Jean-Marc Bernard, Teddy Seidenfeld, & Marco Zaffalon
(Editors)

ISIPTA ’03

Proceedings of the Third International Symposium on
Imprecise Probabilities and Their Applications
Lugano, Switzerland

CARLETON SCIENTIFIC

PROCEEDINGS IN INFORMATICS 18

Published by: Carleton Scientific
Cover design by: Steve Izma
Printed and bound in Canada

Canadian Cataloguing in Publication (pending)

International Symposium on Imprecise Probabilities and Their Applications (3rd : 2003 : Lugano, Switzerland)
ISIPTA ’03 : proceedings of the 3rd International Symposium on Imprecise Probabilities and Their Applications, Lugano, Switzerland)

(Proceedings in informatics ; 18)
Conference held Jul. 14-17, 2003.
Includes bibliographical references.
ISBN 1-894145-17-8

1. Computer networks–Congresses. 2. Electronic data processing–Distributed processing–Congresses. I. Bernard, Jean-Marc II. Title. III. Series.

This book was processed using the $\LaTeX\,2_\varepsilon$ macro package with the Carleton Scientific style package, cs.sty.

Table of Contents

Preface

The ISIPTA meetings are one of the primary international forums to present and discuss new results on the theory and applications of imprecise probabilities. Imprecise probability has a wide scope, being a generic term for the many mathematical or statistical models that measure chance or uncertainty without sharp numerical probabilities. These models include belief functions, Choquet capacities, comparative probability orderings, convex sets of probability measures, fuzzy measures, interval-valued probabilities, possibility measures, plausibility measures, and upper and lower expectations or previsions. Imprecise probability models are needed in inference problems where the relevant information is scarce, vague or conflicting, and in decision problems where preferences may also be incomplete.

A total of 44 papers were presented at ISIPTA '03, covering a wide range of topics, including: new model based inference with imprecise probabilities; computations and foundations of inference with imprecise probabilities; applications of imprecise probabilities in engineering, finance, and medicine; connections with graph theory, belief functions, and fuzzy random variables; and the introduction of new principles and tools for decision theory.

To help promote the exchange of novel ideas, at ISIPTA '03 we continued the conference format begun at ISIPTA '01. Each of the 44 papers were presented both in a 20 minute plenary overview and as part of a poster session. Authors presenting at the plenary sessions were encouraged to use some of their 20 minutes to identify the context for their research, in addition to giving an overview of the research paper appearing in the proceedings. In this way the poster sessions serve, again, as a forum for extended discussions of the papers.

This ISIPTA meeting included three invited contributions from Terrence L. Fine, Irving J. Good, and Patrick Suppes. Copies of their papers were distributed at the conference and are included in the electronic version of the proceedings.

Each paper appearing in these printed proceedings has been the subject of a careful refereeing process. We feel confident that the selection process has resulted in a symposium and proceedings with contributions displaying both very high quality and unusual originality.

We want to thank the contributors for their diligence in preparing their submissions, and for their patience with our selection process. The Program Committee members discharged their refereeing responsibilities effectively and constructively, with the result that all the papers have been improved by the reviewing process. We are grateful to the tutorial leaders (Jean-Marc Bernard: "Imprecise Dirichlet model for multinomial data," Gert de Cooman: "A gentle introduction to imprecise probability models and their behavioral interpretation," Fabio G. Cozman: "Graph-theoretical models for multivariate modeling with imprecise probabilities," Charles F. Manski: "Partial identification of probability distributions,” Sujoy Mukerji: “Imprecise probabilities and ambiguity aversion in economic modeling”), who have contributed their time and talents in such a generous fashion. As with ISIPTA ’01, the Board is exceptionally thankful to Serafín Moral, who has overseen the electronic management of these papers, their submissions and reviews, and who has tirelessly given his time to fixing the inevitable breakdowns that happen in a complicated web-based system. These proceedings simply would not be possible without his expertise.

Jean-Marc Bernard
Teddy Seidenfeld
Marco Zaffalon

ISIPTA ’03 Sponsors

AntOptima
http://www.antoptima.com/

City of Lugano
http://www.lugano.ch/

Dipartimento dell’educazione, della cultura e dello sport del Canton Ticino
http://www.ti.ch/DECS

FINRISK
National Centre of Competence in Research
Financial Valuation and Risk Management

The research programme NCCR FINRISK is gratefully acknowledged.
http://www.nccr-finrisk.unizh.ch/

FNSNF
Swiss National Science Foundation

Grants: 20CO21-101209 and, in part, 2100-067961.02.
http://www.snf.ch/

ISIPTA ’03

Program Committee Board

Jean-Marc Bernard, France
Teddy Seidenfeld, USA
Marco Zaffalon, Switzerland

Program Committee Members

Thomas Augustin, Germany
Salem Benferhat, France
David V. Budescu, USA
Frank P. A. Coolen, UK
Fabio G. Cozman, Brazil
Luis M. de Campos, Spain
Gert de Cooman, Belgium
Dieter Denneberg, Germany
James M. Dickey, USA
Didier Dubois, France
Love Ekenberg, Sweden
Enrico Fagiuoli, Italy
Scott Ferson, USA
Terrence L. Fine, USA
Itzhak Gilboa, Israel
Angelo Gilio, Italy
Michel Grabisch, France
Joseph Y. Halpern, USA
David Harmanec, Czech Republic
Manfred Jaeger, Germany
Jean-Yves Jaffray, France
Edi Karni, USA
Etienne E. Kerre, Belgium
Gernot Kleiter, Austria
George J. Klir, USA
Jürg Kohlas, Switzerland
Igor Kozine, Denmark
Henry Kyburg, USA
Isaac Levi, USA
Thomas Lukasiewicz, Italy
Fabio Maccheroni, Italy
Charles F. Manski, USA
Alfio Marazzi, Switzerland
Glen Meeden, USA
Kalled Mellouli, Tunisia
Serafín Moral, Spain
Sujoy Mukerji, UK
Robert Nau, USA
Klaus Nehring, USA
John Norton, USA
Michael Oberguggenberger, Austria
Endre Pap, Yugoslavia
Andrzej Pownuk, Poland
Henri Prade, France
Urho Pulkkinen, Finland
Marco Ramoni, USA
Giuliana Regoli, Italy
Peter Reichert, Switzerland
Luca Rigotti, USA
David Rios Insua, Spain
Fabrizio Ruggeri, Italy
Paola Sebastiani, USA
Glenn Shafer, USA
Prakash P. Shenoy, USA
Philippe Smets, Belgium
Michael Smithson, Australia
Paul Snow, USA
Claudio Sossai, Italy
Wynn Stirling, USA
Milan Studeny, Czech Republic
Fabio Trojani, Switzerland
Lev V. Utkin, Russia
Paolo Vanini, Switzerland
Barbara Vantaggi, Italy
Jirina Vejnarova, Czech Republic
Paolo Vicig, Italy
Frans Voorbraak, The Netherlands
Kurt Weichselberger, Germany
Nic Wilson, UK
Robert L. Wolpert, USA

Organizing Institutions

IDSIA (USI-SUPSI): Dalle Molle
Institute for Artificial Intelligence
(www.idsia.ch)
IFin USI: Institute of Finance, Lugano
(www.lu.unisi.ch/istfin)
SIPTA: Society for Imprecise
Probability Theory and Applications
(www.sipta.ch)
SUPSI: University of Applied
Sciences of Southern Switzerland
(www.supsi.ch)
USI: University of Lugano
(www.unisi.ch)

Local Organization

Alessandro Antonucci
Marco Zaffalon

Electronic Organization

Serafín Moral

Steering Committee

Jean-Marc Bernard
Gert de Cooman
Serafín Moral
Teddy Seidenfeld
Marco Zaffalon

Maximum of Entropy in Credal Classification*

J. ABELLÁN
Universidad de Granada, Spain

S. MORAL
Universidad de Granada, Spain

Abstract

We present an application of the measure of maximum entropy for credal sets: as a branching criterion for classification trees based on imprecise probabilities. We also justify the use of maximum entropy as a global uncertainty measure for credal sets, and a deduction of this measure, based on the best lower expectation of the logarithmic score, is presented. We have also carried out several experiments in which credal classification trees are built taking a global uncertainty measure as a basis. The results show that there is a lower degree of error when maximum entropy is used as a global uncertainty measure.

Keywords

imprecise probabilities, uncertainty, maximum entropy, imprecision, non-specificity, classification, classification trees, credal sets

1 Introduction

Classification is an important problem in the area of machine learning in which classical probability theory has been extensively used. Basically, we have an incoming set of observations, called the training set, and we want to obtain a set of rules to assign a value of the variable to be classified to any new case. The set used to assess the quality of this set of rules is also called the test set. Classification has notable applications in medicine, recognition of hand-written characters, astronomy, banks, etc. The learned classifier can be represented as a Bayesian network, a neural network, a classification tree, etc. These methods normally use the Theory of Probability to estimate the parameters with a stopping criterion to limit the complexity of the classifier and to avoid overfitting.

*This work has been supported by the Spanish Ministry of Science and Technology, project Elvira II (TIC2001-2973-C05-01).

In some previous papers [4, 5, 6], we have introduced a new procedure to build classification trees based on the use of imprecise probabilities. Classification trees have their origin in Quinlan’s ID3 algorithm [18], and a basic reference is the book by Breiman et al. [8]. We also applied decision trees for classification, but as in Zaffalon [25], the imprecise Dirichlet model is used to estimate the probabilities of belonging to the respective classes defined by the variable to be classified. In classical probabilistic approaches, information gain is used to build the tree, but then other procedures must subsequently be used to prune it, since information gain tends to build structures which are too complex. We have shown that if imprecise probabilities are used and the information gain is computed by measuring the total amount of uncertainty of the associated credal sets (a closed and convex set of probability distributions), then the problem of overfitting disappears and results improve.

In Abellán and Moral [1, 2, 3], we studied how to measure the uncertainty of a credal set by generalizing the measures used in the Theory of Evidence, Dempster [10] and Shafer [20]. We considered two main sources of uncertainty: entropy and non-specificity. We proved that the proposed functions verify the most basic properties of these types of measures (Abellán and Moral [2], Dubois and Prade [12], Klir and Wierman [15]).

We previously proved that by using a global uncertainty measure which is the result of adding an entropy measure and a non-specificity measure, classification results are better than those obtained by the C4.5 classification method, based on Quinlan’s ID3 algorithm. In this paper, we have carried out some experiments in which the maximum entropy of the probability distributions of a credal set is used to measure its uncertainty, and we show that the results obtained are even better. We consider two methods of building classification trees. In the first method, Abellán and Moral [4], we start with an empty tree and in each step, a node and a variable are selected for branching which give rise to a greater decrease in the final entropy of the variable to be classified. In classical probability, a branching always implies a decrease in the entropy. It is necessary to include an additional criterion so as not to create models which are too complex and therefore overfit the data. With credal sets, a branching will produce a lower entropy but, at the same time, a greater non-specificity. Under these conditions, we follow the same procedure as in probability theory, but measuring the total uncertainty of a branching. The stopping criterion is very simple: when every possible branching produces an increment of the total uncertainty.

Finally, in order to carry out the classification given a set of observations, we use a strong dominance criterion to obtain the value of the variable to be classified and a maximum frequency criterion when we want to classify all the cases.

The extended method quantifies the uncertainty of each individual variable in each node in the same way, but also considers the results of adding two variables at the same time. In this way, we aim to discover relationships involving more than two variables that were not seen when investigating the relationships of a single variable with the variable to be classified.

In Section 2, we present the necessary previous concepts on uncertainty on credal sets. We place special emphasis on the maximum of entropy as a global uncertainty measure. In Section 3, we introduce the necessary notation and definitions for our procedure of building classification trees. In Section 4, we describe the methods based on imprecise probabilities. In Section 5, we test our procedure with known data sets used in classification by comparing the use of two global uncertainty measures.

2 Total Uncertainty on Credal Sets

Dempster-Shafer's theory is based on the concept of basic probability assignment (bpa), and it defines a special type of credal set [10, 20]. In this theory, Yager [24] distinguishes two types of uncertainty: one is associated with cases where the information is focused on sets with empty intersections; and the other is associated with cases where the information is focused on sets with a greater than one cardinality. We call these *randomness* and *non-specificity*, respectively. In Abellán [6] we justify that a general convex set of probability distributions (a credal set) may contain the same type of uncertainty as a bpa: we consider similar randomness and non-specificity measures.

In Abellán and Moral [2], we define a measure for non-specificity for convex sets that generalizes Dubois and Prade's measure of non-specificity in the theory of evidence [11]. Using the Möbius inverse function for monotonic capacities [9], we can define:

Definition 1 *Let $\mathcal{P}$ be a credal set on a finite set X. We define the following capacity function,*

$$f_{\mathcal{P}}(A) = \inf_{P \in \mathcal{P}} P(A), \ \forall A \in \wp(X),$$

where $\wp(X)$ is the power set of X. This function is also known as the minimum lower probability which represents $\mathcal{P}$.

Theorem 1 *(Shafer [20]) For any mapping $f_{\mathcal{P}} : \wp(X) \to \boldsymbol{R}$ another mapping $m_{\mathcal{P}} : \wp(X) \to \boldsymbol{R}$ can be associated by*

$$m_{\mathcal{P}}(A) = \sum_{B \subseteq A} (-1)^{|A-B|} f_{\mathcal{P}}(B), \forall A \in \wp(X),$$

Where $|A - B|$ is the cardinal of the set $A - B$. This correspondence is one-to-one, since conversely, we can obtain

$$f_{\mathcal{P}}(A) = \sum_{B \subseteq A} m_{\mathcal{P}}(B), \ \forall A \in \wp(X).$$

These functions, $f_{\mathcal{P}}$ and $m_{\mathcal{P}}$, are Möbius inverses.

Definition 2 *Let $\mathcal{P}$ be a credal set on a frame X, $f_{\mathcal{P}}$ its minimum lower probability as in Definition 1 and let $m_{\mathcal{P}}$ be its Möbius inverse. We say that function $m_{\mathcal{P}}$ is an assignment of masses on $\mathcal{P}$. Any $A \in X$ such that $m_{\mathcal{P}}(A) \neq 0$ will be called a focal element of $m_{\mathcal{P}}$.*

We can now define a general function of non-specificity.

Definition 3 *Let $\mathcal{P}$ be a credal set on a frame X. Let $m_{\mathcal{P}}$ be its associated assignment of masses on $\mathcal{P}$. We define the following function of non-specificity on $\mathcal{P}$:*

$$IG(\mathcal{P}) = \sum_{A \subset X} m_{\mathcal{P}}(A) \ln(|A|).$$

In Abellán and Moral [3], we proposed the following measure of randomness for general credal sets:

$$G^*(\mathcal{P}) = Max \left\{ -\sum_{x \in X} p_x \ln p_x \right\},$$

where the maximum is taken over all probability distributions on $\mathcal{P}$, and $\mathcal{P}$ is a general credal set. This measure generalizes the classical Shannon's measure [21] verifying similar properties. It can be used either as one of the components of a measure of total uncertainty, or as a total uncertainty measure, Harmanec and Klir [14]. We have proved that this function is also a good randomness measure for credal sets and possesses all the basic properties required in Dempster-Shafer's theory [3].

We define a measure of total uncertainty as $TU(\mathcal{P}) = G^*(\mathcal{P}) + IG(\mathcal{P})$. This measure could be modified by the factor introduced in Abellán and Moral [1], but this will not be considered here, due to its computational difficulties (it is a supremum that is not easy to compute). The properties of this measure are studied in Abellán and Moral [2, 3] and these are similar to the properties verified by total uncertainty measures in Dempster-Shafer's theory [17].

In this paper, we shall also consider $G^*(\mathcal{P})$ as a measure of total uncertainty. In the particular case of belief functions, Harmanec and Klir [14] consider that maximum entropy is a measure of total uncertainty. They justify it by using an axiomatic approach: it possesses some basic properties. However, uniqueness is not proved. But perhaps the most compelling reason is given in Walley's book [22]. Walley calls this measure the upper entropy. We start by explaining the case of a single probability distribution, P. If You are subject to the logarithmic scoring rule, that means that You are forced to select a probability distribution Q on X that if the true value is x, then You must pay $-\log(Q(x))$. For example, if You say that $Q(x)$ is very small and finally x is the true value, You must pay a lot. If $Q(x)$ is close to one, then you must pay a small amount. Of course, You should choose Q so that $E_P[-\log(Q(x))]$ is minimum, where E_P is the mathematical expectation with respect to P. This minimum is obtained when $Q = P$ and the value of $E_P[-\log(P(x))]$ is the entropy: the expected loss or the amount that You could accept to be subject to the logarithmic scoring rule. In the case of a credal set, $\mathcal{P}$, we can also have the logarithmic scoring rule, but now we choose Q in such a way that the upper loss $E^*_{\mathcal{P}}[-\log(Q(x))]$ (the supremum of the expectations with respect to the probabilities in $\mathcal{P}$) is minimum. Walley shows that this minimum is obtained for the distribution $P_0 \in \mathcal{P}$ with maximum entropy. Furthermore, $E^*_{\mathcal{P}}[-\log(P_0(x))]$ is equal to the maximum entropy in $\mathcal{P}$: $G^*(\mathcal{P})$. This is the minimum payment You require before being subject to the logarithmic scoring rule. This argument is completely analogous with the probabilistic one, except that we change the expectation for the upper expected loss. This is really a measure of uncertainty, as the better we know the true value of x, then the less we should need to accept the logarithmic scoring rule (lower value of $G^*(\mathcal{P})$). We are not saying that $\mathcal{P}$ can be replaced by the distribution of maximum entropy, only that its uncertainty can be measured by considering maximum entropy in the credal set.

3 Notation and Previous Definitions

For a classification problem we shall consider that we have a data set $\mathcal{D}$ with values of a set $\mathcal{L}$ of discrete and finite variables $\{X_i\}_1^n$. Each variable will take values on a finite set $\Omega_{X_i} = \{x_i^1, x_i^2, ..., x_i^{|\Omega_{X_i}|}\}$. Our aim will be to create a classification tree on the data set $\mathcal{D}$ of one target variable C, with values in $\Omega_C = \{c^1, c^2, ..., c^{|\Omega_C|}\}$.

Definition 4 *A configuration of $\{X_i\}_1^n$ is any m-tuple*

$$(X_{r_1} = x_{r_1}^{t_{r_1}}, X_{r_2} = x_{r_2}^{t_{r_2}}, ..., X_{r_m} = x_{r_m}^{t_{r_m}}),$$

where $x_{r_j}^{t_{r_j}} \in \Omega_{r_j}$, $j \in \{1, ..., m\}$, $r_j \in \{1, ..., n\}$ and $r_j \neq r_h$ with $j \neq h$. That is, a configuration is an assignment of values for some of the variables in $\{X_i\}_1^n$.

If $\mathcal{D}$ is a data set and σ is a configuration, then $\mathcal{D}[\sigma]$ will denote the subset of $\mathcal{D}$ given by the cases which are compatible with configuration σ (cases in which the variables in σ have the same values as the ones assigned in the configuration).

Definition 5 *Given a data set and a configuration σ of variables $\{X_i\}_1^n$ we consider the credal set $\mathcal{P}_C^\sigma$ for variable C with respect to σ defined by the set of probability distributions, p, such that*

$$p_j \in \left[\frac{n_{c^j}^\sigma}{N+s}, \frac{n_{c^j}^\sigma + s}{N+s}\right],$$

for every $j \in \{1,...,|\Omega_C|\}$, where for a generic state $c^j \in \Omega_C$, $n^{\sigma}_{c^j}$ is the number of occurrences of $\{C = c^j\}$ in $\mathcal{D}[\sigma]$, N is the number of cases in $\mathcal{D}[\sigma]$, and $s > 0$ is a parameter.

We denote this interval as

$$\left[\overline{P}(c^j|\sigma), \underline{P}(c^j|\sigma)\right].$$

This credal set is the one obtained on the basis of the imprecise Dirichlet model, Walley [23], applied to the subsample $\mathcal{D}[\sigma]$.

The parameter s determines how quickly the lower and upper probabilities converge as more data become available; larger values of s produce more cautious inferences. Walley [23] suggests a candidate value for s between $s = 1$ and $s = 2$, but no definitive statement is given.

4 Classification Procedure

We have proposed two methods to build a classification tree: the simple method [4] and the double method [5]. Here we describe the double procedure and give the simple as a particular case.

A classification tree is a tree where each interior node is labeled with a variable of the data set X_j with a child for each one of its possible values: $X_j = x_j^t \in \Omega_{X_j}$. In each leaf node, we shall have a credal set for the variable to be classified, $\mathcal{P}_C^{\sigma}$, as defined above, where σ is the configuration with all the variables in the path from the root node to this leaf node, with each variable assigned to the value corresponding to the child followed in the path. We use a measure of total uncertainty to determine how and when to carry out a branching of the tree. The method starts with a tree with a single node, which will have an empty configuration associated. This node will be open. In this node the set of variables $\mathcal{L}^*$ is equal to the list of variables in the database.

I. For each open node already generated, we compute the total uncertainty of the credal set associated with the configuration, σ, of the path from the root node to that node: $TU(\mathcal{P}_C^{\sigma})$. Then we calculate the values of α and β with

$$\alpha = \min_{X_i \in \mathcal{L}^*} \left(\sum_{r \in \{1,..,\left|\Omega_{X_i}\right|\}} \rho_{\{x_i^r\}|\sigma} TU(\mathcal{P}_C^{\sigma \cup (X_i = x_i^r)}) \right)$$

$$\beta = \min_{X_i, X_j \in \mathcal{L}^*} \left(\sum_{r \in \{1,..,\left|\Omega_{X_i}\right|\}, t \in \{1,..,\left|\Omega_{X_j}\right|\}} \rho_{\{x_i^r, x_j^t\}|\sigma} TU(\mathcal{P}_C^{\sigma \cup (X_i = x_i^r, X_j = x_j^t)}) \right),$$

where $\mathcal{L}^*$ is the set of variables of the data set minus those that appear on the path from the actual node to the root node, $\rho_{\{x_i^r\}|\sigma}$ is the relative frequency with which X_i takes the value x_i^r in $\mathcal{D}[\sigma]$, $\rho_{\{x_i^r,x_j^t\}|\sigma}$ is the relative frequency with which X_i and X_j take values x_i^r and x_j^t, respectively, in $\mathcal{D}[\sigma]$, and $\sigma \cup (X_i = x_i^r)$ is the result of adding the value $X_i = x_i^r$ to configuration σ (analogously for $\sigma \cup (X_i = x_i^r, X_j = x_j^t)$).

II. If the minimum of $\{\alpha, \beta\}$ is greater or equal than $TU(\mathcal{P}_C^\sigma)$ (including the case in which $\mathcal{L}^*$ is empty), then the node is closed and the credal set $\mathcal{P}_C^\sigma$ is assigned to it.

III. If the minimum of $\{\alpha, \beta\}$ is smaller than $TU(\mathcal{P}_C^\sigma)$, then if $\alpha \leq \beta$, we choose the variable that attains the minimum in α as branching variable for this node; and if $\alpha > \beta$ we consider the pair of variables X_i, X_j for which the value of β is attained, and select as branching variable that from X_i, X_j with a minimum value of uncertainty (calculated in an individual way as in α computation).

If X_{i_0} is the branching variable we add to this node a child for each one of its possible values. All the children are open nodes.

The simple method does not need β, Abellán and Moral [4]. It only considers α and it carries out a branching if this value is less than or equal to the uncertainty of the actual node ($TU(\mathcal{P}_C^\sigma)$). As above, the branching variable is the one for which the value α is attained. In the double method, we demand that the uncertainty is reduced. However, the double method looks for relationships of two variables with C at the same time. The simple method only considers the information of a single variable about C. In some cases, some multidimensional relationships do not give rise to pairwise relationships between the implied variables, and then they will not be detected by the simple method.

4.1 Decision in the Leaves

In order to classify a new case with observations of all the variables except in the variable to be classified C, we start at the root of the tree and follow the path corresponding to the observed values of the variables in the interior nodes of the tree, i.e. if we are at a node with variable X_i and this variable takes the value x_i^r in this particular case, then we choose the child corresponding to this value. This process is followed until we arrive at a leaf node. We then use the associated credal set about C, $\mathcal{P}_C^\sigma$, to obtain a value for this variable.

We will use a **strong dominance criterion** on C. This criterion generally implies only a partial order, and in some situations, no possible precise classification can be done. We will choose an attribute of the variable $C = c^h$ if $\forall i \neq h$

$$\overline{P}(c^i|\sigma) < \underline{P}(c^h|\sigma)$$

When there is no value dominating all other possible values of C, the output is the set of non-dominated cases (cases c^i for which there is no other case c^h verifying inequality). In this way, we obtain what Zaffalon [26] calls a *credal* classifier, in which, for a set of observations, we obtain a set of possible values for the variable to classify, non-dominated cases, instead of unique prediction. In the experiments, when there is no dominant value, we simply do not classify, without calculating the set of non-dominated attributes. This implies a loss of some valuable information in certain situations.

We want to compare our methods with existing classification methods. These methods classify all the records of the training and test sets, without rejecting any of the cases. In order to carry out a fair comparison with such complete procedures, we also use the **maximum frequency criterion** based on frequency of the data, i.e. we will choose the case with maximum frequency in $\mathcal{D}[\sigma]$ as the attribute of the variable to be classified.

5 Experimentation

We have applied this method to some known data sets, obtained from the *UCI repository of machine learning databases*, which can be found on the following website: http://www.sgi.com/Technology/mlc/db. We use the less conservative parameter $s = 1$, since with $s > 1$, we obtained a high degree of non-classified data in some databases (although with a greater percentage of correct classifications).

We plan to compare the behavior of the two total uncertainty measures we have previously defined:

- $TU1 = G^* + IG$
- $TU2 = G^*$

The data sets are: *Breast*, *Breast Cancer*, *Heart*, *Hepatitis*, *Cleveland*, *Cleveland nominal* and *Pima*(medical); *Australian* (banking); *Monks1* (artificial) and *Soybean-small* (botanical).

These databases were used by Acid [7]. Some of the original data sets have observations with missing values and in some cases, some of the variables are not discrete. The cases with missing values were removed and the continuous variables have been discretized using MLC++ software, available at the website http://www.sgi.com/Technology/mlc. The measure used to discretize them is the entropy. The number of intervals is not fixed and it is obtained following the Fayyad and Irani procedure [13]. Only the training part of the database is used to determine the discretization procedure. In Table 1 there is a brief description of these databases.

In general, when there is no case dominating all the other possible values of the variable to be classified, we simply do not classify this individual.

Data set	N. Tr	N. Ts	N. variables	N. classes
Breast Cancer	184	93	9	2
Breast	457	226	10	2
Heart	180	90	13	2
Hepatitis	59	21	19	2
Cleveland nominal	202	99	7	5
Cleveland	200	97	13	5
Pima	512	256	8	2
Vote1	300	135	15	2
Australian	460	230	14	2
Monks1	124	432	6	2
Soybean-small	31	16	21	4

Table 1: Description of the databases. The column *N. Tr* contains the number of cases of the training set, the column *N. Ts* is the number of cases of the test set, the column *N. variables* is the number of variables in the database and the column *N. classes* is the number of different values of the variable to be classified

Algorithms have been implemented using Java language version 1.1.8. In order to obtain the value of G^* for probability intervals we have used the algorithm proposed in Abellán and Moral [3].

The percentages obtained of correct classifications with the simple model and $TU1$ can be seen in Table 2.

In Table 2, the training column is the percentage of correct classifications in the data set that was used for learning. The $UC(Tr)$ column shows the percentage of rejected cases, i.e. the observations that were not classified by the method due to the fact that no value verifies the strong dominance criterion, and the $UC(Ts)$ column shows the rejected cases in the test set.

In the results presented in Table 2 (Abellán and Moral [4]) there is no overfitting (one of the most common problems of learning procedures): the success of the training set and the test set are very similar.

Only the *Cleveland* database has a high rate of non-classified data. This is the case with the highest number of cases of the variable to be classified and then it is more difficult to obtain a class dominating all the other classes. In this case, we would have obtained more information by changing the output to a set of non-dominated cases. In most of the other databases, the variable to be classified has two possible states and in this situation our classification is equivalent to the set of non-dominated values.

In Table 3, we see the success of other known methods on the same databases, Acid [7]. The NB-columns correspond to the results of the Naive Bayesian classifier on the training set and the test set. Similarly, the C4.5-columns correspond to Quinlan's method [19], based on the ID3 algorithm [18], where a classification

Data set	Training	UC(Tr)	Test	UC(Ts)
Breast Cancer	75.5	0.0	81.7	0.0
Breast	98.0	1.3	96.9	0.9
Heart	92.2	7.2	95.2	6.7
Hepatitis	96.4	5.0	94.7	9.5
Cleveland nominal	62.7	4.4	66.0	5.0
Cleveland	72.8	21.0	69.9	24.7
Pima	79.7	0.2	80.5	0.0
Australian	92.3	3.4	91.0	3.4
Vote1	96.1	6.6	96.9	5.9
Soybean-small	100.0	0.0	100.0	0.0

Table 2: The measured experimental percentages of the simple method and $TU1$. The columns *UC(Tr)* and *UC(Ts)* are the percentages of the rejected cases obtained with the training and the test set respectively.

tree with classical precise probabilities is used. We report the results obtained by Acid [7]. We can see that there is overfitting in these methods, principally in C4.5, being especially notable in certain data sets (*Cleveland nominal*, *Cleveland*, *Hepatitis*).

In Table 4 we can see the results of the simple method with $TU2$ and strong dominance. We have a higher percentage of success and a higher percentage of unclassified cases. This total uncertainty measure obtains larger trees as we can observe for the number of leaves presented in Table 5.

The success of the simple method with all cases classified (0% of rejected cases) with the frequency criterion are presented in Table 6 for the test set, to compare it with the models C4.5 and Naive Bayes. Table 7 shows the results of similar experiments with the double method. We can see the high percentages of correct classifications with $TU2$. These are a little higher than those obtained with $TU1$ and notably higher than the other methods (C4.5 and Naive Bayes).

The results of the simple and double methods are similar (slightly better in the double method). In order to see the potential of the double method we use an artificial database: *Monks1*.

Monks1 is a database with six variables. The variable to be classified has two possible states: $a0$ and $a1$, being $a1$ when the first and the second variables are equal or the fourth variable has the first of its possible four states. This type of dependency is very difficult to find for some classification methods, as this is a deterministic relationship involving more than two variables. The double method should be much better than the simple one.

Table 8 shows the success of the methods C4.5 and Naive Bayes. Table 9 shows the success of the simple and double method with all cases classified.

Data set	NB(Tr)	NB(Ts)	C4.5(Tr)	C4.5(Ts)
Breast Cancer	78.2	74.2	81.5	75.3
Breast	97.8	97.3	97.6	95.1
Cleveland nominal	63.9	57.6	69.3	51.5
Cleveland	78.0	50.5	73.5	54.6
Pima	76.4	74.6	79.9	75.0
Heart	87.8	82.2	83.3	75.6
Hepatitis	96.2	81.5	96.2	85.2
Australian	87.6	86.1	89.3	83.0
Vote1	87.6	88.9	94.5	88.3
Soybean-small	100	93.8	100	100

Table 3: Percentages of another methods

Data set	Training	UC(Tr)	Test	UC(Ts)
Breast Cancer	89.0	16.3	93.5	17.2
Breast	99.1	2.6	98.6	2.6
Cleveland nominal	73.6	21.2	74.4	13.1
Cleveland	82.6	34.0	80.3	31.9
Pima	86.6	15.6	86.2	15.2
Heart	93.9	8.8	93.8	10.0
Hepatitis	96.4	5.0	94.7	9.5
Australian	95.3	6.5	94.4	6.5
Vote1	98.2	5.3	98.4	4.4
Soybean-small	100.0	0.0	100.0	0.0

Table 4: Simple method with TU2 and strong dominance

Data set	TU1	TU2	N of possible leaves
Breast	10	17	512
Cleveland	17	112	635904

Table 5: Number of leaves of the trees obtained with the simple method and $TU1$ and $TU2$

Data set	TU1(Ts)	TU2(Ts)	NB(Ts)	C4.5(Ts)
Breast Cancer	81.7	90.3	74.2	75.3
Breast	96.9	97.8	97.3	95.1
Cleveland nominal	65.7	75.8	57.6	51.5
Cleveland	67.0	80.4	50.5	54.6
Pima	80.5	80.9	74.6	75.0
Heart	93.3	92.2	82.2	75.6
Hepatitis	95.2	95.2	81.5	85.2
Australian	90.9	93.5	86.1	83.0
Vote1	94.8	97.8	88.9	88.3
Soybean-small	100	100	93.8	100

Table 6: Success of the simple method with TU1 and TU2 with the frequency criterion on the test set

Database	TU1(Ts)	TU2(Ts)	NB(Ts)	C4.5(Ts)
Breast Cancer	81.7	91.4	74.2	75.3
Breast	96.9	98.7	97.3	95.1
Cleveland nominal	68.7	74.7	57.6	51.5
Cleveland	67.0	80.4	50.5	54.6
Pima	80.5	82.4	74.6	75.0
Heart	93.3	94.4	82.2	75.6
Hepatitis	95.2	95.2	81.5	85.2
Australian	89.1	91.7	86.1	83.0
Vote1	94.8	98.5	88.9	88.3
Soybean-small	100	100	93.8	100

Table 7: Success of the double method with TU1 and TU2 with the frequency criterion on the test set

Data set	NB(Tr)	NB(Ts)	C4.5(Tr)	C4.5(Ts)
Monks1	79.8	71.3	83.9	75.7

Table 8: C4.5 and Naive Bayes on Monks1

	Simple method		Double method	
Function	Tr	Ts	Tr	Ts
TU1	81.5	80.6	94.4	91.7
TU2	89.5	80.6	96.7	94.4

Table 9: Percentages on *Monks1* of the methods with TU1 and TU2 and all cases classified

We can see some interesting things. There is an appreciable overfitting in C4.5 and Naive Bayes but not in our methods. The percentage obtained with the test set is better in the extended method than in the simple method and there is a difference of 23.1% of the extended method and $TU2$ with respect to Naive Bayes success.

6 Conclusions

In this paper, we have discussed the role of maximum entropy as a total uncertainty measure in credal sets. First, we have revised some decision theoretic justification based on the logarithmic scoring rule. We have carried out a series of experiments in which we compare this measure with the one we had previously used in our experiments. The main conclusion is that, in general, the results are always the same or better when only the maximum entropy is used than when a non-specificity value is added to it (the other total uncertainty measure). And in some cases, the percentages of success are notably better.

Other conclusions from the experiments can be summarized in the following points:

- Imprecise probability methods are outstandingly better than classical probabilistic methods, and also have the option of not classifying difficult cases.
- In general, the double method produces slightly better results than the single one, but in some particular cases the differences can be remarkable.
- Maximum entropy (TU_2) produces larger trees than the other uncertainty measure (TU_1), but even this classifier does not suffer from overfitting.

Acknowledgements

We are very grateful to the anonymous referees for their valuable comments and suggestions.

References

[1] J. Abellán and S. Moral. Completing a Total Uncertainty Measure in Dempster-Shafer Theory. *Int. J. General Systems*, 28:299–314, 1999.

[2] J. Abellán and S. Moral. A Non-specificity Measure for Convex Sets of Probability Distributions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 8:357–367, 2000.

[3] J. Abellán and S. Moral. Maximum entropy for credal sets. To appear in *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2003.

[4] J. Abellán and S. Moral. Using the Total Uncertainty Criterion for Building Classification Trees. *Proceeding of the International Symposium of Imprecise Probabilities and Their Applications*, 1-8, 2001.

[5] J. Abellán and S. Moral. Construcción de árboles de clasificación con probabilidades imprecisas. *Actas de la Conferencia de la Asociación Española para la Inteligencia Artificial*, 2:1035-1044, 2001.

[6] J. Abellán. *Medidas de entropía y distancia en conjuntos convexos de probabilidad: definiciones y aplicaciones*. PhD thesis, Universidad de Granada, 2003.

[7] S. Acid. *Métodos de aprendizaje de Redes de Creencia. Aplicación a la Clasificación*. PhD thesis, Universidad de Granada, 1999.

[8] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth Statistics, Probability Series, Belmont, 1984.

[9] G. Choquet. Théorie des Capacités. *Ann. Inst. Fourier*, 5:131–292, 1953/54.

[10] A.P. Dempster. Upper and Lower Probabilities Induced by a Multivaluated Mapping, *Ann. Math. Statistic*, 38:325–339, 1967.

[11] D. Dubois and H. Prade. A Note on Measure of Specificity for Fuzzy Sets. *BUSEFAL*, 19:83–89, 1984.

[12] D. Dubois and H. Prade. Properties and Measures of Information in Evidence and Possibility Theories. *Fuzzy Sets and Systems*, 24:183–196, 1987.

[13] U.M. Fayyad and K.B. Irani. Multi-valued Interval Discretization of Continuous-valued Attributes for Classification Learning. *Proceeding of the 13th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Mateo, 1022-1027, 1993.

[14] D. Harmanec and G.J. Klir. Measuring Total Uncertainty in Dempster-Shafer Theory: a Novel Approach, *Int. J. General System*, 22:405–419, 1994.

[15] G.J. Klir and M.J. Wierman. *Uncertainty-Based Information*, Phisica-Verlag, 1998.

[16] S. Kullback. *Information Theory and Statistics*, Dover, 1968.

[17] Y. Maeda and H. Ichihashi. A Uncertainty Measure with Monotonicity under the Random Set Inclusion, *Int. J. General Systems* 21:379–392, 1993.

[18] J.R. Quinlan. Induction of decision trees, *Machine Learning*, 1:81–106, 1986.

[19] J.R. Quinlan. *Programs for Machine Learning*. Morgan Kaufmann series in Machine Learning, 1993.

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

[21] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423,623–656, 1948.

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

[23] P. Walley. Inferences from Multinomial Data: Learning about a Bag of Marbles. *J.R. Statist. Soc. B*, 58:3–57, 1996.

[24] R.R. Yager. Entropy and Specificity in a Mathematical Theory of Evidence. *Int. J. General Systems*, 9:249–260, 1983.

[25] M. Zaffalon. A Credal Approach to Naive Classification. *Proceedings of the First International Symposium on Imprecise Probabilities and their Applications*, 405-414, 1999.

[26] M. Zaffalon. The Naive Credal Classifier. *Journal of Statistical Planning and Inference*, 105:5–21, 2002.

J. Abellán is with the Universidad de Granada, Spain.

S. Moral is with the Universidad de Granada, Spain.

Bayesian Robustness with Quantile Loss Functions*

J.P. ARIAS
Universidad de Extremadura, Spain

J. HERNÁNDEZ
Universidad de Extremadura, Spain

J. MARTÍN
Universidad de Extremadura, Spain

A. SUÁREZ
Universidad de Cádiz, Spain

Abstract

Bayes decision problems require subjective elicitation of the inputs: beliefs and preferences. Sometimes, elicitation methods may not perfectly represent the Decision Maker's judgements. Several foundations propose to overlay this problem using robust approaches. In these models, beliefs are modelled by a class of probability distributions and preferences by a class of loss functions. Thus, the solution concept is the set of non-dominated alternatives. In this paper we focus on the computation of the efficient set when the preferences are modelled by a class of convex loss functions, specifically the quantile loss functions. We illustrate the idea with examples and introduce the use of stochastic dominance in this context.

Keywords

Bayesian robustness, non-dominated alternatives, Bayes alternatives, quantile loss functions, stochastic orders, quantile class of prior distributions

1 Introduction

Robust Bayesian analysis arises to avoid demanding an excessively precision in the decision maker's judgements concerning his beliefs and preferences. Thus, the

*This work has been supported in part by a grant of Junta de Extremadura IPR00A075. We thank the referees for their fruitful comments

imprecision in preferences leads to a class of loss functions while the imprecision in beliefs is modelled by a class of prior probability distributions which would be actualized via Bayes Theorem. For some interesting revisions on Bayesian Robustness axiomatic systems see e.g. Ríos Insua and Martín [13], Nau [11], Seidenfeld et al [18] and Weber [19].

In summary, using a class Γ of prior distributions over the set of states Θ and a class $\mathcal{L}$ of loss functions, given $a, b \in \mathcal{A}$, set of alternatives, we say that $b \preceq a$ if and only if

$$T(a,L,\pi) \leq T(b,L,\pi),\ \forall \pi \in \Gamma, \forall L \in \mathcal{L},$$

where $T(a,L,\pi)$ is the posterior expected loss for the action a, L is the loss function, π is the prior and $\preceq$ is the preference relationship between alternatives:

$$T(a,L,\pi) = \frac{\int_\Theta L(a,\theta) l(\theta) d\pi(\theta)}{\int_\Theta l(\theta) d\pi(\theta)},$$

$l(\theta)$ being the likelihood for an experiment x.

This model is similar to a multicriteria optimization problem. The optimal solution is the one that minimizes $T(\cdot,L,\pi)$ for every pair $\pi \in \Gamma, L \in \mathcal{L}$. Unfortunately, in general, that optimal solution does not exist. Thus, the non-dominated set is taken as an starting point. Any dominated alternative must be discarded. See Coello [6] for an excellent discussion on multiobjective optimization. We say that a dominates b if and only if $a \prec b$, (that is, $a \preceq b$ and $\neg(b \preceq a)$). A non-dominated alternative a is such that there is no other alternative b which dominates a. Arias [1] and Arias and Martín [2] provide theoretic results about the existence of such a set and its relationship with the set of Bayes alternatives. Martín and Arias [8] provide a method based on comparing pairs to approximate the non-dominated set. Some references for Bayesian sensitivity are Berger [4], Ríos Insua and Ruggeri [14] and Ríos et al [15].

We study the calculus of the non-dominated set for problems in which the imprecision in preferences is modelled by quantile loss functions. We give general results that we will particularize for classes of quantile prior distributions, see Moreno and Cano [9]. Since we are interested in Bayesian inference, we will consider $\mathcal{A} = \mathbb{R}$ although the results will be easily applicable when $\mathcal{A}$ is an interval of $\mathbb{R}$.

We organize this work as follows. We begin with some results concerning convex loss functions and their implications in the calculus of the non-dominated set. Secondly, we particularize for quantile loss functions, indicating the relationship with the Bayes alternatives in this case. We also consider a quantile class for prior distributions giving some results and an example. Third part of the paper is dedicated to various stochastic orders, only those that hold for the posterior distributions once the priors have been ordered, and how they can be used in order to calculate the non-dominated set.

2 Bayesian Robustness with convex loss functions

We will denote $\mathcal{L}_C$ the class of all convex loss functions in $\mathcal{A}$. Every loss function $L \in \mathcal{L}_C$, verifies for all θ, $a, b \in \mathcal{A}$, and $\lambda \in [0,1]$ that

$$L(\lambda a + (1-\lambda)b, \theta) \leq \lambda L(a,\theta) + (1-\lambda)L(b,\theta). \tag{1}$$

A first useful result, easy to prove, is:

Lemma 1 *Let Γ be any class of prior distributions and $\mathcal{L}_C$ the class of convex loss functions. The function $T(\cdot, L, \pi) : \mathcal{A} \longrightarrow \mathbb{R}$ is convex for every pair $(L,\pi) \in \mathcal{L}_C \times \Gamma$.*

A well known result is that every convex function is continuous in the interior, see Roberts and Varberg [16]. Then considering the set of alternatives, $\mathbb{R}$, the function $T(\cdot, L, \pi)$ is continuous in $\mathbb{R}$ and if it exists the set of Bayes alternatives, this will be a closed interval in $\mathbb{R}$. In the case that, for some pair $(L,\pi) \in \mathcal{L}_C \times \Gamma$ the set of Bayes alternatives is empty, the function $T(\cdot, L, \pi)$ will be increasing or decreasing in $\mathbb{R}$ (strictly increasing or strictly decreasing if the functions are strictly convex).

If the set of Bayes alternatives is not empty, then the function $T(\cdot, L, \pi)$ is strictly decreasing in $(-\infty, a_{(L,\pi)})$ and strictly increasing in $(a^{(L,\pi)}, +\infty)$, being

$$\begin{aligned} a_{(L,\pi)} &= \min_{a \in B_{(L,\pi)}} a, \text{ and} \\ a^{(L,\pi)} &= \max_{a \in B_{(L,\pi)}} a. \end{aligned}$$

Note that the alternatives $a_{(L,\pi)}$ and $a^{(L,\pi)}$ are also Bayes for (L,π).

An immediate result is that the set of non-dominated alternatives is included in the closed interval $[\mu_*, \mu^*]$, being μ_* and μ^*, respectively, the infimum and the supremum of the Bayes alternatives, that is,

$$\begin{aligned} \mu_* &= \inf_{(L,\pi)\in\mathcal{L}\times\Gamma} a_{(L,\pi)}, \text{ and} \\ \mu^* &= \sup_{(L,\pi)\in\mathcal{L}\times\Gamma} a^{(L,\pi)}. \end{aligned}$$

In the Bayesian literature the range of this interval is considered as the robustness measure of the problem, see Berger [4]. However, if we are interested in calculating exactly the set of non-dominated alternatives, we can give a more accurate approximation using the following result due to Arias et al [3]:

Theorem 1 *Let $\mathcal{L} \subseteq \mathcal{L}_C$ be a family of convex loss functions, Γ a class of prior probability distributions, so that, for every pair $(L,\pi) \in \mathcal{L} \times \Gamma$, the set of Bayes* *alternatives $B_{(L,\pi)}$ is not empty and let*

$$\begin{aligned} a_* &= \inf_{(L,\pi)\in\mathcal{L}\times\Gamma} a^{(L,\pi)}. \\ a^* &= \sup_{(L,\pi)\in\mathcal{L}\times\Gamma} a_{(L,\pi)}. \end{aligned}$$

We have

1. *If $a_* < a^*$, then $(a_*, a^*) \subseteq \mathcal{ND}(\mathcal{A}) \subseteq [a_*, a^*]$.*
2. *If $a_* \geq a^*$, then $\mathcal{ND}(\mathcal{A}) = [a^*, a_*]$.*

In order to study the robustness of the problem, it is not necessary to determine whether the alternatives a_* and a^* are dominated or not. Nevertheless, it is interesting to calculate the efficient set in an accurate way. In this paper we will see inference problems modelled by particular classes of loss functions and prior distributions in which we can assure that the extremes of the interval a_* and/or a^* are non-dominated alternatives. If the set of Bayes alternatives is empty for some pair $(L,\pi) \in \mathcal{L}\times\Gamma$ then the result is valid considering $a_* = -\infty$ (when $T(\cdot, L, \pi)$ is increasing) or $a^* = \infty$ (decreasing).

3 Quantile loss functions

Let us consider the case where preferences are modelled by quantile loss functions. A particular case of this type is the absolute value loss function. The class of quantile loss functions is defined as

$$\mathcal{L} = \{L_p : L_p(a,\theta) = |a-\theta| - a(2p-1),\ p \in [0,1]\} \quad (1).$$

Functions equivalent to these have been used in Economy, such as the ones studied by Geweke [7]

$$L(a,\theta) = c_1(a-\theta)I_{(-\infty,a]}(\theta) + c_2(\theta-a)I_{(a,+\infty)}(\theta) \quad (2)$$

where I is the indicator function. Bayes alternatives for this type of function are the quantiles of order $c_2/(c_1+c_2)$ (If $c_1 = c_2$, it coincides with the median) Thus, they are asymmetric functions with different weights on the positive and negative errors. Next example shows the use of this type of functions.

Example 1 *Noortwijt and Gelder [12] studied the Bayes estimators of the optimal dyke height under asymmetric linear loss function. Let us suppose we have to decide the height of the dykes to prevent flooding. The height of the dyke h will be the decision variable and $h_0 = 3.25$ the initial height at the moment when the decision has to be taken. Inundation will occur as soon as the sea water level* *exceeds the hight of the dyke. We assume that the maximal sea levels per year X_i $i = 1,\dots,n$ are conditionally independent, exponentially distributed, with a known location parameter $x_0 = 1.96$ meters and an unknown parameter λ with expected value 0.33 meters. Therefore the likelihood function is*

$$l(x|\lambda) = \prod_{i=1}^{n} f(x_i,\lambda) = \prod_{i=1}^{n} \frac{1}{\lambda} exp\{-\frac{x_i - x_0}{\lambda}\}.$$

The prior density of λ is assumed to be an inverted gamma distribution with scale parameter $\mu > 0$ and shape parameter $\nu > 0$

$$I_g(\lambda,\nu,\mu) = [\mu^{\nu}/\Gamma(\nu)]\lambda^{-(\nu+1)} exp\{-\mu/\lambda\} \quad \lambda > 0.$$

The loss function is (2) with $c_1 = 5.37 \cdot 10^7$ and $c_2 = 1.94 \cdot 10^7$. $\triangle$

This type of loss function have also been used in Forecast Theory, see Capistrán [5] and references therein.

We will use the functions defined in (1) as they only depend on a single parameter. Quantile loss functions are convex in $\mathcal{A}$. The posterior expected loss is

$$T(a,L_p,\pi) = D_{\theta|x}(a) - a(2p-1),$$

and their Bayes alternatives are the quantiles of order p of the posterior distributions, since

$$T'(a,L_p,\pi) = 2F_{\theta|x}(a) - 2p,$$

for every point a where the distribution function is continuous.

Let us recall that it is called quantile of order p of a random variable X, the value $Q_X(p)$ such that

$$P[X \leq Q_X(p)] \geq p \text{ and}$$
$$P[X \geq Q_X(p)] \geq 1-p.$$

As it happened with the absolute value loss function, when using quantile loss functions, the posterior distribution quantiles may not be unique.

Based on theorem 1 we have the following result, when there is precision in DM’s beliefs.

