, Ed.) Springer: United States.

[5] Gilio, A. (1996) Algorithms for Conditional Probability Assessments. In *Bayesian Analysis in Statistics and Econometrics*. D.A.Berry, K.MChaloner, and J.K.Geweke (eds.) John Wiley: New York, pp. 29-39.

[6] Herron, T., T.Seidenfeld, and L.Wasserman (1997) Divisive Conditioning: Further Results on Dilation. *Phil. Sci.*, 64: 411-444.

[7] Levi, I. (1974) On Indeterminate Probabilities. *J.Phil* 71: 391-418.

[8] Levi, I. (1980) *The Enterprise of Knowledge*. MIT Press: Cambridge.

[9] Predd, J., R.Seiringer, E.H.Lieb, D.Osherson, V.Poor, and S.Kulkarni (2009) Probabilistic coherence and proper scoring rules. *IEEE Trans. Information Theory* 55; 4786-4792.

[10] Regazinni, E. (1987) De Finetti's Coherence and Statistical Inference. *Ann. Stat.* **15**: 845–864.

[11] Savage, L.J. (1971) Elicitation of personal probabilities and expectations. *J. Amer. Stat. Assoc.* 66: 783-801.

[12] Schervish, M.J. (1989) A general method for comparing probability assessors. *Ann. Stat.* 17: 1856-1879.

[13] Schervish, M.J., T.Seidenfeld and J.B.Kadane (1990) State-dependent Utilities. *J.A.S.A.* **85**: 840-847.

[14] Schervish, M.J, T.Seidenfeld, and J.B.Kadane (2009) Proper Scoring Rules, Dominated Forecasts, and Coherence. *Decision Analysis* **6**, #4: 202-221.

[15] Seidenfeld, T., M.J.Schervish, and J.B.Kadane (1990) Decisions without Ordering. In *Acting and Reflecting*. W.Sieg (ed.) Kluwer Publishing: Dordrecht, pp. 143-170.

[16] Seidenfeld, T., M.J.Schervish, and J.B.Kadane (2010) Coherent choice functions under uncertainty. *Synthese* 172, #1: 157-176. Presented at ISIPTA-07.

[17] Seidenfeld, T., M.J.Schervish, and J.B.Kadane (05/2011) Dominating Countably Many Forecasts. Technical Report, Statistics Department, CMU.

[18] Smith, C.A.B. (1961) Consistency in Statistical Inference and Decision. *J.R.S.S.* B **23**: 1-25.

[19] Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London.

[20] Williams, P.M. (1975) Notes on conditional previsions. *I. J. ApproximateReasoning* **4**: 366–383.

Never Say "Not:" Impact of Negative Wording in Probability Phrases on Imprecise Probability Judgments

Michael Smithson
The Australian National University
Michael.Smithson@anu.edu.au

David V. Budescu
Fordham University
budescu@fordham.edu

Stephen B. Broomell
Pennsylvania State University
broomell@gmail.com

Han-Hui Por
Fordham University
hanhui.p@gmail.com

Abstract

A reanalysis of Budescu et al.'s (2009) data on numerical interpretations of the Intergovernmental Panel on Climate Change (IPCC 2007) fourth report's verbal probability expressions (PE's) revealed that negative wording has deleterious effects on lay judgements. Budescu et al. asked participants to interpret PE's in IPCC report sentences, by asking them to provide lower, "best" and upper estimates of the probabilities that they thought the authors intended. There were four experimental conditions, determining whether participants were given any numerical guidelines for translating the PE's into numbers.

The first analysis presented here focuses on six sentences in Budescu et al. that used the PE "very likely" or "very unlikely". A mixed beta regression (Verkuilen & Smithson, in press) modelling the three numerical estimates revealed a less regressive mean and less dispersion for positive than for negative wording in all three estimates. Negative wording therefore resulted in more regressive estimates and less consensus regardless of experimental condition.

The second analysis focuses on two statements that were positive-negative duals. Appropriate pairs of responses were assessed for conjugacy and additivity. A large majority of respondents were appropriately super- and sub-additive in their lower and upper probability estimates. A mixed beta regression model of these three variables revealed that the $\underline{P}(A)$ and $\overline{P}(A^c)$ pairs adhered most closely to conjugacy. Also, the greatest dispersion occurred for $\underline{P}(A) + \overline{P}(A^c)$, followed by $P(A) + P(A^c)$. These results were driven by the dispersion in the estimates for the negatively-worded statement. This paper also describes the effects of the experimental conditions on conjugacy and dispersion.

Keywords. subjective probability, probability expression, elicitation, conjugacy, risk communication, climate change.

1 Introduction

The Intergovernmental Panel on Climate Change (IPCC) has provided reports that synthesize and assess information regarding scientific understanding of climate change phenomena and their potential impact. The fourth IPCC (2007) report utilizes verbal phrases to describe the uncertainties affiliated with its major claims. These phrases include positively- and negatively-worded probabilistic expressions (PE's, e.g., "very likely" and "very unlikely"). The guidelines for the IPCC fourth report provided its authors a numerical translation of the seven PE's they recommended for use in the report (Table 1). These guidelines also are included in the assessments and executive summaries.

Table 1: IPCC Probability Phrase Numerical Guides

Phrase	IPCC Range
Virtually certain	$> 99\%$
Extremely likely	$> 95\%$
Very likely	$> 90\%$
Likely	$> 66\%$
More likely than not	$> 50\%$
About as likely as not	$33\% - 66\%$
Unlikely	$< 33\%$
Very unlikely	$< 10\%$
Extremely unlikely	$< 5\%$
Exceptionally unlikely	$< 1\%$

Budescu, Broomell, and Por (2009) conducted an experimental study of lay interpretations of these PE's, using 13 relevant sentences from the IPCC report. Three sentences contained the PE "very likely," three others had "likely," three more had "more likely than not," three had "unlikely," and three used "very unlikely." PE's such as "very likely" are positively-worded PE's, whereas PE's such as "very unlikely" are negatively-worded PE's. Four examples are:

1. It is very likely that hot extremes, heat waves, and heavy precipitation events will continue to become more frequent.
2. Global average sea level in the last interglacial period (about 125,000 years ago) was likely 4 to 6 m higher than during the 20th century, mainly due to the retreat of polar ice.
3. Temperatures of the most extreme hot nights, cold nights and cold days are unlikely to have increased due to factors other than anthropogenic forcing.
4. It is very unlikely that hot extremes, heat waves, and and heavy precipitation events will not continue to become more frequent.

Budescu et al. asked 223 participants to interpret PE's in these sentences by providing lower, "best" and upper estimates of the probabilities that they thought the authors intended. Participants did so by using numerical sliders on a computer screen. Participants were randomly assigned to one of four conditions:

- Control: No numerical guide to the PE's
- Translation: Participants were shown the IPCC numerical translation guide to the PE's
- Wide: Each sentence contained its appropriate IPCC numerical translation guide
- Narrow: Each sentence contained a numerical translation that was a sub-interval of the IPCC translation range

Budescu et al. reported that participants' "best" estimates were more regressive (toward the middle of the unit interval) than the IPCC guidelines' stipulations, although less so in the Narrow and Wide conditions. The Narrow condition provided the largest improvement in the quality of responses over the Control condition.

Budescu et al. ensured that four of their target sentences included negatively-worded PE's, but they did not assess whether the valence of the PE's had any effects on participants' interpretations. Nevertheless, it is apparent from Figures 2-4 in their paper that the negatively-worded PE's yielded a greater spread of responses (i.e., less consensus) than the positively-worded phrases, and the median responses were more regressive. Both possibilities are worthwhile evaluating because of their implications for eliciting and communicating imprecise probability judgments. Indeed there is empirical evidence that "positive" and "negative" PEs induce different actions and interpretations (e.g. Teigen & Brun, 1999).

We model the lower ($\underline{P}(A)$), best ($P(A)$), and upper ($\overline{P}(A)$) probabilities simultaneously, via a mixed GLM for beta-distributed random variables (Smithson & Verkuilen, 2006; Verkuilen & Smithson, in press). A description of and rationale for this model are given in the Appendix, along with explanations of its parameters.

2 Positive Versus Negative Wording Effects

Responses to the three sentences using "very likely" and the three using "very unlikely" from Budescu et al. were modeled, with responses to the "very unlikely" statements subtracted from 1 to render them comparable to those from the "very likely" statements. Figure 1 shows boxplots of the resultant data. They indicate that there are differences in location and dispersion between the positive versus negative PE's, across the lower, best and upper estimates, and between experimental conditions.

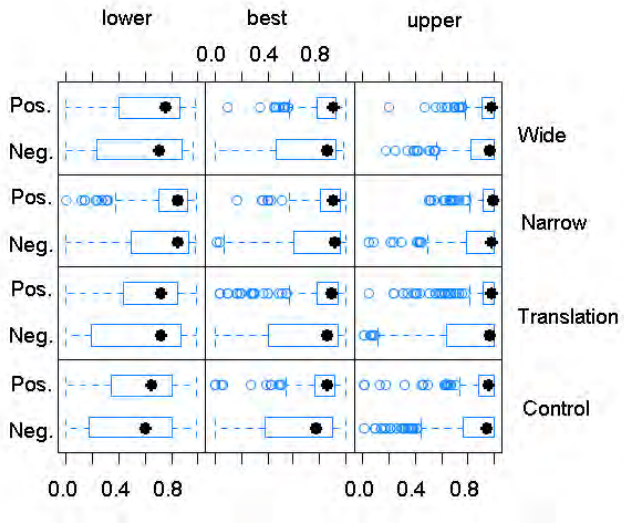


Figure 1: Boxplots of Estimates for Six Questions

We now describe the model of the effects shown in Table 2. The dependent vector consists of six sets of subvectors $\{y_{ij1}, y_{ij2}, y_{ij3}\} = \{\underline{P}(A)_{ij}, P(A)_{ij}, \overline{P}(A)_{ij}\}$, for $j = 1, \ldots, 6$. To respect the ordering $y_{ij1} \leq y_{ij2} \leq y_{ij3}$, we define $x_{i2} = 1$ for $y_{ijk} = y_{ij2}$ or $y_{ijk} = y_{ij3}$ and 0 otherwise, and $x_{i3} = 1$ for $y_{ijk} = y_{ij3}$ and 0 otherwise. We also restrict the regression coefficients for these dummy variables to be non-negative by exponentiating them. The "very likely" versus "very unlikely" predictor is $q_i = 1$ for "very likely" and 0 for "very unlikely". The experimental condition predictors are $t_{i1} = 1$ for the Translation condition, $t_{i2} = 1$ for the Narrow condition, $t_{i3} = 1$ for the Wide con-

dition, and 0 otherwise. Using likelihood-ratio tests and AIC as guides, the best model is

$$\log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) = \beta_0 + x_{2i}e^{\beta_1+\beta_2 q_i} + x_{3i}e^{\beta_3} + \beta_4 q_i + \beta_5 t_{1i} + \beta_6 t_{2i} + \beta_7 t_{3i} + b_i, \tag{1}$$

where $b_i \sim N(0, e^{2u})$, and

$$\log(\phi_{ijk}) = \delta_0 + (\delta_1 + \delta_2 q_i)\, x_{2i} + (\delta_3 + \delta_4 q_i)\, x_{3i} + (\delta_5 + \delta_6 t_{1i} + \delta_7 t_{2i} + \delta_8 t_{3i})\, q_i + \delta_9 t_{1i} + \delta_{10} t_{2i} + \delta_{11} t_{3i}. \tag{2}$$

The coefficients, standard deviations and confidence intervals are shown in Table 2.

Table 2: Mixed Model Parameter Estimates

Param.	Estim.	S.E.	95% Confid. Lower	Interval Upper
		Location	Submodel	
β_0	-0.202	0.096	-0.391	-0.012
β_1	-0.354	0.081	-0.513	-0.196
β_2	0.472	0.089	0.297	0.647
β_3	-0.160	0.054	-0.266	-0.054
β_4	0.369	0.058	0.255	0.482
β_5	0.105	0.124	-0.139	0.349
β_6	0.768	0.139	0.494	1.042
β_7	0.343	0.134	0.078	0.607
u	-0.417	0.054	-0.524	-0.311
		Precision	Submodel	
δ_0	0.526	0.065	0.397	0.654
δ_1	0.319	0.066	0.189	0.448
δ_2	0.576	0.100	0.380	0.772
δ_3	-0.003	0.070	-0.141	0.135
δ_4	-0.272	0.095	-0.458	-0.085
δ_5	0.086	0.091	-0.093	0.265
δ_6	0.365	0.107	0.155	0.575
δ_7	0.707	0.125	0.460	0.953
δ_8	0.466	0.116	0.237	0.696
δ_9	-0.185	0.074	-0.332	-0.039
δ_{10}	0.459	0.087	0.288	0.629
δ_{11}	0.264	0.083	0.100	0.428

The location submodel's β_4 coefficient indicates that the positive statement probabilities were more extreme (less regressive) than their negative statement counterparts. This model's β_2 coefficient also shows that this effect is boosted for the "best" and upper estimates. Significant experimental condition effects occur only in the narrow and wide conditions. In both of those conditions responses are more extreme than in the control condition, and of course this effect is greatest for the narrow condition.

The precision submodel is somewhat more complex. The δ_1 coefficient indicates greater precision for the "best" probability estimates than for the lower probability estimates, and δ_2 suggests this is amplified for the positively-worded statements. However, the negative δ_4 coefficient suggests that this amplification does not hold for the upper estimates.

The positive-negative wording factor moderates the experimental conditions effects in the precision submodel. The interaction effect coefficients δ_7 and δ_8 amplify the greater precision effects from the narrow and wide conditions for the positively-worded sentences, while the δ_6 coefficient negates the lower precision in the translation condition for negatively-worded statements.

The model recovers the mean structure reasonably well. The observed and predicted means are shown in Table 3. The largest inaccuracies are a tendency to under-estimate the lower probability means, and the means for the negative PE's tend to have larger errors (RMS error = .045) than the positive PE's (RMS error = .029).

Table 3: Mixed Model Predicted and Observed Means

	control	treatment	narrow	wide
Negative:	"Very	Unlikely"		
Observed				
lower	.500	.552	.693	.580
best	.652	.686	.775	.702
upper	.825	.798	.863	.866
Predicted				
lower	.450	.476	.638	.535
best	.622	.647	.780	.699
upper	.794	.811	.893	.845
Error				
lower	-.051	-.076	-.055	-.048
best	-.028	-.039	.006	-.003
upper	-.031	.013	.028	-.021
Positive:	"Very	Likely"		
Observed				
lower	.562	.613	.769	.629
best	.809	.816	.856	.828
upper	.905	.912	.930	.927
Predicted				
lower	.542	.568	.718	.625
best	.784	.802	.887	.837
upper	.895	.905	.948	.923
Error				
Lower	-.021	-.045	-.051	-.004
best	-.024	-.015	.031	.009
upper	-.010	-.007	.019	-.003

3 Conjugacy

Two target sentences in Budescu et al. (2009) were positive-negative duals:

- Q1: It is very likely that hot extremes, heat waves, and heavy precipitation events will con-

tinue to become more frequent.

- Q12: It is very unlikely that hot extremes, heat waves, and heavy precipitation events will not continue to become more frequent.

This fact provides an opportunity to examine the relationships among subjective estimates of the lower and upper probabilities of A its complement A^c. Accordingly, this section assesses the responses to this pair of sentences for adherence to superadditivity for lower probabilities, subadditivity for upper proabilities, and the conjugacy rule for lower and upper probabilities.

The superadditivity requirement is $\underline{P}(A)+\underline{P}(A^c) \leq 1$, and the subadditivity requirement is $\overline{P}(A)+\overline{P}(A^c) \geq 1$. A large majority (83.4%) of the respondents' lower probabilities summed to less than 1, and an even larger majority (97.8%) of respondents' upper probabilities summed to more than 1.

Conjugacy is tested via the sums of appropriate pairs of responses, the criteria being
$\underline{P}(A) + \overline{P}(A^c) = 1$,
$P(A) + P(A^c) = 1$, and
$\overline{P}(A) + \underline{P}(A^c) = 1$,
where A^c denotes the complement of event A. Figure 2 shows the boxplots for the three sums and four experimental conditions. The medians all are quite close to 1 (conjugacy). However, there appear to be main effects on dispersion both for experimental conditions and the sums.

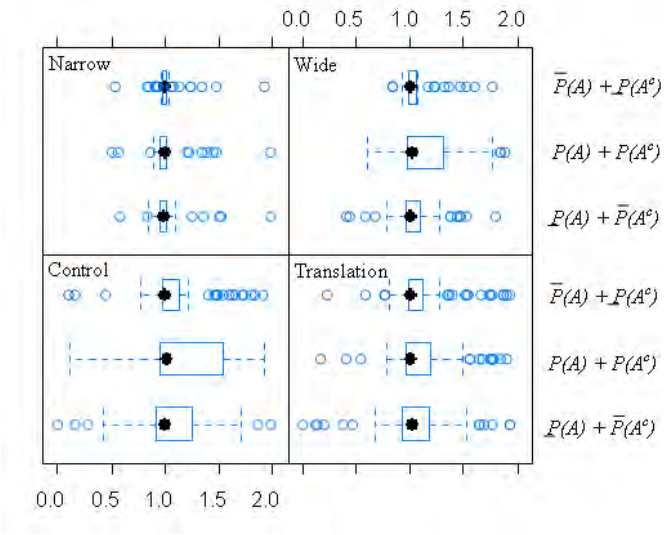


Figure 2: Boxplots of Sums

Turning to a model for the effects, for convenience the three sums described above were divded by 2, so that they lie in the unit interval. The dependent vector $\{y_{ij1}, y_{ij2}, y_{ij3}\}$ consists of the three sums in the order listed above, each divided by 2. We define $x_{i2} = 1$ for $y_{ijk} = y_{ij2}$ and 0 otherwise, and $x_{i3} = 1$ for $y_{ijk} = y_{ij3}$ and 0 otherwise. The experimental condition predictors are defined as before. In terms of likelihood-ratio tests and AIC the best model is

$$\log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) = \beta_0 + \beta_1 x_{2i} + \beta_2 x_{3i} + b_i, \qquad (3)$$

where $b_i \sim N(0, e^{2u})$, and

$$\log(\phi_{ijk}) = \delta_0+\delta_1 x_{2i}+\delta_2 x_{3i}+\delta_3 t_{1i}+\delta_4 t_{2i}+\delta_5 t_{3i}. \quad (4)$$

The coefficients, standard deviations and confidence intervals are shown in Table 4.

Table 4: Conjugacy Model Parameter Estimates

Param.	Estim.	S.E.	95% Confid. Lower	Interval Upper
		Location	Submodel	
β_0	0.140	0.054	0.035	0.246
β_1	0.126	0.036	0.054	0.197
β_2	0.060	0.031	-0.001	0.122
u	-0.388	0.056	-0.499	-0.278
		Precision	Submodel	
δ_0	2.401	0.170	-2.736	-2.066
δ_1	0.382	0.191	0.006	0.759
δ_2	1.148	0.259	0.638	1.657
δ_3	0.301	0.180	-0.053	0.656
δ_4	1.862	0.207	1.454	2.269
δ_5	0.622	0.189	0.249	0.996

The positive β_0 coefficient plus positive β_1 and β_2 show that the closest adherence to conjugacy in the means occurs for lower $\underline{P}(A)+\overline{P}(A^c)$. β_1 is largest so mean conjugacy is worst for $P(A)+P(A^c)$. The large positive δ_2 and moderate positive δ_1 coefficients show that the greatest precision occurs for $\overline{P}(A) + \underline{P}(A^c)$, followed by $P(A)+P(A^c)$. This result is being driven by the imprecision in the $P(A^c)$ estimates.

It turns out that there are no significant experimental condition effects in the location submodel but there are in the precision submodel. The positive δ_4 and δ_5 coefficients suggest that the narrow and wide conditions increase the precision of responses, the narrow condition substantially so.

This model also captures the mean structure well. The location submodel is slightly upward-biased, with the model estimates being about .02 higher than the observed values. However, this bias does not carry over into the differences between the means.

4 Discussion and Conclusions

In their summary and recommendations, Budescu et al. (2009) concluded that access to the IPCC numeri-

Table 5: Conjugacy Model Mean Structure

Conjugacy Sum	observed	predicted
$\underline{P}(A) + \overline{P}(A^c)$	1.052	1.070
$P(A) + P(A^c)$	1.120	1.132
$\overline{P}(A) + \underline{P}(A^c)$	1.076	1.100

cal translation table reduced individual differences in the interpretation of PE's to some degree. Our reanalysis reinforces this claim and their ensuing recommendation. Nevertheless, they also observed that the variability in respondents' estimates in all likelihood is greater than the actual amount of disagreement among the scientists whose views are encompassed by the relevant PE's. Budescu et al. based this assessment on their analysis of the "best" estimates. The reanalysis of the lower and upper probabilities in this paper suggests that the picture is even worse than their summary suggested.

They note, for instance, that 25% of the subjects interpreted "very likely" as having a "best" probability below 70%. The boxplots in Figure 1 show that in three of the four experimental conditions at least 25% of the subjects provided a lower probability of less than 50%. If we turn to "very unlikely" the picture is worse still. The Figure 1 boxplots indicate that in in three of the four experimental conditions about 25% of the subjects returned an upper probability for "very unlikely" greater than 80%!

Our reanalysis provides additional insights. Chief among these is the apparently deleterious impact of negatively-worded PE's on both the regressiveness of people's intuitive numerical translations of these PE's and on the consensus of such translations. Because beta GLMs are naturally heteroscedastic, it is both feasible to separate the effect of a shift in the mean from the effect of a shift in precision on variance. In this setting that separation has important implications regarding our assessment of the amount of variation across individuals in their intuitive numerical translations. More regressive estimates (i.e., further away from 0 or 1) results in greater variability, but that is an artifact of a shift in the mean response. Our results strongly suggest that negatively worded PE's also yield less precision, which results in greater variability that is not attributable to a mean shift.

Two other important findings have emerged regarding precision. First, it is worst for the lower (upper) probability estimates provided for "very likely" ("very unlikely"). But these are translations of the very thresholds identified in the IPCC numerical guides, as shown in Table 1. The effect also was greater for "very unlikely." Second, the narrow and wide conditions not only resulted in less regressive estimates (as Budescu et al. had originally concluded) but they also yielded greater precision, i.e., greater consensus beyond that due to less regressive estimates. This effect was greater for "very likely" than its negative counterpart.

The "pleasant surprise" in our analyses is the fairly strong adherence of subjective estimates to superadditivity, subadditivity, and the conjugacy rules. To our knowledge, only one other empirical assessment of adherence to conjugacy has been reported (Example 2 in Smithson, Merkle & Verkuilen, in press). In our sample, the medians in all conditions and for all three sums deviated no more than .1 from 1, i.e., conjugacy. A substantial majority of these sums were within .2 of 1 (from 52% to 86%). Moreover, both sums involving lower and upper probabilities were closer to conjugacy on average than $P(A) + P(A^c)$, which of course is just binary complementarity. This is striking because while many respondents would have been aware of the binary complementarity rule for classical probabilities, it is very unlikely that they would know about conjugacy. This may be a rather unusual instance where rational prescription coincides with human intuition. However, we urge caution in generalizing from these findings because they are based on only one pair of sentences. A systematic investigation into this matter is needed along the lines suggested below.

At least three avenues of future research are indicated by our findings here. First, the IPCC negatively-worded sentences contained a mixture of negatively-worded PE's and events (of the form "it is very unlikely that A will not occur"). Inspection of the data suggested that at least some respondents many have found these double-negatives especially confusing. Thus, the effect of negatively-worded PE's merits further investigation, most suitably via IPCC report sentences manipulated to incorporate positive and negative wording for various PE's and events crossed in a factorial design, as exemplified in Table 6. It is possible that the greater variability and more regressive means identified with the negatively-worded IPCC sentences are in good part due to double-negatives, but this cannot be determined via the study dealt with here.

Table 6: Factorial Design

Event	Probability phrase	
A	Likely that A	Unlikely that A
A^c	Likely that A^c	Unlikely that A^c

Second, alternative numerical guides could be compared with one another. The IPCC (2007) guides

specified only one bound, leaving the other implicitly at either 0 or 1 as appropriate. For PE's conveying either very high or very low probabilities this seems natural, but for a middling PE such as "likely" an interval from .66 to 1 seems counter-intuitive not only for its width but also because it contains the prescribed interval for "very likely." The IPCC guidelines notwithstanding, it would be worthwhile to ascertain whether there is greater consensus in intuitive translations when the phrases refer to non-overlapping intervals instead of nested ones. Likewise, guides that include prescribed "best" probabilities could be compared with those containing only lower and upper values.

Finally, Budescu et al. suggested several influences on people's intuitive translations. For instance, those convinced about climate change tended to give higher estimates for PE's referring to climate change events or consequences. It is plausible that subjective probability judgments will be subject to confirmation bias, but this has yet to be investigated with respect to subjective imprecise probabilities.

5 Appendix

We begin by describing the mixed GLM employed in this paper. Let $y \in (0,1)$ be distributed Beta$(\mu\phi, (1-\mu)\phi)$, where $\mu = E(y)$ and ϕ is a precision parameter, such that $Var(y) = \mu(1-\mu)/(\phi+1)$ so $\phi = \frac{\mu(1-\mu)}{Var(y)} - 1$. As Smithson and Verkuilen (2006) argue, the Beta distribution is appropriate for modeling a random variable whose support is bounded at both ends, as in this case where the support is the unit interval. While it is not the only such distribution, it is very flexible and also has the attractive property of being parameterized in terms of a mean and a precision parameter. This characteristic renders the Beta distribution especially suitable for modeling the mean response (location) and dispersion simultaneously.

For a two-level model let $i = 1, \ldots, I$ index subjects and $j = 1, \ldots, J$ index observations within the ith subject, so there are $IJ = N$ total observations. A mixed beta GLM contains four matrices of regressors, $\mathbf{X}, \mathbf{Z}, \mathbf{V}, \mathbf{W}$. $\mathbf{X}$ and $\mathbf{V}$ are associated with the location and precision, respectively, so that $\mathbf{x}_i, \mathbf{v}_i$ are their ith row vectors of full rank (Typically they have a column vector $\mathbf{1}$ for an intercept). $\mathbf{Z}$ and $\mathbf{W}$ are the regressors for random effects $\mathbf{b}$ and $\mathbf{d}$, respectively. Then the location and precision submodels are

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{b}, \quad (5)$$

$$\log(\phi_{ij}) = \mathbf{v}_{ij}\boldsymbol{\delta} + \mathbf{w}_{ij}\mathbf{d}. \quad (6)$$

In this paper we restrict the random-effects models to random-intercept models for the location submodel with a normal mixing distribution.

Estimation was by maximum likelihood using the NLMIXED package in SAS 9.2. Maximum likelihood methods enable the use of both likelihood ratio tests for comparing models on the basis of goodness of fit, and Wald t- or z-tests for assessing the significance of individual coefficients in a model. The coefficients' standard errors used in the Wald tests may also be used in constructing confidence intervals for the coefficient estimates.

The location submodel coefficients in this model can be interpreted in a similar way to coefficients in a logistic regression, because the logit link typically is used in both. A positive (negative) β_j is the increase (decrease) in $\log(\mu_{ji}/(1-\mu_{ji}))$ per unit increase (decrease) in its covariate x_{ji}, so e^{β_j} can be interpreted as a multiplier of odds.

In the precision submodel, a positive (negative) δ_j coefficient is the increase (decrease) in $\log(\phi_{ji})$ per unit increase (decrease) in its covariate v_{ji}, so e^{δ_j} can be thought of as a multiplier of precision.

The variance of a Beta random variable is
$\sigma^2 = \mu_{ji}(1-\mu_{ji})/(\phi_{ji}+1)$,
so the variance is influenced both by the mean and precision parameters. This simply reflects the fact that as the mean approaches either 0 or 1, if the precision remains constant then the variance necessarily decreases. However, it is important to bear in mind that modeling precision is not equivalent to modeling the variance. Consequently, interpreting the effect of predictors on the variance may not be straightforward. A positive β_j, for instance, increases variance if it is shifting μ_{ji} from values below .5, but decreases variance if it is shifting μ_{ji} from values above .5.

Acknowledgements

The original survey work by Budescu, Broomell, and Por supported by the National Science Foundation under Grant No. 0345925.

References

[1] D. V. Budescu, S. Broomell, and H.-H. Por. Improving the communication of uncertainty in the reports of the Intergovernmental panel on climate change. *Psychological Science*, 20:299–308, 2009.

[2] Intergovernmental Panel on Climate Change. *Summary for policymakers: Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate*

Change. Retrieved May 2010 from http://www.ipcc.ch/pdf/assessment-report/ar4/wg1/ar4-wg1-spm.pdf, 2007.

[3] M. Smithson and J. Verkuilen. A better lemon-squeezer? Maximum likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11:54-71, 2006.

[4] M. Smithson, Edgar C. Merkle and J. Verkuilen. Beta regression finite mixture models of polarization and priming. *Journal of Educational and Behavioral Statistics*, in press.

[5] K. H. Teigen and W. Brun. The directionality of verbal probability expressions: Effects on decisions, predictions, and probabilistic reasoning. *Organizational Behavior and Human Decision Processes*, 80: 155-190, 1999.

[6] J. Verkuilen and M. Smithson. Mixed and mixture regression models for continuous bounded responses using the beta distribution. *Journal of Educational and Behavioral Statistics*, in press.

Discrete Second-order Probability Distributions that Factor into Marginals

David Sundgren
University of Gävle
dsn@hig.se

Abstract

In realistic decision problems there is more often than not uncertainty in the background information. As for representation of uncertain or imprecise probability values, second-order probability, i.e. probability distributions over probabilities, offers an option. With a subjective view of probability second-order probability would seem to be impractical since it is hard for a person to construct a second-order distribution that reflects his or her beliefs. From the perspective of probability as relative frequency the task of constructing or updating a second-order probability distribution from data is somewhat easier. Here a very simple model for updating lower bounds of probabilities is employed.

But the difficulties in choosing second-order distributions may be further alleviated if structural properties are considered. Either some of the probability values are dependent in some way, e.g. that they are known to be almost equal, or they are not dependent in any other way than what follows from that the values sum to one.

In this work we present the unique family of discrete second-order probability distributions that correspond to the case where dependence is limited. These distributions are shown to have the property that the joint distributions are equal to normalised products of marginal distributions. The distribution family introduced here is a generalisation of a special case of the multivariate Pólya distribution and is shown to be conjugate prior to a compound hypergeometric distribution.

Keywords. Discrete probability, second-order probability, imprecise probability, multivariate Pólya distribution, conjugate prior, compound hypergeometric likelihood.

1 Introduction

In non-trivial decision problems there is often uncertainty about background data. A decision support system or any system that is meant to work with such uncertain data needs a form of representation for uncertain information or else ignore the uncertainty, i.e. allow for false certainty or false precision. Here we are concerned with representation of uncertain or imprecise probability values. Uncertainty and imprecision will be treated the same way, whether a decision maker believes that there is a precise value but is uncertain as regards to what it is, or if imprecision is inherent, the end result is that there is a set of feasible probability values.

Among models for imprecise probability there are interval based approaches, [8, 9, 14, 19, 20, 21], where the probability of an event is represented by two numbers, the lowest and highest possible value. There are also hierarchical models such as those in [11, 10, 23, 6, 4, 2, 22, 18, 17, 5, 12], where each probability value in the interval is weighed. The potential for discrimination that is present in hierarchical models may be utilized to express that some probability values are more reasonable than others. However, this power is difficult to wield since on the one hand local, one-dimensional, changes have global, multi-dimensional, effects that might be hard to grasp, and on the other hand since given some beliefs about imprecise values there appears to be countless sets of weights that are consistent with the beliefs.

1.1 Structural Considerations

The solution might lie in adding structural information, information that is not asked for in traditional models for imprecise probability, but is nonetheless crucially important and not necessarily hard to extract. The importance of structural information in general is argued for in [3]. Here we will focus on one such property, dependency. Dependency is a fun-

damental concept in probability theory. In this paper we work with one particular hierarchical model, second-order probability, where the weights are themselves probabilities. Since second-order probability is a concept that resides fully inside probability theory there is no reason to assume that issues of dependency would be unimportant in that context.

Now the stochastic variables in a second-order probability distribution are probabilities of events in the same outcome space, so the variables are non-negative and sum to one. This fact alone obviously rules out independence. But two first-order probabilities might be dependent beyond the summing to one, it is conceivable that two probability values are almost the same in all situations, or that if one value increases by a certain amount, another value decreases even more. As an example of the first mentioned case we could take the probabilities of three mutually exclusive events A, B and C. We have that $\Pr(B \cup C) = 1 - \Pr(A)$, but assume that $\Pr(B) = \Pr(C) = x$. Then all probability vectors $(\Pr(A), \Pr(B), \Pr(C))$ have the form $(1 - x, x, x)$ and B and C have a higher degree of dependency than what is prescribed by $\Pr(A) + \Pr(B) + \Pr(C) = 1$. Then again, cases where there are no such further dependencies are also conceivable and this is the case that we explore in this paper since independence usually is less complicated than dependence.

Below we suggest a notion that is intended to capture such limited dependency, i.e. that the probabilistic constraint of non-negative variables summing to one is the only source of dependency. Further we demonstrate that such limited dependency means that the joint second-order probability distribution factors into its own marginal distributions, almost as a joint distribution of independent variables. The difference is that since the variables are not really independent the joint distribution is equal to the normalised product of marginal distributions, the product is multiplied with a factor not equal to one.

A family of continuous second-order probability distribution with this property is in [16] shown to be a shifted or contracted variant of the Dirichlet distribution. In fact, the parameters of that version of the Dirichlet distributions are locked to $1/(n-1)$, where n is the number of possible outcomes, instead a new set of parameters a_i, $i = 1, \ldots, n$, are introduced. The a_i are lower bounds of the first-order probability variables. In other words, the lower bounds determine the distribution. The topic of this paper is the discrete counterpart of the contracted Dirichlet distribution that factors into marginals.

Among the reasons for looking at discrete second-order probability distributions as opposed to the shifted Dirichlet distribution are determination and updating of lower bounds. A lower bound for a probability in a continuous second-order distribution can usually not be the result of an observation, but after seeing one black tulip among 20 in a flower shop I know that the probability of a random tulip in the shop being black is at least 1/20. There could also be computational advantages to discrete distributions and in practice the limited resolution of a discrete distribution might be sufficient.

Since an important advantage of discrete second-order distributions as opposed to their continuous counterparts is that they fit nicely into a simple model for updating we consider the conditions under which the distribution considered here are conjugate. Conjugacy is of interest here since it would be important to know whether the structural properties represented by a family of distributions such as that shown here can remain after updating.

The main result of this paper is then twofold; the unique family of discrete second-order probability distributions that factor into marginals and the compound hypergeometric likelihood that is needed for these distributions to be conjugate.

2 Limited Dependency

We assume that all first-order probability values can be written as a ratio k_i/N, where $k_i \geq 0$ and $\sum_{i=1}^{n} k_i = N$. For simplicity we will use the nominators k_i as variables, the denominator N would always be the same. We want to capture and formalise the notion that $\sum_{i=1}^{n} k_i = N, k_i \geq 0$ is the only source of dependency among the variables k_i. When this is the case, the value of a variable would depend on other variables but dependency would only be a function of the sum of variables. For instance, considering the value of k_1 it is important what value the sum of say, k_3 and k_6 holds, but it is irrelevant if k_3 increases and k_6 decreases as long as the sum $k_3 + k_6$ stays the same.

Let $X \not\ni k_i$ be a subset of the set $\{k_1, k_2, \ldots, k_n\}$ of random variables ($\sum_{i=1}^{n} k_i = N$). By definition of conditional probability $p_i(k_i|X) = \frac{p_i(k_i \cup X)}{p(X)}$. The p:s are probability mass functions, indexed where needed to indicate marginal functions. If we wish k_i:s dependency of X to be limited to a function of the sum of variables we should be able to describe $p_i(k_i|X)$ as

$$p_i(k_i|X) = p_i(k_i)\frac{f(k_i + \sum_{k_i \in X} k_j, |X| + 1)}{f(\sum_{k_i \in X} k_j, |X|)}\,. \quad (1)$$

In the functions f we need not only sums of variables but also the number of variables in the sum; the value

of a sum of many variables have more information than a sum of few variables even if the sums are equal.

Since

$$\begin{gathered} p(k_{\pi(1)}, k_{\pi(2)}, \ldots, k_{\pi(n)}) = \\ p_{\pi(1)}(k_{\pi(1)}) p_{\pi(2)}(k_{\pi(2)}|k_{\pi(1)}) \\ p_{\pi(3)}(k_{\pi(3)}|k_{\pi(1)}, k_{\pi(2)}) \\ \vdots \\ p_{\pi(n)}(k_{\pi(n)}|k_{\pi(1)}, k_{\pi(2)}, \ldots, k_{\pi(n-1)}) \end{gathered} \tag{2}$$

for any permutation π, if

$$p_{\pi(i)}(k_i|X) = p_{\pi(i)}(k_i) \frac{f(k_i + \sum_{k_i \in X} k_j, |X|+1)}{f(\sum_{k_i \in X} k_j, |X|)}$$

as in Equation (1) we have that

$$\begin{gathered} p(k_1, \ldots, k_n) = \\ p_{\pi(1)}(k_{\pi(1)}) p_{\pi(2)}(k_{\pi(2)}) \frac{f(k_{\pi(1)} + k_{\pi(2)}, 2)}{f(k_{\pi(1)}, 1)} \\ p_{\pi(3)}(k_{\pi(3)}) \frac{f(k_{\pi(1)} + k_{\pi(2)} + k_{\pi(3)}, 3)}{f(k_{\pi(1)} + k_{\pi(2)}, 2)} \\ \vdots \\ p_{\pi(n)}(k_{\pi(n)}) \frac{f(k_{\pi(1)} + \ldots + k_{\pi(n)}, n)}{f(k_{\pi(1)} + \ldots + k_{\pi(n-1)}, n-1)} = \\ \prod_{i=1}^{n} p_i(k_i) \frac{f(k_{\pi(1)} + \ldots + k_{\pi(n)}, n)}{f(k_{\pi(1)}, 1)} \end{gathered} \tag{3}$$

The numerator $f(k_{\pi(1)} + \ldots + k_{\pi(n)}, n)$ is obviously constant since $\sum_{i=1}^{n} k_i$ is constant equal to N. But the denominator is apparently dependent on the permutation π: if $f(k_i, 1)$ is not constant it is not possible to express the joint probability distribution $p(k_1, \ldots, k_n)$ in this way. On the other hand, if $f(k_i, 1)$ is constant $p(k_1, \ldots, k_n)$ equals the product of marginal distributions multiplied with a constant. That is, if the type of limited dependency described by Equation (1) is achievable the joint probability distribution must factor into marginals.

3 Factoring into Marginals

We have seen that dependence limited to the sum of random variables means that the joint probability density function is proportional to the product of marginal distributions. In the case of discrete second-order probability distributions the limitation is that random variables $k_i, 1 \leq i \leq n$ are such that $k_i \geq 0$ and $\sum_{i=1}^{n} k_i = N$. Note that the k_i/N are probabilities, not the k_i. We could have the rational numbers k_i/N as random variables, but presentation is simplified by dropping the denominator.

Before delving into the calculations, some words about the z transform might be in place. Below we solve the problem at hand by using the convolution property that $\mathcal{Z}\{p_1(k) * p_2(k)\} = \mathcal{Z}p_1(k)\mathcal{Z}p_2(k)$ so that the integrals involved in computing marginal distributions can be computed by eliminating products in a system of equations of products. That we can use convolutions is due to the variables having a fixed sum. The z transform most used below is that of $\frac{\Gamma(k-x+y)}{(k-x)!\Gamma(y)}$ which is $\frac{1}{\left(1-\frac{1}{z}\right)^y z^x}$. In turn, the Gamma function $\Gamma(x)$ is defined as $\int_0^\infty t^{x-1}e^{-t}\,dt$ for complex numbers with positive real parts. For integers it is just the sg hifted factorial, $\Gamma(n) = (n-1)!$. For more on the z transform, see [7] and on the Gamma function, see e.g. [1]

Dependence limited to the sum of k_i being constant equal to N means that

$$p(k_1, k_2, \ldots, k_n) = \frac{1}{K} \prod_{i=1}^{n} p_i(k_i)\,,$$

where p_i is the marginal distribution corresponding to variable k_i. Please observe that $\sum_{i=1}^{n} k_i = N$ throughout the paper.

Then the marginal distribution $p_i(k_i)$ equals

$$\frac{1}{K} p_i(k_i) \ast_{j \neq i} p_j\left(N - k_i\right)\,, \tag{4}$$

where $*_{j\neq i}$ is the $n-1$-fold repeated convolution $p_1 * p_2 * \cdots * p_{i-1} * p_{i+1} * \cdots * p_n$ and $K = *_{i=1}^{n} p_i(N)$.

In the transform domain,

$$\prod_{j \neq i} \mathcal{Z}\left\{p_j(k_j)\right\} = \mathcal{Z}\left\{KH(c_i - k_i)\right\} \tag{5}$$

for all $i, i = 1, \ldots, n$, where H is the Heaviside function and the support of p_i ends at $k_i = c_i$. Cancelling in these n equations in the z domain implies that except for different shifts all marginals p_i are equal.

Since the z transform of a constant K is $\frac{Kz}{z-1}$, if $p_i(k_i)$ is any shifted function $q_i(k_i - a_i)$,

$$\mathcal{Z}\left\{p_i(k_i)\right\} = \left(\frac{Kz}{z-1}\right)^{\frac{1}{n-1}} \frac{1}{z^{a_i}} \tag{6}$$

due to the shift property $\mathcal{Z}\{x(n-k)\} = \mathcal{Z}\{x(n)\}z^{-k}$ and

$$\prod_{j \neq i} \mathcal{Z}\left\{p_j(k_j)\right\} = \frac{Kz}{z-1} \frac{1}{z^{\sum_{j \neq i} a_j}}\,, \tag{7}$$

hence

$$\ast_{j\neq i}p_i(k_i) = \mathcal{Z}^{-1}\left\{\frac{Kz}{z-1}\frac{1}{z^{\sum_{j\neq i}a_j}}\right\}(k_i) = KH\left(k_i - \sum_{j\neq i}a_j\right), \quad (8)$$

giving

$$\ast_{j\neq i}p_i(N-k_i) = KH\left(N - k_i - \sum_{j\neq i}a_j\right) \quad (9)$$

which equals $KH\left(c_i - k_i\right)$ if $c_i = N - \sum_{j\neq i}a_j$, i.e. the upper limit of the support of p_i is $N - \sum_{j\neq i}a_j$, where a_j is the lower limit of the support of marginal distribution p_j.

So

$$p_i(k_i) = \mathcal{Z}^{-1}\left\{\left(\frac{Kz}{z-1}\right)^{\frac{1}{n-1}}\frac{1}{z^{a_i}}\right\}(k_i) = \frac{K^{\frac{1}{n-1}}\Gamma\left(k_i - a_i + \frac{1}{n-1}\right)}{(k_i-a_i)!\Gamma\left(\frac{1}{n-1}\right)}. \quad (10)$$

And

$$K = \ast_{i=1}^{n}p_i(N) = \mathcal{Z}^{-1}\left\{\prod_{i=1}^{n}\mathcal{Z}\left\{p_i(k_i)\right\}\right\}(N) = \mathcal{Z}^{-1}\left\{\prod_{i=1}^{n}\left(\frac{Kz}{z-1}\right)^{\frac{1}{n-1}}\frac{1}{z^{a_i}}\right\}(N) = K^{\frac{n}{n-1}}\mathcal{Z}^{-1}\left\{\left(\frac{z}{z-1}\right)^{\frac{n}{n-1}}\frac{1}{z^{\sum_{i=1}^{n}a_i}}\right\}(N) = \quad (11)$$
$$K^{\frac{n}{n-1}}\mathcal{Z}^{-1}\left\{\left(\frac{z}{z-1}\right)^{\frac{n}{n-1}}\right\}(N - \sum_{i=1}^{n}a_i) = K^{\frac{n}{n-1}}\frac{(N-\sum_{i=1}^{n}a_i)!\Gamma\left(\frac{1}{n-1}\right)}{(n-1)\Gamma\left(N+1-\sum_{i=1}^{n}a_i+\frac{1}{n-1}\right)}$$

That is,

$$K = \left(\frac{(N-\sum_{i=1}^{n}a_i)!\Gamma\left(\frac{1}{n-1}\right)}{(n-1)\Gamma\left(N+1-\sum_{i=1}^{n}a_i+\frac{1}{n-1}\right)}\right)^{n-1} \quad (12)$$

and the marginal distributions are

$$p_i(k_i) = \frac{(N-\sum_{j=1}^{n}a_j)!\Gamma\left(k_i - a_i + \frac{1}{n-1}\right)}{(n-1)\Gamma\left(N+1-\sum_{j=1}^{n}a_j+\frac{1}{n-1}\right)(k_i-a_i)!}, \quad (13)$$

$i = 1, \ldots, n$

The joint distribution is

$$p(k_1, \ldots, k_n) = \frac{(N-\sum_{i=1}^{n}a_i)!\prod_{i=1}^{n}\frac{\Gamma(k_i-a_i+\frac{1}{n-1})}{(k_i-a_i)!}}{(n-1)\Gamma\left(\frac{1}{n-1}\right)^{n-1}\Gamma\left(N+1-\sum_{i=1}^{n}a_i+\frac{1}{n-1}\right)} \quad (14)$$

Going back to Section 2 we have now seen that the form of limited dependency that implies factoring into marginals is possible to realise, in fact by considering the multivariate marginal distributions it can be shown that the functions in Equation (3) have the desired properties, that is $f(k_1, \ldots, k_n, n) = 1/K$ and $f(k_i, 1)$ is constant equal to one. The corresponding reasoning could also justify the constraint of factoring into marginals for the contracted Dirichlet distribution of [16].

3.1 Basic Properties

Since $\Gamma(k+x)/k!$ approaches k^{x-1} as k grows when $x << k$, the discrete distribution described above becomes, appropriately normalised, equal to the shifted Dirichlet distribution of [16] when N tends to infinity. In this, k_i/N and a_i/N of the discrete distribution corresponds to the real-valued first-order probability x_i and a_i in the continuous distribution.

Just as the continuous distribution in [16] is a generalization of a Dirichlet distribution with parameters $1/(n-1)$, the discrete probability distribution considered here is, when the parameters $a_i = 0$, a multivariate Pólya distribution [13] with parameters $1/(n-1)$.

The mean of a marginal probability density function $p_i(k_i)$ of the type described here is

$$a_i + \frac{N - \sum_{i=1}^{n}a_i}{n}, \quad (15)$$

c.f. the mean $a_i + \frac{1-\sum_{i=1}^{n}a_i}{n}$ of the shifted Dirichlet distribution.

The variance is

$$\frac{(n-1)^2(N-\sum_{i=1}^{n}a_i)^2}{n^2(2n-1)} + \frac{(n-1)(N-\sum_{i=1}^{n}a_i)}{n(2n-1)} \quad (16)$$

which approaches N^2 times the variance of the shifted Dirichlet distribution with lower bounds a_i/N.

The multivariate Pólya distribution is obtained by drawing the underlying probabilities p_i from a Dirichlet distribution and integrating out $\mathbf{p} = (p_1, \ldots, p_n)$ from the multinomial distribution. In the same way, if

we compound the Dirichlet distribution with parameters $1/(n-1)$ with the shifted multinomial distribution

$$\frac{(N-\sum_{i=1}^{n} a_i)! \prod_{i=1}^{n} p_i^{k_i-a_i}}{\prod_{i=1}^{n}(k_i-a_i)!} \tag{17}$$

that is used in [15], we have

$$\begin{gathered}\int_{\mathbf{p}} \frac{1}{(n-1)^n \Gamma(n/(n-1))^{n-1} \prod_{i=1}^{n} p_i^{\frac{n-2}{n-1}}} \\ \frac{(N-\sum_{i=1}^{n} a_i)! \prod_{i=1}^{n} p_i^{k_i-a_i}}{\prod_{i=1}^{n}(k_i-a_i)!}\, d\mathbf{p} = \\ \frac{(N-\sum_{i=1}^{n} a_i)!}{(n-1)\Gamma(1/(n-1))^{n-1}\Gamma(N+1+1/(n-1))} \\ \prod_{i=1}^{n} \frac{\Gamma(k_i-a_i+1/(n-1))}{(k_i-a_i)!}\end{gathered} \tag{18}$$

That is, the joint discrete distribution that factors into marginals.