Proposition 1 *Let $\mathcal{L}$ be the class of quantile loss functions*

$$\mathcal{L} = \{L_p : L_p(a,\theta) = |a-\theta| - a(2p-1),\ p \in [p_0,p_1]\}$$

and π a prior distribution so that the posterior distribution quantiles are unique, then

$$\mathcal{ND}_{\pi}(\mathcal{A}) = [Q_{\pi}(p_0), Q_{\pi}(p_1)],$$

where $Q_{\pi}(p_0)$ and $Q_{\pi}(p_1)$ are, respectively, the quantiles of order p_0 and p_1 of the posterior distribution of π.

This result can be generalized in the case that we also have imprecision in the decision maker's beliefs.

Proposition 2 *Let $\mathcal{L}$ be the class of quantile loss functions*

$$\mathcal{L} = \{L_p : L_p(a,\theta) = |a-\theta| - a(2p-1),\ p \in [p_0, p_1]\},$$

a class of distributions

$$\Gamma = \{\pi : \pi(\theta|x) \text{ with posterior quantiles } Q_\pi(p) \text{ unique},\ p \in [0,1]\}$$

and the values

$$\begin{aligned} a_* = \mu_* &= \inf_{\pi\in\Gamma} Q_\pi(p_0) \text{ and} \\ a^* = \mu^* &= \sup_{\pi\in\Gamma} Q_\pi(p_1), \end{aligned}$$

then

$$(\mu_*, \mu^*) \subseteq \mathcal{ND}(\mathcal{A}) \subseteq [\mu_*, \mu^*]\,.$$

If posterior quantiles are not unique we must appeal to theorem 1 with

$$a_* = \inf_{\pi\in\Gamma}\{\sup Q_\pi(p_0)\}$$

$$a_* = \sup_{\pi\in\Gamma}\{\inf Q_\pi(p_1)\}$$

and

$$(a_*, a^*) \subseteq \mathcal{ND}(\mathcal{A}) \subseteq [a_*, a^*]\,.$$

In general it can not be assured that μ_* or μ^* are non-dominated alternatives as we illustrate with the following example.

Example 2 *Let us consider the class of absolute value loss functions and a class of discrete posterior distributions with probability distribution:*

$(n \in \mathbb{N})$

$$\pi_n(\theta) = \begin{cases} \dfrac{2n+3}{4(n+1)} & \text{if } \theta = \dfrac{1}{n}, \\ \\ \dfrac{2n+1}{4(n+1)} & \text{if } \theta = 1. \end{cases}$$

The Bayes alternative for each distribution π_n would be its posterior median $1/n$ and the set of non-dominated alternatives is the interval $(0,1]$. The alternative 0 is dominated by the alternative 1, since for every $n \in \mathbb{N}$

$$T(0, L, \pi_n) = \frac{2n^2+3n+3}{4n\,(n+1)} > \frac{2n^2+n-3}{4n\,(n+1)} = T(1, L, \pi_n)\,.$$

△

Obviously if μ_* and μ^* are unique Bayes alternatives, then they are also non-dominated alternatives.

Note that, if having precision in beliefs, then the range of the non-dominated set is the range between the quantile p_0 and the quantile p_1. Sometimes the non-dominated set can be the same as the (HPD), as in next example. However, so that this happens the elicitation of the class should depend on the posterior distribution.

Example 3 *Let us consider the class of loss functions*

$$\mathcal{L} = \{L_p : L_p(a,\theta) = |a-\theta| - a(2p-1),\ p \in [0,0.8]\}$$

and a Pareto prior distribution with parameters α *and* β*.*

We take a sample $\{X_1,\ldots,X_n\}$ *of a population which is distributed following an uniform distribution with mean* $\theta/2$*. Therefore, the posterior distribution is*

$$P(\alpha+n, \max\left(\beta, X_{(n)}\right)),\ \textit{with}\ X_{(n)} = \max\{X_1,\ldots,X_n\}$$

Then, the posterior quantiles of π *are*

$$Q_\pi(p) = \frac{1}{\sqrt[\alpha+n]{p}} \max\left(\beta, X_{(n)}\right).$$

Thus, the set of non-dominated alternatives would be the closed interval

$$ND(\mathcal{A}) = \left[\max\left(\beta, X_{(n)}\right), Q_\pi(0.8)\right].$$

This means than the non-dominated set coincides with the confidence interval HPD for the parameter θ *at a confidence level of* 80% *.*

The table 1 shows the non-dominated set when $\alpha = 2$ *and* $\beta = 5$ *for various samples.*

n	$X_{(n)}$	$ND(A)$	range
10	3.2	$[5, 5.0938]$	0.0938
50	4.5	$[5, 5.0215]$	0.0215
100	4.8	$[5, 5.011]$	0.011
500	5.9	$[5.9, 5.9026]$	0.0026
1000	4.7	$[5, 5.0011]$	0.0011

Table 1: Non-dominated set for the example 3.

△

3.1 Relationship with the non-dominated set

An important question is the relationship between the Bayes set and the non-dominated set. It is easy to prove that, in general, they are different, see Arias *et al* [1]. In this case, we have the following result:

Proposition 3 *Let $\mathcal{L}$ be the class of quantile loss functions with $p \in [p_0, p_1]$. If the class Γ of prior distributions is convex and $Q_\pi(p)$ are unique for every $\pi \in \Gamma$ and $p \in (p_0, p_1)$, then the set of non-dominated alternatives is the closure of the set of Bayes actions, and the interiors of both sets are the same.*

Proof.

If the set of prior distributions is convex, then the set of posterior distributions is convex as well, see Arias et al[1]. As the quantiles are unique for any π, all bayes actions are non-dominated. So, we only have to prove that given a and b bayes actions for (π, L_1) and(π, L_2), $\alpha a + (1-\alpha)b$ is also bayes for $\alpha \in (0,1)$. Consider $a = Q_{\pi_1}(p')$ and $b = Q_{\pi_2}(p'')$ with $p', p'' \in [p_0, p_1]$ and $a < b$ $p' < p''$.

Then

$$\int_{-\infty}^{\alpha a+(1-\alpha)b} \pi_1(\theta|x)d\theta \quad > p'$$

$$\int_{-\infty}^{\alpha a+(1-\alpha)b} \pi_2(\theta|x)d\theta \quad < p''$$

so, there is β such that

$$\beta \int_{-\infty}^{\alpha a+(1-\alpha)b} \pi_1(\theta|x)d\theta + (1-\beta) \int_{-\infty}^{\alpha a+(1-\alpha)b} \pi_2(\theta|x)d\theta \quad = p \in (p', p'')$$

Then $\alpha a + (1-\alpha)b$ is the $Q_\pi(p)$ with $\pi(\cdot|x) = \beta\pi_1(\cdot|x) + (1-\beta)\pi_2(\cdot|x)$

□

4 Quantile prior distributions

We now consider some classes of prior distributions to model imprecision in beliefs. Let $A_i = \lfloor \theta_i, \theta_{i+1} \rceil$ [1] be a partition of the parameter space and the class of prior distributions:

$$\Gamma_Q = \{\pi : \pi(A_i) = q_i, \quad i = 1 \ldots n, \quad q_i \geq 0 \quad \forall i \quad \sum_i q_i = 1\}$$

This is a particular case of the quantile class, see Moreno and Cano [9] and Moreno and Pericchi [10] and Martín and Ríos Insua [13] among others.

A well known result states that the suprema and the infima of functionals over Γ are attained for discrete distributions. So, we have

Lemma 2 *We have*

$$\max_{\pi_d \in \Gamma_d} \pi(\bigcup_{j=1}^{i} A_i|x) = \frac{\sum_{j=1}^{i} \max_{\theta\in A_j} l(x|\theta)p_j}{\sum_{j=1}^{i} \max_{\theta\in A_j} l(x|\theta)p_j + \sum_{k=i+1}^{n} \min_{\theta\in A_k} l(x|\theta)p_k}$$

[1] by $\lfloor a, b \rceil$ we denote any type of interval in $\mathbb{R}$

Let us denote r_i the second term in (8). Lemma 2 lead us to the following iterative scheme to calculate μ_*

```
$r_0 = 0 \quad i = 0$
while $r_i < p_0 \quad i = i+1$
  compute $r_i$
```

Let A_k be the first interval for which $r_k \geq p_0$. We define now

$$r_k(\theta) = \frac{\sum_{j=1}^{i-1} \max_{\lambda \in A_j} l(x|\lambda) p_j + l(x|\theta) p_k}{\sum_{j=1}^{i-1} \max_{\lambda \in A_j} l(x|\lambda) p_j + l(x|\theta) p_k + \sum_{j=k+1}^{n} \min_{\lambda \in A_j} l(x|\lambda) p_j} \quad \forall \theta \in A_k$$

then

$$\mu_* = \inf\{\theta \in A_k : r_k(\theta) \geq p_0\}$$

By Theorem 1, for the calculus of a_*, we will distinguish two cases, if the last inequality is strict then $\mu_* = a_*$ which is the only quantile of order p. Otherwise, there are several quantiles. For the calculus of a_* we proceed as follows. In A_k we will search a point $a > \mu_*$ for which the inequality is strict. If such point exists then $a_* = \inf\{a \in A_k : r_k(a) > p_0\}$. If there is no point $a > \mu_*$ in A_k such that this is verified then $a_* = \inf\{a \in A_{k+1} : r_{k+1}(a) > p_0\}$ and so on.

Example 4 *The decision maker considers that negative errors are more important than positive, so he uses the class of loss functions:*

$$\mathcal{L} = \{L_p : L_p(a,\theta) = |a-\theta| - a(2p-1),\ p \in [0.45, 0.48]\}.$$

For believes representation he adopts a quantile class with quantiles given in Table 2.

A_i	$[-\infty,-16.44)$	$[-16.44,-5.22)$	$[-5.24,-2.53)$	$[-2.53,-1.25)$	$[-1.25,0)$
p_i	0.05	0.25	0.1	0.05	0.05
A_i	$[0,1.25)$	$[1.25,2.53)$	$[2.53,5.24)$	$[5.24,16.44)$	$[16.44,\infty)$
p_i	0.05	0.05	0.1	0.25	0.05

Table 2: Probabilities for some intervals

Applying the proposed iterative scheme to obtain $\inf Q_\pi(0.45)$ *we get the* r_i *values showed table 3. Therefore* $\min_{\pi\in\Gamma} Q_\pi(0.45) \in A_4 = [-2.533, -1.257]$ *and solving* $\min\{\theta \in A_i$ *such that* $r_k(\theta) \geq 0.45\}$ *we obtain* $\inf Q_\pi(0.45) = -2.533$. *Moreover,* $r_k(-2.5333) > 0.45$ *so* $a_* = -2.533$.

We apply the equivalent algorithm for $sup_{\pi\in\Gamma} Q_\pi(0.48)$ *obtaining 2.533. Then* $\mathcal{ND}(\mathcal{A}) = [-2.533, 2.533]$

i	1	2	3	4	5	6	7	8	9	10
r_i	0.0001	0.28	0.44	0.54	0.65	0.76	0.87	0.99	1	1

Table 3: r_i values

The class of quantiles Γ_Q can be generalized considering bounds over the sets A_i obtaining the class

$$\Gamma_{QG} = \{\pi : \alpha_i \le \pi(A_i) \le \beta_i, \quad 0 \le \alpha_i \le \beta_i \le 1\}$$

In this case the proposed scheme can be modified taking into account that

$$\Gamma_{QG} = \bigcup_{\alpha \le p \le \beta} \{\pi : \pi(A_i) = p_i \quad \sum_i pi = 1\}$$

where $\alpha = (\alpha_1, \dots, \alpha_n)$ $p = (p_1, \dots, p_n)$ $\beta = (\beta_1, \dots, \beta_n)$ and $\alpha \le p \le \beta$ indicates $\alpha_i \le p_i \le \beta_i \quad i = 1, \dots, n.$

Thus, to calculate r_k it is enough to consider sequently the linear problems:

$$\max \sum_{i=1}^{k} \max_{\theta_i \in A_i} l(x|\theta_i) p_i$$

$$s.a. \quad \sum_{i=1}^{n} p_i = 1$$

$$\alpha_i \le p_i \le \beta_i \quad i = 1, \dots, n$$

with optimum $p_1^*, \dots, p_k^*$ and

$$\min \sum_{i=k+1}^{n} \min_{\theta_i \in A_i} l(x|\theta_i) p_i$$

$$s.a. \sum_{i=1}^{k} p_i^* + \sum_{i=k+1}^{n} p_i = 1$$

$$\alpha_i \le p_i \le \beta_i \quad i = 1, \dots, n$$

with optimum $p_{k+1}^*, \dots, p_n^*$ and we replace in the algorithm p_i for p_i^*. A similar modification give us the values μ_* and a_*.

A natural extention of the quantile class for continuous parameters is the class

$$\Gamma_{LU} = \{\pi : L(A) \le \pi(A) \le U(A), \quad \forall A \in \beta\}$$

where β is a σ-field on the state set Θ. This is the class studied, among others, by Moreno and Pericchi [10] who provide the following result for posterior probabilities of sets in β.

Theorem 2 *Let A be an arbitrary set in β. Suppose that $l(x|\theta)$, L and U satisfy $U([l(x|\theta)=z]) = L([l(x|\theta)=z]) = 0$ for any $z \geq 0$ where $[l(x|\theta)=z] = \{\theta : l(x|\theta)=z\}$. Then, we have*

$$\textit{(i)} \quad \textit{if} \quad U(A)+L(A^c) > 1 \quad \textit{then} \sup_{\pi\in\Gamma_{LU}} P^{\pi}(A|x) = P^{\pi_0}(A|x)$$

$$\textit{where} \quad \pi_0(d\theta) = U(d\theta) I_{A\bigcap[l(x|\theta)\geq z_A]}(\theta) + L(d\theta) I_{A\bigcap[l(x|\theta)<z_A]\bigcap A^c}(\theta)$$

$$z_A \quad \textit{being such that} \quad \pi_o(\Theta) = 1$$

$$\textit{(ii)} \quad \textit{if} \quad U(A)+L(A^c) = 1 \quad \textit{then} \sup_{\pi\in\Gamma_{LU}} P^{\pi}(A|x) = P^{\pi_0}(A|x)$$

$$\textit{where} \quad \pi_0(d\theta) = U(d\theta) I_A(\theta) + L(d\theta) I_{A^c}(\theta)$$

$$\textit{(iii)} \quad \textit{if} \quad U(A)+L(A^c) < 1 \quad \textit{then} \sup_{\pi\in\Gamma_{LU}} P^{\pi}(A|x) = P^{\pi_0}(A|x)$$

$$\textit{where} \quad \pi_o(d\theta) = U(d\theta) I_{A\bigcup A^c\bigcap[f(x|\theta)<z_A]}(\theta) + L(d\theta) I_{A^c\bigcap[f(x|\theta)\geq z_A]}(\theta)$$

$$z_A \quad \textit{being such that} \quad \pi_o(\Theta) = 1$$

This resuls allow us to compute $\mathcal{ND}(\mathcal{A})$.

Theorem 3 *Let be the class $\Gamma_{LU} = \{\pi : L(A) \leq \pi(A) \leq U(A), \quad \forall A \in \beta\}$ and $\mathcal{L} = \{L_p : L_p(a,\theta) = |a-\theta| + a(2p-1),\ p \in [p_0, p_1]\}$ and $l(x|\theta)$, L and U satisfy $U([l(x|\theta)=z]) + L([l(x|\theta)=z]) = 0 \quad \forall z \geq 0$ then*

$$\mu_* = \inf\{\theta \in \Theta : \sup_{\pi\in\Gamma_{LU}} P^{\pi}(-\infty,\theta) \geq p_0\},$$

where P^{π} denotes the posterior probability.

Proof. Let $\theta_* = \inf\{\theta \in \Theta : \sup_{\pi\in\Gamma_{LU}} P^{\pi}(-\infty,\theta) \geq p_0\}$. If $\theta < \theta_*$ then $P^{\pi}(-\infty,\theta) < p_0 \quad \forall\pi$. Then by Theorem 1, θ is a dominated alternative. Moreover, if θ_* satisfies $\sup_{\pi\in\Gamma_{LU}}\{P^{\pi}(-\infty,\theta_*) \geq p_0\}$ then there is $\pi_* \in \Gamma_{LU}$ such that $\theta_* \in Q_{\pi_*}(p_0)$. In other case, by previous Theorem 2 there is a sequence of values θ_n such that is exists $\pi_{\theta_n} \in \Gamma_{LU}, \quad \theta_n \in Q_{\pi_{\theta_n}}(p_0)$ with $\theta_n \to \theta_*$ so $\theta_* = \mu_*$. □

Theorem 4 *If L and U verify that $L(A) > 0 \quad U(A) > 0 \quad \forall A \quad$ with $\quad \mu(A) > 0$ then $\mu_* = a_*$.*

Proof. It is easy to prove that π_0 of theorem 2 has unique quantiles. □

These results can be applied using searching methods based on simulation schemes.

5 Stochastic order applied to the calculus of the non-dominated set

The relationship between a prior distribution $\pi(\theta)$ and its corresponding posterior distribution $\pi(\theta|x)$, through Bayes Theorem, is not simple in the sense that, properties in the prior distribution not always hold for the posterior distribution.

In this context, starting from the class Γ of prior distributions which models the decision maker's uncertainty, it would be greatly useful if one could be able to establish order relationships between the posterior distributions from the order relationships already known among the prior distributions. In other words, given two distributions $\pi_1(\theta)$ and $\pi_2(\theta)$ belonging to the class Γ such that $\pi_1(\theta) \prec \pi_2(\theta)$, where $\prec$ is an order relationship between both distributions, it would be of great interest that this relationship remained in the posterior distributions, this is, that it is verified $\pi_1(\theta|x) \prec \pi_2(\theta|x)$. We find the ideal tool for this study in the general theory of stochastic orders. We introduce a brief summary about the concept of stochastic order between distribution functions and the definitions of various orders. The applications of such orders notably simplifies the calculus of the non-dominated set as we will show.

Let Γ be a family of distribution functions over which a binary relationship, which is a partial order, has been defined "$\prec$". Each time we assess the relationship $F \prec G$, we will extend this order to the random variables $X \prec Y$, where F and G are the distribution functions of X and Y respectively.

Definition 1 *The random variable X is said to be stochastically smaller than the random variable Y, we will denote $F \prec_{st} G$, if $F(x) \geq G(x)$ for every x belonging to $\mathbb{R}$ being F and G the corresponding distribution functions.*

This is the most common order in the stochastic distribution theory. If two random variables are stochastically ordered, this implies that all their location parameters are also ordered. Let us remember that, in many examples in decision theory, the Bayes alternatives are the location parameters of the posterior distributions. Besides, it is immediate, from the definition, that the stochastic order between two variables implies the order between their respective quantiles.

Definition 2 *Given X and Y two continuous random variables with density functions f and g respectively, we will say that X is smaller in likelihood ratio than Y, we will denote $X \prec_{lr} Y$, if*

$$\frac{g(t)}{f(t)} \quad \textit{is increasing over the union of the supports of } X \textit{ and } Y,$$

where $a/0$ is consider ∞ every time that a is greater than zero.

Given two random variables X and Y it is verified that

$$X \prec_{lr} Y \quad \Rightarrow \quad X \prec_{st} Y$$

see Shaked and Shantikumar [17] and Whitt [20].

As indicated in the beginning of this section, the relationship between the density function of the prior distribution and the posterior distribution is not easily treatable from a mathematic point of view; although, this relationship is more intuitive when we study properties associated to the quotient of two prior distributions. Due to the form of the posterior density function, it is not difficult to translate these properties to the quotient of the respective posterior distributions. In this way we give the following two propositions, easy to prove, but of great use as we show later with various examples.

Proposition 4 *Let $\pi_1(\theta)$ and $\pi_2(\theta)$ be two prior distributions for an unknown parameter of interest. Let $\pi_1(\theta|x)$ and $\pi_2(\theta|x)$ be the respective posterior distributions of the parameter once the sampling experiment information has been incorporated. Then if $\pi_1(\theta) \prec_{lr} \pi_2(\theta)$ it is verified that $\pi_1(\theta|x) \prec_{lr} \pi_2(\theta|x)$. Particularly, it is verified that $\pi_1(\theta|x) \prec_{st} \pi_2(\theta|x)$.*

Example 5 *Let us consider a decision problem where the decision maker's beliefs are modelled by a parametric class of Pareto distributions with unknown parameter α*

$$\Gamma = \{\pi \sim \mathcal{P}(\alpha,\beta) : \alpha \in [\alpha_1,\alpha_2], \alpha_1,\beta > 0\}$$

and the preferences are modelled by the class of quantile loss functions

$$\mathcal{L} = \{L_p : L_p(a,\theta) = |a-\theta| - a(2p-1),\ p \in [p_1,p_2]\}.$$

The class of Pareto distributions can be ordered in the sense of likelihood ratio, since, for any two distributions $\mathcal{P}(\alpha_1,\beta)$, $\mathcal{P}(\alpha_2,\beta)$,

$$\frac{\pi_2(\theta)}{\pi_1(\theta)} = \frac{\alpha_2}{\alpha_1}\left(\frac{\beta}{\theta}\right)^{\alpha_2-\alpha_1}$$

is an increasing function in $[\beta,+\infty)$, as long as $\alpha_1 > \alpha_2$. Then, it is stochastically ordered and, therefore, all the location parameters are ordered, in particular, the quantiles are ordered. If we take a sample of size n of a population that is distributed according to an uniform distribution of mean $\theta/2$, we have that, by proposition 4, the non-dominated set coincides with the closed interval

$$\left[Q_{\mathcal{P}(\alpha_2,\beta)}(p_1), Q_{\mathcal{P}(\alpha_1,\beta)}(p_2)\right].$$

Table 4 shows the non-dominated set for some samples and supposing that $\beta = 4, \alpha_1 = 2, \alpha_2 = 4, p_1 = \frac{1}{3}$ and $p_2 = \frac{1}{3}$.

References

[1] Arias. J.P., El conjunto no dominado en la teoría de la decisión bayesiana robusta, *Ph.D. Thesis*, Universidad de Extremadura, 2002.

n	$X_{(n)}$	$ND(A)$	range
4	3.2	$[4.208, 4.489]$	0.2819
50	4.5	$[4.534, 4.560]$	0.0264
100	4.8	$[4.819, 4.833]$	0.0139
500	5.9	$[5.905, 5.908]$	0.0034
1000	4.7	$[4.702, 4.703]$	0.0013
10000	6.1	$[6.1002, 6.1004]$	0.00017

Table 4: Non-dominated set for the example 5.

[2] Arias, J.P. and Martín, J., Uncertainty in beliefs and preferences: Conditions for optimal alternatives, *Annals of Mathematics and Artificial Intelligence*, 35, 1-4, 3-10, 2002.

[3] Arias, J.P., Martín, J.R., Suárez, A., Bayesian Robustness with convex loss functions, *Technical Paper*, Universidad de Extremadura, avaliable at jparias@unex.es, 2003.

[4] Berger, J., An overview of robust Bayesian analysis (with discussion), *Test*, 3, 5-124, 1994.

[5] Capistrán Carmona, C., A review of Forecast Theory using generalized loss functions, *Working paper*, University of California, San Diego, 2003.

[6] Coello Coello, C. A., Van Veldhuizen, D. A. and Lamont, G. B., *Evolutionary Algorithms for Solving Multi-Objetive Problems,* Kluber A.P., 2002.

[7] Geweke, J., A brief digression on ancillarity and nuisance parameters, *Lecture Notes in Economics, Applies Econometrics II* 8-205, 1997.

[8] Martín, J. and Arias, J.P., Computing the efficient set in Bayesian decision problems, *Robust Bayesian Analysis*, eds. D. Ríos Insua and F. Ruggeri, Lectures Notes in Statistics, vol 152, Springer, 2000.

[9] Moreno, E. and Cano, J.A., Robust Bayesian analysis for ε-contaminations partially known, *Journal Royal Statistical Society B* 53, 143-155, 1991.

[10] Moreno E. and Pericchi R., Bands of probability measures: a robust bayesian analysis, *Bayesian Statistics 4*, Oxford O.P.707-713 (eds. Bernardo *et al*) , 1992.

[11] Nau, Robert F., The shape of incomplete preferences *Working Paper,* Fuqua School of Bussines, Duke University, 1996.

[12] Van Noortwijk, J. and van Gelder, P., Bayesian estimation of quantiles for the purpose of flood prevention. B.L. Edge, ed., *Proceedings of the 26th International Conference on Coastal Engineering*, Copenhagen, Denmark, 3529-3541, 1998.

[13] Ríos Insua, D. and Martín J., On the foundations of robust Decision Making *Decision Theory and Decision Analysis: Trends and Challenges* Rios, S., Kluwer A.P., 1994.

[14] Ríos Insua, D. and Ruggeri, F., *Robust Bayesian Analysis,* Springer, 2000.

[15] Ríos Insua, D., Ruggeri, F. and Martín, J., Bayesian sensitivity analysis: a review, *Sensitivity Analysis*, (eds. Saltelli *et al*), New York: Wiley, 2000.

[16] Roberts, A.W. and Varberg, A.D., *Convex Functions*, Academic Press. New York, 1973.

[17] Shaked, M. and Shanthikumar, J.G., *Stochastic orders and their applications,* Academic Press, New York, 1994.

[18] Seidenfeld, T., Shervish, M.J., and Kadane, J., A representation of partially ordered preferences, *Annals of Statistics*, 23, 2168-2217, 1995.

[19] Weber, M., Decision Making with incomplete information, *European Journal of Operations Research*, 28, 4-57, 1987.

[20] Whitt, W. Uniform conditional variability ordering of probability distributions, *Journal of Applied Probability,* 22, 619-633, 1985.

Pablo Arias belongs to the Department of Mathematics, Universidad de Extremadura, cáceres, 10071, Spain. E-mail: jparias@unex.es

Javier Hernández belongs to the Department of Mathematics, Universidad de Extremadura, Badajoz, 06006, Spain. E-mail: javierhs@materiales.unex.es

Jacinto Martín belongs to the Department of Mathematics, Universidad de Extremadura, Cáceres, 10071, Spain. E-mail: jrmartin@unex.es

Alfonso Suárez belongs to the Department of Statistics, Universidad de Cádiz, Spain E-mail: alfonso.suarez@uca.es

On the Suboptimality of the Generalized Bayes Rule and Robust Bayesian Procedures from the Decision Theoretic Point of View: A Cautionary Note on Updating Imprecise Priors

THOMAS AUGUSTIN
Ludwig-Maximilians University of Munich, Germany

Abstract

This paper discusses fundamental aspects of inference with imprecise probabilities from the decision theoretic point of view. It is shown why the equivalence of prior risk and posterior loss, well known from classical Bayesian statistics, is no longer valid under imprecise priors. As a consequence, straightforward updating, as suggested by Walley's Generalized Bayes Rule or as usually done in the Robust Bayesian setting, may lead to suboptimal decision functions. As a result, it must be warned that, in the framework of imprecise probabilities, updating and optimal decision making do no longer coincide.

Keywords

decision making, generalized risk, generalized expected loss, imprecise prior risk and posterior loss, robust Bayesian analysis, Generalized Bayes rule

1 Introduction

A powerful method of inference has to provide answers to (at least) the following three questions:

- What is updating?
- How to learn from data? (inference)
- How to make optimal decisions?

The classical Bayesian statistical theory, based on precise probabilities, claims to provide a comprehensive framework to deal with all these aspects simultaneously. For a Bayesian, inference and decision making coincide, and the solution to both tasks is essentially based on updating prior probabilities by means of the Bayes rule. More precisely, Bayesian statistics is based on two paradigms [P1] and [P2], where

[P1] Every uncertainty can adequately be described by a classical probability distribution. This in particular allows to assign a prior distribution $\pi(\cdot)$ on parameter spaces in inferential problems and on the space of states of nature in decision problems.

[P2] After having observed the sample $\{x\}$, the posterior $\pi(\cdot|x)$ contains all the relevant information. Every inference procedure depends on $\pi(\cdot|x)$, and only on $\pi(\cdot|x)$.

There are several strong arguments for [P2], see, for instance, the discussion in [25]. Among them is the decision theoretic foundation by the often so-called 'main theorem of Bayesian decision theory': As discussed below, it says that decision functions with minimal risk under a prior $\pi(\cdot)$ can be constructed from considering optimal actions with respect to the posterior probability $\pi(\cdot|x)$ as an 'updated prior'.

In the last decade a rapidly increasing number of researches have objected against $[P1]$, and so theories of imprecise probabilities and interval probability emerged (see, e.g., the monographs by Walley [33], Kuznetsov [22], Weichselberger [39], the conference proceedings de Cooman, Fine, Moral and Seidenfeld [6] and the web page de Cooman and Walley [7]), offering a comprehensive framework to deal with a more realistic and reliable description of uncertainty. In this context also concepts generalizing conditional probability have been developed, suggesting the straightforward extension of $[P2]$, namely to use imprecise posteriors to update imprecise priors. This approach is discussed, among others, by Levi ([23],[24]), and is rigorously justified by general coherence axioms in Walley's theory ([33]). Moreover, it is even often understood as self-evident, and applied in many cases without a moment of hesitation, for instance, in the robust Bayesian Analysis (e.g., [35, 26]) and in economic applications following Kofler and Menges' [21] approach of decision making under linear partial information.[1]

The self-evidence of this way to proceed is questioned here. From a rigorous decision theoretic point of view, which is taken up in this paper, it is becoming clear without any ifs and buts that – quite surprisingly – such a procedure may be suboptimal: the resulting decision function may have higher risk than the optimal decision function. The present paper wants to illuminate this aspect. To achieve this goal, it proceeds as follows: Section 2 collects basic notions needed

[1] For further references see, e.g., Cozman's survey ([8]) on computational aspects and the references in [41, Section 1].

later from classical decision theory. After recalling some general aspects and terminology from the theory of interval probability in Section 3.1, both ingredients are melt together in Section 3.2, where the general framework for decision making under interval probability developed in [1, 2] is described briefly. Behind this background Section 4 explores the suboptimality of decision functions based on imprecise posteriors, while Section 5 returns to the fundamental questions formulated above and concludes with a short reflection on the consequences to be drawn from the observation made here.

2 Classical Decision Theory

2.1 The Basic Decision Problem and the Data Problem

Classical decision theory provides a formal framework for decision situations under uncertainty. The decision maker aims at choosing an *action* from of a non-empty, finite set $\mathrm{IA} = \{a_1, \ldots, a_i, \ldots, a_n\}$ of possible actions. Apart from trivial border cases, the consequences of every action depend on the true, but unknown *state* of nature $\vartheta \in \Theta = \{\vartheta_1, \ldots, \vartheta_j, \ldots, \vartheta_m\}$. The corresponding outcome is evaluated by a *loss function*

$$\begin{array}{cccc} l & : & (\mathrm{IA} \times \Theta) & \to & \mathrm{I\!R} \\ & & (a, \vartheta) & \mapsto & l(a, \vartheta) \end{array}$$

and by the associated random variable $\mathbf{l}(a)$ on $(\Theta, \mathcal{P}o(\Theta))$ taking the values $l(a, \vartheta)$. For brevity of reference, the relevant components, the set IA of actions, the set Θ of states of nature and the precise loss function[2] $l(\cdot)$, is collected in the triple $(\mathrm{IA}, \Theta, l(\cdot))$, which is called *basic decision problem.*

For many applications it will prove of value to extend the problem by allowing for *randomized actions.* Formally, every randomized action can be identified with a classical probability $\lambda(\cdot)$ on $(\mathrm{IA}, \mathcal{P}o(\mathrm{IA}))$ where $\lambda(\{a\})$, $a \in \mathrm{IA}$, is interpreted as the probability to choose action a. The set of all randomized actions will be denoted by $\Lambda(\mathrm{IA})$. Pure actions, i.e. elements a of IA itself, are identified with the Dirac measure in the point $\{a\}$, and therefore are also understood to be elements of $\Lambda(\mathrm{IA})$. The loss function is extended to the domain $\Lambda(\mathrm{IA}) \times \Theta$ by $l(\lambda, \vartheta_j) := \sum_{i=1}^{n} \lambda(a_i) \cdot l(a_i, \vartheta_j)$. Analogously to $\mathbf{l}(a)$, $\mathbf{l}(\lambda)$ is that random variable which gives the loss of λ in dependence on the true state ϑ.

Quite often it is possible to obtain some information on the states of nature by collecting additional data. Formally, this can be described by an additional ‘experiment’ where the probability $p_\vartheta(\cdot)$ of the outcomes depends on the true state ϑ of nature. Let $\mathcal{X}$ be the sample space of this experiment, and assume

[2] Throughout the paper it is assumed that a (precise) loss function is given. On the *construction* of loss functions in the presence of ambiguity, generalizing the Neumann Morgenstern approach, see, e.g., [14] and the references therein.)

throughout the paper $\mathcal{X}$ to be finite, so that $\mathcal{X} = \{x_1, \dots, x_s, \dots, x_k\}$. The triple $(\mathcal{X}, \mathcal{Po}(\mathcal{X}), (p_\vartheta(\cdot))_{\vartheta \in \Theta})$ is called *sample information*, the basic decision problem together with the sample information *data problem*.

Now the decision problem consists in the choice between *decision functions (strategies)*

$$\begin{array}{rccl} d: & \{x_1, \dots, x_k\} & \rightarrow & \Lambda(\mathrm{I\!A}) \\ & x & \mapsto & d(x) = \lambda\,, \end{array}$$

i.e. functions which map every observation x into a (randomized) action λ which has to be chosen if x occurs. Let $\mathrm{I\!D}$ be the set of all decision functions. Decision functions are compared via their overall expected loss under $p_\vartheta(\cdot)$, i.e. one considers the so called *risk function*

$$R(d, \vartheta) := \sum_{s=1}^{k} l(d(x_s), \vartheta) \cdot p_\vartheta(x_s)\,, \tag{1}$$

which produces, analogous to above, the random variable $\mathbf{R}(d)$.

2.2 Optimality Criteria

If the states of nature are produced by a perfect random mechanism (e.g. an ideal lottery), and the corresponding probability measure $\pi(\cdot)$ on $(\Theta, \mathcal{Po}(\Theta))$ is completely known, the Bernoulli principle is nearly unanimously favored. One chooses that action λ^* which minimizes the expected loss

$$\mathrm{I\!E}_\pi \mathbf{l}(\lambda) = \sum_{j=1}^{m} l(\lambda, \vartheta_j) \cdot \pi(\{\vartheta_j\}) \tag{2}$$

among all $\lambda \in \Lambda(\mathrm{I\!A})$, and that decision function which minimizes the expected risk

$$\mathrm{I\!E}_\pi \mathbf{R}(d) = \sum_{j=1}^{m} R(d, \vartheta_j) \cdot \pi(\{\vartheta_j\}) \tag{3}$$

among all $d \in \mathrm{I\!D}$, respectively.

In most practical applications, however, the true state of nature can not be understood as arising from an ideal random mechanism. And even if so, the corresponding probability distribution will be not known exactly. There are two main directions to proceed in this situation:

Since for a classical subjectivist, or Bayesian, according to $[P1]$, every situation under uncertainty can be described by a single, precise probability measure $\pi(\cdot)$, the lack of such a known random mechanism does not make any important difference to the decision maker. (S)he acts according to *subjective expected loss/risk*. In this context a special terminology became quite common: $\pi(\cdot)$ is called *prior probability*, and the expression in (3) *prior risk*.

In contrast, from the viewpoint of an 'objectivist' it does not make any sense at all to assign a probability on $(\Theta, \mathcal{Po}(\Theta))$. Therefore, the objectivist concludes that the decision maker is completely ignorant about which state of nature will occur; (s)he has to act according to a criterion based on complete ignorance. The most common criterion is the *minimax rule*, which concentrates on the worst state of nature, leading in the basic decision problem to

$$\max_{\vartheta \in \Theta} l(\lambda, \vartheta) \to \min \tag{4}$$

and in the data problem to

$$\max_{\vartheta \in \Theta} R(d, \vartheta) \to \min \ . \tag{5}$$

2.3 The Main Theorem of Bayesian Decision Theory

It is quite an essential characteristic of Bayesian decision theory that an optimal decision function $d^*(\cdot)$ minimizing the prior risk (3) can be obtained by minimizing, for every observation $\{x\}$, the *posterior loss*,

$$\mathbb{E}_{\pi(\cdot|x)}\mathbf{l}(\lambda) = \sum_{j=1}^{m} l(\lambda, \vartheta_j) \cdot \pi(\{\vartheta_j\}|x) \tag{6}$$

where, compared to (2), the prior $\pi(\cdot)$ is replaced by the 'updated prior', i.e., the posterior $\pi(\cdot|x)$. This is the decision theoretic foundation for the usual Bayesian updating (see also $[P2]$ from the Introduction). More precisely this fundamental relation is formulated in

Proposition 1 ("Main theorem of Bayesian decision theory") [3] *Consider a data problem, consisting of a basic decision problem $(\mathbb{A}, \Theta, l(\cdot))$, a sample information $(\mathcal{X}, \mathcal{Po}(\mathcal{X}), (p_\vartheta(\cdot))_{\vartheta \in \Theta})$ and a prior probability $\pi(\cdot)$. For every $s = 1, \ldots, k$, let $\pi(\cdot|x_s)$ be the corresponding posterior given x_s, and λ_s^* be an optimal solution to the basic decision problem with respect to $\pi(\cdot|x_s)$, i.e. an action minimizing (6).*

Then $d^* := (\lambda_1^*, \ldots, \lambda_s^*, \ldots, \lambda_k^*)$ is an optimal decision function minimizing (3).

Remark 1 *The property formulated in Proposition 1 is constitutive for Bayesian decision making. In particular, an analogous reduction of the data problem to basic decision problems is not possible for the maximin criterion (4) and (5).*

3 Decision Making under Interval Probability

It has often been complained that both classical ways to proceed – relying on subjective expected loss as well as acting according to a criterion based on complete ignorance – are inappropriate, because they both distort the *partial* nature of

[3] Compare, for instance, [4, p. 159, Result 1].

the knowledge on the decision maker's hand: The objectivist's criteria treat partial knowledge like complete ignorance, often leading to unsatisfactory, overpessimistic solutions. Subjective utility/loss theory on the other hand identifies partial knowledge with complete probabilistic knowledge. This conflicts with Ellsberg's [11] experiments, which made it perfectly clear that ambiguity (i.e. the deviation from ideal stochasticity) plays a constitutive role in decision making — neglecting it may lead to deceptive conclusions.

Imprecise probabilities and related concepts are understood to provide a powerful language which is able to reflect the partial nature of the knowledge suitably and to express the amount of ambiguity adequately. (See [7] and [39, Ch. 1] for recent reviews on the development in this field.)

3.1 Basic Terminology of Interval Probability

With respect to the intended application the whole consideration is restricted here to the case of a finitely generated algebra $\mathcal{A}$ based on a sample space Ω. Then, without loss of generality, Ω is finite, and $\mathcal{A}$ is the power set of $\Omega = \{\omega_1, \dots, \omega_k\}$.

To distinguish in terminology, every probability measure in the usual sense, i.e. every set function $p(\cdot)$ satisfying Kolmogorov's axioms is called a *classical probability*. The set of all classical probabilities on the measurable space $(\Omega, \mathcal{A})$ will be denoted by $\mathcal{K}(\Omega, \mathcal{A})$.

Axioms for interval-valued probabilities $P(\cdot) = [L(\cdot), U(\cdot)]$ can be obtained by looking at the relation between the non-additive set-function $L(\cdot)$ and $U(\cdot)$ and the set of classical probabilities being in accordance with them. On a finite sample space, as considered throughout this paper, several concepts of interval probability coincide. They all are concerned with set-functions

$$\begin{array}{rcrcl} P(\cdot) & : & \mathcal{A} & \to & \mathcal{Z}_0 := \{[L,U] \mid 0 \le L \le U \le 1\} \\ & & A & \mapsto & P(A) = [L(A), U(A)] \end{array}$$

with

$$\mathcal{M} := \{p(\cdot) \in \mathcal{K}(\Omega, \mathcal{A}) \mid L(A) \le p(A) \le U(A),\ \forall A \in \mathcal{A}\} \neq \emptyset\,. \tag{7}$$

and

$$\left.\begin{array}{l} \inf\limits_{p(\cdot)\in\mathcal{M}} p(A) = L(A) \\ \sup\limits_{p(\cdot)\in\mathcal{M}} p(A) = U(A) \end{array}\right\} \quad \forall A \in \mathcal{A}\,. \tag{8}$$

Such $P(\cdot)$, and the corresponding set functions $L(\cdot)$ and $U(\cdot)$, are called lower and upper probability ([17]), envelopes ([34, 9]), coherent probability ([33]) and F-probability ([37, 38, 39]). In the game theoretic setting $\mathcal{M}$ is the 'core'. Here Weichselberger's terminology is used calling $\mathcal{M}$ *structure*. Note that, by (8), there is a one-to-one correspondence between $P(\cdot)$ and the structure $\mathcal{M}$.

Two-monotone capacities ([17], also called supermodular capacities ([9]) or convex capacities ([18]), as well as belief functions ([28, 42]) are special cases. More general sets of classical probabilities are obtained by the theory of coherent previsions ([33]), i.e. by assigning interval-valued expectations $\mathbb{E}_{\mathcal{M}}(\cdot) := [{}^{\mathrm{L}}\mathbb{E}_{\mathcal{M}}(\cdot), {}^{\mathrm{U}}\mathbb{E}_{\mathcal{M}}(\cdot)]$ on a set $\mathcal{K}$ of random variables on $(\Omega, \mathcal{A})$. By the lower envelope theorem ([33, p.134]) and the fact that classical expectation and classical probabilities uniquely correspond with each other, the definition of coherence can be rewritten in a way similar to (8). Since Walley [33] did not coin a name for the resulting set of classical probabilities, it will be called *structure*, too.

The interval-valued functions or functionals and the structure are dual concepts, they uniquely determine each other. The results obtained in this paper will be given in terms of the structure.

Many concepts of classical probability theory can be generalized appropriately. For decision making the notion of expectation is the most important one. Looking at the structure $\mathcal{M}$, one way how to define expectation for interval probability and how to extend the functional $\mathbb{E}_{\mathcal{M}}$ to random variables $X \notin \mathcal{K}$ suggests itself (see also the natural extension in [33]): Given a structure $\mathcal{M} \subseteq \mathcal{K}(\Omega, \mathcal{A})$

$$\mathbb{E}_{\mathcal{M}}X := \left[{}^{\mathrm{L}}\mathbb{E}_{\mathcal{M}}X, {}^{\mathrm{U}}\mathbb{E}_{\mathcal{M}}X\right] := \Big[\inf_{p(\cdot)\in\mathcal{M}} \mathbb{E}_p X\,,\ \sup_{p(\cdot)\in\mathcal{M}} \mathbb{E}_p X\Big] \tag{9}$$

is the *(interval-valued) expectation of X (with respect to $\mathcal{F}$)*.[4]

3.2 Generalized Expected Loss and Risk

In this section the decision problem as described in the Introduction will be analyzed in the situation where the decision maker's knowledge on the states of nature is ambiguous, expressed by a structure $\mathcal{M}$ of classical probabilities on $(\Theta, \mathcal{P}o(\Theta))$. To focus the argumentation on the essential ideas, it is assumed that the sampling information consists of classical probabilities.[5]

The generalization of the concept of probability now allows to consider generalized prior probabilities describing the decision maker's state of knowledge. With the notion of interval-valued expectation from (9) one immediately obtains the basic element of a generalized decision theory:

Definition 1 *Consider the basic decision problem $(\mathbb{A}, \Theta, l(\cdot))$, a structure $\mathcal{M} \subseteq \mathcal{K}(\Theta, \mathcal{P}o(\Theta))$, and a sample information $(\mathcal{X}, \mathcal{P}o(\mathcal{X}), (p_\vartheta(\cdot))_{\vartheta\in\Theta})$. For every (randomized action) $\lambda \in \Lambda(\mathbb{A})$, and every decision function $d \in \mathbb{D}$, the expectations*

[4] An alternative way to define expectation for non-additive set functions is the *Choquet integral (or fuzzy integral)* (c.f., e.g., [9]). For the case of two-monotone and totally monotone capacities both notions are equivalent (cf., e.g., [9, Prop. 10.3, p. 126]). Therefore, the results developed below are then valid for the Choquet integral, too.

[5] The whole framework can be extended to imprecise sample information without substantial difficulties (cf., also the brief outline in [1]).

$\mathbb{E}_{\mathcal{M}}\mathbf{l}(\lambda)$ *and* $\mathbb{E}_{\mathcal{M}}\mathbf{R}(d)$ *are the* generalized expected loss *and the* generalized expected risk *(with respect to the prior information $\mathcal{M}$), respectively.*

Note that $\mathbb{E}_{\mathcal{M}}\mathbf{l}(\lambda)$ and $\mathbb{E}_{\mathcal{M}}\mathbf{R}(d)$ are interval-valued quantities. In most cases, comparing the generalized expected loss of actions directly will lead only to partial orderings on $\mathbb{A}$ and $\Lambda(\mathbb{A})$. If a linear (complete) ordering of actions is desired, an appropriate *representation* is needed. This is a mapping from $\mathbb{R}\times\mathbb{R}$ to $\mathbb{R}$ which evaluates intervals by real numbers to use the natural ordering on $\mathbb{R}$ for distinguishing optimal actions.

Expressing the probabilistic knowledge by a structure means that inside the structure there is complete ignorance: none of the elements of the structure is ‘more likely’ than another one. Therefore several authors (see the literature cited below) suggested to apply ‘the maximin criterion to the structure’. Then the interval-valued expectations are represented by the upper interval limit alone. Accordingly, an action λ^* or a decision function d^* is optimal iff

$$^{\mathrm{U}}\mathbb{E}_{\mathcal{M}}(\mathbf{l}(\lambda^*)) \leq {}^{\mathrm{U}}\mathbb{E}_{\mathcal{M}}(\mathbf{l}(\lambda)), \quad \forall\lambda\in\Lambda(\mathbb{A}). \tag{10}$$

and

$$^{\mathrm{U}}\mathbb{E}_{\mathcal{M}}(\mathbf{R}(d^*)) \leq {}^{\mathrm{U}}\mathbb{E}_{\mathcal{M}}(\mathbf{R}(d)), \quad \forall d\in\mathbb{D}, \tag{11}$$

respectively. The criterion (10) corresponds, among others, to the Maxmin expected utility model ([15]) and to the MaxEMin criterion considered by Kofler and Menges ([21]; cf. also [20] and the references therein)). (11) is also called Gamma-Minimax principle (e.g. [4, Section 4.7.6],[32]). These criteria will be used in this paper, too.[6]

Remark 2 *It should be noted that the criterion considered here contains the two main classical decision criteria as border cases: If there is perfect probabilistic information and therefore no ambiguity, then $\mathcal{M}$ consists of one single classical prior probability $\pi(\cdot)$ only; (10) and (11) coincide with Bayes optimality with respect to $\pi(\cdot)$. On the other hand, in the case of completely lacking information, the prior information consists of all classical probabilities on $(\Theta,\mathcal{P}o(\Theta)$ (*‘non-selective’ *or* ’vacuous’ *prior). Then it is easily derived that*

$$^{\mathrm{U}}\mathbb{E}_{\mathcal{M}}(l(\lambda)) = \min_{j\in\{1,\ldots,m\}} l(d,\vartheta_j) \quad \textit{and} \quad {}^{\mathrm{U}}\mathbb{E}_{\mathcal{M}}(\mathbf{R}(d)) = \max_{j\in\{1,\ldots,m\}} R(d,\vartheta_j)\,,$$

and (10) as well as (11) lead to the minimax criterion.

[6] This is done, however, without claiming that this is the only appropriate choice. Indeed, already in the seminal paper by Ellsberg [11] there are strong arguments for additionally taking into account other criteria. A convenient and nevertheless flexible choice is a linear combination of lower and upper limits (compare, e.g., with [11, p. 664], [18],[40], [39, Ch. 2.6]).

4 Robust Bayesian Analysis and Generalized Bayes Rule

4.1 Posterior Loss Analysis

The search for a decision function is much more costly than the calculation of optimal actions. Therefore, a natural attempt to solve (11) relies on the idea of the main theorem of Bayesian decision theory (compare Proposition 1): after having observed $\{x\}$, calculate the (now imprecise) posterior to update the imprecise prior, and then determine the action minimizing posterior loss.

Before discussing properties of this way to proceed in detail, the informal description just given has to be made precise:

Definition 2 *Consider the basic decision problem* $(\mathrm{I\!A},\Theta,l(\cdot))$*, a structure* $\mathcal{M} \subseteq \mathcal{K}(\Theta,\mathcal{P}o(\Theta))$*, and a sample information* $(\mathcal{X},\mathcal{P}o(\mathcal{X}),(p_\vartheta(\cdot))_{\vartheta\in\Theta})$*. Assume that* $\pi(\{\vartheta\})>0, \forall\vartheta\in\Theta$*,* $\forall\pi\in\mathcal{M}$*.*

i) Then, for every $x\in\mathcal{X}$*, call*

$$\mathcal{M}_{\cdot|x} = \{\pi(\cdot|x)|\pi\in\mathcal{M}\} \tag{12}$$

the imprecise posterior *given x, and* $\lambda^*\in\Lambda(\mathrm{I\!A})$ *with*

$$^{\mathrm{U}}\mathrm{I\!E}_{\mathcal{M}\cdot|x}(l(\lambda^*,\vartheta_j)) \leq {}^{\mathrm{U}}\mathrm{I\!E}_{\mathcal{M}\cdot|x}(l(\lambda,\vartheta_j)) , \quad \forall\lambda\in\Lambda(\mathrm{I\!A}) , \tag{13}$$

an optimal action with respect to the posterior loss *given x.*[7]

ii) A decision function $\tilde{d}=(\tilde{d}(x_1),\ldots,\tilde{d}(x_s))$ *where, for every* $s=1,\ldots,k$*, the action* $\tilde{d}(x_s)$ *is optimal with respect to the posterior loss given* x_s*, is called* posterior loss optimal decision function.

The imprecise posterior from (12) is the main tool in robust Bayesian analysis (e.g., [35]), and its use is understood as self-evident in the decision theoretic work based on the theory of linear partial information ([21] and subsequent work). Moreover, a strong justification is provided by Walley's [33] theory. The calculation of $\mathcal{M}_{\cdot|x}$ is equivalent to applying his generalized Bayes rule, which is thoroughly derived from general axioms on coherent updating (cf. [33]). And indeed − next to its intuitive plausibility − working with the imprecise posterior has many further appealing properties. For instance, it is a vivid tool to reflect prior-data conflict ([33, p.6]) and it is naturally applied in successive updating where the imprecise posterior serves as an imprecise prior, once additional data are available.[8]

[7] Vidakovic [32] calls such optima *conditional* Gamma-Minimax solutions.

[8] See, however, [41, Section 6].

4.2 Suboptimality of Posterior Loss Optimal Decision Functions

Though this procedure seems to suggest itself, it must, however, be noted that its decision theoretic foundation is lost. As has to be discussed here, the decision function constructed along the lines of Part ii) of Definition 2 may be **suboptimal** with respect to the criterion (11).

A very simple counterexample can be obtained from a border case: Consider the vacuous prior information $\mathcal{K}(\Theta,\mathcal{P}o(\Theta))$. Then, independent of x, also the imprecise posterior is vacuous[9]. Using it as the ‘updated prior’ yields, for every x, according to Remark 2, the maximin solution λ^{mm} of the basic decision problem as the optimal randomized action. In contrast, the optimal decision function coincides with the maximin decision function $d^{mm}(\cdot)$ of the data problem. Typically, $d^{mm}(\cdot)$ has lower risk than the decision function $\tilde{d} = (\lambda^{mm},\lambda^{mm},\ldots,\lambda^{mm})$. Other counterexamples can be obtained, for instance, by considering situations, where the posterior probabilities are dilated (for this phenomenon see: [31, 36]).

The relation to minimax solutions goes far beyond the border case counterexample just given. Indeed, the following representation theorem even shows that optimal actions in the sense of (10) and optimal decision functions according to (11) *are minimax* solutions (in a different decision problem, where the structure serves as the set of states of nature) — except in the case of classical probability where the structure consists of a single element only. Therefore, the optimal solution must share all the (un)pleasant properties of minimax solutions, and so a reduction of the data problem to smaller basic decision problems cannot be expected; the equivalence of optimality with respect to posterior loss and to prior risk has to be given up.[10]

Theorem 1 (Representation Theorem) *Consider the basic decision problem* $(\mathrm{IA},\Theta,l(\cdot))$, *the prior structure* $\mathcal{M} \subseteq \mathcal{K}(\Theta,\mathcal{P}o(\Theta))$, *and a sample information* $(\mathcal{X},\mathcal{P}o(\mathcal{X}),(p_\vartheta(\cdot))_{\vartheta\in\Theta})$. *Then the following equivalences hold:*

i) An action λ^* *is optimal with respect to the criterion (10), iff it is minimax action in the basic decision problem* $(\Lambda(\mathrm{IA}),\mathcal{M},\tilde{l}(\cdot))$ *with*

$$\begin{array}{rclcll} \tilde{l} & : & (\Lambda(\mathrm{IA})\times\mathcal{M}) & \to & \mathrm{IR} & \\ & & (\lambda,\pi) & \mapsto & \tilde{l}(\lambda,\pi) & := \mathrm{IE}_\pi(l(\lambda,\vartheta))\,. \end{array}$$

ii) A decision function $d^*(\cdot)$ *is optimal with respect to the criterion (11), iff* $d^*(\cdot)$ *is minimax solution in the basic decision problem* $(\mathcal{D},\mathcal{M},\tilde{R}(\cdot))$ *with*

$$\begin{array}{rclcll} \tilde{R} & : & (\mathcal{D}\times\mathcal{M}) & \to & \mathrm{IR} & \\ & & (d,\pi) & \mapsto & \tilde{R}(d,\pi) & := \mathrm{IE}_\pi(R(d,\vartheta))\,. \end{array}$$

[9] See, for instance, [33, p.308].

[10] For the same reason also the essential completeness of unrandomized actions, known from classical Bayesian theory, is no longer valid.

Sketch of the proof: For Part i) read the criterion (10)

$$\max_{\pi(\cdot)\in\mathcal{M}} \mathbb{E}_\pi(l(\lambda,\vartheta)) \to \min$$

from the viewpoint of the minimax criterion (4), where Θ has been replaced by $\mathcal{M}$. To show Part ii), analogously rewrite (11) in the light of (5).

The basic idea of this theorem is similar to Schneeweiß' [27] representation of a basic decision problem. A closer study of the proof shows that this theorem can also be directly extended to imprecise sample information and to the Hurwicz-like optimality criteria briefly mentioned in Footnote 6. Moreover, the fact that in this representation the structure $\mathcal{M}$ now serves as the set of states of nature provides straightforwardly a framework for decision making with second order probabilities: in this setting, a prior weighing the states of nature is nothing but a second order distribution.

5 Concluding Remarks

The paper showed that, for imprecise probability, optimality with respect to prior risk and to posterior loss need no longer coincide. Decision functions constructed by collecting, for every potential observation $x \in \mathcal{X}$, the optimal actions given the corresponding imprecise posterior structure may have higher risk than the direct solution to (11). From the computational point of view this means that, in order to calculate the risk minimizing solution, the reduction to small, easy to solve basic decision problems, which is characteristic for the Bayesian approach in the classical setting, is not possible any more; it is indispensable to go the costly way, fraught with difficulty, via the optimal decision *function*. Efficient algorithms solving this challenge in contexts of optimal design and testing are provided by Fandom Noubiap and Seidel [12, 13]. Augustin [1, 3] gives a general algorithm which is, in principle, applicable to arbitrary decision problems on finite spaces.

Concerning the foundations of statistics it is remarkable that, in the area of imprecise probabilities, the intensive debate between frequentists and Bayesians on topics like counterfactual effects and the principle of conditionality, obtains new importance. Should inference be based only on the concrete observation x, or should one take all potential observations $x \in \mathcal{X}$ into account, i.e., evaluate the decision function as a whole? There are sound arguments for both views and, quite evidently, the author is not the one to decide the question definitely. But, at least, it can be said that one should be aware of the fact that in the area of imprecise probability, in contrast to classical theory, now the standpoint matters; it may influence the results substantially. The imprecise posterior does no longer contain all the relevant information to produce optimal decisions. Inference and decision do not coincide any more — just as in every day life, there is a difference between accumulating as much information as possible (inference and updating knowledge) and making optimal decisions. This may lead to a number of paradoxes, since statisticians up to now have been used to phrase estimating and testing problems equivalently as inference as well as decision problems.

Important further insights into the topic should arise from a deeper understanding of the relationship between the result obtained here and the phenomenon of dilation in conditioning imprecise probabilities as described by Seidenfeld and Wasserman [31] and Wasserman and Seidenfeld [36]. There should also be a close and illuminating connection to Jaffray’s [19] observations on sequential decision making, and to Seidenfeld’s paper ([29]) on incoherence in sequential decision making when preferences fail the independence axiom.[11]

Further research may also attempt at reconciling the conditional and the so-to-say global point of view, the more as the debate on appropriately defining conditional imprecise probabilities is far from being closed. An increasing number of results supports the idea that there should be a symbiosis of several concepts of conditional interval probability ([10, 16, 41] and the references provided there.). There may be some hope to find a notion of conditional probability or a meaningful optimality criterion under which both ways to proceed coincide. In such a setting there would be unanimity on the meaning of terms like ‘updating’, ‘inference’ and ‘optimal decision making’, because then, and only then, the posterior would contain all the relevant information for decision making.

Acknowledgement

I am very grateful to numerous colleagues for clarifying discussions on this topic, which has haunted me for quite a long time. Special thanks go to Kurt Weichselberger, Franz Ferschl, Toni Wallner, Bernhard Rüger, Frank Coolen, Gert de Cooman and Teddy Seidenfeld. My appreciation extends to an anonymous member of the program board and to three anonymous referees.

References

[1] Augustin, T. (2001): On decision making under ambiguous prior and sampling information. In [6], 9-16.

[2] Augustin, T. (2002): Expected utility within a generalized concept of probability — a comprehensive framework for decision making under ambiguity. *Statistical Papers* 43, 5-22.

[3] Augustin, T. (2003): Optimal decisions under complex uncertainty — basic notions and a general algorithm for data-based decision making with partial prior knowledge described by interval probability. Submitted. `www.stat.uni-muenchen.de/~thomas/augustin2003d.pdf`

[11] Concerning the relation between the independence axiom and dilation see [30, Section 3.1].

[4] Berger, J.O. (1984): *Statistical Decision Theory and Bayesian Analysis.* (2nd edition). Springer. New York.

[5] de Cooman, G.,Cozman, F.G., Moral, S. and Walley, P. (Eds.) (1999): *ISIPTA'99: Proceedings of the First International Symposium on Imprecise Probabilities and their Applications. Ghent.*

[6] de Cooman, G., Fine, T.L., Moral, S., and Seidenfeld, T. (Eds.) (2001): *ISIPTA 01: Proceedings of the Second International Symposium on Imprecise Probabilities and their Applications. Cornell University, Ithaca (N.Y.),* Shaker, Maastricht.

[7] de Cooman, G., and Walley, P. (Eds.) (2003): *The Imprecise Probability Project.* `http://www.sipta.org`.

[8] Cozman, F.G. (2000): Computing posterior upper expectations. *Internatiopnal Journal of Approximate Reasoning* 24, 191-205.

[9] Denneberg, D. (1994): *Non-Additive Measure and Integral.* Kluwer. Dordrecht.

[10] Dubois, D., and Prade, H. (1994): Focusing versus updating in belief function theory. In: R.R. Yager, M. Fedrizzi and J. Kacprzyk (Eds.), *Advances in the Dempster-Shafer Theory of Evidence*. Wiley, New York, 71-95.

[11] Ellsberg, D. (1961): Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* **75**, 643-669.

[12] Fandom Noubiap, R., and Seidel, W. (2001a): An algorithm for calculating Γ-minimax decision rules under generalized moment conditions. *Annals of Statistics* 29, 1094-1116.

[13] Fandom Noubiap, R., and Seidel, W. (2001b): An efficient algorithm for constructing Γ-minimax tests for finite parameter spaces. *Computational Statistics and Data Analysis* 36, 145-161.

[14] Ghirardato, P. and Marinacci, M. (2001): Risk, ambiguity, and the separation of utility and beliefs. *Mathematics of Operations Research* 26, 864-890.

[15] Gilboa, I., and Schmeidler, D. (1989): Maxmin expected utility with nonunique prior. *Journal of Mathematical Economics* 18, 141–153.

[16] Halpern, J.Y., and Fagin, R. (1992): Two views of belief: belief as generalized probability and belief as evidence. *Artificial Intelligence* 54, 275-317.

[17] Huber, P.J., and Strassen, V. (1973): Minimax tests and the Neyman-Pearson lemma for capacities. *Annals of Statistics* 1, 251-263; Correct.: 2, 223-224.

[18] Jaffray, J.Y. (1989): Linear utility theory and belief functions. *Operations Research Letters* 8, 107–122.

[19] Jaffray, J.Y. (1999): Rational decision making with imprecise probabilities. In [5], 183-188.

[20] Kofler, E. (1989): *Prognosen und Stabilität bei unvollständiger Information*. Campus. Frankfurt/Main.

[21] Kofler, E., and Menges, G. (1976): *Entscheidungen bei unvollständiger Information.* Springer. Berlin.

[22] Kuznetsov, V.P. (1991): *Interval Statistical Methods.* Radio i Svyaz Publ., (in Russian).

[23] Levi, I. (1974): On indeterminate probabilities. *Journal of Philosophy* 71, 391-418.

[24] Levi, I. (1980): The Enterprise of Knowledge. An Essay on Knowledge, Credal Probability and Chance. MIT Press, Cambridge (MA).

[25] Levi, I. (1990): Consequentialism and sequential choice. In: Bacharch, M. and Hurley, S. (Eds.): *Foundations of Decision Theory.* Blackwells, Oxford; 92-122.

[26] Rios Insua, D.R., and Ruggeri, F. (Eds.) (2000): *Robust Bayesian Analysis.* Springer (Lecture Notes in Statistics 152), New York.

[27] Schneeweiß, H. (1964): Eine Entscheidungsregel im Fall partiell bekannter Wahrscheinlichkeiten. *Unternehmensforschung* 10, 86-95.

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

[29] Seidenfeld, T. (1988): Decision theory without ’independence’ or without ’ordering’. Waht is the difference?. *Economics and Philosophy* 4, 267-290.

[30] Seidenfeld, T. (1994): When normal and extensive form decisions differ. In: Prawitz, D., Skyrms, B. and Westerstahl, D. (Eds.): *Logic, Methodology and Philosophy of Science IX (Uppsala, 1991).* Elsevier, Amsterdam, 451-463

[31] Seidenfeld, T., and Wasserman, L. (1993): Dilation for sets of probabilities. *Annals of Statistics* 21, 1139-1154.

[32] Vidakovic, B. (2000): Γ-minimax: A paradigm for conservative robust Bayesians. In: Insua, D.R., and Ruggeri, F. (Eds.): *Robust Bayesian Analysis.* Springer (Lecture Notes in Statistics 152), New York, 241-259.

[33] Walley, P. (1991): *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall. London.

[34] Walley, P., and Fine, T.L. (1982): Towards a frequentist theory of upper and lower probability. *The Annals of Statistics* 10, 741-761.

[35] Wasserman, L. (1997): Bayesian robustness. In: S. Kotz, C.B. Read, D.L. Banks (Eds.): *Encyclopedia of Statistical Sciences. Update Volume 1*. Wiley, New York, pp. 45-51.

[36] Wasserman, L., and Seidenfeld, T. (1994): The dilation phenomenon in robust Bayesian inference. (With discussion). *Journal of Statistical Planning and Inference* 40, 345-356.

[37] Weichselberger, K. (1995): Axiomatic foundations of the theory of interval-probability. In: V. Mammitzsch, and H. Schneeweiß (Eds.): *Symposia Gaussiana Conference B.* de Gruyter, Berlin, 47-64.

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

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

[40] Weichselberger, K., and Augustin, T. (1998): Analysing Ellsberg's Paradox by means of interval-probability. In: R. Galata, and H. Küchenhoff (Eds.): *Econometrics in Theory and Practice. (Festschrift for Hans Schneeweiß.)* Physika. Heidelberg, 291–304.

[41] Weichselberger, K., and Augustin, T. (2003): On the symbiosis of two concepts of conditional interval probability. Conditionally accepted for: Bernard, J.M., Seidenfeld, T., and Zaffalon, M. (Eds.): *ISIPTA 03: Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications, Lugano.*

[42] Yager, R.R., Fedrizzi, M., and Kacprzyk, J. (Eds.) (1994): *Advances in the Dempster-Shafer Theory of Evidence.* Wiley. New York.

T. Augustin is with the Department of Statistics, Ludwig-Maximilians University of Munich, Ludwigstr. 33, D-80539 München, Germany. E-mail: augustin@stat.uni-muenchen.de

Analysis of Local or Asymmetric Dependencies in Contingency Tables using the Imprecise Dirichlet Model

J.-M. Bernard
Université de Paris 5 & CNRS, France

Abstract

We consider the statistical problem of analyzing the association between two categorical variables from cross-classified data. The focus is put on measures which enable one to study the dependencies at a local level and to assess whether the data support some more or less strong association model. Statistical inference is envisaged using an imprecise Dirichlet model.

Keywords

association, logical model, directional association model, statistical inference, upper and lower probabilities, IDM, prior ignorance, Bayesian inference

1 Introduction

1.1 The problem of association in contingency tables

The problem of measuring association in two-way contingency tables arising from cross-classifications has a long tradition in statistical research (see, *e.g.,*the numerous association measures reviewed by Goodman & Kruskal [6]). Though every one agrees on the meaning of "independence", the opposite notion of "complete association" is felt more ambiguous, because there are several directions in which the data may depart from independence. For the simplest case of 2 × 2 tables, Kendall & Stuart [9] make the distinction between "complete association" (one empty cell) and "absolute association" (two empty cells on either diagonal of the table). Although such distinctions are occasionally mentioned in the literature, most statistical research appears to have focused on proposing global measures of association.

The motivation behind this article arise from two (apparently) independent goals. The first one is to provide a local and/or asymmetric approach to the analysis of contingency tables and to define well-suited descriptive indices for that purpose. The second one is to build the inferential part of the analysis on a generalization of the Bayesian framework, the *imprecise Dirichlet model (IDM)*. Let us comment on these two aspects.

1.2 Analysis of local/asymmetric dependencies: two examples

The first aim of this article is to address two related types of statistical issues, that we shall illustrate by two psychological examples.

Example 1 (Stages data, Logical model) *Jamison [8] studied several cognitive tasks related to the Piaget's stage concept. Table 1 gives the levels attained by a group of children in two tasks, A and B, with three levels each. One model predicts that attaining a given level in task A is a prerequisite for attaining the same level in task B, i.e., predicts that cells* $a1b2$, $a1b3$ *and* $a2b3$ *should be empty. This model can also be expressed as the logical expression* $\mathcal{M} = [b3 \Longrightarrow a3 \ \wedge \ b2 \Longrightarrow (a2 \vee a3)]$. *The issue here is to assess whether a conclusion of* quasi-agreement *of the data with model* $\mathcal{M}$*, can be reached or not.*

Table 1: *"Stages" example. Observed counts* $\boldsymbol{x}$ *for* $n = 101$ *children cross-classified according to their performance level in Seriation of lengths (A) and Inclusion of lengths (B), from [8, p. 248]. For each task, children were classified as "preoperational" (*$a1$ *and* $b1$*), "transitional" (*$a2$ *and* $b2$*) or "operational" (*$a3$ *or* $b3$*). Shaded cells are error cells associated to the logical model* $\mathcal{M} = (b3 \Longrightarrow a3 \ \wedge \ b2 \Longrightarrow (a2 \vee a3))$.

	$b1$	$b2$	$b3$
$a1$	14	0	0
$a2$	15	5	2
$a3$	19	20	26

Example 2 (Dyad data, Directional association model) *Another type of problem is the* study of local dependencies *within an* $A \times B$ *table, which aims at showing that a specified group of cells is over- or under-represented. For example, Danis et al. [5] analyzed data about adult-child verbal interactions in a situation of book reading. Each statement produced by either actor was categorized into one of four levels of increasing complexity. Table 2(left) gives one transition matrix (child statement followed by adult statement) for one dyad. One hypothesis of interest here is that some regions of Table 2(left) should be over- or under-represented according to the pattern shown in Table 2(right): over-representation of statements of the adult at the same level as the child's (denoted "*$+$*"), moderate under-representation of statements at an higher level (denoted "*$-$*"), and high under-representation of statements at a lower level (denoted "*$--$*").*

The two types of questions raised by these examples, either asymmetric and expressed in terms of quasi-agreement with a logical model, or local and expressed in terms of over-/under-representation, can be answered using indices of

Table 2: *Dyad data. Counts of transitions from the child’s statement level (A) to the adult statement level (B) for one dyad (left). Expected pattern of over-representations* $(+)$ *and under-representations* $(-$ *or* $--)$ *(right). Levels correspond to increasing cognitive complexity: “perceptual identification” (*$a1$ *and* $b1$*), “perceptual relationship” (*$a2$ *and* $b2$*), “displaced reference” (*$a3$ *and* $b3$*), and “inferential statement” (*$a4$ *and* $b4$*); categories* $a0$ *and* $b0$ *indicate cases in which one of the actors did not speak.*

	*b*0	*b*1	*b*2	*b*3	*b*4
*a*0	0	25	2	8	0
*a*1	6	27	1	3	2
*a*2	2	0	2	0	0
*a*3	13	0	0	20	2
*a*4	0	2	0	0	0

	*b*0	*b*1	*b*2	*b*3	*b*4
*a*0					
*a*1		+	−	−	−
*a*2		−−	+	−	−
*a*3		−−	−−	+	−
*a*4		−−	−−	−−	+

the same family. Hildebrand *et al.* [7], beside the main trend of research sketched previously, proposed a general index, named *Del*, which measures the degree of agreement of cross-classified data to a specified logical model. The building block of the *Del* index is what [10] call the *association rate* between modalities. Our method will be based on these two indices.

1.3 Inference for local/asymmetric analyses

Several difficulties arise when it comes to making inferences about these indices. The inferential methods that were initially proposed were based on the frequentist framework, and, due to the presence of nuisance parameters, relied on asymptotic arguments (see *e.g.,* [7, Chp. 6]), so that the validity conditions of these methods are satisfied neither for small samples, nor for extreme data sets in which some cells are empty or nearly so. These difficulties come in addition to some fundamental shortcomings of the frequentist methods, and, in particular, the fact that they do not obey the *likelihood principle (LP)*.

The Bayesian approach to inference answers most of these problems. However, it also encounters some difficulties when one wants to make inferences from a prior state of ignorance. None of the various solutions which were proposed for that goal simultaneously satisfies some general desirable principles (see [11]), *i.e.,* the LP, and the *representation invariance principle (RIP)* (invariance with respect to how categories are distinguished).

A generalization of the Bayesian framework, involving *imprecise probabilities*, allows one to overcome most, if not all, of the difficulties of the Bayesian approach, while keeping its attractive features (see [11]). In particular, Walley [12] proposed a new method of inference for categorical data based on the *imprecise Dirichlet model (IDM)*. In the IDM, prior uncertainty about the cells’ true frequencies is described by a set of Dirichlet priors, each of which being updated into a Dirichlet posterior using Bayes' theorem. Posterior uncertainty is described by the set of these Dirichlet posteriors. The IDM has several desirable properties as a model for making inferences from a prior state of near ignorance. Firstly, it satisfies both the LP and the RIP. Secondly, the IDM distinguishes between a relative lack of information (high imprecision) and a more substantial state of knowledge (low imprecision). The IDM can also be viewed as a method for making robust inferences.

Our purpose here is to apply the IDM to the problem of studying the association in contingency tables. This article contains relatively few new results about the IDM itself, but we think it is important to face the IDM with several types of applications and data sets, in order to develop more insights about its properties and the scope of its application.

This article is structured as follows. Section 2 defines local or asymmetric association measures. Sections 3 and 4 review the usual Bayesian Dirichlet models and the IDM, respectively. Our main contribution is the study of inferences about association measures from the IDM which is presented in Sections 5 and 6.

2 Descriptive analysis: defining relevant indices

Consider a data set of size n categorized in K categories, with observed counts $\boldsymbol{x} = (x_1, \ldots, x_K)$, with $n = \sum_k x_k$. The observed (relative) frequencies are denoted $\boldsymbol{f} = (f_1, \ldots, f_K)$, with $f_k = x_k/n$. The data $\boldsymbol{x}$ will be considered as a sample from a larger population, characterized by the parent or true frequencies $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, which are the population counterparts of $\boldsymbol{f}$. Both $\boldsymbol{f}$ and $\boldsymbol{\theta}$ belong to the K-dimensional unit simplex $\mathcal{S}(1,K)$. Throughout this paper, the generic expression "*association model*" (or simply "*model*") denotes some summary statement about a frequency-vector, either $\boldsymbol{f}$ or $\boldsymbol{\theta}$, *i.e.*, a statement saying that it belongs to some subset $\mathcal{R} \subset \mathcal{S}(1,K)$. The qualifiers "descriptive" and "inductive" are used for models bearing on $\boldsymbol{f}$ and $\boldsymbol{\theta}$ respectively. At the descriptive level, a model is either true or false, whereas, at the inductive level, the model's truth can only be assessed with some probability.

In this section, we define various indices in terms of which the association models considered in this paper will be defined. Here, these indices are defined as functions of $\boldsymbol{f}$, but each one has its inductive counterpart as a function of $\boldsymbol{\theta}$. The problem of making inferences about parameters $\boldsymbol{\theta}$ (and indices derived from them) will be envisaged in later sections.

2.1 Notation and preliminary definitions

The K categories are obtained here as combinations of modalities of the A and B variables, so we shall use more specific notations: ab or (a,b) for a cell of the contingency table, x_{ab} for its observed count, f_{ab} for its observed frequency; we note $f_a = \sum_b f_{ab}$ and $f_b = \sum_a f_{ab}$ the marginal frequencies of categories $a \in A$ or $b \in B$, and $\widehat{f_{ab}} = f_a f_b$ the *product-frequency* of cell *ab*.[1]

Definition 1 (Local independence) *There is* local independence *between modalities a and b, noted $a \perp\!\!\!\perp b$, whenever $f_{ab} = \widehat{f_{ab}}$.*

Definition 2 (Global independence) *There is* global independence *between variables A and B, noted $A \perp\!\!\!\perp B$, whenever $\forall a \in A, b \in B,\ \ a \perp\!\!\!\perp b$.*

2.2 The association rates as measures of local association

Being interested in the association between variables A and B amounts to being interested in the departures from global independence, *i.e.*, all departures from local independence. This is done by introducing a measure of local association.

Definition 3 (Association rate, [10]) *The association rate between a and b is defined as $t_{ab} = (f_{ab} - \widehat{f_{ab}})/(\widehat{f_{ab}})$.*

The sign of t_{ab} indicates whether there is an *attraction* (case $t_{ab} > 0$), a local independence (case $t_{ab} = 0$), or a *repulsion* (case $t_{ab} < 0$) between a and b. The maximum repulsion is obtained when $t_{ab} = -1$, *i.e.*, when $f_{ab} = 0$, but there is no *a priori* upper limit for t_{ab}. The index t_{ab} can also be interpreted as a *over-* or *under-representation rate* of cell *ab* with respect to the $a \perp\!\!\!\perp b$ case: for example, $t_{ab} = +0.50$ (resp. -0.50), indicates that cell *ab* contains 50% *more* (resp. *less*) observations than in the $a \perp\!\!\!\perp b$ case.

2.2.1 Properties of association rates

As should be clear from properties given below (see also [10, Chp. 7]), the product-frequencies $\widehat{\boldsymbol{f}} = (\widehat{f_{ab}})_{a\in A, b\in B}$ must be considered as a canonical set of weights for $\boldsymbol{t} = (t_{ab})_{a\in A, b\in B}$. In the following, we denote $Mean_R(\boldsymbol{t}, \widehat{\boldsymbol{f}})$ the weighted mean of $\boldsymbol{t}$ (with weights $\widehat{\boldsymbol{f}}$) over $R \subset A \times B$ (R being omitted when $R = A \times B$).

Property 1 *The marginal weighted average of $\boldsymbol{t}$, for any $a \in A$ or any $b \in B$, is equal to* 0, *i.e.*, $\mathrm{Mean}_{\{(a,b), b\in B\}}(\boldsymbol{t}, \widehat{\boldsymbol{f}}) = 0$ *and* $\mathrm{Mean}_{\{(a,b), a\in A\}}(\boldsymbol{t}, \widehat{\boldsymbol{f}}) = 0$.

Corollary 1 *If in any row a (resp. column b) some t_{ab} is positive, then some other $t_{ab'}$ (resp. $t_{a'b}$) is negative: over-representation of some cells implies the existence of some under-represented cells. In particular, for a 2×2 table, $a \perp\!\!\!\perp b$ implies $A \perp\!\!\!\perp B$.*

[1] Throughout this paper, we use K to denote both the set of categories and its cardinal, and similarly for A and B, the distinction being always clear from the context. Unless otherwise stated, all sums over k (resp. a, b) run from 1 to K (resp. A, B).

Property 2 (Pooling) *Consider two applications $A \longrightarrow A^*$ and $B \longrightarrow B^*$ and the pooled table, $A^* \times B^*$, then, $\forall a^* \in A^*, \forall b^* \in B^*$, $t_{a^*b^*} = \text{Mean}_{\{(ab),a\in a^*,b\in b^*\}}(\boldsymbol{t}, \widehat{\boldsymbol{f}})$. In particular, consider cell ab and the pooled table $A^* \times B^*$, where $A^* = \{a, a'\}$ and $B^* = \{b, b'\}$. Then t_{ab} is unchanged, whether it is defined from table $A \times B$ or from $A^* \times B^*$.*

Note 1 (Global independence and $\boldsymbol{t}$) *From Definitions 2 and 3, $A \perp\!\!\!\perp B$ occurs if and only if the t_{ab}'s are all equal to* 0*. Conversely, the departure of any t_{ab} from* 0 *indicates a departure from independence. What is important here is that the precise pattern of the t_{ab}'s departures from* 0 *points to the* direction of association.

2.2.2 Example: Dyad data (continued)

Table 3 gives the t_{ab}'s for all cells of Table 2(left). Descriptively, *(i)* all diagonal cells but one are over-represented, *(ii)* all cells below the diagonal but one are maximally under-represented, and *(iii)* four of the six cells above the diagonal are under-represented (two maximally). Several of these results go in the direction of the pattern of Table 2(right), but this model, if taken at the cell level, is not descriptively satisfied.

2.3 Mean association rate over a region R: index t_R

In order to express the idea that some region $R \subset A \times B$ is over- or under-represented, we shall have recourse to a more global index as in [5].

Definition 4 (Mean association rate) *Given a region $R \subset A \times B$, the* mean association rate *over R is defined as, $t_R = \text{Mean}_R(\boldsymbol{t}, \widehat{\boldsymbol{f}})$.*

The index t_R varies from -1 (all cells in R are empty), to negative values (under-representation of R), to 0 (independence on average in R), to positive values (over-representation of R) without any a priori upper bound.

Table 3: *Dyad data. Observed association rates t_{ab} from data of Table 2.*

	$b0$	$b1$	$b2$	$b3$	$b4$
$a0$	-1.00	0.52	0.31	-0.15	-1.00
$a1$	-0.16	0.47	-0.41	-0.71	0.47
$a2$	1.74	-1.00	10.50	-1.00	-1.00
$a3$	1.03	-1.00	-1.00	1.12	0.64
$a4$	-1.00	1.13	-1.00	-1.00	-1.00

2.3.1 Example: Dyad data (continued)

Consider the Dyad data and the pattern of over-/under-representation of Table 2(right). One possible way to confront the data to this model, at a descriptive level, is to compute the observed mean association rates for the three regions, D for cells on the diagonal, U for cells above and L for cells below it. This yields $t_D = 0.75$, $t_U = -0.50$ and $t_L = -0.91$. A global descriptive summary of the data, which goes in the direction of the expected pattern, is thus: $t_D > 0 > t_U > t_L$.

2.4 The *Del* index, a measure of agreement with a logical model

2.4.1 Quasi-implication for a 2×2 table

Consider a 2×2 table, with binary variables $A = \{a, a'\}$ and $B = \{b, b'\}$. We assimilate a and b to logical propositions, and denote negation by priming, conjunction by concatenation, implication by $\Longrightarrow$, and the false proposition by $\emptyset$. Then the statement $a \Longrightarrow b$ (*i.e.*, any observation of type a is necessarily of type b) is equivalent to $ab' \Longrightarrow \emptyset$, *i.e.*, that cell ab' is empty (cell ab' is an *error cell* for model $a \Longrightarrow b$, see [7]). Bernard [4] weakened the notion of a strict implication $a \Longrightarrow b$ into that of a *quasi-implication*, denoted by $a \longrightarrow b$, by defining the descriptive index $d_{a \Longrightarrow b} = -t_{ab'}$ as a measure of the *degree of agreement* with the logical model $a \Longrightarrow b$. For a given threshold $d_{quasi} > 0$, quasi-implication was defined by: $a \longrightarrow b \iff d_{a \Longrightarrow b} \geq d_{quasi}$.

2.4.2 Generalization to any logical model, the *Del* index

Definition 5 (Del index, [7]) *More generally, consider a logical model $\mathcal{M}$ relative to an $A \times B$ table, and denote by $\mathcal{E}_{\mathcal{M}}$, or $\mathcal{E}$ for short, the set of all error cells that contradict $\mathcal{M}$, i.e., such that $\mathcal{M} = \bigwedge(ab \Longrightarrow \emptyset)_{(a,b) \in \mathcal{E}}$. Let $t_{\mathcal{E}}$ be the mean association rate over region $\mathcal{E}$. Then a global measure of the degree of agreement of the data with $\mathcal{M}$ is the Del index, $d_{\mathcal{M}} = -t_{\mathcal{E}}$.*

Properties of $d_{\mathcal{M}}$ flow from those of (mean) association rates. The index $d_{\mathcal{M}}$ varies in the range $]-\infty, 1]$; $d_{\mathcal{M}} = 0$ in case of independence on average in region $\mathcal{E}$ and $d_{\mathcal{M}} = 1$ when $\mathcal{M}$ is verified. A value of $d_{\mathcal{M}}$ between 0 and 1 can thus be interpreted as a quasi-agreement of the data with $\mathcal{M}$ at degree $d_{\mathcal{M}}$; the closer to 1 its value is, the better the quasi-agreement is.

Property 3 (Equivalent logical models) *Consider two logical models $\mathcal{M}_1$, defined on $A \times B$, and $\mathcal{M}_2$, defined on a table $A^* \times B^*$ obtained by coarsenings of the A and B classifications, such that $\mathcal{M}_1$ and $\mathcal{M}_2$ are logically equivalent. Then, $d_{\mathcal{M}_1} = d_{\mathcal{M}_2}$. This property follows from Property 2.*

As seen from Definition 5, $d_{\mathcal{M}}$ and t_R are equivalent indices. In using t_R, we want to stress the over-/under-representation interpretation and the independence case as a privileged reference ($t_R = 0$), whereas, in using $d_{\mathcal{M}}$, we stress the interpretation in terms of quasi-agreement with a strong/logical model and we point model $\mathcal{M}$ as a privileged reference ($d_{\mathcal{M}} = 1$).

2.4.3 Example: Stages data (continued)

Consider the Stages data in Table 1 and the logical model $\mathcal{M}$ associated with $\mathcal{E} = \{(a1,b2),(a1,b3),(a2,b3)\}$. We see that only two observations fall in region $\mathcal{E}$ and we find $d_{\mathcal{M}} = 0.851$. Descriptively, at threshold, say, $d_{quasi} = 0.80$, we may conclude that the data quasi-agree with model $\mathcal{M}$.

3 Bayesian inference

We now assume that the data $\boldsymbol{x} = (x_1,\dots,x_K)$ is a multinomial sample (with $K = A \times B$ categories) of size n from a population characterized by the unknown parameters $\boldsymbol{\theta} = (\theta_1,\dots,\theta_K)$, the true frequencies of the K categories: $\boldsymbol{x} \sim Mn(n,\boldsymbol{\theta})$. We now want to make inferences about $\boldsymbol{\theta}$, and, more precisely here, about derived parameters such as τ_{ab}, τ_R and $\delta_{\mathcal{M}}$ which are the population counterparts of the descriptive indices t_{ab}, t_R and $d_{\mathcal{M}}$.

3.1 Dirichlet model for $\boldsymbol{\theta}$

In the usual Bayesian conjugate analysis, prior uncertainty about $\boldsymbol{\theta}$ is described by a Dirichlet prior distribution, $\boldsymbol{\theta} \sim Diri(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_K)$ and each hyper-parameter $\alpha_k > 0$. We call the α_k's the *prior strengths* and $\nu = \sum_k \alpha_k$ the *total prior strength*. We shall use an alternative parameterization of the Dirichlet in terms of the *prior frequencies* $\boldsymbol{\varphi} = \boldsymbol{\alpha}/\nu$, where $\boldsymbol{\varphi} \in \mathcal{S}^\star(1,K)$ and $\mathcal{S}^\star(1,K)$ denotes the interior of simplex $\mathcal{S}(1,K)$.[2] The prior expectations are simply $E(\theta_k) = \varphi_k$. The posterior distribution on $\boldsymbol{\theta}$ is then an updated Dirichlet distribution, $\boldsymbol{\theta}|\boldsymbol{x} \sim Diri(\boldsymbol{x}+\boldsymbol{\alpha}) = Diri(\boldsymbol{x}+\nu\boldsymbol{\varphi})$, with posterior expectations given by,

$$E(\theta_k|\boldsymbol{x}) \quad = \quad \frac{x_k + \nu\varphi_k}{n+\nu}. \tag{1}$$

3.2 Objective Bayesian models

For multinomial data, four Dirichlet priors have been proposed as models for prior ignorance about $\boldsymbol{\theta}$. All are symmetric Dirichlet, that is $\varphi_k = 1/K$ for any k, and they only differ in their respective total prior strength ν: $\nu \to 0$ (Haldane), $\nu = 1$ (Perks), $\nu = K/2$ (Jeffreys) and $\nu = K$ (Bayes-Laplace's uniform prior).

Haldane's improper prior leads to some undesirable inferences: when $x_k = 0$, it leads to infer that $\theta_k = 0$, even when n is small. A major difficulty with the other

[2] Walley [12] uses symbols s and t_k in place of ν and φ_k respectively.

three objective Bayesian priors is that inferences they produce depend on how the K categories are distinguished, which is partly arbitrary, and thus they do not satisfy the RIP (see [12]). Jeffreys' prior does not satisfy the LP either. Although it is often claimed that inferences from these priors differ in a negligible way when n is not small, large discrepancies can be obtained for statements bearing on unobserved or rare cells, even with large n.

4 Imprecise Dirichlet model

4.1 Presentation of the model

Walley [12] proposed the *imprecise Dirichlet model (IDM)* as a model for prior ignorance in the case of categorical data. The model consists in describing prior uncertainty about $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ by a set of Dirichlet priors. The prior IDM(ν) is defined as the set of all Dirichlet priors on $\boldsymbol{\theta}$ with a fixed total prior strength $\nu > 0$, *i.e.*, the set $\{Diri(\boldsymbol{\alpha}) : \alpha_k > 0$ for all k, $\sum_k \alpha_k = \nu\}$, or equivalently

$$\{Diri(\nu\boldsymbol{\varphi}) : \boldsymbol{\varphi} \in \mathcal{S}^\star(1,K)\}, \tag{2}$$

where $\mathcal{S}^\star(1,K)$ is the interior of the simplex $\mathcal{S}(1,K)$.

Let $P_{\nu\boldsymbol{\varphi}}(\cdot)$ and $E_{\nu\boldsymbol{\varphi}}(\cdot)$ be respectively a prior probability and a prior expectation provided by a particular $Diri(\nu\boldsymbol{\varphi})$ in the set (2). The uncertainty about any event Z concerning $\boldsymbol{\theta}$ is described by *prior lower and upper probabilities*, denoted by $\underline{P}(Z)$ and $\overline{P}(Z)$, and calculated by minimizing and maximizing $P_{\nu\boldsymbol{\varphi}}(Z)$ with respect to $\boldsymbol{\varphi} \in \mathcal{S}^\star(1,K)$. Similarly, for any real-valued function $\lambda = g(\boldsymbol{\theta})$, *prior lower and upper expectations* $\underline{E}(\lambda)$ and $\overline{E}(\lambda)$ are calculated by minimizing or maximizing the expectation $E_{\nu\boldsymbol{\varphi}}(\lambda)$ with respect to $\boldsymbol{\varphi}$. Inferences about λ can be summarized by the *prior lower and upper cumulative distribution functions (cdf's)*, $\underline{F}_\lambda(l) = \underline{P}(\lambda > l)$ and $\overline{F}_\lambda(l) = \overline{P}(\lambda > l)$.

Each Dirichlet prior in the prior IDM(ν) is updated into a Dirichlet posterior using Bayes' theorem. This updating procedure guarantees coherence of the inferences [11]. Hence the posterior uncertainty about $\boldsymbol{\theta}$ from the IDM(ν) is expressed by the set

$$\{Diri(\boldsymbol{x} + \nu\boldsymbol{\varphi}) : \boldsymbol{\varphi} \in \mathcal{S}^\star(1,K)\} \tag{3}$$

As for the prior IDM, posterior lower and upper probabilities, expectations and cdf's are obtained by minimization or maximization with respect to $\boldsymbol{\varphi} \in \mathcal{S}^\star(1,K)$.

The IDM satisfies several desirable principles of inference, and in particular both the LP and the RIP (see [12]). The RIP states that posterior inferences about any derived parameter $\lambda = g(\boldsymbol{\theta})$ should not depend on the number of categories K used for defining λ. The RIP is satisfied by the IDM in so far as the total prior strength ν is specified independently of K.

4.2 Prior and posterior inferences about θ_k from the IDM

The posterior lower and upper expectations of θ_k are given by

$$\underline{E}(\theta_k|\boldsymbol{x}) = x_k/(n+\nu) \quad \text{and} \quad \overline{E}(\theta_k|\boldsymbol{x}) = (x_k+\nu)/(n+\nu), \tag{4}$$

and are obtained as $\varphi_k \to 0$ and $\varphi_k \to 1$ respectively. The two same limiting values lead to the posterior upper and lower cdf's respectively, $\overline{P}(\theta_k > l|\boldsymbol{x})$ which is the $Beta(x_k, n-x_k+\nu)$ cdf, and $\underline{P}(\theta_k > l|\boldsymbol{x})$ which is the $Beta(x_k+\nu, n-x_k)$ cdf.

By setting $n = x_k = 0$ in (4), we see that prior uncertainty about θ_k is maximal. We have $\underline{E}(\theta_k) = 0$ and $\overline{E}(\theta_k) = 1$, and $\underline{P}(\theta_K > l) = 0$ and $\overline{P}(\theta_k > l) = 1$ for any $0 < l < 1$, that is *vacuous* lower and upper probabilities.

4.3 Choice of ν

The IDM as defined in (2) and (3) depends on the choice of ν. The constant ν determines how fast the lower and upper probabilities converge one towards the other as n increases, and can thus be interpreted as a measure of the caution of the inferences. The larger ν is, the more cautious the inferences are. The most important criterion for the choice of ν is the requirement that the IDM should be cautious enough to encompass frequentist or objective Bayesian alternatives, while not being too cautious to avoid too weak inferences.

The first researches about the IDM lead to several convincing arguments for choosing $1 \leq \nu \leq 2$, but most of these arguments are relative to the binary case ($K = 2$) only (see [2, **?**]. More recent work provides some support for $\nu = 2$ in the case of large K, for non-parametric inference about a mean [3]. In the following, we shall use $\nu = 2$, a value which is also supported by results in Section 5.4.

4.4 Two conjectures about the IDM

Conjecture 1 (Expectation of a derived parameter) *Let $\lambda = g(\boldsymbol{\theta})$ be a real-valued function of $\boldsymbol{\theta}$, and* $\mathrm{E}_{\nu\boldsymbol{\varphi}}(\boldsymbol{\theta})$ *the prior (resp. posterior) expectation of $\boldsymbol{\theta}$ under the prior* $\mathrm{Diri}(\nu\boldsymbol{\varphi})$ *(resp. posterior* $\mathrm{Diri}(\boldsymbol{x}+\nu\boldsymbol{\varphi})$*). Then the upper and lower expectations of $\boldsymbol{\theta}$ under the IDM(ν) are obtained from the (or one of the) Dirichlet prior which maximizes (resp. minimizes)* $g(\mathrm{E}_{\nu\boldsymbol{\varphi}}(\boldsymbol{\theta}))$ *with respect to $\boldsymbol{\varphi}$.*

Conjecture 2 (Cdf of a real-valued derived parameter) *Let $\lambda = g(\boldsymbol{\theta})$ be a real-valued function of $\boldsymbol{\theta}$. Let* $\mathrm{Diri}(\nu\boldsymbol{\varphi})$ *be a Dirichlet prior which provides the lower (resp. upper) prior or posterior expectation of λ under the IDM(ν), then it also provides the prior or posterior upper (resp. lower) cdf of λ.*

The two conjectures hold if $g(.)$ is a linear function of the θ_k's. We don't expect them to be true in the general case (there are simple counter-examples to Conjecture 1). Nevertheless, we suggest that these conjectures actually provide reasonable approximations for the lower and upper expectations and cdf's of λ for most functions $g(.)$. In any case, the procedures they induce necessarily lead to an upper (resp. lower) bound for $\underline{E}(\lambda)$ and $\underline{F}_{\lambda}(.)$ (resp. $\overline{E}(\lambda)$ and $\overline{F}_{\lambda}(.)$).

5 Inference about a single association rate τ_{ij}

We first investigate the properties of the inferences about a single association rate τ_{ab} from the IDM. The following lemma shows that inferences about τ_{ab} can be carried out from the analysis of a simple 2×2 table.

Lemma 1 *Consider the pooled table $A^* \times B^*$, with $A^* = \{a, a'\}$ and $B^* = \{b, b'\}$ and denote τ^*_{ab} the association rate of cell ab from the pooled table. From Property 2, $\tau^*_{ab} = \tau_{ab}$. Further, inferences from the IDM are invariant by such a pooling, since the IDM obeys the RIP. Thus, inferences about any single τ_{ab} only involve the relevant 2×2 table, $A^* \times B^*$.*

5.1 Prior upper and lower expectation and cdf

The prior lower and upper expectation of τ_{ab} are given by $\underline{E}(\tau_{ab}) = -1$ and $\overline{E}(\tau_{ab}) \to +\infty$, and are attained respectively by $\varphi_{ab} = \varphi_{a'b'} \to \frac{1}{2}$, and by $\varphi_{ab} = \lambda$, $\varphi_{a'b'} = 1 - \lambda$, with $\lambda \to 0$. The same limiting values of $\boldsymbol{\varphi}$ also lead to the prior upper and lower cdf's respectively, $\underline{P}(\tau_{ab} > t) = 0$ and $\overline{P}(\tau_{ab} > t) = 1$, for any $0 < t < 1$. These results show that prior inferences about τ_{ab} are vacuous. The prior IDM thus expresses a state of prior ignorance about parameter τ_{ab}.

5.2 Posterior upper and lower expectation and cdf

As in [4], we have recourse to Conjecture 1 in order to find approximate values for the posterior upper and lower expectations of τ_{ab}. Write $\tau_{ab} = g(\boldsymbol{\theta})$ where $g(.)$ is such that $t_{ab} = g(\boldsymbol{f})$ and $g(.)$ is given by Definition 3. Under a single Dirichlet prior, $Diri(\nu\boldsymbol{\varphi})$, the posterior expectation $E_{\nu\boldsymbol{\varphi}}(\tau_{ab}|\boldsymbol{x})$ is approximated by replacing each θ_k in $g(.)$ by $E(\theta_k|\boldsymbol{x})$ given in (1), that is

$$E^{\star}_{\nu\boldsymbol{\varphi}}(\tau_{ab}|\boldsymbol{x}) \quad = \quad \frac{x_{ab} + \nu\varphi_{ab}}{(x_a + \nu\varphi_a)(x_b + \nu\varphi_b)} - 1 \tag{5}$$

where x_a and x_b are the marginal counts of cell ab, and φ_a and φ_b its marginal prior frequencies. Conjecture 1 suggests then to minimize (resp. maximize) $E^{\star}_{\nu\boldsymbol{\varphi}}(\tau_{ab}|\boldsymbol{x})$ with respect to $\boldsymbol{\varphi}$, in order to estimate the posterior lower (resp. upper) expectations of τ_{ab} under the IDM(ν). The minimum value is attained by letting $\varphi_{ab'} \to 1$, $\varphi_{a'b} \to 1$, or $\varphi_{ab'} = \varphi_{a'b} \to 1/2$, whether $f_{ab'}$ is lower than, greater than, or equal to $f_{a'b}$ respectively. The maximum value is attained by letting $\varphi_{ab} \to 1$ or $\varphi_{a'b'} \to 1$ whether $x_a x_b > x_{ab}(x_a + x_b + \nu)$ or not. Following Conjecture 2, we use the same values for finding approximate posterior lower and upper cdf's of τ_{ab}.

5.3 Dyad data: Summary of local inferences

Table 4 gives the lower and upper probabilities of a positive association rate from the IDM with $\nu = 2$, for each cell *ab* concerned by the prediction given in Table 2(right). Three of the four diagonal cells, $(a1,b1)$, $(a2,b2)$ and $(a3,b3)$, can be assessed to be inductively over-represented with a high guarantee, $\underline{P}(\tau_{ab} > 0)$ being at least 0.99 for any of them. For cell $(a4,b4)$, the probability interval, $[0.00;1.00]$ is almost vacuous; uncertainty still dominates, even after observing 115 observations. For the regions off the diagonal, only cells $(a1,b3)$ and $(a3,b1)$ are guaranteed to be under-represented, since, in both cases, $\underline{P}(\tau_{ab} < 0) = 1 - \overline{P}(\tau_{ab} > 0) = 1.00$; cells $(a2,b1)$ and $(a3,b2)$ have a probability of at least 0.79 and 0.61 to be under-represented; uncertainty concerning the 8 remaining off-diagonal cells is even larger, since $\underline{P}(\tau_{ab} < 0) < 0.50$ for each cell.

The first overall conclusion that may be drawn from these results is that the model shown in Table 2(right) cannot not be inductively assessed at the cell level.

Of course, any other reference value for τ_{ab} than 0 can be used in a similar way. For instance, the probability intervals for event $\tau_{ab} > 0.50$ for diagonal cells are: $[0.30;0.50]$ for $(a1,b1)$, $[0.98;1.00]$ for $(a2,b2)$, $[0.99,1.00]$ for $(a3,b3)$ and $[0.00,0.99]$ for $(a4,b4)$. Both cells $(a2,b2)$ and $(a3,b3)$ can be assessed to be over-represented by at least 50% with a high lower probability.

Table 4: *Dyad data. Lower and upper posterior probabilities for event* $\tau_{ab} > 0$, $\underline{\mathrm{P}}(\tau_{ab} > 0|\boldsymbol{x})$ *and* $\overline{\mathrm{P}}(\tau_{ab} > 0|\boldsymbol{x})$*, for cells indexed by* $a1,\dots,a4$ *and* $b1,\dots,b4$ *only, using the IDM(*$\nu = 2$*).*

	*b*0	*b*1	*b*2	*b*3	*b*4
*a*0					
*a*1		1.00;1.00	0.09;0.65	0.00;0.00	0.45;0.95
*a*2		0.00;0.21	0.99;1.00	0.00;0.57	0.00;0.99
*a*3		0.00;0.00	0.00;0.39	1.00;1.00	0.53;0.97
*a*4		0.56;1.00	0.00;0.99	0.00;0.81	0.00;1.00

5.4 Comparison with frequentist and Bayesian approaches

Let us consider the test of the hypothesis $H_0 : \tau_{ab} \leq 0$ versus $H_1 : \tau_{ab} > 0$. Due to Corollary 1, this test is equivalent to $H_0 : \Phi \leq 0$ versus $H_1 : \Phi > 0$, where Φ is the usual contingency coefficient for a 2×2 table.