4 Example

Let $n = 4, N = 8$ and $a_1 = 0, a_2 = 1, a_3 = 3, a_4 = 0$. Then

$$\begin{gathered}p(k_1,k_2,k_3) = \\ \frac{4!\Gamma(k_1+1/3)\Gamma(k_2-1+1/3)}{3\Gamma(1/3)^3\Gamma(5+1/3)k_1!(k_2-1)!} \\ \frac{\Gamma(k_3-3+1/3)\Gamma(8-k_1-k_2-k_3+1/3)}{(k_3-3)!(8-k_1-k_2-k_3)!}\end{gathered}$$

k_1	0	1	2	3	4
	0.534	0.178	0.119	0.0923	00769
k_2	1	2	3	4	5
	0.534	0.178	0.119	0.0923	0.0769
k_3	3	4	5	6	7
	0.534	0.178	0.119	0.0923	0.0769

Table 1: Marginal probability density values for $p_1(k_1), k_1 = 0,\ldots,4, p_2(k_2), k_2 = 1,\ldots,5$ and $p_3(k_3), k_3 = 3,\ldots,7$.

and the marginal distributions are

$$\begin{gathered}p_1(k_1) = \sum_{k_2=1}^{5-k_1} \sum_{k_3=3}^{8-k_1-k_2} p(k_1,k_2,k_3) = \\ \frac{4!\Gamma(k_1+1/3)}{3\Gamma(5+1/3)k_1!}\,,\end{gathered} \tag{19}$$

$$\begin{gathered}p_2(k_2) = \sum_{k_1=0}^{5-k_2} \sum_{k_3=3}^{8-k_1-k_2} p(k_1,k_2,k_3) = \\ \frac{4!\Gamma(k_2-1+1/3)}{3\Gamma(5+1/3)(k_2-1)!}\,,\end{gathered} \tag{20}$$

$$\begin{gathered}p_3(k_3) = \sum_{k_1=0}^{7-k_3} \sum_{k_2=1}^{8-k_1-k_3} p(k_1,k_2,k_3) = \\ \frac{4!\Gamma(k_3-3+1/3)}{3\Gamma(5+1/3)(k_3-3)!}\,,\end{gathered} \tag{21}$$

$$\begin{gathered}p_4(k_4) = p_4(8-k_1-k_2-k_3) = \\ \sum_{k_1=0}^{4-k_4} \sum_{k_2=1}^{5-k_1-k_4} p(k_1,k_2,8-k_1-k_2-k_4) = \\ \frac{4!\Gamma(k_4+1/3)}{3\Gamma(5+1/3)k_4!} = \frac{4!\Gamma(8-k_1-k_2-k_3+1/3)}{3\Gamma(5+1/3)(8-k_1-k_2-k_3)!}\end{gathered} \tag{22}$$

The means of p_1, p_2, p_3 and p_4 are $1, 2, 4$ and 1, respectively, corresponding to mean first-order probabilities of $1/8, 1/4, 1/2$ and $1/8$. Since k_1 and k_4 share the same conditions, their respective marginal probability density functions are equal. We see a table with values of the marginal distribution functions in Table 1. The values reveal that the distributions are essentially equal but differently shifted according to their respective lower bounds of support.

5 Updating

One advantage of treating relative frequencies as first-order probabilities is that updating of lower bounds of probabilities may come about in a natural way. See [15], where this is discussed and exemplified with (shifted) multinomial distributions as prior and posterior distributions and a hypergeometric likelihood. In

this paper we are concerned with a shifted version of the multivariate Pólya distribution where the parameters of the Pólya distribution are locked at $1/(n-1)$ but a new vector $\begin{pmatrix} a_1 & a_2 & \dots & a_n \end{pmatrix}$ of parameters is introduced, where the a_i are lower bounds, i.e. for whatever reason we know that there are at least a_i objects of type i among the total N objects.

As described above in Section 3 this variant of the multivariate Pólya distribution represents a situation where the variables k_i, the number of objects of respective type, are in a sense minimally dependent. This property does not necessarily remain after updating and since the case of further dependencies than those incurred by $\sum_{i=1}^{n} k_i = N$ remains to be investigated we choose to consider the conditions under which updating must be done for the shifted multinomial Pólya distribution to be a conjugate distribution. First though, the model for updating deserves some explanation.

5.1 The Urn and the Plate

Since the lower bounds a_i are the only parameters it is these values that can be affected by updating. The idea behind the model proposed in [15] is that if I observe a_i objects of type i I know with absolute certainty that there were at least a_i such objects to begin with. In terms of the ubiquitous urn, we have N balls with n different colours in an urn and the question is as usual how many balls there are of each colour in the urn. Updating consists of picking a handful $(\sum_{i=1}^{n} a_i)$ of balls from the urn and observing that a_i of them have colour i.

Then we know that there were at least a_i balls with colour i in the urn to begin with. But in terms of probabilities and relative frequencies we are only interested in these numbers in relation to the original number N of balls in the urn, e.g. after observing three green balls from an urn with 20 balls I know that the relative frequency of green balls in the urn was at least 3/20. Thus one might think that replacement is in order so that there remains N balls in the urn. However, if I after replacement pick three green balls again in the next round I have no justification for claiming that there at least six green balls out of 20 since some of the balls might be the same as in the previous updating. One solution could be to mark the already observed balls and ignore them in future updating but then I would not know the results of previous experiments without taking notes. Putting the observed balls on a plate on the side in full sight saves ink and paper and reminds us that observed balls are not simply not replaced in the sense of being discarded. The balls on the plate count but updating is only done by probing the urn.

5.2 Shifted Pólya as Conjugate Prior

First let us observe that since the discrete second-order distributions that are topic of this paper factor into marginals, if prior and posterior are both from this family, the likelihood must also factor into marginals. We look at the one-variable marginal case first for ease of presentation. W.l.o.g. we assume that the prior distribution have parameters $a_i = 0$, i.e. nothing has been observed and apart from a structural assumption of minimal dependency we know nought but N, the total number of objects in the urn, and n, the number of different colours. As described in Section 5.1 above the experiment consists of drawing $\sum_{i=1}^{n} a_i$ balls from the urn and thus rule out the possibility that the number k_i of balls with colour i would be less than a_i.

The i:th marginal of the prior is Beta-binomial with parameters $\alpha = \frac{1}{n-1}$ and $\beta = 1$. i.e.

$$\binom{N}{k_i} \frac{\mathrm{B}\left(k_i + \frac{1}{n-1}, N - k_i + 1\right)}{\mathrm{B}\left(\frac{1}{n-1}, 1\right)}, \tag{23}$$

the i:th marginal of the posterior is Beta-binomial with the same parameters $\alpha = \frac{1}{n-1}, \beta = 1$ as in the prior but k_i replaced with $k_i - a_i$ and N substituted for $N - \sum_{j=1}^{n} a_j$:

$$\binom{N - \sum_{j=1}^{n} a_j}{k_i - a_i} \frac{\mathrm{B}\left(k_i - a_i + \frac{1}{n-1}, N - \sum_{j \neq i} a_i - k_i + 1\right)}{\mathrm{B}\left(\frac{1}{n-1}, 1\right)}. \tag{24}$$

The corresponding likelihood is achieved by a weighted hypergeometric distribution

$$\frac{\binom{N-\sum_{j=1}^{n} a_j}{k_i - a_i}}{\binom{N}{k_i}} p^{-a_i} (1-p)^{a_i - \sum_{j=1}^{n} a_j}, \tag{25}$$

where p is drawn from Beta $\left(k_i + \frac{1}{n-1}, N - k_i + 1\right)$ so that the likelihood is the compound distribution

$$\int_0^1 \frac{\binom{N-\sum_{j=1}^n a_j}{k_i-a_i}}{\binom{N}{k_i}} p^{-a_i}(1-p)^{a_i-\sum_{j=1}^n a_j} \frac{p^{k_i-\frac{n-2}{n-1}}(1-p)^{N-k_i}}{\mathrm{B}\left(k_i+\frac{1}{n-1}, N-k_i+1\right)}\, dp = \frac{\binom{N-\sum_{j=1}^n a_j}{k_i-a_i}}{\binom{N}{k_i}} \frac{\mathrm{B}\left(k_i-a_i+\frac{1}{n-1}, N-\sum_{j=1}^n a_j+a_i-k_i+1\right)}{\mathrm{B}\left(k_i+\frac{1}{n-1}, N-k_i+1\right)} \tag{26}$$

The multivariate likelihood is the weighted hypergeometric distribution

$$\prod_{i=1}^n \frac{\binom{N-\sum_{j=1}^n a_j}{k_i-a_i}}{\binom{N}{k_i}} p_i^{-a_i}, \tag{27}$$

where $\mathbf{p}$ is drawn from the Dirichlet distribution with parameters $k_i + \frac{1}{n-1}$. That is,

$$\prod_{i=1}^n \frac{\binom{N-\sum_{j=1}^n a_j}{k_i-a_i}}{\binom{N}{k_i}} \frac{\Gamma\left(\sum_{i=1}^n k_i+\frac{1}{n-1}\right)}{\prod_{i=1}^n \Gamma\left(k_i+\frac{1}{n-1}\right)} \int_{\mathbf{p}} \prod_{i=1}^n p_i^{k_i-a_i+\frac{1}{n-1}-1}\, d\mathbf{p} = \Gamma\left(N+\frac{1}{n-1}\right)((N-\sum_{j=1}^n a_j)!)^n \prod_{i=1}^n \frac{k_i!(N-k_i)!\Gamma\left(k_i-a_i+\frac{1}{n-1}\right)}{N!\left(N-\sum_{j=1}^n a_j-k_i+a_i\right)!(k_i-a_i)!\Gamma\left(k_i+\frac{1}{n-1}\right)} \tag{28}$$

Admittedly this likelihood function appears rather exotic, particularly in the factors $p_i^{-a_i}$ which mean that it is more likely to draw a larger number of balls of a certain colour. In contrast, as seen in Section 4 the prior and posterior distributions are such that lower values of the number of objects of type i have higher probability. The full implications of this are yet to be considered but one possible interpretation is that such likelihood functions would rarely be seen in nature as it were. In that case the limited dependency of the original proportions in the urn is fragile and easily disturbed when removing objects.

6 Conclusions

Structural properties such as dependency might be worth considering when choosing a second-order distribution for the purpose of expressing imprecise probabilities. Second-order probability distributions have probability values as variables, hence independence is impossible. We have however suggested that joint second-order probability distributions that are equal to the normalised products of their own marginal distributions capture the property of a form of minimal dependency. A continuous family of second-order distributions has been found earlier but here a corresponding discrete family is discovered. This family can be described as a generalisation of a special case of the multivariate Pólya distribution where the parameters are fixed but new parameters in the form of lower bounds on the variables are introduced.

The raison d'être of discrete second-order distributions is that they allow for interpreting relative frequencies as first-order probabilities in a natural way. Such a context makes the interpretation of the meaning of second-order probability values easier in that concrete examples in the form of urn models etc. are readily available. An example is updating where a so-called urn-and-plate model gives a simple description of updating of lower bounds. Discrete second-order distributions are also versatile since they apart from relative frequencies also lend themselves to subjective probabilities. That is as long as the subjective probabilities do not involve statements about irrational numbers such as "I am sure that the probability is at least $1/\pi$". Reasonably the lower bound could be given as $8/25$ or some other rational number instead, infinite precision is meaningless in subjective probability judgements.

The family of distributions discussed here represents a form of limited dependency. We have seen that the family being conjugate requires a rather special likelihood function which suggests that the property of limited dependency is sensitive to the removal of objects that occurs in updating of lower bounds in the plate-and-urn model. Full understanding of the meaning of the parameters of the compound likelihood is however a matter for further investigation.

Acknowledegments

I would like to thank the reviewers for their valuable input. Some of the mistakes they found were important indeed. I am also grateful to my Chinese students who unwittingly led me to see the usefulness of convolution.

References

[1] R. A. Asker and R. Roy. *NIST handbook of mathematical functions*, chapter Gamma function. Cambridge University Press, 2010.

[2] G. De Cooman and P. Walley. A possibilistic hierarchical model for behaviour under uncertainty. *Theory and Decision 52 (4)*, pages 327–374, 2002.

[3] M. Danielson, L. Ekenberg, and D. Sundgren. Structure information in decision trees and similar formalisms. In *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference*, pages 62–67. AAAI Press, 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] P. Gärdenfors and N.-E. Sahlin. Decision, probability and utility: Selected readings. In *Decision, Probability and Utility:Selected Readings*, chapter 16, Unreliable probabilities, risk taking, and decision making, pages 313–334. Cambridge University Press, 1988.

[7] Eliahu Jury. *Theory and application of the z-transform method*. R.E. Krieger Pub. Co, 1973.

[8] B. O. Koopman. The axioms and algebra of intuitive probability. *Annals of Mathematics*, 41:269–292, 1940.

[9] B. O. Koopman. The bases of probability. *Bulletin of the American Mathematical Society*, 46:763–774, 1940.

[10] I. Levi. *The Enterprise of Knowledge*. MIT Press, 1980.

[11] R.D. Luce and H. Raiffa. *Games and Decisions*, chapter Appendix 1, A probabilistic theory of utility, pages 371–384. Dover Publications, 1957.

[12] R. F. Nau. Uncertainty aversion with second-order utilities and probabilities. *Management Science*, 52(1):136–145, 2006.

[13] G. Pólya. Sur quelques points de la théorie des probabilités. *Ann. Inst. Poincaré*, 1:117–161, 1931.

[14] C. A. B. Smith. Consistency in statistical inference and decision. *Journal of the Royal Statistical Society, Series B, xxiii*, pages 1–25, 1961.

[15] D. Sundgren. Expected utility from multinomial second-order probability distributions. *Polibits*, (42):71–75, 2010.

[16] D. Sundgren, L. Ekenberg, and M. Danielson. Shifted dirichlet distributions as second-order probability distributions that factors into marginals. In *Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications*, pages 405–410, 2009.

[17] L. V. Utkin. Imprecise second-order hierarchical uncertainty model. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Volume 11:3*, pages 301–317, 2003.

[18] L. V. 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.

[19] P. Walley. *Statistical reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

[20] P. Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24(2-3):125–148, 2000.

[21] K. Weichselberger. The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24(2-3):149 – 170, 2000.

[22] H.-C. Wu. Fuzzy optimization problems based on ordering cones. *Fuzzy Optimization and Decision Making*, 2 (1):13–29, 2003.

[23] L. A. Zadeh. Fuzzy probabilities. *Information Processing and Management, 20*, pages 363–372, 1984.

Probability boxes on totally preordered spaces for multivariate modelling

Matthias Troffaes
Durham University, Durham, UK
matthias.troffaes@gmail.com

Sebastien Destercke
CIRAD, Montpellier, France
sebastien.destercke@cirad.fr

Abstract

Probability boxes (pairs of cumulative distribution functions) are among the most popular models used in imprecise probability theory. In this paper, we provide new efficient tools to construct multivariate p-boxes and develop algorithms to draw inferences from them. For this purpose, we formalise and extend the theory of p-boxes using lower previsions. We allow p-boxes to be defined on arbitrary totally preordered spaces, hence thereby also admitting multivariate p-boxes. We discuss the construction of multivariate p-boxes under various independence assumptions. An example demonstrates the practical feasibility of our results.

Keywords. p-box, natural extension, multivariate, elicitation, independence, Fréchet, lower prevision

1 Introduction

Imprecise probability [18] refers to uncertainty models applicable in situations where the available information does not allow us to single out a unique probability measure for the random variables involved. They require more complex mathematical tools, such as non-linear functionals. It is therefore of interest to consider models that yield simpler mathematical descriptions, at the expense of generality, but gaining ease of use, elicitation, and representation.

We consider one such model: pairs of lower and upper distribution functions, also called *probability boxes*, or p-boxes [9, 10]. They are often used in risk studies, where cumulative distributions are central. Many theoretical properties and practical aspects of p-boxes have already been studied in the literature. Previous work includes probabilistic arithmetic [20], which provides a very efficient numerical framework for particular inferences with p-boxes (and which we generalise in this paper). In [11], p-boxes are connected to info-gap theory [1]. The relation between p-boxes and random sets was investigated in [14]. Finally, an extension of p-boxes to arbitrary finite spaces [8] yields potential applications to much more general problems.

In this paper, we study p-boxes using lower previsions [19, 18]. From the point of view of lower previsions, p-boxes were studied briefly in [18, Section 4.6.6] and [17]. This has at least two advantages. Firstly, they can be defined on arbitrary spaces. Secondly, they come with a powerful inference tool, called *natural extension*. We will study the natural extension of a p-box, and we derive a number of useful expressions for it, whence providing new numerical tools for exact inferences on arbitrary random quantities and events.

As mentioned, [8] extended p-boxes to finite totally preordered spaces. In this paper, we extend p-boxes further to arbitrary totally preordered spaces, leading to many useful features that classical p-boxes do not have. Firstly, we encompass, in one sweep, p-boxes defined on finite spaces and on closed real intervals. Secondly, as we do not impose anti-symmetry on the ordering, we can also handle product spaces by considering an appropriate total preorder, and thus also admit multivariate non-finite p-boxes, which have not been considered before.[1] Whence, we can specify p-boxes directly on the product space. Contrast this with the usual multivariate approach to p-boxes, such as probabilistic arithmetic [20], that consider one marginal p-box per dimension and draw inferences from a joint model built around some information about variable dependencies. Finally, our approach is also useful in elicitation, as it allows uncertainty to be expressed as probability bounds over any collection of (possibly multivariate) nested sets, because we can always find a total preorder that is compatible with any collection of nested sets.

The paper is organised as follows: Section 2 provides a brief introduction to the theory of coherent lower

[1] We still require the preorder to be total. P-boxes for partially preordered spaces might be interesting, but are not considered in this paper.

previsions. Section 3 introduces and studies the p-box model from the point of view of lower previsions. Section 4 provides an expression for the natural extension of a p-box to all events, and Section 5 studies the natural extension to all gambles. Section 6 studies an important special case of p-boxes whose preorder is induced by a real-valued mapping, as this is a convenient way to specify a multivariate p-box. Section 7 discusses the construction of such multivariate p-boxes from marginal coherent lower previsions under arbitrary dependency models. Section 8 demonstrates the theory with an example.

2 Preliminaries

This section introduces lower previsions, see [2, 19, 18, 15] for details.

The possibility space is Ω. A *gamble* on Ω is a bounded real-valued map on Ω. The set of all gambles on Ω is $\mathcal{L}(\Omega)$, or $\mathcal{L}$ if Ω is evident. A subset of Ω is an *event*. The *indicator* of A is the gamble that is 1 on A and 0 elsewhere: write I_A, or A if confusion fails.

A *lower prevision* $\underline{P}$ is a real-valued map on an arbitrary subset $\mathcal{K}$ of $\mathcal{L}$: for any f in $\mathcal{K}$, $\underline{P}(f)$ represents a subject's supremum buying price for f (see [18] for actual explanation). A lower prevision on a set of indicators of events is a *lower probability*.

$\overline{P}$ denotes the conjugate *upper prevision* of $\underline{P}$: for every $-f \in \mathcal{K}$, $\overline{P}(f) = -\underline{P}(-f)$; it represents a subject's infimum selling price for f.

A real-valued map P on $\mathcal{L}$ satisfying $P(f) \geq \inf f$ and $P(f+g) = P(f) + P(g)$ for all f and $g \in \mathcal{L}$ is a *linear prevision* on $\mathcal{L}$ [18, p. 88, Sec. 2.4.8]. The set of all linear previsions on $\mathcal{L}$ is denoted by $\mathcal{P}$. A linear prevision is essentially an expectation operator.

Of particular interest is the set

$$\mathcal{M}(\underline{P}) = \{Q \in \mathcal{P} \colon (\forall f \in \mathcal{K})(Q(f) \geq \underline{P}(f))\}.$$

If $\mathcal{M}(\underline{P}) \neq \emptyset$, then $\underline{P}$ is said to *avoid sure loss*, in which case the *natural extension* of $\underline{P}$ [18, Sec. 3.4.1]

$$\underline{E}(f) = \min_{Q \in \mathcal{M}(\underline{P})} Q(f) \text{ for all } f \in \mathcal{L}$$

extends $\underline{P}$ to $\mathcal{L}$. Finally, $\underline{P}$ is called *coherent* [19, p. 18] when it coincides with $\underline{E}$ on $\mathcal{K}$.

A lower prevision $\underline{P}$ defined on a lattice of gambles $\mathcal{K}$, i.e., a set of gambles closed under point-wise maximum and point-wise minimum, is called *n-monotone* if for all $p \in \mathbb{N}$, $p \leq n$, and all f, f_1, …, f_p in $\mathcal{K}$ [5]:

$$\sum_{I \subseteq \{1,\ldots,p\}} (-1)^{|I|} \underline{P}\left(f \wedge \bigwedge_{i \in I} f_i\right) \geq 0.$$

A lower prevision which is n-monotone for all $n \in \mathbb{N}$ is called *completely monotone*.

3 P-Boxes

Next, we introduce the formalism of p-boxes defined on totally preordered spaces. In contrast to [9], we do not restrict p-boxes to intervals on the real line.

Let $(\Omega, \preceq)$ be a total preorder: so $\preceq$ is transitive and reflexive and any two elements are comparable. We write $x \prec y$ for $x \preceq y$ and $x \not\succeq y$, $x \succ y$ for $y \prec x$, and $x \simeq y$ for $x \preceq y$ and $y \preceq x$. For any two x, $y \in \Omega$ exactly one of $x \prec y$, $x \simeq y$, or $x \succ y$ holds. We also use the following common notation for intervals in Ω:

$$[x, y] = \{z \in \Omega \colon x \preceq z \preceq y\}$$
$$(x, y) = \{z \in \Omega \colon x \prec z \prec y\}$$

and similarly for $[x, y)$ and $(x, y]$.

For simplicity, we assume that Ω has a smallest element 0_Ω and a largest element 1_Ω (we can always add them to Ω).

A *cumulative distribution function* is a mapping $F : \Omega \to [0, 1]$ which is non-decreasing and satisfies moreover $F(1_\Omega) = 1$. For each $x \in \Omega$, we interpret $F(x)$ as the probability of the interval $[0_\Omega, x]$. We do not impose $F(0_\Omega) = 0$, so we allow $\{0_\Omega\}$ to carry non-zero mass, which happens commonly if Ω is finite. No continuity assumptions are made.

By $\Omega/\simeq$ we denote the quotient set of Ω with respect to the equivalence relation $\simeq$ induced by $\preceq$, that is:

$$[x]_\simeq = \{y \in \Omega \colon y \simeq x\} \text{ for any } x \in \Omega$$
$$\Omega/\simeq = \{[x]_\simeq \colon x \in \Omega\}$$

Because F is non-decreasing, F is constant on elements $[x]_\simeq$ of $\Omega/\simeq$.

Definition 1. A *probability box*, or *p-box*, is a pair $(\underline{F}, \overline{F})$ of cumulative distribution functions from Ω to $[0, 1]$ satisfying $\underline{F} \leq \overline{F}$.

A p-box is interpreted as a lower and an upper cumulative distribution function. In Walley's framework, this means that a p-box is interpreted as a lower probability $\underline{P}_{\underline{F},\overline{F}}$ on the set of events

$$\mathcal{K} = \{[0_\Omega, x] \colon x \in \Omega\} \cup \{(y, 1_\Omega] \colon y \in \Omega\}$$

by

$$\underline{P}_{\underline{F},\overline{F}}([0_\Omega, x]) = \underline{F}(x) \text{ and } \underline{P}_{\underline{F},\overline{F}}((y, 1_\Omega]) = 1 - \overline{F}(y).$$

P-boxes on a totally preordered space $(\Omega, \preceq)$ are coherent (the proof is virtually identical to the one given

in [17, p. 93, Thm. 3.59], which considered p-boxes on $[a, b] \subseteq \mathbb{R}$). We denote by $\underline{E}_{\underline{F},\overline{F}}$ the natural extension of $\underline{P}_{\underline{F},\overline{F}}$ to all gambles.

When $\underline{F} = \overline{F}$, we say that $(\underline{F}, \overline{F})$ is *precise*, and we denote the corresponding lower prevision on $\mathcal{K}$ by $\underline{P}_F$ and its natural extension to $\mathcal{L}$ by $\underline{E}_F$ (with $F := \underline{F} = \overline{F}$).

We end with a useful approximation theorem:

Theorem 2. *Let $\underline{P}$ be any coherent lower prevision defined on $\mathcal{L}$. The least conservative p-box $(\underline{F}, \overline{F})$ on $(\Omega, \preceq)$ whose natural extension is dominated by $\underline{P}$ is*

$$\underline{F}(x) = \underline{P}([0_\Omega, x])), \quad \overline{F}(x) = \overline{P}([0_\Omega, x]), \quad \forall x \in \Omega.$$

4 Natural Extension to All Events

The remainder of this paper is devoted to finding convenient expressions for the natural extension $\underline{E}_{\underline{F},\overline{F}}$ of $\underline{P}_{\underline{F},\overline{F}}$. We start by giving the form of the natural extension on the field of events generated by $\mathcal{K}$.

4.1 Extension to the Field Generated by the Domain

Let $\mathcal{H}$ be the field of events generated by the domain $\mathcal{K}$ of the p-box, i.e., events of the type

$$[0_\Omega, x_1] \cup (x_2, x_3] \cup \cdots \cup (x_{2n}, x_{2n+1}]$$

for $x_1 \prec x_2 \prec x_3 \prec \cdots \prec x_{2n+1}$ in Ω (if n is 0 we simply take this expression to be $[0_\Omega, x_1]$) and

$$(x_2, x_3] \cup \cdots \cup (x_{2n}, x_{2n+1}]$$

for $x_2 \prec x_3 \prec \cdots \prec x_{2n+1}$ in Ω. Clearly, these events form a field: the union and intersection of any two events in $\mathcal{H}$ is again in $\mathcal{H}$, and the complement of any event in $\mathcal{H}$ also is again in $\mathcal{H}$.

To simplify the description of this field, and the expression of natural extension, we introduce an element $0_\Omega-$ such that $0_\Omega- \prec x$ for all $x \in \Omega$ and:

$$F(0_\Omega-) = \underline{F}(0_\Omega-) = \overline{F}(0_\Omega-) = 0$$

So, $(0_\Omega-, x] = [0_\Omega, x]$. With $\Omega^* = \Omega \cup \{0_\Omega-\}$,

$$\mathcal{H} = \{(x_0, x_1] \cup (x_2, x_3] \cup \cdots \cup (x_{2n}, x_{2n+1}] : \quad (1)$$
$$x_0 \prec x_1 \prec \cdots \prec x_{2n+1} \text{ in } \Omega^*\}.$$

To calculate the natural extension of $\underline{P}_{\underline{F},\overline{F}}$ to all gambles, we first consider the extension from $\mathcal{K}$ to $\mathcal{H}$, then to all events, and finally to all gambles.

A precise p-box $\underline{P}_F$ has a unique extension to a finitely additive probability measure on $\mathcal{H}$:

Proposition 3. *$\underline{E}_F$ restricted to $\mathcal{H}$ is a finitely additive probability measure. Moreover, for any $A \in \mathcal{H}$, that is $A = (x_0, x_1] \cup (x_2, x_3] \cup \cdots \cup (x_{2n}, x_{2n+1}]$ with $x_0 \prec x_1 \prec \cdots \prec x_{2n+1}$ in Ω^*, it holds that*

$$\underline{E}_F(A) = \sum_{k=0}^{n} (F(x_{2k+1}) - F(x_{2k})) \quad (2)$$

Proposition 3 extends to p-boxes as follows:

Proposition 4. *For any $A \in \mathcal{H}$, that is $A = (x_0, x_1] \cup (x_2, x_3] \cup \cdots \cup (x_{2n}, x_{2n+1}]$ with $x_0 \prec x_1 \prec \cdots \prec x_{2n+1}$ in Ω^*, it holds that $\underline{E}_{\underline{F},\overline{F}}(A) = \underline{P}^{\mathcal{H}}_{\underline{F},\overline{F}}(A)$, where*

$$\underline{P}^{\mathcal{H}}_{\underline{F},\overline{F}}(A) = \sum_{k=0}^{n} \max\{0, \underline{F}(x_{2k+1}) - \overline{F}(x_{2k})\}. \quad (3)$$

For $\overline{E}_{\underline{F},\overline{F}}$, use $\overline{E}_{\underline{F},\overline{F}}(A) = 1 - \underline{E}_{\underline{F},\overline{F}}(A^c)$.

4.2 Inner Measure

The inner measure $\underline{P}^{\mathcal{H}}_{\underline{F},\overline{F}_*}$ of the coherent lower probability $\underline{P}^{\mathcal{H}}_{\underline{F},\overline{F}}$ defined in Eq. (3) coincides with $\underline{E}_{\underline{F},\overline{F}}$ on all events [18, Cor. 3.1.9, p. 127]:

$$\underline{E}_{\underline{F},\overline{F}}(A) = \underline{P}^{\mathcal{H}}_{\underline{F},\overline{F}_*}(A) = \sup_{C \in \mathcal{H}, C \subseteq A} \underline{P}^{\mathcal{H}}_{\underline{F},\overline{F}}(C). \quad (4)$$

For ease of notation, from now onwards, we denote $\underline{E}_{\underline{F},\overline{F}}$ by $\underline{E}$ when no confusion about the functions $\underline{F}$ and $\overline{F}$ determining the p-box can arise.

In principle, the problem of natural extension to all events is solved: simply calculate the inner measure as in Eq. (4), using Eq. (3) to calculate $\underline{P}^{\mathcal{H}}_{\underline{F},\overline{F}}(C)$ for elements C in $\mathcal{H}$. However, the inner measure still involves calculating a supremum. What we show next is that Eq. (3) can be extended to arbitrary events, by first taking the topological interior with respect to a very simple topology, followed by a (possibly infinite) sum over the so-called full components of this interior.

4.3 The Partition Topology

Consider the *partition topology* on Ω generated by $\tau := \{[x]_\simeq : x \in \Omega\}$. The open sets in this topology are all unions of equivalence classes (or, subsets of $\Omega/\simeq$, if you like). Hence, every open set is also closed. In particular, every interval in $(\Omega, \preceq)$ is clopen.

The topological interior of a set A is given by the union of all equivalence classes contained in A:

$$\text{int}(A) = \bigcup \{[x]_\simeq : [x]_\simeq \subseteq A\} \quad (5)$$

and the topological closure is given by the union of all equivalence classes which intersect with A:

$$\text{cl}(A) = \bigcup \{[x]_\simeq : [x]_\simeq \cap A \neq \emptyset\}. \quad (6)$$

Lemma 5. *For any subset A of Ω, $\underline{E}(A) = \underline{E}(\text{int}(A))$ and $\overline{E}(A) = \overline{E}(\text{cl}(A))$.*

4.4 Additivity on Full Components

Next, we determine a constructive expression of the natural extension $\underline{E}$ on the clopen subsets of Ω.

Definition 6. [16, §4.4] A set $S \subseteq \Omega$ is called *full* if $[a, b] \subseteq S$ for any $a \preceq b$ in S.

What do these full sets look like?

Lemma 7. *Every full set is clopen.*

Under an additional completeness assumption, the full sets are precisely the intervals.

Lemma 8. *If $\Omega/\simeq$ is order complete, that is, if every subset of $\Omega/\simeq$ has a supremum (minimal upper bound) and infimum (maximal lower bound), then every full set is an interval, that is, it can be written as $[x, y]$, $[x, y)$, $(x, y]$, or (x, y), for some x, y in Ω.*

Note that $\Omega/\simeq$ can be made order complete via the Dedekind completion [16, §4.34].

Definition 9. [16, §4.4] Given a clopen set $A \subseteq \Omega$ and an element x of A, the *full component* $C(x, A)$ of x in A is the largest full set S which satisfies $x \in S \subseteq A$.

Lemma 10. *The full components of any clopen set A form a partition of A.*

We can prove that the natural extension $\underline{E}$ is additive on full components. Recall that the sum of a family $(x_\lambda)_{\lambda\in\Lambda}$ of non-negative real numbers is defined as

$$\sum_{\lambda\in\Lambda} x_\lambda = \sup_{\substack{L\subseteq\Lambda \\ L \text{ finite}}} \sum_{\lambda\in L} x_\lambda$$

If the above sum is a finite number, at most countably many of the x_λ's are non-zero [16, 10.40].

Theorem 11. *Let B be a clopen subset of Ω. Let $(B_\lambda)_{\lambda\in\Lambda}$ be the full components of B, and let $(C_\lambda)_{\lambda\in\Lambda'}$ be the full components of B^c. Then*

$$\underline{E}(B) = \sum_{\lambda\in\Lambda} \underline{E}(B_\lambda) \text{ and } \overline{E}(B) \quad = 1 - \sum_{\lambda\in\Lambda'} \underline{E}(C_\lambda)$$

In other words, the natural extension $\underline{E}$ of a p-box is *arbitrarily additive on full components* (but obviously not additive on arbitrary events). Interestingly, additivity on full components is not sufficient for a lower probability to be equivalent to a p-box.

4.5 Practical computations over events

Let us explain how Proposition 4 can be generalized to all events (at least when $\Omega/\simeq$ is order complete).

Consider an arbitrary event A. By Lemma 5, it suffices to find the natural extension of $\text{int}(A)$ or $\text{cl}(A)$. Calculating the interior or closure with respect to the partition topology will usually be trivial (see examples further on). Because the topological interior or closure of a set is always clopen, we only need to know the natural extension of clopen sets.

Now, by Theorem 11, we only need to calculate the natural extension of the (clopen) full components $(B_\lambda)_{\lambda\in\Lambda}$ of $\text{int}(A)$ or the (clopen) full components $(C_\lambda)_{\lambda\in\Lambda}$ of $\text{cl}(A)^c = \text{int}(A^c)$. Finding the full components will often be a trivial operation. By Lemma 8, if $\Omega/\simeq$ is order complete, then each full component is an interval. And for intervals, we immediately infer from Proposition 4 and Eq. (4) that (i.p. standing for immediate predecessor):

$$\underline{E}((x, y]) = \max\{0, \underline{F}(y) - \overline{F}(x)\} \tag{7a}$$

$$\underline{E}((x, y)) = \max\{0, \underline{F}(y-) - \overline{F}(x)\} \tag{7b}$$

$$\underline{E}([x, y]) = \begin{cases} \max\{0, \underline{F}(y) - \overline{F}(x)\} & \text{if } x \text{ has no i.p.} \\ \max\{0, \underline{F}(y) - \overline{F}(x-)\} & \text{if } x \text{ has an i.p.} \end{cases} \tag{7c}$$

$$\underline{E}([x, y)) = \begin{cases} \max\{0, \underline{F}(y-) - \overline{F}(x)\} & \text{if } x \text{ has no i.p.} \\ \max\{0, \underline{F}(y-) - \overline{F}(x-)\} & \text{if } x \text{ has an i.p.} \end{cases} \tag{7d}$$

for any $x \prec y$ in Ω,[2] where $\underline{F}(y-)$ denotes $\sup_{z\prec y} \underline{F}(z)$ and similarly for $\overline{F}(x-)$. The equalities hold because, if $x \prec y$ in Ω, and $x-$ is an immediate predecessor of x, then $[x, y] = (x-, y]$ and $[x, y) = (x-, y)$. Recall also that $\underline{F}(0_\Omega-) = \overline{F}(0_\Omega-) = 0$ by convention. If $\Omega/\simeq$ is finite, then one can think of $z-$ as the immediate predecessor of z in $\Omega/\simeq$.

In other words, we have a simple constructive means of calculating the natural extension of any event.

4.6 Special Cases

The above equations hold for any $(\Omega, \preceq)$ with order complete quotient space. In most cases in practice, either $\Omega/\simeq$ is finite, or $\Omega/\simeq$ is connected, meaning that for any two elements $x \prec y$ in Ω there is a z in Ω such that $x \prec z \prec y$,[3] (this is the case for instance when Ω is a closed interval in $\mathbb{R}$ and $\preceq$ is the usual ordering of reals). Moreover, if $\Omega/\simeq$ is connected, then, in practice, $\underline{F}$ will satisfy $\underline{F}(y-) = \underline{F}(y)$ for all y in Ω. For example, in case Ω is a closed interval in $\mathbb{R}$, this happens precisely when $\underline{F}(0) = 0$ and $\underline{F}$ is left-continuous in the usual sense.

[2] In case $x = 0_\Omega$, evidently, $0_\Omega-$ is the i.p.

[3] This terminology stems from the fact that, in this case, $\Omega/\simeq$ is connected with respect to the order topology [16, §15.46(6)].

If $\Omega/\simeq$ is finite, then every element of Ω has an immediate predecessor (remember, we take the immediate predecessor of 0_Ω to be $0_\Omega-$), and if $\Omega/\simeq$ is connected, then no element except 0_Ω has an immediate predecessor. So:

Corollary 12. *If $\Omega/\simeq$ is finite, then every full set $B \subseteq \Omega$ is of the form $[a, b]$ and for every event $A \subseteq \Omega$,*

$$\underline{E}(A) = \sum_{\lambda\in\Lambda} \max\{0, \underline{F}(b_\lambda) - \overline{F}(a_\lambda-)\}$$
$$\overline{E}(A) = 1 - \sum_{\lambda\in\Lambda'} \max\{0, \underline{F}(b'_\lambda) - \overline{F}(a'_\lambda-)\}$$

where $([a_\lambda, b_\lambda])_{\lambda\in\Lambda}$ are the full components of $\operatorname{int}(A)$, and $([a'_\lambda, b'_\lambda])_{\lambda\in\Lambda'}$ are the full components of $\operatorname{int}(A^c) = \operatorname{cl}(A)^c$.

Corollary 13. *If $\Omega/\simeq$ is order complete and connected, and $\underline{F}(y-) = \underline{F}(y)$ for all y in Ω, then*

$$\underline{E}(A) = \sum_{\lambda\in\Lambda} \max\{0, \underline{F}(\sup B_\lambda) - \overline{F}(\inf B_\lambda)\}$$
$$\overline{E}(A) = 1 - \sum_{\lambda\in\Lambda'} \max\{0, \underline{F}(\sup C_\lambda) - \overline{F}(\inf C_\lambda)\}$$

where $(B_\lambda)_{\lambda\in\Lambda}$ are the full components of $\operatorname{int}(A)$ and $(C_\lambda)_{\lambda\in\Lambda'}$ are the full components of $\operatorname{int}(A^c) = \operatorname{cl}(A)^c$.

Beware of $\underline{F}(0_\Omega) = \underline{F}(0_\Omega-) = 0$ in the last corollary.

4.7 Example

Let's investigate a particular type of p-boxes on the unit square $[0, 1]^2$. First, we must specify a preorder on Ω. A natural yet naive way of doing so is, for instance, saying that $(x_1, y_1) \preceq (x_2, y_2)$ whenever $x_1 + y_1 \leq x_2 + y_2$. Consider a p-box $(\underline{F}, \overline{F})$ on $([0, 1]^2, \preceq)$. Since $\underline{F}$ is required to be non-decreasing with respect to $\preceq$, it follows that $\underline{F}(x, y)$ is constant on elements of $[0, 1]^2/\simeq$, which means that $F(x_1, y_1) = F(x_2, y_2)$ whenever $x_1 + y_1 = x_2 + y_2$. Thus, we may think of $\underline{F}(x, y)$ as a function of a single variable $z = x + y$, and we write $\underline{F}(z)$. Similarly, we write $\overline{F}(z)$.

So, our p-box specifies bounds on the probability of right-angled triangles (restricted to $[0, 1]^2$) whose hypothenuses are orthogonal to the diagonal:

$$\underline{F}(z) \leq p(\{(x, y) \in [0, 1]^2 \colon x + y \leq z\}) \leq \overline{F}(z) \quad (8)$$

Observe that the p-box is given directly on the two-dimensional product space, without the need to define marginal p-boxes for each dimension. The base τ for our partition topology is given by

$$\tau = \{\{(x, y) \in [0, 1]^2 \colon x + y = z\} \colon z \in [0, 2]\}$$

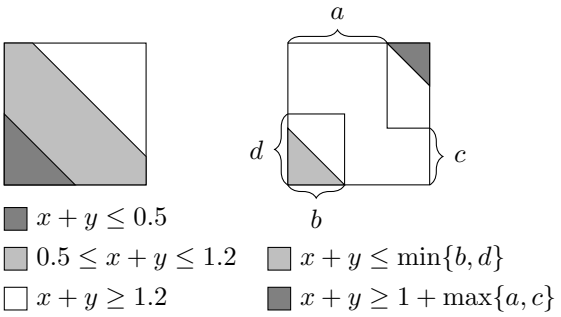


Figure 1: Shape of intervals induced by $\preceq$, and calculation of the topological interior.

For example, the topological interior of a rectangle $A = [a, b] \times [c, d]$ is empty, unless $a = c = 0$ or $b = d = 1$, because in all other cases, no element of τ is a subset of A. In the cases where $a = c = 0$ and $\min\{b, d\} < 1$, or $\max\{a, c\} > 0$ and $b = d = 1$ (if $a = c = 0$ and $b = d = 1$ then the interior is Ω), respectively, we have:

$$\operatorname{int}([0, b]\times[0, d]) = \{(x, y) \in [0,1]^2 \colon x+y \leq \min\{b, d\}\}$$
$$\operatorname{int}([a, 1]\times[c, 1]) = \{(x, y) \in [0,1]^2 \colon x+y \geq 1+\max\{a, c\}\}$$

Consequently, $\underline{E}(A) = 0$ for all rectangles A, except

$$\underline{E}([0, b] \times [0, d]) = \underline{F}(\min\{b, d\})$$
$$\underline{E}([a, 1] \times [c, 1]) = 1 - \overline{F}(1 + \max\{a, c\})$$

Fig. 1 illustrates the situation. So, for the purpose of making inferences about the lower probability of events that are rectangles, the ordering $\preceq$ was obviously poorly chosen. In general, *one should choose $\preceq$ in a way that $\Omega/\simeq$ contains good approximations for all events of interest.*

For example, a strategy would be to start from a reference point (e.g., an elicited modal value) and then to choose the ordering $\preceq$ such that intervals correspond to concentric regions of interests around the reference point. Again, all of this is possible because our theory concerns p-boxes on arbitrary totally preordered spaces, and is not limited to the real line with its natural ordering. More realistic examples in which such concentric regions are used are given in Section 8.

5 Natural Extension to All Gambles

Next, we establish that the natural extension of p-boxes to all gambles can be expressed as a Choquet integral. We further simplify the calculation of this Choquet integral via the lower and upper oscillation of gambles with respect to the partition topology introduced earlier.

5.1 Choquet Integral Representation

Extending previous results [8] where the relation between p-boxes and complete monotonicity was established for finite spaces, we can show that the natural extension of p-boxes on totally pre-ordered spaces are completely monotone. Let $\underline{P}^{\mathcal{H}}_{\underline{F},\overline{F}}$ denote the restriction of $\underline{E}_{\underline{F},\overline{F}}$ to $\mathcal{H}$, given by Proposition 4:

Theorem 14. *$\underline{P}^{\mathcal{H}}_{\underline{F},\overline{F}}$ is completely monotone.*

This allows us to characterise the natural extension on all gambles:

Theorem 15. *The natural extension $\underline{E}$ of $\underline{P}_{\underline{F},\overline{F}}$ is given by the Choquet integral*

$$\underline{E}(f) = \inf f + \int_{\inf f}^{\sup f} \underline{E}(\{f \geq t\})\,\mathrm{d}t$$

for every gamble f. Moreover, $\underline{E}$ is completely monotone on all gambles. Similarly,

$$\overline{E}(f) = \inf f + \int_{\inf f}^{\sup f} \overline{E}(\{f \geq t\})\,\mathrm{d}t.$$

5.2 Lower and Upper Oscillation

By Lemma 5, to turn Theorem 15 in an effective algorithm, we must calculate $\mathrm{int}(\{f \geq t\})$ for every t. Fortunately, there is a very simple way to do this.

For any gamble f on Ω and any topological base τ, define its *lower oscillation* as the gamble

$$\underline{osc}(f)(x) = \sup_{C \in \tau\colon x \in C} \inf_{y \in C} f(y)$$

For the partition topology which we introduced earlier, this simplifies to

$$\underline{osc}(f)(x) = \inf_{y \in [x]_\simeq} f(y) \tag{9}$$

The upper oscillation is:

$$\overline{osc}(f)(x) = -\underline{osc}(-f)(x) = \sup_{y \in [x]_\simeq} f(y) \tag{10}$$

For a subset A of Ω, the lower oscillation of I_A is $I_{\mathrm{int}(A)}$, so the lower oscillation is the natural generalisation of the topological interior to gambles. Similarly, the upper oscillation of I_A is $I_{\mathrm{cl}(A)}$.

Proposition 16. *For any gamble f on Ω,*

$$\mathrm{int}(\{f \geq t\}) = \{\underline{osc}(f) \geq t\}$$
$$\mathrm{cl}(\{f \geq t\}) = \{\overline{osc}(f) \geq t\}$$

so, in particular,

$$\underline{E}(f) = \inf \underline{osc}(f) + \int_{\inf \underline{osc}(f)}^{\sup \underline{osc}(f)} \underline{E}(\{\underline{osc}(f) \geq t\})\,\mathrm{d}t$$

$$\overline{E}(f) = \inf \overline{osc}(f) + \int_{\inf \overline{osc}(f)}^{\sup \overline{osc}(f)} \overline{E}(\{\overline{osc}(f) \geq t\})\,\mathrm{d}t$$

Concluding, to calculate the natural extension of any gamble, in practice, we must simply determine the full components of the cut sets of its lower or upper oscillation, and calculate a simple Riemann integral of a monotonic function.

Examples will be given in Section 8.

6 P-Boxes Whose Preorders are Induced by a Real-Valued Function

In practice, a convenient way to specify a preorder $\preceq$ on Ω such that $\Omega/\simeq$ is order complete and connected is by means of a bounded real-valued function $Z\colon \Omega \to \mathbb{R}$. For instance, in the example in Section 4.7, we used $Z(x, y) = x + y$. Also see [1, 12]. Let us assume from now onwards that Z is a surjective mapping from Ω to $[0, 1]$.

For any x and y in Ω, define $x \preceq y$ whenever $Z(x) \leq Z(y)$. Because Z is surjective, $\Omega/\simeq$ is order complete and connected. In particular, Ω has a smallest and largest element, for which $Z(0_\Omega) = 0$ and $Z(1_\Omega) = 1$. Moreover, we can think of any cumulative distribution function on $(\Omega, \preceq)$ as a function over a single variable $z \in [0, 1]$. Consequently, we can think of any p-box on $(\Omega, \preceq)$ as a p-box on $([0, 1], \leq)$. In particular, for any subset I of $[0, 1]$ we write $\underline{E}(I)$ for $\underline{E}(Z^{-1}(I))$. For example, for a, b in $[0, 1]$, and $A = Z^{-1}((a, b]) \subseteq \Omega$, we have that

$$\underline{E}(A) = \underline{E}((a, b]) = \max\{0, \underline{F}(a) - \overline{F}(b)\}$$

by Proposition 4. Similar expressions for other types of intervals follow from Eq. (7).

The topological interior and closure can be related to the so-called *lower and upper inverse* of Z^{-1}. Indeed, consider the multi-valued mapping $\Gamma := Z^{-1}\colon [0, 1] \to \wp(\Omega)$. Because for every x in Ω, it holds that $[x]_\simeq = \Gamma(Z(x))$, it follows that, for any subset A of Ω, $\mathrm{int}(A) = \Gamma(\Gamma_*(A))$, and $\mathrm{cl}(A) = \Gamma(\Gamma^*(A))$, where Γ_* and Γ^* denote the lower and upper inverse of Γ respectively, that is [7]

$$\Gamma_*(A) = \{z \in [0, 1]\colon \Gamma(z) \subseteq A\}, \text{ and}$$
$$\Gamma^*(A) = \{z \in [0, 1]\colon \Gamma(z) \cap A \neq \emptyset\}.$$

Theorem 17. *Let A be any subset of Ω. Then*

$$\underline{E}(A) = \sum_{\lambda \in \Lambda} \underline{E}(I_\lambda)$$
$$\overline{E}(A) = 1 - \sum_{\lambda \in \Lambda'} \underline{E}(J_\lambda)$$

where $(I_\lambda)_{\lambda \in \Lambda}$ are the full components of $Z(\mathrm{int}(A)) = \Gamma_*(A)$ and $(J_\lambda)_{\lambda \in \Lambda'}$ are the full components of $Z(\mathrm{int}(A^c)) = Z(\mathrm{cl}(A)^c) = \Gamma_*(A^c) = (\Gamma^*(A))^c$.

If, in addition, $\underline{F}$ is left-continuous as a function of $z \in [0,1]$ and $\underline{F}(0) = 0$, then

$$\underline{E}(A) = \sum_{\lambda \in \Lambda} \max\{0, \underline{F}(\sup I_\lambda) - \overline{F}(\inf I_\lambda)\}$$
$$\overline{E}(A) = 1 - \sum_{\lambda \in \Lambda'} \max\{0, \underline{F}(\sup J_\lambda) - \overline{F}(\inf J_\lambda)\}$$

For gambles, the lower oscillation is constant on equivalence classes. So, we may also consider $\underline{osc}(f)$ and $\overline{osc}(f)$ in Proposition 16 as functions of $z \in [0,1]$.