In the frequentist framework, the usual corresponding test is Fisher's exact test for a 2×2 table. The one-sided level p_{inc} of this test is usually computed as the probability of the *observations or more extreme cases* (*inclusive* test) under H_0. However, as argued by [2], this choice is a matter of convention and one could also envisage the *exclusive* alternative with level p_{exc} involving *more extreme cases* only. The following lemma shows that both these frequentist tests can actually be reinterpreted in a Bayesian way.

Lemma 2 *Let p_{exc} and p_{inc} by the exclusive and the inclusive levels (one-sided) of Fisher's exact test of $H_0 : \Phi \leq 0$ versus $H_1 : \Phi > 0$ for a 2×2 table with counts $\boldsymbol{x}$. Let $\mathrm{P}_{\nu\boldsymbol{\varphi}}(.)$ be a Bayesian probability obtained from the prior $\mathrm{Diri}(\nu\boldsymbol{\varphi})$ on $\boldsymbol{\theta}$. Then, $p_{exc} = \mathrm{P}_{\nu\boldsymbol{\varphi}}(H_1|\boldsymbol{x})$ with $\nu = 2$ and $\boldsymbol{\varphi} = (0, \frac{1}{2}, \frac{1}{2}, 0)$, and $p_{inc} = \mathrm{P}_{\nu\boldsymbol{\varphi}}(H_1|\boldsymbol{x})$ with $\nu = 2$ and $\boldsymbol{\varphi} = (\frac{1}{2}, 0, 0, \frac{1}{2})$. The former prior allocates non-null strengths evenly to cells (a, b') and (a', b), the latter to cells (a, b) and (a', b').*

Lemma 3 *Under the same assumptions, the probability $\mathrm{P}_{\nu\boldsymbol{\varphi}}(\tau_{ab} > 0|\boldsymbol{x})$ from any of the four symmetric ($\boldsymbol{\varphi}$ constant) objective Bayesian priors, i.e., $\nu \to 0$, $\nu = 1$, $\nu = 2$ and $\nu = 4$, are in the interval $[p_{exc}; p_{inc}]$.*

Proof. Lemmas 2 and 3 can be readily deduced from results in [1, Sec. 3]. □

Theorem 1 *For any cell (a, b), the posterior lower and probabilities of event $\tau_{ab} \leq 0$ from the IDM with $\nu = 2$ encompass (i) Fisher's exact probabilities for $H_0 : \tau_{ab} \leq 0$ versus $H_1 : \tau_{ab} > 0$ using either the exclusive or the inclusive convention and (ii) the Bayesian posterior probabilities of the same event under the objective priors of Haldane, Perks, Jeffreys and Bayes-Laplace (the latter two being defined on the relevant specific 2×2 table).*

Proof. The proof follows from (i) the equivalence between $\tau_{ab} > 0$ and $\Phi > 0$ for the pooled $\{a, a'\} \times \{b, b'\}$ table, (ii) the two Lemmas 2 and 3, and (iii) from the fact that the two Bayesian priors equivalent to p_{exc} and p_{inc} are such that $\nu = 2$ and thus belong to the IDM($\nu = 2$). □

Note 2 *In analyzing a 2×2 table, Walley et al. [13, Sec. 5.4] advocate the use of two independent IDM's with same prior strength ν_1, one for each line of the table. They note that the value $\nu_1 = 1$ leads to $\overline{\mathrm{P}}(H_0|\boldsymbol{x}) = p_{inc}$, a result which is only half of what Lemma 2 says. Here, we propose a more cautious model, a single IDM with $\nu = 2\nu_1 = 2$ for the whole table, which encompasses Walley's model. As Theorem 1 implies, our model has the advantage of producing inferences that encompass inferences from alternative objective models* for all cells of the table simultaneously. *The IDM($\nu = 2$) is the smallest IDM having this property.*

5.5 Absent or rare cells

For some cells, posterior uncertainty is still quite large. As an example, consider the unobserved cell $(a2, b4)$ for which the posterior probability interval for $\tau_{ab} > 0$ is almost vacuous, [0.00; 0.986] (see Table 4). Such a wide interval results from the rareness of both $a2$ and $b4$ ($f_{a2} = f_{b4} = 4/115$). Even if $a2$ and $b4$ were locally independent, the expected number of observations in cell $(a2, b4)$ would be extremely small, $\widehat{x_{a2b4}} = n\widehat{f_{a2b4}} = 16/115$, far less than one observation. Thus, despite the extreme descriptive result $t_{a2b4} = -1$, both the hypotheses $a2 \perp\!\!\!\perp b4$ ($\tau_{a2b4} = 0$) and $a2b4 \Longrightarrow \emptyset$ ($\tau_{a2b4} = -1$) are compatible with the data. A similar result was found by [4]. This uncertainty is also reflected in the large differences between the alternative objective models: $P(\tau_{a2b4} > 0)$ ranges from 0 (Haldane), 0.350 (Perks), 0.571 (Jeffreys), to 0.802 (Bayes-Laplace), and the corresponding probability from Fisher's exact tests are 0 (exclusive) and 0.866 (inclusive).

6 Inference about a mean association rate τ_R

Without loss of generality (see Property 1), we consider a non-empty region R which does not contain any full row or a full column of the $A \times B$ table. It is easy to find a Dirichlet prior within the IDM for which the prior lower expectation of τ_R is -1 ($\forall (a,b) \in R, \varphi_{ab} \to 0$ with strengths of cells outside R carefully chosen). This limiting value for $\boldsymbol{\varphi}$ also provides the prior upper cdf, $\overline{P}(\tau_R > t) = 1$ for $0 < t < 1$. We believe that the prior upper expectation and lower cdf of τ_R lead to vacuous inferences about τ_R, but we have no formal proof of that.

6.1 Posterior inferences about a single τ_R

Let $\tau_R = g(\boldsymbol{\theta})$, with $t_R = g(\boldsymbol{f})$ as given in Definition 4. We shall assume that allocating ν to a single cell suffices to attain the lower or upper expectation or cdf of τ_R. This assumption actually appears to be true in most cases we tested, but is certainly not true in all cases. However, we shall consider that it provides a reasonable approximation for inferences about τ_R from the IDM. As a second level of approximation, we use the same argument as in Section 5.2 using Conjectures 1 and 2. Define $E^{\star}_{\nu\boldsymbol{\varphi}}(\tau_R|\boldsymbol{x}) = g(E_{\nu\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{x}))$ and $E_{\nu\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{x})$ is given by (1).

Theorem 2 *Denote by r_{ab} the indicator variable of $(a,b) \in R$ and R' the complement of R in $A \times B$. Compute $m_{ab} = \sum_{i=1}^{A} r_{ib} x_i + \sum_{j=1\dots B} r_{aj} x_j$ for each cell (a,b). Then $\mathrm{E}^{\star}_{\nu\boldsymbol{\varphi}}(\tau_R|\boldsymbol{x})$ is minimized by letting $\varphi_{ab} \to 1$ for cell $(a,b) \in R'$ maximizing m_{ab}. (We have no simple formula for maximization of $\mathrm{E}^{\star}_{\nu\boldsymbol{\varphi}}(\tau_R|\boldsymbol{x})$.) Proof involves tedious but rather simple algebra.*

6.2 Stages data: Inference on $\delta_{\mathcal{M}}$

Consider the Stages data (Table 1) and the model $\mathcal{M}$ defined therein. We found $d_{\mathcal{M}} = 0.851$ and we now want to make inferences about parameter $\delta_{\mathcal{M}}$ using the IDM($\nu = 2$). For various statements about $\delta_{\mathcal{M}}$, we find the following probability intervals: $[1.00; 1.00]$ for $\delta_{\mathcal{M}} > 0$, $[0.95; 1.00]$ for $\delta_{\mathcal{M}} > 0.50$, $[0.84; 0.98]$ for $\delta_{\mathcal{M}} > 0.60$ and $[0.62; 0.93]$ for $\delta_{\mathcal{M}} > 0.70$. We thus may assess that the data quasi-agree with $\mathcal{M}$ at threshold $d_{quasi} = 0.50$, with probability at least 0.95.

6.3 Inferences about a complex directional association model

The IDM can be applied to study any kind of complex model expressed as a conjunction of constraints about association rates of specific cells or regions of an $A \times B$ table. Consider the Dyad data in Table 2(left) and the two models $\mathcal{M}_1 = \tau_L < \tau_U < 0 < \tau_D$ and $\mathcal{M}_2 = \tau_L < -0.70 < \tau_U < 0 < 0.50 < \tau_D$. Both try to express the expected pattern shown in Table 2(right), in a more or less strong way. Computing the posterior lower and upper probabilities of $\mathcal{M}_1$ or $\mathcal{M}_2$ can be done numerically by minimization/maximization over the set of Dirichlet posteriors. Using the IDM with $\nu = 2$, we find $\underline{P}(\mathcal{M}_1) = 0.98$, $\overline{P}(\mathcal{M}_1) = 1.00$, and $\underline{P}(\mathcal{M}_2) = 0.84$, $\overline{P}(\mathcal{M}_2) = 0.96$. Model $\mathcal{M}_1$ only is supported by the data with a sufficiently high lower probability.

Of course, models $\mathcal{M}_1$ and $\mathcal{M}_2$ are only two candidates amongst the possible inductive summaries of the data. The task of model selection (which is not addressed here) would require taking into account, not only the (lower) probability of each model, but also the degree of specificity or generality of each model.

7 Concluding remarks

This paper proposes a method for analyzing local or asymmetric dependencies in a contingency table, by focusing on previously suggested indices — (mean) association rates [5, 10] and *Del* index [7] —, which, we believe, are simple and natural, and yet provide means to define a wide variety of association models.

We showed how the imprecise Dirichlet model (IDM) can be applied to assess whether the data support such association models or not. Several results provide approximate solutions to the minimizing/maximizing problems required by the IDM. Further research would be needed to develop exact solutions or to measure the accuracy of our approximate procedures.

The exact comparison between the IDM and alternative frequentist or objective Bayesian models, carried out in Section 5.4 (see especially Theorem 1), provides a new argument for choosing $\nu = 2$ in the IDM, for a problem involving a possibly large number of categories (see also [4]). The large discrepancies which can be obtained in the inferences from these various alternative models are translated as a high imprecision in the IDM (see an example in Section 5.5). Section 5.4 shows that this phenomenon occurs whenever the frequentist *probability of the observed data* (under some particular null hypothesis) is not negligible.

References

[1] ALTHAM, P. M. E. Exact Bayesian analysis of a 2×2 contingency table and Fisher’s exact significance test. *J. Roy. Statist. Soc. Ser. B 31*, 2 (1968), 261–269.

[2] BERNARD, J.-M. Bayesian interpretation of frequentist procedures for a Bernoulli process. *The American Statistician 50*, 1 (1996), 7–13.

[3] BERNARD, J.-M. Non-parametric inference about an unknown mean using the imprecise Dirichlet model. In *Proceedings of the 2nd International Symposium on Imprecise Probabilities and their Applications (ISIPTA'01)* (Maastricht, 2001), G. de Cooman, T. Fine, and T. Seidenfeld, Eds., Shaker Publishing BV, pp. 40–50.

[4] BERNARD, J.-M. Implicative analysis for multivariate binary data using an imprecise Dirichlet model. *J. Statist. Plann. Inference 105* (2002), 83–103.

[5] DANIS, A., BERNARD, J.-M., AND LEPROUX, C. Shared picture-book reading: A sequential analysis of adult-child verbal interactions. *British Journal of Developmental Psychology 18* (2000), 369–388.

[6] GOODMAN, L. A., AND KRUSKAL, W. H. Measures of association for cross classifications. II: Further discussion and references. *J. Amer. Statist. Assoc. 54* (1959), 123–163.

[7] HILDEBRAND, D. K., LAING, J. D., AND ROSENTHAL, H. *Prediction Analysis of Cross Classifications*. John Wiley & sons, 1977.

[8] JAMISON, W. Developmental inter-relationships among concrete operational tasks: An investigation of Piaget's stage concept. *Journal of Experimental Child Psychology 24* (1977), 235–253.

[9] KENDALL, M. G., AND STUART, A. *The Advanced Theory of Statistics, Vol. 2: Inference and Relationship*, 3rd ed. Griffin, 1973.

[10] ROUANET, H., BERT, M.-P., AND LE ROUX, B. *Statistique en Sciences Humaines: Procédures Naturelles*. Dunod, Paris, 1987.

[11] WALLEY, P. Statistical reasoning with imprecise probabilities. In *Monographs on Statistics and Applied Probability*, vol. 42. Chapman & Hall, London, 1991.

[12] WALLEY, P. Inferences from multinomial data: learning about a bag of marbles. *J. Roy. Statist. Soc. Ser. B 58* (1996), 3–57.

[13] WALLEY, P., GURRIN, L., AND BURTON, P. Analysis of clinical data using imprecise prior probabilities. *The Statistician 45* (1996), 457–485.

Jean-Marc Bernard is with the Laboratoire de Psychologie Environnementale, Université de Paris 5 & CNRS UMR 8069, 71 avenue Edouard Vaillant, 92774 Boulogne-Billancourt Cedex, France. E-mail: jmbernard@psycho.univ-paris5.fr

Some Results on Generalized Coherence of Conditional Probability Bounds

V. BIAZZO
Università di Catania, Italy

A. GILIO
Università "La Sapienza," di Roma, Italy

G. SANFILIPPO
Università di Catania, Italy

Abstract

Based on the coherence principle of de Finetti and a related notion of generalized coherence (g-coherence), we adopt a probabilistic approach to uncertainty based on conditional probability bounds. Our notion of g-coherence is equivalent to the ”avoiding uniform loss” property for lower and upper probabilities (a la Walley). Moreover, given a g-coherent imprecise assessment by our algorithms we can correct it obtaining the associated coherent assessment (in the sense of Walley and Williams). As is well known, the problems of checking g-coherence and propagating tight g-coherent intervals are $NP-$ and $FP^{NP}-$complete, respectively, and thus $NP-$hard. Two notions which may be helpful to reduce computational effort are those of non relevant gain and basic set. Exploiting them, our algorithms can use linear systems with reduced sets of variables and/or linear constraints. In this paper we give some insights on the notions of non relevant gain and basic set. We consider several families with three conditional events, obtaining some results characterizing g-coherence in such cases. We also give some more general results.

Keywords

uncertain knowledge, coherence, g-coherence, imprecise probabilities, conditional probability bounds, lower and upper probabilities, non relevant gains, basic sets

1 Introduction

Among the many symbolic or numerical approaches to the management of uncertain knowledge, the probabilistic treatment of uncertainty by means of precise or imprecise assessments is a well known formalism often applied in real situations.

A general framework which allows a consistent management of probabilistic assessments is obtained by resorting to de Finetti's coherence principle ([2], [7], [8], [11]), or suitable generalizations of it given for upper and lower probabilities ([20], [19]). In our approach we adopt the notion of g-coherence (i.e. generalized coherence) introduced in [1] (see also [10]), which is weaker than the notion of coherence given in [19]. Actually, the notion of g-coherence is equivalent to the property of ”avoiding uniform loss” given in [19]. Within our framework, a given g-coherent assessment can be corrected, obtaining the associated coherent one, and possibly extended to further conditional events. As is well known, if we discard the case of conditioning events with zero probability the probabilistic reasoning can be reduced to a linear optimization problem (we also point out that g-coherent probabilistic reasoning generally does not coincide with probabilistic reasoning as in, e.g., [12], [14], when the conditioning event has a non-zero probability). When conditioning events may have zero lower/upper probability, the methods presented in the literature (our one too) usually exploit sequences of linear programs. Among them, a ”dual” approach for the extension of lower and upper previsions, explicitly based on random gains, has been developed in [20]. With the aim of improving the method given in [20], an interesting technique for computing lower conditional expectations through sequences of pivoting operations has been proposed in [9]. Roughly speaking, probabilistic reasoning can be developed by local approaches, based on the iteration of suitable inference rules, and global ones (the issue of local versus global approaches has been examined especially in [17], [18]). We recall that probabilistic reasoning based on a global approach tends to become intractable. Hence, it is worthwhile to examine any method which try to eliminate or reduce computational difficulties, possibly finding efficient special-case algorithms. This problem has been faced by many authors (see, e.g., [5], [7], [8], [9], [12], [14], [20]). Many aspects concerning the complexity of probabilistic reasoning under coherence have been studied in [3]. The relationship between coherence-based and model-theoretic probabilistic reasoning has been widely explored in [4]. In [16] an efficient procedure has been proposed for families of *conjunctive* conditional events. Such procedure can be characterized in the framework of coherence introducing suitable notions of non relevant gains and basic sets ([2]). Exploiting such notions, our algorithms for g-coherence checking and propagation of conditional probability bounds can use linear systems with reduced sets of variables and/or constraints. In this paper we illustrate the notions of non relevant gain and basic set, by examining several examples of families constituted by three conditional events. We obtain some theoretical results which characterize g-coherence in such particular cases. In this way, the characterization of g-coherence in the case of larger families of conditional events should be facilitated. We obtain some necessary and sufficient conditions for the g-coherence of lower probability bounds. We also give some more general results. Notice that the case of families with three conditional events may have a specific importance, e.g., in the field of default reasoning where many inference rules consist of two premises and one consequence. We also recall that coherence-based probabilistic reasoning can be reduced to standard reasoning tasks in model-theoretic probabilistic logic, using concepts from default reasoning ([4]). The rest of the paper is organized as follows. In Section 2 we recall some preliminary concepts. In Section 3 we illustrate the notions of non relevant gain and basic set and we recall some theoretical results. In Section 4 we consider several cases of families constituted by three conditional events and we give some necessary and sufficient conditions of g-coherence. In Section 5 we give some more general results. Finally, in Section 6 we give some conclusions and an outlook on further developments.

2 Some preliminary concepts

For each integer n, we set $J_n = \{1,\dots,n\}$. Given any event E, we denote by the same symbol its indicator and by E^c its negation. Given a further event H, we denote by EH (resp. $E \vee H$) the conjunction (resp. disjunction) of E and H. Let P be a conditional probability assessment defined on a family of conditional events $\mathcal{K}$. Given a finite subfamily $\mathcal{F}_n = \{E_1|H_1,\dots,E_n|H_n\} \subseteq \mathcal{K}$, let $\mathcal{P}_n$ be the vector $(p_1,\dots,p_n)$, where $p_i = P(E_i|H_i), i \in J_n$. With the pair $(\mathcal{F}_n,\mathcal{P}_n)$ we associate the random quantity $G_n = \sum_{i\in J_n} s_i H_i(E_i - p_i)$, with $s_1,\dots,s_n$ arbitrary real numbers. Moreover, we denote by $G_n|\mathcal{H}_n$ the restriction of G_n to $\mathcal{H}_n = H_1 \vee \cdots \vee H_n$. Then, based on the *betting scheme*, we have

Definition 1 The probability assessment P on $\mathcal{K}$ is said coherent if, for every integer $n = 1,2,\dots$, for every subfamily $\mathcal{F}_n \subseteq \mathcal{K}$ and for every real numbers $s_1,\dots,s_n$, the condition $Max\ G_n|\mathcal{H}_n \geq 0$ is satisfied.

We denote by $\mathcal{A}_n$ a vector $(\alpha_1,\dots,\alpha_n)$ of lower probability bounds on $\mathcal{F}_n$. We say that the pair $(\mathcal{F}_n,\mathcal{A}_n)$ is associated with the set J_n.

Definition 2 The vector of lower bounds $\mathcal{A}_n$ is g-coherent iff there exists a coherent probability assessment $\mathcal{P}_n = (p_1,\dots,p_n)$ on $\mathcal{F}_n$ such that $p_i \geq \alpha_i, \forall\, i \in J_n$.

By expanding the expression $\bigwedge_{i\in J_n}(E_iH_i \vee E_i^cH_i \vee H_i^c)$, we obtain the constituents associated with $\mathcal{F}_n$. We denote by $C_1,\dots,C_m$, where $m \leq 3^n - 1$, the constituents contained in $\mathcal{H}_n = \bigvee_{j\in J_n} H_j$. A further constituent (if it is not impossible) is $C_0 = \mathcal{H}_n^c = H_1^c \cdots H_n^c$.
Remark: With the family $\mathcal{F}_n$ we associate a set L which describe the logical relationships among the events $E_i, H_i, i \in J_n$. Then, the set of constituents is the set of those conjunctions $\chi_1\cdots\chi_n$, with $\chi_i \in \{E_iH_i, E_i^cH_i, H_i^c\}, \forall\, i \in J_n$, which satisfy the set of logical relations L. Notice that, if $L = \emptyset$, then $m = 3^n - 1$ and $C_0 \neq \emptyset$, i.e. the number of constituents is 3^n.

For each constituent $C_r, r \in J_m$, we introduce a vector $V_r = (v_{r1},\dots,v_{rn})$, where for each $i \in J_n$ it is respectively $v_{ri} = 1$, or $v_{ri} = 0$, or $v_{ri} = \alpha_i$, according to whether $C_r \subseteq E_iH_i$, or $C_r \subseteq E_i^cH_i$, or $C_r \subseteq H_i^c$. With the pair $(\mathcal{F}_n, \mathcal{A}_n)$ we associate the random gain $G_n = \sum_{i \in J_n} s_i H_i (E_i - \alpha_i)$, where $s_i \geq 0$, $\forall\, i \in J_n$. Moreover, we denote by

$$g_h = G_n(V_h) = \sum_{i \in J_n} s_i (v_{hi} - \alpha_i) = \sum_{i : C_h \subseteq H_i} s_i (v_{hi} - \alpha_i) \tag{1}$$

the value of $G_n | \mathcal{H}_n$ associated with C_h. We denote by $(\mathcal{S}_n)$ the following system in the unknowns λ_r's.

$$\sum_{r \in J_m} \lambda_r v_{ri} \geq \alpha_i,\ \ i \in J_n; \qquad \sum_{r \in J_m} \lambda_r = 1; \qquad \lambda_r \geq 0,\ \ \forall\, r \in J_m. \tag{2}$$

Remark: The solvability of $(\mathcal{S}_n)$ means that there exists a non negative vector $(\lambda_r\,;\ r \in J_m)$, with $\sum_{r \in J_m} \lambda_r = 1$, such that $\sum_{r \in J_m} \lambda_r V_r \geq \mathcal{A}_n$. In other words, in the convex hull of the points V_r's there exists a point $V^* = \sum_{r \in J_m} \lambda_r V_r$ such that $V^* \geq \mathcal{A}_n$ (this geometrical approach will be used in the proof of Theorem 4).

As shown in [10], a set of lower bounds $\mathcal{A}$ defined on $\mathcal{K}$ is g-coherent iff, for every n and for every $\mathcal{F}_n \subseteq \mathcal{K}$, the system (2) is solvable. Moreover, based on a suitable alternative theorem, it can be shown ([2]) that the solvability of system (2) is equivalent to the following condition

$$Max\ G_n | \mathcal{H}_n \geq 0. \tag{3}$$

Then, we have

Proposition 1 A set of lower bounds $\mathcal{A}$ defined on a family of conditional events $\mathcal{K}$ is g-coherent iff $\forall\, n, \forall\, \mathcal{F}_n \subseteq \mathcal{K}$, and $\forall\, s_i \geq 0\,, i \in J_n$, it is $Max\ G_n | \mathcal{H}_n \geq 0$.

We remark that, if the case of zero probability for conditioning events is discarded, then to check g-coherence of the assessment $\mathcal{A}_n$ on $\mathcal{F}_n$ it is enough to check solvability of system (2). However, in our coherence-based approach, some (or possibly all) conditioning events may have zero probability. Then, to check g-coherence we should study the solvability of a very large number of systems, like (2). Actually, we can exploit algorithms which only check (the solvability of) a small number of linear systems (see, e.g., [1], [2], [5]).

3 Non relevant gains and basic sets

In this section we illustrate the notions of non relevant gain and basic set. Exploiting such notions, the algorithms for g-coherence checking and propagation of conditional probability bounds can use linear systems with reduced sets of variables and/or constraints. We recall some theoretical conditions given in [2].

Definition 3 Let $\mathcal{G} = \{g_j\}_{j \in J_m}$ be the set of possible values of the random gain $G_n | \mathcal{H}_n$. Then, a value $g_r \in \mathcal{G}$ is said *"not relevant for the checking of condition (3)"*, or in short *"not relevant"*, if there exists a set $T_r \subseteq J_m \setminus \{r\}$ such that:

$$Max\ \{g_j\}_{j \in T_r}\ < 0 \Longrightarrow g_r\ < 0\,. \tag{4}$$

Remark: Notice that, in the previous definition, it wouldn’t be equivalent to use the condition $T_r = J_m \setminus \{r\}$ instead of $T_r \subseteq J_m \setminus \{r\}$. In fact, it may happen that (4) holds with $T_r \subset J_m \setminus \{r\}$, so that g_r is not relevant, while at the same time it may be $Max\,\{g_j\}_{j\in J_m\setminus\{r\}} > 0$.

Definition 4 A set $\mathcal{G}_\Gamma = \{g_r\}_{r\in\Gamma}$, with $\Gamma \subset J_m$, is said *not relevant* if, $\forall r \in \Gamma$, there exists a set $T_r \subseteq J_m \setminus \Gamma$ such that (4) is satisfied.

Definition 5 A set $\mathcal{T} \subset J_m$ is said *basic* if the following property holds:
Basic Property. For every $r \in J_m \setminus \mathcal{T}$ there exists a set $T_r \subseteq \mathcal{T}$ such that the condition (4) is satisfied.
A basic set $\mathcal{T}$ is said *minimal* if, for every $T \subset \mathcal{T}$, the set T is not basic.

We observe that $Max\ G_n|\mathcal{H}_n \ = \ Max\ \{g_j\}_{j\in J_m}$. Then, we have

Theorem 1 Let $\mathcal{T} \subset J_m$ be a *basic* set. Then

$$Max\,\{g_j\}_{j\in J_m}\ \geq\ 0 \quad\Longleftrightarrow\quad Max\,\{g_j\}_{j\in\mathcal{T}}\ \geq\ 0\,. \tag{5}$$

Remark: We point out that, given a subset $\mathcal{T}$, if there exists $r \notin \mathcal{T}$ such that, for every $T_r \subseteq \mathcal{T}$, the condition (3) is not satisfied, then $\mathcal{T}$ is not a basic set. Moreover, we observe that the condition (5) is trivially satisfied for $\mathcal{T} = J_m$. Then, as for $\mathcal{T} = J_m$ the set $J_m \setminus \mathcal{T}$ is empty, we can enlarge the class of basic sets by including in it J_m too.

Given $r \in J_m$ and a set $\mathcal{T}_r \subseteq J_m \setminus \{r\}$, let us consider the following condition

$$g_r \leq \sum_{j\in\mathcal{T}_r} a_j g_j\ ; \quad a_j > 0\ ,\ \forall j \in \mathcal{T}_r\,. \tag{6}$$

By Definition 3 one has that, if the above condition is satisfied, then g_r is not relevant. The condition (6) can be exploited in general to reduce the number of variables. The basic idea is illustrated by the following theorem ([2], [5]).

Theorem 2 Let $\mathcal{T}$ be a strict subset of the set J_m such that for every $r \notin \mathcal{T}$ there exists $T_r \subseteq \mathcal{T}$ satisfying the condition (6). Then:

$$Max\,\{g_j\}_{j\in J_m} \geq 0 \Longleftrightarrow Max\,\{g_j\}_{j\in\mathcal{T}} \geq 0\,. \tag{7}$$

Based on the previous result and on suitable alternative theorems, in order to check g-coherence we can replace $(\mathcal{S}_n)$ by an equivalent system $(\mathcal{S}_n^{\mathcal{T}})$, which has a reduced vector of unknowns $\Lambda_T = (\lambda_r; r \in \mathcal{T})$. We denote by $S_{\mathcal{T}}$ the set of solutions of $(\mathcal{S}_n^{\mathcal{T}})$. Moreover, for each $j \in J_n$, we consider the function $\Phi_j^{\mathcal{T}}(\Lambda_T) = \sum_{r\in\mathcal{T}:C_r\subseteq H_j}\lambda_r$. We denote by $I_0^{\mathcal{T}}$ the (strict) subset of J_n defined as

$$I_0^{\mathcal{T}} = \{j \in J_n\ :\ M_j = Max_{\Lambda_{\mathcal{T}}\in S_{\mathcal{T}}}\Phi_j^{\mathcal{T}}(\Lambda_{\mathcal{T}}) = 0\} \tag{8}$$

and by $(\mathcal{F}_0^{\mathcal{T}}, \mathcal{A}_0^{\mathcal{T}})$ the pair associated with $I_0^{\mathcal{T}}$. Then, to check g-coherence of $\mathcal{A}_n$, we can exploit the following result ([2]).

Theorem 3 The imprecise assessment $\mathcal{A}_n$ on $\mathcal{F}_n$ is g-coherent if and only if: 1) the system $(\mathcal{S}_n^{\mathcal{I}})$ is solvable; 2) if $I_0^{\mathcal{I}} \neq \emptyset$, then $\mathcal{A}_0^{\mathcal{I}}$ is g-coherent.

Note that, if $|I_0^{\mathcal{I}}| = 1$, say $I_0^{\mathcal{I}} = \{h\}$, then $\mathcal{A}_0^{\mathcal{I}} = (\alpha_h)$ and the g-coherence of $\mathcal{A}_0^{\mathcal{I}}$ simply amounts to the condition: $\alpha_h \leq 1$.

4 Some results on g-coherence of lower probability bounds for families of three conditional events

In this section we will illustrate the notions of non relevant gain and basic set by examining several examples which concern particular families of three conditional events.

Remark: We recall that such kind of families may be relevant in the field of default reasoning, where many inference rules are associated with two premises and one conclusion. As an example, with the following basic inference rules of System *P* ([15])

$$A \mid\!\sim B,\ A \mid\!\sim C \implies A \mid\!\sim BC\,, \qquad (And),$$

$$A \mid\!\sim C,\ A \mid\!\sim B \implies AB \mid\!\sim C\,, \qquad (Cautious\, Monotonicity),$$

$$A \mid\!\sim C,\ B \mid\!\sim C \implies A \vee B \mid\!\sim C\,, \qquad (Or),$$

are associated, respectively, the following families of conditional events

$$\{B|A,\ C|A, BC|A\}\,; \quad \{C|A, B|A, C|AB\}\,; \quad \{C|A, C|B, C|(A \vee B)\}\,.$$

We also note that the theoretical results obtained in the case $n = 3$ may be useful in establishing more general results when $n > 3$.

In what follows, to avoid the analysis of trivial or particular cases, we assume

$$\emptyset \subset E_i H_i \subset H_i\,, \quad 0 < \alpha_i < 1\,, \quad \forall\, i\,.$$

Then, for each $r \in J_m$, as $\alpha_i < 1$, if $v_{ri} = 1$ for some i, it follows $C_r \subseteq E_i H_i$.
Let $\mathcal{A}_3 = (\alpha_1, \alpha_2, \alpha_3)$ be a vector of lower bounds on $\mathcal{F}_3 = \{E_1|H_1, E_2|H_2, E_3|H_3\}$.
Given the set $\mathcal{V} = \{V_1, \ldots, V_m\}$, we define

$$\mathcal{W} = \{V_r \in \mathcal{V} : v_{ri} \neq 0, \forall\, i \in J_n\} \tag{9}$$

and, for each $V_r \in \mathcal{W}$,

$$N_r = \{i \in J_n : C_r \subseteq H_i^c\}\;. \tag{10}$$

Of course, $N_r \subset J_n$. Then, we define

$$\mathcal{V}_h = \{V_r \in \mathcal{W} : |N_r| = h\},\ h = 0, 1, \ldots, n-1\;. \tag{11}$$

With each $V_r \in \mathcal{V}$, $r \in J_m$, we associate the set N_r defined in (10) and the set

$$M_r = \{i \in J_n : v_{ri} = 0\} \,. \tag{12}$$

Then, introducing the set $I = \{(h,k) : h = 0,\ldots,n-1;\, k = 1,\ldots,n\}$, we define the sets

$$\mathcal{U}_{h,k} = \{V_r \in \mathcal{V} : |N_r| = h\,,\ |M_r| = k\}, \quad (h,k) \in I\,. \tag{13}$$

We observe that, if the sets $\mathcal{U}_{h,0}$ were defined, then recalling (11) we would have $\mathcal{V}_h = \mathcal{U}_{h,0}$. Then, recalling (9), we have

$$\mathcal{V} = \mathcal{W} \cup \Big(\bigcup_{(h,k)\in I} \mathcal{U}_{h,k}\Big) = \Big(\bigcup_{h=0}^{n-1} \mathcal{V}_h\Big) \cup \Big(\bigcup_{h,k} \mathcal{U}_{h,k}\Big)\,. \tag{14}$$

As $n = 3$, the set of vectors $\mathcal{V} = \{V_1,\ldots,V_m\}$, where $m \leq 26$, is a subset of the set

$$\{(1,1,1),(1,1,\alpha_3),(1,\alpha_2,1),(\alpha_1,1,1),\ldots,(\alpha_1,0,0),(0,\alpha_2,0),(0,0,\alpha_3),(0,0,0)\}\,.$$

By (14), we have

$$\mathcal{V} = \mathcal{V}_0 \cup \mathcal{V}_1 \cup \mathcal{V}_2 \cup \mathcal{U}_{0,1} \cup \mathcal{U}_{1,1} \cup \mathcal{U}_{0,2} \cup \mathcal{U}_{2,1} \cup \mathcal{U}_{1,2} \cup \mathcal{U}_{0,3}\,, \tag{15}$$

where

$\mathcal{V}_0 \subseteq \{(1,1,1)\}\,, \quad \mathcal{V}_1 \subseteq \{(1,1,\alpha_3),(1,\alpha_2,1),(\alpha_1,1,1)\}\,,$

$\mathcal{V}_2 \subseteq \{(1,\alpha_2,\alpha_3),(\alpha_1,1,\alpha_3),(\alpha_1,\alpha_2,1)\}\,, \quad \mathcal{U}_{0,1} \subseteq \{(1,1,0),(1,0,1),(0,1,1)\}\,,$

$\mathcal{U}_{1,1} \subseteq \{(1,\alpha_2,0),(1,0,\alpha_3),(\alpha_1,1,0),(0,1,\alpha_3),(\alpha_1,0,1),(0,\alpha_2,1),\}\,,$

$\mathcal{U}_{0,2} \subseteq \{(1,0,0),(0,1,0),(0,0,1)\}\,, \quad \mathcal{U}_{2,1} \subseteq \{(\alpha_1,\alpha_2,0),(\alpha_1,0,\alpha_3),(0,\alpha_2,\alpha_3)\}\,,$

$\mathcal{U}_{1,2} \subseteq \{(\alpha_1,0,0),(0,\alpha_2,0),(0,0,\alpha_3)\}\,, \quad \mathcal{U}_{0,3} \subseteq \{(0,0,0)\}\,.$

Remark: Notice that each given set of logical relationships L among the events $E_i, H_i, i = 1,2,3$, determines a particular representation (15) for the set of vectors $\mathcal{V}$. Then, in what follows, instead of assigning the set L, we directly assume some hypotheses on the subsets $\mathcal{V}_h$'s and $\mathcal{U}_{h,k}$'s. We list below some sufficient conditions, proved in [6], for g-coherence of the vector of lower bounds $\mathcal{A}_3$ on $\mathcal{F}_3$.
1. $|\mathcal{V}_0| = 1$; 2. $\mathcal{V}_0 = \emptyset,\ |\mathcal{V}_1| \geq 1$; 3. $\mathcal{V}_0 = \mathcal{V}_1 = \emptyset, |\mathcal{V}_2| \geq 2$;
4. $\mathcal{V}_0 = \mathcal{V}_1 = \emptyset,\ \mathcal{V}_2 = \{(1,\alpha_2,\alpha_3)\}\,,\ E_2H_2E_3H_3 \vee E_2H_2H_3^c \vee H_2^cE_3H_3 \neq \emptyset\,.$
Some further conditions obtained in [6] are given below.
5. If $\mathcal{V}_0 = \mathcal{V}_1 = \emptyset,\ \mathcal{V}_2 = \{(1,\alpha_2,\alpha_3)\}\,,\ E_2H_2E_3H_3 = E_2H_2H_3^c = H_2^cE_3H_3 = \emptyset\,,$ then $\mathcal{A}_3$ is g-coherent iff $\alpha_2 + \alpha_3 \leq 1\,.$
6. $\mathcal{V}_0 = \mathcal{V}_1 = \mathcal{V}_2 = \emptyset,\ \ \alpha_1 + \alpha_2 + \alpha_3 > 2 \implies \mathcal{A}_3$ not g-coherent.

7. If $\mathcal{V}_0 = \mathcal{V}_1 = \mathcal{V}_2 = \emptyset$, $|\mathcal{U}_{0,1}| = 3$, $\alpha_i < 1, \forall\, i$, then: a) there exists a basic set $\mathcal{T}$, with $|\mathcal{T}| = 3$; b) $\mathcal{A}_3$ is g-coherent iff $\alpha_1 + \alpha_2 + \alpha_3 \leq 2$.
8. If $\mathcal{V}_0 = \mathcal{V}_1 = \mathcal{V}_2 = \mathcal{U}_{0,1} = \mathcal{U}_{1,1} = \emptyset$, $\mathcal{U}_{0,2} = \{(1,0,0),(0,1,0),(0,0,1))\}$, $\alpha_i < 1, \forall\, i$, then:
a) if $\alpha_1 + \alpha_2 \leq 1, \alpha_1 + \alpha_3 \leq 1, \alpha_2 + \alpha_3 \leq 1$, then $\mathcal{T} = \{1,2,3\}$ is a basic set;
b) $\mathcal{A}_3$ is g-coherent iff $\alpha_1 + \alpha_2 + \alpha_3 \leq 1$.

Now we give further results concerning the case $n = 3$. Besides providing a better understanding of the notions of basic set and non relevant gain, these results permit in particular the deepening of the condition (6). In next theorem the hypotheses concerning the set of logical relations L specify that the conjunctions

$$\begin{array}{llll} E_1H_1E_2H_2E_3H_3\,, & E_1H_1E_2H_2H_3^c\,, & E_1H_1H_2^cE_3H_3\,, & H_1^cE_2H_2E_3H_3\,, \\ E_1H_1H_2^cH_3^c\,, & H_1^cE_2H_2H_3^c\,, & H_1^cH_2^cE_3H_3\,, & E_1^cH_1E_2H_2E_3H_3 \end{array}$$

are impossible, while the conjunctions

$$E_1H_1E_2H_2E_3^cH_3\,,\;\; E_1H_1E_2^cH_2E_3H_3\,,\;\; E_1^cH_1E_2H_2H_3^c\,,\;\; E_1^cH_1H_2^cE_3H_3$$

are possible. Then, concerning the number m of unknowns in the system $(\mathcal{S}_3)$, one has: $4 \leq m \leq 18$. Actually, we will use a system $(\mathcal{S}_3^{\mathcal{T}})$ with only 3 or 4 unknowns.

Theorem 4 If $\mathcal{V}_0 = \mathcal{V}_1 = \mathcal{V}_2 = \emptyset$, $\mathcal{U}_{0,1} = \{V_1,V_2\} = \{(1,1,0),(1,0,1)\}, \{V_3,V_4\} = \{(0,1,\alpha_3),(0,\alpha_2,1)\} \subseteq \mathcal{U}_{1,1}$, $0 < \alpha_i < 1, \forall\, i$, then one has:
a) for every $r > 4$, the gain g_r is not relevant;
b) if $\alpha_1 + \alpha_2 \leq 1$, or $\alpha_2 + \alpha_3 \leq 1$, or $\alpha_1 + \alpha_3 \leq 1$, then there exists a basic set $\mathcal{T}$, with $|\mathcal{T}| \leq 3$, and $\mathcal{A}_3$ is g-coherent;
c) if $\alpha_1 + \alpha_2 > 1, \alpha_2 + \alpha_3 > 1, \alpha_1 + \alpha_3 > 1$, then $\mathcal{A}_3$ is g-coherent iff

$$\alpha_1\alpha_3 + \alpha_2 \leq 1\,, \quad \text{or} \quad \alpha_1\alpha_2 + \alpha_3 \leq 1.$$

Proof. a) by the hypotheses, it follows that for each $V_r \in \mathcal{V}$, with $r > 4$, there exists $h \in \{1,2,3,4\}$ such that $V_r \leq V_h$; hence g_r is not relevant. Then, $\mathcal{T} = \{1,2,3,4\}$ is a basic set.
In order to study the g-coherence of $\mathcal{A}_3$, we first determine the gains associated with the vectors V_1,V_2,V_3,V_4. Recalling (1), these gains are respectively

$$\begin{array}{ll} g_1 = s_1(1-\alpha_1) + s_2(1-\alpha_2) - s_3\alpha_3\,, & g_2 = s_1(1-\alpha_1) - s_2\alpha_2 + s_3(1-\alpha_3)\,, \\ g_3 = -s_1\alpha_1 + s_2(1-\alpha_2)\,, & g_4 = -s_1\alpha_1 + s_3(1-\alpha_3)\,. \end{array}$$

We also need the equations of the planes $\pi_1, \pi_2, \pi_3, \pi_4$, containing respectively the triangles $V_1V_2V_3$, $V_1V_2V_4$, $V_1V_3V_4$, $V_2V_3V_4$, which are given below

$$\begin{array}{ll} \pi_1 :\; \alpha_3 x + y + z = 1 + \alpha_3\,; & \pi_3 :\; \alpha_3(1-\alpha_2)x + (1-\alpha_3)y + (1-\alpha_2)z = 1 - \alpha_2\alpha_3\,; \\ \pi_2 :\; \alpha_2 x + y + z = 1 + \alpha_2\,; & \pi_4 :\; \alpha_2(1-\alpha_3)x + (1-\alpha_3)y + (1-\alpha_2)z = 1 - \alpha_2\alpha_3\,. \end{array}$$

The intersection points of the segment (x,α_2,α_3), $0 \le x \le 1$, with the planes π_1 and π_2, are respectively $V_x^* = (\frac{1-\alpha_2}{\alpha_3},\alpha_2,\alpha_3)$ and $V_x^{**} = (\frac{1-\alpha_3}{\alpha_2},\alpha_2,\alpha_3)$. Moreover,

$$V_x^* \ge \mathcal{A}_3 \iff \alpha_1\alpha_3+\alpha_2 \le 1\,; \qquad V_x^{**} \ge \mathcal{A}_3 \iff \alpha_1\alpha_2+\alpha_3 \le 1\,.$$

The intersection point of the segment (α_1,y,α_3), $0 \le y \le 1$, with the plane π_3 is

$$V_y^* = (\alpha_1, \tfrac{1-\alpha_3-\alpha_1\alpha_3(1-\alpha_2)}{1-\alpha_3}, \alpha_3) \ge \mathcal{A}_3\,, \qquad \forall\, \alpha_2 \in [0,1]\,.$$

The intersection point of the segment (α_1,α_2,z), $0 \le z \le 1$, with the plane π_4 is

$$V_z^* = (\alpha_1,\alpha_2, \tfrac{1-\alpha_2-\alpha_1\alpha_2(1-\alpha_3)}{1-\alpha_2}) \ge \mathcal{A}_3\,, \qquad \forall\, \alpha_3 \in [0,1]\,.$$

b.1) assume that $\alpha_1+\alpha_2 \le 1$ and consider the set

$$S = \{(a,b)\,:\, a \ge \frac{1-\alpha_2}{1-\alpha_1-\alpha_2}\,,\ 1+\frac{\alpha_2}{1-\alpha_2}a \le b \le \frac{1-\alpha_1}{\alpha_1}a - \frac{1-\alpha_1}{\alpha_1}\}\,.$$

We have: $a>0\,,\ b>0\,,\ ag_2+bg_3 \ge g_1\,,\ \forall\,(a,b)\in S$. Then, g_1 is not relevant and $\mathcal{T}=\{2,3,4\}$ is a basic set. Moreover, $V_z^* = \lambda_2V_2+\lambda_3V_3+\lambda_4V_4$, with

$$\lambda_2=\alpha_1\,,\quad \lambda_3=\frac{\alpha_1\alpha_2}{1-\alpha_2}\,,\quad \lambda_4=\frac{1-\alpha_1-\alpha_2}{1-\alpha_2}\,.$$

We recall that $0<\alpha_i<1$, $i=1,2,3$, so that $\lambda_2>0\,,\ \lambda_3>0\,,\ \lambda_4\ge 0$. Then, the vector $(\lambda_2,\lambda_3,\lambda_4)$ is a solution of the system $(\mathcal{S}_3^{\mathcal{T}})$, with $|I_0^{\mathcal{T}}| \le 1$, and hence, by Theorem 3, $\mathcal{A}_3$ is g-coherent.
b.2) assume that $\alpha_2+\alpha_3 \le 1$ and consider the sets

$$S_1 = \{(a,b)\,:\, 0<a\le \frac{1-\alpha_2-\alpha_3+\alpha_2\alpha_3}{1-\alpha_2-\alpha_3}\,,\ \frac{\alpha_3}{1-\alpha_3}a \le b \le \frac{1-\alpha_2}{\alpha_2}a - \frac{1-\alpha_2}{\alpha_2}\}\,;$$

$$S_2 = \{(\gamma,\delta)\,:\, 0<\gamma\le \frac{\alpha_2\alpha_3(1-\alpha_3)}{1-\alpha_2-\alpha_3}\,,\ 1+\frac{\alpha_3}{1-\alpha_3}\gamma \le \delta \le \frac{1-\alpha_2}{\alpha_2}\gamma\}\,.$$

For each $(a,b)\in S_1$, $(\gamma,\delta)\in S_2$, one has

$$a>0\,,\quad b>0\,,\quad \gamma>0\,,\quad \delta>0\,,\quad ag_1+bg_2 \ge g_3\,,\quad \gamma g_1+\delta g_2 \ge g_4\,.$$

Then, g_3 and g_4 are not relevant and $\mathcal{T}=\{1,2\}$ is a basic set. Moreover, defining $V^*=(1,\alpha_2,1-\alpha_2)$, $\lambda_1=\alpha_2$, $\lambda_2=1-\alpha_2$, one has

$$V^* \ge (1,\alpha_2,\alpha_3) \ge \mathcal{A}_3\,;\quad V^*=\lambda_1V_1+\lambda_2V_2\,,\ \ \lambda_1>0\,,\ \lambda_2>0\,,\ \lambda_1+\lambda_2=1\,.$$

Then, the vector (λ_1,λ_2) is a solution of the system $(\mathcal{S}_3^{\mathcal{T}})$, with $I_0^{\mathcal{T}}=\emptyset$, and hence, by Theorem 3, $\mathcal{A}_3$ is g-coherent.
b.3) assume that $\alpha_1+\alpha_3 \le 1$ and consider the set

$$S = \{(a,b)\,:\, a \ge \frac{1-\alpha_3}{1-\alpha_1-\alpha_3}\,,\ 1+\frac{\alpha_3}{1-\alpha_3}a \le b \le \frac{1-\alpha_1}{\alpha_1}a - \frac{1-\alpha_1}{\alpha_1}\}\,.$$

We have: $a > 0\,,\ b > 0\,,\ \ ag_1 + bg_4 \ \geq\ g_2\,,\ \ \forall\,(a,b) \in S\,$. Then, g_2 is not relevant and $\mathcal{T} = \{1,3,4\}$ is a basic set. Moreover, $V_y^* = \lambda_1 V_1 + \lambda_3 V_3 + \lambda_4 V_4\,$, with

$$\lambda_1 = \alpha_1 > 0\,,\quad \lambda_3 = \frac{1-(\alpha_1+\alpha_3)}{1-\alpha_3} \geq 0\,,\quad \lambda_4 = \frac{\alpha_1\alpha_3}{1-\alpha_3} > 0\,.$$

Then, the vector $(\lambda_1, \lambda_3, \lambda_4)$ is a solution of the system $(\mathcal{S}_3^{\mathcal{T}})$, with $|I_0^{\mathcal{T}}| \leq 1$, and hence, by Theorem 3, $\mathcal{A}_3$ is g-coherent. Therefore, under the condition

$$\alpha_1 + \alpha_2 \leq 1\,,\quad \text{or}\quad \alpha_2 + \alpha_3 \leq 1\,,\quad \text{or}\quad \alpha_1 + \alpha_3 \leq 1\,,$$

$\mathcal{A}_3$ is g-coherent.
c) assume that $\alpha_1 + \alpha_2 > 1$, $\alpha_2 + \alpha_3 > 1$, $\alpha_1 + \alpha_3 > 1$.
c.1) if $\alpha_1\alpha_3 + \alpha_2 \leq 1$, then $V_x^* \geq \mathcal{A}_3$. Moreover, $V_x^* = \lambda_1 V_1 + \lambda_2 V_2 + \lambda_3 V_3$, with

$$\lambda_1 = \frac{(1-\alpha_2)(1-\alpha_3)}{\alpha_3} > 0\,,\quad \lambda_2 = 1-\alpha_2 > 0\,,\quad \lambda_3 = \frac{\alpha_2+\alpha_3-1}{\alpha_3} > 0\,.$$

Then, considering the basic set $\mathcal{T} = \{1,2,3,4\}$, the vector $(\lambda_1, \lambda_2, \lambda_3, 0)$ is a solution of the system $(\mathcal{S}_3^{\mathcal{T}})$, with $I_0^{\mathcal{T}} = \emptyset$, and hence, by Theorem 3, $\mathcal{A}_3$ is g-coherent.
c.2) if $\alpha_1\alpha_2 + \alpha_3 \leq 1$, then $V_x^{**} \geq \mathcal{A}_3$. Moreover, $V_x^{**} = \lambda_1 V_1 + \lambda_2 V_2 + \lambda_4 V_4$, with

$$\lambda_1 = 1-\alpha_3 > 0\,,\quad \lambda_2 = \frac{(1-\alpha_2)(1-\alpha_3)}{\alpha_2} > 0\,,\quad \lambda_4 = \frac{\alpha_2+\alpha_3-1}{\alpha_2} > 0\,.$$

Then, considering the basic set $\mathcal{T} = \{1,2,3,4\}$, the vector $(\lambda_1, \lambda_2, 0, \lambda_4)$ is a solution of the system $(\mathcal{S}_3^{\mathcal{T}})$, with $I_0^{\mathcal{T}} = \emptyset$, and hence, by Theorem 3, $\mathcal{A}_3$ is g-coherent.
c.3) assume that $\alpha_1\alpha_2 + \alpha_3 > 1$, $\alpha_1\alpha_3 + \alpha_2 > 1$, and let us make the (absurd) hypothesis that $\mathcal{A}_3$ were g-coherent. Then, considering the basic set $\mathcal{T} = \{1,2,3,4\}$, the system $(\mathcal{S}_3^{\mathcal{T}})$ should be solvable and hence, for suitable non negative values $\lambda_1, \ldots, \lambda_4$, with $\lambda_1 + \cdots + \lambda_4 = 1$, defining

$$V^* \,=\, \lambda_1 V_1 + \lambda_2 V_2 + \lambda_3 V_3 + \lambda_4 V_4 \,=\, (\lambda_1 + \lambda_2\,, \lambda_1 + \lambda_3 + \alpha_2\lambda_4\,, \lambda_2 + \alpha_3\lambda_3 + \lambda_4)\,,$$

it should be: $V^* \,\geq\, \mathcal{A}_3\,$, that is

$$\lambda_1 + \lambda_2 \,\geq\, \alpha_1\,;\quad \lambda_1 + \lambda_3 \,\geq\, \alpha_2 - \alpha_2\lambda_4\,;\quad \lambda_2 + \lambda_4 \,\geq\, \alpha_3(\lambda_1 + \lambda_2 + \lambda_4)\,, \qquad (16)$$

or, equivalently

$$\lambda_1 + \lambda_2 \,\geq\, \alpha_1\,;\quad \lambda_1 + \lambda_3 \,\geq\, \alpha_2(\lambda_1 + \lambda_2 + \lambda_3)\,;\quad \lambda_2 + \lambda_4 \,\geq\, \alpha_3 - \alpha_3\lambda_3\,. \qquad (17)$$

Then, assuming $\alpha_3 - \alpha_2 \,\geq\, 0$ and recalling that $\alpha_1\alpha_3 + \alpha_2 > 1$, by summing the last two inequalities in (16) we would obtain

$$1 \,\geq\, \alpha_3(\lambda_1 + \lambda_2) + \alpha_2 + (\alpha_3 - \alpha_2)\lambda_4 \,\geq\, \alpha_1\alpha_3 + \alpha_2 + (\alpha_3 - \alpha_2)\lambda_4 \,>\, 1\,,$$

which is absurd. On the other hand, assuming $\alpha_3 - \alpha_2 < 0$ and recalling that $\alpha_1\alpha_2 + \alpha_3 > 1$, by summing the last two inequalities in (17) we would obtain

$$1 \geq \alpha_2(\lambda_1 + \lambda_2) + \alpha_3 + (\alpha_2 - \alpha_3)\lambda_3 \geq \alpha_1\alpha_2 + \alpha_3 + (\alpha_2 - \alpha_3)\lambda_3 > 1\,,$$

which is absurd too. Hence, $(\mathcal{S}_3^{\mathcal{T}})$ is not solvable and $\mathcal{A}_3$ is not g-coherent. □

We observe that the hypotheses concerning the set of logical relations L can be modified in many ways. Then, by the same reasoning as in Theorem 4, we obtain many similar results, which we give without proof in the remaining part of this section (the proofs of these results can be found in [6]).