7 Constructing Multivariate P-Boxes from Marginals

Next, we construct a multivariate p-box from marginal lower previsions under arbitrary rules of combination. We then focus on two special joint models: the first without any assumptions about dependence between variables (using the Fréchet-Hoeffding bounds [13]), and the second assuming epistemic independence between all variables (using the factorization property [3]). Finally, we derive Williamson and Downs's [20] probabilistic arithmetic as a special case of our framework.

Specifically, consider n variables $X_1, \ldots, X_n$ assuming values in $\mathcal{X}_1, \ldots, \mathcal{X}_n$, and marginal lower previsions $\underline{P}_1, \ldots, \underline{P}_n$ for each variable. Each $\underline{P}_i$ is a coherent lower prevision on $\mathcal{L}(\mathcal{X}_i)$.

7.1 Multivariate P-Boxes

First, we must define a mapping Z to induce a preorder $\preceq$ on $\Omega = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. The following choice works perfectly for our purpose:

$$Z(x_1, \ldots, x_n) = \max_{i=1}^{n} Z_i(x_i)$$

where each Z_i is a surjective mapping from $\mathcal{X}_i$ to $[0,1]$ and hence, also induces a marginal preorder $\preceq_i$ on $\mathcal{X}_i$. Each $\underline{P}_i$ can be approximated by a p-box $(\underline{F}_i, \overline{F}_i)$ on $(\mathcal{X}_i, \preceq_i)$, defined by

$$\underline{F}_i(z) = \underline{P}_i(Z_i^{-1}([0,z])) \quad \overline{F}_i(z) = \overline{P}_i(Z_i^{-1}([0,z]))$$

This approximation is the best possible one, by Theorem 2.

Beware that even though different choices of Z_i may induce the same total preorder $\preceq_i$, they might lead to a different total preorder $\preceq$ induced by Z. Roughly speaking, the Z_i specify how the marginals scale relative to one another. This means that our choice of Z_i affects the precision of our inferences: a good choice will ensure that any event of interest can be well approximated by elements of $\Omega / \simeq$. Of course, nothing prevents us, at least in theory, to consider the set of all Z_i which induce some given marginal total preorders $\preceq_i$, and whence to work with a set of p-boxes. In Section 7.4, we will see an example where this approach is feasible.

Anyway, with this choice of Z, we can easily find the p-box which represents the joint as accurately as possible, under any rule of combination of coherent lower previsions:

Theorem 18. *Consider any rule of combination $\odot$ of coherent lower and upper previsions, mapping the marginals $\underline{P}_1, \ldots, \underline{P}_n$ to a joint coherent lower prevision $\bigodot_{i=1}^{n} \underline{P}_i$ on all gambles. Suppose there are functions ℓ and u for which:*

$$\bigodot_{i=1}^{n} \underline{P}_i \left(\prod_{i=1}^{n} A_i \right) = \ell(\underline{P}_1(A_1), \ldots, \underline{P}_n(A_n)) \text{ and}$$
$$\bigodot_{i=1}^{n} \overline{P}_i \left(\prod_{i=1}^{n} A_i \right) = u(\overline{P}_1(A_1), \ldots, \overline{P}_n(A_n)),$$

for all $A_1 \subseteq \mathcal{X}_1, \ldots, A_n \subseteq \mathcal{X}_n$. Then, the couple $(\underline{F}, \overline{F})$ defined by

$$\underline{F}(z) = \ell(\underline{F}_1(z), ..., \underline{F}_n(z)); \overline{F}(z) = u(\overline{F}_1(z), ..., \overline{F}_n(z))$$

is the least conservative p-box on $(\Omega, \preceq)$ whose natural extension $\underline{E}_{\underline{F},\overline{F}}$ is dominated by the combination $\bigodot_{i=1}^{n} \underline{P}_i$ of $\underline{P}_1, \ldots, \underline{P}_n$.

7.2 Natural Extension: The Fréchet Case

The *natural extension* $\boxtimes_{i=1}^{n} \underline{P}_i$ of $\underline{P}_1, \ldots, \underline{P}_n$ is the lower envelope of all joint distributions whose marginal distributions are compatible with the given marginal lower previsions. So, the model is completely vacuous about the dependence structure. We refer to for instance [4, p. 120, §3.1] for a rigorous definition. In this paper, we only need to use the Fréchet bounds (see [21, p. 131]), in which case the functions ℓ and u of Theorem 18 are respectively the Lukasiewicz and the minimum t-norms.

Theorem 19. *The p-box $(\underline{F}, \overline{F})$ defined by*

$$\underline{F}(z) = \max\left\{0, 1 - n + \sum_{i=1}^{n} \underline{F}_i(z)\right\} \quad \overline{F}(z) = \min_{i=1}^{n} \overline{F}_i(z)$$

is the least conservative p-box on $(\Omega, \preceq)$ whose natural extension $\underline{E}_{\underline{F},\overline{F}}$ is dominated by the natural extension $\boxtimes_{i=1}^{n} \underline{P}_i$ of $\underline{P}_1, \ldots, \underline{P}_n$.

It is easily seen that the joint lower prevision $\boxtimes_{i=1}^{n} \underline{P}_i$ is in general not completely monotone, hence the joint p-box of Theorem 19 is in general only an outer approximation.

7.3 Independent Natural Extension

In contrast, the *independent natural extension* $\otimes_{i=1}^{n} \underline{P}_i$ of $\underline{P}_1, \ldots, \underline{P}_n$ models epistemic independence between $X_1, \ldots, X_n$. We refer to [3] for a rigorous definition and properties. In this paper we only need the factorization property, which implies that the functions ℓ and u of Theorem 18 are the product rule.

Theorem 20. *The p-box* $(\underline{F}, \overline{F})$ *defined by*

$$\underline{F}(z) = \prod_{i=1}^{n} \underline{F}_i(z) \qquad \overline{F}(z) = \prod_{i=1}^{n} \overline{F}_i(z)$$

is the least conservative p-box on $(\Omega, \preceq)$ *whose natural extension* $\underline{E}_{\underline{F},\overline{F}}$ *is dominated by the indepedent natural extension* $\otimes_{i=1}^{n} \underline{P}_i$ *of* $\underline{P}_1, \ldots, \underline{P}_n$.

Again, the joint p-box will only be an outer approximation of the actual joint lower prevision.

7.4 Special Case: Probabilistic Arithmetic

Let $Y = X_1 + X_2$ with X_1 and X_2 real-valued random variables. Probabilistic arithmetic [21] estimates $\underline{P}_Y([-\infty, y]) = \underline{F}_Y(y)$ and $\overline{P}_Y([-\infty, y]) = \overline{F}_Y(y)$ for any $y \in \mathbb{R}$ under the assumptions that the uncertainty on X_1 and X_2 is given by p-boxes $(\underline{F}_1, \overline{F}_1)$ and $(\underline{F}_2, \overline{F}_2)$, with $\preceq_1$ and $\preceq_2$ the natural ordering of real numbers, and the dependence structure is completely unknown. Williamson and Downs [20] provide explicit formulae for common arithmetic operations, making inferences from marginal p-boxes very easy.

Let us show, for the particular case of addition, that their results are captured by our joint p-box proposed in Theorem 19. Cases of other arithmetic operators, not treated here to save space, follow from almost identical reasoning. The lower cumulative distribution function $\underline{F}_{X_1+X_2}(y)$ resulting from probabilistic arithmetic is, for any $y \in \mathbb{R}$,

$$\sup_{x_1, x_2 \colon x_1 + x_2 = y} \max\{0, \underline{F}_1(x_1) + \underline{F}_2(x_2) - 1\}. \quad (11)$$

Without much loss of generality, assume that both X_1 and X_2 lie in a bounded interval $[a, b]$.

Let Z_1 and Z_2 be any surjective maps $[a, b] \to [0, 1]$ which induce the usual ordering on $[0, 1]$ (so both must be continuous and strictly increasing).

To apply Theorem 19, consider the total preorder $\preceq$ on $\Omega = [a, b]^2$ induced by $Z(x_1, x_2) = \max\{Z_1(x_1), Z_2(x_2)\}$. Figure 2 illustrates the event[4] $\{X_1 + X_2 \leq y\}$, with $y \in [2a, 2b]$, as well as the largest interval $Z^{-1}([0, z])$ included in it. For z such that

[4] $\{X_1 + X_2 \leq y\}$ is $\{(x_1, x_2) \in [0, 1]^2 \colon x_1 + x_2 \leq y\}$.

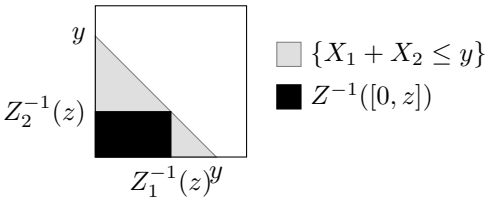


Figure 2: The event $\{X_1 + X_2 \leq y\}$, and the largest interval $Z^{-1}([0, z])$ included in it.

$Z_1^{-1}(z) + Z_2^{-1}(z) = y$, we achieve the largest interval $Z^{-1}([0, z])$ which is still included in $\{X_1 + X_2 \leq y\}$. There is always a unique such z because also $Z_1^{-1} + Z_2^{-1}$ is continuous and strictly increasing.

Using Theorems 19 and 17, we find that

$$\begin{aligned} \underline{E}_{\underline{F},\overline{F}}(\{X_1 + X_2 \leq y\}) &= \underline{F}(Z^{-1}(z)) \\ &= \max\{0, \underline{F}_1(Z_1^{-1}(z)) + \underline{F}_2(Z_2^{-1}(z)) - 1\} \end{aligned}$$

But, this holds for every valid choice of Z_1 and Z_2, whence $\underline{P}_1 \boxtimes \underline{P}_2(\{X_1 + X_2 \leq y\})$ dominates Eq. (11).

8 Example

Next, we investigate an example in which p-boxes are used to model uncertainty around some parameters.

We aim to estimate the minimal required dike height h along a stretch of river, using a model proposed in [6]. Although this model is quite simple, it provides a realistic industrial application. Skipping technical details, the model results in the following relationship:

$$h(q, k, u, d) = \begin{cases} \left(\frac{q}{k\sqrt{\frac{u-d}{\ell}}b} \right)^{\frac{3}{5}} & \text{if } q \geq 0 \\ 0 & \text{otherwise,} \end{cases} \quad (12)$$

with b and ℓ the river width and length, q the river flow rate, k the Strickler coefficient and u, d respectively the upriver and downriver water levels.

For this case study, the river width is $b = 300m$ and the length is $\ell = 6400m$. The remaining parameters are uncertain. Expert assessment leads to the following distributions.

The river flow rate q has a Gumbel distribution with location and scale parameters $\mu = 1335m^3s^{-1}$ and $\beta = 716m^3s^{-1}$. To simplify calculations, we introduce a variable r satisfying $q = \mu - \beta \ln(-\ln(r))$. If r is uniform over $[0, 1]$, then q is Gumbel with parameters μ and β. So, after transformation,

$$h(r, k, u, d) = \begin{cases} \left(\frac{\mu - \beta \ln(-\ln(r))}{k\sqrt{\frac{u-d}{\ell}}b} \right)^{\frac{3}{5}} & \text{if } q \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

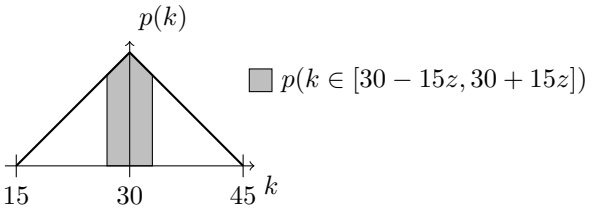


Figure 3: Derivation of the p-box for a triangular distribution.

The Strickler coefficient k has a symmetric triangular distribution over the interval $[15m^{1/3}s^{-1}, 45m^{1/3}s^{-1}]$.

Upper and downriver water levels u and d are uncertain due to sedimentary conditions. Measured values are $u^* = 55m$ and $d^* = 50m$, with measurement error definitely less than $1m$. These are also modelled by symmetric triangular distributions, on $[54m, 56m]$ and $[49m, 51m]$ respectively.

A natural choice for Z is the distance between the expected values $(r^* = 1/2, k^* = 30, u^* = 55, d^* = 50)$ and the actual values (r, k, u, d):

$$Z(r,k,u,d) = \max\{2|r - \tfrac{1}{2}|, \tfrac{|k-30|}{15}, |u-55|, |d-50|\}.$$

The scale of the distances has been chosen such that $Z(r,k,u,d) \leq 1$ for all points of interest. Equivalence classes $[(r,k,u,d)]_{\sim}$ are borders of 4-dimentional boxes with vertices (with $z = Z(r,k,u,d)$)

$$((1 \pm z)/2, 30 \pm 15z, 55 \pm z, 50 \pm z).$$

The marginal p-boxes are, for r:

$$\underline{F}_1(z) = \overline{F}_1(z) = p(2|r - 1/2| \leq z) = z$$

because r is uniformly distributed over $[0, 1]$. For k:

$$\underline{F}_2(z) = \overline{F}_2(z) = p(|k - 30|/15 \leq z) = 1 - (1-z)^2$$

(see Fig. 3). Similarly, for u and d, it is easily verified that $\underline{F}_3(z) = \overline{F}_3(z) = \underline{F}_4(z) = \overline{F}_4(z) = 1 - (1-z)^2$.

Next, $\underline{osc}(h)$ and $\overline{osc}(h)$ are:

$$\underline{osc}(h)(z) = \inf_{(r,k,u,d):\ Z(r,k,u,d)=z} h(r,k,u,d) = o(-z)$$
$$\overline{osc}(h)(z) = \sup_{(r,k,u,d):\ Z(r,k,u,d)=z} h(r,k,u,d) = o(z)$$

with

$$o(z) = \begin{cases} \left(\frac{\mu - \beta \ln(-\ln((1+z)/2))}{(30-15z)\sqrt{\frac{5-2z}{\ell}}\, b} \right)^{\frac{3}{5}} & \text{if } \cdots \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

The function $o(z)$ is increasing, with $o(-1) = 0$, $o(0) = 3.032$, and $o(1) = +\infty$.

Hence, $\underline{osc}(h)(z)$ and $\overline{osc}(h)(z)$ are decreasing and increasing in z, respectively. So, the full components of the events

$$L_t = \{z \in [0,1] \colon \underline{osc}(h)(z) \geq t\} = \{z \in [0,1] \colon o(-z) \geq t\}$$
$$U_t = \{z \in [0,1] \colon \overline{osc}(h)(z) \geq t\} = \{z \in [0,1] \colon o(z) \geq t\}$$

are of the form $L_t = [0, \ell_t]$ and $U_t = [u_t, 1]$, with

$$\ell_t = -o^{-1}(t) \text{ for } t \leq o(0) \quad u_t = o^{-1}(t) \text{ for } t \geq o(0)$$

With unknown dependence, using Theorem 19,

$$\underline{F}(z) = \max\{0, -3 + z + 3(1 - (1-z)^2)\}$$

and whence

$$\underline{E}(h) = \int_0^{o(0)} \underline{F}(-o^{-1}(t))\, \mathrm{d}t = 1.515$$
$$\overline{E}(h) = o(0) + \int_{o(0)}^{+\infty} \left(1 - \underline{F}(o^{-1}(t))\right) \mathrm{d}t = 6.423$$

Therefore, we should consider average overflowing heights of at least $6.5m$. For comparison, using traditional methods and assuming independence between all variables, h has expectation $3.2m$, which lies between our lower and upper expectation, as expected. Note that the imprecision has two sources: we have reduced a multivariate problem to a univariate one and we have not made any assumption of independence.

Calculations were relatively simple due to the monotonicity of the target function with respect to the uncertain variables. This may not be the case in general.

9 Conclusions

We studied inferences (lower and upper expectations) from p-boxes on arbitrary totally preordered spaces. For this purpose, we represented p-boxes as coherent lower previsions, and studied their natural extension. Defining p-boxes on totally pre-ordered spaces allowed us to unify p-boxes on finite spaces and on real intervals, and to extend the theory to the multivariate case.

One interesting result is a practical means of calculating the natural extension of a p-box in this general setting: we proved that it suffices to calculate the full components of the cut sets of the lower oscillation, followed by a simple Riemann integral (Proposition 16).

As examples of how this model can be used in practice, we have detailed the cases of p-boxes whose preorders are induced by a real-valued mapping, and of joint p-boxes built from marginals under various combination rules. We demonstrated our methodology on inference about a river dike assessment, showing that calculations are generally straightforward.

Of course, many open problems regarding p-boxes remain. For instance, can the dependency model inform the choice of preorder, to arrive at tighter bounds? Our choice led to simple expressions, but other choices giving more precise inference could be investigated. Also, the connection of p-boxes with other uncertainty models, such as possibility measures and clouds, deserves further investigation.

Acknowledgements

We are particularly grateful to Enrique Miranda for the many very fruitful discussions, extremely useful suggestions, and various contributions to this paper. We also thank Gert de Cooman and Didier Dubois for their help with a very early draft of this paper. Finally, we thank both reviewers for their valuable comments and suggestions.

References

[1] Y. Ben-Haim. *Info-gap decision theory: decisions under severe uncertainty*. Academic Press, London, 2006.

[2] G. Boole. *An investigation of the laws of thought on which are founded the mathematical theories of logic and probabilities*. Walton and Maberly, London, 1854.

[3] G. de Cooman, E. Miranda, and M. Zaffalon. Independent natural extension. In E. Hüllermeier, R. Kruse, and F. Hoffmann, editors, *Computational Intelligence for Knowledge-Based Systems Design*, Lecture Notes in Computer Science, pages 737–746. Springer, 2010.

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

[5] G. de Cooman, M. C. M. Troffaes, and E. Miranda. n-Monotone exact functionals. *Journal of Mathematical Analysis and Applications*, 347(1):143–156, 2008.

[6] E. de Rocquigny, N. Devictor, and S. Tarantola. *Uncertainty in Industrial Practice: A guide to Quantitative Uncertainty Management*, chapter 10. Partial Safety Factors to Deal with Uncertainties in Slope Stability of River Dykes. Wiley, 2008.

[7] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[8] S. Destercke, D. Dubois, and E. Chojnacki. Unifying practical uncertainty representations: I. Generalized p-boxes. *International Journal of Approximate Reasoning*, 49(3):649–663, 2008.

[9] S. Ferson, V. Kreinovich, L. Ginzburg, D. S. Myers, and K. Sentz. Constructing probability boxes and Dempster-Shafer structures. Technical Report SAND2002–4015, Sandia National Laboratories, January 2003.

[10] S. Ferson and W. Tucker. Sensitivity analysis using probability bounding. *Reliability engineering and system safety*, 91(10-11):1435–1442, 2006.

[11] S. Ferson and W. Tucker. Probability boxes as info-gap models. In *Proceedings of the Annual Meeting of the North American Fuzzy Information Processing Society*, New York (USA), 2008.

[12] M. Fuchs and A. Neumaier. Potential based clouds in robust design optimization. *Journal of Statistical Theory and Practice, Special Issue on Imprecision*, 3(1):225–238, 2008.

[13] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the Americal Statistical Association*, 58:13–30, 1963.

[14] E. Kriegler and H. Held. Utilizing belief functions for the estimation of future climate change. *International Journal of Approximate Reasoning*, 39:185–209, 2005.

[15] E. Miranda. A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 48(2):628–658, 2008.

[16] E. Schechter. *Handbook of Analysis and Its Foundations*. Academic Press, San Diego, CA, 1997.

[17] M. C. M. Troffaes. *Optimality, Uncertainty, and Dynamic Programming with Lower Previsions*. PhD thesis, Ghent University, Ghent, Belgium, March 2005.

[18] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[19] P. M. Williams. Notes on conditional previsions. Technical report, School of Mathematical and Physical Science, University of Sussex, UK, 1975.

[20] R. C. Williamson and T. Downs. Probabilistic arithmetic I: Numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning*, 4:89–158, 1990.

[21] R.C. Williamson. *Probabilistic Arithmetic*. PhD thesis, University of Queensland, Australia, 1989.

Robust detection of exotic infectious diseases in animal herds: A comparative study of two decision methodologies under severe uncertainty

Matthias C. M. Troffaes
Durham University
United Kingdom
matthias.troffaes@gmail.com

John Paul Gosling
The Food and Environment Research Agency (FERA)
United Kingdom
johnpaul.gosling@fera.gsi.gov.uk

Abstract

When animals are transported and pass through customs, some of them may have dangerous infectious diseases. Typically, due to the cost of testing, not all animals are tested: a reasonable selection must be made. How to test effectively, yet avoid cataclysmic events? First, we extend a model proposed in the literature for the detection of invasive species to suit our purpose. Secondly, we explore and compare two decision methodologies on the problem at hand, namely, info-gap theory and imprecise probability theory, both of which are designed to handle severe uncertainty. We show that, under rather general conditions, every info-gap solution is maximal with respect to a suitably chosen imprecise probability model, and that therefore, perhaps surprisingly, the set of maximal options can be inferred at least partly—and sometimes entirely—from an info-gap analysis.

Keywords. exotic disease, lower prevision, info-gap, maximality, minimax, robustness, inspection, protocol

1 Introduction

This paper concerns the inspection of imported herds of animals for signs of known or unknown major exotic infectious diseases. On the one hand, imports and exports of animals represent a significant contribution to the UK economy. On the other hand, there is a real risk of animal diseases being introduced. Imports are therefore subject to strict controls at the UK border under EU and national rules. Fèvre et al. [6] review the problems associated with animal movement and the spread of disease.

We will build further on the work of Moffitt et al. [10], who study inspection protocols for shipping containers of invasive species, employing info-gap theory [1] to model the severely uncertain number of infested items. The aim of their study is to realistically take into account economical considerations (actual costs of testing, and of invasive species passing through customs), whilst also soundly handling the enormous uncertainty.

A key feature of their, and also our, problem is that exact probabilities of the constituent events are very hard to come by [9]. This motivates the use of robust uncertainty models and decision tools, such as info-gaps [1] (i.e. robust satisficing) as in the original study, but also imprecise probabilities [12], as we will do in this paper.

Our study, using both decision methodologies, leads us to surmise a connection between info-gap analysis and imprecise probability theory (Γ-minimax and maximality in particular). We prove that the perceived connection is no coincidence, and we establish a rigorous theoretical link between the two approaches.

The paper is organised as follows. Section 2 introduces the problem of animal inspection, defines the model, discusses various uncertainties involved, and derives an expression for the expected loss under a simple binomial model for infection. Section 3 solves the inspection problem, first using an info-gap model, and then using an imprecise probability model (with maximality). These results are discussed in Section 4, where we formally define an info-gap model based on a nested set of imprecise probability models, and establish the theoretical connections between info-gap, Γ-minimax, and maximality. Section 5 concludes the paper.

2 Animal Herd Testing

In this section, we extend a model, proposed by [10] for the detection of invasive species, to suit our purpose:

- we explicitly take specificity and sensitivity into account in order to allow for imperfect testing,

- we take into account an additional cost term for terminating the herd in case an infection is detected, and
- we model the occurrence of diseased animals in the herd as a binomial process, under a worst-case assumption of independence of infections between animals.

2.1 Model Description

Consider a herd of n animals, of which m are tested—the problem is to choose m optimally. The uncertain number of diseased animals in the herd is denoted by d. The test has sensitivity—the probability that a diseased animal tests positive—equal to p, and specificity—the probability that a healthy animal tests negative—equal to q.

Testing m animals costs $c(m)$ utiles. If d diseased animals pass inspection undetected, we incur a cost of $a(d)$ utiles. When at least one diseased animal is detected, then, typically, the whole herd is terminated, costing $t(n)$ utiles.

Following [10, p. 295, Sec. 3], in the numerical examples that follow, we take

$$c(m) = 1000 - 2000m + 1000m^2 \quad (m \geq 1)$$

$$a(d) = \begin{cases} 0 & \text{if } d = 0 \\ a & \text{if } d \geq 1 \end{cases} \qquad (a = 10\,000\,000)$$

Moffitt et al. [10] consider n between 250 and 2 500, do not need to consider the cost of termination ($t(n) = 0$), and assume perfect testing ($p = q = 1$). For our problem, we take

$$\begin{aligned} n &= 250 \\ t(n) &= 400n = 100\,000 \\ p &= 0.9999 \\ q &= 0.999 \end{aligned}$$

so we assume that a diseased animal tests positive with probability 0.9999, and a healthy animal tests negative with probability 0.999. For reference, if $q = 0.999$, then probability that all animals in a healthy herd of size $n = 250$ test negative is $q^n = 0.78$. These values for p and q are reasonable in so far that, in practice, things would be really bad if they were any lower.

2.2 Model Uncertainties

Obviously, many of these values are rather uncertain. The only values we are pretty certain of are the number of animals n in the herd, the cost of testing $c(n)$, and the cost of termination $t(n)$.

Due to the necessity that the herd must have valid health documentation, we would expect that the number of infected animals d would be low. Additional inspection by veterinary officials is costly and depends on the inspecting official's ability to spot signs of infectious disease like pathological lesions and abnormal behaviour. Of course, the level of experience and competency will vary from official to official, but the testing procedure should be thorough enough for us to be confident of both a high sensitivity, p, and specificity, q. In addition to this, the government would prefer the most sensitive test possible (within budgetary constraints), even if specificity was slightly compromised, because a rare false positive would be better for the prevention of disease entry than a rare false negative. Hence, we would expect $p > q$. Further discussion of this can be found in [15].

Regarding the cost a of an infection passing through customs, some historical data is available. For example, instances of major disease outbreaks in the last couple of decades include BSE where public spending was over £5 billion, and the foot and mouth outbreak in 2001 which costed the UK government £2.6 billion [4]. These experiences show that there is great variation in the level of costs of exotic disease outbreaks. Due to the exceptional nature of the outbreaks, there is limited evidence on which to base cost assessments. Therefore, there is great uncertainty about what may happen in the future.

Outbreaks of any particular exotic disease are generally rare or may never have occurred at all. Also, diseases change as new strains develop, and the possibility of new diseases arriving into the UK can change rapidly. For example, until a few years ago, bluetongue was considered extremely unlikely, but now we expect an outbreak every one to two years in the UK.

In late 2009, an elicitation exercise was carried out with government experts to help quantify the average annual costs to the UK government of exotic infectious disease outbreaks and the uncertainty about those estimates [8]. In that exercise, it was clear that the costs are severely uncertain even when the disease was known (for example, foot and mouth is an exotic infectious disease). A major contributor to the uncertainty about the overall cost was the possibility of an outbreak of an unknown infectious disease, which could cost anywhere from £0.5 billion to £6 billion.

The scale and costs of an outbreak will depend on the length of time between the diseased animal entering circulation and the disease's presence being confirmed, and the speed and effectiveness of the government's response. The eventual costs are influenced

by any public health implications and the effects of disease controls on other industries. The main elements of the costs due to control measures include: the disposal of and payments for culled animals; the tracing, testing and diagnosis of animals; the cleaning and disinfection of infected premises; and administrative costs in managing the outbreak. The size of these costs will vary according to the scale of the outbreak with key factors being the number of infected premises, the numbers of animals culled, and the duration of the outbreak. These types of factors are considered in greater detail in [4] and [7].

A serious study of how all uncertainties involved could be taken into account in the model would of course be extremely interesting, but is beyond the goal of this paper. Instead, in this initial study, following [10] and many others, for now we will focus on the main uncertainty, that is, the number of diseased animals d, and simply assume reasonable values for the remaining parameters.

2.3 Expected Loss

First, we derive the expected loss, in case all parameters of the problem are perfectly known, including the number of diseased animals d. Clearly, conditional on d, the expected loss is:

$$\begin{aligned} &L(m,d,p,q,c,a,t) \\ &\qquad = c(m) + t(n)\Pr(T|d) + a(d)\Pr(T^c|d) \end{aligned}$$

where T denotes termination of the herd, that is, the event that at least one diseased animal is detected, and T^c denotes its complement, that is, the event that the herd passes inspection.

Let us deduce $\Pr(T^c|d)$. First, if the test group of size m is sampled randomly and without replacement, then the probability of exactly z diseased animals in the test group follows a hypergeometric distribution:

$$\Pr(z|d) = \frac{\binom{d}{z}\binom{n-d}{m-z}}{\binom{n}{m}}.$$

Next, we calculate the probability of non-termination given z diseased animals in the test group, that is $\Pr(T^c|d,z)$. If $d = 0$, then the probability of non-termination is the probability of all healthy animals in the sample testing negative, so $\Pr(T^c|0,z) = q^m$. If $d \geq 1$, then given z diseased animals in the sample, non-termination occurs when none of the z diseased animals tests positive and all of the $m - z$ healthy animals test negative. Hence, in all cases,

$$\Pr(T^c|d,z) = (1-p)^z q^{m-z}. \tag{1}$$

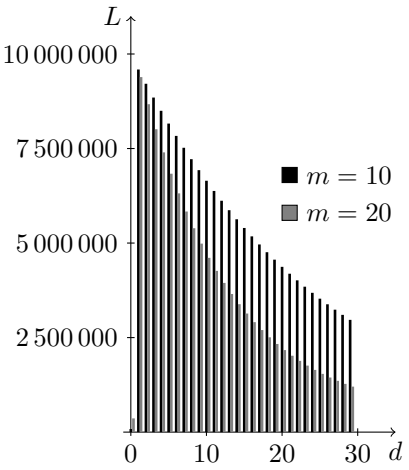


Figure 1: Loss as a function of the number of diseased animals for $m = 10$ and $m = 20$.

By the law of total probability,

$$\begin{aligned} \Pr(T^c|d) &= \sum_{z=0}^{d} \Pr(T^c|d,z)\Pr(z|d) \\ &= \sum_{z=0}^{d} (1-p)^z q^{m-z} \frac{\binom{d}{z}\binom{n-d}{m-z}}{\binom{n}{m}}. \end{aligned} \tag{2}$$

Now we have all the ingredients to calculate the total expected loss if we choose to test m out of n animals:

$$\begin{aligned} &L(m,d,p,q,c,a,t) \\ &= c(m) + t(n) + (a(d) - t(n))\Pr(T^c|d) \end{aligned}$$

or, if $a'(n,d) = a(d) - t(n)$ denotes the termination adjusted cost of apocalypse,

$$= c(m) + t(n) + a'(n,d)\Pr(T^c|d)$$

where $\Pr(T^c|d)$ is given by Eq. (2). Figure 1 depicts the expected loss for a few typical cases.

2.4 A Binomial Model for Infection

Moffitt et al. [10] consider an info-gap model directly over the number of diseased animals d, which leads to a rather tricky optimisation problem. Instead, we will consider the (highly uncertain) probability r that an animal is infected, and derive the expected loss as a function of r. Although we do not explore this topic further in this paper, this also paves the way to modelling spatial dependencies between infections in the herd, leading to more optimal testing strategies.

So, assume that each animal has a probability r of being infected; for simplicity, for now, we assume that

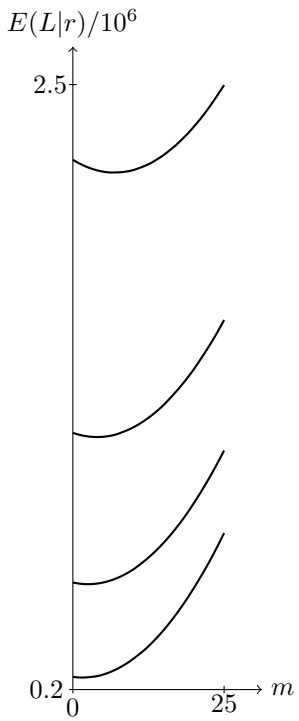


Figure 2: Expected loss $L(m|r)$ as a function of the test group size m, for $r = 0.00010$, $r = 0.00025$, $r = 0.00050$, and $r = 0.00100$, from bottom to top.

one animal being diseased does not affect another animal being diseased. Obviously, this will generally not be satisfied, and more realistically, we would expect a positive correlation, resulting in diseased animals being clustered together in the herd. Assuming independence essentially amounts to a worst case study: at the other extreme end, if one diseased animal would immediately infect the whole herd, then it would be sufficient to test only a single animal, as $d = 0$ and $d = n$ would be the only two possibilities.

Under the worst case assumption of independence, the probability of having d out of n animals infected is:

$$\Pr(d|r) = \binom{n}{d} r^d (1-r)^{n-d} \tag{3}$$

The expected loss is:

$$E(L(m,\cdot,p,q,c,a,t)|r) = \sum_{d=0}^{n} L(m,d,p,q,c,a,t) \Pr(d|r) \tag{4}$$

From now onwards, we will simply write $L(m|r)$ instead of $E(L(m,\cdot,p,q,c,a,t)|r)$ in order to simplify notation. Figure 2 depicts $L(m|r)$ as a function of m for a few typical situations.

3 Decision Analysis

In this section, we explore and compare two decision methodologies, designed for severe uncertainty, on the problem at hand. In particular,

- we accommodate the info-gap approach suggested by [10] to our extended model,
- we investigate possible ways of constructing sets of probabilities (i.e. imprecise probability models) which are in some sense equivalent to the proposed info-gap model, and
- we compare the decisions that these various models lead to.

3.1 Info-Gap Analysis

One approach to solve our decision problem, under severe uncertainty about the exact probability r of a single animal being viciously infected, is to select that decision which meets a given performance criterion, L_c, under the largest possible range of r. Given that we have almost no information about r, this simple model seems to suffice for our purpose. Obviously, one could define many other more refined info-gap models—and our choice of model is just one example among many. For a much more detailed account, see [1].

Specifically, for a given value of L_c, the largest possible range $[0, h]$ of r for which we meet our performance criterion is characterised by

$$\hat{h}(m, L_c) = \max_{h \geq 0} \left\{ h \colon \underbrace{\max_{r \in [0,h]} L(m|r)}_{M(m,h)} \leq L_c \right\}$$

The value $\hat{h}(m, L_c)$, as a function of L_c, is called the *robustness curve*: it tells us how uncertain about r we can be for our decision m still to meet a given level of performance L_c.

A quick Poisson approximation reveals that as long as $\exp(-nh)$ is sufficiently close to 1 (and this holds for sufficiently small values of nh) the inner maximum over $r \in [0, h]$ is achieved at $r = h$ (also see Figure 2: the cost increases as r increases), so

$$M(m, h) = L(m|h)$$

Obviously, $M(m, h)$ increases as the horizon of uncertainty h increases, whence $\hat{h}(m, L_c)$ as a function of L_c is simply the inverse of $M(m, h)$ as a function of h. In other words, plotting $M(m, h)$ as a function

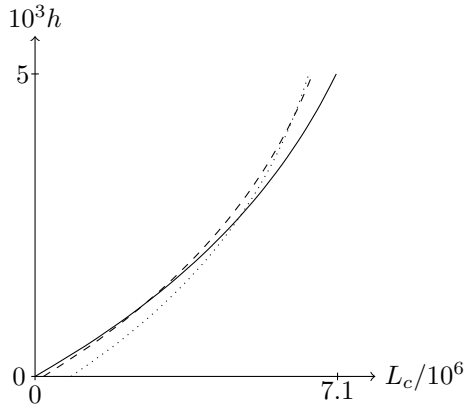


Figure 3: Robustness curves $\hat{h}(m, L_c)$ as a function L_c for test group sizes $m = 1$ (solid), $m = 15$ (dashed), and $m = 30$ (dotted).

$L_c/10^6$	m^*	$10^3\hat{h}(m^*, L_c)$
0.5	2	0.207
1.5	5	0.661
2.5	8	1.184
3.5	11	1.803

Table 1: Info-gap choice of m, and corresponding horizon of uncertainty, for various values of the critical cost L_c.

of h for different values of m effectively gives us the robustness curves. Figure 3 depicts them.

The choices of m which maximise robustness, for various values of the critical cost L_c, are tabulated in Table 1. For example, at a cost of at most $L_c = 2\,500\,000$, we can safeguard against any probability of infection $r \in [0, 0.001\,184]$, by testing 8 animals in the herd.

3.2 Imprecise Probability Analysis: Maximality

There are several ways one might go about constructing an imprecise probability model for our problem. As we have just seen, the info-gap approach hinges on the idea of satisficing. We may start out with a level of minimum performance that we hope to achieve, and the analysis tells us how much uncertainty we can account for, at this price. One might also interpret it conversely: for a given level of uncertainty, the analysis tells us how much we might potentially pay, if it comes to the worst.

Typical decision models for imprecise probabilities studied in the literature do not relate to satisficing, yet, they do incorporate an idea similar to the info-gap horizon of uncertainty: the imprecision of our model. Concretely, consider the set $\mathcal{M}_h$ of all probability densities over r that are zero outside $[0, h]$.[1] We say that a choice m *dominates* a choice m', and we write $m \succ m'$ whenever the expected loss under m is strictly less than the expected loss under m' over all densities p in $\mathcal{M}_h$, that is, whenever

$$\int_0^\infty L(m|r)p(r)dr + \epsilon \leq \int_0^\infty L(m'|r)p(r)dr$$

for all probability densities p in $\mathcal{M}_h$ and some $\epsilon > 0$.

This happens exactly when

$$\min_{r\in[0,h]} [L(m'|r) - L(m|r)] > 0$$

Note that the $\min_{r\in[0,h]}$ operator can be thought of as a lower expectation operator, or *lower prevision* $\underline{P}_h$—we will come back to this in Section 4.

One can easily prove that $\succ$ is a partial order, whence, a sensible way to choose m is to pick one which is not dominated by any other option, or in other words, which is *maximal*. The idea of choosing undominated options goes back at least to Condorcet [3, pp. lvj–lxix, 4.^{e} Exemple]; also see [11, p. 55, Eq. (1)], [13, Sections 3.7–3.9], and [12] for further discussion.

Given our partial order, one can easily show that an option m is maximal if and only if

$$\min_{m'\in\{0,1,\ldots,n\}} \max_{r\in[0,h]} [L(m'|r) - L(m|r)] \geq 0 \qquad (5)$$

The inner maximum is almost always achieved at either $r = 0$ or $r = h$, simplifying practical calculations substantially. Table 2 depicts these values for all choices of m, and varying values of h. For ease of comparison with the info-gap solution, we have chosen the same values of h as those listed in Table 1.

4 Discussion

Interestingly, info-gap and maximality give essentially the same result, with maximality refining the picture slightly: for a given horizon of uncertainty h, the maximal solutions are $\{1, \ldots, m^*\}$, where m^* is the info-gap solution. The most notable result is that all info-gap solutions are maximal. Is this a coincidence? Formulating info-gap theory in terms of lower previsions, we show that this holds under fairly general circumstances.

[1] The adventurous reader may take all finitely additive probability measures μ on $[0, +\infty]$ with $\mu([0, h]) = 1$. We do without this complication: because all functions involved are continuous, those additional measures make no difference.

	$10^3 h$			
m	0.207	0.661	1.184	1.803
0	−0.9	−0.9	−0.9	−0.9
1	1.1	1.1	1.1	1.1
2	1.4	3.1	3.1	3.1
3	−0.6	4.9	5.1	5.1
4	−3.1	2.9	7.1	7.1
5	−7.7	0.9	7.0	9.1
6	−14.3	−1.1	5.0	11.1
7	−22.9	−4.3	2.9	9.9
8	−33.4	−9.5	0.9	7.9
9	−46.0	−16.6	−1.1	5.8
10	−60.6	−25.9	−4.3	3.7
11	−77.2	−37.1	−9.5	1.7
12	−95.8	−50.3	−16.8	−0.4
13	−116.4	−65.6	−26.1	−2.9
14	−139.1	−82.9	−37.4	−7.4
15	−163.7	−102.2	−50.8	−14.1

Table 2: Result of Eq. (5) (divided by a factor 10^3 for everything to fit in the table). A positive value means that the corresponding choice of m is optimal for the given horizon of uncertainty h.

4.1 Info-Gaps for Imprecise Probabilities

Let $\omega \in \Omega$ be an uncertain parameter of interest—Ω can be an arbitrary set. We must select a decision d from a finite set D. The loss function $L(d, \omega)$ represents the loss (in utiles) if we choose d and ω obtains.

Info-gap theory starts out with a family of nested sets U_h of Ω, where h is a non-negative parameter called the *horizon of uncertainty* and $U_h \subseteq U_{h'}$ whenever $h \leq h'$. In our example, U_h was simply $[0, h]$. Following that example, we saw that a very natural way to model these nested sets U_h in terms of sets of probabilities goes by way of a *vacuous model* $\mathcal{M}_h$, that is, the set of all probability densities that are zero outside U_h.

If we denote the upper expectation induced by $\mathcal{M}_h$ by $\overline{P}_h$, then, formally, we define the info-gap solution $D^*(L_c) \subseteq D$ at satisficing level L_c as:

$$\hat{h}(d, L_c) = \max\left\{h \colon \overline{P}_h(L(d,\cdot)) \leq L_c\right\}$$
$$D^*(L_c) = \arg\max_{d \in D} \hat{h}(d, L_c)$$

Note that $D^*(L_c)$ will usually be a singleton (or, the empty set).

Also note that the first equation may not have a solution: this happens when $\overline{P}_0(L(d,\cdot)) > L_c$, that is, when d is infeasible even if we are as certain as can be ($h = 0$).

Now, from the point of view of imprecise probability, there is no compelling reason to restrict ourselves to vacuous models. In fact, we can allow $\mathcal{M}_h$ to be any set of probability densities on Ω, under one restriction: a close inspection of the theory reveals that a crucial property that the info-gap model relies on is that the worst case cost, $\overline{P}_h(L(d,\cdot))$ is increasing as the horizon of uncertainty h increases. Whence, we logically impose that $\mathcal{M}_h \subseteq \mathcal{M}_{h'}$ whenever $h < h'$.

So, instead of starting out from a family of nested subsets U_h of Ω, we start out from a family of nested sets $\mathcal{M}_h$ of probability densities on Ω. One can of course interpret this again as an info-gap model, where the uncertain parameter is now the probability density over Ω—also see [2, pp. 1062–1063] for an informal discussion of this approach. The imprecise Dirichlet model [14] is an example of such family (with $h = 1/s$). For another example, see [5] for a discussion of nested sets of p-boxes and the resulting info-gap analysis.

4.2 Main Result

The next result links the info-gap solution to the so-called Γ-minimax[2] solution (see [2, p. 1061, Fig. 14] for an informal discussion of a very similar equivalence between info-gap and minimax):

Theorem 1. *The info-gap solution $D^*(L_c)$ coincides with Γ-minimax solution with respect to $\overline{P}_h$, that is,*

$$D^*(L_c) = \arg\min_{d \in D} \overline{P}_h(L(d,\cdot)),$$

whenever the following conditions are satisfied:

(i) for all $d \in D$, $\overline{P}_h(L(d,\cdot))$ is strictly increasing as a function of h, and

(ii) it holds that

$$L_c = \min_{d \in D} \overline{P}_h(L(d,\cdot)). \tag{6}$$

Proof. By definition, $d^* \in D^*(L_c)$ whenever, for all $d \in D$,

$$\hat{h}(d^*, L_c) \geq \hat{h}(d, L_c)$$

By definition of $\hat{h}(d, L_c)$, this is equivalent to saying that

$$\left\{h' \colon \overline{P}_{h'}(L(d^*,\cdot)) \leq L_c\right\} \supseteq \cup_{d \in D} \left\{h' \colon \overline{P}_{h'}(L(d,\cdot)) \leq L_c\right\}$$

Rewriting the above expression, we have, equivalently,

$$\left\{h' \colon \overline{P}_{h'}(L(d^*,\cdot)) \leq L_c\right\} \supseteq \left\{h' \colon \min_{d \in D} \overline{P}_{h'}(L(d,\cdot)) \leq L_c\right\}$$

[2] Γ-minimax minimises the upper expectation of the loss.

But, by Eq. (6), $L_c = \min \overline{P}_h(L(d,\cdot))$, and $\overline{P}_h(L(d,\cdot))$ is strictly increasing for all d as a function of h, whence its minimum over d is strictly increasing as well. Concluding, the set on the right hand side is a fancy way of writing $[0,h]$. Therefore, the above is equivalent to

$$\overline{P}_h(L(d^*,\cdot)) \leq L_c$$

Once more by Eq. (6), this is equivalent to saying that d^* is a Γ-minimax solution with respect to $\overline{P}_h$. □

Interestingly, for given L_c such that

$$\min_{d\in D} \overline{P}_0(L(d,\cdot)) \leq L_c \leq \min_{d\in D} \overline{P}_\infty(L(d,\cdot))$$

it holds that Eq. (6) has a unique solution for $h \geq 0$ whenever all $\overline{P}_h(L(d,\cdot))$ are strictly increasing *and continuous* in h. It is given by:

$$h = \max\left\{h' \colon \min_{d\in D} \overline{P}_h(L(d,\cdot)) \leq L_c\right\} \tag{7}$$

This means that we are effectively free to choose L_c under the additional assumption of continuity. To see why we are not free to choose L_c when continuity is not satisfied, imagine for instance that:

$$\overline{P}_h(L(d_1,\cdot)) = \begin{cases} x & \text{if } h \leq 1 \\ 3+x & \text{if } h > 1 \end{cases}$$

$$\overline{P}_h(L(d_2,\cdot)) = \begin{cases} 1+x & \text{if } h \leq 1 \\ 4+x & \text{if } h > 1 \end{cases}$$

Then, for $L_c = 3$, we have that $D^*(2) = \{d_1, d_2\}$ because $\hat{h}(d,2) = 1$ for both d_1 and d_2, yet obviously d_1 is Γ-minimax (it could even be uniformly dominated by d_2). Effectively, this is simply a technical limitation of the info-gap model, as any reasonable person would probably agree with the Γ-minimax solution.

Now, it is well known that every Γ-minimax solution is also maximal (see for instance [12]), whence, we conclude:

Theorem 2. *Suppose that, for all $d \in D$, $\overline{P}_h(L(d,\cdot))$ is strictly increasing as a function of h. Let*

$$L_c(h) = \min_{d\in D} \overline{P}_h(L(d,\cdot)) \tag{8}$$

Then, for all $h' \leq h$, every info-gap decision $d^* \in D^*(L_c(h'))$ is maximal with respect to $\overline{P}_h$:

$$\bigcup_{0\leq h'\leq h} D^*(L_c(h')) \subseteq \{d \in D \colon (\forall d' \in D)(\overline{P}_h(L(d',\cdot) - L(d,\cdot)) \geq 0)\}$$

Proof. Use the preceding theorem, and note that every Γ-minimax with respect to $\underline{P}_{h'}$ is maximal with respect to $\underline{P}_h$, provided that $h' \leq h$. □

Again, if in addition all $\overline{P}_h(L(d,\cdot))$ are continuous in h, then the range for L_c in the above theorem is simply an interval:

$$\{L_c(h') \colon h' \leq h\} = \left[\min_{d\in D} \overline{P}_0(L(d,\cdot)), \min_{d\in D} \overline{P}_h(L(d,\cdot))\right].$$

Summarising, Theorem 1 provides sufficient conditions[3] for the info-gap solution, for fixed values of L_c and h, to be equivalent to a Γ-minimax solution: proponents of either approach must reconcile.

Theorem 2 shows that a full fledged info-gap analysis, varying the horizon of uncertainty along an interval $[0,h]$, yields an elegant approach to capture maximal solutions. In our example, we actually find *all* maximal options—in general this may not be the case. Still, it shows the that an info-gap analysis can be of value even if maximality is the final goal:

- an info-gap analysis might give a rough idea of the size of the maximal set (in particular, it provides a lower bound for it),
- the analysis can be an appealing way to represent the maximal solution graphically, and
- as robustness curves show the trade-off between uncertainty and cost, they are also obviously useful in the process of elicitation.

5 Conclusion

We constructed a simple model for inspecting animal herds for dangerous exotic infections, building further on the work of Moffitt et al. [10]. We solved the problem using two popular decision methodologies suited for dealing with severe uncertainty: info-gap analysis, and imprecise probability theory (maximality and Γ-minimax). We found that, in this example, the solutions of both models essentially coincide, although the way they arrive at it is very different.

We explored the theoretical link between info-gap theory, Γ-minimax, and maximality. We established that, under rather general conditions, every info-gap solution is maximal. Therefore, the set of maximal options can be inferred at least partly, and sometimes wholly, from an info-gap analysis. Consequently, robustness curves also make sense in an imprecise probability context, for exploring maximal options, and for elicitation, when studying the trade-off between uncertainty and cost that is often of interest to decision makers.

[3] We have not yet investigated in how far they are also necessary.

Acknowledgements

The authors thank Kirsty Hinchliff and Ben Powell, who have been involved with an embryonic draft of this paper. The paper has also benefited greatly from discussions with Yakov Ben-Haim and Frank Coolen, to who we extend our sincerest thanks. Finally, we thank both reviewers for their valuable comments and useful suggestions.

References

[1] Y. Ben-Haim, *Information gap decision theory: Decisions under severe uncertainty*, Academic Press, 2001.

[2] Y. Ben-Haim, C. C. Dacso, J. Carrasco, and N. Rajan, *Heterogeneous uncertainties in cholesterol management*, International Journal of Approximate Reasoning **50** (2009), 1046–1065.

[3] Marquis de Condorcet, *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, L'Imprimerie Royale, Paris, 1785.

[4] Defra, *Impact assessment of an independent body for animal health in England*, 2009, `http://www.defra.gov.uk/corporate/consult/newindependent-body-ah/impact-assessment.pdf`.