Theorem 5 If $\mathcal{V}_0 = \mathcal{V}_1 = \mathcal{V}_2 = \emptyset,\ \mathcal{U}_{0,1} = \{V_1, V_2\} = \{(1,1,0),(1,0,1)\}, V_3 = (0,1,\alpha_3) \in \mathcal{U}_{1,1}, (0,\alpha_2,1) \notin \mathcal{U}_{1,1},\ \alpha_i < 1\,, \forall\, i$, then one has:
a) for every $r > 3$, the gain g_r is not relevant;
b) if $\alpha_2 + \alpha_3 \leq 1$, then there exists a basic set $\mathcal{T}$, with $|\mathcal{T}| = 2$, and $\mathcal{A}_3$ is g-coherent;
c) if $\alpha_2 + \alpha_3 > 1$, then $\mathcal{A}_3$ is g-coherent iff $\alpha_1\alpha_3 + \alpha_2 \leq 1$.

Theorem 6 If $\mathcal{V}_0 = \mathcal{V}_1 = \mathcal{V}_2 = \emptyset,\ \mathcal{U}_{0,1} = \{V_1\} = \{(1,1,0)\}, \{V_2, V_3, V_4, V_5\} = \{(\alpha_1,0,1),(0,\alpha_2,1),(1,0,\alpha_3),(0,1,\alpha_3)\} \subseteq \mathcal{U}_{1,1},\ \alpha_i < 1\,, \forall\, i$, then one has:
a) for every $r > 5$, the gain g_r is not relevant;
b) if $\alpha_1 + \alpha_2 \leq 1$, then there exists a basic set $\mathcal{T}$, with $|\mathcal{T}| = 4$, and $\mathcal{A}_3$ is g-coherent.
c) if $\alpha_1 + \alpha_2 > 1$, then $\mathcal{A}_3$ is g-coherent iff

$$\alpha_3 \leq Max\{\frac{\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2}{\alpha_1 + \alpha_2 - \alpha_1\alpha_2}\,,\ 1 - \alpha_1 + \alpha_1\alpha_3 - \alpha_1\alpha_2\alpha_3\,,\ 1 - \alpha_1\alpha_3\}\,.$$

Remark: We observe that, by suitably modifying the hypotheses in Theorems 4, 5, and 6, we obtain similar results on non relevant gains and basic sets, with further conditions characterizing the g-coherence of the assessment $\mathcal{A}_3$ on $\mathcal{F}_3$. As an example, by Theorem 4, still assuming $\mathcal{V}_0 = \mathcal{V}_1 = \mathcal{V}_2 = \emptyset$, under the hypotheses

$$\mathcal{U}_{0,1} = \{V_1, V_2\} = \{(1,1,0),(0,1,1)\},\ \{V_3, V_4\} = \{(1,0,\alpha_3),(\alpha_1,0,1)\} \subseteq \mathcal{U}_{1,1}\,,$$

we obtain a new result, which is similar to such theorem, and so on.

Theorem 7 If $\mathcal{V}_0 = \mathcal{V}_1 = \mathcal{V}_2 = \mathcal{U}_{0,1} = \emptyset,\ \mathcal{U}_{1,1} = \{V_1, V_2, V_3, V_4, V_5, V_6\} = \{(1,\alpha_2,0),(1,0,\alpha_3),(\alpha_1,1,0),(0,1,\alpha_3),(\alpha_1,0,1),(0,\alpha_2,1))\},\ \alpha_i < 1, \forall\, i$, then one has:
a) for every $r > 6$, the gain g_r is not relevant;
b) if $\alpha_1 + \alpha_2 \leq 1$, or $\alpha_1 + \alpha_3 \leq 1$, or $\alpha_2 + \alpha_3 \leq 1$, then there exists a basic set $\mathcal{T}$, with $|\mathcal{T}| = 4$, and $\mathcal{A}_3$ is g-coherent.
c) if $\alpha_1 + \alpha_2 > 1$, $\alpha_1 + \alpha_3 > 1$, $\alpha_2 + \alpha_3 > 1$, then $\mathcal{A}_3$ is not g-coherent.

Theorem 8 If $\mathcal{V}_0 = \mathcal{V}_1 = \mathcal{V}_2 = \mathcal{U}_{0,1} = \emptyset$, $\mathcal{U}_{1,1} = \{V_1, V_2, V_3, V_4\} = \{(1,\alpha_2,0),(\alpha_1,1,0),(\alpha_1,0,1),(0,\alpha_2,1))\}$, $\alpha_i < 1\,\forall\, i$, then one has:
a) for every $r > 4$, the gain g_r is not relevant;
b) if $\alpha_1 + \alpha_3 \leq 1$, or $\alpha_2 + \alpha_3 \leq 1$, then there exists a basic set $\mathcal{T}$, with $|\mathcal{T}| = 2$, and $\mathcal{A}_3$ is g-coherent.
c) if $\alpha_1 + \alpha_3 > 1$, $\alpha_2 + \alpha_3 > 1$, then $\mathcal{A}_3$ is not g-coherent.

Theorem 9 If $\mathcal{V}_0 = \mathcal{V}_1 = \mathcal{V}_2 = \mathcal{U}_{0,1} = \emptyset$, $\mathcal{U}_{1,1} = \{V_1, V_2, V_3, V_4\} = \{(1,0,\alpha_3),(0,1,\alpha_3),(\alpha_1,1,0),(\alpha_1,0,1))\}$, $\alpha_i < 1\,\forall\, i$, then one has:
a) for every $r > 4$, the gain g_r is not relevant;
b) if $\alpha_1 + \alpha_2 \leq 1$, or $\alpha_2 + \alpha_3 \leq 1$, then there exists a basic set $\mathcal{T}$, with $|\mathcal{T}| = 2$, and $\mathcal{A}_3$ is g-coherent.
c) if $\alpha_1 + \alpha_2 > 1$, $\alpha_2 + \alpha_3 > 1$, then $\mathcal{A}_3$ is not g-coherent.

Theorem 10 If $\mathcal{V}_0 = \mathcal{V}_1 = \mathcal{V}_2 = \mathcal{U}_{0,1} = \emptyset$, $\mathcal{U}_{1,1} = \{V_1, V_2, V_3, V_4\} = \{(1,0,\alpha_3),(0,1,\alpha_3),(1,\alpha_2,0),(0,\alpha_2,1))\}$, $\alpha_i < 1\,\forall\, i$, then one has:
a) for every $r > 4$, the gain g_r is not relevant;
b) if $\alpha_1 + \alpha_2 \leq 1$, or $\alpha_1 + \alpha_3 \leq 1$, then there exists a basic set $\mathcal{T}$, with $|\mathcal{T}| = 2$, and $\mathcal{A}_3$ is g-coherent.
c) if $\alpha_1 + \alpha_2 > 1$, $\alpha_1 + \alpha_3 > 1$, then $\mathcal{A}_3$ is not g-coherent.

5 Some general results

In this section we give some theorems on g-coherence of a vector of lower probability bounds $\mathcal{A}_n$ defined on a family of n conditional events $\mathcal{F}_n$. Notice that detailed proofs of all theorems presented in this section are given in [6].
In the next theorem we generalize the condition 6 in Remark 4. In such theorem the set of logical relations L specifies that the conjunctions

$$\begin{array}{l} E_1H_1 \cdots E_nH_n\,, \quad E_1H_1 \cdots E_{n-1}H_{n-1}H_n^c\,, \quad \ldots\,, \quad H_1^cE_2H_2 \cdots E_nH_n\,, \\ E_1H_1 \cdots E_{n-2}H_{n-2}H_{n-1}^cH_n^c\,, \quad \ldots\,, \quad H_1^cH_2^cE_3H_3 \cdots E_nH_n\,, \quad \ldots \;\; \ldots\,, \\ E_1H_1H_2^c \cdots H_n^c\,, \quad \ldots\,, \quad H_1^c \cdots H_{n-1}^cE_nH_n \end{array}$$

are impossible. Then, under such hypotheses, the condition $\alpha_1 + \cdots + \alpha_n \leq n-1$ is necessary for the g-coherence of $\mathcal{A}_n$.

Theorem 11 If $\mathcal{V}_0 = \mathcal{V}_1 = \cdots = \mathcal{V}_{n-1} = \emptyset$ and $\alpha_1 + \cdots + \alpha_n > n-1$, then $\mathcal{A}_n$ is not g-coherent.

In the next theorem we generalize the condition 7 given in Remark 4.

Theorem 12 If $\mathcal{V}_0 = \cdots = \mathcal{V}_{n-1} = \emptyset$, $|\mathcal{U}_{0,1}| = n$, $0 < \alpha_i < 1\,\forall\, i$, then one has:
a) there exists a basic set $\mathcal{T}$, with $|\mathcal{T}| = n$;
b) $\mathcal{A}_n$ is g-coherent iff $\alpha_1 + \cdots + \alpha_n \leq n-1$.

We denote by $\mathcal{Z}$ the set defined as

$$\mathcal{Z} = \{(h,k) \,:\, h+k = n-1\,,\; h>0\} \cup \{(h,k) \,:\, h+k < n-1\}\,.$$

Then, we have

Theorem 13 If $\mathcal{V}_0 = \cdots = \mathcal{V}_{n-1} = \emptyset$, $\mathcal{U}_{h,k} = \emptyset$ for each $(h,k) \in \mathcal{Z}$, and $\alpha_1 + \cdots + \alpha_n > 1$, then $\mathcal{A}_n$ is not g-coherent.

The next result generalizes the condition 8 in Remark 4.

Theorem 14 If $\mathcal{V}_0 = \cdots = \mathcal{V}_{n-1} = \emptyset$, $\mathcal{U}_{h,k} = \emptyset$, for each pair $(h,k) \in \mathcal{Z}$, $|\mathcal{U}_{0,n-1}| = n$, $0 < \alpha_i < 1\; \forall\, i$, then one has:
a) if, for every $j \in J_n$, it is $\sum_{i \in J_n \setminus \{j\}} \alpha_i \leq 1$, then $\mathcal{T} = J_n$ is a basic set;
b) $\mathcal{A}_n$ is g-coherent iff $\alpha_1 + \cdots + \alpha_n \leq 1$.

6 Conclusions

Exploiting the coherence principle of de Finetti and the related notion of g-coherence, we illustrated a probabilistic approach to uncertain reasoning based on lower probability bounds. We examined the notions of non relevant gain and basic set which may be helpful, in g-coherence checking and propagation of conditional probability bounds, to reduce the sets of variables and/or constraints in the linear systems used in our algorithms. We observe that such notions and in particular the condition (6), in the form $g_r = \sum_{j \in \mathcal{T}_r} g_j$, have been used in ([3], Theorem 5.6) to characterize in term of random gains an efficient procedure proposed in [16] for families of conjunctive conditional events. To provide a better understanding of these notions, we examined several examples of families constituted by three conditional events. This case may have a specific importance, e.g., in default reasoning where many inference rules consist of two premises and one conclusion. We obtained some necessary and sufficient conditions of g-coherence and we also generalized some theoretical results. Further work should allow to extend the results of this paper to the case of families of n conditional events, with $n > 3$.

Acknowledgments
The authors are grateful to the referees for the valuable criticisms and suggestions.

References

[1] Biazzo V., and Gilio A.: A generalization of the fundamental theorem of de Finetti for imprecise conditional probability assessments. International Journal of Approximate Reasoning **24**: 251-272 (2000)

[2] Biazzo V., Gilio A.: On the linear structure of betting criterion and the checking of coherence. Annals of Mathematics and Artificial Intelligence **35**: 83-106 (2002)

[3] Biazzo V., Gilio A., Lukasiewicz T., and Sanfilippo G.: Probabilistic Logic under Coherence: Complexity and Algorithms. Proc. of The Second International Symposium on Imprecise Probabilities and their Applications (ISIPTA '01), Ithaca, USA, June 26 - 29, 51-61 (2001)

[4] Biazzo V., Gilio A., Lukasiewicz T., and Sanfilippo G.: Probabilistic Logic under Coherence, Model-Theoretic Probabilistic Logic, and Default Reasoning in System P. Journal of Applied Non-Classical Logics **12**(2): 189-213 (2002)

[5] Biazzo V., Gilio A., and Sanfilippo G.: Coherence Checking and Propagation of Lower Probability Bounds. Soft Computing **7:** 310-320 (2003)

[6] Biazzo V., Gilio A., and Sanfilippo G.: On the checking of g-coherence of conditional probability bounds. Preprint (2002) (available at *http://www.dmmm.uniroma1.it/∼ gilio/publications/BGS-1.pdf*).

[7] Capotorti A., and Vantaggi B.: Locally strong coherence in inference processes. Annals of Mathematics and Artificial Intelligence **35**: 125-149 (2002)

[8] Coletti G., and Scozzafava R.: Conditioning and inference in intelligent systems. Soft Computing **3**(3): 118-130 (1999)

[9] Cozman F. G.: Algorithms for Conditioning on Events of Zero Probability. Fifteenth International Florida Artificial Intelligence Society Conference. Pensacola, Florida, 248-252 (2002)

[10] Gilio A.: Probabilistic consistency of conditional probability bounds. Advances in Intelligent Computing, Lecture Notes in Computer Science 945 (B. Bouchon-Meunier, R. R. Yager, and L. A. Zadeh, Eds.), Springer-Verlag, Berlin Heidelberg, 200-209 (1995)

[11] Gilio A.: Probabilistic reasoning under coherence in System P. Annals of Mathematics and Artificial Intelligence **34**: 5-34 (2002)

[12] Hansen, P.; Jaumard, B.; Nguetsé, G.-B. D.; and de Aragão M. P.: Models and algorithms for probabilistic and Bayesian logic. Proc. of IJCAI-95, 1862-1868 (1995)

[13] Hansen P., Jaumard B., de Aragao M.P., Chauny F., and Perron S.: Probabilistic satisfiability with imprecise probabilities. International Journal of Approximate Reasoning **24**(2-3): 171-189 (2000)

[14] Jaumard, B.; Hansen, P.; and de Aragão, M. P.: Column generation methods for probabilistic logic. ORSA J. Comput. **3**: 135-147 (1991)

[15] Kraus, K.; Lehmann, D.; and Magidor, M.: Nonmonotonic reasoning, preferential models and cumulative logics. Artificial Intelligence **44**: 167-207 (1990)

[16] Lukasiewicz T.: Efficient Global Probabilistic Deduction from Taxonomic and Probabilistic Knowledge-Bases over Conjunctive Events. Proc. of the 6th International Conference on Information and Knowledge Management, ACM Press, 75-82 (1997)

[17] Lukasiewicz, T.: Local probabilistic deduction from taxonomic and probabilistic knowledge-bases over conjunctive events. Intern. J. Approx. Reason. **21**: 23-61 (1999)

[18] Lukasiewicz, T.: Probabilistic deduction with conditional constraints over basic events. Journal of Artificial Intelligence Research **10**: 199-241 (1999)

[19] Walley P.: Statistical reasoning with imprecise probabilities, Chapman and Hall, London (1991)

[20] Walley P., Pelessoni R., and Vicig P.: Direct Algorithms for Checking Coherence and Making Inferences from Conditional Probability Assessments. Quaderni del Dipartimento di Matematica Applicata alle Scienze Economiche, Statistiche e Attuariali ”B. de Finetti” 6 (1999)

V. Biazzo & G. Sanfilippo are with the Dipartimento di Matematica e Informatica, Università di Catania, Italy. E-mail: vbiazzo@dmi.unict.it, gsanfilippo@dmi.unict.it

A. Gilio is with the Dipartimento di Metodi e Modelli Matematici, Università “La Sapienza,” Roma, Italy. E-mail: gilio@dmmm.uniroma1.it

The Maximal Variance of Fuzzy Interval[*]

A.G. BRONEVICH
Taganrog State University of Radio-Engineering, Russia
Bremen University, Germany

Abstract

The paper gives the solution of calculating maximal variance of fuzzy interval in the scope of the theory of imprecise probabilities. As it appears, this problem is more difficult than analogous one connected with evaluation of lower and upper expectations of fuzzy interval. This paper gives some contribution to possibility theory in the framework of probability approach.

Keywords

possibility measure, upper and lower probabilities, maximal variance

1 Introduction

There is a well-known interpretation of fuzzy interval in the framework of the theory of imprecise probabilities [1, 2]. To get this, we associate with any fuzzy interval a possibility or necessity measure, and then consider that values of the pointed measures give us lower or upper assessments of probabilities. This interpretation was discussed in detail in [3], and there it is proposed to use upper and lower expectations for evaluating uncertainty of such intervals. These characteristics and other crude moments of order k can be easily calculated by Choquet integral. However, to calculate upper and lower central moments is more difficult as it is shown in investigations, presented below.

Throughout the paper we will use the following notations: 1) $E[\xi]$ is an ordinary expectation of the random variable ξ, i.e. $E[\xi] = \int\limits_{-\infty}^{+\infty} x dP(x)$, where P is a probability measure associated with the random variable ξ; 2) $\sigma^2[\xi]$ is an ordinary variance of the random variable ξ, i.e. $\sigma^2[\xi] = E\left[\xi^2\right] - (E[\xi])^2$.

[*] I would like to express my sincere thanks to the German Academic Exchange Service (DAAD), Bremen University and Prof. Dr. Dieter Denneberg for the research opportunity provided.

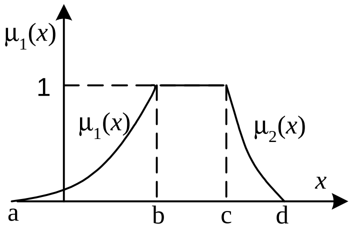


Figure 1: Membership function of a fuzzy interval

2 Basic definitions and problem statement

We will consider fuzzy intervals with a form (fig.1). The function μ is assumed to be continuous, and the functions μ_1 and μ_2 are differentiable on the intervals (a,b) and (c,d) correspondingly.

$$\mu(x) = \begin{cases} 0, & x \le a \ \ or \ \ x \ge d, \\ \mu_1(x), & a < x < b, \\ 1, & b \le x \le c, \\ \mu_2(x), & c < x < d. \end{cases} \tag{1}$$

In addition, μ_1 is increasing on (a,b), μ_2 is decreasing on (c,d).

In possibility theory, for each fuzzy interval, a possibility measure $\Pi(A) = \sup\limits_{x\in A} \mu(x)$ and a necessity measure $N(A) = \inf\limits_{x\notin A}[1-\mu(x)]$ are introduced, and can be considered as lower or upper estimation of probability of the event $A \in \Im$ (where $\Im$ is Borel algebra of real axis). Taking this into account, possibility measure Π and necessity measure N define a family of probability measures $\Xi = \{P\,|N(A) \le P(A) \le \Pi(A)\,\}$, and the problem arises, how to calculate digital characteristics of such family, in particular, the maximal variance $\overline{\sigma^2}(\mu) = \sup\limits_{P_i\in\Xi} \sigma^2[\xi_i]$. In the last expression, it is assumed that the probability measure P_i determines a random value ξ_i. For the fuzzy interval, the value $\overline{\sigma^2}(\mu)$ can serve as some characteristic of uncertainty.

3 The research of possibilistic inclusion

Theorem 1 [4, 5]. *Let P be a probability measure, Ξ a family of probability measures, generated by a fuzzy interval with a membership function μ. Then $P \in \Xi$ iff $P\{A(p)\} \le p$ for all $p \in [0,1]$, where $A(p) = \{x \in R\,|\mu(x) \le p\,\}$.*

Theorem 1 can be reformulated by using standard terms for random values as follows.

Theorem 1*. *Let we use the same notations as in theorem 1, and the random value ξ is described by the probability measure P on $\Im$. Consider also a random value $\eta = \mu(\xi) \in [0,1]$. Then $P \in \Xi$ iff $F_\eta(y) \leq y$, where $F_\eta(y) = P\{\eta \leq y\}$.*

Remark. The function F_η is a distribution function of η, whenever η is continuous.

4 The solution of the optimization problem

Theorem 2. *Let ξ be a random value, described by a probability measure $P \in \Xi$, in addition, $\sigma^2[\xi] = \overline{\sigma^2}(\mu)$. Then we have $P\{(b,c)\} = 0$ for fuzzy interval (1).*

Proof. Suppose that the coordinate system has been chosen in a way that $E[\xi] = 0$. Assume also that $b < 0$, and the condition of the theorem is not fulfilled, i.e. $P(b,0] > 0$. The theorem is valid if one can find such a measure P^* that $P^* \in \Xi$ and $\sigma^2[\xi^*] > \sigma^2[\xi]$. We will search the probability measure P^* in a form:

$$P^*(A) = \begin{cases} P(A)\varepsilon, & A \subseteq (b,0], \\ P\{b\} + P(b,0](1-\varepsilon), & A = \{b\}, \\ P(A), & A \cap [b,0] = \emptyset. \end{cases}$$

It is obvious that P^* extends on $\Im$ uniquely and $P^* \in \Xi$. Calculate derivative of

$$\sigma^2[\xi^*] = \int_{-\infty}^{+\infty} x^2 dP^*(x) - \left[\int_{-\infty}^{+\infty} x dP^*(x)\right]^2$$

w.r.t. ε at the point $\varepsilon = 1$. Since $\int_{-\infty}^{+\infty} x dP^*(x) = 0$ at the point $\varepsilon = 1$,

$$\frac{d}{d\varepsilon}\left(\sigma^2[\xi^*]\right)_{\varepsilon=1} = \frac{d}{d\varepsilon}\left[\int_{-\infty}^{+\infty} x^2 dP^*(x)\right]_{\varepsilon=1}.$$

Describe the last expression in detail.

$$\int_{-\infty}^{+\infty} x^2 dP^*(x) = \int_{R\setminus[b,0]} x^2 P(x) + b^2\left(P\{b\} + P(b,0](1-\varepsilon)\right) + \varepsilon \int_{(b,0]} x^2 dP(x).$$

Therefore,

$$\frac{d}{d\varepsilon}\left(\sigma^2[\xi^*]\right)_{\varepsilon=1} = -b^2 P(b,0] + \int_{(b,0]} x^2 dP(x) < 0.$$

It means that there exists $\varepsilon < 1$ that $\sigma^2[\xi^*] > \sigma^2[\xi]$. For the complete proof of the theorem, we must consider also a case, where $c > 0$ and $P[0,c) > 0$.

Corollary. *Let* $P \in \Xi$, $\sigma^2[\xi] = \overline{\sigma^2}(\mu)$ *as in theorem 2, in addition,* $E[\xi] = 0$. *Then*

1) $P[0,c) = 0$ *if* $c > 0$*;*

2)$P(b,0) = 0$ *if* $b < 0$.

Theorem 3. *Let* $P \in \Xi$, $\sigma^2[\xi] = \overline{\sigma^2}(\mu)$ *as in theorem 2. Then the random value* η *has a distribution function* $F_\eta(y) = y$.

Proof. We will assume that the coordinate system has been chosen in a way that $E[\xi] = 0$. Suppose the contrary assumption, that for ξ from the theorem, $F_\eta(y) \neq y$. The theorem will be proved, if under this condition, there exists a random value ξ^* associated with a probability measure $P^* \in \Xi$ for which $\sigma^2[\xi^*] > \sigma^2[\xi]$. The random value ξ^* will be searched for a certain $\alpha \in [0,1]$, using the expression:

$$\xi^* = \begin{cases} \mu_1^{-1}\left[\alpha F_\eta\left(\mu(\xi)\right) + (1-\alpha)\mu(\xi)\right], & \xi \in [a,b], \\ \mu_2^{-1}\left[\alpha F_\eta\left(\mu(\xi)\right) + (1-\alpha)\mu(\xi)\right], & \xi \in [c,d]. \end{cases}$$

Hence, we need to find $\alpha \in [0,1]$ such that $\sigma^2[\xi^*] > \sigma^2[\xi]$. But at first, check that ξ^* generates the probability measure $P^* \in \Xi$. To do this, we need to confirm that the inequality

$$F_{\eta^*}(y) = P\left\{\mu(\xi^*) \leq y\right\} \leq y$$

is valid. Actually,

$$\eta^* = \mu(\xi^*) = \alpha F_\eta\left(\mu(\xi)\right) + (1-\alpha)\mu(\xi),$$

$$F_{\eta^*}(y) = P\left\{\alpha F_\eta\left(\mu(\xi)\right) + (1-\alpha)\mu(\xi) \leq y\right\}.$$

Since $F_\eta(y) \leq y$, then $\left\{\alpha F_\eta\left(\mu(\xi)\right) + (1-\alpha)\mu(\xi) \leq y\right\} \subseteq \left\{F_\eta\left(\mu(\xi)\right) \leq y\right\}$. Therefore,

$$F_\eta\left(y\right) \leq P\left\{F_\eta\left(\mu(\xi)\right) \leq y\right\} = P\left\{\mu(\xi) \leq F_\eta^{-1}(y)\right\} = F_\eta\left(F_\eta^{-1}(y)\right) = y.$$

Thus, it has been shown that $P^* \in \Xi$. Further we will prove that $\sigma^2[\xi^*] > \sigma^2[\xi]$ for a certain $\alpha \in [0,1]$. To do this, calculate derivative of

$$\frac{d}{d\alpha}\sigma^2\left[\xi^*\right] = \frac{d}{d\alpha}\left(E\left[(\xi^*)^2\right] - \left(E\left[\xi^*\right]\right)^2\right)_{\alpha=0}$$

at the point $\alpha = 0$. Since $E\left[\xi^*\right]_{\alpha=0} = E\left[\xi\right] = 0$,

$$\left.\frac{d}{d\alpha}\sigma^2\left[\xi^*\right]\right|_{\alpha=0} = \left.\frac{d}{d\alpha}E\left[(\xi^*)^2\right]\right|_{\alpha=0} =$$

$$\frac{d}{d\alpha}\left(\int_a^b \left[\mu_1^{-1}\left[\alpha F_\eta\left(\mu(x)\right) + (1-\alpha)\mu(x)\right]\right]^2 dP(x)\right)_{\alpha=0} +$$

$$+\frac{d}{d\alpha}\left(\int_c^d \left[\mu_2^{-1}\left[\alpha F_\eta\left(\mu(x)\right)+(1-\alpha)\mu(x)\right]\right]^2 dP(x)\right)_{\alpha=0}.$$

Taking derivative w.r.t. α at the point $\alpha = 0$, we get

$$\left.\frac{d}{d\alpha}\sigma^2\left[\xi^*\right]\right|_{\alpha=0} = \int_a^b 2x\left.\frac{d}{dy}\mu_1^{-1}(y)\right|_{y=\mu(x)}\left(F_\eta\left(\mu(x)\right)-\mu(x)\right)dP(x)+$$

$$+\int_c^d 2x\left.\frac{d}{dy}\mu_2^{-1}(y)\right|_{y=\mu(x)}\left(F_\eta\left(\mu(x)\right)-\mu(x)\right)dP(x).$$

Analyze signs of factors stating in the integrands.

1) $F_\eta\left(\mu(x)\right)-\mu(x)\le 0$, in addition, since according to our supposition $F_\eta(y)\ne y$, $y\in[0,1]$, there exists a non-empty set of points, in which $F_\eta\left(\mu(x)\right)-\mu(x)<0$. Since F_η is continuous, increasing function, $P\left\{F_\eta\left(\mu(\xi)\right)-\mu(\xi)<0\right\}>0$.

2) The function μ_1 is increasing on $[a,b]$, therefore, $\left.\frac{d}{dy}\mu_1^{-1}(y)\right|_{y=\mu(x)}>0$ if $x\in(a,b)$.

3) According to the corollary of theorem 2, $P[0,c)=0$. It enables to exchange the area of integration to $(a,\min\{b,0\})$. Notice that $(2x)<0$ if x is in this interval.

4) The function μ_2 is decreasing on $[c,d]$, thus, $\left.\frac{d}{dy}\mu_2^{-1}(y)\right|_{y=\mu(x)}<0$, whenever $x\in(c,d)$.

5) According to the corollary of theorem 2, $P[d,0)=0$. It enables to exchange the area of integration to $(\max\{c,0\},d)$ in the second integral. Note that the factor $(2x)>0$ in this interval.

Analyzing signs of integrals, one can confirm that value of each of them is non-negative; in addition, one of them is strictly positive. Hence, $\left.\frac{d}{d\alpha}\sigma^2\left[\xi^*\right]\right|_{\alpha=0}>0$. It means that one can find $\alpha>0$ that $\sigma^2[\xi^*]>\sigma^2[\xi]$, i.e. the supposition has been made is wrong, and it implies $F_\eta(y)=y$.

The proved theorems enable to make some simplifying of our optimization problem. To do this, introduce into consideration the functions

$$F_{\eta_1}(y)=P\left\{\mu_1(\xi)\le y\,|\xi\in[a,b]\right\},F_{\eta_2}(y)=P\left\{\mu_2(\xi)\le y\,|\xi\in[c,d]\right\}.$$

It is clear that $\eta_1=\mu(\xi_1)$, $\eta_2=\mu(\xi_2)$, and also the random value ξ_1 is associated with the probability measure $P\{*\,|\xi\in[a,b]\}$, ξ_2 with the probability measure $P\{*\,|\xi\in[c,d]\}$. Let $P\in\Xi$ and $\sigma^2[\xi]=\overline{\sigma^2}(\mu)$, then $P\{R\backslash[a,b]\cup[c,d]\}=0$, and, using formula of composite probability, one can write:

$$F_\eta(y)=F_{\eta_1}(y)P\{\xi\in[a,b]\}+F_{\eta_2}(y)P\{\xi\in[c,d]\}.$$

By theorem 3, $F_\eta(y) = y$. Assume that functions $F_{\eta_1}(y)$ and $F_{\eta_2}(y)$ are differentiable, then the calculation of probability is transformed to Riemannian integral:

$$P\{\xi \in A\} = P\{\xi \in [a,b]\} \int\limits_{\mu\{A\cap[a,b]\}} dF_{\eta_1}(y) + P\{\xi \in [c,d]\} \int\limits_{\mu\{A\cap[c,d]\}} dF_{\eta_2}(y). \tag{2}$$

By analogy, using the expression $\xi = \begin{cases} \mu_1^{-1}(\eta_1), \xi \in [a,b], \\ \mu_2^{-1}(\eta_2), \xi \in [c,d], \end{cases}$ one can write the formula for calculating moments:

$$E\left[\xi^k\right] = P\{\xi \in [a,b]\} \int\limits_0^1 \left[\mu_1^{-1}(y)\right]^k dF_{\eta_1}(y) + P\{\xi \in [c,d]\} \int\limits_0^1 \left[\mu_2^{-1}(y)\right]^k dF_{\eta_2}(y).$$

Introduce the following notations:

$$h_1(y) = P\{\xi \in [a,b]\} F'_{\eta_1}(y), \qquad h_2(y) = P\{\xi \in [c,d]\} F'_{\eta_2}(y).$$

Then

$$E\left[\xi^k\right] = \int\limits_0^1 \left[\mu_1^{-1}(y)\right]^k h_1(y)dy + \int\limits_0^1 \left[\mu_2^{-1}(y)\right]^k h_2(y)dy. \tag{3}$$

It is clear, that functions h_1, h_2 have to be non-negative in $[0,1]$, in addition, $h_1(y) + h_2(y) = 1$ by theorem 3.

Theorem 4. *Let ξ be associated with a probability measure P, $P \in \Xi$, $\sigma^2[\xi] = \overline{\sigma^2}(\mu)$, $E[\xi] = 0$, and $E\left[\xi^k\right]$ is calculated by formula (3). In addition, functions h_1, h_2 are piecewise continuous. Then in the range of h_i continuity the following formula is valid:*

$$h_1(y) = \begin{cases} 1, \left|\mu_1^{-1}(y)\right| > \left|\mu_2^{-1}(y)\right|, \\ 0, \left|\mu_1^{-1}(y)\right| < \left|\mu_2^{-1}(y)\right|, \end{cases} \qquad h_2(y) = 1 - h_1(y). \tag{4}$$

Proof. Assume, on the contrary, that the condition of the theorem takes place, but formula (4) is not valid at least for one point of $h_i(y)$ continuity. The theorem is valid if, for this case, we can find a random value ξ^*, associated with a probability measure $P^* \in \Xi$ such that $\sigma^2[\xi^*] > \sigma^2[\xi]$. To do this, introduce into consideration the following functions:

$$g_1(y) = \begin{cases} 1, & \left|\mu_1^{-1}(y)\right| > \left|\mu_2^{-1}(y)\right|, \\ 0, & \left|\mu_1^{-1}(y)\right| < \left|\mu_2^{-1}(y)\right|, \\ h_1(y), & \left|\mu_1^{-1}(y)\right| = \left|\mu_2^{-1}(y)\right|, \end{cases} \qquad g_2(y) = 1 - g_1(y),$$

and also

$$h_1^*(y) = g_1(y)\alpha + h_1(y)(1-\alpha), \quad h_2^*(y) = g_2(y)\alpha + h_2(y)(1-\alpha), \ \ y \in [0,1].$$

It is assumed that functions h_1^*, h_2^* generate the probability distribution of ξ^* by the formula:

$$P\{\xi^* \in A\} = \int\limits_{\mu\{A\cap[a,b]\}} h_1^*(y)dy + \int\limits_{\mu\{A\cap[c,d]\}} h_2^*(y)dy.$$

It is clear that the last formula is an analog of formula (2), and the random value ξ^* generates the probability measure $P^* \in \Xi$ for all values $\alpha \in [0,1]$. Calculate derivative of $\frac{d}{d\alpha}\sigma^2[\xi^*] = \frac{d}{d\alpha}\left(E\left[(\xi^*)^2\right] - (E[\xi^*])^2\right)_{\alpha=0}$ at the point $\alpha = 0$. Since $E[\xi^*] = 0$ for $\alpha = 0$, we get $\left.\frac{d}{d\alpha}\sigma^2[\xi^*]\right|_{\alpha=0} = \left.\frac{d}{d\alpha}E\left[(\xi^*)^2\right]\right|_{\alpha=0}$. Then

$$\frac{d}{d\alpha}E\left[(\xi^*)^2\right] = \frac{d}{d\alpha}\int\limits_0^1 \left[\mu_1^{-1}(y)\right]^2 [g_1(y)\alpha + h_1(y)(1-\alpha)]\,dy +$$

$$+\frac{d}{d\alpha}\int\limits_0^1 \left[\mu_2^{-1}(y)\right]^2 [g_2(y)\alpha + h_2(y)(1-\alpha)]\,dy =$$

$$= \int\limits_0^1 \left[\mu_1^{-1}(y)\right]^2 [g_1(y) - h_1(y)]\,dy + \int\limits_0^1 \left[\mu_2^{-1}(y)\right]^2 [g_2(y) - h_2(y)]\,dy.$$

Since $g_1(y) - h_1(y) = h_2(y) - g_2(y)$, we get at last

$$\left.\frac{d}{d\alpha}\sigma^2[\xi^*]\right|_{\alpha=0} = \int\limits_0^1 \left(\left[\mu_1^{-1}(y)\right]^2 - \left[\mu_2^{-1}(y)\right]^2\right)[g_1(y) - h_1(y)]\,dy.$$

Analyze signs of integrands factors.

1) Let $g_1(y) > h_1(y)$, then $g_1(y) = 1$, i.e. $\mu_1^{-1}(y) > \mu_2^{-1}(y)$ by formula (4).

2) Let $g_1(y) < h_1(y)$, then $g_1(y) = 0$. i.e. $\mu_1^{-1}(y) < \mu_2^{-1}(y)$ by formula (4).

From this, one can make a conclusion that the integrand on $[0,1]$ is non-negative. In addition, by our assumption, there is a point in the range of $h_1(y)$ continuity such that $g_1(y) \neq h_1(y)$. It implies $\left.\frac{d}{d\alpha}\sigma^2[\xi^*]\right|_{\alpha=0} > 0$. It means, there is a point $\alpha > 0$ such that $\sigma^2[\xi^*] > \sigma^2[\xi]$, i.e. the assumption made is wrong. It proves the theorem in the whole.

Theorem 5. *Let a set $\{\xi_i\}$ of random values with maximal variance $\sigma^2[\xi_i] = \overline{\sigma^2}(\mu)$, $E[\xi_i] = 0$, be in a fuzzy interval F with the membership function μ. Then there is a random value $\xi^* = \{\xi_i\}$ such that*

$$h_1^*(y) = \begin{cases} 1, \left|\mu_1^{-1}(y)\right| > \left|\mu_2^{-1}(y)\right|, \\ 0, \left|\mu_1^{-1}(y)\right| < \left|\mu_2^{-1}(y)\right|, \\ \alpha, \left|\mu_1^{-1}(y)\right| = \left|\mu_2^{-1}(y)\right|, \end{cases} \qquad h_2^*(y) = 1 - h_1^*(y), \quad \alpha \in [0,1].$$

Proof. By theorem 4, the set $\{\xi_i\}$ includes a random value ξ such that

$$h_1(y) = \begin{cases} 1, \left|\mu_1^{-1}(y)\right| > \left|\mu_2^{-1}(y)\right|, \\ 0, \left|\mu_1^{-1}(y)\right| < \left|\mu_2^{-1}(y)\right|, \end{cases} \qquad h_2(y) = 1 - h_1(y).$$

Denote $A = \left\{y \in [0,1] | \left|\mu_1^{-1}(y)\right| = \left|\mu_2^{-1}(y)\right|\right\}$. For the random value ξ^*, choose parameter $\alpha \in [0,1]$ as follows. Under the condition, $E[\xi] = E[\xi^*] = 0$, in addition, $h_i(y) = h_i^*(y)$ if $y \in \bar{A}$. Therefore,

$$E[\xi^*] - E[\xi] = \int_A \mu_1^{-1}(y)h_1(y)dy + \int_A \mu_2^{-1}(y)h_2(y)dy -$$

$$-\left(\int_A \mu_1^{-1}(y)h_1^*(y)dy + \int_A \mu_2^{-1}(y)h_2^*(y)dy\right).$$

For $y \in A$, $\mu_1^{-1}(y) = -\mu_2^{-1}(y)$, thus,

$$\int_A \mu_1^{-1}(y)h_1(y)dy + \int_A \mu_2^{-1}(y)h_2(y)dy =$$

$$= \int_A \mu_1^{-1}(y)\left(h_1(y) - h_2(y)\right)dy = \beta \int_A \mu_1^{-1}(y)dy,$$

where $\beta \in [0,1]$. The last equality is obtained with the help of mean-value theorem. By analogy,

$$\int_A \mu_1^{-1}(y)h_1^*(y)dy + \int_A \mu_2^{-1}(y)h_2^*(y)dy =$$

$$= \int_A \mu_1^{-1}(y)\left(h_1^*(y) - h_2^*(y)\right)dy = (2\alpha - 1)\int_A \mu_1^{-1}(y)dy.$$

Thus, $E[\xi] = E[\xi^*] = 0$ if $\beta = 2\alpha - 1$. Let us show that $\sigma^2[\xi^*] = \sigma^2[\xi]$ in this case. Actually,

$$\sigma^2[\xi] - \sigma^2[\xi^*] = E\left[\xi^2\right] - E\left[(\xi^*)^2\right] =$$

$$-\left(\int_A \left[\mu_1^{-1}(y)\right]^2 h_1^*(y)dy + \int_A \left[\mu_2^{-1}(y)\right]^2 h_2^*(y)dy\right) =$$

$$= \int_A \left[\mu_1^{-1}(y)\right]^2 dy - \int_A \left[\mu_2^{-1}(y)\right]^2 dy = 0.$$

The theorem is proved.

5 The practical calculation of maximal variance

Theorem 6. *Let the function $\mu_1^{-1}+\mu_2^{-1}$ be increasing. Then functions h_i for calculating the maximal variance have a form:*

$$h_1(y)=\begin{cases}1, y<\alpha,\\ 0, y>\alpha,\end{cases}\qquad h_2(y)=1-h_1(y),\quad y,\alpha\in[0,1]. \tag{5}$$

Proof. Let ξ be associated with a probability measure $P\in\Xi$ and $\sigma^2[\xi_i]=\overline{\sigma^2}(\mu)$. Suppose that $E[\xi]=m$, then by theorem 4,

$$h_1(y)=\begin{cases}1, \left|\mu_1^{-1}(y)-m\right|>\left|\mu_2^{-1}(y)-m\right|,\\ 0, \left|\mu_1^{-1}(y)-m\right|<\left|\mu_2^{-1}(y)-m\right|.\end{cases}$$

Thus, we need to solve the inequality, $\left|\mu_1^{-1}(y)-m\right|>\left|\mu_2^{-1}(y)-m\right|$. One can consider that $\mu_1^{-1}(y)-m<0$ and $\mu_2^{-1}(y)-m>0$ (see corollary of theorem 2). Therefore, the last inequality is transformed to a form:

$$\mu_1^{-1}(y)+\mu_2^{-1}(y)<2m.$$

Let the number $2m$ belong to the range of values of the function $\mu_1^{-1}+\mu_2^{-1}$, g be an inverse function to this median, then, since g is increasing function, we get that $y<g(2m)=\alpha$. The cases, where $2m$ does not belong to the range of median values, are also described by formula (5).

Corollaries of theorem 6. *Let we use notations of theorem 6. Then*

1) $h_1(y)=\begin{cases}1, \mu_1^{-1}(y)+\mu_2^{-1}(y)<0,\\ 0, \mu_1^{-1}(y)+\mu_2^{-1}(y)>0,\end{cases}$ $h_2(y)=1-h_1(y)$, *if* $E[\xi]=0$.

2) Let the function $\mu_1^{-1}+\mu_2^{-1}$ be increasing on $[0,1]$, *then there is a certain* $\alpha\in[0,1]$ *such that* $h_1(y)=\begin{cases}1, y>\alpha,\\ 0, y<\alpha,\end{cases}$ $h_2(y)=1-h_1(y), y,\alpha\in[0,1]$.

Theorem 7. *Let ξ belong to a fuzzy interval F with a membership function μ and* $\sigma^2[\xi]=\overline{\sigma^2}(\mu)$. *Then* $E[\xi]\in\left\{\mu_1^{-1}(y)+\mu_2^{-1}(y)\,|y\in[0,1]\right\}$.

Proof. Assume that the condition of the theorem is not satisfied. Then, using corollary 1 of theorem 6, we get that either $h_1(y)\equiv 1$ or $h_2(y)\equiv 1$. For the sake of determinacy, let $E[\xi]=0$. 1) Let $h_1(y)\equiv 1$, then $E[\xi]<b$. It means that $P[0,c)>0$, but this contradicts to the corollary of theorem 2. 2) Let $h_2(y)\equiv 1$, then $E[\xi]>c$. It means that $P(b,0]>0$, but this contradicts to the corollary of theorem 2. The contradictions found prove the truth of the theorem.

Corollary. *Let a fuzzy interval F be symmetric, i.e.* $\mu_1^{-1}(y)+\mu_2^{-1}(y)=const$, *and, for the sake of determinacy,* $const=0$. *Then* $\overline{\sigma^2}(\mu)=\int\limits_0^1\left[\mu_1^{-1}(y)\right]^2dy$.

Proof. According to theorem 7, for ξ with maximal variance, the value $E[\xi]$

belongs to the range of $\mu_1^{-1}+\mu_2^{-1}$ values, i.e. $E[\xi]=const=0$. Therefore,

$$\sigma^2[\xi]=\int_0^1\left[\mu_1^{-1}(y)\right]^2 h_1(y)dy+\int_0^1\left[\mu_2^{-1}(y)\right]^2 h_2(y)dy.$$

Since $\left[\mu_1^{-1}(y)\right]^2=\left[\mu_2^{-1}(y)\right]^2$ and $h_1(y)+h_2(y)=1$, we get $\sigma^2[\xi]=\int_0^1\left[\mu_1^{-1}(y)\right]^2 dy$. The corollary is proved.

Theorem 8. *Let functions $h_i(y)$ in the formula (3) for calculating maximal variance have a form:*

$$h_1(y)=\begin{cases}1, y<\alpha,\\ 0, y>\alpha,\end{cases}\qquad h_2(y)=1-h_1(y),\quad y,\alpha\in[0,1]. \tag{6}$$

Then α can be found from the equality:

$$\frac{\mu_1^{-1}(\alpha)+\mu_2^{-1}(\alpha)}{2}+\int_0^\alpha\left[\mu_2^{-1}(y)-\mu_1^{-1}(y)\right]dy-\frac{1}{2}\int_0^1\left[\mu_2^{-1}(y)-\mu_1^{-1}(y)\right]dy=0, \tag{7}$$

if the coordinate system is chosen such that $\int_0^1\left[\mu_2^{-1}(y)+\mu_1^{-1}(y)\right]dy=0$. There is a unique solution if the function $\mu_1^{-1}+\mu_2^{-1}$ is increasing.

Proof. Let the functions h_i have a form (6). Then

$$\overline{\sigma^2}[\mu]=E\left[\xi^2\right]-E^2[\xi]=\int_0^\alpha\left[\mu_1^{-1}(y)\right]^2 h_1(y)dy+$$

$$+\int_\alpha^1\left[\mu_2^{-1}(y)\right]^2 h_2(y)dy-\left(\int_0^\alpha\mu_1^{-1}(y)dy+\int_\alpha^1\mu_2^{-1}(y)dy\right)^2.$$

Taking derivative w.r.t. α and using the necessity condition for extremum, we get the equality:

$$\left[\mu_1^{-1}(\alpha)\right]^2-\left[\mu_2^{-1}(\alpha)\right]^2-2\left[\mu_1^{-1}(\alpha)-\mu_2^{-1}(\alpha)\right]\left[\int_0^\alpha\mu_1^{-1}(y)dy+\int_\alpha^1\mu_2^{-1}(y)dy\right]=0.$$

Since $\alpha<1$ and $\mu_1^{-1}(\alpha)-\mu_2^{-1}(\alpha)<0$, then we can reduce this factor. As result,

$$\mu_1^{-1}(\alpha)+\mu_2^{-1}(\alpha)-2\left[\int_0^\alpha\mu_1^{-1}(y)dy+\int_\alpha^1\mu_2^{-1}(y)dy\right]=0. \tag{8}$$

Transform the expression:

$$2\left[\int\limits_0^\alpha \mu_1^{-1}(y)dy+\int\limits_\alpha^1 \mu_2^{-1}(y)dy\right]=\int\limits_0^\alpha \mu_1^{-1}(y)dy-\int\limits_0^\alpha \mu_2^{-1}(y)dy+\int\limits_0^1 \mu_2^{-1}(y)dy$$

$$+\int\limits_0^1 \mu_1^{-1}(y)dy-\int\limits_\alpha^1 \mu_1^{-1}(y)dy++\int\limits_\alpha^1 \mu_2^{-1}(y)dy=\left(\int\limits_\alpha^1 \left[\mu_2^{-1}(y)-\mu_1^{-1}(y)\right]dy-\right.$$

$$\left.-\int\limits_0^\alpha \left[\mu_2^{-1}(y)-\mu_1^{-1}(y)\right]dy\right)+\int\limits_0^1 \left[\mu_2^{-1}(y)+\mu_1^{-1}(y)\right]dy.$$

By the supposition, $\int\limits_0^1 \left[\mu_2^{-1}(y)+\mu_1^{-1}(y)\right]dy=0$, in addition, the first item in the last expression can be transformed to a form:

$$\int\limits_0^1 \left[\mu_2^{-1}(y)-\mu_1^{-1}(y)\right]dy-2\int\limits_0^\alpha \left[\mu_2^{-1}(y)-\mu_1^{-1}(y)\right]dy.$$

Taking this into account, the equality (8) is written as follows:

$$\mu_1^{-1}(\alpha)+\mu_2^{-1}(\alpha)+2\int\limits_0^\alpha \left[\mu_2^{-1}(y)-\mu_1^{-1}(y)\right]dy--\int\limits_0^1 \left[\mu_2^{-1}(y)-\mu_1^{-1}(y)\right]dy=0, \tag{9}$$

i.e. we really prove the truth of equation (7).

Denote the left part of equation (9) by $f(\alpha)$. Let the function $\mu_1^{-1}+\mu_2^{-1}$ be increasing, then, since by supposition $\int\limits_0^1 \left[\mu_2^{-1}(y)+\mu_1^{-1}(y)\right]dy=0$, it is obvious that $\mu_1^{-1}(0)+\mu_2^{-1}(0)\leq 0$ and $\mu_1^{-1}(1)+\mu_2^{-1}(1)\geq 0$. Taking this into our account, analyze signs of $f(\alpha)$ at the ends of $[0,1]$:

$$f(0)=\mu_1^{-1}(0)+\mu_2^{-1}(0)-S,$$

$$f(1)=\mu_1^{-1}(1)+\mu_2^{-1}(1)+S,$$

where S is an area of the fuzzy interval. Therefore, $f(0)<0$ and $f(1)>0$, i.e. the equation has at least one root. Analyze the sign of

$$f'(\alpha)=\frac{d}{d\alpha}\left[\mu_1^{-1}(\alpha)+\mu_2^{-1}(\alpha)\right]+2\left[\mu_2^{-1}(\alpha)-\mu_1^{-1}(\alpha)\right].$$

It is obvious, that $f'(\alpha)>0$ for $\alpha\in[0,1]$. Thus, the equality $f(\alpha)=0$ has only one root, and this root is a point of maximum (you should remind, that for obtaining equality (8), we reduce the expression by the negative factor $\left(\mu_1^{-1}(\alpha)-\mu_2^{-1}(\alpha)\right)$). Thus, the theorem is proved in the whole.

Remarks.

1) Theorem 8 is easily generalized for the case, where

$$h_1(y)=\begin{cases}1, y>\alpha,\\ 0, y<\alpha,\end{cases} \quad h_2(y)=1-h_1(y),\ \ y,\alpha\in[0,1],$$

and the function $\mu_1^{-1}+\mu_2^{-1}$ is decreasing. In this case α can be found from the equation:

$$\frac{\mu_1^{-1}(\alpha)+\mu_2^{-1}(\alpha)}{2}-\int_0^{\alpha}\left[\mu_2^{-1}(y)-\mu_1^{-1}(y)\right]dy+\frac{1}{2}\int_0^{1}\left[\mu_2^{-1}(y)-\mu_1^{-1}(y)\right]dy=0.$$

We also suppose that $\int_0^1\left[\mu_2^{-1}(y)+\mu_1^{-1}(y)\right]dy=0$.

2) The equation (7) has a geometrical interpretation (fig. 2).

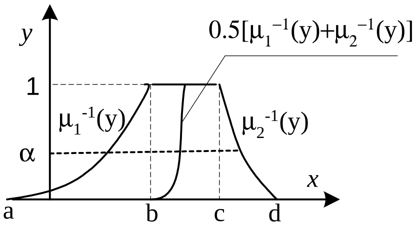


Figure 2: Fuzzy interval: inverse functions

a) $0.5(\mu_1^{-1}+\mu_2^{-1})$ is the median of the fuzzy interval;

b) $\int_0^1\left[\mu_2^{-1}(y)-\mu_1^{-1}(y)\right]dy$ is the area of the fuzzy interval;

c) $\int_0^{\alpha}\left[\mu_2^{-1}(y)-\mu_1^{-1}(y)\right]dy$ is the area of the part of the fuzzy interval that is below of α level ;

d) $\mu_2^{-1}(\alpha)-\mu_1^{-1}(\alpha)$ is the length of the level line $y=\alpha$ for the fuzzy interval.

3) Introduce into consideration functions

$$m(y)=\frac{\mu_1^{-1}(y)+\mu_2^{-1}(y)}{2},\ w(y)=\frac{\mu_2^{-1}(y)-\mu_1^{-1}(y)}{2},\ F(\alpha)=\int_0^{\alpha}\left[w(y)+\frac{m'(y)}{2}\right]dy.$$

Then equation (7) can be transformed to a form:

$$2F(\alpha)-F(1)=0. \tag{7*}$$

Example. Consider, how to calculate the maximal variance for the fuzzy interval having a form of trapezium (fig. 3). In this case, the functions *m*, *w* are linear.

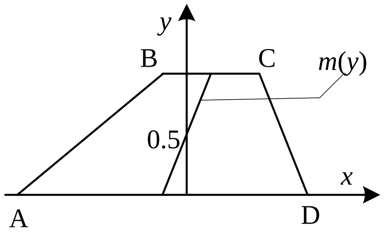


Figure 3: Fuzzy interval with a form of trapezium

We assume that *m* is increasing and $\int\limits_0^1 m(y)dy = 0$. In this case, one can easily show that $m(y) = k(y-0.5)$, where $k > 0$. The function *w* is expressed through lengths of the trapezium sides $l_1 = |BC|$ and $l_2 = |AD|$. Since $w(0) = 0.5l_2$ and $w(1) = 0.5l_1$, then $w(y) = 0.5[l_2 - (l_2 - l_1)y]$. The parameter α can be found from equation (7*). Then $F(\alpha) = 0.5(k+l_2)\alpha - 0.25(l_2 - l_1)\alpha^2$, and we need to solve the equation:

$$(l_2 - l_1)\alpha^2 - 2(k+l_2)\alpha + \frac{l_1 + l_2 + 2k}{2} = 0.$$

Solving it, we get

$$\alpha = \frac{(k+l_2) - \sqrt{0.5\left[(k+l_2)^2 + (k+l_1)^2\right]}}{l_2 - l_1},$$

in addition, $\alpha \in [0,1]$. The precise value of $\overline{\sigma^2}(\mu)$ can be calculated by formula (3). Namely, according to the form of h_1, h_2 we can write

$$\overline{\sigma^2}(\mu) = \int\limits_0^\alpha \left[\mu_1^{-1}(y)\right]^2 dy + \int\limits_\alpha^1 \left[\mu_2^{-1}(y)\right]^2 dy - \left(\int\limits_0^\alpha \mu_1^{-1}(y)dy + \int\limits_\alpha^1 \mu_2^{-1}(y)dy\right)^2,$$

where

$$\mu_1^{-1}(y) = [k + 0.5(l_2 - l_1)](y - 0.5) - 0.25(l_2 + l_1),$$

$$\mu_2^{-1}(y) = [k - 0.5(l_2 - l_1)](y - 0.5) + 0.25(l_2 + l_1).$$

Let $l_1 = 1,\ l_2 = 3,\ k = 0.5$, then $\alpha = 0.404$, $\overline{\sigma^2}(\mu) = 1.342$. Fig. 4 shows this fuzzy interval, and probability distribution function F of the extreme random value ξ, being in the fuzzy interval, for which $\overline{\sigma^2}(\mu) = \sigma^2[\xi]$.

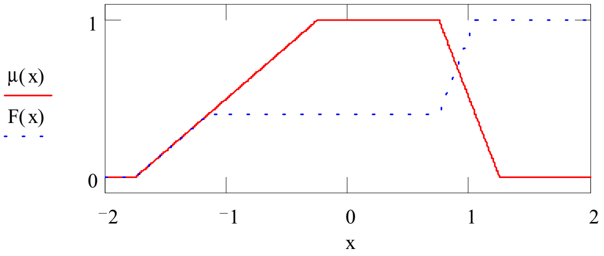


Figure 4: Numerical example

References

[1] P. Walley. Statistical reasoning with imprecise probabilities. *Chapman and Hall, London*, 1991.

[2] D. Dubois, and H. Prade. Possibility theory. *Plenum Press, New-York*, 1988.

[3] D. Dubois, and H. Prade. The mean value of fuzzy number. In *Fuzzy sets and systems*, 24: 279-300, 1987.

[4] D. Dubois, and H. Prade. When upper probabilities are possibility measures. In *Fuzzy sets and systems*, 49: 65-74, 1992.

[5] A.G. Bronevich, and A.N. Karkishchenko. Statistical classes and fuzzy set theoretical classification of possibility distributions. In *Statistical modelling, analysis and management of fuzzy data. Heidelberg. New-York: Physica-Verl.*, 173-195, 2002.

Andrew G. Bronevich is with the Laboratory of Mathetimatical Problems in Artificial Intelligence, Taganrog State University of Radio-Engineering, Nekrasovskiy street, 44, Russia, 347928. E-mail: brone@mail.ru

Inter-personal Communication of Precise and Imprecise Subjective Probabilities[*]

D. V. BUDESCU
University of Illinois at Urbana-Champaign, USA

T. M. KARELITZ
University of Illinois at Urbana-Champaign, USA

Abstract

We analyze communication of uncertainty among individuals as a function of the parties' preference for modes of communication. We assume that different individuals may prefer precise *Numerical* probabilities, *Ranges* of probabilities or *Verbal* descriptions of probabilities, and consider all possible pairings of communicators and receivers under this classification. We propose a general criterion of optimal conversion among the various modalities, describe several instantiations tailored to fit the special features of the various modalities, and illustrate the efficacy of the proposed procedures with empirical results from several experiments.

Keywords

subjective probability, judgment, inter-personal translation, verbal probabilities

1 Introduction

Consider a situation where two individuals communicate about stochastic events. The two are equally interested and motivated to communicate as efficiently and precisely as possible. This paper is concerned with procedures that can be employed to address the individuals' different preferences for modality of communicating probabilistic opinions. Although many decision analysts and orthodox Bayesians consider precise numerical probabilities to be *the* language of uncertainty, many people (layman and experts, alike) prefer to use probability phrases (e.g. review by Budescu and Wallsten [6]) or other imprecise variants of probability. In this paper we propose ways to achieve the highest possible level of accuracy in communication while accommodating these individual preferences.

[*]This work has been supported by the US National Science Foundation under Award No. SES 9975360. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the US National Science Foundation.

1.1 Reasons for preferences of specific communication modes

Spontaneous preferences for one particular mode may be due to several factors:

The perceived *nature of the uncertainty to be communicated*–Budescu and Wallsten [6] have speculated, and Olson and Budescu [14] have documented empirically that most individuals prefer to use precise numerical estimates to communicate uncertainty about repeated events with aleatory uncertainty, but tend to use more imprecise methods when communicating the probabilities of unique events with epistemic uncertainty.

The perceived *strength of the available information*–The responses to the survey conducted by Wallsten, Budescu, Zwick, and Kemp [18], indicate that people would gravitate towards more precise modes of communication, if they perceive the available information to be firmer, reliable and valid.

The person’s *role in the communication*–In the same survey Wallsten et al. [18] have found that most people prefer to use imprecise terms when they communicate to others, but prefer others to communicate to them in precise terms, if possible (see also, Brun and Teigen [2] and Erev and Cohen [8]).

In addition to these systematic factors, preferences may be due to plain *individual differences* that reflect one’s lifetime experiences in dealing with, and communicating, uncertainties.

1.2 The problem

The need to communicate probabilities arises in a variety of situations. A common case is when both individuals have prior opinions, have access to some relevant (possibly overlapping) information, and wish to exchange information to further refine their respective estimates. In this *symmetric* case the designation of communicator and receiver is arbitrary, as the two individuals can act in both capacities. For example, think of two friends who talk about the chances of their favorite team to win a game. The other prototypical case involves *asymmetric* communication: only one individual, the Forecaster (**F** for short), has access to, or possesses the necessary expertise to make sense of, the relevant information for the probability estimation. The second individual, the Decision Maker (or **DM**) needs to make a choice or decision on the basis of the F’s estimate, and without the benefit of his, or her, own probability assessment. For example, think of an investor (the DM) who gets from his, or her, favorite financial advisor estimates of the likelihood that certain investment policies will succeed.

The two situations are similar in many respects but the former is more complex because a complete analysis should take into account the processes that govern the combination of one’s own opinions with estimates obtained form others (Yaniv and Kleinberger [19]). To simplify the analysis, we will focus on the second case. In the same spirit, we will not consider the case where one needs to aggregate multiple forecasts from various sources (Budescu, Rantilla, Yu and

Karelitz [5], Wallsten, Budescu and Tsao [17]).

To summarize, we analyze an asymmetric dyadic communication situation where one F and one DM share a common interest in optimizing communication, but they may have different preferences for modality of communicating probabilistic opinions.

1.3 A typology of communication preferences in a dyad

We distinguish between three modes of communication: precise (point) **N**umerical probability estimates (e.g., 0.45), precise **R**anges of numerical values (e.g., 0.3 - 0.55), and **V**erbal phrases (e.g., good chance). Ranges with precise end points exclude implicit vague ranges such as ”in the forties” or ”at least 0.80”, but such expressions can be analyzed as verbal terms.

The three modes can be ranked from the most precise (N) to the most vague (V). In fact, the more precise modes can be represented as special cases of the more vague modes: clearly an N is an R where the lower and upper limit coincide, and we will show later how N and R can be viewed as special cases of V under a particular representation of the probability phrases. This typology implies 9 distinct dyadic patterns of dyadic preferences for modes of communication that will be denoted by ordered pairs, where the first character in the pair refers to the F’s preference.

2 The translation process

The problem we wish to address is deceptively simple – How to *best* convert a judgment originally expressed in the F’s favorite response mode (N, R or V), to an estimate in the DM’s favorite mode (N, R or V).

The criterion of optimality is the level of (dis)similarity between the F’s judgments translated into the DM’s favorite mode, and the DM’s spontaneous (and independent) judgments of the same events in his, or her, favorite mode. For example, assume that the F prefers numbers and the DM prefers verbal terms (i.e., an [N,V] dyad). If both had the same prior probability distribution and could access the same information pertaining to the target event, X_i, their spontaneous and independent judgments would be $n_F(X_i)$, and $v_{DM}(X_i)$, respectively.

Any mapping of the F’s spontaneous judgment into the DM’s favorite communication mode is a *translation*. For example, $v_{DM}[n_F(X_i)]$ is the verbal translation of the F’s original numerical judgments. An *optimal translation* is one that maximizes the similarity between the translation of the F’s term into the DM’s favorite mode, and the DM’s spontaneous judgment of the target event (assuming he/she has the same priors and could access the same information).

Note that (dis)similarity is measured in the scale of the target modality (i.e., the one that is favored by the DM), so it always relies on commeasurable units or entities. On the other hand, these entities vary as a function of the DM's favorite modality. Next we describe some sensible choices for the dissimilarity metrics. Our goal in this paper is to provide a general framework for the translation process and illustrate the feasibility of the approach. We make no claim of optimality, or uniqueness on behalf of these choices, and realize that other metrics could be used in this context.

Dissimilarity between two numbers, n_{DM} and n_F, is defined as the distance between them:

$$DS_n\{n_{DM}, n_F\} = |n_{DM} - n_F|. \tag{1}$$

Dissimilarity between two ranges, r_{DM} and r_F, is a function of their respective lengths, and their overlap. Consider two ranges, $r1$ (ranging from $l1$ to $u1$) and $r2$ (from $l2$ to $u2$). The width of the range over which the two overlap is $OV_{12} = Max\{0, [Min(u1, u2) - Max(l1, l2)]\}$, and the joint range of values they span is $JR_{12} = [Max(u1, u2) - Min(l1, l2)]$. We define the dissimilarity between the two ranges as:

$$DS_r\{r_{DM}, r_F\} = JR_{DM,F} - OV_{DM,F}. \tag{2}$$

This measure is zero if, and only if, the two ranges coincide. For any pair of ranges, r_{DM} and r_F, the index is maximal when they are disjoint.

Dissimilarity between two verbal terms, v_{DM} and v_F, is defined in the context of a particular representation of such phrases. Wallsten, Budescu, Rapoport, Zwick, and Forsyth [16] suggested that probability phrases are fuzzy concepts and proposed using Membership Functions (MFs) over the $[0, 1]$ probability interval to represent their vague meanings (see Zadeh [20]). A phrase's MF assigns to each probability a real number that represents the (non-negative) degree of its membership in the concept defined by the phrase. These values are scaled between 0 and 1 (Norwich and Turksen [13]), such that memberships of 0 denote probabilities that are absolutely not in the concept and memberships of 1 denote elements that are perfect exemplars of the concept. All other positive values represent intermediate degrees of membership. MFs can be estimated directly (non-parametrically) based on the participants' direct or indirect judgments (see Budescu and Wallsten [6], Wallsten et al. [16]). Alternatively, one can fit MF using specific families of functions, such as polynomials (Budescu, Karelitz and Wallsten [4]), or trapezoidal functions.

Let $\mu_{v_{DM}}(p)$ and $\mu_{v_F}(p)$ be the MFs representing the two words being compared. The similarity between the two words should reflect the closeness between their respective MFs. There are many possible single-valued indices of closeness between the two functions (see review by Zwick, Carlstein, and Budescu [21]), and we will only list two of them here (these are not necessarily monotonically related). The first measure is the total absolute distance between the two functions.

Formally, we can write[1]:

$$DS_{v_\mu}\{v_{DM}, v_F\} = \int_{p=0}^{1} |\mu_{v_{DM}}(p) - \mu_{v_F}(p)|dp. \tag{3}$$

The second index is the distance between the peaks of the two functions. Assume that both $\mu_{v_{DM}}(p)$ and $\mu_{v_F}(p)$ are single peaked (see Budescu and Wallsten [6] on this point). Let $\pi(v)$ be the probability (or the center of the range of probabilities) at which the function $\mu_v(p)$ reaches its maximal value. We define, a second measure of dissimilarity as:

$$DS_{v_\pi}\{v_{DM}, v_F\} = |\pi(v_{DM}) - \pi(v_F)|. \tag{4}$$

2.1 General comments on the measures of dissimilarity

The various measures may appear at first glance to be unrelated and, somewhat arbitrary, so a few comments and clarifications are in order. First, we should point out that all the dissimilarity indices are *distances*. In all cases they assign to every pair of (N,R or V) judgments a non-negative real number ($DS = 0$ only if the two members of the pair are identical). The measures are symmetric, satisfy the triangle inequality and induce a weak order over all pairs.

One could invoke other metrics for these comparisons. A particularly elegant approach would be to use the same metric for all modalities. Technically, this is feasible since numbers can be represented by point MFs (membership of 0 everywhere, and 1 for the chosen number) and ranges can be represented by flat MFs (membership of 0 everywhere outside, and 1 everywhere within the chosen range), and treated in the same fashion as the MFs obtained for verbal terms. However, we believe that the metrics identified above are better suited for our purposes because they are more in line with the particular level of (im)precision implied by the three modalities.

The last comment is subtler. Our definition of similarity relies on a counterfactual scenario that gives rise to a hypothetical entity - the DM's spontaneous judgment of the target event if he, or she, had the same prior probability distribution and could access the same information that was used by the F as a basis for his/her judgment. Strictly speaking, this definition is meaningful only in those cases where it makes sense to assume that a person's *judgment depends only on the specific* information presented. This implies that the relevant information is unambiguous and does not lend itself to different (subjective) interpretations. In other words, the observed variability among probabilities assigned to a target event by different individuals can be attributed solely to different response styles and/or random factors within the judges. This formulation makes perfect sense

[1] In most empirical applications the MFs are approximated by a set of n points over $[0,1]$, so a discrete version of this measure can be used to approximate it.

for repeatable and exchangeable events, but not for unique events where subjective probabilities rely on internal epistemic uncertainty that can vary systematically across individuals. (Ariely, Au, Bender, Budescu, Dietz, Gu, Wallsten and Zauberman [1] and Wallsten, Budescu, Erev and Diederich [15], discuss various facets of this key distinction).

For example, it is quite unlikely that if we were to present anti-smoking activists and tobacco lobbyists with the results of a new study on the effects of second-hand smoking, they would agree in their estimation of the probabilities that second-hand smoking has serious public health consequences. The differences between their estimates would reflect (a) their different prior probabilities, and (b) their differential assessment of the quality, reliability and validity of the new data. Clearly, no translation method can be expected to reconcile disagreements of this type. Despite these irreconcilable differences in their opinions, we can still take advantage of optimal translation schemes derived for various pairs of communicators based on their judgments of a standard set of exchangeable events. When these translation methods are applied they can reduce the effect of other sources of variability among the participants and provide *the most accurate representation of the F’s assessment in the DM’s favorite communication mode*, where accuracy is measured by one of the dissimilarity metrics discussed above.

2.2 Methods of translation

We return now to our original question: how to *best* convert a judgment originally expressed in the F’s favorite response mode (N, R or V), to an estimate in the DM’s favorite mode (N, R or V). Before we discuss translation schemes for each of the 9 cases, it makes sense to classify them into three distinct groups:

Common modalities - In three cases ([N,N], [R,R] and [V,V]) both individuals share a common preference for mode of communication, so there is no need to worry about differential precision. Conversions may be employed to account for inter-personal differences in the way the relevant terms are chosen and used.

Resolving vagueness - In three cases ([R,N], [V,R] and [V,N]) the DM prefers a more precise mode of communication than the F. Thus, the challenge is to find a translation that resolves the vagueness implicit in the F’s judgment to achieve the higher level of precision required by the DM.

Imputing vagueness - In the other three cases ([N,R], [N,V] and [R,V]) the DM prefers a more vague mode of communication than the F. Thus, the challenge is to find a translation that replaces the precision implicit in the F’s judgment to reflect the higher level of vagueness expected by the DM.

We will discuss the three classes separately. In each case we describe and justify a translation method designed to optimize our stated goal and, when appropriate, we review and discuss relevant results from several empirical studies that are described in the next section.

2.3 The data

Over the last two years we have conducted four experiments designed to test the efficacy and accuracy of various translation methods of probability phrases (V). The studies vary in many specific details (Budescu and Karelitz [3] and Karelitz and Budescu [11]) but they share a set of common features that allow us to analyze some of their results jointly. The focus on the N and V responses is neither accidental, nor arbitrary. Subjects rarely communicate their probabilities by means of ranges even when offered the opportunity (e.g. references in Budescu and Wallsten [6]). For example, in one of the studies analyzed below when this option was present, it was used in less that 7.5% of the cases. Thus, we will not present any empirical results concerning translations involving Rs.

The four studies involved a total of 128 individuals (all students at the University of Illinois in Urbana Champaign, and most of them native English speakers[2]). All the experiments were computer controlled, and included the following three tasks: (1) Selection of a personal verbal probability lexicon including 5-11 phrases (In a few cases some, or all, the phrases were selected by the experimenters based on previous research); (2) Elicitation of MFs for all the phrases; and (3) Numerical and verbal estimation of probabilities of a common set of events.

Subjects created their lists by selecting combinations of words and semantic operators (modifiers, intensifiers, etc.) from two lists, or typing in phrases. They were instructed to select phrases that span the whole probability range, and they tend to use regularly. Membership functions were elicited using a method validated by Budescu et al. [4]. Each phrase was presented with a set of eleven probabilities ranging from 0 to 1 in increments of 0.1. The subjects judged the degree to which the target phrase captured the intended meaning of each of the eleven numerical probabilities by using a bounded scale, anchored by the terms '*not at all*' and '*absolutely*'. In the last task, the participants saw a series of circular, partially shaded, targets. Their task was to assess the likelihood that a dart, aimed at the center of the target, would hit the shaded area. The shaded areas varied from one trial to another and covered the full (0,1) range. On separate presentations these probabilities were judged numerically (by selecting one value from a list of 21 probabilities, ranging from 0 to 1 in increments of 0.05), or verbally by selecting (in some cases up to four) phrases from their lexicons.

2.4 Common modalities

[**N**,**N**] This is the "gold standard" case of Bayesian decision analysis. Presumably numbers are universal and everyone understands, interprets and uses them in identical fashion. Therefore, no transformation is required. There is, however, evidence that people's mapping of their internal feelings of uncertainty into

[2] One of the studies was concerned with translation of probability phrases across languages and we recruited native speakers of French, German, Spanish, Russian and Turkish.

numbers is imperfect. In particular, most people over-(under-) estimate low(high) probabilities (e.g. references in Erev, Wallsten and Budescu [9]), and it is conceivable that there are systematic differences in the degree to which individuals tend to avoid (or favor) the extreme values. In principle, one could quantify this tendency and apply appropriate *stretching (or contracting) transformations*. To illustrate this point consider multiple judges $(1,2,\ldots,j,j',\ldots,J)$ who judge a set of stochastic events $(1,2,\ldots,i,\ldots,I)$. Assume that: (1) All judges have access to the same amount of information, implying that differences in their judgments are due only to (a) differences in their use of the response scales and (b) random components. (2) All judges spontaneously recognize events that are impossible (probability = 0), certain (probability = 1), and as likely as not (probability = 0.5). (3) Assume an "ideal judge" who is perfectly calibrated (no biases) and accurate (no random component). Thus his/her judgments, $p_1, p_2, \ldots p_i$, coincide with the events' "objective probabilities".

The probability assigned by judge j to event i is denoted by p_{ij}, and can be expressed as a function of the objective probability, p_i, his/her bias parameter, α_j, and the random component, e_{ij} which we assume is distributed with $\mu_e = 0$ and (finite) σ_e. We use a variation of Karmarkar's [10] model, that assumes that the logit of the judged probabilities is a linear function of the logit of their objective counterparts:

$$Log\left[\frac{p_{ij}}{(1-p_{ij})}\right] = \alpha_j \cdot Log\left[\frac{p_i}{(1-p_i)}\right] + e_{ij} \tag{5}$$

Individual differences between judges are captured by the parameter α_j, which is bounded from below by 0 (when all events are assigned a probability of 0.5). An unbiased judge should have an α_j of 1, but we expect that most individuals would have parameters between 0 and 1 that are consistent with the regressive model described above. We used a least-squares procedure to estimate the individual parameter, α_j, in model 5. The model fits the data well for almost all subjects (median $R^2 = 0.98$, median $MSE = 0.13$). The distribution of the individual parameters matches our expectations: 64 values (50%) are between 0.55 and 0.98, 45 participants (35.2%) are almost perfectly calibrated ($0.99 \leq \alpha_j \leq 1.01$), and only 19 individuals (14.8%) have parameters values above 1. To verify that these differences reflect systematic individual differences rather than pure random error, we performed two additional analyses: (a) we compared these results with a model where the parameter, α_j, was constrained to be 1 (thus the model includes only random error). A comparison of the two models in terms of R^2_{adj} favors slightly the fitted model. The modal difference (34% of the cases) is 0, but there is a clear majority (43% vs. 23%) of cases where the fitted model fits better (mean difference in fit = 0.02), even after we account for its extra parameter. Significance tests comparing the fit of the two models (separately for each subject) revealed that this differences was significant at the traditional 0.05 level, for 25.2% of the subjects. (b) We re-analyzed two of the studies in which all subjects judged all the displays twice, so we could obtain two estimates of the parameter, α_j, for each person. In both studies (involving a total of 55 subjects) the between-subjects variance component was considerably larger than the within-subject component (in fact the within-subject component was not significantly greater then 0).

In principle, one could convert numerical estimates from one person to another in an optimal fashion by applying simple stretching (contracting) transformation based on the estimates of the individual parameters, α_j, $\alpha j'$.

[**R**,**R**] The use of precise ranges instead of simple point estimates reflects one's perceived level of imprecision in his or her estimate of the probability of the target event. Clearly, the arguments invoked in the [N,N] case regarding the nature of the numbers, apply here as well. This would suggest that no transformations are indicated. It is conceivable, however, that there are systematic differences in the degree of imprecision perceived by different individuals and this would induce systematic differences in the widths of their ranges. One could quantify this tendency and apply appropriate *imprecision equating* transformations.

[**V**,**V**] This situation is, probably, the most interesting and it has been the focus of much of our recent research. This case is qualitatively different from the previous two for several reasons. There is a large literature indicating that (a) spontaneously, people tend to use highly different and diverse lexicons, and (b) the numerical meanings (as well as other forms of representation) associated with these words vary dramatically across people (e.g. review in Budescu and Wallsten [6]). Thus, one cannot assume that everyone is equally comfortable with, or interprets identically terms such as "*likely*", "*poor chance*", etc. For this case we advocate the following multi-stage procedure that is sensitive to these empirical findings: (a) each participant selects his/her own subjective lexicon; (b) MFs are elicited for all the terms in the list; (c) the MFs of the words selected by the F and the DM are placed on a common probability scale and are matched according to the criterion of choice (DS_{v_μ} or DS_{v_p}). Occasionally this procedure does not yield a unique solution, i.e., one of the F's words can be translated equally well into several of the DM's words. Of course, all these words are equally valid translations of the F's judgment. If practical considerations prevent one-to-many translations, one of them can be selected randomly (or by some other sensible tie-breaking procedure).

We have done quite a lot of empirical work documenting the efficacy of this approach (Karelitz and Budescu [11], Karelitz, Dhami, Budescu, and Wallsten [12]). In each of our studies we compared the level of agreement in assignment of verbal phrases to the same events among numerous pairs of distinct individuals. We hypothesized that the lowest level of agreement would be observed with spontaneous (un-aided), verbal discourse, and the best level of agreement would be found in the case of numerical communication. Most importantly, we expect that communication with converted phrases would be superior to un-aided verbal communication, and closer in quality to the numerical case. To quantify the level of inter-personal agreement we defined two indices of co-assignment. We use two measures because some of the events were judged more than once and yielded different responses from the subjects. Both measures range from 0 to 1 (higher values indicate stronger agreement), and can be interpreted as measures of the accuracy of inter-personal communication of imprecise opinions.

PIA- **P**roportion of **I**dentical **A**ssignments- the proportion of *comparisons* where both participants assigned the **same** phrase to a given event.

PMA- **P**roportion of **M**inimal **A**greement - the proportion of *stimuli* for which both participants assigned **at least one** common phrase to a given event.

In the interest of brevity we only report results based on PIA, which is a more stringent measure than PMA (PIA $\leq$ PMA) because it weighs the agreement by the number of comparisons made (The PMA results are very similar in a qualitative sense). Table 1 summarizes the results of 4 studies (details in Karelitz and Budescu [11]). Each cell presents the mean (and SD) PIA in the various modes, and across all pairs of subjects analyzed.

Table 1: Summary of agreement indices from 4 studies

Study	No. of pairs	Translation Criterion			
		VJ: Unaided Verbal Judgments	DS_{v_μ} (Eq. 3)	DS_{v_π} (Eq. 4)	NJ: Unaided Numerical Judgments
1	306	0.05 (0.03)	0.23 (0.12)	0.22 (0.10)	0.29 (0.07)
2	90	0.04 (0.04)	0.22 (0.11)	0.19 (0.09)	0.36 (0.09)
3	86	0.04 (0.07)	0.35 (0.16)	0.35 (0.16)	0.40 (0.15)
4*	509	0.06 (0.09)	0.34 (0.15)	0.35 (0.14)	0.40 (0.13)

* Experiment 4 involves translations of words across various languages. VJ is based of the subjects’ spontaneous translation of words from their native languages to English.

The results clearly support our predictions: unaided VJ had the lowest values for both indices in all the studies and NJ had the largest values. The two translation criteria clearly outperformed the unaided verbal communication[3].

2.5 Resolving vagueness

[**R**,**N**] In principle, any sensible person should be able to infer a single N value from his/her partner range without invoking any translation scheme. The individual differences discussed in the [N,N] and [R,R] cases apply here as well. In principle, one could improve the quality of communication by (a) inferring the F’s best guess (presumably, the center of the reported range) and, if necessary, (b) applying the appropriate *stretching (or contracting) transformation*.

[**V**,**N**] Recall that every word in the F’s lexicon has a (single-peaked) MF defined over the [0,1] interval that describes the degree to which the various prob-

[3] Dhami and Wallsten [7] and Karelitz and Budescu [11] report similar results with several other translation methods.

abilities match the intended meaning of that particular word. The MF's peak, $\pi(v)$, is the single numerical probability that is most representative of the word's meaning and is the translation of choice. Occasionally, the MF does not have a unique maximum, so all probabilities within a given range can be considered to be equally good representations of the word's meaning. In these cases it is convenient to translate the word into the mid-range of these probabilities. To illustrate the potential accuracy of this approach we compared the peaks of the 977 verbal phrases used by 113 of the subjects in our experiments with the mean of their numerical judgments when judging the same events. We found a remarkable similarity between the two sets: (a) the median within-subject difference between the two is 0.006 and the median absolute difference between them is 0.097; (b) the median within-subject rank order correlation between the peaks of the words and the mean numerical judgments is 0.89; and (c) the two sets are almost perfectly related linearly with a median within-subject intercept of -0.022, a median within subject slope of 1.06, and a median R^2_{adj} of 0.90. These results indicate that the translation procedure can map with high accuracy the intended meaning of the words and predict accurately the numerical probabilities used to describe the same events.

[**V**,**R**] Every MF is, essentially, a collection of ranges since every level of membership, ν ($0 \leq \nu \leq 1$), defines a range of values, $R(\nu)$, such that $\mu(\nu) \geq \nu$. Typically, as ν increases, $R(\nu)$ becomes narrower indicating the range of values that possess that (higher) level of membership is more restrictive. Thus, the translation from a V to a R boils down to the issue of which threshold, ν, to choose. Presumably, there are systematic differences in the ”typical” threshold that individuals tend to use in these circumstances, so one could quantify this tendency and identify the most appropriate range for each individual. We are not aware of any studies that have collected both verbal and upper and lower numerical bounds of the probabilities of the same events, so we are not in a position to assess the efficacy of the proposed approach.

2.6 Imputing vagueness

[**N**,**R**] If numbers are the universal language of uncertainty and everyone interprets them identically, any sensible DM would infer that the F's single N is the center of a range that describes his/her opinions, but there is no clue regarding the implied imprecision of the F's opinion. One could improve the quality of communication by reversing the procedure described for [R,N], i.e., (a) applying the appropriate *stretching (or contracting) transformation* for the DM, and (b) imputing the DM's typical band of imprecision. We are not familiar with any empirical work along these lines.

[**N**,**V**] Recall that all the words in the F's lexicon have single-peaked MFs defined over the [0,1] interval. These functions describe the degree to which any given probability matches the intended meaning of the various phrases. The proposed translation rule calls for the choice of that phrase that has the highest membership at the N in question. This procedure is not guaranteed to yield a unique solution, i.e. there could be several words with equally high membership at that probability, and all these words should be considered equally valid translations of the numerical judgment. If necessary, one of these words can be selected randomly (or by some other tie-breaking procedure). In analyzing our studies we looked at responses from 118 subjects who used an average of 14.85 distinct numerical judgments. We analyzed the verbal responses that were assigned by the subjects to the events to which they assigned a certain numerical response. On the average, each set of events that were judged to be equally probable (in the numerical mode) generated 1.81 distinct verbal phrases, and in 68% of the cases at least one[4] of these verbal responses had, indeed, the maximal membership for that probability. Another look at the same data indicates that for 59.7% of the numerical judgments at least one of the verbal terms used was predicted from the MFs. Interestingly, we found large individual differences: 30 subjects (25.4%) are at, or below a 40% success rate, while for 27 subjects (22.9%) the rate of accurate translation is greater than, or equal to 75%. Not surprisingly, the level of agreement is considerably higher for the extreme (0 and 1), and the central (0.5) numerical probabilities.

[**R**,**V**] All the words in the F's lexicon have (single-peaked and continuous) MFs defined over the [0,1] interval. For any fixed range of numerical probability these functions describe the degree to which the probabilities in that range match the intended meaning of the various phrases. The proposed translation rule calls for the choice of that verbal term that has the highest average membership over the R in question. It is possible that there would be several words with equally high membership over that range probabilities. All these words should be considered equally valid translations. We are not aware of studies that have collected the relevant data for the empirical evaluation of this procedure.

3 General Discussion

In this paper we proposed a unifying conceptual framework for optimal interpersonal translation of probabilistic information for the 9 distinct cases we identified. We discussed the 9 scenarios at different levels of details, and provided extensive empirical support for some of them. Although the cases are not encountered with similar frequency in applied settings, we decided to review all of them to illustrate the generality, feasibility and flexibility of the overall approach.

This line of research is part of an effort to create a general Linguistic Probability Translator (*LiProT*, for short) that could serve both as a useful research tool, and a general decision aid. *LiProT* would facilitate communication of subjec-

[4] In 14.5% of the cases more than one word tied for the highest membership at a given probability. The mean number of words tied for maximal membership was 1.14.

tive uncertainties between participants in various decision situations - forecasters, judges, experts and decision makers - by reducing the dangers of miscommunication of probabilities among the various members of the group.

To fix ideas consider a group of experts (physicians, intelligence officers, financial forecasters, etc.) who communicate with each other, possibly electronically form various locations. As part of this process they need to exchange probabilistic information based on the evidence available to them and reflecting their own unique expertise. If various people in this group have differential preferences for modes of communicating probabilities to others and receiving information from others, then each of the 9 cases discussed above may be relevant for some of the pairs. The procedures described and partially tested in this paper provide a foundation for such a system. Before the meeting, the participants' preferred modes of communication are ascertained, their verbal probability lexicons are mapped, and *LiProT* derives the appropriate translation scheme for each dyad. During the meeting, every probability (N, R or V) used by each of the experts is instantly converted optimally to the favorite modality (N, R or V) of each of the other participants.

For example, assume that participant *A* prefers to communicate and to receive numbers, participant *B* has a universal preference for Vs, and participant *C* prefers to communicate with V, but to receive Ns (the modal pattern according to Wallsten et al. [18]). Every uncertainty judgment provided by *A* (using Ns) will be translated by *LiProT* into the closest V in judge *B* lexicon (using the [N,V] module), and into the most appropriate N for judge *C* (using the [N,N] module). Similarly, the verbal uncertainty judgments provided by *C* will be translated into the closest N for judge *A* (using the [V,N] module), and into the most appropriate V in *B*'s lexicon (using the [V,V] module). Thus, all judges communicate their opinions and receive information in their respective preferred modes. This approach may be too restrictive, since preferences for a particular mode may vary as a function of the situation, the nature of the target event and its underlying uncertainty. A good translator should allow the receiver of the communication to choose the mode of communication. For example, judge *B* may choose to have judge *A* numerical translated by *LiProT* into the closest V in judge in most cases, but occasionally he/she may opt for a simpler, and more direct, translation into the most appropriate N.

In closing we emphasize that this work has focused on communication of uncertainty, and has not addressed the issue of the efficacy of the proposed translations in the context of specific decision situations. We are now conducting empirical work that seeks to determine the degree to which these translation rules, which were shown to improve the inter-personal communication of uncertainties, could also improve the quality of the ultimate decisions involving these uncertainties.

4 Acknowledgements

We thank Professor Thomas S. Wallsten for many insightful comments on an earlier version of this paper.

References

[1] D. Ariely, W. T. Au, R. H. Bender, D. V. Budescu, C. Dietz, H. Gu, T. S. Wallsten, and G. Zauberman. The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, 6:130–147, 2000.

[2] W. Brun, and K. H. Teigen. Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes*, 41:390–414, 1988.

[3] D. V. Budescu, and T. M. Karelitz Improving the quality of decisions through interpersonal translation of probability phrases. *In preparation.*

[4] D. V. Budescu, T. M. Karelitz, and T. S. Wallsten. Predicting the directionality of probability words from their membership functions. *Journal of Behavioral Decision Making*, in press, 2003.

[5] D. V. Budescu, A. K. Rantilla, H. Yu, and T. M. Karelitz. The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes*, in press, 2003.

[6] D. V. Budescu, and T. S. Wallsten. Processing linguistic probabilities: General principles and empirical evidence. In *Decision Making from a Cognitive Perspective*, J. Busemeyer, D. L. Medin, and R. Hastie (Eds.), 275-318, 1995. San Diego, CA: Academic Press.

[7] M. K. Dhami, and T. S. Wallsten. Interpersonal comparison of subjective probability and subjective probability phrases. *Submitted for publication*.

[8] I. Erev, and B. L. Cohen. Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes*, 45:1–18, 1990.

[9] I. Erev, T. S. Wallsten, and D. V. Budescu. Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, 101:519–527, 1994.

[10] U. S. Karmarkar. Subjectively weighted utility: A descriptive extension of expected utility model. *Organizational Behavior and Human Performance* 21:61–72, 1978.

[11] T. M. Karelitz, and D. V. Budescu. You say probable and I say likely: Improving interpersonal communication with probability phrases. *Submitted to publication*.

[12] T. M. Karelitz, M. K. Dhami, D. V. Budescu., and T. S. Wallsten. Toward a universal translator of verbal probabilities. In *Proceedings of the 15th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*, 298-503, 2002.

[13] A. M. Norwich, and I. B. Turksen. A model for the measurement of membership and the consequences of its empirical implementation. *Fuzzy Sets and Systems*, 12:1–25, 1984.

[14] M. J. Olson, and D. V. Budescu. Patterns of preference for numerical and verbal probabilities. *Journal of Behavioral Decision Making*, 10:117–131, 1997.

[15] T. S. Wallsten, D. V. Budescu, I. Erev, and A. Diederich. Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, 10:189–209, 1997.

[16] T. S. Wallsten, D. V. Budescu, A. Rapoport, R. Zwick, and B. Forsyth. Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, 115:348–365, 1986.

[17] T. S. Wallsten, D. V. Budescu, and C-Y. Tsao. Combining linguistic probabilities. *Psychologische Beitraege*, 39:27–55, 1997.

[18] T. S. Wallsten, D. V. Budescu, R. Zwick, and S. M. Kemp. Preferences and reasons for communicating probabilistic information in numerical or verbal terms. *Bulletin of the Psychonomic Society*, 31:135–138, 1993.

[19] I. Yaniv, and E. Kleinberger. Advice taking in decision-making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83:260–281, 2000.

[20] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

[21] R. Zwick, E. Carlstein, and D. V. Budescu. Measures of similarity among fuzzy concepts: A comparative analysis. *International Journal of Approximate Reasoning*, 1:221–242, 1987.

David V. Budescu is a Professor in the Quantitative division of the Department of Psychology, University of Illinois, Urbana-Champaign, IL, USA 61820. E-mail: dbudescu@uiuc.edu

Tzur M. Karelitz is a graduate student in the Quantitative division of the Department of Psychology, University of Illinois, Urbana-Champaign, IL, USA 61820. E-mail: karelitz@uiuc.edu

Relevance of Qualitative Constraints in Diagnostic Processes

A. CAPOTORTI
Università di Perugia, Italy

Abstract

This paper reviews recent results obtained in the medical diagnosis field by adding to a coherent inference process qualitative constraints. Such further considerations turn out to be significant whenever a basic lower-upper conditional probability assessment induces extension bounds too vague to take any decision. Three general types of qualitative judgements are proposed and fully described. They do not constitute a "panacea" to solve any problematic situation, but their application can considerably improve inferences results in specific fields, as two practical applications show.

Keywords

coherent inference, conditional exchangeability, qualitative constraints, diagnosis procedures

1 Introduction

In many practical applications, and in particular in the medical field, there is the problem that the information at hand is not so fully detailed and sound to adopt sophisticated statistical tools. This happens especially whenever information is based on data collected from different sources or by heterogenous samples. In these cases a *genuinely* probabilistic reasoning can anyway help to reach considerable results about relevant statements. Of course, with such approach, answers differ from usual *uniquely determined* statistical results, having, in general, interval-based conclusions. Unluckily, there is the widespread bad habit of avoiding not unique answers by forcing in the model *artificial* assumptions, such as independence, and this can bring to misleading inferences. On the other hand, it is true that, especially if information is very limited , results could be so vague that it is impossible to make any reasonable decision. Hence, it is reasonable to search for further properties that can help us to reach sharper conclusions. This can be obtained by a deeper analysis of the problem and also by further qualitative judgements. Of particular importance are conditional exchangeability assumptions, which are more general and reasonable than those of independence, comparisons between conditional probabilities, which are apt to capture expert convictions not numerically expressible, and restrictions on the admissible class of agreeing conditional measures, which are induced by indirect considerations on some statement not considered at the beginning.

In this paper we will explicitly show how such further considerations can be formalized and operationally adopted in general inference processes. Moreover we will have an idea of their relevance by applying them on two medical diagnostic procedures: a *median decision* process for the asbestosis diagnosis based on X-ray film's readings and a reliability judgement of a *GIST (gastrointestinal stromal tumor)* diagnosis based on istochemical results.

2 Coherent Inferences with Limited Information

As already sketched out in the Introduction, whenever a problem does not allow a description by usual statistical models, a simple probabilistic approach can anyway be adopted to extrapolate which are the bounds induced by the available information. This is possible by embedding the problem at hand in a coherent setting, i.e. representing the relevant entities through conditional events endowed with numerical values or bounds and looking for some class of conditional measures agreeing with them. Once a class has been detected, it can be used to make inference on relevant quantities (usually called "indexes").

With such approach, we have, on one side, the peculiarity of a direct introduction of conditional probability assessments, hence they are not derived as sub-products of joints and marginal evaluations, on the other hand we are aware of working with imprecise tools (interval assessments, classes of distributions, bounds for conclusions, etc.). The wide range of subjects covered in the previous ISIPTAs symposia ([8, 9]) testifies of the meaningfulness and soundness of the last aspect, while appropriateness and usefulness, both from a theoretical and a practical point of view, of the first are contained in the work started in [6] and recently fully described in Coletti & Scozzafava's book [7].

2.1 Preliminaries

Let us now introduce a proper formalization to operate with the framework depicted before. For the sake of simplicity we will use conditional and unconditional events, but everything can be easily generalized to (finite) random variables, conditional or not (see for example what it has been done about conditional previsions in [4]). The initial information, usually a knowledge and/or rule base, is represented trough a conditional lower-upper probability assessment. Hence we will have a generic list of n conditional events $\mathcal{F} = (S_1|C_1, \dots, S_n|C_n)$, where each $S_i|C_i$ represents some macro-situation S_i (i.e. some combination of events) considered in some particular hypothetical circumstances C_i (usually the C_is represent different scenarios).

Incompleteness of the information can have two origins: the S_is do not describe all possible combinations and the different circumstances C_is can *overlap* or do not cover all possibilities. For this, it is crucial to know which are the relationships of incompatibility, implication, coincidence or whatever, among the events $\mathcal{U}_{\mathcal{F}} = \{C_1, \ldots, C_n, S_1, \ldots, S_n\}$ because they represent constraints that any model must fulfill. Moreover they limit which are the possible *atoms*. The atoms are elementary events obtained by full combinations of affirmed or negated events in $\mathcal{U}_{\mathcal{F}}$ [1].

We will generally denote by $\mathcal{L}_C$ the set of such logical constraints and we will refer only to atoms A_r, with $r = 1, \ldots, a$, spanned by $\mathcal{U}_{\mathcal{F}}$ and inside the disjunction $\bigvee_{i=1}^{n} C_i$. In the sequel we will also need to use the characteristic vectors of the events, i.e. vectors whose components are 1 or 0 depending if the corresponding atom implies or not the event, and we will denote them with the same letter of the event but in boldface lower-cases. Hence, for example, $\mathbf{s_i}$ and $\mathbf{c_i}$ will denote the characteristic vectors of S_i and C_i, respectively, while their juxtaposition $\mathbf{s_i c_i}$ will represent the characteristic vector of the conjunction S_iC_i (for the sake of simplicity we will omit the usual conjunction operator $\wedge$). To complete this notational parenthesis, in the following we will use the logical operator $\neg$ to denote negations.

The last component of an assessment is represented by numerical bounds $\mathbf{p} = ([lb_1, ub_1], \ldots, [lb_n, ub_n])$, each closed interval $[lb_i, ub_i]$ associated to the corresponding conditional event $S_i|C_i$, and usually estimated by expert believes, literature reports or by collected data.

Note that some $S_i|C_i$ could be actually unconditional (i.e. the situation S_i is considered independently from any specific circumstance) and in such case C_i will coincide with the sure event Ω. Moreover some of the numerical bounds $[lb_i, ub_i]$ could degenerate in a single value p_i, representing a precise assessment.

2.2 Coherence

If we don't want, or we cannot, adopt for the domain $(\mathcal{F}, \mathcal{L}_C, \mathbf{p})$ a unique probabilistic model, it is just possible to search for a class $\mathbb{P}_{\mathcal{F}}$ of conditional probability distributions, such that $\mathbf{p}$ coincides with the restriction to $\mathcal{F}$ of the closed envelop of $\mathbb{P}_{\mathcal{F}}$. This can be operationally checked by the satisfiability of a *class of sequences* of linear systems. *Sequences* of linear systems are necessary to allow the possibility that conditioning events C_is have induced probability not bounded away from 0. Hence there could be the need of classifying the conditional events in different *zero layers*. On the other hand, a *class* of linear systems is required because, to be sure $\mathbf{p}$ agrees with a *closed envelope*, each bound lb_j or ub_j must be cyclically forced to be strictly fulfilled as an equality (for a deeper exposition

[1] In some discipline atoms are called *possible worlds*.

of both aspects refer again to [7], in particular to chapt. 12 and 15).