[5] S. Ferson and W. T. Tucker, *Probability boxes as info-gap models*, Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS 2008), 2008, Article number 4531314.

[6] E. M. Fèvre, B. M. de C. Bronsvoort, K. A. Hamilton, and S. Cleaveland, *Animal movements and the spread of infectious diseases*, Trends in Microbiology **14** (2006), no. 3, 125–131.

[7] M. G. Garner and M. B. Lack, *Modelling the potential impact of exotic diseases on regional Australia*, Australian Veterinary Journal **72** (1995), no. 3, 81–87.

[8] J. P. Gosling, A. Hart, D. Mouat, M. Sabirovic, and A. Simmons, *Quantifying uncertainty about the cost of exotic diseases using expert elicitation*, Tech. report, The Food and Environment Research Agency, 2011, Submitted to Risk Analysis. Available at `http://www.jpgosling.co.uk/Pub/Disease%20cost.pdf`.

[9] L. J. Moffitt and C. D. Osteen, *Prioritizing invasive species threats under uncertainty*, Agricultural and Resource Economics Review **35** (2006), no. 1, 41–51.

[10] L. J. Moffitt, J. K. Stranlund, and C. D. Osteen, *Robust detection protocols for uncertain introductions of invasive species*, Journal of Environmental Management **89** (2008), 293–299.

[11] A. Sen, *Social choice theory: A re-examination*, Econometrica **45** (1977), no. 1, 53–89.

[12] M. C. M. Troffaes, *Decision making under uncertainty using imprecise probabilities*, International Journal of Approximate Reasoning **45** (2007), no. 1, 17–29.

[13] P. Walley, *Statistical reasoning with imprecise probabilities*, Chapman and Hall, London, 1991.

[14] ———, *Inferences from multinomial data: Learning about a bag of marbles*, Journal of the Royal Statistical Society, Series B **58** (1996), no. 1, 3–34.

[15] D. H. Zeman, *The "best" diagnostic test*, Swine Health and Production **5** (1997), no. 4, 159–160.

Robustness of Natural Extension

Matthias C. M. Troffaes
Durham University
United Kingdom
matthias.troffaes@gmail.com

Robert Hable
Universität Bayreuth
Germany
Robert.Hable@uni-bayreuth.de

Abstract

How sensitive is the natural extension of an upper prevision against small perturbations in the assessments? We revise some basic results from the theory of systems of linear inequalities and equalities, and linear programming, and apply them to the theory of upper previsions. We find that stability is most easily characterized through a regularity condition on the constraints of the primal problem. We then study stability, and the existence of stable representations, in detail. We find necessary and sufficient conditions for the usual representations of natural extension to be stable, and necessary and sufficient conditions for natural extension to have a stable representation at all. We show that, by arbitrary small perturbation, we can force stability of the usual representations.

1 Introduction

Brevity pertains—see [8] for more about upper previsions.

Let Ω be any finite possibility space. A *gamble* is a real-valued function on Ω. The set of all such gambles is denoted by $\mathcal{L}$, so $\mathcal{L} = \mathbb{R}^\Omega$.

We are uncertain about the true value ω in Ω. A popular way of modeling our uncertainty about ω goes by means of an *upper prevision* $\overline{P}$. Specifically, assume that for each gamble g from a finite set $\mathcal{K} \subseteq \mathcal{L}$, we can specify an upper bound $\overline{P}(g)$ on its expectation. We limit ourselves to upper bounds, without loss of generality: a lower bound $\underline{P}(g)$ for g simply translates into an upper bound $\overline{P}(-g) = -\underline{P}(g)$ for $-g$.

A *probability mass function* x on Ω incurs a special kind of upper prevision, namely, one that fixes the expectation exactly, as $x(f) = -x(-f)$:[1]

$$x(f) = \sum_{\omega \in \Omega} x(\omega) f(\omega),$$

noting that, for convenience, we denote the expectation with respect to a probability mass function x also by x. We call x, as a function of gambles, a *linear prevision*. The set of all linear previsions on $\mathcal{L}$ is denoted by C, and it is a subset of the set P of all positive linear functionals (those x for which $x(\omega) \geq 0$ for all ω but not necessarily $x(1) = 1$) on $\mathcal{L}$:

$$C = \{x \in P \colon x(1) = 1\}.$$

For a general upper prevision $\overline{P}$, its *natural extension* $\overline{E}$ is of particular interest [8, §3.4.1]:

$$\overline{E}(f) = \max\{x(f) \colon x \in P,\, x(1) = 1,\, x \leq \overline{P}\} \quad (1)$$

Here, $x \leq \overline{P}$ means that $x(g) \leq \overline{P}(g)$ for all $g \in \mathcal{K}$. Basically, $\overline{E}$ tells us how to accomplish inference from $\overline{P}$: given the bounds specified by $\overline{P}$, it gives us bounds for all other gambles.

The problem of natural extension in Eq. (1) is easily seen to be a linear programming problem. If it has a solution, then $\overline{P}$ is said to *avoid sure loss*. If $\overline{E}$ coincides with $\overline{P}$ on $\mathcal{K}$, then $\overline{P}$ is said to be *coherent*.

Its dual is (abusing notation for brevity) [8, §3.1.3(e)]:

$$\overline{E}(f) = \min \left\{ a + \textstyle\sum_{g \in \mathcal{K}} \lambda_g \overline{P}(g) \colon (a, \lambda_\mathcal{K}) \in Q^*, \; a + \textstyle\sum_{g \in \mathcal{K}} \lambda_g g \geq f \right\} \quad (2)$$

where $Q^* = \{(a, \lambda_\mathcal{K}) \colon a \in \mathbb{R} \text{ and } \lambda_g \in \mathbb{R}^+\}$.[2]

[1] The notation 'x' for a probability mass function follows the usual convention in the linear programming literature, where x usually denotes the variable over which we optimize.

[2] Technically, $\lambda_\mathcal{K} \in (\mathbb{R}^+)^\mathcal{K}$, and we denote $\lambda_\mathcal{K}(g)$ by λ_g.

For the purpose of numerical analysis, but also for elicitation, it is important to know whether the solution is sensitive to perturbations in the assessments embodied by $\overline{P}$. The main purpose of this paper is to characterize those upper previsions that are insensitive to such perturbations. We investigate under what conditions a stable representation exists, and how to find this stable representation.

We extend, and to some extent, also simplify, earlier work by Hable, in particular, [2, pp. 118–125, Sec. 5.2] and [3, Sec. 2]. Doing so, we rely on well-known results about the stability of systems of linear inequalities and equalities.

The paper is structured as follows. Section 2 introduces and demonstrates the problem of instability of natural extension by means of a few simple examples. Section 3 reviews the theory of stability of systems of linear inequalities and equalities. Section 4 applies these results on the theory of lower previsions, and natural extension in particular. Section 5 concludes the paper.

2 Examples

Before we venture into the realm of the theory of systems of linear inequalities and equalities, we present some straightforward, yet insightful, examples. Although these examples present an oversimplified and naive view of the notion of stability of linear programs, they do capture the key aspects of the discussion that will follow.

2.1 Instability of Avoiding Sure Loss

We start with a special case of instability of natural extension, namely, when small perturbations cause the lower prevision to incur sure loss.

Consider $\Omega = \{\omega_1, \omega_2\}$, and the following assessments:[3]

$$\overline{P}(I_{\omega_2}) = 2/3 \qquad \overline{P}(I_{\omega_1}) = 1/3$$

By Eq. (1), it follows that we can calculate the natural extension $\overline{E}$ of for instance $I_{\omega_1} + 2I_{\omega_2}$ by the following linear program:

$$\text{maximize } \begin{bmatrix}1 & 2\end{bmatrix} \begin{bmatrix}x_1 \\ x_2\end{bmatrix}$$

subject to

$$x_1 \geq 0,\ x_2 \geq 0,\ x_1 + x_2 = 1 \qquad \text{(C)}$$

$$\begin{bmatrix}0 & 1 \\ 1 & 0\end{bmatrix} \begin{bmatrix}x_1 \\ x_2\end{bmatrix} \leq \begin{bmatrix}2/3 \\ 1/3\end{bmatrix} \qquad \text{(S)}$$

[3] By I_ω we denote the gamble which is 1 at ω and zero elsewhere.

Clearly, (C) + (S) have a non-empty feasible set: it includes the probability mass function x with $x(\omega_1) = 1/3$ and $x(\omega_2) = 2/3$ (in fact, this is the only element of the feasible set).

However, (C) + (S_ϵ), with

$$\begin{bmatrix}0 & 1 \\ 1 & 0\end{bmatrix} \begin{bmatrix}x_1 \\ x_2\end{bmatrix} \leq \begin{bmatrix}2/3 - \epsilon \\ 1/3\end{bmatrix} \qquad (\text{S}_\epsilon)$$

has an empty feasible set, for any $\epsilon > 0$. If a feasible system of constraints has no solution for some (but not necessarily all) arbitrary small perturbations, then we say that these *constraints are unstable*. Obviously, in such a case, the linear program is deemed unstable as well.

The above example shows that carelessly designed linear programming algorithms may fail to solve even this simple problem due to simple rounding errors.

In practice, implementations of linear programming get around this limitation by transforming to a so-called stable representation. Indeed, by identifying implicit linearities, the program becomes stable, at least in this case. Concretely, the modified system (C) + (S')

$$\begin{bmatrix}0 & 1\end{bmatrix} \begin{bmatrix}x_1 \\ x_2\end{bmatrix} = \begin{bmatrix}2/3\end{bmatrix} \qquad \text{(S')}$$

has the same feasible region as original problem. But, now, unlike the original system, all perturbations to the modified assessments:

$$\begin{bmatrix}0 \pm \epsilon & 1 \pm \delta\end{bmatrix} \begin{bmatrix}x_1 \\ x_2\end{bmatrix} = \begin{bmatrix}2/3 \pm \eta\end{bmatrix} \qquad (\text{S'}_{\epsilon,\delta,\eta})$$

have a solution for every sufficiently small ϵ, δ, η. In other words, the modified constraints are feasible for every sufficiently small perturbation, and so the modified system constraints is stable: we say that the original system has a *stable representation*. Moreover, the solution to the perturbed problem

$$x_2 = \frac{2/3 + \eta - \epsilon}{1 + \delta - \epsilon}$$

remains close to the original solution $x_2 = 2/3$. Whence, the linear program, under the stable representation, is stable too.

2.2 Instability of Natural Extension

The following example is adapted from an example given by Robinson [6, p. 443]. Consider $\Omega = \{a, b, c, d\}$, and the following assessments:

$$\overline{P}(I_a + 2I_b/3 + 2I_d) = 1/2 \quad \overline{P}(I_b + 3I_c) = 3/2$$

By Eq. (1), it follows that we can calculate the natural extension $\overline{E}$ of for instance $2I_b + 2I_c$ by the following linear program:

$$\text{maximize } \begin{bmatrix}0 & 2 & 2 & 0\end{bmatrix} \begin{bmatrix}x_a \\ x_b \\ x_c \\ x_d\end{bmatrix}$$

subject to

$$\begin{gathered} x_a \geq 0,\, x_b \geq 0,\, x_c \geq 0,\, x_d \geq 0 \\ x_a + x_b + x_c + x_d = 1 \end{gathered} \tag{C}$$

$$\begin{bmatrix}1 & 2/3 & 0 & 2 \\ 0 & 1 & 3 & 0\end{bmatrix} \begin{bmatrix}x_a \\ x_b \\ x_c \\ x_d\end{bmatrix} \leq \begin{bmatrix}1/2 \\ 3/2\end{bmatrix} \tag{S}$$

Clearly, (C) + (S) have a non-empty feasible set: it consists of all the probability mass functions of the form (with $\alpha \in [0,1]$)

$$\alpha \begin{bmatrix}1/2 \\ 0 \\ 1/2 \\ 0\end{bmatrix} + (1-\alpha) \begin{bmatrix}0 \\ 3/4 \\ 1/4 \\ 0\end{bmatrix}$$

so $\overline{E}(2I_b + 2I_c) = 2$.

However, (C) + (S_ϵ), with

$$\begin{bmatrix}1 & 2/3 - \epsilon & 0 & 2 \\ 0 & 1 & 3 & 0\end{bmatrix} \begin{bmatrix}x_a \\ x_b \\ x_c \\ x_d\end{bmatrix} \leq \begin{bmatrix}1/2 \\ 3/2\end{bmatrix} \tag{S_ϵ}$$

has a very different feasible set, for any $\epsilon > 0$. Indeed, regardless of how small ϵ is chosen, the feasible set of the perturbed system contains only one probability mass function:

$$\begin{bmatrix}1/2 \\ 0 \\ 1/2 \\ 0\end{bmatrix}$$

so, now, $\overline{E}(2I_b + 2I_c) = 1$. An arbitrary small perturbation can lead to an unproportionally large variation in the solution of the natural extension.

One can easily check that the system has perturbations that incur sure loss, for instance, by reducing the upper prevision of the first gamble to $1/2 - \epsilon$. We will prove that the natural extension is unstable if and only if there are perturbations which push the system into incurring sure loss (or equivalently, that the natural extension is stable if and only if all sufficiently small perturbations avoid sure loss).

Observe that the dual problem has an unbounded optimal solution:

$$\begin{aligned}\begin{bmatrix}0 & 2 & 2 & 0\end{bmatrix} \geq 2 &+ \lambda_1 \left(1/2 - \begin{bmatrix}1 & 2/3 & 0 & 2\end{bmatrix}\right) \\ &+ \lambda_2 \left(3/2 - \begin{bmatrix}0 & 1 & 3 & 0\end{bmatrix}\right)\end{aligned}$$

for all non-negative λ_1 and λ_2 such that $\lambda_1 = 3\lambda_2$. We will see that this is also tightly related to the instability of the primal problem.

Finally, it is unclear whether the system has a stable representation or not. Intuitively, it seems not; we will prove this later. For now, we present next a much simpler example which has clearly no stable representation.

2.3 Unrepairable Instability

As suggested already, not every upper prevision has a stable representation. Consider for instance the upper prevision defined on I_{ω_2} by

$$\overline{P}(I_{\omega_2}) = 0$$

To calculate its natural extension, we must consider the constraints (C) + (S2), with

$$\begin{bmatrix}0 & 1\end{bmatrix} \begin{bmatrix}x_1 \\ x_2\end{bmatrix} \leq \begin{bmatrix}0\end{bmatrix} \tag{S2}$$

The feasible region is non-empty: it contains the probability mass function x with $x(\omega_1) = 1$ and $x(\omega_2) = 0$ (in fact, here again, this is the only element of the feasible set). However, the perturbation

$$\begin{bmatrix}0 & 1\end{bmatrix} \begin{bmatrix}x_1 \\ x_2\end{bmatrix} \leq \begin{bmatrix}-\epsilon\end{bmatrix} \tag{$S2_\epsilon$}$$

has an empty feasible region, no matter how small $\epsilon > 0$. In fact, even after recognizing the implicit linearities, the system remains unstable under perturbations. In conclusion, it seems that there is no stable representation.

2.4 Main Issues

Assuming that we can generalize the above observations to arbitrary problems of natural extension, we are left with the following important questions:

1. For the stability of natural extension, does it matter whether we consider the primal or the dual representation?

2. In order to establish the stability of natural extension, is it sufficient to establish stability of the constraints of the primal linear program?

3. Under what conditions are the constraints of the usual representation of the primal linear program stable?

4. If this usual representation is not stable, under what conditions can it be transformed to a stable representation? In other words, when does a stable representation exist?

5. If a stable representation exists, how to find it?

3 Stability of Linear Programming

Robinson [5] characterizes stability of systems of linear inequalities and equalities, and [6] relates this characterization to the stability of natural extension. Here we quickly summarize his results. Also see [10] and [4].

3.1 Stability of Linear Systems of Inequalities and Equalities

Let X and Y be real Banach spaces, let Q be a non-empty convex cone in Y, let P be a non-empty convex set (usually, but not always, assumed to be a convex cone) in X, let b be a point in Y, and let A be a continuous linear operator from X into Y. For two points y_1 and y_2 in Y, we write $y_1 \leq_Q y_2$ if $y_2 - y_1 \in Q$. The cone is used to treat equalities and inequalities homogeneously. Distinguishing between them is crucial when studying stability.[4]

The solution set to

$$Ax \leq_Q b, \quad x \in P, \tag{*}$$

is denoted by F, and for the time being, we are interested in the stability of F with regard to perturbations in A and b.

3.1.1 Definition of Stability

Note that $x \in P$ is a solution of the above system of inequalities if and only if $b - Ax$ is in Q (this is immediate by the definition of $\leq_Q$). Hence, for any arbitrary $x \in P$, we can take the distance between $b - Ax$ and Q as a measure of how much x deviates from a solution of the system, or, if you like, as a measure of infeasibility with respect to the system.

$$\rho(x) = d(b - Ax, Q) = \inf_{q \in Q} \|b - Ax - q\|$$

The distance will be zero exactly when x satisfies the system.

[4] For example, $x = 0$ is obviously stable, but $\{x \geq 0, x \leq 0\}$ is obviously not (for instance, perturb the first inequality to $x \geq \epsilon$ for some $\epsilon > 0$).

Definition 1 (Robinson [5, p. 755])**.** *The system* (*) *is said to be* stable *if there is a positive number β, such that for each $x_0 \in F$ and for any continuous linear operator $A' \colon X \to Y$ and any $b' \in Y$, sufficiently close to A and b respectively, the distance from x_0 to the solution set of the perturbed system*

$$A'x \leq_Q b', \quad x \in P,$$

is not greater than $\beta\rho'(x_0)$, where

$$\rho'(x) = d(b' - A'x, Q) = \inf_{q \in Q} \|b' - A'x - q\|$$

is the distance between $b' - A'x$ and Q.

Note that stability implicitly demands that the original system is feasible, and that all (sufficiently small) perturbations of the original system are feasible.

In order to understand the reasoning behind Robinson's stability condition, let us rewrite the distance condition into something we can easily interpret:

$$\begin{aligned}
d(x_0, F') &\leq \beta\rho'(x_0) \\
&= \beta \inf_{q \in Q} \|b' - A'x - q\| \\
&\leq \beta \inf_{q \in Q} \big(\|b' - b - (A'x_0 - Ax_0)\| \\
&\qquad + \|b - Ax_0 - q\|\big) \\
&= \beta\big(\|b' - b - (A'x_0 - Ax_0)\| \\
&\qquad + \inf_{q \in Q} \|b - Ax_0 - q\|\big) \\
&= \beta\|(b' - A'x_0) - (b - Ax_0)\|
\end{aligned}$$

which we can further bound by

$$\begin{aligned}
&= \beta\|b' - b - (A'x_0 - Ax_0)\| \\
&\leq \beta(\|b' - b\| + \|A'x_0 - Ax_0\|) \\
&\leq \beta(\|b' - b\| + \|A' - A\|\|x_0\|)
\end{aligned}$$

Roughly speaking, the condition implies that any solution x_0 of the original system, is also a solution of the perturbed system up to an error that is proportional to the size of the perturbation and $\|x_0\|$.

3.1.2 Stability Criterion

Next, Robinson identifies a simple necessary and sufficient criterion for stability.

Definition 2 (Robinson [5, Def. 1])**.** *The system* (*) *is called* regular *if $b \in \operatorname{int}(AP + Q)$.*

Theorem 3 (Robinson [5])**.** *The system* (*) *is stable if and only if it is regular.*

Proof. As discussed in [5, p. 755, last paragraph], this follows immediately from [5, Thm. 1]. □

The following interesting result is an immediate consequence of [5, Thm. 1] (also see [6, Lem. 3]):

Theorem 4. *The system* (*) *is stable if and only if there is an* $\epsilon > 0$ *such that, for all* A' *and* b' *satisfying* $\max\{\|A - A'\|, \|b - b'\|\} < \epsilon$*, the system*

$$A'x \leq_Q b', \quad x \in P,$$

is feasible.

3.1.3 Stable Representation Criterion

In finite dimensions, we have the following result as well, where $\operatorname{ri} P$ denotes the topological interior of P relative to its affine span.

Theorem 5 (Robinson [5, Thm. 3])**.** *The system*

$$Gx \leq g,\ Hx = h, \quad x \in P \tag{3}$$

is representable as a regular system of inequalities and equalities over P *with the same solution set* F *if and only if* $F \cap \operatorname{ri} P \neq \emptyset$*. If the condition is satisfied, then the system can be made regular by changing certain inequalities to equalities and deleting certain redundant equalities.*

3.2 Stability of Linear Programming

Robinson's stability criterion for systems of linear inequalities and equalities does *not* say that the *Hausdorff distance* (see [7, Sec. 3] for a study of this metric in the context of credal sets) between the solution sets is small: it only says that the solution set of the perturbed system is contained, up to a small error, in the solution set of the original system. In fact, the solution set of the original system could be much larger (we hinted already at an example of this earlier, once realized that the dual constraints for natural extension are always stable).

Confusingly, when considering the primal constraints for natural extension, it turns out that stability of these constraints *do* imply that the Hausdorff distance between the credal sets of the original and perturbed systems is small. One of the underlying reasons for this is that the set C of probability mass functions is bounded.

The following result summarizes the relationship between stability of systems of linear inequalities and equalities and the stability of linear programs.

Note that we say that a linear program is *solvable* whenever it has an optimal solution, and that the dual $Q^* \subseteq \mathbb{R}^n$ of a cone $Q \subseteq \mathbb{R}^n$ is defined as

$$Q^* = \{z \in \mathbb{R}^n \colon (\forall x \in Q)(zx \geq 0)\}$$

where zx denotes the dot product of z and x.

Definition 6. *Consider a finite dimensional linear program (P) and its dual (D):*

maximize cx	*subject to*	$Ax \leq_Q b$	$x \in P$
minimize ub	*subject to*	$uA \geq_{P^*} c$	$u \in Q^*$

where P *and* Q *are convex cones. The following conditions are equivalent. If any (and whence, all) of them are satisfied, then we say that the linear program (P) is* stable*.*

(A) The constraints of (P) and (D) are regular.

(B) The sets of optimal solutions of (P) and (D) are non-empty and bounded.

(C) For all sufficiently small perturbations (P')—with corresponding dual (D')—of the linear program (P), both (P') and (D') are solvable.

Proof of equivalence. See Robinson [6, Theorem 1]. □

Robinson [6, Theorem 1] also shows that, whenever a linear program is stable in the above sense, every optimal solution of (P') and (D') remains close to the the optimal solution set of (P) and (D). This obviously implies that the optimal value will not deviate much, which is exactly what we are after for the stability of natural extension. We refer to [6, Theorem 1] for a rigorous statement of what is meant by "sufficiently small" and "remains close" (we have omitted it here to keep the exposition as non-technical as possible).

3.3 Examples Revisited

Before we apply the above results to the specific problem of natural extension, we check stability and stable representability on the earlier examples.

For the first example, again look at Eq. (S), which we demonstrated to be unstable. The cone Q, in this case, is simply the set of non-negative gambles. Let us check that $b \notin \operatorname{int}(AC + Q)$, where $Ax \leq_Q b$ embodies the constraints $x \leq \overline{P}$ of Eq. (1), for $x \in C$.

Note that this turns out to be equivalent to checking that $b \notin \operatorname{int}(AP + Q)$, where $Ax \leq_Q b$ corresponds to the system *including* the constraint $x(1) = 1$, but $x \in P$ (see Theorem 7 further on).

A parametric representation of the set $AC + Q$ follows readily:

$$AC + Q = \left\{ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \right\}$$

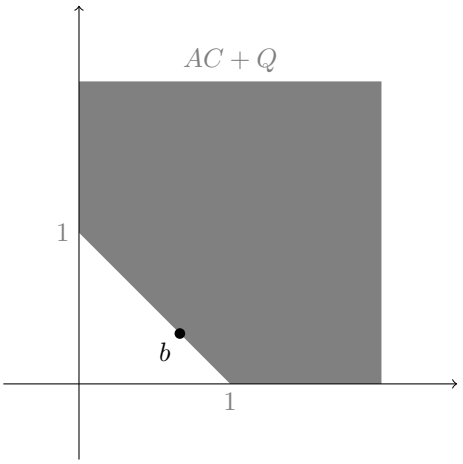


Figure 1: The region $AC+Q$ for (C) + (S). The vector b lies on the border, so the system is not stable.

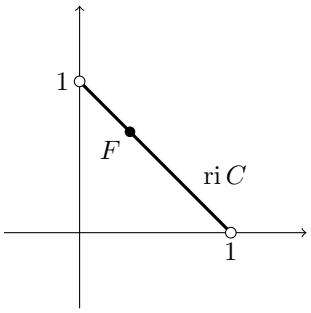


Figure 2: The relative interior of C, and solution set F, for (C) + (S). The solution set F has non-empty intersection with the relative interior of C, so the system has a stable representation.

over all $x_1, x_2, y_1, y_2 \geq 0$ such that $x_1 + x_2 = 1$. The vector

$$b = \begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix}$$

lies on the border of this set, but not in its interior (see Fig. 1). Whence, the system is not stable.

However, it has a stable representation: the solution set

$$F = \left\{ \begin{bmatrix} 1/3 \\ 2/3 \end{bmatrix} \right\}$$

intersects with the relative interior of the set C of all probability mass functions (see Fig. 2).

For the second example, one can similarly show that it does not satisfy the stability criterion. It is easy to show that it does not have a stable representation. Indeed, the feasible set lies on the edge of the set C of all probability mass functions, because $x_d = 0$ everywhere in the feasible region. So F does not intersect with the relative interior of C, and therefore there is no stable representation.

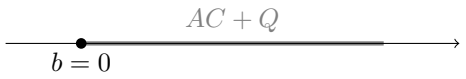


Figure 3: The region $AC + Q$ for (C) + (S2). The vector b lies on the border, so the system is not stable.

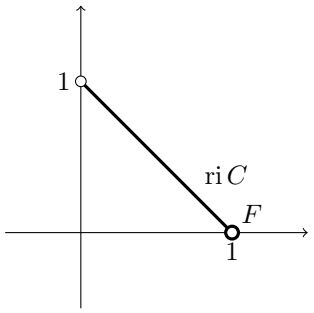


Figure 4: The relative interior of C, and solution set F, for (C) + (S2). The solution set F has empty intersection with the relative interior of C, so the system has no stable representation.

Let us now revisit the third example. Inspect Eq. (S2). A parametric representation of the region $AC + Q$ is

$$AC + Q = \left\{ \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + y_1 \right\}$$

over all $x_1, x_2, y_1 \geq 0$ such that $x_1 + x_2 = 1$, which reduces to

$$= \{y_1 : y_1 \geq 0\}$$

that is, the set of non-negative real numbers. The vector

$$b = \begin{bmatrix} 0 \end{bmatrix}$$

lies on the border of this set, but not in its interior (see Fig. 3). Whence, the system is not stable.

Moreover, we can now prove our earlier intuition that it has no stable representation: the solution set

$$F = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$$

does not intersect with the relative interior of the set C of all probability mass functions (see Fig. 4).

4 Stability of Natural Extension

4.1 Canonical Representations

We now rewrite the primal and dual forms of natural extension using the notation of the previous section

on linear programming. The primal linear program, Eq. (1), is:

$$\text{maximize } c_f x \text{ subject to } A_{\overline{P}} x \leq_Q b_{\overline{P}},\ x \in P \quad (\mathbf{P})$$

with

$$c_f = \begin{bmatrix} f(\omega_1) & \dots & f(\omega_n) \end{bmatrix}$$

$$A_{\overline{P}} = \begin{bmatrix} 1 & \dots & 1 \\ g_1(\omega_1) & \dots & g_1(\omega_n) \\ \vdots & \ddots & \vdots \\ g_k(\omega_1) & \dots & g_k(\omega_n) \end{bmatrix} \quad b_{\overline{P}} = \begin{bmatrix} 1 \\ \overline{P}(g_1) \\ \vdots \\ \overline{P}(g_k) \end{bmatrix}$$

$$Q = \left\{ \begin{bmatrix} 0 \\ y_1 \\ \vdots \\ y_k \end{bmatrix} : y_1, \dots, y_k \in \mathbb{R}^+ \right\}$$

$$P = \left\{ \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} : x_1, \dots, x_n \in \mathbb{R}^+ \right\}$$

Note that the feasible region F is exactly the credal set of $\overline{P}$.

We call the linear program (**P**) the *canonical representation* of the natural extension of $\overline{P}$.

If we omit the constraint $x(1) = 1$ from the system of inequalities, and consider the reduced optimization problem over $x \in C$ (as we did before in the examples), then we arrive at the *reduced canonical representation* of the natural extension of $\overline{P}$:

$$\text{maximize } c_f x \text{ subject to } A^-_{\overline{P}} x \leq b^-_{\overline{P}},\ x \in C \quad (\mathbf{P}^-)$$

where $A^-_{\overline{P}}$ is $A_{\overline{P}}$ without the first row, and $b^-_{\overline{P}}$ is $b_{\overline{P}}$ without the first element.

Studying the stability of this reduced system simply means that we do not consider perturbations in the normalization constraint $x(1) = 1$, which in fact seems a natural thing to do. However, the theory of stability of linear programs demands that the linear program has a dual, and (**P**$^-$) does not have a dual, because the set C is not a cone. Fortunately, as we shall prove, stability properties are independent of whether we allow perturbations in $x(1) = 1$ or not.

Of course, (**P**) does have a dual program, given earlier by Eq. (2):

$$\text{minimize } u b_{\overline{P}} \text{ subject to } u A_{\overline{P}} \geq_{P^*} c_f,\ u \in Q^* \quad (\mathbf{D})$$

with $P^* = \{z^T : z \in P\}$ and

$$Q^* = \left\{ \begin{bmatrix} a & \lambda_1 & \dots & \lambda_k \end{bmatrix} : a \in \mathbb{R},\ \lambda_1, \dots, \lambda_k \in \mathbb{R}^+ \right\}$$

The linear program (**D**) is the *canonical dual representation* of the natural extension of $\overline{P}$.

4.2 Stability of the Canonical Representation of Natural Extension

It will follow from our discussion in Section 3 that, to determine stability of natural extension in its canonical representation, it suffices to determine the regularity (or, stability) of the system of linear inequalities and equalities (**P**) or equivalently, of (**P**$^-$).

First, we need one more definition: a *linear-vacuous* mixture is any coherent upper prevision of the form

$$(1 - \alpha)x + \alpha \sup_{\omega \in \Omega}$$

for some $\alpha \in [0, 1]$ and $x \in C$. We say that this linear-vacuous mixture is non-linear whenever $\alpha > 0$.

Theorem 7. *Let $\overline{P}$ be any upper prevision. The following conditions are equivalent.*

*(A) The linear program (**P**) is stable.*

*(B) The linear program (**D**) is stable.*

*(C) The system of linear inequalities and equalities of (**P**) is regular.*

*(D) The system of linear inequalities and equalities of (**P**$^-$) is regular.*

(E) All sufficiently small perturbations of $\overline{P}$ avoid sure loss, that is, there is an $\epsilon > 0$ such that all $\overline{P}'$ on $\mathcal{K}$ satisfying $\overline{P}(g) - \epsilon \leq \overline{P}'(g) \leq \overline{P}(g)$ avoid sure loss.

(F) There is a linear prevision x such that $\overline{P}(g) > x(g)$ for all g in $\mathcal{K}$.

(G) $\overline{P}$ dominates a non-linear linear-vacuous mixture.

(H) $\overline{P}$ avoids sure loss and $\underline{E}(g) < \overline{E}(g)$ for all g in $\mathcal{K}$.

Proof. (A) and (B) are equivalent by Definition 6(A).

(A) and (C) are equivalent, again by Definition 6(A), once established that the system of linear inequalities and equalities of (**D**) is *always* regular. Indeed, it suffices to show that

$$c_f \in \operatorname{int}(Q^* A - P^*)$$

This holds trivially because

$$Q^* A - P^* = \left\{ a + \textstyle\sum_{g \in \mathcal{K}} \lambda_g g - p^* : \dots \right\} = \mathbb{R}^n$$

as we vary over all $a \in \mathbb{R}$ and all $p^* \in P^*$.

(C) implies (D), by Theorem 4. [One can also quickly see that (F) implies (D) by [5, Theorem 2]—also see the discussion at [6, p. 444].]

Equivalence between (D) and (E) follows from Theorem 4, once noted that we only need to consider perturbations in $\overline{P}$ because probabilities sum one—whence every small perturbation in $A^-_{\overline{P}}$ and $b^-_{\overline{P}}$ can be bounded by a proportionally small perturbation in $b^-_{\overline{P}}$ only—and the usual properties of avoiding sure loss with respect to dominating upper previsions.

Equivalence between (E), (F), (G), and (H) follows trivially from the usual properties of lower previsions.

Finally, we establish equivalence between (C) and (F).

We rely on Robinson's regularity condition, $b_{\overline{P}} \in \operatorname{int}(A_{\overline{P}}P + Q)$. It is satisfied if and only if there is an $\epsilon > 0$ such that

$$b_{\overline{P}} + \epsilon B \subseteq A_{\overline{P}}P + Q$$

where B is the closed unit ball in $Y = \mathbb{R}^{\mathcal{K}}$, that is, the set $\{b \in Y : \sup|b| \leq 1\}$. Equivalently, now in matrix notation, we need that

$$\begin{bmatrix} 1 \\ \overline{P}(g_1) \\ \vdots \\ \overline{P}(g_k) \end{bmatrix} + \epsilon B \subseteq \left\{ \begin{bmatrix} x(1) \\ x(g_1) + y_1 \\ \vdots \\ x(g_k) + y_k \end{bmatrix} : x \in P,\, y \in Q \right\}.$$

Equivalently, there must be some $\epsilon > 0$ such that, for every $b \in B$ (that is, $b_i \in [-1, 1]$), there is an $x \in P$ and a $y \in Q$ such that

$$\begin{aligned} 1 + b_0\epsilon &= x(1) \\ \overline{P}(g_i) + b_i\epsilon &= x(g_i) + y_i \text{ for all } i \in \{1, \ldots, k\}. \end{aligned}$$

If the above is satisfied, take $b_0 = 0$ and $b_1 = \cdots = b_n = 1$ to find that $\overline{P}(g_i) > x(g_i)$ for all i, and note that $x \in C$ because $b_0 = 0$.

Conversely, if there is some x' such that $\overline{P}(g_i) > x'(g_i)$ for all i, then the above is satisfied for sufficiently small ϵ. Indeed, fix any $0 < \epsilon < 1$, and let $x = (1 + b_0\epsilon)x'$—obviously $x \in P$, and the first equality is satisfied. The second equality can be satisfied as well, because

$$\overline{P}(g_i) - \epsilon \geq \max_{b'_0 \in \{-1,1\}} (1 + \epsilon b'_0)x'(g_i)$$

can always be achieved for small enough ϵ, because $\overline{P}(g_i) > x'(g_i)$, whence, for such ϵ,

$$\begin{aligned} \overline{P}(g_i) + b_i\epsilon &\geq \overline{P}(g_i) - \epsilon \\ &\geq \max_{b'_0 \in \{-1,1\}} (1 + \epsilon b'_0)x'(g_i) \\ &\geq (1 + \epsilon b_0)x'(g_i) = x(g_i) \end{aligned}$$

which concludes the proof. □

Informally, the canonical representation is stable if and only if $\overline{P}$ is inherently imprecise. This also means that we can always enforce stability by perturbation, for any upper prevision that avoids sure loss: simply mix $\overline{P}$ with a stable one, such as the vacuous upper prevision:

$$(1 - \alpha)\overline{P} + \alpha \sup_{\omega \in \Omega}$$

is *always* stable, for any $\alpha \in (0, 1]$. So, every upper prevision that avoids sure loss has arbitrarily close stable approximations.

Note that the natural extension of the above perturbation will not necessarily behave nicely as a function of α, particularly when $\overline{P}$ is unstable. For instance, in the perturbed example of Section 2.2, $\overline{E}(2I_b + 2I_c) = 1$ if $\alpha \ll \epsilon$ and $\overline{E}(2I_b + 2I_c) = 2$ if $\alpha \gg \epsilon$. In essence, one should pick α large enough to counter any (presumably unintended) implicit linearities, or near linearities.

If, for some reason, approximation is not an option, we have to find a stable representation. The conditions under which this is possible are uncovered in the next section.

4.3 Necessary and Sufficient Conditions for Stable Representations of Natural Extension

Definition 8. *A system of linear inequalities and equalities is said to be a* representation *of another system if it has the same feasible region F as that system.*

Definition 9. *A linear program is said to be a* representation *of another linear program if it has the same feasible region F and objective function as that linear program.*

Theorem 10. *Let $\overline{P}$ be any upper prevision. The following conditions are equivalent.*

(A) The linear program (**P**) *has a stable representation.*

(B) The linear program (**D**) *has a stable representation.*

(C) The system of linear inequalities and equalities of (**P**) *has a regular representation.*

(D) The system of linear inequalities and equalities of ($\mathbf{P}^-$) *has a regular representation.*

(E) There is a linear prevision x in the credal set F of $\overline{P}$ such that $x(\omega) > 0$ for all $\omega \in \Omega$.

(F) $\overline{P}$ avoids sure loss and $\overline{E}(I_\omega) > 0$ for all $\omega \in \Omega$.

Proof. The first part of the proof is similar to the proof of Theorem 7(A)&(B)&(C): again, the key observation is that the system of the dual is always regular. We also rely on the fact that the dual of a representation is a representation of the dual.

(C) $\iff$ (E). Such x belongs precisely to $F \cap \operatorname{ri} P$. Apply Theorem 5.

(D) $\iff$ (E). Such x belongs precisely to $F \cap \operatorname{ri} C$. Apply Theorem 5.

(E) $\implies$ (F). Immediate, because

$$\overline{E}(I_\omega) = \sup_{x' \in F} x'(\omega) \geq x(\omega) > 0.$$

(F) $\implies$ (E). Condition (F) implies that, for every ω, there is an x_ω in F such that $x_\omega(\omega) > 0$. Take any convex mixture x of x_ω with non-zero coefficients. Because F is convex, x belongs to F. Clearly, $x(\omega) > 0$ for all ω in F. □

The condition for having a stable representation is clearly much weaker than the one for stability: in essence, we only need to ensure that no singleton has zero upper probability. Again, it is obvious that this can be achieved by an arbitrary small perturbation, for any upper prevision that avoids sure loss: simply mix $\overline{P}$ with a linear prevision x that satisfies $x(\omega) > 0$ for all ω, such as the uniform one:

$$(1-\alpha)\overline{P} + \alpha \frac{1}{n} \sum_{\omega \in \Omega}$$

where n is the cardinality of Ω, *always* has a stable representation, for any $\alpha \in (0, 1]$. So, every upper prevision that avoids sure loss has arbitrarily close approximations that admit stable representations, and whose canonical representation is stable if and only if the canonical representation of $\overline{P}$ is stable (indeed, by Theorem 7!).

4.4 Finding the Stable Representation

Every reasonably advanced application for working with systems of linear inequalities and equalities has routines for finding all redundant constraints and all implicit linearities (see for instance [1]), effectively recovering the stable representation, when it exists.

5 Discussion and Conclusion

We have linked Robinson's stability criterion for systems of linear inequalities and equalities, and for linear programming, to the theory of upper previsions. We found a range of interesting necessary and sufficient conditions for the usual canonical representations of natural extension to be robust against perturbations, that is, to be stable. Thereby, we provided theoretical guarantees for small changes in the assessments not to have a large impact on any inferences made.

This is obviously rather useful in elicitation: if a subject makes assessments which violate stability, then the subject should at least be made aware of this. We provided a simple tool to fix unstable assessments, through perturbation with a vacuous model.

In case of instability of the canonical constraints, a subject could be unhappy to perturb with a vacuous model, for instance because she insists on certain assessments to be precise. We found that a stable representation may still exist after removal of redundant constraints and recognition of implicit linearities. Tools for doing so are readily available in the literature. Of course, it is *mandatory* to check that the subject actually agrees with the reduced system, and particularly that any linearities, or near linearities, are in agreement with her beliefs. When in doubt, we recommend the vacuous mixture.

In case the reduced system is still unstable, we found that it can be made stable via perturbation with for instance a uniform probability mass function—this may be preferred over vacuous perturbation in case the subject insists on particular assessments to remain precise.

In conclusion, we characterized the robustness of natural extension in a variety of ways, and we provided straightforward ways to work around instabilities by means of perturbation.

Many open problems remain, including the extension to non-finite spaces, and conditional lower previsions, which are typically solved by sequences of linear programs [9], and thus for which stability may be much harder to characterize.

Acknowledgments

The authors thank Thomas Augustin for hosting us in Munich during the summer of 2009, enabling the authors to collaborate on this topic. We also thank Nathan Huntley, who has been involved with a very early draft of the paper. Finally, we thank both reviewers for their valuable suggestions.

References

[1] K. Fukuda and A. Prodon. Double description method revisited. *Combinatorics and Computer*

Science, 1120:91–111, 1996. `http://www.ifor.math.ethz.ch/~fukuda/cdd_home/`.

[2] Robert Hable. *Data-Based Decisions under Complex Uncertainty*. PhD thesis, Ludwig-Maximilians-Universität München, 2008.

[3] Robert Hable. Finite approximations of data-based decision problems under imprecise probabilities. *International Journal of Approximate Reasoning*, 50(7):1115–1128, 2009.

[4] B. Jansen, J. J. de Jong, C. Roos, and T. Terlaky. Sensitivity analysis in linear programming: just be careful! *European Journal of Operational Research*, 101:15–28, 1997.

[5] Stephen M. Robinson. Stability theory for systems of inequalities. Part I: Linear systems. *SIAM Journal on Numerical Analysis*, 12(5):754–769, October 1975.

[6] Stephen M. Robinson. A characterization of stability in linear programming. *Operations Research*, 25(3):435–447, 1977.

[7] Damjan Škulj and Robert Hable. Coefficients of ergodicity for imprecise Markov chains. In Thomas Augustin, Frank P. A. Coolen, Serafín Moral, and Matthias C. M. Troffaes, editors, *ISIPTA'09: Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications*, pages 377–386, Durham, UK, July 2009. SIPTA.

[8] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[9] Peter Walley, Renato Pelessoni, and Paolo Vicig. Direct algorithms for checking consistency and making inferences from conditional probability assessments. *Journal of Statistical Planning and Inference*, 126:119–151, 2004.

[10] S. Zlobec. Stability in linear programming models: An index set approach. *Annals of Operations Research*, 101:363–382, 2001.

Interval-valued regression and classification models in the framework of machine learning

Lev V. Utkin
Department of Computer Science
St.Petersburg State Forest Technical Academy
St.Petersburg, Russia
lev.utkin@mail.ru

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

Abstract

We present a new approach for constructing regression and classification models for interval-valued data. The risk functional is considered under a set of probability distributions, resulting from the application of a chosen inferential method to the data, such that the bounding distributions of the set depend on the regression and classification parameter. Two extreme ('pessimistic' and 'optimistic') strategies of decision making are presented. The method is applicable with many inferential methods and risk functionals. The general theory is presented together with the specific optimisation problems for several scenarios, including the extension of the support vector machine method for interval-valued data.

Keywords. belief functions, classification, interval-valued observations, machine learning, p-box, regression, risk functional, support vector machines.

1 Introduction

A main goal of statistical machine learning is prediction of an unobserved output value y based on an observed input vector $\mathbf{x}$, which requires estimation of a predictor function f from training data consisting of pairs $(\mathbf{x}, y)$. Two major topics in statistics which fit into the statistical machine learning framework are regression analysis and classification. In regression analysis, one typically aims at estimation of a real-valued function based on a finite set of observations with random noise. In classification, the output variable is in one of a finite number of classes[1] and the main task is to classify the output y corresponding to each input $\mathbf{x}$ into one of the classes by means of a discriminant function. Many methods have been proposed for solving machine learning problems, but these are mostly based on rather restrictive assumptions, for example assuming the availability of a large amount of training data, known probability distribution for the random noise, or that all observations are point-valued ('precise'). Such assumptions are typically not fully satisfied in applications. For example, data often include interval-valued ('imprecise') observations, which may result from imperfection of measurement tools or imprecision of expert information if used as data. There may also be (partially) missing data, for example in classification problems the input vector ('pattern') $\mathbf{x}$ is often not fully observed. Many methods for dealing with such features use additional assumptions. In this paper, a general framework is presented that allows such important aspects to be incorporated in machine learning problems without additional assumptions, instead it uses the framework of imprecise probability [34] and it can be used for a wide variety of inferences, models and real-world situations.

Many methods have been presented for regression and classification with interval-valued data [11, 16, 23]. In some methods for machine learning, interval-valued observations are replaced by precise values based on some (often ad-hoc) additional assumptions, for example by taking middle points of the intervals [14]. Also, they may not be suitable if an observation is not restricted to an interval of finite length. This is an important restriction, as frequently it may only be known that an observation is larger (or smaller) than a specific value while the support of the corresponding random quantity is not finite. The method presented in this paper can deal with such information without additional assumptions and allows infinite support[2], including the use of $(-\infty, \infty)$ for missing elements of the input vector $\mathbf{x}$. Machine learning methods have been presented which use standard interval analysis and provide predictor functions with interval-valued parameter [2, 9, 26], and construction of second-order machine learning models for interval-valued patterns

[1] Often two classes, to which attention is restricted in this paper; generalization is possible but not addressed here.

[2] It should be noted that the support of elements of vector $\mathbf{x}$ can be arbitrary. Without loss of generality, we assume it to be $(-\infty, \infty)$.

was proposed in [4]. Although many methods have been presented for dealing with interval-valued data [23], these are mostly based on interval extension of the empirical risk functional [33] without benefiting from, or even considering, an imprecise probabilistic framework in direct relation to imprecise statistical data.

Pelckmans et al. [17] presented a detailed analysis of different methods and models for dealing with missing data in classification. Many methods do so by imputation of (partially) missing patterns, where missing (precise) values are replaced by some preferable values. Imputation using intervals, including the full support in case of missing elements of $\mathbf{x}$ has also been presented. De Cooman and Zaffalon [5] studied the classification problem with missing data in the framework of imprecise probability theory. An interesting approach for regression analysis with interval-valued and fuzzy data using belief functions and evidence theory has been proposed by Petit-Renaud and Denoeux [18]. One of the possible approaches to regression analysis is to consider a set of probability distributions for the random noise instead of a single distribution. This approach can be realized in the framework of imprecise probability theory [34] and has been developed by Walter et al. [36].

The novel approach for constructing a class of machine learning models and methods proposed in this paper uses risk functionals as in [18] and sets of probability distributions as in [36]. The starting point is a set of probability distributions related to the training data, which can just be a small amount of data or imprecise data, and this set can be generated by a variety of inferential methods and is assumed to be bounded by some lower and upper CDFs. Such sets of probability distributions are also called p-boxes [7]. In the regression and classification applications considered in this paper, these bounds for the set of probability distributions depend on the unknown parameter of the regression or discriminant function, because the sets of probability distributions considered are for the random residuals and as such they depend on the model parameter. It should be noted that the considered set of distributions is not the set of parametric distributions having the same parametric form as the bounding distributions, but it is the set of all possible distributions restricted by the lower and upper bounds. This is an important feature of the proposed approach in this paper.

Traditionally, machine learning methods have used a variety of simplifying assumptions in order to maintain acceptable computational effort required for implementation. The fact that the bounds for the set of probability distributions considered in the regression and classification problems depend on the model parameter makes it clear that any optimisation of risk functionals over the whole set of probability distributions is likely to require an enormous computational effort. It will be illustrated that, for a wide range of popular risk functions, computational is feasible due to new results for the optimisation. In addition to introduction of the general theory, the approach will be illustrated by presenting the resulting optimisation problem formulations for several combinations of loss functions and sets of probability distributions.

Generally, the parameter of a regression model is computed by minimising a risk functional defined by the combination of a certain loss function and a probability distribution for the random noise [10, 33]. When using a set of probability distributions instead of a precise distribution, we can choose a single distribution from this set which minimises or maximises the risk functional; the probability distribution maximising (minimising) the risk functional corresponds to the minimax (minimin) strategy. These cases can be called the 'pessimistic' and 'optimistic' decisions, respectively. The main problem in finding these two ('extreme' or 'optimal') precise distributions is that, like the bounds of the corresponding set of distributions, they depend on the unknown regression and classification model parameter which has to be computed. We will identify these optimal probability distributions as functions of the unknown parameter only, which enables us to substitute them into the expression for the risk functional and to compute the optimal model parameter by minimising the risk measure over the set of possible values for the parameter.

The sets of probability distributions can be constructed from training data by a variety of statistical inference methods, including imprecise ('generalized') Bayesian inference models [19, 34, 35], nonparametric predictive inference [3] or belief functions [1, 6, 7, 13, 22]. The approach has recently been used in regression modelling with precise statistical data using Kolmogorov-Smirnov (KS) confidence bounds [30] and also includes imprecise Bayesian normal regression [28]. In this paper, there is special attention to the use of extended support vector machines (SVMs) [10, 33] to construct sets of probability distributions in case of interval-valued data, as SVMs are popular tools in machine learning. It will be interesting to implement the general approach presented here with a wide range of methods for constructing the sets of probability distributions and to compare the resulting inferences, for example also with regard to the effect of parameters such as the chosen confidence level if KS bounds are used; this is left as an important topic for future research.

2 Regression and classification in the machine learning framework

The standard learning problem can be formulated as follows [10, 33]. We select the best available function $f(\mathbf{x}, \alpha_{\text{opt}})$ from the set of functions $f(\mathbf{x}, \alpha)$ parameterized by parameter $\alpha \in \Lambda$ (this parameter is typically multi-dimensional), so the function $f(\mathbf{x}, \alpha_{\text{opt}})$ is considered to be the best approximation of the system response. The selection of the desired function is based on a (training) set of n observations $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$, assumed to be independent (conditionally on the assumed model) and identically distributed with probability density function (PDF) $p(\mathbf{x}, y) = p(y \mid \mathbf{x})p(\mathbf{x})$ and CDF $F(\mathbf{x}, y)$. Here $\mathbf{x} \in \mathbb{R}^m$ is a multivariate input and y is a scalar output which takes values from $\mathbb{R}$ for the regression model and from the set $\{-1, 1\}$ for the classification model[3]. The regression and classification models can be regarded as special cases of the general learning problem, the method presented here is widely applicable.

The quality of an approximation $f(\mathbf{x}, \alpha)$ in a regression model is measured by the loss function $L(y, f(\mathbf{x}, \alpha))$ which typically depends on the difference $z = y - f(\mathbf{x}, \alpha)$. Therefore, we use the notation $L(z) = L(y, f(\mathbf{x}, \alpha))$. Common and convenient loss functions are the quadratic loss $L(z) = z^2$, the linear loss $L(z) = |z|$, and the so-called 'ε-insensitive' [33] and 'pinball' loss functions [12]. In classification models, commonly used loss functions are the indicator loss function $L(\mathbf{x}, y) = \mathbf{1}\{\text{sgn}(f(\mathbf{x}, \alpha)) \neq y\}$, the logistic loss, the hinge loss, the squared hinge loss and the least square loss functions [21]. All these loss functions can be implemented in the general approach presented in this paper.

The main goal of learning is to find the optimal parameter α_{opt} which minimises the following risk functional over the parametrized class of functions $f(\mathbf{x}, \alpha)$, $\alpha \in \Lambda$:

$$R(\alpha) = \int\int L(y - f(\mathbf{x}, \alpha))p(\mathbf{x}, y)\mathrm{d}\mathbf{x}\mathrm{d}y.$$

A commonly made assumption for regression models is that the random error (noise) Z, which takes the values $z = y - f(\mathbf{x}, \alpha)$, has mean zero and PDF $p(z \mid \alpha) = p(y \mid \mathbf{x})$, leading to

$$R(\alpha) = \int L(z \mid \alpha)p(z \mid \alpha)\mathrm{d}z.$$

If the joint density $p(\mathbf{x}, y)$ is unknown (or no specific form of it has been assumed), then the risk functional $R(\alpha)$ can be replaced by the *empirical risk functional*

$$R_{\text{emp}}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} L(y - f(\mathbf{x}, \alpha)). \tag{1}$$

If $p(\mathbf{x}, y)$ is known or of an assumed parametric form, then a common technique for computing α_{opt} is the maximum likelihood estimation method [33].

In this paper we assume that the function f is linear,

$$f(\mathbf{x}, \alpha) = \alpha_0 + \langle \alpha\varphi(\mathbf{x}) \rangle$$

with $\langle\cdot\rangle$ the canonical dot product notation. In particular we consider the function with $\varphi_i(x_i) = x_i$, which corresponds to many popular models in learning. The use of more general functions f will be discussed elsewhere.

3 Regression with a set of distributions

Suppose that we do not know the precise CDF of Z, but we know that it belongs to a set $\mathcal{F}(\alpha)$ bounded by lower CDF $\underline{F}(z \mid \alpha)$ and upper CDF $\overline{F}(z \mid \alpha)$ which depend on the parameter α. As mentioned before, these bounds can result from the use of a wide range of inferential methods applied to the observations $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$. It is important to emphasize that the dependence of the lower and upper CDFs on the parameter α is an important feature of the proposed approach. When we have a set of probability distributions instead of a single one, we can construct a corresponding set of regression models. For decision making, it is important to choose some of these models[4], we consider the use of the minimax ('pessimistic') and minimin ('optimistic') strategies to judge the quality of an estimator and hence of the corresponding regression model.

3.1 The minimax strategy

The minimax strategy can be motivated as follows. We do not know (or wish to assume) a precise CDF F and every CDF in $\mathcal{F}(\alpha)$ could be selected. Therefore, we should take the 'worst' distribution providing the largest value of the risk functional. The minimax criterion can be interpreted as an insurance against the worst case because it aims at minimising the expected loss in the least favorable case [20]. The upper risk functional $\overline{R}(\alpha)$ for α is defined as

$$\overline{R}(\alpha) = \max_{F(z \mid \alpha) \in \mathcal{F}(\alpha)} \int L(z \mid \alpha)\mathrm{d}F(z \mid \alpha). \tag{2}$$

[3] Generally, y might take values in any finite set, the restriction to binary classification is due to space limitations.

[4] Alternative methods for dealing with the set of regression models can be of interest but are not investigated here.

It can be regarded as the upper expectation of the loss function. The optimal parameter α_{opt} is computed by minimising the upper risk functional over the set Λ.

Most loss functions in regression models have one minimum at point 0. Utkin and Destercke [31, 32] have shown that the optimal CDF from the set $\mathcal{F}(\alpha)$ providing the upper bound for $R(\alpha)$ in case of such loss functions is of the form

$$F_U(z) = \begin{cases} \overline{F}(z), & z \le \overline{F}^{-1}(\tau), \\ \tau, & \overline{F}^{-1}(\tau) < z < \underline{F}^{-1}(\tau), \\ \underline{F}(z), & z \ge \underline{F}^{-1}(\tau), \end{cases} \tag{3}$$

where τ is one of the roots of the equation

$$L\left(\overline{F}^{-1}(\tau)\right) = L\left(\underline{F}^{-1}(\tau)\right).$$

If the loss function is symmetric about 0, then τ can be derived from the equation $\underline{F}^{-1}(\tau) + \overline{F}^{-1}(\tau) = 0$. Using this optimal CDF, the upper risk functional $\overline{R}(\alpha)$ is

$$\overline{R}(\alpha) = \int_{-\infty}^{\overline{F}^{-1}(\tau)} L(z \mid \alpha) \mathrm{d}\overline{F}(z \mid \alpha) + \int_{\underline{F}^{-1}(\tau)}^{\infty} L(z \mid \alpha) \mathrm{d}\underline{F}(z \mid \alpha). \tag{4}$$

The optimal value of parameter α according to the minimax strategy can be derived by minimising $\overline{R}(\alpha)$ over $\alpha \in \Lambda$.

3.2 The minimin strategy

The minimin strategy can be interpreted as corresponding to an 'optimistic' decision, namely a CDF $F(z \mid \alpha) \in \mathcal{F}(\alpha)$ is used which provides the smallest value for the risk functional $R(\alpha)$ for arbitrary values of α. The corresponding lower risk functional for α is defined as

$$\underline{R}(\alpha) = \min_{F(z \mid \alpha) \in \mathcal{F}(\alpha)} \int L(z \mid \alpha) \mathrm{d}F(z \mid \alpha). \tag{5}$$

It can be regarded as the lower expectation of the loss function. The optimal parameter α_{opt} is computed by minimising the lower risk functional over the set Λ.

The optimal CDF from the set $\mathcal{F}(\alpha)$ providing the lower bound for the expectation is

$$F_L(z) = \begin{cases} \underline{F}(z), & z \le 0, \\ \overline{F}(z), & z > 0. \end{cases} \tag{6}$$

Using this optimal CDF, which has a jump at point $z = 0$, the lower risk functional $\underline{R}(\alpha)$ is

$$\underline{R}(\alpha) = \int_{-\infty}^{0} L(z \mid \alpha) \mathrm{d}\underline{F}(z \mid \alpha) + \int_{0}^{\infty} L(z \mid \alpha) \mathrm{d}\overline{F}(z \mid \alpha). \tag{7}$$

The optimal value of parameter α according to the minimin strategy can be derived by minimising $\underline{R}(\alpha)$ over $\alpha \in \Lambda$.

4 Regression with interval-valued observations

Suppose that the training set consists of n independent observations $(\mathbf{x}_i, \mathcal{Y}_i)$, $i = 1, ..., n$, with intervals $\mathcal{Y}_i = [\underline{y}_i, \overline{y}_i]$ instead of point-valued observations[5]. This implies that the random noise Z takes values in intervals $\mathcal{Z}_i(\alpha)$ such that $y - f(\mathbf{x}_i, \alpha) \in \mathcal{Z}_i(\alpha)$ for all $y \in \mathcal{Y}_i$. The question that needs to be addressed is how to proceed with the interval-valued training set in the framework of predictive learning.

There are several ways in which one could deal with such an interval-valued data set. In this paper, we construct the lower and upper CDFs for a set of probability distributions corresponding to the available information through a chosen inferential method out of a wide range of possibilities, as discussed before. This set depends on the parameter α because the intervals $\mathcal{Z}_i(\alpha)$, $i = 1, ..., n$ are functions of α. With such intervals $\mathcal{Z}_i(\alpha)$, the same approach as proposed by Utkin and Coolen [30], who used p-boxes corresponding to Kolmogorov-Smirnov bounds, can be applied for parameter optimisation in the regression model under the minimax and minimin scenarios. Denoting the boundary points of intervals $\mathcal{Z}_i(\alpha)$ by $\underline{\mathcal{Z}}_i(\alpha) = \underline{y}_i - f(\mathbf{x}_i, \alpha)$ and $\overline{\mathcal{Z}}_i(\alpha) = \overline{y}_i - f(\mathbf{x}_i, \alpha)$, a p-box can be constructed from the observed intervals in the framework of Dempster-Shafer theory [6, 22]. If we assume for simplicity that every observation interval occurs only once, then

$$\underline{F}(z \mid \alpha) = \text{Bel}((-\infty, z]) = n^{-1} \sum_{i: \overline{\mathcal{Z}}_i(\alpha) \le z} 1,$$

$$\overline{F}(z \mid \alpha) = \text{Pl}((-\infty, z]) = n^{-1} \sum_{i: \underline{\mathcal{Z}}_i(\alpha) \le z} 1.$$

If some intervals occur more than once then the corresponding CDFs follow straightforwardly. These lower and upper CDFs, which depend on the parameter α, can be used for dealing with interval-valued y in regression, as is illustrated next for several scenarios.

[5] The method presented in this paper can also deal with interval-valued input variables $\mathbf{x}_i$. Due to space limitations, for regression the presentation is restricted to point-valued input variables, but for classification (Section 5) interval-valued input variables are used. Throughout, the intervals are not restricted, hence they can be any interval of possible values upto the whole of $(-\infty, \infty)$.

4.1 The minimax strategy

With the lower and upper CDFs corresponding to the interval-valued observations as discussed above, the upper risk functional in (4) is

$$\overline{R}(\alpha) = n^{-1} \sum_{i:\underline{\mathcal{Z}}_i(\alpha) \leq \overline{F}^{-1}(\tau)} L(\underline{\mathcal{Z}}_i(\alpha)) + n^{-1} \sum_{i:\overline{\mathcal{Z}}_i(\alpha) \geq \underline{F}^{-1}(\tau)} L(\overline{\mathcal{Z}}_i(\alpha)).$$

with τ such that $\underline{F}^{-1}(\tau) = -\overline{F}^{-1}(\tau)$. Note that this upper risk functional uses, for every α, only the boundary points $\underline{\mathcal{Z}}_i(\alpha)$ and $\overline{\mathcal{Z}}_i(\alpha)$ of the intervals $\mathcal{Z}_i(\alpha)$. This feature is important as it significantly simplifies computation of the optimal parameter α_{opt}.

The upper risk functional for the minimax strategy with a fixed α can be written as the upper expectation corresponding to basic probability assignments [15, 24], giving

$$\overline{R}(\alpha) = n^{-1} \sum_{i=1}^{n} \max_{z \in [\underline{\mathcal{Z}}_i(\alpha), \overline{\mathcal{Z}}_i(\alpha)]} L(z).$$

We also concluded that this upper risk functional is achieved at boundary points of intervals $\mathcal{Z}_i$, with

$$\overline{R}(\alpha) = n^{-1} \sum_{i=1}^{n} \max \left\{ L(\underline{\mathcal{Z}}_i(\alpha)), L(\overline{\mathcal{Z}}_i(\alpha)) \right\}.$$

It should be pointed out that, if all observations are precise ('point-valued'), so $\underline{y}_i = \overline{y}_i = y_i$, this upper risk functional is equal to the standard empirical risk functional (1). We can now consider some of the most important loss function in regression, where the optimal parameter α_{opt} under minimax can be obtained by minimising $\overline{R}(\alpha)$ over all $\alpha \in \Lambda$.

4.1.1 Quadratic loss function

We consider the quadratic loss function $L(z) = z^2$, the most popular one in classical regression theory and applications. To minimise the corresponding upper risk functional we have to solve the optimisation problem functional:

$$\min_{\alpha} \left(\sum_{i=1}^{n} \max \left\{ \underline{\mathcal{Z}}_i{}^2(\alpha), \overline{\mathcal{Z}}_i{}^2(\alpha) \right\} \right). \tag{8}$$

Introducing new optimisation variables G_i, $i = 1, ..., n$, such that $G_i^2 = \max \left\{ \underline{\mathcal{Z}}_i{}^2(\alpha), \overline{\mathcal{Z}}_i{}^2(\alpha) \right\}$, problem (8) can be rewritten as

$$\min_{\alpha, G_i} \sum_{i=1}^{n} G_i^2, \tag{9}$$

subject to

$$\begin{aligned} &G_i \geq \underline{\mathcal{Z}}_i(\alpha), \quad G_i \geq \overline{\mathcal{Z}}_i(\alpha), \\ &G_i \geq -\underline{\mathcal{Z}}_i(\alpha), \; G_i \geq -\overline{\mathcal{Z}}_i(\alpha), \; i = 1, ..., n. \end{aligned} \tag{10}$$

The third and fourth constraints take into account the fact that residuals may be negative. If we assume that the function $f(\mathbf{x}, \alpha)$ is linear, i.e., $f(\mathbf{x}, \alpha) = \alpha_0 + \langle \alpha \mathbf{x} \rangle$, then the optimisation problem specified by (9) and (10) is a well-known quadratic programming problem with the optimisation variables α and G_i, $i = 1, ..., n$, which can be solved by means of standard methods.

4.1.2 Linear and pinball loss function

The pinball loss function with parameter $\tau \in [0, 1]$ is given by [12]

$$L_\tau(z) = \begin{cases} \tau z, & z > 0, \\ (\tau - 1)z, & z \leq 0. \end{cases}$$

The linear loss function is the special case of the pinball loss function with $\tau = 1$. We consider calculation of the optimal parameter of the regression model using the minimax criterion with the pinball loss function. We introduce new optimisation variables G_i, $i = 1, ..., n$, such that $G_i = \max \left\{ L_\tau(\underline{\mathcal{Z}}_i(\alpha)), L_\tau(\overline{\mathcal{Z}}_i(\alpha)) \right\}$. The condition $z \geq 0$ implies the condition $G_i \geq \tau \cdot z$. However, if $G_i \geq \tau \cdot z$ and $z \geq 0$, then $G \geq \tau \cdot z - z$. On the other hand, the condition $z < 0$ implies the condition $G_i \geq (\tau - 1) \cdot z = \tau \cdot z - z$. However, if $G_i \geq \tau \cdot z - z$ and $z < 0$, then $G_i \geq \tau \cdot z$. Finally, the condition $G_i \geq L_\tau(z)$ can be represented by means of two constraints $G_i \geq \tau \cdot z$ and $G_i \geq \tau \cdot z - z$, which simultaneously 'cover' all possible values of z. This implies that the optimisation problem for computing the optimal regression parameter can be written as

$$\min_{\alpha, G_i} \sum_{i=1}^{n} G_i, \tag{11}$$

subject to

$$\begin{aligned} &G_i \geq \tau \cdot \underline{\mathcal{Z}}_i(\alpha), \; G_i \geq \tau \cdot \overline{\mathcal{Z}}_i(\alpha), \\ &G_i \geq (\tau - 1) \cdot \underline{\mathcal{Z}}_i(\alpha), \\ &G_i \geq (\tau - 1) \cdot \overline{\mathcal{Z}}_i(\alpha), \; i = 1, ..., n. \end{aligned} \tag{12}$$

If we assume that the function $f(\mathbf{x}, \alpha)$ is linear, then this is a well-known linear programming problem.

4.2 SVM

Let us return to the case with the linear loss function and the minimax strategy, and compare the obtained optimisation problem with the popular SVM approach [10, 21, 33] which in regression is also called

'support vector regression'. The ε-insensitive loss function is applied in the corresponding regression models [33]. If all observations are point-valued, so $\underline{y}_i = \overline{y}_i = y_i$, then according to the standard SVR approach, parameter α is determined by the quadratic programming problem

$$\min_{\alpha} \left(\frac{1}{2} \langle \alpha, \alpha \rangle + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \right) \tag{13}$$

subject to

$$\begin{aligned} &\xi_i \geq 0,\ \xi_i + \varepsilon \geq (\langle \alpha \mathbf{x}_i \rangle + \alpha_0) - y_i, \\ &\xi_i^* \geq 0,\ \xi_i^* + \varepsilon \geq y_i - (\langle \alpha \mathbf{x}_i \rangle + \alpha_0), i = 1, .., n. \end{aligned} \tag{14}$$

Here C is a constant 'cost' parameter, ξ_i, ξ_i^*, $i = 1, ..., n$, are slack variables, and $\frac{1}{2} \langle \alpha, \alpha \rangle$ is the Tikhonov regularization term (the most popular penalty or smoothness term) [27] which enforces uniqueness by penalizing functions with wild oscillation and effectively restricting the space of admissible solutions [8]. The positive slack variables ξ_i, ξ_i^* represent the distance from y_i to the corresponding boundary values of the ε-tube.

The constraints (12) and (14) coincide if the variables G_i coincide with the slack variables ξ_i, ξ_i^* and $\underline{y}_i = \overline{y}_i$, $\varepsilon = 0$, $\tau = 1$. Consequently, the proposed approach for constructing the regression model with interval-valued data, supplemented by the regularization term and the constant 'cost' parameter C, can be regarded as an extension of the SVM approach to the case of interval-valued data, i.e. we have the same objective function and the following constraints in terms of SVR for every $i = 1, ..., n$:

$$\begin{aligned} &\xi_i \geq 0, \ \ \xi_i + \varepsilon \geq (\langle \alpha \mathbf{x}_i \rangle + \alpha_0) - \underline{y}_i, \\ &\xi_i \geq 0, \ \ \xi_i + \varepsilon \geq (\langle \alpha \mathbf{x}_i \rangle + \alpha_0) - \overline{y}_i, \\ &\xi_i^* \geq 0, \ \ \xi_i^* + \varepsilon \geq \underline{y}_i - (\langle \alpha \mathbf{x}_i \rangle + \alpha_0), \\ &\xi_i^* \geq 0, \ \ \xi_i^* + \varepsilon \geq \overline{y}_i - (\langle \alpha \mathbf{x}_i \rangle + \alpha_0). \end{aligned}$$

Now the slack variables ξ_i, ξ_i^* are additionally constrained and represent the largest distance from $\underline{y}_i$ and $\overline{y}_i$ to the corresponding boundary values of the ε-tube, respectively. This implies that the minimax strategy searches for the largest residuals (or 'margins' in terms of classification) from all residuals in every interval $\mathcal{Z}_i$, $i = 1, ..., n$. The corresponding dual optimisation problem is

$$\begin{aligned} \max \Bigg(&-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (Q_i - T_i)(Q_j - T_j) \langle \mathbf{x}_i \mathbf{x}_j \rangle \\ &- \varepsilon \sum_{i=1}^{n} (Q_i + T_i) - \sum_{i=1}^{n} \underline{y}_i (Q_i - T_i) \\ &+ \sum_{i=1}^{n} (\overline{y}_i - \underline{y}_i)(\varphi_i^* - \varphi_i) \Bigg), \end{aligned}$$

subject to

$$\sum_{i=1}^{n} (Q_i - T_i) = 0,\ 0 \leq Q_i \leq C,\ 0 \leq T_i \leq C.$$

Here $\psi_i, \psi_i^*, \varphi_i, \varphi_i^*$ are Lagrange multipliers and $Q_i = \psi_i + \varphi_i$, $T_i = \psi_i^* + \varphi_i^*$.

It can be seen from this dual optimisation problem that in the regression model we use a point in every observation interval which is a linear combination of its bounds $\underline{y}_i$ and $\overline{y}_i$ with coefficients determined by the values of the Lagrange multipliers. If $\overline{y}_i = \underline{y}_i$ we get the dual optimisation problem of the standard SVM method with variables Q_i and T_i.

If the quadratic loss function is used instead of the ε-insensitive loss function, then the proposed regression model (optimisation problem (9)-(10)) is the 'least squares SVM' approach [25] which is solved through a system of linear equations.

4.3 The minimin strategy

Using the lower and upper CDFs corresponding to the interval-valued observatons, as discussed at the start of this section, we can rewrite the lower risk functional (7) as

$$\begin{aligned} \underline{R}(\alpha) = n^{-1} &\sum_{i: \overline{\mathcal{Z}}_i(\alpha) \leq 0} L(\overline{\mathcal{Z}}_i(\alpha)) \\ &+ n^{-1} \sum_{i: \underline{\mathcal{Z}}_i(\alpha)) \geq 0} L(\underline{\mathcal{Z}}_i(\alpha)). \end{aligned}$$

As in the case of the upper risk function, this lower risk functional is, for every α, defined only by the boundary points $\underline{\mathcal{Z}}_i(\alpha)$ and $\overline{\mathcal{Z}}_i(\alpha)$ of the intervals $\mathcal{Z}_i(\alpha)$. However, not all observation intervals contribute to the lower risk functional because the optimal CDF has a jump at point 0.

The lower risk functional for the minimin strategy with a fixed α can be written as the lower expectation corresponding to basic probability assignments [15, 24], giving

$$\underline{R}(\alpha) = n^{-1} \sum_{i=1}^{n} \min_{z \in [\underline{\mathcal{Z}}_i(\alpha), \overline{\mathcal{Z}}_i(\alpha)]} L(z). \tag{15}$$

It follows that the risk measure is 0 if there exist one or more values of α such that $0 \in [\underline{\mathcal{Z}}_i(\alpha), \overline{\mathcal{Z}}_i(\alpha)]$ for every $i = 1, ..., n$. If this is the case for multiple vectors α, one can consider to have found several 'perfect fits' to the available data, which either could be considered all together (this would be in line with some fundamental ideas behind imprecise probability) or which could be compared by a secondary criterion (the same comment applies generally if there are multiple optimal vectors α). This is an interesting topic for future research, for now let us assume that a unique best estimate of α can be obtained and that the corresponding lower risk functional is positive (so there is no 'perfect fit'). A term in the objective function is non-zero if one of the following two conditions holds

$$\overline{\mathcal{Z}}_i(\alpha) < 0, \quad \underline{\mathcal{Z}}_i(\alpha) > 0.$$

Let us consider the pinball loss function for this situation. Introducing new optimisation variables H_i, $i = 1, ..., n$, it is easy to prove that the optimisation problem can be written as

$$\min_{\alpha, H_i} \sum_{i=1}^{n} H_i,$$

subject to

$$H_i \geq \tau \underline{\mathcal{Z}}_i(\alpha), \; H_i \geq (\tau - 1)\overline{\mathcal{Z}}_i(\alpha),$$
$$H_i \geq 0, \; i = 1, ..., n.$$

The quadratic loss function leads to the similar problem with H_i replaced by G_i^2, minimisation over G_i, and $\tau = 1$. These are well-known optimisation problems that can be solved efficiently by standard methods.

4.4 SVM

We consider the case with the linear loss function under the minimin strategy and derive the optimisation problem in the SVM framework. By using the standard Tikhonov regularization term, we can formulate the following convex optimisation problem

$$\min_{\alpha} \left(\frac{1}{2} \langle \alpha, \alpha \rangle + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \right),$$

subject to

$$\xi_i \geq 0, \; \xi_i + \varepsilon \geq (\langle \alpha \mathbf{x}_i \rangle + \alpha_0) - \overline{y}_i, \; i = 1, ..., n,$$
$$\xi_i^* \geq 0, \; \xi_i^* + \varepsilon \geq \underline{y}_i - (\langle \alpha \mathbf{x}_i \rangle + \alpha_0), \; i = 1, ..., n.$$

This is a quadratic programming problem, with slack variables ξ_i, ξ_i^* representing the distance from $\overline{y}_i$ and $\underline{y}_i$ to the corresponding lower and upper boundary values of the ε-tube, respectively. The minimin strategy searches for the smallest residuals in each interval $\mathcal{Z}_i$, $i = 1, ..., n$, under condition that there are positive residuals. As in Subsection 4.2, the corresponding dual optimisation problem provides further insights into the optimal solution, a detailed analysis will be presented elsewhere.

5 Classification with interval-valued observations

We consider classification problems where the system output y is restricted to two values, the proposed method can be generalized to more possible values. The input variables (patterns) $\mathbf{x}$ may be interval-valued. Suppose that we have a training set $(\mathcal{X}_i, y_i)$, $i = 1, ..., n$. Here $\mathcal{X}_i \subset \mathbb{R}^m$ is the Cartesian product of m intervals $[\underline{x}_k^{(i)}, \overline{x}_k^{(i)}]$, $k = 1, ..., m$, which again are not restricted so could even include intervals $(-\infty, \infty)$, and $y_i \in \{-1, 1\}$. Let the $n_{-1} = r$ observations $\mathcal{X}_i$ with $i = 1, \ldots, r$ correspond to the class (with) $y = -1$ and the $n_{+1} = n - r$ observations $\mathcal{X}_i$ with $i = r+1, \ldots, n$ correspond to the class $y = 1$.

The risk functional can be written as $R(\alpha) = R_{-1}(\alpha) + R_{+1}(\alpha)$, with

$$R_y(\alpha) = \int_{\mathbb{R}^n} L(\mathbf{x}, y) \mathrm{d}F(\mathbf{x}, y)$$
$$= \pi_y \int_{\mathbb{R}^n} L(\mathbf{x}, y) \mathrm{d}F(\mathbf{x} \mid y),$$

where $\pi_y = p(y)$ is a prior probability[6] for class y. Suppose that the CDFs $F(\mathbf{x} \mid y)$ are unknown. As discussed before, a wide range of inferential methods can be chosen to, in combination with the dataset containing interval-valued observations, produce a set of CDFs $F(\mathbf{x} \mid y)$. One additional obstacle due to the interval-valued input variables is that $\mathbf{x}$ is a vector so now p-boxes of multivariate distributions must be constructed. We propose that this problem can be resolved as follows. Note that interval-valued data $\mathbf{x}$ lead to an interval-valued discriminant function $f(\mathbf{x}, \alpha)$ whose parameter α is unknown and has to be determined. Therefore, in contrast to many alternative approaches in classification, we propose to consider the CDF $F(f \mid y)$ instead of the multivariate CDF $F(\mathbf{x} \mid y)$. This is briefly discussed further below, detailed explanation and illustrations will be presented elsewhere. With this change, the risk functional becomes

$$R_y(\alpha) = \pi_y \int_{\mathbb{R}} L(f \mid y) \mathrm{d}F(f \mid y).$$

[6] Choice of prior probabilities is not addressed here. However, it is worth noting that generalization to allow imprecise prior probabilities is possible.

But we allowed explicitly the use of a set of CDFs, so now consider the set $\mathcal{F}(y)$ of probability distributions produced by lower CDF $\underline{F}(f \mid y)$ and upper CDF $\overline{F}(f \mid y)$ CDFs , i.e.

$$\mathcal{F}(y) = \{F(f) \mid \forall f \in \mathbb{R}, \underline{F}(f|y) \leq F(f) \leq \overline{F}(f|y)\}.$$

It is important to emphasize that, although we have not explicitly included α in the notation for these distribution sets, $\mathcal{F}(y)$ depends on the parameter α because f is a function of α and the lower and upper CDFs depend on α. We introduce notation

$$f_L = \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \alpha), \quad f_U = \max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \alpha).$$

If the function f is linear, then the lower and upper bounds for the discriminant functions are determined only by the bounds of pattern intervals, i.e.

$$f_L = \min_{x_k \in \{\underline{x}_k, \overline{x}_k\},\ k=1,...,m} f(\mathbf{x}, \alpha),$$
$$f_U = \max_{x_k \in \{\underline{x}_k, \overline{x}_k\},\ k=1,...,m} f(\mathbf{x}, \alpha).$$

This property is also valid for arbitrary monotone discriminant functions. For every interval-valued observation $(\mathcal{X}_i, y_i)$, we have the interval $\mathbf{f}_i = [f_{L,i}, f_{U,i}]$ of values of the discriminant function. These intervals depend on the parameter α, so the bounds $f_{L,i}$ and $f_{U,i}$ cannot be computed explicitly, but inference is again possible in many important scenarios through specification of the optimisation problems involved, and the use of standard algorithms to solve such problems. We illustrate this next for the minimax strategy, methods for the minimim strategy can be developed similarly and will be presented elsewhere.

5.1 The minimax strategy

According to the minimax strategy, we select a probability distribution from the set $\mathcal{F}(-1)$ and a probability distribution from the set $\mathcal{F}(+1)$ such that the risk measures $R_{-1}(\alpha)$ and $R_{+1}(\alpha)$ achieve their maxima for every fixed α. It must be emphasized that the 'optimal' probability distributions may be different for different values of parameter α, which implies that the corresponding 'optimal' probability distributions depend on α. Since the sets $\mathcal{F}(-1)$ and $\mathcal{F}(1)$ are obtained independently for $y = -1$ and $y = 1$, the upper risk functional with respect to the minimax strategy is of the form

$$\overline{R}(\alpha) = \max_{F(f|-1) \in \mathcal{F}(-1)} R_{-1}(\alpha) + \max_{F(f|1) \in \mathcal{F}(1)} R_{+1}(\alpha).$$

For many popular loss functions in such classification the loss function $L(f, -1)$ is increasing. If this is the case, then the upper bound for $R_{-1}(\alpha)$ is achieved at the distribution $\underline{F}(f, -1)$, hence

$$\overline{R}_{-1}(\alpha) = \int_{\mathbb{R}} L(f, -1) \mathrm{d}\underline{F}(f, -1).$$

In this case the function $L(f, 1)$ is decreasing, so

$$\overline{R}_{+1}(\alpha) = \int_{\mathbb{R}} L(f, 1) \mathrm{d}\overline{F}(f, 1).$$

The upper expectation $\overline{R}_{-1}(\alpha)$ corresponding to given basic probability assignments $m(\mathbf{f}_i) = r^{-1}$ for intervals $\mathbf{f}_i$, $i = 1, ..., r$, can be derived for fixed α by [15, 24]

$$\overline{R}_{-1}(\alpha) = r^{-1} \sum_{i=1}^{r} \max_{f \in [f_{L,i}(\alpha), f_{U,i}(\alpha)]} L(f, -1)$$
$$= r^{-1} \sum_{i=1}^{r} L(f_{U,i}(\alpha), -1).$$

And similarly, the corresponding upper expectation $\overline{R}_{+1}(\alpha)$ is

$$\overline{R}_{+1}(\alpha) = (n - r)^{-1} \sum_{i=r+1}^{n} L(f_{L,i}(\alpha), 1).$$

Finally, we minimise $\overline{R}(\alpha)$ to compute α_{opt}, with

$$\overline{R}(\alpha) = \frac{\pi_-}{r} \sum_{i=1}^{r} L(f_{U,i}(\alpha), -1)$$
$$+ \frac{\pi_+}{n - r} \sum_{i=r+1}^{n} L(f_{L,i}(\alpha), 1).$$

Further steps towards the solution of the problem depend on the chosen loss function, we briefly consider one important special case. For the hinge loss function $L(\mathbf{x}, y) = \max(1 - yf, 0)$,

$$\overline{R}(\alpha) = \frac{\pi_-}{r} \sum_{i=1}^{r} \max\left(0, 1 + f_{U,i}(\alpha)\right)$$
$$+ \frac{\pi_+}{n - r} \sum_{i=r+1}^{n} \max\left(0, 1 - f_{L,i}(\alpha)\right).$$

After simple modifications, we get the linear problem

$$\min_{\alpha} \left(\frac{\pi_-}{r} \sum_{i=1}^{r} G_i + \frac{\pi_+}{n - r} \sum_{i=r+1}^{n} G_i \right) \tag{16}$$

subject to

$$G_i \geq 1 - y_i \left(\langle \alpha \mathbf{x}_i \rangle + \alpha_0 \right), \ \forall x_k^{(i)} \in \{\underline{x}_k^{(i)}, \overline{x}_k^{(i)}\},$$
$$G_i \geq 0, \ i = 1, ..., n. \tag{17}$$

By adding the standard Tikhonov regularization term to the objective function, we get the SVM classifier with cost parameters $C_- = \pi_-/r$ and $C_+ = \pi_+/(n-r)$. We introduce notation

$$Q_i = \sum_{k \in J_i} \psi_{ik},\ T_j(i) = \sum_{k \in J_i(j)} \psi_{ik} x_j^{(i,k)}$$

where the set $J_i(j)$ is a 'projection' of the set of indices on the j-th element of the vector $\mathbf{x}_i$. Then the dual optimisation problem is

$$\max \left(\sum_{i=1}^{n} Q_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \left(\sum_{v=1}^{m} T_v(i) T_v(j) \right) \right),$$

subject to

$$\sum_{i=1}^{n} y_i Q_i = 0,\ 0 \le Q_i \le C_-,\ i = 1, ..., r,$$
$$0 \le Q_i \le C_+,\ i = r+1, ..., n.$$

This is the SVM classification approach with interval-valued data under the minimax strategy. Space restrictions prevent further details, illustration or discussion of this result and related results for different loss functions and for the minimin strategy. However, it is clear that the general approach presented in this paper leads to a wide variety of attractive methods for machine learning, with relatively straightforward inclusion of interval-valued observations.

6 Concluding remarks

In this paper, a new class of imprecise regression and classification models has been proposed which are capable to deal with interval-valued data as frequently occur in practice. The class has been illustrated for several important specific cases, and it has been shown that the resulting inference problems can be formulated as standard optimisation problems, so the method can be implemented using readily available software. This new method has several important features. First, it has a clear explanation and justification in the decision making framework. Secondly, it allows a wide variety of inferential methods for constructing the p-boxes. For example, imprecise ('generalized') Bayesian inference models [19] can be used and these provide an exciting opportunity for developing learning models for a wide range of different applications. Thirdly, the method can deal with (partly) missing data as the intervals for observations are not restricted, which is important as complete data sets are the exception in practice. Finally[7], resulting statistical inferences are similar some well-known robust statistics methods, for which the current approach provides formal justifications and interpretations in a decision theoretic framework. Detailed study of these aspects, and development of further models and corresponding inferences, is ongoing. The main disadvantage of the proposed approach is that it is often not straightforward how the bounding CDFs can be explicitly defined as functions of the regression or classification parameter, which may add to computational complexity but the results show that the approach can be developed to allow real-world applications. A main strength of the proposed method is the link with the popular SVM approach. A key feature of SVMs is the use of kernels which are functions that transform the input data to a high-dimensional space where the learning problem is solved. Such kernel functions can be linear or nonlinear, which will allow us to significantly extend the class of regression or discriminant functions that can be used. Our approach directly showed how the regular SVM approach can be generalized for dealing with interval-valued observations.

There are interesting possibilities for combining corresponding 'minimin' and 'minimax' strategies. For example, the method for cautious decision making proposed by Utkin and Augustin [29], which uses the extreme points of a set of probability distributions produced by imprecise data, can be applied. In our approach, the values of the extreme points are determined from the optimal CDFs (3) and (6) for the minimax and minimin strategies, respectively. Detailed analysis of this cautious strategy and the possibility to arrive at set-based predictions and related final decisions on the basis of our model outputs, are interesting topics for future research, together with dealing with imprecise input variables for the object to predict, imprecision in the dependent variables and of course comparison with more established methods.

[7] This was discussed by Utkin and Coolen [30] for p-boxes based on Kolmogorov-Smirnov bounds

Acknowledgements

We thank two referees for detailed comments and suggestions which all have improved this paper and will guide future research.

References

[1] A. Arequi, T. Denoeux. Constructing predictive belief functions from continuous sample data using confidence bands. *ISIPTA'07*[8], pp 11-20, 2007.

[2] C. Angulo, D. Anguita, L. Gonzalez-Abril, J.A. Ortega. Support vector machines for interval discriminant analysis. *Neurocomputing*, 71:1220–1229, 2008.

[8] Proceedings ISIPTA conferences available from *www.sipta.org*

[3] T. Augustin, F.P.A. Coolen. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124:251–272, 2004.

[4] E. Carrizosa, J. Gordillo, F. Plastria. Classification problems with imprecise data through separating hyperplanes. Technical Report MOSI/33, Vrije Universiteit Brussel, 2007.

[5] G. de Cooman, M. Zaffalon. Updating beliefs with incomplete observations. *Artificial Intelligence*, 159:75–125, 2004.

[6] A.P. Dempster. Upper and lower probabilities induced by a multi-valued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[7] S. Destercke, D. Dubois, E. Chojnacki. Unifying practical uncertainty representations - i: Generalized p-boxes. *International Journal of Approximate Reasoning*, 49:649–663, 2008.

[8] T. Evgeniou, T. Poggio, M. Pontil, A. Verri. Regularization and statistical learning theory for data analysis. *Computational Statistics & Data Analysis*, 38:421–432, 2002.

[9] P.Y. Hao. Interval regression analysis using support vector networks. *Fuzzy Sets and Systems*, 60:2466–2485, 2009.

[10] T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, 2001.

[11] H. Ishibuchi, H. Tanaka, N. Fukuoka. Discriminant analysis of multi-dimensional interval data and its application to chemical sensing. *International Journal of General Systems*, 16:311–329, 1990.

[12] R. Koenker, G. Bassett. Regression quantiles. *Econometrica*, 46:33–50, 1978.

[13] E. Kriegler, H. Held. Utilizing belief functions for the estimation of future climate change. *International Journal of Approximate Reasoning*, 39:185–209, 2005.

[14] E.A. Lima Neto, F.A.T. de Carvalho. Centre and range method to fitting a linear regression model on symbolic interval data. *Computational Statistics and Data Analysis*, 52:1500–1515, 2008.

[15] H.T. Nguyen, E.A. Walker. On decision making using belief functions. In: R.Y. Yager, M. Fedrizzi, J. Kacprzyk (Eds), *Advances in the Dempster-Shafer Theory of Evidence*. Wiley, New York, pp 311–330, 1994.

[16] P. Nivlet, F. Fournier, J.J. Royer. Interval discriminant analysis: An efficient method to integrate errors in supervised pattern recognition. *ISIPTA'01*, pp 284-292, 2001.

[17] K. Pelckmans, J. De Brabanter, J.A.K. Suykens, B. De Moor. Handling missing values in support vector machine classifiers. *Neural Networks*, 18:684 – 692, 2005.

[18] S. Petit-Renaud, T. Denoeux. Nonparametric regression analysis of uncertain and imprecise data using belief functions. *International Journal of Approximate Reasoning*, 35:1–28, 2004.

[19] E. Quaeghebeur, G. de Cooman. Imprecise probability models for inference in exponential families. *ISIPTA'05*, pp 287–296, 2005.

[20] C.P. Robert. *The Bayesian Choice*. Springer, New York, 1994.

[21] B. Scholkopf, A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, 2002.

[22] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[23] A. Silva, P. Brito. Linear discriminant analysis for interval data. *Computational Statistics*, 21:289–308, 2006.

[24] T.M. Strat. Decision analysis using belief functions. *International Journal of Approximate Reasoning*, 4:391–418, 1990.

[25] J.A.K. Suykens, J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9:293–300, 1999.

[26] H. Tanaka, H. Lee. Interval regression analysis by quadratic programming approach. *IEEE Transactions on Fuzzy Systems*, 6:473–481, 1998.

[27] A.N. Tikhonov, V.Y. Arsenin. *Solution of Ill-Posed Problems*. W.H. Winston, Washington DC, 1977.

[28] L.V. Utkin. Regression analysis using the imprecise Bayesian normal model. *International Journal of Data Analysis Techniques and Strategies*, 2:356–372, 2010.

[29] L.V. Utkin, T. Augustin. Efficient algorithms for decision making under partial prior information and general ambiguity attitudes. *ISIPTA'05*, pp 349–358, 2005.

[30] L.V. Utkin, F.P.A. Coolen. On reliability growth models using Kolmogorov-Smirnov bounds. *International Journal of Performability Engineering*, 7:5–19, 2011.

[31] L.V. Utkin, S. Destercke. Computing expectations with p-boxes: two views of the same problem. *ISIPTA'07*, pp 435–444, 2007.

[32] L.V. Utkin, S. Destercke. Computing expectations with continuous p-boxes: Univariate case. *International Journal of Approximate Reasoning*, 50:778 – 798, 2009.

[33] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[34] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[35] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. (With discussion) *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996.

[36] G. Walter, T. Augustin, A. Peters. Linear regression analysis under sets of conjugate priors. *ISIPTA'07*, pp 445–455, 2007.

Conditioning, Conditional Independence and Irrelevance in Evidence Theory

Jiřina Vejnarová
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
vejnar@utia.cas.cz

Abstract

The goal of the paper is to reveal the relationships between recently introduced concept of conditional independence in evidence theory and those (dependent on the choice of conditioning rule) of conditional irrelevance.

Keywords. Evidence theory, multidimensional models, conditioning rules, conditional independence, conditional irrelevance.

1 Introduction

When applying models of artificial intelligence to any practical problem one must cope with two basic problems: uncertainty and multidimensionality. The most widely used models managing these issues are, at present, so-called *probabilistic graphical Markov models*.

The problem of multidimensionality is solved in these models with the help of the notion of conditional independence, which enables factorization of a multidimensional probability distribution into small parts, usually marginal or conditional low-dimensional distributions (e.g. in *Bayesian networks*), or generally into low-dimensional factors (e.g. in *decomposable models*). Such a factorization not only decreases the storage requirements for representation of a multidimensional distribution but it usually also induces efficient computational procedures allowing inference from these models.

It is easy to realize that if we need efficient methods for representation of probability distributions (requiring an exponential number of parameters), the greater is the 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 its representation are superexponential. To solve this problem, in [9, 15] we proposed a new concept of conditional independence in evidence theory, proved its formal properties and showed [16] in which sense it is superior to the previous one [3].

However, another problem appears when one tries to construct an evidential counterpart of Bayesian network: problem of conditioning, which is not sufficiently solved in evidence theory. There exist many conditioning rules [6], but is any of them compatible with our conditional independence concept? In other words, if one is interested in Bayesian-networks-like evidential models, he/she will need rather the concept of conditional irrelevance. Therefore, it is also necessary to find the relationship between conditional independence and irrelevance. It is not necessary for Bayesian networks, as in (precise) probability framework the difference between conditional independence and irrelevance is only subtle.

The contribution is organized as follows. After a short overview of necessary terminology and notation (Section 2), in Section 3 we recall two conditioning rules (suggested for conditioning of events) and introduce their generalizations for variables. In Section 4 the above-mentioned concept of conditional independence is recalled and a new concept of (conditional) irrelevance is presented. In Section 5 the relationship between (conditional) independence and (conditional) irrelevance is studied.

2 Basic Concepts

In this section we will briefly recall basic concepts from evidence theory [11] concerning sets, set functions and marginalization.

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 *multidi-*

mensional 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 usually called a cylindrical extension. The *cylindrical extension* of $A \subset \mathbf{X}_K$ to $\mathbf{X}_L$ $(K \subset L)$ is the set

$$A^{\uparrow L} = \{x \in \mathbf{X}_L : x^{\downarrow K} \in A\}.$$

Clearly

$$A^{\uparrow L} = A \times \mathbf{X}_{L \setminus K}.$$

A more complicated case is to make common extension of two sets, which will be called a join. By a *join*[2] of two sets $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_L$ $(K, L \subseteq N)$ we will understand a set

$$A \bowtie B = \{x \in \mathbf{X}_{K \cup L} : x^{\downarrow K} \in A \ \ \& \ \ x^{\downarrow L} \in B\}.$$

Let us note that for any $C \subseteq \mathbf{X}_{K \cup L}$ naturally $C \subseteq C^{\downarrow K} \bowtie C^{\downarrow L}$, but generally $C \neq C^{\downarrow K} \bowtie C^{\downarrow L}$.

Let us also note that if K and L are disjoint, then the join of A and B is just their Cartesian product $A \bowtie B = A \times B$, if $K = L$ then $A \bowtie B = A \cap B$. If $K \cap L \neq \emptyset$ and $A^{\downarrow K \cap L} \cap B^{\downarrow K \cap L} = \emptyset$ then also $A \bowtie B = \emptyset$. Generally,

$$A \bowtie B = (A \times \mathbf{X}_{L \setminus K}) \cap (B \times \mathbf{X}_{K \setminus L}),$$

i.e. a join of two sets is the intersection of their cylindrical extensions.

[1] Let us remark that we do not exclude situations when $M = \emptyset$. In this case $A^{\downarrow \emptyset} = \emptyset$.

[2] This term and notation are taken from the theory of relational databases [1].

2.2 Set functions

In evidence theory [11] (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. ,

$$m : \mathcal{P}(\mathbf{X}_N) \longrightarrow [0, 1],$$

where $\mathcal{P}(\mathbf{X}_N)$ is power set of $\mathbf{X}_N$ and

$$\sum_{A \subseteq \mathbf{X}_N} m(A) = 1.$$

Furthermore, we assume that $m(\emptyset) = 0$.

A set $A \in \mathcal{P}(\mathbf{X}_N)$ is a *focal element* if $m(A) > 0$. Let $\mathcal{F}$ denote the set of all focal elements, a focal element $A \in \mathcal{F}$ is called an $m-atom$ if for any $B \subseteq A$ either $B = A$ or $B \notin \mathcal{F}$. In other words, $m-atom$ is a setwise-minimal focal element.

Let us note that atomicity of a focal element is not closed with respect to either marginalization or extension.

Belief and *plausibility measures* are defined for any $A \subseteq \mathbf{X}_N$ by the equalities

$$Bel(A) = \sum_{B \subseteq A} m(B). \tag{1}$$

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \tag{2}$$

respectively.

It is well-known (and evident from these formulae) that for any $A \in \mathcal{P}(\mathbf{X}_N)$

$$Bel(A) \leq Pl(A), \tag{3}$$

$$Pl(A) = 1 - Bel(A^C), \tag{4}$$

where A^C is the set complement of $A \in \mathcal{P}(\mathbf{X}_N)$. Furthermore, basic assignment can be computed from belief function via Möbius inversion:

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} Bel(B), \tag{5}$$

i.e. any of these three functions is sufficient to define values of the remaining two.

2.3 Marginalization

For a basic assignment m on $\mathbf{X}_K$ and $M \subset K$, a *marginal basic assignment* of m on $\mathbf{X}_M$ is defined (for each $A \subseteq \mathbf{X}_M$):

$$m^{\downarrow M}(A) = \sum_{\substack{B \subseteq \mathbf{X}_K \\ B^{\downarrow M} = A}} m(B). \tag{6}$$

Analogously we will denote by $Bel^{\downarrow M}$ and $Pl^{\downarrow M}$ marginal belief and plausibility measures on $\mathbf{X}_M$, respectively.

The following simple lemma concerning marginal beliefs and plausibilities will be used in the next section.

Lemma 1 *Let m be a basic assignment on $\mathbf{X}_N$, Bel and Pl corresponding beliefs and plausibilities and $K \subset N$. Then for any $A \subset \mathbf{X}_K$*

$$Bel^{\downarrow K}(A) = Bel(A^{\uparrow N}), \tag{7}$$
$$Pl^{\downarrow K}(A) = Pl(A^{\uparrow N}). \tag{8}$$

Proof. Using (1) and (6) one obtains

$$\begin{aligned} Bel^{\downarrow K}(A) &= \sum_{\substack{B\subseteq\mathbf{X}_K\\ B\subseteq A}} m^{\downarrow K}(B) \\ &= \sum_{\substack{B\subseteq\mathbf{X}_K\\ B\subseteq A}} \sum_{\substack{C\subseteq\mathbf{X}_N\\ C^{\downarrow K}=B}} m(C) \\ &= \sum_{\substack{C\subseteq\mathbf{X}_N\\ C^{\downarrow K}\subseteq A}} m(C) \\ &= \sum_{\substack{C\subseteq\mathbf{X}_N\\ C\subseteq A^{\uparrow N}}} m(C) \\ &= Bel(A^{\uparrow N}), \end{aligned}$$

where we used the fact that $C^{\downarrow K} \subseteq A$ if and only if $C \subseteq A^{\uparrow N}$ for any $C \subseteq \mathbf{X}_N$ and $A \subseteq \mathbf{X}_K$.