Such linear systems will anyway have a common structure like

$$\left\{ \begin{array}{rcl} \mathbf{E}\cdot\mathbf{x} & = & 0 \\ \mathbf{L}\cdot\mathbf{x} & \geq & 0 \\ \mathbf{U}\cdot\mathbf{x} & \leq & 0 \\ \mathbf{x}\geq 0 & , & \mathbf{x}\neq 0 \end{array} \right. \tag{1}$$

where $\cdot$ represents the row-column matrix product, $\mathbf{x}$ is a column vector of unknowns, with each component x_r associated to an atom A_r, $r = 1,\dots,a$, while $\mathbf{E}$, $\mathbf{L}$ and $\mathbf{U}$ are matrices that reflect the numerical constraints induced by $\mathbf{p}$. Hence in $\mathbf{E}$ a generic row is of the form

$$(\mathbf{s_i c_i} - p_i \mathbf{c_i})$$

for each $S_i|C_i$ with a precise assessment p_i and cyclically for one $S_k|C_k$ with an imprecise assessment and forcing p_k to be equal to lb_k or to ub_k. On the other hand, in $\mathbf{L}$ and $\mathbf{U}$ there are, respectively, rows like

$$(\mathbf{s_j c_j} - lb_j \mathbf{c_j})$$

and

$$(\mathbf{s_j c_j} - ub_j \mathbf{c_j})$$

for each $S_j|C_j$ with probability bounds lb_j and ub_j different from the chosen p_k.

Through the set of solutions $\mathbf{x}$, it is possible to represent the searched class $\mathbb{P}_{\mathcal{F}}$.

2.3 Extension

Once coherence of the assessment $(\mathcal{F},\mathcal{L}_C,\mathbf{p})$ has been assured, and in practical application this turns out to be a compulsory step whenever information comes from different sources, it is possible to perform inference on any conditional event $H|E$ judged important to reach conclusions on the problem. Usually H represents some hypothesis to test on the basis of some fact E.

In this context, inference reduces to compute the coherent extension of $\mathbf{p}$ to $H|E$, obtainable as the closed envelop $[lb_{H|E}, ub_{H|E}]$ of the values $P(H|E)$ with $P \in \mathbb{P}_{\mathcal{F}}$. Operationally we need to perform sequences of optimizations of the form

$$\begin{array}{l} \text{minimize/maximize } \mathbf{he}\cdot\mathbf{x} \\ \text{s.t.} \\ \quad \begin{array}{rcl} \mathbf{E}\cdot\mathbf{x} & = & 0 \\ \mathbf{L}\cdot\mathbf{x} & \geq & 0 \\ \mathbf{U}\cdot\mathbf{x} & \leq & 0 \\ \mathbf{e}\cdot\mathbf{x} & = & 1 \\ \mathbf{x}\geq 0 & & \end{array} \end{array} \tag{2}$$

where the normalization constraint $\mathbf{e}\cdot\mathbf{x}=1$ permits the optimization problem to be linear instead of fractional.

The main difficulty of such procedure is the usually huge number a of atoms but, thanks to a smart use of null probabilities, in [2, 5] this complexity problem has been tackled and mainly solved for practical applications.

3 Results Improvement by Qualitative Constraints

Extension bounds $[lb_{H|E}, ub_{H|E}]$ are what, from a pure probabilistic point of view, our information implies on $H|E$ but, sometimes, they could result too wide to take any decision. Anyway, it is possible, maintaining a *model free* approach, to shrink the reference conditional probability class $\mathbb{P}_{\mathcal{F}}$ adding qualitative (i.e. not numerically expressed) considerations to the numerical constraints $\mathbf{p}$. Of course there are several possible different kinds of constraints to introduce, but we will focus on few of them, either because they are quite natural or because by them we have reached quite satisfactory results.

3.1 Conditional Exchangeability vs Independence

As already mentioned, a widespread tool for restricting the variability of the conclusions is to adopt some assumption of independence. And it is actually a powerful restriction, but usually it is a too strong assumption, not supported by the problem. It is in fact usually confused with the information that some evaluations are made *independently* (i.e. one given without knowing the others), while it should be used to model situations whose measure of uncertainty cannot be *modified* by simply taking into account some other aspect. Moreover its formalization and use in a context of partial information should be done with the awareness of all its implications, that are deeper then the simple factorization of some joint probabilities (for more details see once more [7], chapt. 17).

In the presence of strong symmetries, like for example assessment on the same statement made *independently* by different experts with similar skills (see for example Lad et al. [11]), it is more suitable to introduce some kind of *exchangeability*. This is opportune whenever it is relevant *how many* instead of *which* events realize, or, in other words, whenever it possible to identify a *sum* as a sufficient statistic (for a detailed explanation refer to [10], sect. 3.9). In particular, whenever the assessment is mainly conditional, the judgement of *conditional exchangeability* could be the more suitable and it is formulated as follows:
if there is a group of k events $E_1,\dots,E_k$ regarded exchangeable *under a specific scenario* C_j, then any conjunction of the E_is with the same number of affirmed and negated events must be equally evaluated. In other words, for any fixed number $s\in\{0,\dots,k\}$ there must be a constant c_s such that

$$P(E_{i_1}\dots E_{i_s}\neg E_{i_{s+1}}\dots\neg E_{i_k}|C_j)=c_s \tag{3}$$

for any permutation of the indexes $i_1, \ldots, i_k$.

Conditions like (3) actually reduce the "degree of freedom" for the unknowns **x** respect the constraints (1) of the original assessment, restricting "de facto" the admissible class of conditional measures $\mathbb{P}_{\mathcal{F}}$ and, possibly, shrinking some extension bounds.

Since (3) refers to a fixed conditioning event C_j, restriction of this type are easily reported as linear constraints. In fact, denoting with π_s and π'_s the characteristic vectors of two different permutations of the combination $E_{i_1} \ldots E_{i_s} \neg E_{i_{s+1}} \ldots \neg E_{i_k}$, extensions with the further conditional exchangeability requirement obtain by adding to (2) pairwise equalities of the form

$$(\pi_s \mathbf{c_j} - \pi'_s \mathbf{c_j}) \cdot \mathbf{x} = 0 \tag{4}$$

for each pair of permutations $\pi_\mathbf{s}$ and $\pi'_\mathbf{s}$ and each $s = 1, \ldots, k-1$ (note that extreme cases $s = 0$ and $s = k$ do not actually constitute any constraint).

3.2 Conditional Probabilities Comparison

Sometimes there are conditional events which an expert believes more than some other, but he/she cannot express neither precise nor imprecise probability assessments on them, being only capable to compare them.

This is immediately interpretable as

$$P(S_i|C_i) \geq k^+ P(S_j|C_j) \tag{5}$$

for some constant value k^+.

Anyway, if none of the conditional probabilities present in (5) is uniquely constrained, its direct representation by vectors would be

$$\mathbf{x}^T \cdot [(\mathbf{s_i c_i})^T \cdot \mathbf{c_j} - (k^+ \mathbf{s_j c_j})^T \cdot \mathbf{c_i}] \cdot \mathbf{x} \geq 0 \tag{6}$$

that has the drawback of being quadratic. This increases the difficulties for the computation of the extension bounds. In fact, to deal with quadratically constrained optimization problems there are specific Operational Research's techniques, like interior-point algorithms [13] or duality bound methods [14], but they are not so safe and confirmed like those for linear programming problems.

That is why we propose an approximation of (5) that, even being a weaker constraint, has the advantage of leaving the extension problem in a linear form. The idea is of expressing (5) in a parametric way and introducing further unknowns that can capture the basic structure of the parameterization.

If we focus our attention on one of the two conditional probabilities in (5), let us say $P(S_j|C_j)$, we can take it as an *inference target* and compute its extension bounds $[lb_{S_j|C_j}, ub_{S_j|C_j}]$ as it has been illustrated in Subsection 2.3. We can now introduce new variables y_i, $i = 1, \ldots, a$, representing the quantities $P(S_j|C_j)x_i$, so that the inequality (5) can be represented by

$$\mathbf{s_i c_i} \cdot \mathbf{x} - k^{+}\mathbf{c_i} \cdot \mathbf{y} \geq 0; \tag{7}$$

the link by new and old variables by

$$\mathbf{s_j c_j} \cdot \mathbf{x} - \mathbf{c_j} \cdot \mathbf{y} = 0; \tag{8}$$

while the variability bounds for $P(S_j|C_j)$ imply the constraints

$$lb_{S_j|C_j} x_i \leq y_i \leq ub_{S_j|C_j} x_i \qquad \text{for } i = 1, \ldots, a. \tag{9}$$

These constraints are all implied by (5), while the vice versa does not hold in general. Hence, if the minimization/maximization of $\mathbf{he} \cdot \mathbf{x}$ is performed with constraints (2), (7), (8) and (9) we are not guaranteed to have obtained the coherent extension for $P(H|E)$ of $\mathbf{p}$ plus (5), but just an interval containing it. However, once such optimal solutions $\mathbf{x}$ are obtained, they can be substituted in (6) to check if the interval $[lb_{H|E}, ub_{H|E}]$ is coherent. If not, the left-hand-side of (6) will result a negative value that can be adopted as a *measure of violation* of (5).

Of course it is not needed to add sequences of optimizations to cyclically impose equalities in (7) and (9) because they must be fulfilled as they are by each $P \in \mathbb{P}_{\mathcal{F}}$.

Anyway, (7), (8) and (9) increase significantly the space complexity of the optimization procedure. Hence, before to adopt them it would be better to check if the optimal solutions of the original linear program (2) already satisfy (6). If it is the case, it means that the qualitative comparison (5) is redundant because it actually does not restrict the class $\mathbb{P}_{\mathcal{F}}$.

3.3 Selectors Restriction

We introduce now a consideration that will result more technical than the previous ones. It will be less intuitive and also more debatable, hence it should be used more carefully and it will anyway need an *interpretation process* before being presented to a field’s expert for its acceptance.

Analyzing the inference procedure for some conditional event $H|E$, it could happen to notice that results are mainly influenced by the possible variability of some other $K|F$. As usual $K|F$ can be conditioned to a proper F or unconditional, i.e. with $F = \Omega$. If $K|F$ does not belong to the initial list of conditional events $\mathcal{F}$, the induced bounds $[lb_{K|F}, ub_{K|F}]$ for its conditional probability could be extremely vague, and usually this is not noted at the beginning because $K|F$ could be of no direct interest.

However, it could be impossible to assess bounds for $P(K|F)$ either because the data on which $\mathbf{p}$ was built are not available anymore or because there is not direct information on $K|F$. Anyway, an *indirect* consideration is possible.

Variability range $[lb_{K|F}, ub_{K|F}]$ results from the union of all the extensions, say $[lb^j_{K|F}, ub^j_{K|F}]$, with $1 \leq j \leq n$, of the *extreme* conditional distributions $\mathcal{P}_j \subset \mathbb{P}_{\mathcal{F}}$. With *extreme distribution* we mean those $P \in \mathbb{P}_{\mathcal{F}}$ that reach at least one the lower or upper bounds (lb_j or ub_j) of the assessment $\mathbf{p}$. It could happen that some of the $[lb^j_{K|F}, ub^j_{K|F}]$ is narrow enough to drastically influence $P(H|E)$, showing that not all the admissible distributions play the same role for the inference.

Hence, adopting a more *restrictive* attitude and thanks also to some extra consideration, it is possible to select only some of the admissible $\mathcal{P}_j \subset \mathbb{P}_{\mathcal{F}}$ by choosing more informative lower-upper bounds for $P(K|F)$ (possibly coinciding with the narrower interval $[lb^j_{K|F}, ub^j_{K|F}]$) so that the initial assessment can be updated and a new inference on $H|E$ performed.

4 Two Medical Applications

We will show now how the procedures described before can be applied on practical problems. In particular we will illustrate the results we recently attained for two different medical diagnostic processes. The first problem will show how to apply and the relevance of the conditional exchangeability assumptions and of the conditional probabilities comparisons as depicted in subsections 3.1 and 3.2. On the contrary, with the second one we will show the importance of a preliminary check of coherence whenever information comes from different sources and the influence in the results of selector restrictions, in line with subsections 2.2 and 3.3, respectively.

4.1 Accuracy Rates for an Asbestosis Median Decision Procedure

In [3] we re-examined the procedure of median decision making in the context of radiological determination of asbestosis. Median decision applies whenever there is a pool of experts, usually equivalent in skill, examining the same patients and each single case is finally diagnosed on the basis of the agreement of the majority of judgments.

In particular, in a recent paper [12], Tweedie and Mergersen analyze a previous case-report about incidence of asbestosis among a group of people with a similar history of asbestos exposure. Opinions of three radiologists are based on X-ray films readings, and the authors have rather limited information about the median decision procedure. Anyway, they are able to propose a tricky methodology to retrieve some conclusion about the probability of the diagnosis being correct.

However, the authors' analysis deeply relies on a assumption of independence for the experts' assessments and they adopt it because X-ray films are read *in-*

dependently by the radiologists. But this consideration should pertain to *experts' assessment procedure*, not to *our belief* about information's influence one expert opinion *could* have on an other. Actually, since the experts have similar skills, the response of one of them is already a significant indicator of what we could expect from an other.

Tweedie and Mergersen are aware of the inadequacy of the independence presumption, but they wonder how it could be replaced. The fact is that they ”need” to introduce independence to maintain uniqueness of the agreeing conditional probability distribution. On the other hand, the information that the three experts are judged equivalently because of their similar skill cannot be ignored. As we have underlined in Subsection 3.1, assumptions of conditional exchangeabilities could be an appropriate answer to this need.

To make a synthesis (a full description can be obviously found in the cited papers), we can formalize the problem as it follows.

First of all we introduce events that refer to a generic patient with a X-ray film available:

label	*description*
F	asbestosis (*fibrosis*) presence
$D_i \quad , i=1,2,3$	i-th expert positive asbestosis judgment
D^*	positive median decision diagnosis
S^*	positive median decision with a splitting vote

Since the similarity among radiologists, their *sensitivities* for the films' reading process $P(D_i|F)$, $i=1,2,3$, are thought to be equal.

On the basis of recorded data on 642 patients and of specific literature references, the following conditional probability assessment $\mathbf{p}$ on $\mathcal{F}=(D_1|F, D_2|F, D_3|F, D^*, S^*|D^*)$ is considered[2]:

$$\begin{aligned} &P(D_i|F) = .82 \qquad i=1,2,3 \\ &P(D^*) = .12 \\ &P(S^*|D^*) = .42 \end{aligned}$$

The first probability $P(D_i|F)$ comes from literature results on sensitivity analyses performed by comparing radiological and histopathological evaluations. The other two $P(D^*)$ and $P(S^*|D^*)$ derives from the only data reported in [12]. In particular, $P(D^*)$ is directly estimated by the ratio $77/642$ of positive median diagnoses, while $P(S^*|D^*)$ is attained indirectly by the three individual 82%, 86% and 90% positive assessments through the formula

$$P(S^*|D^*) = (100-82)\% + (100-86)\% + (100-90)\% = 42\%.$$

To complete the assessment we must explicitly give which are the possible logical relations $\mathcal{L}_C$ among the unconditional events $\mathcal{U}_{\mathcal{F}} = \{F, D_1, D_2, D_3, D^*, S^*\}$.

[2] In [12] and [3] several assessments with different sensitivity values are examined, here we report only the first one as prototype

By the problem description we can pick out logical dependencies among the median decisions, with or without splitting vote, and individual experts' diagnosis

$$\begin{array}{rcl} S^* & = & (D_1D_2\neg D_3) \vee (D_1 \neg D_2D_3) \vee (\neg D_1D_2D_3) \\ D^* & = & S^* \vee (D_1D_2D_3) \end{array}$$

It is easy to check that the numeric assessment $\mathbf{p}$ is coherent and that, even being a precise conditional probability assessment, the admissible class $\mathbb{P}_{\mathcal{F}}$ is not a single conditional distribution, as it will appear in the sequel.

We can consider the assessment $(\mathcal{F}, \mathcal{L}_{\mathcal{C}}, \mathbf{p})$ as a partial knowledge base whose main "lack" is the absence of an estimate for the expert's *specificity* $P(\neg D_i|\neg F)$. Anyhow, thanks to the conditional independence assumptions

$$P(D_i|D_jF) = P(D_i|F) \quad \text{and} \quad P(D_i|D_j\neg F) = P(D_i|\neg F) \tag{10}$$

and thanks to some algebraic manipulation involving Bayes' Theorem, Tweedie and Mergersen uniquely determine probability values for the usual accuracy indexes *specificity*, *positive predictive value*, *negative predictive value* and estimate the *true positive proportion*. We can compare their results with what we obtained firstly without any assumption, secondly adopting the method of Subsection 3.1 to incorporate the following conditions of conditional exchangeability[3]

$$\begin{array}{rcl} P(D_1D_2\neg D_3|F) = & P(D_1\neg D_2D_3|F) & = P(\neg D_1D_2D_3|F) \\ P(D_1\neg D_2\neg D_3|F) = & P(\neg D_1\neg D_2D_3|F) & = P(\neg D_1D_2\neg D_3|F) \\ \\ P(D_1D_2\neg D_3|\neg F) = & P(D_1\neg D_2D_3|\neg F) & = P(\neg D_1D_2D_3|\neg F) \\ P(D_1\neg D_2\neg D_3|\neg F) = & P(\neg D_1\neg D_2D_3|\neg F) & = P(\neg D_1D_2\neg D_3|\neg F) \end{array} \tag{11}$$

and finally using considerations of Subsection 3.2 to consider the following conditional probabilities' comparisons that arise from the formalization of an interview with a further physician[4]:

[3] With respect to the notation of Subs.3.1 we have $k = 3$, $E_i = D_i$ and C_j equal at first to F and after to $\neg F$

[4] These comparisons are the result of the formalization of a long and detailed analysis of the influence of the knowledge of the answers of some expert on the behaviors of the others. It has been performed with a physician extraneous to the rest of the work

$$\frac{P(D_3|D_1D_2F)}{P(\neg D_3|D_1D_2F)} \geq 3/2\frac{P(D_1|F)}{P(\neg D_1|F)}$$
(linear)

$$\frac{P(D_3|\neg D_1\neg D_2F)}{P(\neg D_3|\neg D_1\neg D_2F)} \leq 2/3\frac{P(D_3|F)}{P(\neg D_3|F)}$$
(linear)

$$\frac{P(D_3|\neg D_1\neg D_2\neg F)}{P(\neg D_3|\neg D_1\neg D_2\neg F)} \leq 2/3\frac{P(D_3|\neg F)}{P(\neg D_3|\neg F)}$$
(quadratic)

$$P(D_3|D_1\neg D_2F) \in [.5, .5 + (P(D_3|F) - .5)]$$
(linear)

$$P(D_3|D_1\neg D_2\neg F) \in [.5 - (P(D_3|F) - .5), .5]$$
(linear)

$$P(D_2|D_1F) \geq P(D_2|F)$$
(linear)

$$P(D_3|D_2D_1F) \geq P(D_3|D_1F)$$
(quadratic)

$$P(D_2|\neg D_1\neg F) \leq P(D_2|\neg F)$$
(quadratic)

$$P(D_3|\neg D_2\neg D_1\neg F) \leq P(D_3|\neg D_1\neg F)$$
(quadratic)

Note that such relations, even being similar in structure (the first three actually reflect odds ratios comparisons), are distinguished, by labels, between those of them that are actually linear constraints since some quantity is uniquely determined and those that are properly quadratic and need the proposed linear approximation.

We cannot go into technical details, but it is important to mention just one computational feature: the number of atoms in this problem is 16, but conditional independence assumptions (10) reduce at two the degrees of freedom for their probabilities, i.e. everything is fully determined once the experts’ sensitivity $P(D_i|F)$ and specificity $P(\neg D_i|\neg F)$ could be selected, while with conditional exchangeabilities (11) we have only a reduction at 8 degrees of freedom.

Here we report the different inferences performed on several accuracy indexes, specifying the particular assumptions adopted

index	*description*	extension bounds under			
		cond. idep.	*no ass.*	*cond. exch.*	*qual. comp.*
$P(\neg D_i\|\neg F)$	experts’ specificity	.957	[0 , 1]	[.603 , 1]	[.820 , .970]
$P(F\|D^*)$	positive predict. val.	.961	[0 , 1]	[0 , 1]	[0 , .779]
$P(\neg F\|\neg D^*)$	negative predict. val.	.988	[.970 , 1]	[.971 , 1]	[.979 , 1]
$P(F)$	asbestosis incidence	.126	[0 , .130]	[0 , .130]	[0 , .106]
$P(D^*\|F)$	med. dec. sensitivity	.994	[.730 , 1]	[.730 , 1]	[.820 , .878]
$P(\neg D^*\|\neg F)$	med. dec. specificity	.995	[.880 , 1]	[.880 , 1]	[.954 , .970]

Whenever conditional exchangeability cannot help on limiting vague inference bounds, the further qualitative probabilistic comparisons are determinant. In fact, apart from the positive predictive value, all the intervals in the last column are tight enough to judge the procedure. About the only ”vague” interval $[0, .779]$, even it does not bound from below the positive predictive value, it gives an interesting upper limitation for such index.

Moreover, note that some interval of the last column do not contain the corresponding values obtained by Tweedie and Mergersen. This because the further constraints go in the opposite direction of independence, bringing some kind of correlation but leaving ”untouched” the conditional exchangeability framework.

Our computations needed to solve several liner programming problems, but what we obtained is really based on reasonable probabilistic statements and not on tricky manipulation that have the only justification of bringing to single values instead of intervals.

4.2 Reliability of GIST Diagnosis Based on Partial Information

Other prototypes of applications of inference with a not fully detailed model are the medical diagnostic procedures where there is not a *golden standard* protocol to follow. This happens when new advances in the understanding of the biology are done or new techniques are discovered. In such situations, different opinions appear in scientific literature and they are based on disparate case studies, each one with its peculiarity and heterogeneity of data.

In particular, in [1] we analysed a diagnostic process for *gastrointestinal stromal tumors* (GISTs) where only recently a new and reliable phenotypic marker (the KIT protein CD117) for these neoplasm has been introduced.

The diagnosis path consist mainly of two stages: at first a histological analysis is done and later an immunohistochemical schema is adopted to confirm cases previously suspected to be GISTs. What we have done was to numerically evaluate the quality of the first discrimination and it was possible by matching information from a personal case study[5] and immunohistochemical behaviors reported in the relevant literature.

The problem can be synthesized as it follows: we have selected as relevant for a lesion the events

label	*description*
DIAGNOSIS	lesion is histologically suspected to be a GIST
GIST	lesion is really a GIST
CD117	KIT protein expression
CD34	Hematopoietic progenitor cell antigen expression
SMA	Muscle actin expression
DESM	Desmin expression
S100	S-100 protein expression

where the first two distinguish the suspected tumors by those actually belonging to the GIST's family, while the others represent the positivity for specific immunohistochemical markers.

[5] Data was collected at *Istituto di Anatomia e Istologia Patologica - Divisione di ricerca sul cancro - Universit degli Studi di Perugia - Italy* during the period Jan.1998–Sept.2002

We had only the following logical restriction due to the extreme specificity of the KIT marker

$$\text{CD117} \subseteq \text{GIST}.$$

By the personal case study we estimated (by observed frequencies) the following “knowledge base”

statement	*cond. prob.*
DIAGNOSIS	.510
CD117 CD34 ¬DESM ¬S100 | DIAGNOSIS	.308
¬SMA ¬CD117 CD34 DESM ¬S100 | DIAGNOSIS	.077
¬SMA CD117 CD34 ¬DESM S100 | DIAGNOSIS	.077
SMA ¬CD117 CD34 ¬DESM ¬S100 | DIAGNOSIS	.077
SMA CD117 ¬CD34 ¬S100 | DIAGNOSIS	.231
SMA CD117 ¬CD34 ¬DESM S100 | DIAGNOSIS	.077
¬SMA CD117 ¬CD34 ¬DESM S100 | DIAGNOSIS	.077

but it turned out to be incoherent with the “rule base” we derived by collecting different literature sources

statement	*expected frequencies bounds*
CD34 | CD117	[.60 , .70]
SMA | CD117	[.30 , .40]
S100 | CD117	[.096 , .105]
DESM | CD117	[.01 , .02]

A deeper analysis of the observed results has shown that there were two cases with dubious S100 positivity and they have judged as the cause of the inconsistency. In fact, performing an inference based only on the knowledge base, we obtain that the percentage for S100 | CD117 results between 13% and 70%, while it should be around 10% as indicated in the rule base.

Revising these two judgements, we have obtained a different knowledge base consistent with the literature rule base

statement	*cond. prob.*
DIAGNOSIS	.510
CD117 CD34 ¬DESM ¬S100 | DIAGNOSIS	.380
¬SMA ¬CD117 CD34 DESM ¬S100 | DIAGNOSIS	.077
SMA ¬CD117 CD34 ¬DESM ¬S100 | DIAGNOSIS	.077
SMA CD117 ¬CD34 ¬S100 | DIAGNOSIS	.077
SMA CD117 ¬CD34 ¬DESM S100 | DIAGNOSIS	.077
¬SMA CD117 ¬CD34 ¬DESM ¬S100 | DIAGNOSIS	.077

Further considerations has induced us to add the further constraint $P(\text{CD117}|\text{GIST}) \in [0.95, 0.99]$ for the sensitivity of the KIT marker.

Putting together all these assessments, they force the usual accuracy indexes to be in the following bounds

index	*description*	*extension bounds*
$\mathrm{P}(\mathrm{DIAGNOSIS} \mid \mathrm{GIST})$	sensitivity	[.47 , .76]
$\mathrm{P}(\neg\mathrm{DIAGNOSIS} \mid \neg\mathrm{GIST})$	specificity	[0 , .88]
$\mathrm{P}(\mathrm{GIST} \mid \mathrm{DIAGNOSIS})$	positive predictive value	[.85 , .94]
$\mathrm{P}(\neg\mathrm{GIST} \mid \neg\mathrm{DIAGNOSIS})$	negative predictive value	[0 , 69]

that, apart from the positive predictive value, reflect a weak ”influence” of the constraint considered.

Adding to the assessment the probabilistic comparison $P(\text{DIAGNOSIS} \mid \text{GIST}) \geq P(\text{DIAGNOSIS} \mid \neg\text{GIST})$ we have not obtained appreciable improvements.

On the contrary, reasoning as described in Subsection 3.3, we have focused the attention on the ”a priori” values of GIST’s incidence. In fact, its coherent bounds result $P(\text{GIST}) \in [.59, .97]$ while one extreme sub-class of the admissible conditional probabilities induce the more restrictive lower bound of .81. Since the pathologist judged as reasonable a variability around 81% of the GISTS’s incidence, we have added to the whole assessment the restriction $P(\text{GIST}) \in [.806, .815]$ obtaining the more relevant results

index	*description*	*extension bounds*
$\mathrm{P}(\mathrm{DIAGNOSIS} \mid \mathrm{GIST})$	sensitivity	[.53 , .59]
$\mathrm{P}(\neg\mathrm{DIAGNOSIS} \mid \neg\mathrm{GIST})$	specificity	[.58 , .80]
$\mathrm{P}(\mathrm{GIST} \mid \mathrm{DIAGNOSIS})$	positive predictive value	[.85 , .93]
$\mathrm{P}(\neg\mathrm{GIST} \mid \neg\mathrm{DIAGNOSIS})$	negative predictive value	[.22 , .32]

that confirm a good positive predictive performance of the diagnostic procedure, while they express a really bad reliability in the case of a negative diagnosis. This, in a way, reverses the role that the KIT marker should have. Instead of being used as a *confirmatory* tool in already suspected cases, it should have a crucial role for the right diagnosis of lesion at first not suspected to be GISTs.

References

[1] A. Capotorti, S. Fagundes Leite. Reliability of GIST diagnosis based on partial information. accepted for the pubblication on *Statistical Inference in Human Biology*, (M. Di Bacco, G. D’Amore, F. Scalfari editors), Kluwer Academic Publishers, Norwell, MA (USA), 2003.

[2] 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), 280-87, 2003.

[3] A. Capotorti, F. Lad. Reassessing accuracy rates from median decision procedures. submitted to *The American Statistician*, 2003.

[4] A. Capotorti, T. Paneni. An operational view of coherent conditional previsions. *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Benferhat, S., Besnard, P. (eds): ECSQUARU 2001, LNAI 2143, 132–143, 2001.

[5] A. Capotorti, B. Vantaggi. Locally strong coherence in an inference process. *Annals of Mathematics and Artificial Intelligence*, 35(1-4), 125- 149, 2002.

[6] G. Coletti, R. Scozzafava. Characterization of Coherent Conditional Probabilities as a Tool for their Assessment and Extension. *Int. Journ. of Uncertainty, Fuzziness and Knowledge-Based Systems*, 4(2): 103-127, 1996.

[7] G. Coletti, R. Scozzafava. *Probabilistic Logic in a Coherent Setting*, Dordrecht: Kluwer, Series “Trends in Logic”, 2002.

[8] ISIPTA’99. *Proocedings of the First International Symposium on Imprecise Probabilities and Their Applications.* de Cooman, G., Cozman, F.G., Moral, S., Walley, P. (eds.), Imprecise Probabilities Project, Universiteit Gent, Belgium, 1999.

[9] ISIPTA’01. *Proocedings of the Second International Symposium on Imprecise Probabilities and Their Applications.* held at the Robert Purcell Community Center of Cornell University Ithaca, NY, USA, 26-29 June 2001, de Cooman, G., Fine, T. L., Seinfeld, T. (eds.), The Nederlands: Shaker Publishing.

[10] F. Lad. *Operational Subjective Statistical Methods: a mathematical, philosophical, and historical introduction*, New York: John Wiley, 1996.

[11] F. Lad, M. Di Bacco. Assessing the value of a second opinion: the role and structure of exchangeability. *Annals of Mathematics and Artificial Intelligence*, 35:227–252, 2002.

[12] R. Tweedie, K. Mengersen. Calculating accuracy rates from multiple assessors with limited information. *The American Statistician* , 53, 233–238, 1999.

[13] R.J. Vanderbei, D.F. Shanno. An interior-point algorithm for nonconvex nonlinear programming. Tecnical report SOR-97-21, *Statistics and Operations Research*, Princeton University, 1997.

[14] N.V. Thoai. Duality bound method for the general quadratic programming problem with quadratic constraints. *Jour. of Optimization Theory and Applications*, 107(2): 331–354, 2000.

Andrea Capotorti is with the Dipartimento di Matematica e Informatica at Università degli Studi di Perugia, via Vanvitelli 1, 06123 Perugia, Italy. E-mail: capot@dipmat.unipg.it

Expected Utility with Multiple Priors[*]

E. CASTAGNOLI
Università Bocconi, Italy

F. MACCHERONI
Università Bocconi and ICER, Italy

M. MARINACCI
Università di Torino and ICER, Italy

Abstract

Let $\succsim$ be a preference relation on a convex set F. Necessary and sufficient conditions are given that guarantee the existence of a set $\{u_l\}$ of affine utility functions on F such that $\succsim$ is represented by $U(f) = u_l(f)$ if $f \in F_l$; where each F_l is a convex subset of F. The interpretation is simple: facing a "non-homogeneous" set F of alternatives, a decision maker splits it into "homogeneous" subsets F_l, and acts as a standard expected utility maximizer on each of them.

In particular, when F is a set of simple acts, each u_l corresponds to a subjective expected utility with respect to a finitely additive probability P_l; while when F is a set of continuous acts, each probability P_l is countably additive.

Keywords

preference representation, subjective probability, nonexpected utility, integral representation, multiple priors, countable additivity

1 Introduction

Given a preference relation $\succsim$ on a convex set F, we provide necessary and sufficient conditions that guarantee the existence of a set $\{u_l\}$ of affine utility functions on F such that $\succsim$ is represented by

$$U(f) = u_l(f) \quad \text{if } f \in F_l,$$

[*] We gratefully acknowledge the financial support of CNR and MIUR. First version: November 1999.

where each F_l is a convex subset of F. This representation has a simple interpretation: facing a “non-homogeneous” set of alternatives F, a decision maker splits it into “homogeneous” subsets F_l and, on each of them, she behaves as a standard expected utility maximizer. For example, the F_l can be commodities traded in a local market l and F be the global market, or the F_l can be sets of lotteries on which the decision maker feels she has the same information.

The idea underlying these results is close to the one of Castagnoli and Maccheroni (2000), but the difference of setups heavily reflects on the techniques we use in the proofs.

In particular, if F is a convex set of objective lotteries, the model falls in the class of lottery dependent utility (see, e.g., Maccheroni, 2002, and the references therein).

While, when F is a set of simple (resp. continuous) acts, each u_l corresponds to a subjective expected utility with respect to a finitely additive (resp. countably additive) probability P_l. This time we are in the spirit of multiple priors models: for example, Choquet Expected Utility of Schmeidler (1989) and Maxmin Expected Utility of Gilboa and Schmeidler (1989) are particular cases of the proposed model when the family $\{F_l\}_{l\in L}$ consists of sets of comonotone and affinely related acts, respectively. In fact, many recent papers focus on specific cases of the model obtained here, and provide interesting interpretations on the derived family of probabilities. See, e.g. Nehring (2001), Ghirardato, Maccheroni, and Marinacci (2002), Kopylov (2002), Siniscalchi (2003). In particular, the latter work builds on a similar idea and looks for conditions ensuring the uniqueness of the subjective probability used to evaluate the expected utility of each act; furthermore, differently from us, the sets F_l are elicited from the preference.

2 A general representation result

Let F be a convex subset of a vector space, X a nonempty convex subset of F, $\{F_l\}_{l\in L}$ a family of convex subsets of F such that $F = \bigcup_{l\in L} F_l$ and $X \subseteq \bigcap_{l\in L} F_l$, and $\succsim$ a binary relation on F. As usual, we denote by $\succ$ and $\sim$ the asymmetric and the symmetric parts of $\succsim$. In the sequel we will make use of the following assumptions on $\succsim$.

Weak Order (WO): For all f_1 and f_2 in F: $f_1 \succsim f_2$ or $f_2 \succsim f_1$. For all f_1, f_2, and f_3 in F: if $f_1 \succsim f_2$ and $f_2 \succsim f_3$, then $f_1 \succsim f_3$.

Local Independence (LI): For all $l \in L$, all f_1, f_2, and f_3 in F_l, and all α in $(0,1)$: $f_1 \succsim f_2$ implies $\alpha f_1 + (1-\alpha) f_3 \succsim \alpha f_2 + (1-\alpha) f_3$. When L is a singleton this property is the standard *Independence (I)*.

Local Continuity (LC): For all $l \in L$ and all f_1, f_2, and f_3 in F_l: if $f_1 \succ f_2$ and $f_2 \succ f_3$, then there exist α and β in $(0,1)$ such that $\alpha f_1 + (1-\alpha) f_3 \succ f_2$

and $f_2 \succ \beta f_1 + (1-\beta) f_3$. When L is a singleton this property is the standard *Continuity (C)*.

Boundedness (B): For all f in F: there exist $x_1, x_2 \in X$ such that $x_1 \succsim f$ and $f \succsim x_2$.

Quasiconcavity (Q): For all f_1 and f_2 in F and all α in $(0,1)$: $f_1 \sim f_2$ implies $\alpha f_1 + (1-\alpha) f_2 \succsim f_1$.

As suggested by Siniscalchi (2003), a natural way to elicit the sets F_l from the preference is to look for the maximal convex subsets of F on which it satisfies the standard assumptions of expected utility. Next theorem shows that the first four properties are necessary and sufficient to yield a piecewise affine representation of $\succsim$.

Theorem 1 *Given a binary relation $\succsim$ on F, the following conditions are equivalent:*

(i) $\succsim$ satisfies WO, LI, LC, and B.

(ii) There exists a family $\{u_l\}$ of affine functionals on F such that the functional

$$U(f) = u_l(f) \quad \text{if } f \in F_l \tag{1}$$

represents $\succsim$ on F and $U(X) = U(F)$.

Moreover, U is unique up to positive affine transformations.

Ghirardato, Maccheroni, and Marinacci (2002), show that under suitable topological assumptions, the closed and convex hull of the family $\{u_l\}$ is the Clarke subdifferential of U.

Next we show that the quasiconcavity assumption Q implies concavity of the representation.

Corollary 1 *Let $\succsim$ be a binary relation represented by (1). Then, $\succsim$ satisfies Q if and only if $\{u_l\}$ can be chosen such that*

$$U(f) = \min_{l \in L} u_l(f)$$

for all $f \in F$.

It is easy to see that the assumptions WO, LI, LC, B, and Q are independent. Moreover, the Example on page 216 of Castagnoli and Maccheroni (2000) with $F = \mathbb{R}^2$ and $X = \{0\}$ shows that WO, LI, and LC are not sufficient to obtain a representation like (1). Further notice that $U(\alpha f + (1-\alpha) x) = \alpha U(f) + (1-\alpha) U(x)$ for all $\alpha \in [0,1]$, $f \in F$, and $x \in X$. We call this property *X-affinity.*

A special case of interest is the one in which only F and X are *a priori* given and, for all $f \in F - X$, F_f is the convex hull $\mathrm{co}\{f, X\}$ of $\{f\}$ and X. In this case LI and LC can be restated with no explicit reference to the family F_f, moreover they can be replaced by

X-Independence (X-I): For all f_1, f_2 in F, all x in X, and all α in $(0,1)$: $f_1 \succsim f_2$ iff $\alpha f_1 + (1-\alpha)x \succsim \alpha f_2 + (1-\alpha)x$.

X-Continuity (X-C): For all $x_1, x_2 \in X$ and all f in F: if $x_1 \succ f$ and $f \succ x_2$, then there exist α and β in $(0,1)$ such that $\alpha x_1 + (1-\alpha)x_2 \succ f$ and $f \succ \beta x_1 + (1-\beta)x_2$.

The previous Theorem takes the following form.

Corollary 2 *Let $\succsim$ be a binary relation on F, and $F_f = \mathrm{co}\{f, X\}$ for all $f \in F - X$. The following statements are equivalent:*

(i) $\succsim$ satisfies WO, LI, LC, and B.

(ii) $\succsim$ satisfies WO, X-I, X-C, and B.

(iii) There exists an X-affine functional $U : F \to \mathbb{R}$ representing $\succsim$ and such that $U(X) = U(F)$.

Moreover, U is unique up to positive affine transformations.

In this case, $\succsim$ satisfies Q iff there exists a family $\mathcal{U}$ of affine functionals on F, all of which are concordant on X, such that

$$U(f) = \min_{u \in \mathcal{U}} u(f).$$

We think that the above general results shed some light on the common traits of several well-known particular results in the literature. As an exemplification in the next section we apply them to a problem of choice under uncertainty. We are confident that they can be fruitfully employed to the study of different problems; e.g., decision models in which F is the convex set of all (closed and convex) sets of lotteries over a finite set Z of outcomes, and its elements are considered as menus of alternatives available to a decision maker (see, e.g., Dekel, Lipman, and Rustichini, 2001).

3 The Anscombe - Aumann setup

We now consider the special case in which F is a set of acts; more precisely, we focus on two possible settings.

The first one is the classical Anscombe - Aumann setup. S is a nonempty set of *states of the world*, Σ an algebra of subsets of S called *events*, X a convex set of *outcomes*. A *simple act* is just an X-valued, simple and Σ-measurable function; $F = F^s$ is the set of all simple acts. In this setting a *probability* on Σ is a finitely additive set function $P : \Sigma \to [0,1]$ such that $P(\emptyset) = 0$ and $P(S) = 1$.

The second one is a topological variation of the first. S is a compact metric set, Σ its Borel σ-field, and X a finite dimensional simplex. A *continuous act* is just an X-valued, continuous function; $F = F^c$ is the set of all continuous acts. In this setting a *probability* on Σ is a countably additive set function $P : \Sigma \to [0,1]$ such that $P(\emptyset) = 0$ and $P(S) = 1$.

For every $f_1, f_2 \in F$ and $\alpha \in [0,1]$ we denote by $\alpha f_1 + (1-\alpha) f_2$ the act in F which yields $\alpha f_1(s) + (1-\alpha) f_2(s) \in X$ for every $s \in S$. With a slight abuse of notation, we identify X with the set of all constant acts (thus making it a convex subset of F).

We will replace assumption B with the mildly stronger conditions:

Monotonicity (M): For all f_1 and f_2 in F: if $f_1(s) \succsim f_2(s)$ on S, then $f_1 \succsim f_2$.

Nondegeneracy (N): Not for all f_1 and f_2 in F, $f_1 \succsim f_2$.

Let $G \supseteq X$ be a subset of F, a functional $U : G \to \mathbb{R}$ is said to be *monotone* if $g_1(s) \succsim g_2(s)$ on S implies $U(g_1) \geq U(g_2)$; *automonotone* if $U(g_1(s)) \geq U(g_2(s))$ on S implies $U(g_1) \geq U(g_2)$ (that is, if U is monotone with respect to the pointwise dominance relation it induces on G). Next lemma is a little variation on the von Neumann - Morgenstern Theorem to yield a subjective probability result *à la* Anscombe and Aumann (1963). In particular, the lemma guarantees an expected utility representation for any preference $\succsim$ on G that satisfies WO, I, C, M, and N.

Lemma 1 *Let $G \supseteq X$ be a convex subset of F, $U : G \to \mathbb{R}$ a nonconstant, automonotone, affine functional, and u the restriction of U to X.*[1] *There exists a probability P on Σ such that*

$$U(g) = \int_S (u \circ g)\, dP$$

for all $g \in G$.

We are now ready to state the anticipated result.

Theorem 2 *Given a binary relation $\succsim$ on F, the following conditions are equivalent:*

(i) $\succsim$ satisfies WO, LI, LC, M, and N.

[1] More precisely: denoted by x_S the constant act taking value x for all $s \in S$, u is the function defined by $u(x) = U(x_S)$; the shorter expression we adopted derives from the identification of X with the set of all constant acts.

(ii) *There exists a family $\{P_l\}_{l\in L}$ of probabilities on Σ, and an affine nonconstant function u on X, such that the functional*

$$U(f) = \int_S (u \circ f)\, dP_l \quad \text{if } f \in F_l \tag{2}$$

represents $\succsim$ on F and it is monotone.

Moreover, U is unique up to positive affine transformations.

In the next corollary we consider the special case when the quasiconcavity axiom Q holds.

Corollary 3 *Let $\succsim$ be a binary relation represented by (2). Then, $\succsim$ satisfies Q if and only if $\{P_l\}_{l\in L}$ can be chosen such that*

$$U(f) = \min_{l\in L} \int_S (u \circ f)\, dP_l$$

for all $f \in F$.

The counterpart of Corollary 2 for $F = F^s$ is Theorem 1 of Gilboa and Schmeidler (1989), and we explicitly state it only in the case $F = F^c$. Here, the set of all probability measures is endowed with the weak* topology.

Corollary 4 *A binary relation $\succsim$ on F^c satisfies WO, X-I, X-C, M, N, and Q iff there exist an affine function $u : X \to \mathbb{R}$ and a compact and convex set $\mathcal{C}$ of probability measures, such that*

$$f \succsim g \Leftrightarrow \min_{P\in\mathcal{C}} \int_S (u \circ f)\, dP \geq \min_{P\in\mathcal{C}} \int_S (u \circ g)\, dP$$

for all $f, g \in F^c$. $\mathcal{C}$ is unique and u is unique up to a positive linear transformation.

Differently from the Gilboa and Schmeidler (1989) result, the set of priors $\mathcal{C}$ consists of countably additive probability measures. This way of obtaining countable additivity is alternative to that used by Marinacci, Maccheroni, Chateauneuf, and Tallon (2002); in fact, we add assumptions on the structure of the model rather than assumptions on the preference.

4 Proofs

Next Lemma is a minor variation on the Hahn - Banach Extension Theorem. Its proof is part of the one of Lemma 4 p. 829-830 in Maccheroni (2002).

Lemma 2 *Let $F \supseteq G \supseteq X$ be nonempty convex subsets of a vector space. If a functional $U : F \to \mathbb{R}$ is X-affine, concave, and $U_{|G}$ is affine, then there exists an affine functional $u : F \to \mathbb{R}$ such that $u \geq U$ and $u_{|G} = U_{|G}$.*

The following is a topological version of the previous one. We refer to Aliprantis and Border (1999) Chapter 5 for the basic notation and results on topological vector spaces' theory.

Lemma 3 *Let E be a vector space, E' be a total subspace of its algebraic dual, and K be a $\sigma(E',E)$-compact subset of E'. Set*

$$I(e)=\min_{e'\in K}\langle e,e'\rangle$$

for all $e\in E$. If I is affine on a convex subset C of E, there exists an extreme point e'_C of K such that $I_{|C}=e'_C$, i.e.

$$e'_C\in \text{Argmin}_{e'\in K}\langle e,e'\rangle$$

for all $e\in C$.

Proof of Lemma 3. For all $e\in C$, $\text{Argmin}_{e'\in K}\langle e,e'\rangle$ is a $\sigma(E',E)$-closed subset of K. By compactness of K, it is enough to show that $\bigcap_{j=1}^{n}\text{Argmin}_{e'\in K}\langle e_j,e'\rangle\neq\emptyset$ for any $e_1,e_2,...,e_n\in C$. Choose $w'\in\text{Argmin}_{e'\in K}\left\langle\Sigma_{j=1}^{n}\frac{1}{n}e_j,e'\right\rangle$.

$$I\left(\Sigma_{j=1}^{n}e_j\right)=nI\left(\Sigma_{j=1}^{n}\frac{1}{n}e_j\right)=n\left\langle\Sigma_{j=1}^{n}\frac{1}{n}e_j,w'\right\rangle=\Sigma_{j=1}^{n}\langle e_j,w'\rangle$$

but, since I is affine on C

$$I\left(\Sigma_{j=1}^{n}e_j\right)=nI\left(\Sigma_{j=1}^{n}\frac{1}{n}e_j\right)=n\Sigma_{j=1}^{n}\frac{1}{n}I(e_j)=\Sigma_{j=1}^{n}\min_{e'\in K}\langle e_j,e'\rangle.$$

We can conclude that $w'\in K$ and

$$\Sigma_{j=1}^{n}\langle e_j,w'\rangle=\Sigma_{j=1}^{n}\min_{e'\in K}\langle e_j,e'\rangle.$$

Hence

$$\langle e_j,w'\rangle=\min_{e'\in K}\langle e_j,e'\rangle$$

for all $j=1,2,...,n$, that is $w'\in\bigcap_{j=1}^{n}\text{Argmin}_{e'\in K}\langle e_j,e'\rangle$.

Moreover, $\bigcap_{e\in C}\text{Argmin}_{e'\in K}\langle e,e'\rangle$ is a nonempty intersection of compact extreme sets, hence it is a compact extreme set, and it contains an extreme point. **Q.E.D.**

Proof of Theorem 1 and Corollary 1.[2] By the von Neumann - Morgenstern Theorem for all $l\in L$ there exists an affine functional

$$u_l:F_l\to\mathbb{R}$$

[2] The proofs are not separated to avoid duplicate notation.

representing $\succsim$ on F_l. We still denote by u_l an arbitrarily fixed affine extension of u_l to F. Since $u_{l|X}$ is an affine representation of $\succsim$ on X, it is unique up to positive affine transformations. Fix arbitrarily $m \in L$ and set $u = u_{m|X}$. For all $l \in L$ choose u_l so that $u_{l|X} = u$.

By B, for all $f \in F$ there exist $x_1, x_2 \in X$ such that $x_1 \succsim f \succsim x_2$. Therefore $u(x_1) = u_l(x_1) \geq u_l(f) \geq u_l(x_2) = u(x_2)$, for all $l \in L$ such that $f \in F_l$, and there exists $\alpha \in [0,1]$ such that

$$\begin{aligned} u_l(f) &= \alpha u(x_1) + (1-\alpha) u(x_2) \\ &= u(\alpha x_1 + (1-\alpha) x_2) \\ &= u_l(\alpha x_1 + (1-\alpha) x_2), \end{aligned}$$

therefore $u_l(f)$ does not depend on the choice of $l \in L$ such that $f \in F_l$. Moreover, the argument above shows that there exists $x_f \in X$ (i.e. $\alpha x_1 + (1-\alpha) x_2$) such that $x_f \sim f$ and

$$u_l(f) = u(x_f)$$

for all $l \in L$ such that $f \in F_l$.

We set

$$U(f) = u_l(f) \quad \text{if } f \in F_l.$$

What precedes guarantees that U is well defined, and $U(f) = u(x_f) = U(x_f)$ implies $U(F) = U(X)$. For all $f_1, f_2 \in F$, let $f_i \sim x_i \in X$ to obtain

$$f_1 \succsim f_2 \Leftrightarrow x_1 \succsim x_2 \Leftrightarrow u(x_1) \geq u(x_2) \Leftrightarrow U(f_1) \geq U(f_2).$$

If $U' : F \to \mathbb{R}$ is affine on F_l for all $l \in L$ and represents $\succsim$, then $u' = U'_{|X} = au + b$ for some $a > 0$ and $b \in \mathbb{R}$; for all $f \in F$, let $f \sim x_f \in X$ to obtain

$$U'(f) = u'(x_f) = au(x_f) + b = aU(f) + b.$$

This concludes the proof of Theorem 1.

For any $\alpha \in [0,1]$, $f \in F$, and $x \in X$, choose $l \in L$ such that $f \in F_l$ to obtain

$$\begin{aligned} U(\alpha f + (1-\alpha) x) &= u_l(\alpha f + (1-\alpha) x) \\ &= \alpha u_l(f) + (1-\alpha) u_l(x) \\ &= \alpha U(f) + (1-\alpha) U(x), \end{aligned}$$

this shows that U is X-affine.