Similarly, using (2), (6) and the fact that $D^{\downarrow K} \subseteq \mathbf{X}_K, D^{\downarrow K} \cap B \neq \emptyset$ if and only if $D \subseteq \mathbf{X}_N, D \cap B^{\uparrow N} \neq \emptyset$

$$\begin{aligned} Pl^{\downarrow K}(B) &= \sum_{\substack{C\subseteq\mathbf{X}_K\\ C\cap B\neq\emptyset}} m^{\downarrow K}(C) \\ &= \sum_{\substack{C\subseteq\mathbf{X}_K\\ C\cap B\neq\emptyset}} \sum_{\substack{D\subseteq\mathbf{X}_N\\ D^{\downarrow K}=C}} m(D) \\ &= \sum_{\substack{D\subseteq\mathbf{X}_N\\ D\cap B^{\uparrow L}\neq\emptyset}} m(D) \\ &= Pl(B^{\uparrow N}), \end{aligned}$$

as desired. □

3 Conditioning

Conditioning belongs to the most important topics of any theory dealing with uncertainty. From the viewpoint of construction of Bayesian-network-like multidimensional models it seems to be inevitable.

3.1 Conditioning of Events

In evidence theory the "classical" conditioning rule is so-called *Dempster's rule of conditioning* defined for any $\emptyset \neq A \subseteq \mathbf{X}_N$ and $B \subseteq \mathbf{X}_N$ such that $Pl(B) > 0$ by the formula

$$m(A|B) = \frac{\sum_{C\subseteq\mathbf{X}_N : C\cap B = A} m(C)}{Pl(B)} \tag{9}$$

and $m(\emptyset|B) = 0$.

Let us note that formula (9) is special case of Dempster's rule of combination, when combining basic assignment m with another m_B such that $m_B(B) = 1$.

From this formula one can immediately obtain:

$$\begin{aligned} Bel(A|B) &= \frac{Bel(A \cup B^C) - Bel(B^C)}{1 - Bel(B^C)}, \\ Pl(A|B) &= \frac{Pl(A \cap B)}{Pl(B)}. \end{aligned} \tag{10}$$

This is not the only possibility how to make conditioning, another — in a way symmetric — conditioning rule is the following one called *focusing* defined for any $\emptyset \neq A \subseteq \mathbf{X}_N$ and $B \subseteq \mathbf{X}_N$ such that $Bel(B) > 0$ by the formula

$$m(A||B) = \begin{cases} \dfrac{m(A)}{Bel(B)} & \text{if} \quad A \subseteq B, \\ 0 & \text{otherwise.} \end{cases}$$

From the following two equalities one can see, in which sense are these two conditioning rules symmetric:

$$\begin{aligned} Bel(A||B) &= \frac{Bel(A \cap B)}{Bel(B)}, \\ Pl(A||B) &= \frac{Pl(A \cup B^C) - Pl(B^C)}{1 - Pl(B^C)}. \end{aligned} \tag{11}$$

These rules are based on different philosophy. Focusing assigns positive values only to those elements which are subsets of B, while Dempster's rule of conditioning to those which have nonempty intersection with it.

It is evident, that focusing is applicable in less cases than Dempster's rule, because of relation (3), hence from this point of view the latter seems to be more advantageous.

On the other hand, from the computational viewpoint the latter is more suitable, as it produces less focal elements (and in any of them a bigger "mass" is contained; cf. also Example 1). Due to this fact it may seem that focusing produces bigger intervals

than Dempster's rule (and it is very often true), but it is not generally satisfied, as can be seen again from Example 1.

Formulae (10) and (11) are, in a way, evidential counterparts of conditioning in probabilistic framework. Let us note that seemingly "natural" way of conditioning

$$m(A|_p B) = \frac{m(A \cap B)}{m(B)} \tag{12}$$

is not possible, since $m(A|_p B)$ need not be a basic assignment, as can be seen from the following simple example. It is caused by a simple fact that m, in contrary to Bel and Pl is not monotonous with respect to set inclusion.

Example 1 Let $\mathbf{X} = \{a, b, c\}$ and m on $\mathbf{X}$ be defined as follows:

$$\begin{aligned} m(\{a\}) = m(\{b\}) = m(\{c\}) &= \frac{1}{4}, \\ m(\{a, b\}) = m(\mathbf{X}) &= \frac{1}{8}. \end{aligned}$$

Using (12) one would obtain

$$m(\{a\}|_p\{a, b\}) = m(\{b\}|_p\{a, b\}) = 2,$$

which is out of the framework of evidence theory.

Let us use this example also for demonstrating the difference between Dempster's rule of conditioning and focusing. For this purpose let us compute

$$Bel(\{b, c\} = \frac{1}{2} \quad \text{and} \quad Pl(\{b, c\} = \frac{3}{4}.$$

Then we have

$$\begin{aligned} m(\{b\}|\{b, c\}) = \frac{m(\{b\}) + m(\{a, b\})}{Pl(\{b, c\})} &= \frac{1}{2}, \\ m(\{c\}|\{b, c\}) = \frac{m(\{c\})}{Pl(\{b, c\})} &= \frac{1}{3}, \\ m(\{b, c\}|\{b, c\}) = \frac{m(\mathbf{X})}{Pl(\{b, c\})} &= \frac{1}{6}, \end{aligned}$$

as $\{a, b\} \cap \{b, c\} = \{b\}$ and $\mathbf{X} \cap \{b, c\} = \{b, c\}$, while

$$\begin{aligned} m(\{b\}||\{b, c\}) = \frac{m(\{b\})}{Bel(\{b, c\})} &= \frac{1}{2}, \\ m(\{c\}||\{b, c\}) = \frac{m(\{c\})}{Bel(\{b, c\})} &= \frac{1}{2}, \end{aligned}$$

as $\{b\}$ and $\{c\}$ are the only subsets of $\{b, c\}$. ♢

Nevertheless, rather than in conditional beliefs and plausibilities of events we are interested in conditioning by variables. This problem will be in the center of our attention in the next subsection.

3.2 Conditional Variables

Definition 1 Let X_K and X_L $(K \cap L = \emptyset)$ be two groups of variables with values in $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively. Then the *conditional basic assignment according to Dempster's conditioning rule* of X_K given $X_L \in B \subseteq \mathbf{X}_L$ (for B such that $Pl(B) > 0$) is defined as follows:

$$m_{X_K|X_L}(A|B) = \frac{\sum_{C \subseteq \mathbf{X}_{K \cup L} : (C \cap B^{\uparrow K \cup L})^{\downarrow K} = A} m(C)}{Pl(B)} \tag{13}$$

for $A \neq \emptyset$ and $m_{K|L}(\emptyset|B) = 0$. Similarly, the *conditional basic assignment according to focusing* of X_K given $X_L \in B \subseteq \mathbf{X}_L$ (for B such that $Bel(B) > 0$) is defined by the equality

$$m_{X_K||X_L}(A||B) = \frac{\sum_{C \subseteq \mathbf{X}_{K \cup L} : C \subseteq B^{\uparrow K \cup L} \& C^{\downarrow K} = A} m(C)}{Bel(B)} \tag{14}$$

for any $A \neq \emptyset$ and $m_{K||L}(\emptyset||B) = 0$.

Now, let us prove that the definition is correct.

Theorem 1 *Set functions $m_{X_K|X_L}$ and $m_{X_K||X_L}$ defined for any fixed $B \subseteq \mathbf{X}_L$, such that $Pl(B) > 0$ and $Bel(B) > 0$, respectively, by Definition 1 are basic assignments on $\mathbf{X}_K$.*

Proof.

(i) Let $B \subseteq \mathbf{X}_L$ be such that $Pl(B) > 0$. As nonnegativity of $m_{X_K|X_L}(A|B)$ for any $A \subseteq \mathbf{X}_K$ and the fact that $m_{X_K|X_L}(\emptyset|B) = 0$ follow directly from the definition, to prove that $m_{X_K|X_L}$ is a basic assignment it is enough to show that

$$\sum_{A \subseteq \mathbf{X}_K} m_{X_K|X_L}(A|B) = 1.$$

To check it, let us sum the values of the numerators in (13)

$$\begin{aligned} \sum_{A \subseteq \mathbf{X}_K} \sum_{\substack{C \subseteq \mathbf{X}_{K \cup L} \\ (C \cap B^{\uparrow K \cup L})^{\downarrow K} = A}} m(C) &= \sum_{\substack{A \subseteq \mathbf{X}_K \\ A \neq \emptyset}} \sum_{\substack{C \subseteq \mathbf{X}_{K \cup L} \\ (C \cap B^{\uparrow K \cup L})^{\downarrow K} = A}} m(C) \\ &= \sum_{\substack{C \subseteq \mathbf{X}_{K \cup L} \\ C \cap B^{\uparrow L} \neq \emptyset}} m(C) \\ &= Pl(B^{\uparrow L}). \end{aligned}$$

To finish the proof it is enough to realize that $Pl(B^{\uparrow K\cup L}) = Pl^{\downarrow L}(B)$ for any $B \subseteq \mathbf{X}_L$ (by (8) of Lemma 1).

(ii) Analogously we will show that $m_{X||Y}$ is defined correctly. Let $B \subseteq \mathbf{X}_L$ be such that $Bel(B) > 0$. To prove that $m_{X||Y}$ is a basic assignment, it is again enough to check that

$$\sum_{A\subseteq\mathbf{X}_K} m_{X_K||X_L}(A||B) = 1.$$

To do so, let us compute

$$\begin{aligned}\sum_{A\subseteq\mathbf{X}_K} \sum_{\substack{C\subseteq\mathbf{X}_{K\cup L}:\\ C\subseteq B^{\uparrow K\cup L}:C^{\downarrow K}=A}} m(C) &= \sum_{\substack{C\subseteq\mathbf{X}_{K\cup L}:\\ C\subseteq B^{\uparrow L}A}} m(C)\\ &= Bel(B^{\uparrow L}).\end{aligned}$$

The rest of the proof, i.e. validity of $Bel(B^{\uparrow K\cup L}) = Bel^{\downarrow L}(B)$ follows directly from (7) of Lemma 1. □

4 Conditional Independence and Irrelevance

4.1 Conditional Independence and Irrelevance in Probability Theory

Independence and irrelevance need not be (and usually are not) distinguished in the probabilistic framework, as they are almost equivalent to each other.

Supposing X_K, X_L and X_M are groups of random variables with a joint probability distribution P we say that X_K is *conditionally independent* of X_L given X_M with respect to P if the equality

$$\begin{aligned}&P(x_K, x_L, x_M)\cdot P^{\downarrow M}(x_M)\\ &\quad = P^{\downarrow K\cup M}(x_K, x_M)\cdot P^{\downarrow L\cup M}(x_L, x_M)\end{aligned}$$

(where $P_{X_KX_M}, P_{X_LX_M}, P_{X_M}$ denote corresponding marginal distributions) holds for every value (x_K, x_L, x_M) of the variables X_K, X_L, X_M. It means that in every situation when the value of X_M is known the values of X_K and X_L are completely unrelated (from the stochastic point of view).

There exist several equivalent definitions of stochastic conditional independence, e.g.

$$P_{X_K|X_LX_M}(x_K|x_L, x_M) = P_{X_K|X_M}(x_K|x_M),$$

but this definition may be used only in the situation when $P^{\downarrow L\cup M}(x_L, x_M)$ is positive.

Similarly, in possibilistic framework adopting De Cooman's measure-theoretical approach [7] (particularly his notion of almost everywhere equality) we proved that analogous definitions are equivalent (for more details see [13]).

4.2 Independence

When constructing graphical models in any framework, (conditional) independence concept plays an important role. In evidence theory the most common notion of independence is that of random set independence [5].

It has already been proven [14] that it is also the only sensible one, as e.g. application of strong independence to two bodies of evidence may generally lead to a model which is beyond the framework of evidence theory. Epistemic independence and irrelevance were not taken into consideration, as none of them seem to be a suitable tool for factorization of multidimensional models. Furthermore, they require conditioning, so their application is also problematic from this point of view.

Definition 2 Let m be a basic assignment on $\mathbf{X}_N$ and $K, L \subset N$ be disjoint. We say that groups of variables X_K and X_L are *independent with respect to basic assignment* m (in notation $K \perp\!\!\!\perp L\ [m]$) if

$$m^{\downarrow K\cup L}(A) = m^{\downarrow K}(A^{\downarrow K})\cdot m^{\downarrow L}(A^{\downarrow L})$$

for all $A \subseteq \mathbf{X}_{K\cup L}$ for which $A = A^{\downarrow K} \times A^{\downarrow L}$, and $m(A) = 0$ otherwise.

This notion can be generalized in various ways [3, 12, 15]; the concept of conditional non-interactivity from [3], based on conjunction combination rule, is used for construction of directed evidential networks in [4]. In this paper we will use the concept introduced in [9, 15], as we consider it more suitable (the arguments can be found in [15]).

Definition 3 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

$$\begin{aligned}&m^{\downarrow K\cup L\cup M}(A)\cdot m^{\downarrow M}(A^{\downarrow M})\\ &\quad = m^{\downarrow K\cup M}(A^{\downarrow K\cup M})\cdot m^{\downarrow L\cup M}(A^{\downarrow L\cup M})\end{aligned}$$

holds for any $A \subseteq \mathbf{X}_{K\cup L\cup M}$ such that $A = A^{\downarrow K\cup M} \bowtie A^{\downarrow L\cup M}$, and $m(A) = 0$ otherwise.

It has been proven in [15] that this conditional independence concept satisfies so-called semi-graphoid

properties taken as reasonable to be valid for any conditional independence concept (see e.g. [10]) and it has been shown in which sense this conditional independence concept is superior to previously introduced ones [3, 12].

4.3 Irrelevance

Irrelevance is usually considered to be a weaker notion than independence (see e.g. [5]). It expresses the fact that a new piece of evidence concerning one variable cannot influence the evidence concerning the other variable, in other words is irrelevant to it. More formally: group of variables X_L is *irrelevant* to X_K ($K \cap L = \emptyset$) if for any $B \subseteq \mathbf{X}_L$ such that $Pl(B) > 0$

$$m_{X_K|X_L}(A|B) = m(A) \tag{15}$$

for any $A \subseteq \mathbf{X}_K$.[3]

It follows from the definition of irrelevance that it need not be a symmetric relation. Its symmetrized version is sometimes taken as a definition of independence. Let us note, that in the framework of evidence theory even in cases when the relation is symmetric, it does not imply independence, as can be seen from Examples 2 and 3.

Generalization of this notion to conditional irrelevance may be done as follows. Group of variables X_L is *conditionally irrelevant* to X_K given X_M (K, L, M disjoint, $K \neq \emptyset \neq L$) if for any $B \subseteq \mathbf{X}_L$ and $C \subseteq \mathbf{X}_M$ such that $Pl(B \times C) > 0$

$$m_{X_K|X_L X_M}(A|B \times C) = m_{X_K|X_M}(A|C) \tag{16}$$

for any $A \subseteq \mathbf{X}_K$.

Remark. This is not the only way of generalization of the irrelevance concept, e.g. we could allow for conditioning by general sets and not only by rectangles on the left side of (16), i.e. the equality

$$m_{X_K|X_L X_M}(A|B) = m_{X_K|X_M}(A|B^{\downarrow M}) \tag{17}$$

is satisfied for any $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_{L \cup M}$. This definition is evidently more general, but it seemingly has little sense, as the "interesting" sets from the viewpoint of (conditional) independence are rectangles, or, more generally, joins.

Let us note that the conditioning in equalities (15) and (16) stands for an abstract conditioning rule (any of those mentioned in the previous section or some other [6]). Nevertheless, the validity of (15) and (16) may depend on the choice of conditioning rule. To demonstrate it let us present two simple examples.

[3] Let us note that somewhat weaker definition of irrelevance one can found in [2], where equality is substituted by proportionality. This notion has been later generalized using conjunctive combination rule [3].

Example 2 Let X_1 and X_2 be two binary variables (with values in $\mathbf{X}_i = \{a_i, \bar{a}_i\}$) with joint basic assignment m defined as follows:

$$\begin{aligned} m(\{(a_1, a_2)\}) &= \frac{1}{2}, \\ m(\mathbf{X}_1 \times \mathbf{X}_2 \setminus \{(\bar{a}_1, \bar{a}_2)\}) &= \frac{1}{4}, \\ m(\mathbf{X}_1 \times \mathbf{X}_2) &= \frac{1}{4}. \end{aligned}$$

From these values one can obtain

$$m^{\downarrow 2}(\{a_2\}) = m^{\downarrow 2}(\mathbf{X}_2) = \frac{1}{2},$$

and therefore

$$Bel^{\downarrow 2}(\{a_2\}) = \frac{1}{2}, \qquad Bel^{\downarrow 2}(\{\bar{a}_2\}) = 0.$$
$$Pl^{\downarrow 2}(\{a_2\}) = 1, \qquad Pl^{\downarrow 2}(\{\bar{a}_2\}) = \frac{1}{2}.$$

Computing conditional basic assignments (according to Dempster's conditioning rule) one can easily see that

$$\begin{aligned} m_{X_1|X_2}(\{a_1\}|\{a_2\}) &= m_{X_1|X_2}(\{a_1\}|\{\bar{a}_2\}) \\ &= \frac{1}{2} = m^{\downarrow 1}(\{a_1\}), \\ m_{X_1|X_2}(\{\bar{a}_1\}|\{a_2\}) &= m_{X_1|X_2}(\{\bar{a}_1\}|\{\bar{a}_2\}) \\ &= 0 = m^{\downarrow 1}(\{\bar{a}_1\}), \\ m_{X_1|X_2}(\mathbf{X}_1|\{a_2\}) &= m_{X_1|X_2}(\mathbf{X}_1|\{\bar{a}_2\}) \\ &= \frac{1}{2} = m^{\downarrow 1}(\mathbf{X}_1), \end{aligned}$$

i.e. X_1 and X_2 are irrelevant (with respect to Dempster's conditioning rule). On the other hand, as e.g.

$$\begin{aligned} &m_{X_1||X_2}((\{a_1\}||(\{a_2\})) \\ &\quad = \frac{m(\{(a_1, a_2)\}))}{Bel(\{a_2\})} = 1 \neq \frac{1}{2} = m^{\downarrow 1}(\{a_1\}), \end{aligned}$$

they are not irrelevant with respect to focusing. ♢

Example 3 Let X_1 and X_2 be two binary variables (with values in $\mathbf{X}_i = \{a_i, \bar{a}_i\}$) with joint basic assignment m defined as follows:

$$\begin{aligned} m(\{(a_1, a_2)\}) &= \frac{1}{4}, \\ m(\{a_1\} \times \mathbf{X}_2) &= \frac{1}{4}, \\ m(\mathbf{X}_1 \times \{a_2\}) &= \frac{1}{4}, \\ m(\mathbf{X}_1 \times \mathbf{X}_2 \setminus \{(\bar{a}_1, \bar{a}_2)\}) &= \frac{1}{4}. \end{aligned}$$

From these values one can obtain

$$m^{\downarrow 2}(\{a_2\}) = m^{\downarrow 2}(\mathbf{X}_2) = \frac{1}{2},$$

and therefore

$$Bel^{\downarrow 2}(\{a_2\}) = \frac{1}{2}, \qquad Bel^{\downarrow 2}(\{\bar{a}_2\}) = 0,$$
$$Pl^{\downarrow 2}(\{a_2\}) = 1, \qquad Pl^{\downarrow 2}(\{\bar{a}_2\}) = \frac{1}{2}.$$

Evidently, it is not possible to condition by $\{\bar{a}_2\}$ and we have to confine ourselves to conditioning by $\{a_2\}$:

$$\begin{aligned} m_{X_1||X_2}(\{a_1\}||\{a_2\}) &= \tfrac{1}{2} = m^{\downarrow 1}(\{a_1\}), \\ m_{X_1||X_2}(\{\bar{a}_1\}||\{a_2\}) &= 0 = m^{\downarrow 1}(\{\bar{a}_1\}), \\ m_{X_1||X_2}(\mathbf{X}_1||\{a_2\}) &= \tfrac{1}{2} = m^{\downarrow 1}(\mathbf{X}_1), \end{aligned}$$

i.e. X_1 and X_2 are irrelevant (under focusing). On the other hand, as e.g.

$$\begin{aligned} & m_{X_1|X_2}((\{a_1\}|(\{\bar{a}_2\})) \\ &= \frac{m(\{a_1\} \times \mathbf{X}_2) + m(\mathbf{X}_1 \times \mathbf{X}_2 \setminus \{(\bar{a}_1, \bar{a}_2)\})}{Pl(\{a_2\})} \\ &= 1 \neq \frac{1}{2} = m^{\downarrow 1}(\{a_1\}), \end{aligned}$$

they are not irrelevant with respect to Dempster's conditioning rule. ♢

5 Relationship Between Independence and Irrelevance

As we demonstrated at the end of preceding section, different conditioning rules lead to different irrelevance concepts. Therefore we will study the relationships between independence and irrelevance separately for Dempster's conditioning rule and for focusing.

5.1 Dempster's rule of conditioning

For (unconditional) independence and irrelevance the following assertion holds true.

Theorem 2 *Let X_K and X_L ($K \cup L = \emptyset$) be independent groups of variables (under joint basic assignment m defined on $\mathbf{X}_{K\cup L}$). Then X_L are irrelevant to X_K with respect to Dempster's conditioning rule.*

Proof. Let X_K and X_L be independent. Then

$$m(A) = m^{\downarrow K}(A^{\downarrow K}) \cdot m^{\downarrow L}(A^{\downarrow L})$$

for any $A \subseteq \mathbf{X}_{K\cap L}$ for which $A = A^{\downarrow K} \times A^{\downarrow L}$, and $m(A) = 0$ otherwise, i.e. the only focal elements of m are rectangles. Therefore we have for arbitrary $A \subseteq \mathbf{X}_{K\cup L}$

$$\begin{aligned} Pl(A) &= \sum_{C:C\cap A\neq\emptyset} m(C) \\ &= \sum_{C:C\cap A\neq\emptyset} m^{\downarrow K}(C^{\downarrow K}) \cdot m^{\downarrow L}(C^{\downarrow L}) \\ &= \sum_{D:D\cap A^{\downarrow K}\neq\emptyset} m^{\downarrow K}(D) \cdot \sum_{E:E\cap A^{\downarrow L}\neq\emptyset} m^{\downarrow L}(E) \\ &= Pl^{\downarrow K}(A^{\downarrow K}) \cdot Pl^{\downarrow L}(A^{\downarrow L}). \end{aligned}$$

From this equality we immediately obtain that for all A such that $Pl^{\downarrow L}(A^{\downarrow L}) > 0$ equality

$$\frac{Pl(A)}{Pl^{\downarrow L}(A^{\downarrow L})} = Pl^{\downarrow K}(A^{\downarrow K})$$

is satisfied. But the left side of this equality is equal to $Pl_{X_K|X_L}(A^{\downarrow K}|A^{\downarrow L})$. As both conditional and marginal basic assignments can be obtained from corresponding plausibilities using the equality (4) and Möbius inversion (5), we immediately obtain that also for any fixed $B \subseteq \mathbf{X}_L$ such that $Pl^{\downarrow L}(B) > 0$

$$m_{K|L}(A|B) = m^{\downarrow K}(A)$$

for any $A \subseteq \mathbf{X}_K$, i.e. X_K and X_L are irrelevant. □

The reverse implication does not hold in general. To demonstrate it let us recall Example 2.

Example 2 *(Continued)* We have already shown that X_1 and X_2 are irrelevant (with respect to Dempster's conditioning rule). But they are not independent, as the focal elements are not rectangles, which contradicts Definition 2. ♢

Unfortunately, a generalization of Theorem 2 to conditional independence and conditional irrelevance does not hold, as can be seen from the following simple example.

Example 4 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:

$$m(\{(x_1, x_2, x_3)\}) = \tfrac{1}{16},$$
$$m(\mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3) = \tfrac{1}{2},$$

for $x_i = a_i, \bar{a}_i$, values of m on the remaining sets being 0, i.e. we have 9 focal elements — 8 singletons and the whole frame of discernment. Its marginal basic assignments on $\mathbf{X}_1 \times \mathbf{X}_3, \mathbf{X}_2 \times \mathbf{X}_3$ and $\mathbf{X}_3$ are

$$m^{\downarrow 13}(\{(x_1, x_3)\}) = \tfrac{1}{8},$$
$$m^{\downarrow 13}(\mathbf{X}_1 \times \mathbf{X}_3) = \tfrac{1}{2},$$
$$m^{\downarrow 23}(\{(x_2, x_3)\}) = \tfrac{1}{8},$$
$$m^{\downarrow 23}(\mathbf{X}_2 \times \mathbf{X}_3) = \tfrac{1}{2},$$

and

$$m^{\downarrow 3}(\{x_3\}) = \tfrac{1}{4},$$
$$m^{\downarrow 3}(\mathbf{X}_3) = \tfrac{1}{2},$$

respectively (values of m of remaining subsets being 0, again). It is easy (but somewhat time-consuming) to show that for any $A \subseteq \mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3$ such that $A = A^{\downarrow 13} \bowtie A^{\downarrow 23}$

$$m(A) \cdot m^{\downarrow 3}(A^{\downarrow 3}) = m^{\downarrow 13}(A^{\downarrow 13}) \cdot m^{\downarrow 23}(A^{\downarrow 23}),$$

the values of remaining sets being zero, i.e. $\{1\} \perp\!\!\!\perp \{2\}|\{3\}$ $[m]$ holds.

Now, let us show, that X_2 is not irrelevant to X_1 given X_3. To do so, we have to compute $m_{X_1|X_2X_3}$ and $m_{X_1|X_3}$. First, let us take into account that

$$Pl(\{(x_2, x_3)\}) = \frac{5}{8}$$

for any $x_i = a_i, \bar{a}_i, i = 2, 3$ and

$$Pl(\{x_3\}) = \frac{3}{4}$$

for both $x_3 = a_3, \bar{a}_3$ and that

$$(\mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3 \cap \{(a_2, a_3)\}^{\uparrow 123})^{\downarrow 1} = \mathbf{X}_1$$

and similarly

$$(\mathbf{X}_1 \times \mathbf{X}_3 \cap \{a_3\}^{\uparrow 13})^{\downarrow 1} = \mathbf{X}_1.$$

Then we have

$$m_{X_1|X_2X_3}(\{a_1\}|\{(a_2, a_3)\}) = \frac{m(\{(a_1, a_2, a_3)\})}{Pl(\{(a_2, a_3)\})} = \frac{1}{10},$$

$$m_{X_1|X_2X_3}(\{\bar{a}_1\}|\{(a_2, a_3)\}) = \frac{m(\{(\bar{a}_1, a_2, a_3)\})}{Pl(\{(a_2, a_3)\})} = \frac{1}{10},$$

$$m_{X_1|X_2X_3}(\mathbf{X}_1|\{(a_2, a_3)\}) = \frac{m(\mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3)}{Pl(\{(a_2, a_3)\})} = \frac{4}{5},$$

while

$$m_{X_1|X_3}(\{a_1\}|\{a_3\}) = \frac{m(\{(a_1, a_3)\})}{Pl(\{a_3\})} = \frac{1}{6},$$

$$m_{X_1|X_3}(\{\bar{a}_1\}|\{a_3\}) = \frac{m(\{(\bar{a}_1, a_3)\})}{Pl(\{a_3\})} = \frac{1}{6},$$

$$m_{X_1|X_3}(\mathbf{X}_1|\{a_3\}) = \frac{m(\mathbf{X}_1 \times \mathbf{X}_3)}{Pl(\{a_3\})} = \frac{2}{3},$$

i.e. $m_{X_1|X_2X_3} \neq m_{X_1|X_3}$. ♢

5.2 Focusing

In this subsection we will investigate mutual relationship between (conditional) independence and irrelevance based on the latter conditioning rule introduced in Section 3.

Theorem 3 *Let X_K and X_L ($K \cap L = \emptyset$) be independent groups of variables (under joint basic assignment m on $\mathbf{X}_{K \cup L}$). Then X_K and X_L are irrelevant with respect to focusing.*

Proof. Let X_K and X_L be independent. Then

$$m(A) = m^{\downarrow K}(A^{\downarrow K}) \cdot m^{\downarrow L}(A^{\downarrow L})$$

for any $A \subseteq \mathbf{X}_{K \cup L}$ for which $A = A^{\downarrow K} \times A^{\downarrow L}$, and $m(A) = 0$ otherwise, i.e. the only focal elements of m are rectangles. Therefore we have for arbitrary $A \subseteq \mathbf{X}_K$

$$\begin{aligned} Bel(A) &= \sum_{C \subseteq A} m(C) \\ &= \sum_{C \subseteq A} m^{\downarrow K}(C^{\downarrow K}) \cdot m^{\downarrow L}(C^{\downarrow L}) \\ &= \sum_{D: D \subseteq A^{\downarrow K}} m^{\downarrow K}(D) \cdot \sum_{E: E \subseteq A^{\downarrow L}} m^{\downarrow L}(E) \\ &= Bel^{\downarrow K}(A^{\downarrow K}) \cdot Bel^{\downarrow L}(A^{\downarrow L}). \end{aligned}$$

From this equality we immediately obtain that for all A such that $Bel^{\downarrow L}(A^{\downarrow L}) > 0$ equality

$$\frac{Bel(A)}{Bel^{\downarrow L}(A^{\downarrow L})} = Bel^{\downarrow K}(A^{\downarrow K})$$

is satisfied. But the left side of this equality is equal to $Bel_{X_K||X_L}(A^{\downarrow K}||A^{\downarrow L})$. As the both conditional and marginal basic assignments can be obtained from corresponding beliefs using Möbius inversion (5) we immediately obtain that also for any fixed $B \subseteq \mathbf{X}_L$ such that $Bel(B) > 0$

$$m_{X_K||X_L}(A||B) = m^{\downarrow K}(A)$$

for any $A \subseteq \mathbf{X}_K$, i.e. X_K and X_L are irrelevant. □

The reverse implication does not hold again, as can be seen from the following simple example (continuation of Example 3).

Example 3 *(Continued)* We have already proven that X_1 and X_2 are irrelevant (under focusing). But they are not independent, as the focal elements are not rectangles, which again contradicts Definition 2.♢

Up to now the results presented in this subsection have been exactly the same as in the preceding one.

Now, let us study the problem of the relationship between conditional independence and irrelevance. For this purpose, let us recall Example 4.

Example 4 *(Continued)* We have already shown that although X_1 and X_2 are conditionally independent given X_3, X_2 is not irrelevant to X_1 given X_3 under Dempster's rule of conditioning.

Now, let us us check whether X_2 is irrelevant to X_1 given X_3 under focusing. To do so, we have to compute $m_{X_1||X_2X_3}$ and $m_{X_1||X_3}$. Again, we have to take into account that

$$Bel(\{(x_2, x_3)\}) = \frac{1}{8}$$

for any $x_i = a_i, \bar{a}_i, i = 2, 3$ and

$$Bel(\{x_3\}) = \frac{1}{4}$$

for both $x_3 = a_3, \bar{a}_3$ and the fact that there does not exist any focal element A of m such that $A \subseteq \{(x_2, x_3)\}^{\uparrow 123}$ (for any pair (x_2, x_3)) and $A^{\downarrow 1} = \mathbf{X}_1$ and similarly there does not exist any focal element B of $m^{\downarrow 13}$ such that $B \subseteq \{x_3\}^{\uparrow 13}$ (for any x_3) and $B^{\downarrow 1} = \mathbf{X}_1$. Therefore we have

$$\begin{aligned} &m_{X_1||X_2X_3}(\{a_1\}||\{(x_2, x_3)\}) \\ &\qquad = \frac{m(\{(a_1, x_2, x_3)\})}{Bel(\{(x_2, x_3)\})} = \frac{1}{2}, \\ &m_{X_1||X_2X_3}(\{\bar{a}_1\}||\{(x_2, x_3)\}) \\ &\qquad = \frac{m(\{(\bar{a}_1, x_2, x_3)\})}{Bel(\{(x_2, x_3)\})} = \frac{1}{2}, \\ &m_{X_1||X_2X_3}(\mathbf{X}_1||\{(x_2, x_3)\}) = 0, \end{aligned}$$

for any pair $(x_2, x_3) \in \mathbf{X}_2 \times \mathbf{X}_3$ and

$$\begin{aligned} m_{X_1||X_3}(\{a_1\}||\{x_3\}) &= \frac{m(\{(a_1, x_3)\})}{Bel(\{x_3\})} = \frac{1}{2}, \\ m_{X_1||X_3}(\{\bar{a_1}\}||\{x_3\}) &= \frac{m(\{(\bar{a_1}, x_3)\})}{Bel(\{x_3\})} = \frac{1}{2}, \\ m_{X_1||X_3}(\mathbf{X}_1||x_3) &= 0, \end{aligned}$$

for any $x_3 \in \mathbf{X}_3$, i.e. $m_{X_1||X_2X_3} = m_{X_1||X_3}$ when conditioning by singletons, which is quite different from the previous case, based on Dempster's conditioning rule.

Nevertheless, to demonstrate that X_2 is irrelevant to X_1 given X_3 we have also to check the validity of equality (16) for a general rectangle $B \times C$ such that $Bel(B \times C) > 0$. As both X_2 and X_3 are binary, only three situations may happen:

- $B = \mathbf{X}_2$ **and** $C = \mathbf{X}_3$**:** in this case equality (16) is trivially satisfied, as conditional basic assignments on both sides are, in fact, marginal basic assignments on X_1, and therefore identical;
- $B = \mathbf{X}_2$ **and** $C = \{x_3\}$ **for** $x_3 = a_3, \bar{a}_3$**:** in this case equality (16) is again satisfied, as conditional basic assignment on the left side is, in fact, the same as that on the right side;
- $B = \{x_2\}$ **for** $x_2 = a_2, \bar{a}_2$ **and** $C = \mathbf{X}_3$**:** this is the nontrivial case, corresponding to unconditional irrelevance (15); nevertheless, its validity need not be checked, since X_1 and X_2 are not (unconditionally) independent, as can be easily checked.

Therefore X_2 is irrelevant to X_1 given X_3 (under focusing). ♢

Let us finish the section with a partial generalization of Theorem 3, which, maybe surprisingly, proves that conditioning by sets which are not rectangles is sensible.

Theorem 4 *Let X_K and X_L be conditionally independent groups of variables given X_M under joint basic assignment m on $\mathbf{X}_{K \cup L \cup M}$ (K, L, M disjoint, $K \neq \emptyset \neq L$). Then*

$$m_{X_K||X_LX_M}(A||B) = m_{X_K||X_M}(A||B^{\downarrow M}) \qquad (18)$$

for any $m^{\downarrow L \cup M}$-atom $B \subseteq \mathbf{X}_{L \cup M}$ such that $B^{\downarrow M}$ is $m^{\downarrow M}$-atom and $A \subseteq \mathbf{X}_K$.

Proof. Let X_K and X_L be conditionally independent given X_M. Then

$$\begin{aligned} &m(C) \cdot m^{\downarrow M}(C^{\downarrow M}) \\ &\quad = m^{\downarrow K \cup M}(C^{\downarrow K \cup M}) \cdot m^{\downarrow L \cup M}(C^{\downarrow L \cup M}) \end{aligned}$$

holds for any $C \subseteq \mathbf{X}_{K \cup L \cup M}$ such that $C = C^{\downarrow K \cup M} \bowtie C^{\downarrow L \cup M}$, and $m(C) = 0$ otherwise. From this equality we immediately obtain that for all C such that $m^{\downarrow L}(C^{\downarrow L}) > 0$ equality

$$\frac{m(C)}{m^{\downarrow L \cup M}(C^{\downarrow L \cup M})} = \frac{m^{\downarrow K \cup M}(C^{\downarrow K \cup M})}{m^{\downarrow M}(C^{\downarrow M})}$$

is satisfied. If $C^{\downarrow L \cup M}$ is an atom, then $m^{\downarrow L \cup M}(C^{\downarrow L \cup M}) = Bel^{\downarrow L \cup M}(C^{\downarrow L \cup M})$ (and analogously $m^{\downarrow M}(C^{\downarrow M}) = Bel^{\downarrow M}(C^{\downarrow M})$ if $C^{\downarrow M}$ is an atom) and this equality may be rewritten into the form

$$\frac{m(C)}{Bel^{\downarrow L \cup M}(C^{\downarrow L \cup M})} = \frac{m^{\downarrow K \cup M}(C^{\downarrow K \cup M})}{Bel^{\downarrow M}(C^{\downarrow M})}.$$

If we denote $C^{\downarrow L \cup M}$ by B, we obtain

$$m_{X_K||X_LX_M}(C^{\downarrow K}||B) = m_{X_K||X_M}(C^{\downarrow K}||B^{\downarrow M}).$$

If $A \neq C^{\downarrow K}$, then $m(A^{\uparrow K \cup L \cup M} \cap B^{\uparrow K \cup L \cup M}) = 0$ and therefore equality (18) is trivially satisfied. □

From this theorem it is evident, that conditions under which conditional independence implies conditional irrelevance are rather restrictive.

The requirement in Theorem 4 for B being an atom is substantial, as can be seen from the following simple example (again continuation of Example 4).

Example 4 *(Continued)* Let us consider a set $B = \{(a_2, a_3), (\bar{a}_2, \bar{a}_3)\} \subseteq \mathbf{X}_2 \times \mathbf{X}_3$. One can easily compute that $Bel(B) = \frac{1}{4}$ and therefore

$$\begin{aligned} &m_{X_1|X_2X_3}(\{a_1\}|B) \\ &\quad = \frac{m(\{a_1, a_2, a_3\}) + m(\{a_1, \bar{a}_2, \bar{a}_3\})}{Bel(B)} = \frac{1}{2}, \end{aligned}$$

$$m_{X_1|X_2X_3}(\{\bar{a}_1\}|B) = \frac{m(\{\bar{a}_1, a_2, a_3\}) + m(\{\bar{a}_1, \bar{a}_2, \bar{a}_3\})}{Bel(B)} = \frac{1}{2},$$
$$m_{X_1|X_2X_3}(\mathbf{X}_1|B) = 0,$$

while,

$$m_{X_1|X_3}(\{a_1\}|B^{\downarrow 3}) = m^{\downarrow 1}(\{a_1\}) = \frac{1}{4},$$
$$m_{X_1|X_3}(\{\bar{a}_1\}|B^{\downarrow 3}) = m^{\downarrow 1}(\{\bar{a}_1\}) = \frac{1}{4},$$
$$m_{X_1|X_2X_3}(\mathbf{X}_1|B^{\downarrow 3}) = m^{\downarrow 1}\ (\mathbf{X}_1) = \frac{1}{2}.$$

as $B^{\downarrow 3} = \mathbf{X}_3$. ♢

6 Conclusions

We presented two conditional rules for basic assignment and studied the relationship between (conditional) independence and (conditional) irrelevance (based on these conditioning rules) in evidence theory.

While in unconditional case independence implies irrelevance and not vice versa (as expected), for conditional independence such an implication does not hold, in general. Therefore, it is necessary to be cautious when constructing Bayesian-network-like models in evidence theory, as the mutual relationship is more complicated than in probabilistic framework.

It may be of some interest to study another way of conditioning presented in [8], however, its application will be more complicated, as conditional basic assignment must be obtained via Möbius transform from conditional beliefs. Furthermore, we are somewhat sceptic about the result.

Acknowledgements

The work of the author was supported by the grant GA ČR 201/09/1891. The author would like to express her gratitude to both referees for their inspiring comments.

References

[1] C. Beeri, R. Fagin, D. Maier, M. Yannakakis, On the desirability of acyclic database schemes, *J. of the Association for Computing Machinery,* **30** (1983), 479–513.

[2] B. Ben Yaghlane, Ph. Smets and K. Mellouli, Belief functions independence: I. the marginal case. *Int. J. Approx. Reasoning*, **29** (2002), 47–70.

[3] B. Ben Yaghlane, Ph. Smets and K. Mellouli, Belief functions independence: II. the conditional case. *Int. J. Approx. Reasoning*, **31** (2002), 31-75.

[4] B. Ben Yaghlane, Ph. Smets and K. Mellouli, Directed evidential networks with conditional belief functions. *Proceedings of ECSQARU 2003,* eds. T. D. Nielsen, N. L. Zhang, 291–305.

[5] 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.

[6] M. Daniel, Belief conditioning Rules for classic belief functions, *Proceedings of WUPES'09,* eds. T. Kroupa, J. Vejnarová, J., 46–56.

[7] G. de Cooman, Possibility theory I – III. *Int. J. General Systems* **25** (1997), pp. 291–371.

[8] R. Fagin and J. Y. Halpern, A new approach to updating beliefs, *Uncertainty in artifiacial intelligence,* eds. Bonissone et al., vol. VI, pp. 347–374, Elsevier, 1991.

[9] R. Jiroušek and J. Vejnarová, Compositional models and conditional independence in Evidence Theory, *Int. J. Approx. Reasoning,* **52** (2011), 316-334.

[10] S. L. Lauritzen, *Graphical Models.* Oxford University Press, 1996.

[11] G. Shafer, *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, New Jersey, 1976.

[12] P. P. Shenoy, Conditional independence in valuation-based systems. *Int. J. Approx. Reasoning,* **10** (1994), 203-234.

[13] J. Vejnarová, Conditional independence relations in possibility theory. *Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 8, pp. 253–269, 2000.

[14] 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.

[15] J. Vejnarová, On conditional independence in evidence theory, *Proceedings of ISIPTA'09,* eds. T. Augustin, F. P. A. Coolen, S. Moral and M. C. M. Troffaes,Durham, UK, 2009, pp. 431–440.

[16] J. Vejnarová, A thorough comparison of two conditional independence concepts for belief functions, *Proceedings of Workshop on the Theory of Belief Functions,* Brest, France, 2010.

On Prior-Data Conflict in Predictive Bernoulli Inferences

Gero Walter, Thomas Augustin
Department of Statistics
Ludwig-Maximilians-Universität München (LMU)
{gero.walter; thomas}@stat.uni-muenchen.de

Frank P.A. Coolen
Department of Mathematics
Durham University
frank.coolen@durham.ac.uk

Abstract

By its capability to deal with the multidimensional nature of uncertainty, imprecise probability provides a powerful methodology to sensibly handle prior-data conflict in Bayesian inference. When there is strong conflict between sample observations and prior knowledge the posterior model should be more imprecise than in the situation of mutual agreement or compatibility. Focusing presentation on the prototypical example of Bernoulli trials, we discuss the ability of different approaches to deal with prior-data conflict.

We study a generalized Bayesian setting, including Walley's Imprecise Beta-Binomial model and his extension to handle prior data conflict (called pdc-IBBM here). We investigate alternative shapes of prior parameter sets, chosen in a way that shows improved behaviour in the case of prior-data conflict and their influence on the posterior predictive distribution. Thereafter we present a new approach, consisting of an imprecise weighting of two originally separate inferences, one of which is based on an informative imprecise prior whereas the other one is based on an uninformative imprecise prior. This approach deals with prior-data conflict in a fascinating way.

Keywords. Bayesian inference; generalized iLUCK-models; imprecise Beta-Binomial model; imprecise weighting; predictive inference; prior-data conflict.

1 Introduction

Imprecise probability has shown to be a powerful methodology to cope with the multidimensional nature of uncertainty [8, 2]. Imprecision allows the quality of information, on which probability statements are based, to be modeled. Well supported knowledge is expressed by comparatively precise models, while highly imprecise (or even vacuous) models reflect scarce (or no) knowledge on probabilities. This flexible, multidimensional perspective on uncertainty modeling has intensively been utilized in generalized Bayesian inference to overcome the criticism of the arbitrariness of the choice of single prior distributions in traditional Bayesian inference. In addition, only imprecise probability models react reliably to the presence of prior-data conflict, i.e. situations where "the prior [places] its mass primarily on distributions in the sampling model for which the observed data is surprising" [9, p. 894]. Lower and upper probabilities allow a specific reaction to prior-data conflict and offer reasonable inferences if the analyst wishes to stick to his prior assumptions: starting with the same level of ambiguity in the prior specification, wide posterior intervals can reflect conflict between prior and data, while no prior-data conflict will lead to narrow intervals. Ideally the model could provide an extra 'bonus' of precision if prior assumptions are very strongly supported by the data. Such a model would have the advantage of (relatively) precise answers when the data confirm prior assumptions, while still rendering more cautionary answers in the case of prior-data conflict, thus leading to cautious inferences if, and only if, caution is needed.

Although Walley [18, p. 6] explicitly emphasizes this possibility to express prior-data conflict as one of the main motivations for imprecise probability, it has received surprisingly little attention. Rare exceptions include two short sections in [18, p. 6 and Ch. 5.4] and [14, 7, 23]. The popular IDM [19, 3] and its generalization to exponential families [15] do not reflect prior-data conflict. [21] used the basic ideas of [18, Ch. 5.4] to extend the approach of [15] to models that show sensitivity to prior-data conflict.

In this paper a deeper investigation of the issue of prior-data conflict is undertaken, focusing on the prototypic special case of predictive inference in Bernoulli trials: We are interested in the posterior predictive probability for the event that a future Bernoulli random quantity will have the value 1, also called a 'success'. This event is not explicitly included in the nota-

tion, i.e. we simply denote its lower and upper probabilities by $\underline{P}$ and $\overline{P}$, respectively. This future Bernoulli random quantity is assumed to be exchangeable with the Bernoulli random quantities whose observations are summarized in the data, consisting of the number n of observations and the number s of these that are successes. In our analysis of this model, we will often consider s as a a real-valued observation in $[0, n]$, keeping in mind that in reality it can only take on integer values, but the continuous representation is convenient for our discussions, in particular in our predictive probability plots (PPP), where for given n, $\underline{P}$ and $\overline{P}$ are discussed as functions of s.

Section 2.1 describes a general framework for generalized Bayesian inference in this setting. The method presented in [18, Ch. 5.4.3], called 'pdc-IBBM' in this paper, is considered in detail in Section 2.2 and we show that its reaction to prior-data conflict can be improved by suitable modifications of the underlying imprecise priors. A basic proposal along these lines is discussed in Section 2.3 with further alternatives sketched in Section 2.4. Section 3 addresses the problem of prior-data conflict from a completely different angle. There we combine two originally separate inferences, one based on an informative imprecise prior and one on an uninformative imprecise prior, by an imprecise weighting scheme. The paper concludes with a brief comparison of the different approaches.

2 Imprecise Beta-Binomial Models

2.1 The Framework

The traditional Bayesian approach for our basic problem is the Beta-Binominal model, which expresses prior beliefs about the probability p of observing a 'success' by a Beta distribution. With[1] $f(p) \propto p^{n^{(0)}y^{(0)}-1}(1-p)^{n^{(0)}(1-y^{(0)})-1}$, $y^{(0)} = \mathrm{E}[p]$ can be interpreted as prior guess of p, while $n^{(0)}$ governs the concentration of probability mass around $y^{(0)}$, also known as 'pseudo counts' or 'prior strength'.[2] These denominations are due to the role of $n^{(0)}$ in the update step: With s successes in n draws observed, the posterior parameters are[3]

$$n^{(n)} = n^{(0)} + n, \qquad y^{(n)} = \frac{n^{(0)}y^{(0)} + s}{n^{(0)} + n}. \tag{1}$$

Thus $y^{(n)}$ is a weighted average of the prior parameter $y^{(0)}$ and the sample proportion s/n, and potential prior data conflict is simply averaged out.

[1] Our notation relates to [18]'s as $n^{(0)} \leftrightarrow s_0$, $y^{(0)} \leftrightarrow t_0$.

[2] $^{(0)}$ denotes prior parameters; $^{(n)}$ posterior parameters.

[3] The model is prototypic for conjugate Bayesian analysis in canonical exponential families, for which updating of the parameters $n^{(0)}$ and $y^{(0)}$ can be written as (1).

Overcoming the dogma of precision, formulating generalized Bayes updating in this setting is straightforward. By Walley's Generalized Bayes Rule [18, Ch. 6] the imprecise prior $\mathcal{M}^{(0)}$, described by convex sets of precise prior distributions, is updated to the imprecise posterior $\mathcal{M}^{(n)}$ obtained by updating $\mathcal{M}^{(0)}$ elementwise. In particular, the convenient conjugate analysis used above can be extended: One specifies a prior parameter set $\Pi^{(0)}$ of $(n^{(0)}, y^{(0)})$ values and takes as imprecise prior the set $\mathcal{M}^{(0)}$ consisting of all convex mixtures of Beta priors with $(n^{(0)}, y^{(0)}) \in \Pi^{(0)}$. In this sense, the set of Beta priors corresponding to $\Pi^{(0)}$ gives the set of extreme points for the actual convex set of priors $\mathcal{M}^{(0)}$. Updating $\mathcal{M}^{(0)}$ with the Generalized Bayes' Rule results in the convex set $\mathcal{M}^{(n)}$ of posterior distributions that conveniently can be obtained by taking the convex hull of the set of Beta posteriors, which in turn are defined by the set of updated parameters $\Pi^{(n)} = \{(n^{(n)}, y^{(n)}) \mid (n^{(0)}, y^{(0)}) \in \Pi^{(0)}\}$. This relationship between the sets $\Pi^{(0)}$ and $\Pi^{(n)}$ and the sets $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(n)}$ will allow us to discuss different models $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(n)}$ by depicting the corresponding parameter sets $\Pi^{(0)}$ and $\Pi^{(n)}$. When interpreting our results, care will be needed with respect to convexity. Although $\mathcal{M}^{(0)}$ and $\mathcal{M}^{(n)}$ are convex, the parameter sets $\Pi^{(0)}$ and $\Pi^{(n)}$ generating them need not necessarily be so. Indeed, convexity of the parameter set is not necessarily preserved in the update step: Convexity of $\Pi^{(0)}$ does not imply convexity of $\Pi^{(n)}$.

Throughout, we are interested in the posterior predictive probability $[\underline{P}, \overline{P}]$ for the event that a future draw is a success. In the Beta-Bernoulli model, this probability is equal to $y^{(n)}$, and we get[4]

$$\underline{P} = \underline{y}^{(n)} := \min_{\Pi^{(n)}} y^{(n)} = \min_{\Pi^{(0)}} \frac{n^{(0)}y^{(0)} + s}{n^{(0)} + n}, \tag{2}$$

$$\overline{P} = \overline{y}^{(n)} := \max_{\Pi^{(n)}} y^{(n)} = \max_{\Pi^{(0)}} \frac{n^{(0)}y^{(0)} + s}{n^{(0)} + n}. \tag{3}$$

2.2 Walley's pdc-IBBM

Special imprecise probability models are now obtained by specific choices of $\Pi^{(0)}$. If one fixes $n^{(0)}$ and varies $y^{(0)}$ in an interval $[\underline{y}^{(0)}, \overline{y}^{(0)}]$, Walley's [18, Ch. 5.3] model with learning parameter $n^{(0)}$ is obtained, which typically is used in its near-ignorance form $[\underline{y}^{(0)}, \overline{y}^{(0)}] \to (0, 1)$, denoted as the imprecise Beta (Binomal/Bernoulli) model (IBBM)[5], which is a special case of the popular Imprecise Dirichlet (Multinomial) Model [19, 20]. Unfortunately, in this basic form with fixed $n^{(0)}$ the model is insensitive to prior-

[4] [15, 21, 22] use the prototypical character of (1) underlying (2) and (3) to generalize this inference to models based on canonical exponential families.

[5] We use 'IBBM' also for the model with prior information.

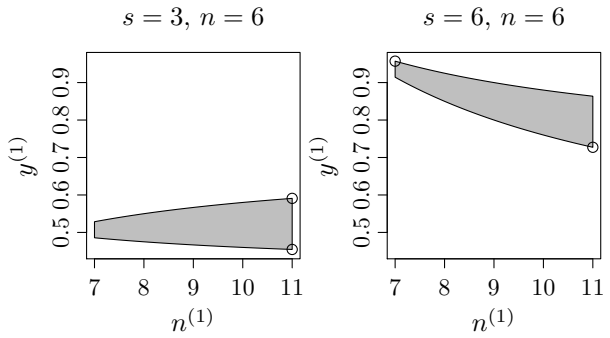


Figure 1: Posterior parameter sets $\Pi^{(n)}$ for rectangular $\Pi^{(0)}$. Left: *spotlight* shape; right: *banana* shape.

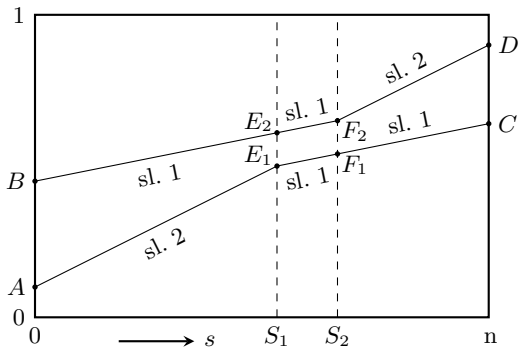


Figure 2: $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ for models in Sections 2.2 and 2.3.

data conflict [21, p. 263]. Walley [18, Ch. 5.4] therefore generalized this model by additionally varying $n^{(0)}$. In his extended model, called pdc-IBBM in this paper, the set of priors is defined via the set of prior parameters $\Pi^{(0)} = [\underline{n}^{(0)}, \overline{n}^{(0)}] \times [\underline{y}^{(0)}, \overline{y}^{(0)}]$, being a two-dimensional interval, or a rectangle set. Studying inference in this model, it is important to note that the set of posterior parameters $\Pi^{(n)}$ is not rectangular anymore. The resulting shapes are illustrated in Figure 1: For the prior set $\Pi^{(0)} = [1, 5] \times [0.4, 0.7]$—thus assuming a priori the fraction of successes to be between 40% and 70% and rating these assumptions with at least 1 and at most 5 pseudo observations—the resulting posterior parameter sets $\Pi^{(n)}$ are shown for data consisting of 3 successes in 6 draws (left) and with all 6 draws successes (right). We call the left shape *spotlight*, and the right shape *banana*. In both graphs, the elements of $\Pi^{(n)}$ yielding $\underline{y}^{(n)}$ and $\overline{y}^{(n)}$, and thus $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$, are marked with a circle.

The transition point between the *spotlight* and the *banana* shape in Figure 1 is the case when $\frac{s}{n} = \overline{y}^{(0)}$. Then $\overline{y}^{(n)}$, being a weighted average of $\overline{y}^{(0)}$ and $\frac{s}{n}$, is attained for all $n^{(0)} \in [\underline{n}^{(0)}, \overline{n}^{(0)}]$, and the top border of $\Pi^{(n)}$ in the graphical representation of Figure 1 is constant. Likewise, $\underline{y}^{(n)}$ is constant if $\frac{s}{n} = \underline{y}^{(0)}$. Therefore, (2) and (3) can be subsumed as

$$\underline{\mathrm{P}} = \begin{cases} \frac{\overline{n}^{(0)}\underline{y}^{(0)}+s}{\overline{n}^{(0)}+n} & \text{if } s \geq n \cdot \underline{y}^{(0)} =: S_1 \\ \frac{\underline{n}^{(0)}\underline{y}^{(0)}+s}{\underline{n}^{(0)}+n} & \text{if } s \leq n \cdot \underline{y}^{(0)} =: S_1 \end{cases},$$

$$\overline{\mathrm{P}} = \begin{cases} \frac{\overline{n}^{(0)}\overline{y}^{(0)}+s}{\overline{n}^{(0)}+n} & \text{if } s \leq n \cdot \overline{y}^{(0)} =: S_2 \\ \frac{\underline{n}^{(0)}\underline{y}^{(0)}+s}{\underline{n}^{(0)}+n} & \text{if } s \geq n \cdot \overline{y}^{(0)} =: S_2 \end{cases}.$$

The interval $[S_1, S_2]$ gives the range of expected successes $[n \cdot \underline{y}^{(0)}, n \cdot \overline{y}^{(0)}]$ and will be called 'Total Prior-Data Agreement' interval, or TPDA. For s in the TPDA, we are 'spot on': $\underline{y}^{(n)}$ and $\overline{y}^{(n)}$ are attained for $\overline{n}^{(0)}$ and $\Pi^{(n)}$ has the *spotlight* shape. But if the observed number of successes is outside TPDA, $\Pi^{(n)}$ goes *bananas* and either $\underline{\mathrm{P}}$ or $\overline{\mathrm{P}}$ is calculated with $\underline{n}^{(0)}$.

To summarize, the predictive probability plot (PPP), displaying $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ for $s \in [0, n]$, is given in Figure 2. For the pdc-IBBM, the specific values are

$$A = \frac{\underline{n}^{(0)}\underline{y}^{(0)}}{\underline{n}^{(0)}+n} \qquad C = \frac{\overline{n}^{(0)}\underline{y}^{(0)}+n}{\overline{n}^{(0)}+n}$$

$$B = \frac{\overline{n}^{(0)}\overline{y}^{(0)}}{\overline{n}^{(0)}+n} \qquad D = \frac{\underline{n}^{(0)}\overline{y}^{(0)}+n}{\underline{n}^{(0)}+n}$$

$$\text{sl. } 1 = \frac{1}{\overline{n}^{(0)}+n} \quad E_1 = \underline{y}^{(0)} \quad E_2 = \frac{\overline{n}^{(0)}\overline{y}^{(0)}+n\underline{y}^{(0)}}{\overline{n}^{(0)}+n}$$

$$\text{sl. } 2 = \frac{1}{\underline{n}^{(0)}+n} \quad F_2 = \overline{y}^{(0)} \quad F_1 = \frac{\overline{n}^{(0)}\underline{y}^{(0)}+n\overline{y}^{(0)}}{\overline{n}^{(0)}+n}.$$

As noted by [18, p. 224], the posterior predictive imprecision $\Delta = \overline{\mathrm{P}} - \underline{\mathrm{P}}$ can be calculated as

$$\Delta = \frac{\overline{n}^{(0)}(\overline{y}^{(0)} - \underline{y}^{(0)})}{\overline{n}^{(0)}+ n} + \frac{\overline{n}^{(0)} - \underline{n}^{(0)}}{(\underline{n}^{(0)}+ n)(\overline{n}^{(0)}+ n)}\Delta(s, \Pi^{(0)}),$$

where $\Delta(s, \Pi^{(0)}) = \inf\{|s - ny^{(0)}| : y^{(0)} \in [\underline{y}^{(0)}, \overline{y}^{(0)}]\}$ is the distance of s to the TPDA. If $\Delta(s, \Pi^{(0)}) \neq 0$, we have an effect of additional imprecision as desired, increasing linearly in s, because $\Pi^{(n)}$ is going *bananas*. However, when considering the fraction of observed successes instead of s, the onset of this additional imprecision immediately if $\frac{s}{n} \notin [\underline{y}^{(0)}, \overline{y}^{(0)}]$ seems very abrupt. Moreover, and even more severe, it happens irrespective of the number of trials n. When updating successively, this means that all single Bernoulli observations, being either 0 or 1, have to be treated as if being in conflict (except if $\overline{y}^{(0)} = 1$ and $s = n$ or if $\underline{y}^{(0)} = 0$ and $s = 0$). Furthermore, regarding $s/n = 7/10$ as an instance of prior-data conflict when $\overline{y}^{(0)} = 0.6$ had been assumed seems somewhat picky. To explore possibilities to amend this behaviour, alternative approaches are explored next.

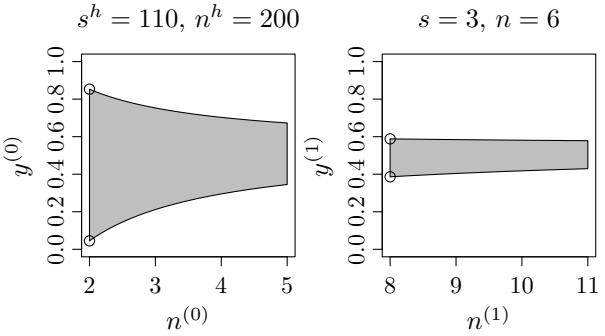


Figure 3: $\Pi^{(0)}$ and $\Pi^{(n)}$ for the *anteater* shape.

2.3 *Anteater* Shape Prior Sets

Choosing a two-dimensional interval $\Pi^{(0)}$ seems logical but the resulting inference is not fully satisfactory in case of prior data conflict. Recall that $\Pi^{(0)}$ is used to produce $\mathcal{M}^{(0)}$, which then is processed by the Generalized Bayes rule. Any shape can be chosen for $\Pi^{(0)}$, including the composure of single pairs $(n^{(0)}, y^{(0)})$. In this section we investigate an alternative shape, with $y^{(0)}$ a function of $n^{(0)}$, aiming at a more advanced behaviour in the case of prior-data conflict. To elicit $\Pi^{(0)}$, one could consider a thought experiment[6]: Given the hypothetical observation of s^h successes in n^h trials, which values should $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ take? In other words, what would one like to learn from data s^h/n^h in accordance with prior beliefs? As a simple approach, we can define $\Pi^{(0)}$ such that $\underline{\mathrm{P}} = \underline{c}$ and $\overline{\mathrm{P}} = \overline{c}$ are constants in $n^{(n)} = n^{(0)} + n^h$. Then, the lower and upper bounds for $y^{(0)}$ must be

$$\begin{aligned} \underline{y}^{(0)}(n^{(0)}) &= \big((n^h + n^{(0)})\underline{c} - s^h\big)/n^{(0)}\,, \\ \overline{y}^{(0)}(n^{(0)}) &= \big((n^h + n^{(0)})\overline{c} - s^h\big)/n^{(0)}\,, \end{aligned} \tag{4}$$

for $n^{(0)}$ in an interval $[\underline{n}^{(0)}, \overline{n}^{(0)}]$ derived by the range $[\underline{n}^{(n)}, \overline{n}^{(n)}]$ one wishes to attain for $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ given the n^h hypothetical observations.[7] The resulting shape of $\Pi^{(0)}$ is as in Figure 3 (left) and called *anteater* shape. Rewriting (4), $\Pi^{(0)}$ is now defined as

$$\Big\{(n^{(0)}, y^{(0)}) \mid n^{(0)} \in [\underline{n}^{(0)}, \overline{n}^{(0)}],\ y^{(0)}(n^{(0)}) \in \Big[\underline{c} - \frac{n^h}{n^{(0)}}\Big(\frac{s^h}{n^h} - \underline{c}\Big),\ \overline{c} + \frac{n^h}{n^{(0)}}\Big(\overline{c} - \frac{s^h}{n^h}\Big)\Big]\Big\}\,.$$

With the reasonable choice of $\underline{c}$ and $\overline{c}$ such that $\underline{c} \le s^h/n^h \le \overline{c}$, $\Pi^{(0)}$ can be interpreted as follows: The range of $y^{(0)}$ protrudes over $[\underline{c}, \overline{c}]$ on either side far enough to ensure $\underline{\mathrm{P}} = \underline{c}$ and $\overline{\mathrm{P}} = \overline{c}$ if updated with $s = s^h$ for $n = n^h$, the amount of protrusion decreasing in $n^{(0)}$ as the movement of $y^{(0)}(n^{(0)})$ towards s^h/n^h is slower for larger values of $n^{(0)}$. As there is a considerable difference in behaviour if $n > n^h$ or $n < n^h$, these two cases are discussed separately.

If $n > n^h$, the PPP graph in Figure 2 holds again, now with the values

$$A = \tfrac{\underline{c}(\underline{n}^{(0)} + n^h) - s^h}{\underline{n}^{(0)} + n} \quad S_1 = s^h + \underline{c}(n - n^h) \quad E_1 = \underline{c}$$
$$B = \tfrac{\overline{c}(\overline{n}^{(0)} + n^h) - s^h}{\overline{n}^{(0)} + n} \quad S_2 = s^h + \overline{c}(n - n^h) \quad F_2 = \overline{c}$$
$$C = \tfrac{\underline{c}(\overline{n}^{(0)} + n^h) - s^h + n}{\overline{n}^{(0)} + n} \qquad \text{sl. } 1 = 1/(\overline{n}^{(0)} + n)$$
$$D = \tfrac{\overline{c}(\underline{n}^{(0)} + n^h) - s^h + n}{\underline{n}^{(0)} + n} \qquad \text{sl. } 2 = 1/(\underline{n}^{(0)} + n)$$

$$E_2 = \underline{c} + \frac{\overline{n}^{(0)} + n^h}{\overline{n}^{(0)} + n}(\overline{c} - \underline{c}) = \overline{c} - \frac{n - n^h}{\overline{n}^{(0)} + n}(\overline{c} - \underline{c})$$
$$F_1 = \overline{c} - \frac{\overline{n}^{(0)} + n^h}{\overline{n}^{(0)} + n}(\overline{c} - \underline{c}) = \underline{c} + \frac{n - n^h}{\overline{n}^{(0)} + n}(\overline{c} - \underline{c})\,.$$

As for the pdc-IBBM, the TPDA boundaries S_1 and S_2 mark the transition points where either $\underline{y}^{(n)}$ or $\overline{y}^{(n)}$ are constant in $n^{(0)}$. We now have

$$\frac{S_1}{n} = \underline{c} + \frac{n^h}{n}\Big(\frac{s^h}{n^h} - \underline{c}\Big), \quad \frac{S_2}{n} = \overline{c} - \frac{n^h}{n}\Big(\overline{c} - \frac{s^h}{n^h}\Big),$$

so this TPDA is a subset of $[\underline{c}, \overline{c}]$. The *anteater* shape is, for $n > n^h$, even more strict than the pdc-IBBM, as, e.g., $\underline{y}^{(0)}(\overline{n}^{(0)}) = \underline{c} - \frac{n^h}{\overline{n}^{(0)}}\big(\frac{s^h}{n^h} - \underline{c}\big) < \frac{S_1}{n}$.

The situation for $n < n^h$ is illustrated in Figure 4, where A, B, C, D, E_1, F_2 and slopes 1 and 2 are the same as for $n > n^h$, but

$$E_2 = \underline{c} + \frac{\underline{n}^{(0)} + n^h}{\underline{n}^{(0)} + n}(\overline{c} - \underline{c}) = \overline{c} + \frac{n^h - n}{\underline{n}^{(0)} + n}(\overline{c} - \underline{c})\,,$$
$$F_1 = \overline{c} - \frac{\underline{n}^{(0)} + n^h}{\underline{n}^{(0)} + n}(\overline{c} - \underline{c}) = \underline{c} - \frac{n^h - n}{\underline{n}^{(0)} + n}(\overline{c} - \underline{c})\,.$$

Note that now $S_2 < S_1$, so the TPDA is $[S_2, S_1]$. In this interval, $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ are now calculated with $\underline{n}^{(0)}$; for $s \notin [S_2, S_1]$ the same situation as for $n > n^h$ applies, with the bound nearer to s/n calculated with $\underline{n}^{(0)}$ and the other with $\overline{n}^{(0)}$.

The upper transition point S_1 can now be between $\overline{y}^{(0)}(\underline{n}^{(0)})$ and $\overline{y}^{(0)}(\overline{n}^{(0)})$, and having S_1 decreasing in n now makes sense: the smaller n, the larger S_1, i.e. the more tolerant is the *anteater* set. The switch over S_1 (with s/n increasing) is illustrated in the three graphs in Figures 3 (right) and 5 (left, right): First, $\Pi^{(0)}$ from Figure 3 (left) is updated with $s/n = 3/6 < S_1/n$, leading again to an *anteater* shape, and so we get $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ from the elements of $\Pi^{(n)}$ at $\underline{n}^{(n)}$, as marked with circles. Second, the transition point is reached for $s = S_1 = 4.27$, and now $\underline{\mathrm{P}}$ is attained for any $n^{(n)} \in [\underline{n}^{(n)}, \overline{n}^{(n)}]$, as emphasized by the arrow. Third, as soon as s exceeds S_1 (in the graph:

[6] AKA 'pre-posterior' analysis in the Bayesian literature.

[7] For the rest of the paper, we tacitly assume that n^h, s^h, $n^{(0)}$ and $\underline{c}/\overline{c}$ are chosen such that $\underline{y}^{(0)} \ge 0$ resp. $\overline{y}^{(0)} \le 1$ to generate Beta distributions as priors.

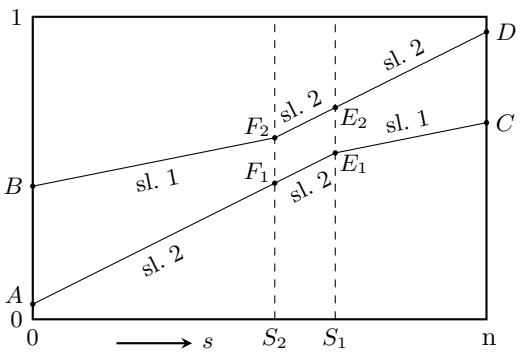


Figure 4: $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ for the *anteater* shape if $n < n^h$.

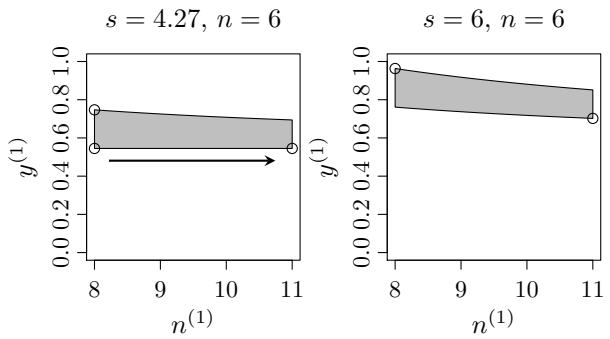


Figure 5: Posterior parameter sets $\Pi^{(n)}$ for *anteater* prior sets $\Pi^{(0)}$. Left: the transition point where $\underline{y}^{(n)}$ is attained for all $n^{(n)}$, right: the *banana* shape.

$s/n = 6/6$), it holds that $\underline{y}^{(n)}(\underline{n}^{(n)}) > \underline{y}^{(n)}(\overline{n}^{(n)})$, and $\underline{\mathrm{P}}$ is now attained at $\overline{n}^{(n)}$. As for the pdc-IBBM, for s outside the TPDA $\Pi^{(n)}$ goes *bananas*, leading to additional imprecision. The imprecision $\Delta = \overline{\mathrm{P}} - \underline{\mathrm{P}}$ if $n < n^h$ is

$$\Delta = \frac{\underline{n}^{(0)} + n^h}{\underline{n}^{(0)} + n}(\overline{c} - \underline{c}) + \frac{\overline{n}^{(0)} - \underline{n}^{(0)}}{(\underline{n}^{(0)} + n)(\overline{n}^{(0)} + n)}\Delta(s, n, \mathbf{c}),$$

where $\Delta(s, n, \mathbf{c}) = n\left|c^* - \frac{s}{n}\right| - n^h\left|c^* - \frac{s^h}{n^h}\right|$ and $c^* = \arg\max_{c \in [\underline{c}, \overline{c}]} \left|\frac{s}{n} - c\right|$ is the boundary of $[\underline{c}, \overline{c}]$ with the largest distance to s/n. For $s \in [S_2, S_1]$, $\Delta(s, n, \mathbf{c}) = 0$, giving a similar structure as for the pdc-IBBM except that $\Delta(s, n, \mathbf{c})$ does not directly give the distance of s/n to $\Pi^{(0)}$ but is based on $[\underline{c}, \overline{c}]$. The imprecision increases again linearly with s, but now also with n. The distance of s/n to the opposite bound of $[\underline{c}, \overline{c}]$ (weighted with n) is discounted by the distance of s^h/n^h to the same bound (weighted with n^h). In essence, $\Delta(s, n, \mathbf{c})$ is thus a reweighted distance of s/n to s^h/n^h. The more dissimilar these fractions are, the larger the posterior predictive imprecision is.

For $n = n^h$, $S_1 = S_2 = s^h$ so the TPDA is reduced to a single point. In this case, the *anteater* shape can be considered as an equilibrium point, with any $s \neq s^h$ leading to increased posterior imprecision. In this case, the weights in $\Delta(s, n, \mathbf{c})$ coincide, and so the posterior imprecision depends directly on $|s - s^h|$.

$n > n^h$	$s < S_1$	$s \in [S_1, S_2]$	$s > S_2$
	banana	spotlight	banana
$n = n^h$	$s < s^h$	$s = s^h$	$s > s^h$
	banana	rectangular	banana
$n < n^h$	$s < S_2$	$s \in [S_2, S_1]$	$s > S_1$
	banana	anteater	banana

Table 1: Shapes of $\Pi^{(n)}$ if $\Pi^{(0)}$ has the *anteater* shape.

For $n > n^h$ the transition behaviour is as for the pdc-IBBM: As long as $s \in [S_1, S_2]$, $\Pi^{(n)}$ has the *spotlight* shape, where both $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ are calculated with $\overline{n}^{(n)}$; Δ for $s \in [S_1, S_2]$ is thus calculated with $\overline{n}^{(n)}$ as well. If, e.g., $s > S_2$, $\overline{\mathrm{P}}$ is attained with $\underline{n}^{(n)}$, and $\Delta(s, n, \mathbf{c})$ gives directly the distance of s/n to s^h/n^h, the part of which is inside $[\underline{c}, \overline{c}]$ is weighted with n, and the remainder with n^h. Table 1 provides an overview of the possible shapes of $\Pi^{(n)}$.

2.4 Intermediate Résumé

Despite the (partly) different behaviour inside the TPDA, both pdc-IBBM and the *anteater* shape display only two different slopes in their PPPs (Figures 2 and 4), with either $\underline{n}^{(n)}$ or $\overline{n}^{(n)}$ used to calculate $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$. It is possible to have shapes such that for some s other values from $[\underline{n}^{(n)}, \overline{n}^{(n)}]$ are used. As a toy example, consider $\Pi^{(0)} = \{(1, 0.4), (3, 0.6), (5, 0.4)\}$, so consisting only of three parameter combinations $(n^{(0)}, y^{(0)})$. $\overline{\mathrm{P}}$ is then derived as $\overline{y}^{(n)} = \max\{\frac{0.4+s}{1+n}, \frac{1.8+s}{3+n}, \frac{2+s}{5+n}\}$, leading to

$$\overline{y}^{(n)} = \begin{cases} \frac{0.4+s}{1+n} & \text{if } s > 0.7n + 0.3 \\ \frac{1.8+s}{3+n} & \text{if } 0.1n - 1.5 < s < 0.7n + 0.3 \\ \frac{2+s}{5+n} & \text{if } s < 0.1n - 1.5 \end{cases}.$$

So, in a PPP we would observe the three different slopes $1/(1+n)$, $1/(3+n)$ and $1/(5+n)$ depending on the value of s. Our conjecture is therefore that with carefully tailored sets $\Pi^{(0)}$, an arbitrary number of slopes is possible, and so even smooth curvatures. Using a thought experiment as for the *anteater* shape, $\Pi^{(0)}$ shapes can be derived to fit any required behaviour. Another approach for constructing a $\Pi^{(0)}$ that is more tolerant with respect to prior-data conflict could be as follows: As the onset of additional imprecision in the pdc-IBBM is caused by the fact that $\overline{y}^{(n)}(\underline{n}^{(n)}) > \overline{y}^{(n)}(\overline{n}^{(n)})$ as soon as $s/n > \overline{y}^{(0)}$, we could define the $y^{(0)}$ interval at $\underline{n}^{(0)}$ to be narrower than the $y^{(0)}$ interval at $\overline{n}^{(0)}$, so that the *banana* shape results only when s/n exceeds $\overline{y}^{(0)}(\overline{n}^{(0)})$

far enough. Having a narrower $y^{(0)}$ interval at $\underline{n}^{(0)}$ than at $\overline{n}^{(0)}$ could also make sense from an elicitation point of view: We might be able to give quite a precise $y^{(0)}$ interval for a low prior strength $\underline{n}^{(0)}$, whereas for a high prior strength $\overline{n}^{(0)}$ we must be more cautious with our elicitation of $y^{(0)}$, i.e. giving a wider interval. The rectangular shape for $\Pi^{(0)}$ as discussed in Section 2.2 seems thus somewhat peculiar. One could also argue that if one has substantial prior information but acknowledges that this information may be wrong, one should not reduce the weight of the prior $n^{(0)}$ on the posterior while keeping the same informative interval of values of $y^{(0)}$.

Generally, the actual shape of a set $\Pi^{(0)}$ influences the inferences, but for a specific inference only a few aspects of the set are relevant. So, while a detailed shape of a prior set may be very difficult to elicit, it may not even be that relevant for a specific inference. A further general issue seems unavoidable in the generalized Bayesian setting as developed here, namely the dual role of $n^{(0)}$. On the one hand, $n^{(0)}$ governs the weighting of prior information $y^{(0)}$ with respect to the data s/n, as mentioned in Section 2.1: The larger $n^{(0)}$, the more $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ are dominated by $\underline{y}^{(0)}$ and $\overline{y}^{(0)}$. On the other hand, $n^{(0)}$ governs also the degree of posterior imprecision: the larger $n^{(0)}$, the larger c.p. Δ. A larger $n^{(0)}$ thus leads to more imprecise posterior inferences, although a high weight on the supplied prior information should boost the trust in posterior inferences if s in the TPDA, i.e. the prior information turned out to be appropriate. In the next section, we thus develop a different approach separating these two roles: Now, two separate models for predictive inference, each resulting in different precision as governed by $n^{(0)}$, are combined with an imprecise weight α taking the role of regulating prior-data agreement.

3 Weighted Inference

We propose a variation of the Beta-Binomial model that is attractive for prior-data conflict and has small yet fascinating differences with the models in Sections 2.2 and 2.3. We present a basic version of the model in Section 3.1, followed by an extended version in Section 3.2. Opportunities to generalize the model are mentioned in Section 3.3.

3.1 The Basic Model

The idea for the proposed model is to combine the inferences based on two models, each part of an imprecise Bayesian inferential framework using sets of prior distributions, although the inferences can also result from alternative inferential methods. The combination is not achieved by combining the two sets of prior distributions into a single set, but by combining the posterior predictive *inferences* by imprecise weighted averaging. When the weights assigned to the two models can vary over the whole range $[0, 1]$ we actually return to imprecise Bayesian inference with a prior set, as considered in this subsection. In Section 3.2 we restrict the values of the model weights. The basic model turns out to be relevant from many perspectives, in particular to highlight similarities and differences with the methods presented in Sections 2.2 and 2.3, and it is a suitable starting point for more general models. These aspects will be discussed in Subsection 3.3.

We consider the combination of the imprecise posterior predictive probabilities $[\underline{\mathrm{P}}^i, \overline{\mathrm{P}}^i]$ and $[\underline{\mathrm{P}}^u, \overline{\mathrm{P}}^u]$ for the event that the next observation is a success with

$$\underline{\mathrm{P}}^i = \frac{s^i + s}{n^i + n + 1} \quad \text{and} \quad \overline{\mathrm{P}}^i = \frac{s^i + s + 1}{n^i + n + 1}\,, \tag{5}$$

$$\underline{\mathrm{P}}^u = \frac{s}{n+1} \quad \text{and} \quad \overline{\mathrm{P}}^u = \frac{s+1}{n+1}\,. \tag{6}$$

The superscript i indicates 'informative', in the sense that these lower and upper probabilities relate to an 'informative' prior distribution reflecting prior beliefs of similar value as s^i successes in n^i observations. The superscript u indicates 'uninformative', which can be interpreted as absence of prior beliefs. These lower and upper probabilities can for example result from Walley's IBBM, with $\underline{\mathrm{P}}^i$ and $\overline{\mathrm{P}}^i$ based on the prior set with $n^{(0)} = n^i + 1$ and $y^{(0)} \in \left[\frac{s^i}{n^i+1}, \frac{s^i+1}{n^i+1}\right]$, and $\underline{\mathrm{P}}^u$ and $\overline{\mathrm{P}}^u$ on the prior set with $n^{(0)} = 1$ and $y^{(0)} \in [0, 1]$. There are other methods for imprecise statistical inference that lead to these same lower and upper probabilities, including Nonparametric Predictive Inference for Bernoulli quantities [4][8], where the s^i and n^i would only be included if they were actual observations, for example resulting from a second data set that one may wish to include in the 'informative' model but not in the 'uninformative' model.

The proposed method combines these lower and upper predictive probabilities by imprecise weighted averaging. Let $\alpha \in [0, 1]$, we define

$$\underline{\mathrm{P}}_\alpha = \alpha\underline{\mathrm{P}}^i + (1-\alpha)\underline{\mathrm{P}}^u, \quad \overline{\mathrm{P}}_\alpha = \alpha\overline{\mathrm{P}}^i + (1-\alpha)\overline{\mathrm{P}}^u, \tag{7}$$

and as lower and upper predictive probabilities for the event that the next Bernoulli random quantity is a success[9]

$$\underline{\mathrm{P}} = \min_{\alpha\in[0,1]} \underline{\mathrm{P}}_\alpha \quad \text{and} \quad \overline{\mathrm{P}} = \max_{\alpha\in[0,1]} \overline{\mathrm{P}}_\alpha\,.$$

[8] See also `www.npi-statistics.com`.

[9] While in (2) and (3), prior and sample information are imprecisely weighted, here informative and uninformative models are combined.

Allowing α to take on any value in $[0,1]$ reduces this method to the IBBM with a single prior set, as discussed in Section 2, with the prior set simply generated by the union of the two prior sets for the 'informative' and the 'uninformative' models as described above. For all s these minimum and maximum values are obtained at either $\alpha = 0$ or $\alpha = 1$. With switch points $S_1 = (n+1)\frac{s^i}{n^i} - 1$ and $S_2 = (n+1)\frac{s^i}{n^i}$, they are equal to

$$\underline{\mathrm{P}} = \begin{cases} \underline{\mathrm{P}}^u = \frac{s}{n+1} & \text{if } s \leq S_2 \\ \underline{\mathrm{P}}^i = \frac{s^i+s}{n^i+n+1} & \text{if } s \geq S_2\,, \end{cases}$$

$$\overline{\mathrm{P}} = \begin{cases} \overline{\mathrm{P}}^i = \frac{s^i+s+1}{n^i+n+1} & \text{if } s \leq S_1 \\ \overline{\mathrm{P}}^u = \frac{s+1}{n+1} & \text{if } s \geq S_1\,. \end{cases}$$

The PPP graph for this model is displayed in Figure 6. The upper probability for $s = S_1$ and the lower probability for $s = S_2$ are both equal to $\frac{s^i}{n^i}$. The TPDA contains only a single possible value of s (except if S_1 and S_2 are integer), namely the one that is nearest to $\frac{s^i}{n^i}$. The specific values for this basic case are

$$A = 0 \qquad B = \frac{s^i+1}{n^i+n+1} \qquad C = \frac{s^i+n}{n^i+n+1}$$

$$D = 1 \qquad E = \frac{s^i}{n^i} - \frac{1}{n+1} \qquad F = \frac{s^i}{n^i} + \frac{1}{n+1}$$

$$\text{sl. } 1 = \frac{1}{n^i+n+1} \qquad \text{sl. } 2 = \frac{1}{n+1}\,.$$

If s is in the TPDA it reflects optimal agreement of the 'prior data' (n^i, s^i) and the (really observed) data (n, s), so it may be a surprise that both the lower and upper probabilities in this case correspond to $\alpha = 0$, so they are fully determined by the 'uninformative' part of the model. This is an important aspect, it will be discussed in more detail and compared to the methods of Section 2 in Subsection 3.3. For s in the TPDA both $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ increase with slope $\frac{1}{n+1}$ and $\Delta = \frac{1}{n+1}$.

Figure 6, with the specific values for this basic case given above, illustrates what happens for values of s outside this TPDA. Moving away from the TPDA in either direction, the imprecision increases as was also the case in the models in Section 2. For s decreasing towards 0, this is effectively due to the smaller slope of the upper probability, while for s increasing towards 1 it is due to the smaller slope of the lower probability. For $s \in [0, S_1]$, the imprecision is $\Delta = \frac{s^i+1}{n^i+n+1} - \frac{sn^i}{(n^i+n+1)(n+1)}$. For $s \in [S_2, n]$ the imprecision is $\Delta = \frac{1}{n+1} - \frac{s^i}{n^i+n+1} + \frac{sn^i}{(n^i+n+1)(n+1)}$. For the two extreme possible cases of prior data conflict, with either $s^i = n^i$ and $s = 0$ or $s^i = 0$ and $s = n$, the imprecision is $\Delta = \frac{n^i+1}{n^i+n+1}$. For this combined model with $\alpha \in [0,1]$, we have $\underline{\mathrm{P}} \leq \frac{s}{n} \leq \overline{\mathrm{P}}$ for all s, which is attractive from the perspective of objective inference.

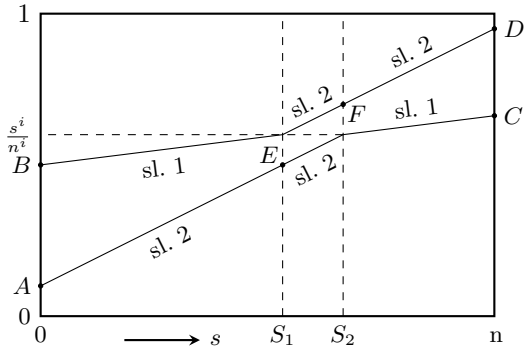


Figure 6: $\underline{\mathrm{P}}$ and $\overline{\mathrm{P}}$ for the weighted inference model.

3.2 The Extended Model

We extend the basic model from Subsection 3.1, perhaps remarkably by reducing the interval for the weighting variable α. We assume that $\alpha \in [\alpha_l, \alpha_r]$ with $0 \leq \alpha_l \leq \alpha_r \leq 1$. We consider this an extended version of the basic model as there are two more parameters that provide increased modelling flexibility. It is important to remark that, with such a restricted interval for the values of α, this weighted model is no longer identical to an IBBM with a single set of prior distributions. One motivation for this extended model is that the basic model seemed very cautious by not using the informative prior part if s is in the TPDA. For $\alpha_l > 0$, the informative part of the model influences the inferences for all values of s, including the one in the TPDA. As a consequence of taking $\alpha_l > 0$, however, the line segment $(s, \frac{s}{n})$ with $s \in [0, n]$ will not always be in between the lower and upper probabilities anymore, specifically not at, and close to, $s = 0$ and $s = n$, as follows from the results presented below.

The lower and upper probabilities resulting from the two models that are combined by taking an imprecise weighted average are again as given by formulae (5)-(6), with the weighted averages $\underline{\mathrm{P}}_\alpha$ and $\overline{\mathrm{P}}_\alpha$, for any $\alpha \in [\alpha_l, \alpha_r]$, again given by (7). This leads to the lower and upper probabilities for the combined inference

$$\underline{\mathrm{P}} = \min_{\alpha \in [\alpha_l, \alpha_r]} \underline{\mathrm{P}}_\alpha \qquad \text{and} \qquad \overline{\mathrm{P}} = \max_{\alpha \in [\alpha_l, \alpha_r]} \overline{\mathrm{P}}_\alpha\,.$$

The lower and upper probabilities have, as function of s, the generic forms presented in Figure 6, with $[S_1, S_2] = \left[(n+1)\frac{s^i}{n^i} - 1, (n+1)\frac{s^i}{n^i}\right]$ as in Section 3.1. The specific values for Figure 6 are

$$A = \frac{\alpha_l s^i}{n^i+n+1} \qquad B = \frac{1}{n+1} + \frac{\alpha_r[s^i(n+1)-n^i]}{(n^i+n+1)(n+1)}$$

$$D = 1 - \frac{\alpha_l(n^i-s^i)}{n^i+n+1} \qquad C = \frac{n}{n+1} - \frac{\alpha_r[(n^i-s^i)(n+1)-n^i]}{(n^i+n+1)(n+1)}$$

sl. 1 $= \frac{n^i+n+1-\alpha_r n_i}{(n^i+n+1)(n+1)}$ $\quad E = \frac{s^i}{n^i} - \frac{1}{n+1}\left[1 - \frac{\alpha_l n^i}{n^i+n+1}\right]$

sl. 2 $= \frac{n^i+n+1-\alpha_l n_i}{(n^i+n+1)(n+1)}$ $\quad F = \frac{s^i}{n^i} + \frac{1}{n+1}\left[1 - \frac{\alpha_l n^i}{n^i+n+1}\right]$.

The increase in imprecision when s moves away from the TPDA can again be considered as caused by the informative part of the model, which is logical as the uninformative part of the model cannot exhibit prior-data conflict.

The possibility to choose values for α_l and α_r provides substantially more modelling flexibility compared to the basic model presented in Section 3.1. One may, for example, wish to enable inferences solely based on the informative part of the model, hence choose $\alpha_r = 1$, but ensure that this part has influence on the inferences in all situations, with equal influence to the uninformative part in case of TPDA. This latter aspect can be realized by choosing $\alpha_l = 0.5$. When compared to the situation in Section 3.1, this choice moves, in Figure 6, A and D away from 0 and 1, respectively, but does not affect B and C. It also brings E and F a bit closer to the corresponding upper and lower probabilities, respectively, hence reducing imprecision in the TPDA.

3.3 Weighted Inference Model Properties

The basic model presented in Section 3.1 is fits in the Bayesian framework, but its use of prior information is different to the usual way in Bayesian statistics. The lower and upper probabilities are mainly driven by the uninformative part, which e.g. implies that $\underline{\mathrm{P}} \leq \frac{s}{n} \leq \overline{\mathrm{P}}$ for all values of s. While in (imprecise, generalized) Bayesian statistics any part of the model that uses an informative prior can be regarded as adding information to the data, the informative part of the basic model leads to more careful inferences when there is prior-data conflict. Figure 6 shows that, for the basic case of Section 3.1, the points A and D are based only on the uninformative part of the model, but the points B and C are based on the informative part of the model.

Prior-data conflict can be of different strength, one would expect to only talk about 'conflict' if consideration is required, hence the information in the prior and in the data should be sufficiently strong. The proposed method in Section 3.1 takes as starting point inference that is fully based on the data, it uses the informative prior part of the model to widen the interval of lower and upper probabilities in the direction of the value $\frac{s^i}{n^i}$. For example, if one observed $s = 0$, the upper probability of a success at the next observation is equal to $\frac{s^i+1}{n^i+n+1}$, which reflects inclusion of the information in the prior set for the informative part of the model that is most supportive for this event, equivalent to $s^i + 1$ successes in $n^i + 1$ observations. As such, the effect of the prior information is to weaken the inferences by increasing imprecision in case of prior-data conflict.

One possible way in which to view this weighted inference model is as resulting from a multiple expert or information source problem, where one wishes to combine the inferences resulting individually from each source. The basic model of Section 3.1 leads to the most conservative inference such that no individual model or expert disagrees, while the restriction on weights provides a guaranteed minimum level for the individual contributions to the combined inference.

It should be emphasized that the weighted inference model has wide applicability. The key idea is to combine, by imprecise weighting, the actual inferences resulting from multiple models, and as such there is much scope for the use and further development of this approach. The individual models could even be models such as those described in Sections 2.2 and 2.3, although that would lead to more complications. If the individual models are coherent lower and upper probabilities, i.e. provide separately coherent inferences, then the combined inference via weighted averaging and taking the lower and upper envelopes is also separately coherent[10].

In applications, it is often important to determine a sample size (or more general design issues) before data are collected. If one uses a model that can react to prior-data conflict, this is likely to lead to a larger data requirement. One very cautious approach is to choose n such that the maximum possible resulting imprecision does not exceed a chosen threshold. In the models presented in this paper, this maximum imprecision will always occur for either $s = 0$ or $s = n$, whichever is further away from the TPDA. In such cases, a preliminary study has shown an attractive feature if one can actually sample sequentially. If some data are obtained with success proportion close to s^i/n^i, the total data requirement (including these first observations) to ensure that the resulting maximum imprecision cannot exceed the same threshold level is substantially less than had been the case before any data were available. This would be in line with intuition, and further research into this and related aspects is ongoing, including of course the further data need in case first sampled data is in conflict with (n^i, s^i), and the behaviour of the models of Section 2 in such cases.

The weighted inference method combines the inferences based on two models, and can be generalized to allow more than two models and different inferential methods. It is also possible to allow more impreci-

[10] This follows from e.g. [18, 2.6.3f]

sion in each of the models that are combined, leading to more parameters in the overall model that can be used to control the behaviour of the inferences. Similar post-inference combination via weighted averaging, but with precise weights, has been presented in the frequentist statistics literature [11, 13], where the weights are actually determined based on the data and a chosen optimality criterion for the combined inference. In Bayesian statistics, estimation or prediction inferences based on different models can be similarly combined using Bayes factors [12], which are based on both the data (via the likelihood function) and prior weightings for the different models. In our approach, we do not use the data or prior beliefs about the models to derive precise weights for the models, instead we cautiously base our combined lower and upper predictive probabilities on those of the individual models with a range of possible weights. This range is set by the analyst and does not explicitly take the data or prior beliefs into account, but it provides flexibility with regard to the relative importance given to the individual models.

4 Insights and Challenges

We have discussed two different classes of inferential methods to handle prior-data conflict in the Bernoulli case. These can be generalized to the multinomial case corresponding to the IDM. It also seems possible to extend the approaches to continuous sampling models like the normal or the gamma distribution, by utilizing the fact that the basic form of the updating of $n^{(0)}$ and $y^{(0)}$ in (1) underlying (2) and (3) is valid for arbitrary canonical exponential families [15, 21]. Further insight into the weighting method may also be provided by comparing it to Generalized Bayesian analysis based on sets of conjugate priors consisting of nontrivial mixtures of two Beta distributions. There, however, the posterior mixture parameter depends on the other parameters. For a deeper understanding of prior-data conflict it may also be helpful to extend our methods to coarse data, in an analogous way to [17] and [16], and to look at other model classes of prior distributions, most notably at contamination neighbourhoods. Of particular interest here may be to combine both types of prior models, considering contamination neighbourhoods of our exponential family based-models with sets of parameters, as developed in the Neyman-Pearson setting by [1, Section 5].

The models presented here address prior-data conflict in different ways, either by fully utilizing the prior information in a way that is close to the traditional Bayesian method, where this information is added to data information, or by not including them initially as in Section 3. All these models show the desired increase of imprecision in case of prior-data conflict. It may be of interest to derive methods that explicitly respond to (perhaps surprisingly) strong prior-data agreement. One possibility to achieve this with the methods presented here is to consider the TPDA as this situation of strong agreement in which one wants imprecision reduced further than compared to an 'expected' situation, and to choose the prior set (Section 2) or the two inferential models (Section 3) in such a way to create this effect. This raises interesting questions for elicitation, but both approaches provide opportunities for this and we consider it as an important topic for further study.

Far beyond further extensions one has, from the foundational point of view, to be aware that there are many ways in which people might react to prior-data conflict, and we may perhaps at best hope to catch some of these in a specific model and inferential method. This is especially important when the conflict is very strong, and indeed has to be considered as full contradiction of modeling assumptions and data, which may lead to a revision of the whole system of background knowledge in the light of surprising observations, as Hampel argues.[11] In this context applying the weighting approach to the NPI-based model for categorical data [6] may provide some interesting opportunities, as it explicitly allows to consider not yet observed and even undefined categories [5].

There is another intriguing way in which one may react to prior-data conflict, namely by considering the combined information to be of less value than either the real data themselves or than both information sources. Strong prior beliefs about a high success rate could be strongly contradicted by data, as such leading to severe doubt about what is actually going on. The increase of imprecision in case of prior-data conflict in the methods presented in this paper might be interpreted as reflecting this, but there may be other opportunities to model such an effect. It may be possible to link these methods to some popular approaches in frequentist statistics, where some robustness can be achieved or where variability of inferences can be studied by round robin deletion of some of the real observations.This idea may open up interesting research challenges for imprecise probability models, where the extent of data reduction could perhaps be related to the level of prior-data conflict. Of course, such approaches would only be of use in situations with substantial amounts of real data, but as mentioned before these are typically the situations where prior-data conflict is most likely to be of sufficient relevance to take its modelling seriously. As

[11] See in particular the discussion of the structure and role of background knowledge in [10].

(imprecise, generalized) Bayesian methods all work essentially by *adding* information to the real data, it is unlikely that such new methods can be developed within the Bayesian framework, although there may be opportunities if one restricts the inferences to situations where one has at least a pre-determined number of observations to ensure that posterior inferences are proper. For example, one could consider allowing the prior strength parameter $n^{(0)}$ in the IBBM to take on negative values, opening up a rich field for research and discussions.

Acknowledgements

We thank the referees for very helpful comments.

References

[1] 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(1):149–173, 2002.

[2] T. Augustin, F.P.A. Coolen, S. Moral, and M.C.M. Troffaes, editors. *ISIPTA '09: Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications*. SIPTA, 2009.

[3] J.-M. Bernard. Special Issue on the Imprecise Dirichlet Model. *International Journal of Approximate Reasoning*, 50:201–268, 2009.

[4] F. P. A. Coolen. Low structure imprecise predicitive inference for Bayes' problem. *Statistics & Probability Letters*, 36:349–357, 1998.

[5] F.P.A. Coolen and T. Augustin. Learning from multinomial data: a nonparametric predictive alternative to the Imprecise Dirichlet Model. In F.G. Cozman, R. Nau, and T. Seidenfeld, editors, *ISIPTA '05: Proc. of the Fourth International Symposium on Imprecise Probabilities and their Applications*, pages 125–135, 2005.

[6] 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.

[7] Frank P. A. Coolen. On Bernoulli experiments with imprecise prior probabilities. *The Statistician*, 43:155–167, 1994.

[8] G. de Cooman, J. Vejnarová, and M. Zaffalon, editors. *ISIPTA '07: Proceedings of the Fifth International Symposium on Imprecise Probabilities and Their Applications*. SIPTA, 2007.

[9] M. Evans and H. Moshonov. Checking for prior-data conflict. *Bayesian Analysis*, 1:893–914, 2006.

[10] F. Hampel. How can we get new knowledge? In T. Augustin, F.P.A. Coolen, S. Moral, and M.C.M. Troffaes, editors, *ISIPTA '09: Proc. of the Sixth International Symposium on Imprecise Probabilities: Theories and Applications*, pages 219–227, 2009.

[11] N.L. Hjort and G. Claeskens. Frequentist model average estimators. *Journal of the American Statistical Association*, 98:879–899, 2003.

[12] R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[13] N.T. Longford. An alternative to model selection in ordinary regression. *Statistics and Computing*, 13:67–80, 2003.

[14] L. P. Pericchi and P. Walley. Robust Bayesian credible intervals and prior ignorance. *International Statistical Review*, 58:1–23, 1991.

[15] 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. Proc. of the Fourth International Symposium on Imprecise Probabilities and Their Applications*, pages 287–296, 2005.

[16] M.C.M. Troffaes and F.P.A. Coolen. Applying the imprecise Dirichlet model in cases with partial observations and dependencies in failure data. *International Journal of Approximate Reasoning*, 50(2):257–268, 2009.

[17] L.V. Utkin and T. Augustin. Decision making under imperfect measurement using the imprecise Dirichlet model. *International Journal of Approximate Reasoning*, 44(3):322–338, 2007.

[18] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[19] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996.

[20] P. Walley and J.-M. Bernard. Imprecise probabilistic prediction for categorical data. Technical Report CAF-9901, Paris 8, 1999.

[21] G. Walter and T. Augustin. Imprecision and prior-data conflict in generalized Bayesian inference. *Journal of Statistical Theory and Practice*, 3:255–271, 2009.

[22] G. Walter, T. Augustin, and A. Peters. Linear regression analysis under sets of conjugate priors. In G. de Cooman, J. Vejnarová, and M. Zaffalon, editors, *ISIPTA '07: Proc. of the Fifth International Symposium on Imprecise Probabilities: Theories and Applications*, pages 445–455, 2007.

[23] K.M. Whitcomb. Quasi-Bayesian analysis using imprecise probability assessments and the generalized Bayes' rule. *Theory and Decision*, 58:209–238, 2005.

Utility-Based Accuracy Measures to Empirically Evaluate Credal Classifiers

Marco Zaffalon
IDSIA, Switzerland
zaffalon@idsia.ch

Giorgio Corani
IDSIA, Switzerland
giorgio@idsia.ch

Denis Mauá
IDSIA, Switzerland
denis@idsia.ch

Abstract

Predictions made by imprecise-probability models are often indeterminate (that is, set-valued). Measuring the quality of an indeterminate prediction by a single number is important to fairly compare different models, but a principled approach to this problem is currently missing. In this paper we derive a measure to evaluate the predictions of credal classifiers from a set of assumptions. The measure turns out to be made of an objective component, and another that is related to the decision-maker's degree of risk-aversion. We discuss when the measure can be rendered independent of such a degree, and provide insights as to how the comparison of classifiers based on the new measure changes with the number of predictions to be made. Finally, we empirically study the behavior of the proposed measure.

Keywords. Credal classification, indeterminacy, empirical evaluations, discounted accuracy, utility, risk-aversion.

1 Introduction

When we use an imprecise-probability model to make predictions, we meet one of the most striking differences of imprecise probability in comparison to precise probability: the imprecise-probability model can issue indeterminate predictions. That is, among the set of possible options, the model may drop some of them as sub-optimal, while keeping the entire remaining set as its prediction. The prediction is generally indeterminate as such a set is not necessarily a singleton. Indeterminate predictions are a crucially important feature of imprecise-probability models: they allow credible, and reliable, predictions to be obtained no matter how scarce is the information available to build a model.

Yet, we should have a way to *measure* how good is an indeterminate prediction. A major reason is that we need to compare imprecise- with precise-probability models: we should have a clear, simple, and possibly shared, way to say which one is better, in a given application. The same consideration applies, of course, when we compare two imprecise-probability models. Ideally, we would like to be able to reward each, determinate or indeterminate, prediction by a single number. Most probably this would speed up progress in the field, as it would enable comparisons to be automatized over a large number of test applications.

In the case of precise-probability models, there are well-consolidated measures to do this. Let us consider the field of *pattern classification* [4], which is the focus of this paper (Section 2 gives a brief introduction to classification problems). In this case, the predictive models are called (precise) *classifiers*. A classifier predicts one out of a finite set $\mathscr{C}$ of so-called *classes*. In this case, correct predictions may be rewarded with 1 and incorrect ones with 0, thus giving rise to the measure of performance called the *predictive accuracy* of a classifier: i.e., the proportion of correct predictions it makes.

The situation is very different with *credal classifiers*, that is, classifiers that issue set-valued predictions. One of the very few proposals to evaluate an indeterminate prediction by a single number can be found in [2]: a prediction made of a set $\mathscr{K}$ of k classes is rewarded with $1/k$ if it contains the actual class, and with 0 otherwise. This gives rise to the measure called *discounted accuracy*, which was borrowed from the field of multi-label classification [11]. The problem here is that no justification is given for discounted accuracy, as the work in [2] points out. In [7], classifiers which return indeterminate classifications are evaluated through the F-metric, originally designed for information retrieval problems; but also here the measure is not justified. Other than these, the proposals are either explicitly non-numerical, as the rank test in [2], or require a vector of parameters to evaluate the performance, as in [1]. The latter approach is actually meaningful, but was conceived to compare credal with precise classifiers, and cannot be easily generalized to the more general case; moreover, it is a method that needs supervision so that it does not easily lend itself to be run automatically on many test cases.

In our view, the scarcity of principled numerical evaluation methods for credal classifiers is not accidental: in fact, it is not easy to assign a single number to an indeterminate prediction. Consider the following case: there is a *vacuous*

classifier, which every time predicts the set of all classes $\mathscr{C}$, and a *random* one, which picks up a class from $\mathscr{C}$ through the uniform distribution. If $\mathscr{C}$ is made of two classes (we say that the classification problem is binary), and we use the predictive accuracy, the random classifier has an expected reward equal to $1/2$. What should be the expected reward of the vacuous classifier? Both classifiers do not know how to predict the class, but only the vacuous classifier declares it. From this, one might argue that the latter should be rewarded with more than $1/2$. On the other hand, it is clear also that the vacuous classifier cannot predict the class better than the random one, so that one might argue that it should be rewarded with $1/2$ too.

In the attempt to address these kinds of problems in the most objective way, we found it useful to regard classifiers as bettors. In the betting framework introduced in Section 3, we assume we only know how to value determinate predictions, in particular by 0-1 rewards. In Section 4, we extend the framework, in a kind of least-committal way, to credal classifiers: we show that, under certain assumptions, indeterminate predictions should be valued according to discounted accuracy.

Note that, in the previous example, discounted accuracy would value the vacuous and the random classifiers the same. This kind of (questionable) effect can be traced back to having deliberately avoided introducing subjective considerations in the evaluation. Still, subjective preferences should be accounted for: we introduce in Section 5 a decision-maker in charge of selecting the 'best' classifier in the next bet, and show that preferences can enter the picture through his utility, as a function of discounted accuracy. This defines the utility-based accuracy measure we propose to evaluate credal classifiers. More generally, this shows in a very definite sense how the reliability of a classifier is tightly related to the variability of its predictions, and that the aversion to this variability is what makes some people prefer credal classifier to precise ones.

In Section 6 we discuss an important case where the evaluation can still be made in quite an objective way despite the decision-maker's preferences, and we relate this to the amount of indeterminacy produced by a credal classifier.

In Section 7 we analyze how the picture changes if we focus on evaluating classifiers in the next $m \geq 1$ bets. We show that the difference between precise and credal classifiers decreases with growing m, so that the relative benefits of credal classification are less important with large m.

Finally, in Sections 8 and 9 we make some empirical analysis of our utility-based measure. We compare *naive Bayes* [3] and *naive credal classifier* [1] on binary problems. We show that the decision-maker's utility can be defined very easily in this case, and that the credal classifier becomes superior to the precise one even with relatively small preferences of the decision-maker towards reliable predictions.

2 Classification Problems

A classification problem is made of objects described by *attribute* (or *feature*) variables, which we group into the single variable A, and a class variable C. The class variable represents the object's category. There are finitely many possible categories, which we identify with their indexes to simplify notation: $\{1,\ldots,n\} =: \mathscr{C}$. We denote by c the generic element of $\mathscr{C}$. The attribute variable represents some characteristics of the object that are related to the class. Variable A takes values in the set $\mathscr{A}$; we denote by a its generic element. As an example, objects might be patients; A would represent information about a patient, such as personal information as well as outcomes of medical tests; $\mathscr{C}$ would index the patient's possible diseases.

Usually, some values of (A,C) are sampled in an independent and identically distributed way according to a law that is not known a priori. The so-called *learning set* $\mathscr{L}$ records those values, which are also called *instances* of (A,C). The goal of classification is to learn from the learning set a function that maps attributes into classes. We call this function a (*precise*) *classifier*.

A classifier is applied to predict the class of new objects based on their attributes. Predictions are rewarded through a *reward matrix* $\mathbb{R}$. This is an $n \times n$ matrix whose generic element r_{ij} is a number representing the reward obtained by predicting class i when the actual class is j. Equivalently, we can regard the reward matrix as a set of *gambles* (i.e., bounded random variables) $\mathbb{R}_i$, $i = 1,\ldots,n$, each one corresponding to a row of $\mathbb{R}$: gamble $\mathbb{R}_i$ represents the uncertain reward obtained by predicting class i and is defined by $\mathbb{R}_i(j) := r_{ij}$, with $j \in \mathscr{C}$. The reward matrix is an input of the classification problem, in the sense that it is given.

In classification, at least with respect to the machine learning practice, rewards are usually measured in a linear utility scale: although this point is often left implicit, we can deduce it from the observation that the performance of a classifier is usually identified with its expected reward.

The most frequent practice consists also in using just a 0-1 valued reward matrix, which we denote by $\mathbb{I}$. In this case, the gamble corresponding to the i-th row of the matrix coincides with the indicator function of set $\{i\}$, which yields $\mathbb{I}_i(i) = 1$, and $\mathbb{I}_i(j) = 0$ for $i \neq j$. Accordingly, the performance of a classifier corresponds to the probability of predicting the actual class. Such a probability is called the *predictive accuracy* (or simply the *accuracy*) of a classifier.

The term 'accuracy' is used also for the sample estimate of such a probability. In fact, a classification problem usually comes with a test set $\mathscr{T}$. This set contains a number of sampled instances of (A,C) that are used to evaluate the classifier's predictive performance by measuring its accuracy on them. And in fact the predictive accuracy is by far the most frequently used empirical index to compare classi-

fiers, even though a careful elicitation of rewards would arguably lead in many cases to a reward matrix more general than $\mathbb{I}$. Such a widespread use has probably been favored by the simple interpretation of predictive accuracy; a more substantial reason could be that the predictive accuracy is particularly convenient to make extensive comparisons of classifiers over many data sets, which is a key component of the machine learning practice. Accordingly, in this paper we focus on the 0-1 valued reward matrix $\mathbb{I}$.

So far we have introduced the traditional view of classification, where the predictions issued by (precise) classifiers are made of single classes. This view has been generalized through the introduction of credal classifiers [13, 14]. A *credal classifier* is also a function learned from set $\mathscr{L}$, but it maps the attributes of an instance into a set $\mathscr{K} \subseteq \mathscr{C}$ of $k := |\mathscr{K}|$ classes in general. We call this a set-valued classification. We also say that the classification is *determinate* when $k = 1$, and *indeterminate* otherwise. When a classification is fully indeterminate, that is, when $\mathscr{K} = \mathscr{C}$, we call it *vacuous*. Similarly, the *vacuous classifier* is the one that always issues vacuous predictions. To each credal classifier it is possible to associate a determinate classifier that outputs predictions by choosing every time a class uniformly at random[1] from the output set $\mathscr{K}$ of the credal classifier. We call this the *$\mathscr{K}$-random classifier*; when the related credal classifier is the vacuous one, we just call it the *random classifier*.

Evaluating a credal classifier can be regarded as the problem of defining an 'extended' reward matrix, which associates a reward gamble to each non-empty subset of classes.

3 Introducing the Betting Framework

In order to make the comparison of credal classifiers as objective as possible, we introduce the idea of a betting framework. We define the framework for a traditional problem of classification, where classifiers issue determinate predictions. In Section 4 we will extend the framework to credal classification.

In the framework under consideration, we have two classifiers, which we would like to compare, that have already been inferred from data (so that there is no further learning, only an evaluation stage). These classifiers are regarded as bettors. Bets correspond to instances of the problem of classification: a bet is set up by sampling an instance of the problem. Classifiers are required to bet by predicting the actual class of the instance, and are rewarded according to matrix $\mathbb{I}$. The process is repeated for ever, and the performance of classifiers is taken to be their predictive accuracy.

Let us make the betting framework more precise by describing the two types of actors that play a role there:

Bettors: each of the two classifier we aim at comparing is regarded as a bettor.

House: rewards are delivered to bettors by an artificial entity that we call House. House only accepts determinate bets, which are rewarded according to matrix $\mathbb{I}$.

These actors are characterized by clarifying their relationship with the rewards, that is, with the utility scale involved. To start with, based on the discussion made in Section 2, we can readily state our first assumption concerning the betting framework:

(A1) Utility of bettors is linear in the rewards.

This assumption simply states explicitly what is current practice in classification.

The second assumption concerns House. We want to model House as an agent whose only aim is to reward correct predictions. In other words, House should not introduce any subjective bias in the process of rewarding bettors because of a risk-averse or risk-seeking attitude; it should just be risk-neutral:

(A2) Utility of House is linear in the rewards.

4 Betting with Credal Classifiers

Now we would like to extend the betting framework to credal classifiers. The crucial point here is that House only accepts determinate bets, while a credal classifier outputs set-valued classifications in general. Therefore, if we want to allow a credal classifier to play, we should find a way to extend the reward matrix to set-valued classifications in a way that both House and bettor find acceptable.

The first step in this direction is to recognize that any negotiation between the credal classifier and House can be made only on the basis of determinate bets, which is the only language that House understands. In order to enable the credal classifier to play as a determinate bettor, we state the following assumption:

(A3) The credal bettor accepts betting on any single class from its set-valued prediction, if forced to make a determinate bet, and on no class outside that set.

This assumption is satisfied whenever the classes in the output set of the credal classifier are incomparable, and the other ones represent dominated options. This is the case when credal classifiers are obtained using sets of probabilities and decision criteria like maximality or e-admissibility

[1] Throughout the paper we use the word 'random' to mean *uniformly* random.

(see, e.g., [12, Section 3.9]). We state the assumption explicitly in order to allow the framework to be used also by credal classifiers created in a different way.

The next assumption formalizes the idea that the framework is run for ever:

(A4) Every possible bet is repeated infinitely many times in the betting framework by sampling the problem instances.

This assumption, together with the previous one, enable the credal classifier to actually adopt a randomized strategy over the k classes in its output set $\mathscr{K}$. A randomized strategy is a mass function $\sigma = (\sigma_i)_{i\in\mathscr{K}}$ that represents the (determinate) betting behavior of the credal classifier in the limit.

At this point House knows that the credal classifier has the freedom to implement any randomized betting strategy: this means that the credal classifier can actually force House to undergo any expected loss that can follow from the choice of the strategy.

Let us call a prediction $\mathscr{K}$ 'successful' if the actual class belongs to $\mathscr{K}$. We restrict the attention to successful predictions as they determine House's expected loss: in fact, an unsuccessful prediction always yields a zero loss, by definition of $\mathbb{I}$, irrespective of the randomized strategy adopted. Let $\theta = (\theta_j)_{j\in\mathscr{C}}$ be the vector of chances, that is, the population proportions, for the classes conditional on the prediction being successful (this means that $\theta_j = 0$ if $j \notin \mathscr{K}$). House's expected loss conditional on a successful predictions equals

$$\sum_{i\in\mathscr{K}} \sum_{j\in\mathscr{C}} \mathbb{I}_i(j)\sigma_i\theta_j = \sum_{i\in\mathscr{K}} \sigma_i\theta_i,$$

where we are assuming that the strategy is chosen independently of the chances.

The loss depends on σ, which is chosen by bettor, and on θ. The latter models the specific problem under consideration. But House knows that the betting system will be applied, in principle, to every possible problem. House should then be enabled to consider every possible scenario:

(A5) In the determination of the expected loss, House has the freedom to choose any value for θ.

At this point we are ready to derive the extended reward matrix (as described at the end of Section 2):

Theorem 1. *Let $\mathscr{K} \subseteq \mathscr{C}$ be a set-valued prediction made of k classes, $\mathbb{I}_{\mathscr{K}}$ be the indicator function of set $\mathscr{K}$, and j the actual class. The corresponding value in the extended reward matrix that is uniquely consistent with (A1)–(A5) is the discounted accuracy:*

$$\frac{\mathbb{I}_{\mathscr{K}}(j)}{k}. \tag{1}$$

Proof. If $\mathscr{K}$ is unsuccessful, then any randomized strategy will yield a zero loss. Let us focus on successful predictions. Let Δ be the $n-1$ probability simplex. We formulate the problem in a game-theoretic setting. The two players are just bettor and House. Bettor can choose $\sigma \in \Delta$, while House can choose $\theta \in \Delta$. What we get is a zero-sum game with a gain for bettor defined by $\sum_{i\in\mathscr{K}} \sigma_i\theta_i$. This is a continuous linear function in σ for all $\theta \in \Delta$, as well as in θ for all $\sigma \in \Delta$, and moreover Δ is a compact convex set. The minimax theorem (see, e.g., [10, Theorem 6.7.3]) allows us to deduce that there is an optimal solution to the game with expected reward equal to $\max_{\sigma\in\Delta} \min_{\theta\in\Delta} \sum_{i\in\mathscr{K}} \sigma_i\theta_i$. It is easy to see that that is equal to $1/k$: once a strategy σ is fixed, the minimum is achieved by setting $\theta_{i_*} := 1$ on any $i_* = \operatorname{argmin}_{i\in\mathscr{K}} \sigma_i$; then the problem becomes $\max_{\sigma\in\Delta} \min_{i\in\mathscr{K}} \sigma_i = 1/k$. The related optimal strategy σ^* is uniform, $\sigma^*_i := 1/k$ for all $i \in \mathscr{K}$; this means that bettor and House agree that credal bettor should act like the $\mathscr{K}$-random classifier.

Now remember that, according to (A1)–(A2), both bettor and House are risk-neutral. This means they agree that an unsuccessful prediction is rewarded by the certain value 0 and a successful one by the certain value $1/k$. This is achieved by setting the reward equal to the discounted accuracy. □

It is useful to comment on this result from a few different viewpoints.

One thing is that the discounted accuracy implements a kind of least-committal reward system for House, in the sense that House gives bettor only what is certainly due to it. In fact, if the credal bettor does implement strategy σ^*, the expected reward that it achieves is indeed $1/k$, irrespective of the chances. Therefore the established reward is what House knows already that bettor can make for sure. For the same reason, it would be implausible to expect that credal bettor accepts any smaller reward. It is also interesting to observe that playing as the $\mathscr{K}$-random bettor (i.e., classifier) is the only way for credal bettor to have a sure reward.

The next consideration is again based on the observation that credal bettor is evaluated exactly as the $\mathscr{K}$-random bettor. This has important implications for the comparison of classifiers through the discounted accuracy: the main point is that the $\mathscr{K}$-random bettor is actually taken as a baseline to compare classifiers. Consider, for the sake of explanation, a determinate classifier whose output class is always contained in that of a certain credal classifier. The determinate classifier will be evaluated better than the credal classifier as soon as it exploits, to any (even a very tiny) degree, the credal classifier's set of output classes better than the $\mathscr{K}$-random one. Looking at this from another side, it means that the credal classifier can be better than the determinate one only if the latter behaves worse than the $\mathscr{K}$-random classifier! In practical applications, this will imply that a credal classifier will almost *never* be superior to a determinate classifier whose output is included in the credal's one. This discussion should make clear that the discounted accuracy, although it is a reasonable criterion, is probably

the most unfavorable way (among the reasonable ones) to evaluate credal classifiers, as a credal classifier cannot do better than isolating a set of classes that is impossible to compare.

This points to an aspect of the evaluation that the discounted accuracy certainly fails to capture. Let us focus on the simplest possible setup, using the following example. You are trying to evaluate two physicians based on some recorded diagnostic performance of theirs. In your records, the first physician always issues a vacuous diagnosis, that is, the entire set $\mathscr{C}$ of possible diseases. The second always issues a determinate diagnosis. But when you measure the second physician's predictive accuracy, you realize that his predictions are random. In this case, the discounted accuracy values the two physicians the same: $1/n$. But it is clear that the first physician provides you with something more than the second, because, in a sense, he delivers what he promises. How to precisely value this 'something more' appears to be quite a subjective matter. In this sense, it should not be too surprising that discounted accuracy does not value it at all, as it has been created trying to keep subjectivity out of consideration. And yet, subjectivity matters, and should be taken into account. The next section shows that this can be done in a very natural way.

5 Comparing Credal Classifiers

We have two classifiers f, g. We focus on selecting the classifier whose expected performance in the next instance (i.e., next bet) is greater than the other's. In the previous section we have measured performance by discounted accuracy. In this section, we want to make the method of comparison more flexible by allowing subjectivity to enter the picture, so as to be able to deal with the issues discussed at the end of the previous section. To this end, we start identifying classifiers with gambles: gambles f and g yield the discounted-accuracy reward achieved by classifiers f and g, respectively, in the next instance. There is uncertainty about these gambles because we assume that the instance has yet to be sampled.

The comparison of gambles f and g needs a (rational) decision-maker, whom we call 'you'. By definition of the gambles, you will compare them based on discounted-accuracy rewards. We model your attitude towards these rewards through the following assumption:

(A6) Your utility function[2] $u(\cdot)$ is concave in the discounted-accuracy rewards,

which means that you are risk-averse, or at most neutral, in these rewards.[3]

This seems to be quite a reasonable assumption, at least in the common setup where the original rewards (the ones used to define the 0-1 reward matrix $\mathbb{I}$) are measured in a utility scale that is linear for you. In fact, imagine that you are explicitly asked to extend the reward matrix to take into account your attitude towards set-valued classifications. Can we say something about the values you would use to define such an extended matrix? On the one hand, we argue that the rewards you would put there should be greater than or equal to the discounted-accuracy rewards. This follows from the discussion at the end of Section 4, which shows that it would be unreasonable to use values smaller than the discounted accuracy. On the other hand, values strictly greater than that would be reasonable: these allow you to express a preference in favor of a set-valued classification in comparison to the related $\mathscr{K}$-random prediction. These considerations imply that your utility function is in general non-linear in the discounted-accuracy (that is, discounted accuracy can be regarded as defining a new utility scale out of the original one). We take your utility in particular to be concave to express a consistent preference for set-valued classifications in comparison to the related $\mathscr{K}$-random predictions (note that this includes the extreme case of a linear utility function, in which the two options are equally valued).

Going back to the comparison of classifiers, it follows immediately from (A6) and decision-theoretic arguments that you will choose the one with maximum expected utility: $h^* := \operatorname{argmax}_{h\in\{f,g\}} E[u(h)]$.

Re-consider the example of the vacuous and the random classifier, discussed at the end of Section 4, as they are emblematic of the differences that arise in the evaluation of credal and precise classifiers when using utility.

Proposition 2. *The random and the vacuous classifiers have the same expected reward on the next instance, but the expected utility of the vacuous is greater under any strictly concave utility function.*

Proof. Denote the random classifier by r, and the vacuous classifier by v. As usual, we identify the classifiers with the corresponding gambles, which represent uncertain discounted-accuracy rewards for the next bet. The vacuous classifier gets on any instance the deterministic reward $1/n$. Thus, under any utility function:

$$E[u(v)] = u\left(\frac{1}{n}\right) = u(E[v])\,.$$

The random classifier r samples the predicted class from $\mathscr{C}$ according to the uniform mass function σ^*, independently of the actual class. Let us denote, as usual, by $\theta = (\theta_j)_{j\in\mathscr{C}}$ the vector of chances for the actual classes. We obtain that

$$E[r] = \sum_{i\in\mathscr{C}}\sum_{j\in\mathscr{C}} \mathbb{I}_i(j)\sigma_i^*\theta_j = \sum_{i\in\mathscr{C}} \sigma_i^*\theta_i = 1/n\,.$$

[2] We assume that the usual regularity conditions for utility hold, and in particular that it is strictly increasing, and that it has first and second derivatives (see, e.g., [9]).

[3] Note that House is not affected by your entering the picture, as it keeps on delivering discounted accuracy rewards as before. What changes is the explicit introduction of a decision-maker and his perception of the value of these rewards, as modeled by your risk-aversion.

This shows that $E[v] = E[r]$. In addition, using Jensen's inequality leads to

$$E[u(r)] < u(E[r]) = u(1/n) = E[u(v)],$$

whenever u is a strictly concave function. □

To better analyze this point, it is useful to approximate the expected utility by a second-order Taylor series. Let h be a generic classifier (and hence, a gamble):

$$\begin{aligned} E[u(h)] &\simeq u(E[h]) + \overbrace{u'(E[h])E(h-E[h])}^{=0} + \\ &+ \frac{1}{2}u''(E[h])E\left[(E[h]-h)^2\right] = \\ &= u(E[h]) + \frac{1}{2}u''(E[h])Var[h], \end{aligned} \tag{2}$$

where u', u'' are the first and second derivative of the utility function, and $Var[h]$ denotes the variance of h. Well-known papers in finance [6, 8] have shown that this is a very accurate approximation.

Remember that $u''(E[h]) \leq 0$ for every concave utility function (moreover, $u''(\cdot)$ is related to the degree of risk aversion of the utility assessor). Therefore what Equation (2) tells us is that the expected utility increases by increasing the expectation of rewards and decreasing their variance. It is clear now why the vacuous classifier, with variance equal to zero, is preferred to the random one. In other words, the 'something more' that the vacuous classifier is providing is its inherent reliability in earning rewards, which, using discounted accuracy, has a very clear numerical counterpart in its variance. The value that you give to this is indeed personal, and is formalized through your utility function. In the extreme case when you are risk-neutral in the discounted-accuracy rewards, the value is zero, and in this case there seems to be little room for credal classifiers in your interests. Bigger values express stronger preferences for reliable predictions.

It is also interesting to briefly consider the case where you are risk-averse in the original rewards defining $\mathbb{I}$. This would most probably be the case if those rewards represented amounts of money. In particular, if the discounted-accuracy rewards were the actual money payed by a betting system, then you would be 'natively' risk-averse in them; as a natural byproduct, you would prefer the more reliable (i.e., less variable) credal classifier to its $\mathscr{K}$-random counterpart.

All the above considerations can be turned into a remarkably simple procedure to empirically compare credal classifiers in practice. Remember that in a classification problem we usually have a test set $\mathscr{T}$, that is, a collection of instances used to evaluate the performance of a classifier. We need to estimate $E[u(h)]$ for a certain classifier h. Let us denote by $\mathscr{U}$ the set of values that gamble $u(h)$ can take. Set $\mathscr{U}$ has $(2^n-1)\times n$ elements at most, as the values are in one-to-one correspondence with the elements of the reward matrix extended through discounted accuracy. If we estimate the chance of a value $u_h \in \mathscr{U}$ by its sample proportion $\#(u_h)/|\mathscr{T}|$ in the test set, we obtain:

$$E[u(h)] \simeq \sum_{u_h \in \mathscr{U}} u_h \frac{\#(u_h)}{|\mathscr{T}|} = \frac{1}{|\mathscr{T}|} \sum_{(a,c)\in\mathscr{T}} u(h(a,c)).$$

This is equivalent to evaluating the performance of a credal classifier using the $(2^n-1)\times n$ reward matrix obtained by applying function $u(\cdot)$ point-wise to the matrix extended through discounted accuracy. In other words, what is done in practice is to change the 'discounting' factor in the discounted accuracy by means of the concave utility function.

A final consideration is that the comparison can be, perhaps more conveniently, made also using $u^{-1}(E[u(h)])$, the so-called *certainty equivalent*. This brings the performance index back to the range $[0,1]$ so that it can still be interpreted as a predictive accuracy, although one that is distorted through the utility function.

6 The Case for an Objective Winner

Equation (2) is useful because it gives us a very accurate approximation to the expected utility while releasing us from having our considerations narrowed down by the specific form of the utility function considered. To this end, in the following, we will repeatedly refer to (2) as if it were our actual expected utility.

In particular, an interesting consideration suggested by Equation (2) is that in one case the comparison of classifiers can be done by minimizing subjective considerations: when the two classifiers have equal expected reward. In this case, the classifier with minimum variance wins under every strictly concave utility function: that is, no matter how tiny (but non-zero) is your degree of risk-aversion. This can be implemented in practice by defining a range where the difference of the expected rewards is deemed irrelevant, and estimating their variances from the test set.

In the following, we investigate whether we can relate the variance of a classifier with its *determinacy*, that is, with a measure of the amount of imprecision in the output. Intuitively, we expect such a relationship to exist because both measures are related to the reliability of a classifier, and moreover, we expect that larger indeterminacy corresponds to smaller variance.

The gamble h corresponding to a classifier's performance in the next bet can be decomposed in two other gambles h_D and h_I such that $h = h_D + h_I$ and $h_D h_I = 0$ (element-wise). Intuitively, h_D and h_I represent the rewards for f when it returns, respectively, a determinate and an indeterminate classification. The following relationships follow from the

decomposition under discounted accuracy:

$$E[h^2] = E[h_D^2] + E[h_I^2], E[h_D^2] = E[h_D], E[h_I] \geq E[h_I^2],$$

where in the last expression we have the equality only if $E[h_I] = E[h_I^2] = 0$, which implies that either h is a precise classifier or that indeterminate predictions of h contain the actual class with probability zero.

Let f and g denote two generic classifiers with the same expected discounted accuracy: $E[f] = E[g]$. Using the identities above, one can show that the difference of variances is thus

$$\Delta Var := Var[g] - Var[f] = E[g_D] + E[g_I^2] - E[f_D] - E[f_I^2]. \quad (3)$$

Let us start by considering the important case where we compare a credal classifier with a precise one:

Proposition 3. *Consider a credal classifier and a precise classifier with the same expected reward. Then the credal classifier is preferable to the precise classifier under any strictly concave utility function.*

Proof. Let us denote by f the credal classifier and by g the precise one. We know by Equation (2) that we prefer the classifier with smaller variance under any strictly concave utility function. Thus, it suffices to show that $\Delta Var \geq 0$. Since $E[f_I^2] \leq E[f_I]$, it follows from Equation (3) that $\Delta Var = E[g_D] - E[f_D] - E[f_I^2]$ so that

$$\Delta Var \geq E[g_D] - E[f_D] - E[f_I] = E[g] - E[f],$$

which equals zero, since f and g have equal expected reward. Note the inequality is strict (i.e., there is strict preference) if the credal classifier is not always determinate and its indeterminate predictions are successful with positive probability. □

Now, let H_D be the event that equals 1 when the generic classifier h is determinate on the next instance, and 0 otherwise. We define the *determinacy* of classifier h as the probability that h is determinate: $P(H_D)$. This definition allows us to settle the problem for the next case:

Proposition 4. *Consider two credal classifiers that are vacuous whenever they are indeterminate and that have the same expected reward. Then the more indeterminate classifier is preferable under any strictly concave utility function.*

Proof. Let us denote by f and g the two credal classifiers, assuming f to be more indeterminate than g: $P(G_D) > P(F_D)$. It suffices to show that $\Delta Var > 0$. Any generic classifier h that is vacuous whenever it is indeterminate is rewarded with $1/n$ for any indeterminate prediction. Hence,

$$E[h_I] = \frac{1 - P(H_D)}{n}, \qquad E[h_I^2] = \frac{E[h_I]}{n}.$$

From these identities and Equation (3) we have that

$$\begin{aligned} \Delta Var &= E[g_D] + E[g_I]/n - E[f_D] - E[f_I]/n \\ &= -E[g_I] + E[g_I]/n + E[f_I] - E[f_I]/n \\ &= \frac{n-1}{n}\left(-E[g_I] + E[f_I]\right) = \frac{n-1}{n^2}\left(P(G_D) - P(F_D)\right), \end{aligned}$$

which is strictly positive by the initial assumptions. □

This proposition is particularly useful as it allows us to solve the problem in the case of binary classification problems, where any indeterminate prediction is necessarily vacuous.

One might be tempted to think that the previous result extends to non-vacuous classifiers as well, that is, the more determinate a classifier the higher its variance (and therefore the less preferable it is). Unfortunately, this is not the case, as the following example shows.

Example 1. *Consider a three-class classification problem. Let H_k denote the event that equals 1 if the generic classifier h returns a set of k classes that contains the actual one, and 0 otherwise. Likewise, let H_k^c be the event that equals 1 if h outputs k incorrect classes, and 0 otherwise. Note that $\sum_{k=1}^3 H_k + H_k^c = 1$ and $H_3^c = 0$. We can define the relevant expectations in terms of H_k, H_k^c:*

$$P(D_h) = P(H_1) + P(H_1^c), \qquad E[h] = \sum_{k=1}^{3} \frac{1}{k} P(H_k),$$

$$E[h^2] = \sum_{k=1}^{3} \frac{1}{k^2} P(H_k), \qquad 1 = \sum_{k=1}^{3} P(H_k) + P(H_k^c).$$

Assume that $P(F_1) = P(G_1) + \varepsilon$, $P(G_1^c) = P(F_1^c) + 2\varepsilon$, $P(G_2) = P(F_2) + 2\varepsilon$, $P(F_2^c) = P(G_2^c) + 3\varepsilon$, and $P(F_3) = P(G_3)$, for some small $\varepsilon > 0$. Then we have from the identities above that $E[f] = E[g]$. Similarly, we have that $E[f^2] = E[g^2] + \frac{\varepsilon}{2}$. Hence, $\Delta Var = E[g^2] - E[f^2] < 0$, and g is preferred over f even though g is more determinate than f: $P(D_f) = P(D_g) - \varepsilon$.

Alternatively, we might measure the indeterminacy of a classifier h by the expected number of classes it outputs: $\sum_{k=1}^n k\left[P(H_k) + P(H_k^c)\right]$. Thus, in the example, we would have

$$\sum_{k=1}^{n} k\left[P(F_k) + P(F_k^c)\right] = \sum_{k=1}^{n} k\left[P(G_k) + P(G_k^c)\right] + 4\varepsilon,$$

and g is preferred over f even though the former has a smaller expected number of output classes than the latter. ♦

7 Comparison Over the Next m Bets

So far, we have considered the expected reward and utilities for the next *single* classification; this setting fits for instance the case of a patient, who asks a doctor for a diagnosis and who is concerned only about the utility generated by the very next classification (his diagnosis). Conversely, an online trader, who performs m trading operations every day, might accept to lose some money in the very next transaction, provided that the set of m transactions generated at the end of day has high enough utility. In this case, expected rewards and expected utilities should be computed over the next m bets. In the following, we compare the random classifier r and the vacuous classifier v on the next m bets; we denote by v_m and r_m the rewards of the vacuous and the random ones over the next m instances.

Gamble v_m has deterministic value m/n and thus:

$$E[u(v_m)] = u\left(\frac{m}{n}\right).$$

To compute $E[u(r_m)]$, let us consider that classifier r yields utility $u(\ell)$ when it correctly predicts ℓ outcomes in the next m bets; considering that classifier r issues a correct classification with probability $1/n$ (see Proposition 2), the probability of correctly predicting ℓ instances out of the next m is the binomial:

$$Bin(\ell, m, \frac{1}{n}) = \binom{\ell}{m} \frac{1}{n}^{\ell} \left(1 - \frac{1}{n}\right)^{m-\ell}.$$

The expected utility produced by the random classifier over the next m bets is thus:

$$E[u(r_m)] = \sum_{\ell=1}^{m} u(\ell) Bin(\ell, m, \frac{1}{n}). \tag{7}$$

It is not immediate to compare the expected utilities of the random and vacuous classifiers using Equation (7); a clear understanding can be obtained through the second-order approximation given by Equation (2). In the following, we analyze in this way the logarithmic and the exponential utility. The second-order approximation of both the logarithmic and the exponential utility is very good, having relative absolute error consistently smaller than 1%.

7.1 Logarithmic Utility

The logarithmic utility is $u(x) := \log(1+x)$, whence $u''(x) = -\frac{1}{(1+x)^2}$; applying Equation (2), we get:

$$u(E[r_m]) + \frac{1}{2}u''(E[r_m])Var(r_m) =$$
$$u(E[r_m]) - \frac{Var(r_m)}{2(E[r_m]+1)^2} =$$
$$u\left(\frac{m}{n}\right) - \frac{m\frac{1}{n}\left(1-\frac{1}{n}\right)}{2(\frac{m}{n}+1)^2},$$

where in the last passage we introduced the analytical expression of the variance for a binomial distribution.

Thus, the (approximated) difference between the expected utility of the random and the vacuous over the next m bets is

$$d(m) = E[u(v_m)] - E[u(r_m)] = \frac{\frac{m}{n}\left(1-\frac{1}{n}\right)}{2(\frac{m}{n}+1)^2} \propto \frac{m}{(\frac{m}{n}+1)^2}, \tag{8}$$

where in the last passage we removed the proportionality constant $\frac{1}{2n}\left(1-\frac{1}{n}\right) > 0$. Function $d(m)$ is shown in Fig. 1.

The first derivative of $d(m)$ is:

$$d'(m) = \frac{1}{(\frac{m}{n}+1)^2} - 2\frac{\frac{m}{n}}{(\frac{m}{n}+1)^3} \propto 1 - \frac{m}{n}, \tag{9}$$

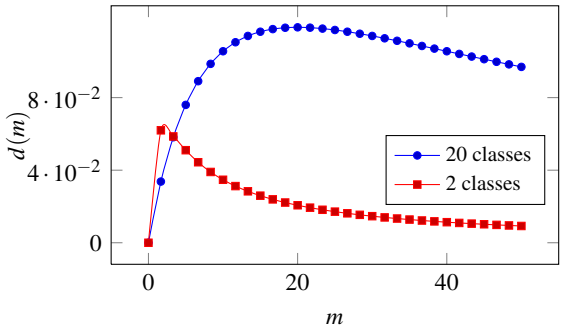


Figure 1: Function $d(m)$ for *logarithmic* utility, under different number of classes.

where the last passage is obtained considering that $(\frac{m}{n}+1)^3 > 0$. From Equations (8) and (9), we can figure out that $d(m)$ will monotonically increase up to $m < n$ (*inversion point*), to then indefinitely decrease, so that $d(m) \to 0$ for $m \to \infty$; if expectations of utilities are computed over a long enough number of bets, the expected utility produced by the two classifiers is the same. It also follows that increasing n delays the convergence of the expected utilities to the same value, as also shown in Fig. 1.

7.2 Exponential Utility

The exponential utility is $u(x) := 1 - \exp(-ax)$, where a is a coefficient of risk-aversion. Noting that $u''(x) = -a^2\exp(-ax)$, the second-order approximation yields:

$$u(E[r_m]) + \frac{1}{2}u''\left(\frac{m}{n}\right)Var(r_m) =$$
$$u\left(\frac{m}{n}\right) - \frac{1}{2}a^2\exp\left(-a\frac{m}{n}\right)m\frac{1}{n}\left(1-\frac{1}{n}\right),$$

whence

$$d(m) = -\frac{1}{2}a^2\exp\left(-a\frac{m}{n}\right)m\frac{1}{n}\left(1-\frac{1}{n}\right) \propto$$
$$\propto -\exp\left(-a\frac{m}{n}\right)m,$$

where the proportionality constant is $\frac{a^2}{2}\frac{1}{n}\left(1-\frac{1}{n}\right) > 0$.

We have

$$d'(m) = \exp\left(-a\frac{m}{n}\right)\cdot\left(a\frac{m}{n}-1\right).$$

Function $d(m)$ has qualitatively the same behavior of the logarithmic case, but the inversion point is now located at $m = \frac{n}{a}$. Moreover, the difference between the expected utility of the two classifiers depends also on the risk-aversion coefficient a; higher risk-aversion delays the convergence of the expected utilities, thus emphasizing the difference in favor of the vacuous on small m.

8 Experiments on Artificial Data Sets

In the following, we denote the naive Bayes classifier as NBC [3] and the naive credal classifier as NCC [1]. We compare the utility generated by NBC and NCC on the next *single* bet. In a first set of experiments, we generated artificial data sets, considering a binary class and 10 binary features; we set the marginal chances of classes as uniform, while we drew the conditional chances of the features under the constraint $|\theta_{i1\ell} - \theta_{i2\ell}| \geq 0.1\ \forall i, j$, where $\theta_{ij\ell}$ denotes the chance of feature A_i to be in state ℓ when $C = j$; the constraint forced each feature to be truly dependent on the class. We drew θ 80 times uniformly at random and we consider the sample sizes: $s \in \{25, 50, 100\}$. We did not consider larger sample sizes, under which NCC would have been almost completely determinate, and thus not really different from NBC. For each pair (θ, s) we generated 50 training sets; we then evaluate the trained classifiers on a test set of 10000 instances. In the following, the instances indeterminately classified by NCC are referred to as the *area of ignorance*. We denote as NBC(NCC-I) the accuracy of NBC on the area of ignorance. For each sample size, we thus perform $80\theta \times 50$ trials $= 4000$ training/test experiments.

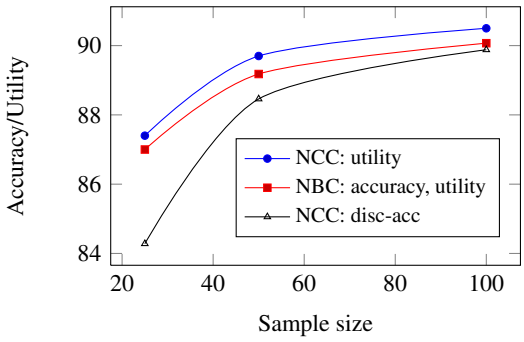


Figure 2: Experimental results with artificial data; each point shows the median over 4000 experiments, performed with the same sample size s. For NBC, accuracy and utility coincide. For NCC, the curve of utility rises the values of discounted accuracy.

We set the utility of a determinate and successful classification as $u(1) := 1$; the utility of a non-successful classification (determinate or indeterminate) as $u(0) := 0$. This is the case, for instance, if you are risk-neutral in the scale the original rewards are measured. It remains to set the utility $u(0.5)$ of an indeterminate classification (notice that for a data set with two classes, an indeterminate classification has necessarily discounted accuracy of 0.5). We think that in general the value of $u(0.5)$ could reasonably lie between 0.6 and 0.8; in our experiments, we set $u(0.5) := 0.65$. As a term of comparison, determinate and indeterminate classifiers have been compared in [7] through the F_1 metric, which is widely used in information retrieval. Under the F_1 metric, on a dataset with 2 classes, the vacuous classifier gets the same score of a precise classifier with 66% accuracy; this gives further support to our choice.

As expected, NBC has higher discounted accuracy than NCC (see Fig. 2); this means that, on the area of ignorance, it is doing better than the $\mathscr{K}$-random guesser. Yet, NCC produces slightly higher utility than NBC at each sample size. The determinacy of NCC rises steadily with the sample size; interestingly, at the same time the value of NBC(NCC-I) decreases; this means that NCC is getting better at identifying instances which are really hard to classify. For instance, NBC(NCC-I) is 64% for $s = 25$, and 54% for $s = 100$; this explains why the gap of utility tends to slightly increase with the sample size. Note however that the restriction of the area of ignorance (20% for $s = 25$, and only 4% for $s = 100$) works against enlarging the gap between NCC and NBC. Results similar to those shown here are obtained also using logarithmic utility; however we find it clearer in this simple setting to reason about the only point to elicit, $u(0.5)$, rather about the whole utility function.

9 Experiments on the kr-kp Data Set

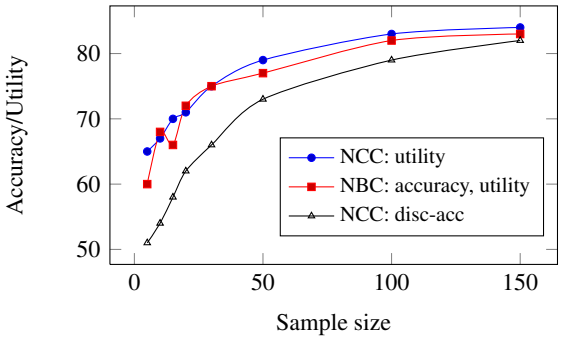


Figure 3: Utility and discounted accuracy generated by NBC and NCC for downsampled versions of kr-kp.

We then performed some experiments on the kr-kp data set (2 classes, 36 binary features, 3200 instances) from the UCI repository. To evaluate the sensitivity of the performance on the sample size, we worked by *downsampling* the kr-kp data set. In particular, we generated training sets of size $s \in \{5, 10, 15, 20, 30, 50, 100, 150\}$; for each sample size, we generated 100 different training sets; for each training set, the corresponding test set is given by the instances left in the original data set. All training and test sets are *stratified*, namely the proportion among the two classes matches that of the original data set. For each sample size, we report the average results over all splits; the results are shown in Tab. 1 and Fig. 3. The determinacy of NCC steadily increases with the sample size, as well its discounted accuracy and the accuracy of NBC. For NBC, notice that accuracy and utility have the same value. For very small s

s	NCC: Determ (%)	NBC: NBC(NCC-I) (%)
5	2	59
10	10	65
15	25	60
20	29	64
30	41	64
50	60	62
100	78	60
150	85	59

Table 1: Results for the kr-kp experiment; Determ. indicates the % of instances determinately classified by NCC.

(e.g., $s = 5$), NCC is almost always indeterminate; in this case, its utility corresponds to $u(0.5)$ and thus is 0.65; in the same situation, NBC is almost randomly guessing, and thus its utility is close to 50%. Both the utility of NBC and NCC smoothly increases with s; the utility of NCC remains however slightly superior. In fact, under a data set with two classes, whether NCC or NBC produces a higher utility can be realized by comparing NBC(NCC-I) with $u(0.5)$; if $u(0.5) <$NBC(NCC-I), then NCC produces higher utility than NBC, and vice versa. However, the outcome of the comparison would be slightly in favor of NBC by (conservatively) setting $u(0.5) = 0.6$, as can be deduced from Tab. 1; in fact, once utility is introduced in the evaluation of the classifiers, it also plays a role in the final decision about which of the considered classifiers is better. This also implies that to generate sensible results when using utility-based metrics, it is fundamental to *carefully* elicit the decision maker's utility.

10 Conclusions

In this paper, we have tried to define in a principled way a measure to empirically evaluate credal classifiers. In our proposal, any such measure is made of two main components: the discounted accuracy, which represents a kind of objective performance of a classifier, and its variance, which represents the unreliability of the classifier, and whose contribution to the overall measure has to be weighted through subjective considerations of risk-aversion. Our measure can be implemented very easily in practice, and in fact is shown to empirically lead to some interesting results. Future work could (i) explore generalizations to rewards more general than 0-1 ones; (ii) exploit what appear to be natural connections between our measure and finance, in order to evaluate credal classifiers (some recent work connecting utility and machine learning, that could be useful to consider in that respect, has also recently appeared [5]); and also (iii) deepen the empirical study in order to verify the possibility to define some kind of 'general purpose' utility functions for machine learning aims.

Acknowledgements

The research in this paper has been partially supported by the Swiss NSF grants n. 200020_134759 / 1, 200020-121785 / 1, 200020-132252 and by the Hasler foundation grant n. 10030.

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] G. Corani and M. Zaffalon. Lazy naive credal classifier. In J. Pei, L. Getoor, and A. de Keijzer, editors, *First ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data*, pages 30–37. ACM, 2009.

[3] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

[4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, 2001. 2nd edition.

[5] C. Friedman and S. Sandow. *Utility-Based Learning from Data*. Chapman & Hall/CRC, Boca Raton, Fl, 2011.

[6] W. Hlawitschka. The empirical nature of Taylor-series approximations to expected utility. *The American Economic Review*, 84(3):713–719, 1994.

[7] J. Jose del Coz and A. Bahamonde. Learning nondeterministic classifiers. *Journal of Machine Learning Research*, 10:2273–2293, 2009.

[8] H. Levy and H.M. Markowitz. Approximating expected utility by a function of mean and variance. *The American Economic Review*, 69(3):308–317, 1979.

[9] D. G. Luenberger. *Investment Science*. Oxford University Press, New York, 1998.

[10] J. Stoer and C. Witzgall, editors. *Convexity and Optimization in Finite Dimensions*. Springer-Verlag, Berlin, 1970.

[11] G. Tsoumakas and I. Vlahavas. Random k-label sets: an ensemble method for multilabel classification. In J. N. Kok, J. Koronacki, R. López de Mántaras, S. Matwin, D. Mladenic, and A. Skowron, editors, *Proceedings of ECML 2007, 18th European Conference on Machine Learning*, volume 4701 of *Lecture Notes in Computer Science*, pages 406–417. Springer, 2007.

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

[13] M. Zaffalon. A credal approach to naive classification. 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*, pages 405–414, Universiteit Gent, Belgium, 1999.

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

Index