Next we prove Corollary 1. If U is constant, the result is trivial. If U is not constant, there exist $f_1, f_2 \in F$ such that $f_1 \succ f_2$ and, by B, there exist $x_1^*, x_{-1}^* \in X$ such that $x_1^* \succ x_{-1}^*$. W.l.o.g. assume $x_{-1}^* = -x_1^*$ (so that $0 \in X$) and $U(x_1^*) = 1$, $U(x_{-1}^*) = -1$, whence $U(0) = U\left(\frac{1}{2}x_1^* + \frac{1}{2}x_{-1}^*\right) = 0$. Then U is positively homogeneous. The (unique) positively homogeneous extension of U to the convex cone H generated by F is the functional defined by

$$V(\gamma f) = \gamma U(f)$$

if $f \in F$ and $\gamma > 0$. Let $h \in H$ and y in the convex cone Y generated by X, there exist $\gamma > 0$, $f \in F$ and $x \in X$ such that $h = \gamma f$ and $y = \gamma x$, whence

$$\begin{aligned} \frac{1}{2}V(h+y) &= \frac{1}{2}V(\gamma(f+x)) \\ &= \gamma V\left(\frac{1}{2}(f+x)\right) \\ &= \gamma U\left(\frac{1}{2}f+\frac{1}{2}x\right) \\ &= \gamma\left(\frac{1}{2}V(f)+\frac{1}{2}V(x)\right) \\ &= \frac{1}{2}(V(h)+V(y)), \end{aligned}$$

that is $V(h+y) = V(h)+V(y)$.

Let $h_1, h_2 \in H$; there exist $\gamma > 0$, $f_1, f_2 \in F$ such that $h_i = \gamma f_i$. If $V(h_1) = V(h_2)$, $U(f_1) = V(f_1) = V(f_2) = U(f_2)$, so that $f_1 \sim f_2$ and $U\left(\frac{1}{2}f_1+\frac{1}{2}f_2\right) \geq U(f_1) = \frac{1}{2}U(f_1)+\frac{1}{2}U(f_2)$, that is $V(h_1+h_2) \geq V(h_1)+V(h_2)$. Else if $V(h_1) > V(h_2)$, there exists $y \in Y$ such that $V(y) = V(h_1)-V(h_2)$ (take $(V(h_1)-V(h_2))x_1^*$), then

$$\begin{aligned} V(h_1+h_2)+V(y) &= V(h_1+h_2+y) \\ &\geq V(h_1)+V(h_2+y) \\ &= V(h_1)+V(h_2)+V(y). \end{aligned}$$

That is, V is superlinear and U is concave. Now using Lemma 2 for each F_l we can choose v_l such that $v_l : F \to \mathbb{R}$ is affine, $v_l \geq U$ and $v_{l|F_l} = U_{|F_l}$. Replace the u_l chosen at the beginning of the proof with v_l to obtain

$$U(f) = v_l(f) = \min_{i \in L} v_i(f)$$

if $f \in F_l$. The rest is trivial. **Q.E.D.**

Given Theorem 1 and Corollary 1, the **proof of Corollary 2** is a long, simple exercise.

We denote by $B_0(S,\Sigma)$ the vector space of all real valued, simple and Σ-measurable functions, endowed with the supnorm topology. If S is a compact metric set, we denote by $C(S)$ the vector space of all real valued, continuous functions, endowed with the supnorm topology. It is well known that the topological dual of $B_0(S,\Sigma)$ (resp. $C(S)$) is the vector space $ba(S,\Sigma)$ of all bounded, finitely additive set functions on Σ (resp. the vector space $ca(S)$ of all countably additive set functions on Σ): the duality being

$$\langle \varphi, \mu \rangle = \int_S \varphi d\mu$$

for all $\varphi \in B_0(S,\Sigma)$ and $\mu \in ba(S,\Sigma)$ (resp. $\varphi \in C(S)$ and $\mu \in ca(S)$). If $k \in \mathbb{R}$, the constant element of $B_0(S,\Sigma)$ or $C(S)$ taking value k on S will be denoted again by k. A functional I on a subset of $B_0(S,\Sigma)$ or $C(S)$ is *monotone* if $\varphi_1 \geq \varphi_2$ implies $I(\varphi_1) \geq I(\varphi_2)$. A monotone linear functional I on $B_0(S,\Sigma)$ or $C(S)$ corresponds to a positive set function μ.

Proof of Lemma 1. Let $u = U_{|X}$; obviously u is affine, (and continuous if $F = F^c$). For all $g \in G$, let $\overline{x} \in g(S)$ be such that $u(\overline{x}) \geq u(g(s))$ for all $s \in S$ and $\underline{x} \in g(S)$ be such that $u(\underline{x}) \leq u(g(s))$ for all $s \in S$. The existence of such $\overline{x}$ and $\underline{x}$ descends from the finiteness of $g(S)$ if $F = F^s$, from the continuity of g and u if $F = F^c$. Then $U(\underline{x}) \leq U(g) \leq U(\overline{x})$, and there exists $x_g \in X$ such that $U(x_g) = U(g)$. Hence $U(G) = U(X)$ and there exists $x_*, x^* \in \operatorname{int} U(X)$ with $U(x_*) < U(x^*)$. Assume first $-U(x_*) = U(x^*) = 1$. Automonotonicity of U yields that $g_1, g_2 \in G$ and $u \circ g_1 = u \circ g_2$ imply $U(g_1) = U(g_2)$. It is easy to see that $\Phi = \{u \circ g : g \in G\}$ is a convex subset of $B_0(S,\Sigma)$ or $C(S)$ containing the constant functions 1 and -1.

Define $I : \Phi \to \mathbb{R}$ by

$$I(\varphi) = U(g)$$

if $\varphi = u \circ g$. I is monotone, affine, $I(0) = 0$ and $I(1) = 1$. It is routine to extend I to the vector subspace $\langle \Phi \rangle$ of $B_0(S,\Sigma)$ or $C(S)$ generated by Φ and obtain a linear, monotone functional $\hat{I} : \langle \Phi \rangle \to \mathbb{R}$ such that $\hat{I}(0) = 0$ and $\hat{I}(1) = 1$. A classical extension result of Kantorovich (see, e.g., Aliprantis and Border, 1999, Lemma 7.31) guarantees that there exists a linear, monotone extension $\tilde{I}$ of $\hat{I}$ to the whole $B_0(S,\Sigma)$ or $C(S)$. We can conclude that there exists a probability P on Σ such that

$$U(g) = I(u \circ g) = \int_S (u \circ g)\, dP$$

for all $g \in G$.

Finally, if it is not the case that $-U(x_*) = U(x^*) = 1$, there exist $a > 0$ and $b \in \mathbb{R}$ such that $-(aU(x_*) + b) = (aU(x^*) + b) = 1$, and the proposed technique yields

$$aU(g) + b = \int_S (a(u \circ g) + b)\, dP$$

for all $g \in G$, as wanted. **Q.E.D.**

Proof of Theorem 2 and Corollary 3.[3] M implies B. If $F = F^s$, for any act f take $\overline{x} \in f(S)$ such that $\overline{x} \succsim f(s)$ for all $s \in S$ and $\underline{x} \in f(S)$ such that $f(s) \succsim \underline{x}$ for all $s \in S$ to obtain $\overline{x} \succsim f$ and $f \succsim \underline{x}$. If $F = F^c$, let $v : X \to \mathbb{R}$ be an affine function that represents $\succsim$ on X; for any act f, there exists $\underline{s}$ and $\overline{s}$ such that $v(f(\overline{s})) \geq v(f(s)) \geq v(f(\underline{s}))$ for all $s \in S$, then M guarantees that $f(\overline{s}) \succsim f \succsim f(\underline{s})$. By Theorem 1 there exists a functional $U : F \to \mathbb{R}$, affine on F_l for all $l \in L$, that represents $\succsim$ (and for all $f \in F$ there exists $x_f \in X$ such that $x_f \sim f$).

[3] The proofs are not separated to avoid duplicate notation.

M also implies that U is automonotone on F (*a fortiori* on F_l for all $l \in L$). In fact, $U(f_1(s)) \geq U(f_2(s))$ on S implies $f_1(s) \succsim f_2(s)$ on S and $f_1 \succsim f_2$, whence $U(f_1) \geq U(f_2)$. M and N imply that U is nonconstant on F_l for all $l \in L$ (just take $f_1^* \succ f_{-1}^*$, and $x_1^*, x_{-1}^* \in X$ with $x_i^* \sim f_i$ to have $U(x_1^*) > U(x_{-1}^*)$). Apply Lemma 1 to F_l for each $l \in L$ to obtain a family $\{P_l\}_{l \in L}$ of probabilities on Σ such that

$$U(f) = \int_S (u \circ f)\, dP_l \quad \text{if } f \in F_l,$$

where $u : X \to \mathbb{R}$ is the restriction of U to X. This proves Theorem 2.

Next we prove Corollary 3. Assuming Q holds, then U is concave.

If $F = F^s$, w.l.o.g. $u(X) \supseteq [-1,1]$, and $\{u \circ f : f \in F\}$ is the set $B_0(S, \Sigma, u(X))$ of simple, Σ measurable functions from S to $u(X)$.

Else if $F = F^c$, w.l.o.g. $u(X) = [-1,1]$, and $\{u \circ f : f \in F\}$ is the set $C(S, u(X))$ of continuous functions from S to $u(X)$.[4]

For all $\varphi \in B_0(S, \Sigma, u(X))$ or $C(S, u(X))$, set

$$I(\varphi) = U(f)$$

if $\varphi = u \circ f$. I is monotone, $u(X)$-affine, concave, $I(0) = 0$ and $I(1) = 1$. Therefore, its positive homogeneous extension $\hat{I}$ to $B_0(S, \Sigma)$ or $C(S)$ is monotone, superlinear, and such that $\hat{I}(\varphi + k) = \hat{I}(\varphi) + k$ for all $\varphi \in B_0(S, \Sigma)$ or $C(S)$ and all $k \in \mathbb{R}$. Moreover, being bounded on $B_0(S, \Sigma, [-1,1])$ or $C(S, [-1,1])$, $\hat{I}$ is continuous in the supnorm. Standard convex analysis results guarantee that there exists a unique convex and weak* compact set $\mathcal{C}$ of probabilities such that

$$\hat{I}(\varphi) = \min_{P \in \mathcal{C}} \int_S \varphi\, dP$$

(just take as $\mathcal{C}$ the superdifferential of $\hat{I}$ at 0). The functional $\hat{I}$ is affine on the convex set $\Phi_l = \{u \circ f : f \in F_l\}$ for all $l \in L$. In fact, for all $l \in L$ and all $\varphi_i = u \circ f_i$ with $f_i \in F_l$, and $\alpha \in [0,1]$ we have

$$\begin{aligned} \hat{I}(\alpha\varphi_1 + (1-\alpha)\varphi_1) &= I(u \circ (\alpha f_1 + (1-\alpha) f_2)) \\ &= U(\alpha f_1 + (1-\alpha) f_2) \\ &= \alpha U(f_1) + (1-\alpha) U(f_2) \\ &= \alpha \hat{I}(\varphi_1) + (1-\alpha)\hat{I}(\varphi_2). \end{aligned}$$

By Lemma 3, there exist $P_l' \in \mathcal{C}$ such that

$$\hat{I}(\varphi) = \int_S \varphi\, dP_l'$$

[4] Let $x_1^* \in u^{-1}(1)$ and $x_{-1}^* \in u^{-1}(-1)$. The restriction ν of u to $[x_{-1}^*, x_1^*]$ is an homeomorphism between $[x_{-1}^*, x_1^*]$ and $[-1,1]$; so if $\varphi : S \to [-1,1]$ is continuous, $f = \nu^{-1} \circ \varphi : S \to [x_{-1}^*, x_1^*] \subseteq X$ is a continuous act such that

$$u(f(s)) = u(\nu^{-1}(\varphi(s))) = \nu(\nu^{-1}(\varphi(s))) = \varphi(s).$$

for all $\varphi \in \Phi_l$. Therefore for all $l \in L$ and all $f \in F_l$

$$U(f) = I(u \circ f) = \min_{P \in C} \int_S (u \circ f)\, dP = \int_S (u \circ f)\, dP'_l = \min_{m \in L} \int_S (u \circ f)\, dP'_m.$$

The rest is trivial. **Q.E.D.**

The **proof of Corollary 4** is immediate.

References

[1] Aliprantis, C.D., and K.C. Border (1999). *Infinite Dimensional Analysis*, 2nd edition. Springer, New York.

[2] Anscombe, F.J., and R.J. Aumann (1963). A definition of subjective probability, *Annals of Mathematical Statistics* **34**, 199-205.

[3] Castagnoli, E., and F. Maccheroni (2000). Restricting independence to convex cones, *Journal of Mathematical Economics* **34**, 215-223.

[4] Dekel, E., B.L. Lipman, and A. Rustichini (2001). Representing preferences with a unique subjective state space, *Econometrica* **69**, 891-934.

[5] Ghirardato, P., F. Maccheroni, and M. Marinacci (2002). Ambiguity from the differential viewpoint, ICER Working Paper 2002/17.

[6] Gilboa, I., and D. Schmeidler (1989). Maxmin expected utility with non-unique prior, *Journal of Mathematical Economics* **18**, 141-153.

[7] Kopylov, I. (2002). α-maxmin expected Utility, mimeo, University of Rochester, Department of Economics.

[8] Maccheroni, F. (2002). Maxmin under risk, *Economic Theory* **19**, 823-831.

[9] Marinacci, M., F. Maccheroni, A. Chateauneuf, and J-M. Tallon (2002). Monotone continuous multiple priors, mimeo, Università di Torino, Dipartimento di Statistica e Matematica Applicata.

[10] Nehring, K. (2001). Ambiguity in the context of probabilistic beliefs, mimeo, University of California, Davis, Department of Economics.

[11] Siniscalchi, M. (2003). A behavioral characterization of plausible priors, mimeo, Northwestern University, Department of Economics.

[12] Schmeidler, D. (1989), Subjective probability and expected utility without additivity, *Econometrica* **57**, 571–587.

Combining Belief Functions Issued from Dependent Sources

MARCO E.G.V. CATTANEO
ETH Zürich, Switzerland

Abstract

Dempster's rule for combining two belief functions assumes the independence of the sources of information. If this assumption is questionable, I suggest to use the least specific combination minimizing the conflict among the ones allowed by a simple generalization of Dempster's rule. This increases the monotonicity of the reasoning and helps us to manage situations of dependence. Some properties of this combination rule and its usefulness in a generalization of Bayes' theorem are then considered.

Keywords

belief functions, propositional logic, combination, Dempster's rule, independence, conflict, monotonicity, nonspecificity, idempotency, associativity, Bayes' theorem

1 Introduction

In the theory of belief functions, Dempster's rule allows us to pool the information issued from several sources, if we assume that these are independent. In his original work [2], Dempster based the independence concept on the usual statistical one and underlined the vagueness of its real world meaning. Shafer reinterpreted Dempster's work and in his monograph [8] defined a belief function without assuming an underlying probability space, making so the independence assumption even more problematic.

In probability theory, the independence concept refers to classes of events or to random variables, with respect to a single probability distribution (this kind of independence for belief functions is studied for instance in Ben Yaghlane, Smets and Mellouli [1]). On the contrary, the concept considered here refers to several sources of information issuing several belief functions over the same frame of discernment. The assumption of the independence of the sources can be justified only by analogies with other situations in which this assumption proved to be sensible (cf. Smets [10]).

Following Dubois and Prade [3], I consider a generalization of Dempster's rule which allows the sources of information to be dependent. This general rule just assigns to a pair of belief functions a set of possible combinations, compelling us to make a choice. If the independence of the sources of information is doubtful (that is, we cannot adequately justify its assumption), I suggest to choose the least specific combination minimizing the conflict. This increases the monotonicity of the reasoning (in particular, complete monotonicity is assured if it does not entail incoherence) and helps us to manage situations of dependence (in particular, idempotency is assured).

2 Setting and Notation

It is assumed that the reader has a basic knowledge of the Dempster-Shafer theory and of classical propositional logic (refer for instance to Shafer [8] and to Epstein [4], respectively).

Let $\mathcal{U}$ be a finite set of propositional variables, which represents the topic considered. $\mathcal{L}_{\mathcal{U}}$ denotes the language of propositional logic built over the alphabet $\mathcal{U} \cup \{\top, \neg, \vee, \wedge, \rightarrow\}$, where $\top$ is the tautology. $V_{\mathcal{U}}$ denotes the set of (classical) valuations of $\mathcal{L}_{\mathcal{U}}$, i.e. the consistent assignments $v : \mathcal{L}_{\mathcal{U}} \longrightarrow \{t, f\}$ of truth values to the formulas of $\mathcal{L}_{\mathcal{U}}$ (thus $|V_{\mathcal{U}}| = 2^{|\mathcal{U}|}$). The mapping

$$\begin{array}{rrcl} T_{\mathcal{U}}: & \mathcal{L}_{\mathcal{U}} & \longrightarrow & 2^{V_{\mathcal{U}}} \\ & \varphi & \longmapsto & \{v \in V_{\mathcal{U}} : v(\varphi) = t\} \end{array}$$

assigns to each formula of $\mathcal{L}_{\mathcal{U}}$ the set of its models, i.e. the valuations for which the formula is true.[1]

Definition 1 *A basic belief assignment (bba) is a function*

$$m : 2^{V_{\mathcal{U}}} \longrightarrow [0,1] \textit{ such that } m(\emptyset) = 0 \textit{ and } \sum_{A \subseteq V_{\mathcal{U}}} m(A) = 1.^{2}$$

$\mathcal{M}_{\mathcal{U}}$ is the set of bbas on $2^{V_{\mathcal{U}}}$.

The belief and the plausibility about $\mathcal{U}$ with bba m are the functions

$$\begin{array}{rrcl} bel: & \mathcal{L}_{\mathcal{U}} & \longrightarrow & [0,1] \\ & \varphi & \longmapsto & \displaystyle\sum_{A \subseteq T_{\mathcal{U}}(\varphi)} m(A), \\ pl: & \mathcal{L}_{\mathcal{U}} & \longrightarrow & [0,1] \\ & \varphi & \longmapsto & \displaystyle\sum_{A \cap T_{\mathcal{U}}(\varphi) \neq \emptyset} m(A). \end{array}$$

[1] $T_{\mathcal{U}}$ is not injective ($\mathcal{L}_{\mathcal{U}}$ is redundant) but it is surjective ($\mathcal{L}_{\mathcal{U}}$ is sufficient).

[2] The beliefs are normalized, since the “open-world assumption” (see for instance Smets [9]) does not make sense in the setting of classical propositional logic: a formula and its negation cannot both be false.

Consider two finite sets of propositional variables $\mathcal{U} \subseteq \mathcal{V}$. If *bel* is a belief about $\mathcal{V}$, the belief $bel\downarrow_{\mathcal{U}}$ about $\mathcal{U}$ is the restriction of *bel* to $\mathcal{L}_{\mathcal{U}}$. If *bel* is a belief about $\mathcal{U}$, the belief $bel\uparrow^{\mathcal{V}}$ about $\mathcal{V}$ is the vacuous extension of *bel* to $\mathcal{L}_{\mathcal{V}}$, i.e. the minimal belief about $\mathcal{V}$ whose restriction to $\mathcal{L}_{\mathcal{U}}$ is *bel* (where minimal means that if bel' is a belief about $\mathcal{V}$ satisfying $bel'\downarrow_{\mathcal{U}}= bel$, then $bel\uparrow^{\mathcal{V}} \leq bel'$).[3]

Definition 2 *A joint belief assignment (jba) with marginal bbas $m_1, m_2 \in \mathcal{M}_{\mathcal{U}}$ is a function*

$$\underline{m} : 2^{V_{\mathcal{U}}} \times 2^{V_{\mathcal{U}}} \longrightarrow [0,1] \text{ such that}$$
$$\sum_{B \subseteq V_{\mathcal{U}}} \underline{m}(A,B) = m_1(A) \text{ for all } A \subseteq V_{\mathcal{U}} \text{ and}$$
$$\sum_{A \subseteq V_{\mathcal{U}}} \underline{m}(A,B) = m_2(B) \text{ for all } B \subseteq V_{\mathcal{U}}.$$

$\underline{\mathcal{M}}_{\mathcal{U}}^{m_1,m_2}$ *is the set of jbas with marginal bbas $m_1, m_2 \in \mathcal{M}_{\mathcal{U}}$.*

The conflict of a jba $\underline{m}$ is the quantity

$$c(\underline{m}) = \sum_{A \cap B = \emptyset} \underline{m}(A,B).$$

For any $m_1, m_2 \in \mathcal{M}_{\mathcal{U}}$, the function $\underline{m}_D$ on $2^{V_{\mathcal{U}}} \times 2^{V_{\mathcal{U}}}$ defined by

$$\underline{m}_D(A,B) = m_1(A)\, m_2(B)$$

is a jba with marginal bbas m_1 and m_2 (it is the jba which corresponds to the independence assumption). Thus $\underline{\mathcal{M}}_{\mathcal{U}}^{m_1,m_2}$ cannot be empty.

In the following, bel_1 and bel_2 will denote two beliefs about $\mathcal{U}$ with bbas m_1 and m_2, respectively (and pl_1 and pl_2 will denote the respective plausibilities). If $\underline{m} \in \underline{\mathcal{M}}_{\mathcal{U}}^{m_1,m_2}$ with $c(\underline{m}) < 1$, the function m on $2^{V_{\mathcal{U}}}$ defined by $m(\emptyset) = 0$ and

$$m(A) = \frac{1}{1 - c(\underline{m})} \sum_{B \cap C = A} \underline{m}(B,C) \text{ if } A \neq \emptyset$$

is a bba. The belief about $\mathcal{U}$ with bba m is called combination of bel_1 and bel_2 with respect to $\underline{m}$, and is denoted by $bel_1 \otimes_{\underline{m}} bel_2$. The rule $\otimes$ generalizes Dempster's one $\oplus$, since the latter is the combination with respect to the particular jba $\underline{m}_D$, or symbolically $\oplus = \otimes_{\underline{m}_D}$.

3 Monotonicity and Conflict

A reasoning process is called monotonic if the acquisition of new information does not compel us to give up some of our beliefs; otherwise it is called non-monotonic. In the Dempster-Shafer theory, the reasoning process consists in the

[3] If m is the bba associated with bel, then the bba associated with $bel\uparrow^{\mathcal{V}}$ is the function m' on $2^{V_{\mathcal{V}}}$ defined by $m'(T_{\mathcal{V}}(\varphi)) = m(T_{\mathcal{U}}(\varphi))$ for all $\varphi \in \mathcal{L}_{\mathcal{U}}$, and $m'(A) = 0$ if $A \notin T_{\mathcal{V}}(\mathcal{L}_{\mathcal{U}})$.

combination of beliefs. That is, the reasoning would be monotonic only if

$$bel_1 \otimes_{\underline{m}} bel_2 \geq \max\left(bel_1, bel_2\right),$$

which does not always hold (cf. Yager [12]). Proposition 1 gives the best possible lower bound for $bel_1 \otimes_{\underline{m}} bel_2(\varphi)$ based only on the knowledge of $bel_1(\varphi)$, $bel_2(\varphi)$ and $c(\underline{m})$.

Proposition 1 *If $\underline{m} \in \underline{\mathcal{M}}_{\mathcal{U}}^{m_1,m_2}$ with $c(\underline{m}) < 1$, and $\varphi \in \mathcal{L}_{\mathcal{U}}$, then*

$$bel_1 \otimes_{\underline{m}} bel_2(\varphi) \geq \max\left(\frac{bel_1(\varphi) - c(\underline{m})}{1 - c(\underline{m})}, \frac{bel_2(\varphi) - c(\underline{m})}{1 - c(\underline{m})}, 0\right).$$

Proof. $$(1 - c(\underline{m}))\, bel_1 \otimes_{\underline{m}} bel_2(\varphi) = \sum_{\emptyset \neq (A \cap B) \subseteq T_{\mathcal{U}}(\varphi)} \underline{m}(A,B) \geq$$
$$\geq \sum_{A \subseteq T_{\mathcal{U}}(\varphi)} \sum_{B \subseteq V_{\mathcal{U}}} \underline{m}(A,B) - \sum_{A \cap B = \emptyset} \underline{m}(A,B) = bel_1(\varphi) - c(\underline{m}).$$

Similarly, $(1 - c(\underline{m}))\, bel_1 \otimes_{\underline{m}} bel_2(\varphi) \geq bel_2(\varphi) - c(\underline{m})$. □

From Proposition 1 it follows that if $\underline{m}$ has no conflict (i.e. $c(\underline{m}) = 0$), then we have monotonicity. But if $\underline{m}$ has some conflict (i.e. $c(\underline{m}) > 0$), then the monotonicity is assured only for the formulas φ such that $\max(bel_1(\varphi), bel_2(\varphi)) = 1$. In general we can affirm that the more $\underline{m}$ has conflict, the more we have nonmonotonicity.

The monotonicity is admissible only if there is a belief bel about $\mathcal{U}$ with $bel \geq \max(bel_1, bel_2)$. If there is a formula φ with $bel_1(\varphi) > pl_2(\varphi)$,[4] then the monotonicity is not admissible, since $bel \geq \max(bel_1, bel_2)$ implies that

$$bel(\top) \geq bel(\varphi) + bel(\neg\varphi) \geq bel_1(\varphi) + bel_2(\neg\varphi) > 1.$$

Proposition 2 assures that if the monotonicity is admissible, then it is feasible (that is, there is a jba without conflict).

Proposition 2

$$\min_{\underline{m} \in \underline{\mathcal{M}}_{\mathcal{U}}^{m_1,m_2}} c(\underline{m}) = \max_{\varphi \in \mathcal{L}_{\mathcal{U}}} \left(bel_1(\varphi) - pl_2(\varphi)\right).$$

Proof. Let $\underline{m}$ be a jba minimizing the conflict (such a jba certainly exists since $\underline{\mathcal{M}}_{\mathcal{U}}^{m_1,m_2} \subset \mathbb{R}^{2^{2|V_{\mathcal{U}}|}}$ is compact and not empty).

If $A_1, A_2, B_1, B_2 \subseteq V_{\mathcal{U}}$ with $A_1 \cap B_1 = \emptyset$, $A_1 \cap B_2 \neq \emptyset$, $A_1 \neq A_2$, $\underline{m}(A_1, B_1) > 0$ and $\underline{m}(A_2, B_2) > 0$, then $A_2 \cap B_1 = \emptyset$ and $A_2 \cap B_2 \neq \emptyset$, and without loss of generality we may assume that $\underline{m}(A_2, B_1) > 0$.

[4] Notice that $bel_2(\psi) - pl_1(\psi) = bel_1(\varphi) - pl_2(\varphi)$ with $\varphi = \neg\psi$.

To prove this, consider the function $\underline{m}'$ on $2^{V_{\mathcal{U}}} \times 2^{V_{\mathcal{U}}}$ defined by

$$\underline{m}'(A,B) = \begin{cases} \underline{m}(A,B) - \varepsilon & \text{if } (A,B) \in \{(A_1,B_1),(A_2,B_2)\}, \\ \underline{m}(A,B) + \varepsilon & \text{if } (A,B) \in \{(A_1,B_2),(A_2,B_1)\}, \\ \underline{m}(A,B) & \text{otherwise,} \end{cases}$$

for an ε such that $0 < \varepsilon < \min(\underline{m}(A_1,B_1), \underline{m}(A_2,B_2))$. It is easily verified that $\underline{m}' \in \underline{\mathcal{M}}_{\mathcal{U}}^{m_1,m_2}$ and $c(\underline{m}') \leq c(\underline{m})$, with equality only if $A_2 \cap B_1 = \emptyset$ and $A_2 \cap B_2 \neq \emptyset$.

Let $\mathcal{A} = \{A \subseteq V_{\mathcal{U}} : \exists\, B \subseteq V_{\mathcal{U}} \;\; A \cap B = \emptyset, \underline{m}(A,B) > 0\}$ and $\underline{\mathcal{A}} = \bigcup_{A \in \mathcal{A}} A$.

If $B \cap \underline{\mathcal{A}} \neq \emptyset$, then $m_2(B) = \sum_{A \in \mathcal{A}, A \cap B \neq \emptyset} \underline{m}(A,B)$.

This can be proven as follows. Since $B \cap \underline{\mathcal{A}} \neq \emptyset$, there is an $A_1 \in \mathcal{A}$ with $A_1 \cap B \neq \emptyset$. Since $A_1 \in \mathcal{A}$, there is a $B_1 \subseteq V_{\mathcal{U}}$ with $A_1 \cap B_1 = \emptyset$ and $\underline{m}(A_1,B_1) > 0$. If $A_2 \subseteq V_{\mathcal{U}}$ with $A_1 \neq A_2$ and $\underline{m}(A_2,B) > 0$, then we are in the situation considered above (with $B_2 = B$). Therefore $A_2 \in \mathcal{A}$ (since $A_2 \cap B_1 = \emptyset$ and $\underline{m}(A_2,B_1) > 0$) and $A_2 \cap B \neq \emptyset$. Thus $\underline{m}(A,B) > 0$ implies $A \in \mathcal{A}$ and $A \cap B \neq \emptyset$.

Let $\varphi \in \mathcal{L}_{\mathcal{U}}$ with $T_{\mathcal{U}}(\varphi) = \underline{\mathcal{A}}$. Then

$$\begin{aligned} c(\underline{m}) = \sum_{A \in \mathcal{A}, A \cap B = \emptyset} \underline{m}(A,B) = \sum_{A \in \mathcal{A}} m_1(A) - \sum_{A \in \mathcal{A}, A \cap B \neq \emptyset} \underline{m}(A,B) = \\ = \sum_{A \in \mathcal{A}} m_1(A) - \sum_{B \cap \underline{\mathcal{A}} \neq \emptyset} m_2(B) \leq bel_1(\varphi) - pl_2(\varphi). \end{aligned}$$

On the other hand, for any $\psi \in \mathcal{L}_{\mathcal{U}}$ (let $C = T_{\mathcal{U}}(\psi)$ and $\overline{C} = V_{\mathcal{U}} \setminus C$),

$$\begin{aligned} c(\underline{m}) \geq \sum_{A \subseteq C, B \subseteq \overline{C}} \underline{m}(A,B) = \sum_{A \subseteq C} m_1(A) - \sum_{A \subseteq C, B \not\subseteq \overline{C}} \underline{m}(A,B) \geq \\ \geq \sum_{A \subseteq C} m_1(A) - \sum_{B \not\subseteq \overline{C}} m_2(B) = bel_1(\psi) - pl_2(\psi) \end{aligned}$$

□

Let $c_{\min}^{m_1,m_2}$ denote the value of $\min_{\underline{m} \in \underline{\mathcal{M}}_{\mathcal{U}}^{m_1,m_2}} c(\underline{m})$, and let bel_1 and bel_2 be called compatible if $bel_1 \leq pl_2$. Proposition 2 enables us to determine $c_{\min}^{m_1,m_2}$ and to prove Corollary 1.

Corollary 1 *The following assertions are equivalent.*

- *The monotonicity of the combination of bel_1 and bel_2 is admissible.*
- *bel_1 and bel_2 are compatible.*
- *$c_{\min}^{m_1,m_2}$=0.*

4 The Choice of a Combination Rule

The only case in which the marginal bbas uniquely determine the jba is the conditioning of a belief. The conditioning on $\varphi \in \mathcal{L}_{\mathcal{U}}$ of a belief *bel* about $\mathcal{U}$ is the result of its combination with $bel_{\mathcal{U}}^{\varphi}$, where $bel_{\mathcal{U}}^{\varphi}$ denotes the minimal belief about $\mathcal{U}$ assigning the value 1 to the formula φ.[5] It is easily verified that if one of the two beliefs which have to be combined has the form $bel_{\mathcal{U}}^{\varphi}$, then the jba is unique (i.e. $\underline{\mathcal{M}}_{\mathcal{U}}^{m_1,m_2} = \{\underline{m}_D\}$). Thus in the case of conditioning the general rule $\otimes$ reduces itself to Dempster’s one.

Generally, in order to combine two beliefs bel_1 and bel_2 about $\mathcal{U}$, we must choose a jba $\underline{m} \in \underline{\mathcal{M}}_{\mathcal{U}}^{m_1,m_2}$. Sometimes we can analyse in detail the situation and base our choice on specific assumptions about the nature of the dependence of the sources of information, but usually we can at most assume their independence. Thus there is little loss of generality in considering only the two usual cases: the one in which the independence is assumed, and the one in which nothing is assumed about the sources. In both cases we need a combination rule; that is, we need an operator $\star$ assigning to every pair of bbas $m_1, m_2 \in \mathcal{M}_{\mathcal{U}}$ a jba $m_1 \star m_2 \in \underline{\mathcal{M}}_{\mathcal{U}}^{m_1,m_2}$, for any finite set of propositional variables $\mathcal{U}$. Such an operator can be sensible only if it satisfies the following basic requirements (the first two make the combination rule independent of the particular logical formalization, whereas the third one is a technical necessity).

- The influence of $\mathcal{U}$ on $\star$ must be limited to the cardinality of $V_{\mathcal{U}}$. That is, if $\mathcal{V}$ is a set of propositional variables and $f : V_{\mathcal{V}} \longrightarrow V_{\mathcal{U}}$ is a bijection, then
$$(m_1 \circ f) \star (m_2 \circ f)(A,B) = m_1 \star m_2 (f(A), f(B)) \text{ for all } A,B \subseteq V_{\mathcal{V}}.$$

- The operator $\star$ must be “equivariant” with respect to the vacuous extensions. That is, if $\mathcal{V}$ is a finite set of propositional variables with $\mathcal{U} \subseteq \mathcal{V}$ and m_1', m_2' are the bbas associated with $bel_1 \uparrow^{\mathcal{V}}$ and $bel_2 \uparrow^{\mathcal{V}}$, respectively, then
$$m_1' \star m_2' (T_{\mathcal{V}}(\varphi), T_{\mathcal{V}}(\psi)) = m_1 \star m_2 (T_{\mathcal{U}}(\varphi), T_{\mathcal{U}}(\psi)) \text{ for all } \varphi, \psi \in \mathcal{L}_{\mathcal{U}}.$$

- The combination with respect to $m_1 \star m_2$ must be defined as often as possible. That is, if $c_{\min}^{m_1,m_2} < 1$, then $c(m_1 \star m_2) < 1$.

It is easily verified that the operator which corresponds to Dempster’s rule ($m_1 \star m_2 = \underline{m}_D$) satisfies these basic requirements. Thus if in the considered situation the assumption of the independence of the sources of information is sensible, we should employ Dempster’s rule. But if the independence is doubtful, employing this rule can be hazardous, since the conflict is in general pretty high (even if

[5] The bba associated with $bel_{\mathcal{U}}^{\varphi}$ is the function m on $2^{V_{\mathcal{U}}}$ defined by $m(T_{\mathcal{U}}(\varphi)) = 1$ and $m(A) = 0$ if $A \neq T_{\mathcal{U}}(\varphi)$. In particular, $bel_{\mathcal{U}}^{\top}$ is the vacuous belief about $\mathcal{U}$.

the combined beliefs are exactly the same) and this means unnecessary nonmonotonicity.

In order to reduce the unnecessary nonmonotonicity, I suggest to choose the jba which minimizes the conflict (with this choice the monotonicity is assured if it is admissible). If this is not unique, it seems natural to me to choose the least specific one. This is the jba whose respective combination of beliefs maximizes the well established measure of nonspecificity (see for instance Klir and Wierman [7]) among the combinations with respect to the jbas with minimal conflict.

Definition 3 *If bel is a belief about $\mathcal{U}$ with bba m, the measure of nonspecificity of bel is the quantity*

$$N(bel) = \sum_{A \neq \emptyset} m(A) \log_2 |A| .$$

Thus if $c_{\min}^{m_1,m_2} < 1$, I suggest to choose as $m_1 \star m_2$ a jba $\underline{m}$ maximizing $N\big(bel_1 \otimes_{\underline{m}} bel_2\big)$ among the $\underline{m} \in \underline{\mathcal{M}}_{\mathcal{U}}^{m_1,m_2}$ with $c(\underline{m}) = c_{\min}^{m_1,m_2}$ (if $c_{\min}^{m_1,m_2} = 1$, the choice of a jba is useless, since anyway we cannot combine bel_1 and bel_2). From Proposition 3 follows that the task of finding such a $\underline{m}$ is a problem of linear programming.[6]

Proposition 3 *If $m_1, m_2 \in \mathcal{M}_{\mathcal{U}}$, $c_{\min}^{m_1,m_2} < 1$ and $f : \mathbb{N} \longrightarrow \mathbb{R}$ with $f(0) < -|\mathcal{U}|$ and $f(n) = \log_2 n$ for all $n > 0$, then*

$$\arg \max_{\substack{\underline{m} \in \underline{\mathcal{M}}_{\mathcal{U}}^{m_1,m_2} \\ c(\underline{m}) = c_{\min}^{m_1,m_2}}} N\big(bel_1 \otimes_{\underline{m}} bel_2\big) = \arg \max_{\underline{m} \in \underline{\mathcal{M}}_{\mathcal{U}}^{m_1,m_2}} \sum_{A,B \subseteq V_{\mathcal{U}}} \underline{m}(A,B)\, f(|A \cap B|) .$$

Proof. Let $F(\underline{m}) = \sum_{A,B \subseteq V_{\mathcal{U}}} \underline{m}(A,B)\, f(|A \cap B|)$. If $c(\underline{m}) = c_{\min}^{m_1,m_2}$, then

$$F(\underline{m}) = c_{\min}^{m_1,m_2} f(0) + \big(1 - c_{\min}^{m_1,m_2}\big)\, N\big(bel_1 \otimes_{\underline{m}} bel_2\big) .$$

Therefore it suffices to show that if $\underline{m}$ maximizes $F(\underline{m})$, then $c(\underline{m}) = c_{\min}^{m_1,m_2}$. In the proof of Proposition 2 it is shown that $c(\underline{m}) = c_{\min}^{m_1,m_2}$ is implied by the following property: if $A_1, A_2, B_1, B_2 \subseteq V_{\mathcal{U}}$ with $A_1 \cap B_1 = \emptyset$, $A_1 \cap B_2 \neq \emptyset$, $A_1 \neq A_2$, $\underline{m}(A_1,B_1) > 0$ and $\underline{m}(A_2,B_2) > 0$, then $A_2 \cap B_1 = \emptyset$ and $A_2 \cap B_2 \neq \emptyset$.

Assume that $\underline{m}$ maximizes $F(\underline{m})$ and consider the transformation $\underline{m} \longmapsto \underline{m}'$ defined in the proof of Proposition 2. If the hypothesis of the property stated above holds, we have

$$\begin{aligned} F\big(\underline{m}'\big) &= F(\underline{m}) + \varepsilon\big(f(|A_1 \cap B_2|) + f(|A_2 \cap B_1|) - f(|A_1 \cap B_1|) - f(|A_2 \cap B_2|)\big) > \\ &> F(\underline{m}) + \varepsilon\big(f(|A_2 \cap B_1|) + |\mathcal{U}| - f(|A_2 \cap B_2|)\big) . \end{aligned}$$

[6] The proof of Proposition 3 suggests an iteration algorithm for solving this problem: start for instance from $\underline{m}_D$ and recursively apply a transformation of the form $\underline{m} \longmapsto \underline{m}'$ in order to increase the value of the linear functional $\sum \underline{m}(A,B)\, f(|A \cap B|)$. I have not studied the properties of such an algorithm yet.

Therefore $F(\underline{m}') \leq F(\underline{m})$ implies $f(|A_2 \cap B_1|) < 0$ and $f(|A_2 \cap B_2|) \geq 0$; that is, $A_2 \cap B_1 = \emptyset$ and $A_2 \cap B_2 \neq \emptyset$. □

The least specific jba minimizing the conflict is not always unique, thus $m_1 \star m_2$ is not always defined. Consider first the set $\mathcal{S}$ of pairs (m_1, m_2) for which the operator $\star$ is defined: the following properties can be easily verified. In $\mathcal{S}$ the operator $\star$ satisfies the three basic requirements stated above (notice that if $(m_1, m_2) \in \mathcal{S}$, then $(m_1 \circ f, m_2 \circ f) \in \mathcal{S}$ and $(m_1', m_2') \in \mathcal{S}$). If $(m_1, m_2) \in \mathcal{S}$, then $(m_2, m_1) \in \mathcal{S}$ and $m_1 \star m_2\,(A,B) = m_2 \star m_1\,(B,A)$ for all $A, B \subseteq V_{\mathcal{U}}$. If $m \in \mathcal{M}_{\mathcal{U}}$, then $(m,m) \in \mathcal{S}$ and $m \star m\,(A,A) = m(A)$ for all $A \subseteq V_{\mathcal{U}}$. The last two properties imply commutativity and idempotency for the respective combinations of beliefs.

Commutativity is a necessary requirement in symmetrical situations where the two sources of information have the same importance and credibility. In other situations we can prefer that one of the two beliefs has a prominent role in the combination: since these cases can be worked out with other methods (such as discounting), I shall consider commutativity as necessary.

For any pair of bbas $m_1, m_2 \in \mathcal{M}_{\mathcal{U}}$, the least specific jbas minimizing the conflict form a convex polytope (i.e. the bounded intersection of a finite number of closed half-spaces) in $\mathbb{R}^{2^{2|V_{\mathcal{U}}|}}$. Therefore the completion of the definition of the operator $\star$ consists in a rule for assigning to every pair of bbas a point of the respective convex polytope, in such a way that commutativity and the first two basic requirements remain satisfied (the third one being trivially satisfied). Symmetry considerations could lead to the choice of the centre of the polytope (that is, the barycentre with respect to the uniform mass density): this choice fulfills the requirements. Another possibility fulfilling them is for instance the selection of the point of the polytope which minimizes the Euclidean distance from $\underline{m}_D$. I think that the choice of a rule should be based not only on its theoretical properties, but also on considerations about the computational complexity of possible implementations of this rule; since I have not analysed this aspect yet, I leave the question of the completion of the definition of $\star$ open. The contents of the rest of this paper are independent of any particular completion of this definition (such that the above requirements are fulfilled): simply let $\star$ be the obtained operator and let $\odot$ be the respective rule for the combination of beliefs.

Both rules $\oplus$ and $\odot$ satisfy the three basic requirements (to be precise, the corresponding operators satisfy them) and commutativity; $\oplus$ is associative, while $\odot$ is idempotent. Dempster's one is perhaps the only rule of the form $\otimes$ with the four common properties and associativity;[7] anyway, Example 1 shows that associativity and idempotency are two incompatible properties for rules of this form, even if we abandon every other assumption.

[7] The axiomatic derivations of Dempster's rule in Klawonn and Schwecke [5] and Smets [9] do not allow an answer to this question, since both sets of axioms contain a property which is stronger than the ones considered here; while Klawonn and Smets [6] consider a framework which is more restrictive than the one used here.

Example 1 *Let $q \in \mathcal{U}$ and $\frac{1}{2} < \alpha < 1$. Let bel_1 and bel_2 be the minimal beliefs with $bel_1(q) = \alpha$ and $bel_2(\neg q) = \alpha$, respectively. That is, $m_1(Q) = m_2\left(\overline{Q}\right) = \alpha$ and $m_1(V_{\mathcal{U}}) = m_2(V_{\mathcal{U}}) = 1 - \alpha$, with $Q = T_{\mathcal{U}}(q)$ and $\overline{Q} = V_{\mathcal{U}} \backslash Q$.*

Then the bba m associated with $bel_1 \otimes bel_2$ satisfies $m(Q) = m\left(\overline{Q}\right) = \beta$ and $m(V_{\mathcal{U}}) = 1 - 2\beta$, for a β such that $0 \leq \beta \leq \frac{1}{2}$ (the value of β depends on the choice of a jba).

If we assume idempotency and associativity, we obtain

$$bel_1 \otimes bel_2 = (bel_1 \otimes bel_1) \otimes bel_2 = bel_1 \otimes (bel_1 \otimes bel_2).$$

That is, there is a jba $\underline{m} \in \underline{\mathcal{M}}_{\mathcal{U}}^{m_1,m}$ with

$$m\left(\overline{Q}\right) = \frac{\underline{m}\left(V_{\mathcal{U}}, \overline{Q}\right)}{1 - c(\underline{m})} = \frac{m\left(\overline{Q}\right) - \underline{m}\left(Q, \overline{Q}\right)}{1 - \underline{m}\left(Q, \overline{Q}\right)}.$$

Therefore $c(\underline{m}) = \underline{m}\left(Q, \overline{Q}\right) = 0$, and from Proposition 1 follows that

$$\beta = bel_1 \otimes bel_2(q) \geq bel_1(q) = \alpha,$$

which is a contradiction to $\beta \leq \frac{1}{2} < \alpha$. Thus idempotency and associativity are incompatible (if $|\mathcal{U}| \geq 1$).

In order to combine two beliefs without assuming the independence of the sources, I suggest the rule $\odot$. This can be considered as the most conservative rule of the form $\otimes$: it conserves as much as possible of both beliefs (it has minimal conflict, i.e. maximal monotonicity) without adding anything (it has minimal specificity among the rules with minimal conflict). It is idempotent, thus it cannot be associative. It can be easily verified (for instance by considering epistemic probabilities, defined in Example 3) that associativity is incompatible also with the minimization of the conflict (which is the basic feature of the rule $\odot$).

Idempotency is only a particular case of the following property of the rule $\odot$: if bel_2 is a specialization of bel_1 (i.e. m_2 can be obtained through redistribution of $m_1(A)$ to the non-empty sets $B \subseteq A$, for all $A \subseteq V_U$), then $bel_1 \odot bel_2 = bel_2$. This property is important if strong dependence is possible: if bel_2 is a specialization of bel_1, the information encoded by bel_1 can be part of the information encoded by bel_2, in which case the result of pooling the information is actually bel_2.

Associativity is important because (with commutativity) it implies that the result of the combination of n beliefs is independent of the order in which these beliefs are combined. In a sense, this independence of the order can be obtained also for the rule $\odot$: if we have to combine n beliefs simultaneously, we can consider the set of n-dimensional jbas and extend our rule for the selection of a jba to the n-dimensional case. An interesting problem could be the search for an analogue of Proposition 2 for the n-dimensional case.

Example 2 and Example 3 illustrate the differences between the two rules $\oplus$ and $\odot$ in two simple situations.

Example 2 *Consider the situation of Example 1. Since bel_1 and bel_2 are not compatible, the monotonicity of their combination is not admissible. In fact, for both $\varphi \in \{q, \neg q\}$ we have $\max(bel_1, bel_2)(\varphi) = \alpha > \frac{1}{2}$, while $bel_1 \otimes bel_2(\varphi) = \beta \leq \frac{1}{2}$. Using $\oplus$ we obtain $\beta = \frac{\alpha}{\alpha+1} < \frac{1}{2}$, whereas using $\odot$ we obtain $\beta = \frac{1}{2}$. Thus, unlike the rule $\oplus$, the rule $\odot$ allows only the necessary nonmonotonicity.*

In Example 1 we have seen that no rule of the form $\otimes$ can satisfy both equations $bel_1 \otimes bel_1 = bel_1$ and $(bel_1 \otimes bel_1) \otimes bel_2 = bel_1 \otimes (bel_1 \otimes bel_2)$. Obviously, $\oplus$ satisfies the second one, whereas $\odot$ satisfies the first one. If we want to combine the three beliefs of the second equation in a unique way with the rule $\odot$, we can extend it to the 3-dimensional case. The 3-dimensional jba minimizing the conflict is unique and the respective combination of the three beliefs is the one that we obtain by using the rule $\odot$ in the left-hand side of the equation: $bel_1 \odot bel_2$.

Example 3 *The beliefs bel_1 and bel_2 considered in Example 2 are consonant. In some senses, at the opposite extreme from consonant beliefs we find the epistemic probabilities. A belief about $\mathcal{U}$ with bba m is an epistemic probability if $m(A) = 0$ for all $A \subseteq V_{\mathcal{U}}$ with $|A| \neq 1$. Such a belief is completely defined by the $r = |V_{\mathcal{U}}|$ values $p_1, \ldots, p_r$ that m assigns to the $A \subseteq V_{\mathcal{U}}$ with $|A| = 1$ (it suffices to decide an order for the elements of $V_{\mathcal{U}}$).*

Let bel_1 and bel_2 be two epistemic probabilities defined by $p_1^{(1)}, \ldots, p_r^{(1)}$ and $p_1^{(2)}, \ldots, p_r^{(2)}$, respectively. Then their combination $bel_1 \otimes bel_2$ is still an epistemic probability; let it be defined by $p_1, \ldots, p_r$. The monotonicity is admissible only if $bel_1 = bel_2$, and to assure this monotonicity a rule must be idempotent. Using $\oplus$ we obtain that $p_i = b p_i^{(1)} p_i^{(2)}$ for each $i \in \{1, \ldots, r\}$, where $b \geq 1$ is a normalizing constant. Using $\odot$ we obtain that $p_i = c \min\left\{p_i^{(1)}, p_i^{(2)}\right\} \geq \min\left\{p_i^{(1)}, p_i^{(2)}\right\}$ for each $i \in \{1, \ldots, r\}$, where $c \geq 1$ is a normalizing constant (notice that the inequality is strict unless $bel_1 = bel_2$).

If we want to simultaneously combine n epistemic probabilities defined, respectively, by $p_1^{(j)}, \ldots, p_r^{(j)}$ (for each $j \in \{1, \ldots, n\}$), we can easily extend the rule $\odot$ to the n-dimensional case. The result of the combination is the epistemic probability defined by $p_1, \ldots, p_r$, with $p_i = d \min\left\{p_i^{(1)}, \ldots, p_i^{(n)}\right\}$ for each $i \in \{1, \ldots, r\}$, where $d \geq 1$ is a normalizing constant.

5 A Generalization of Bayes' Theorem

Now I present a situation in which a combination rule minimizing the conflict is especially sensible and in which we can get many results without need to consider the whole combination of beliefs: it suffices to know the value of the conflict between them (which for a combination rule minimizing the conflict can be determined thanks to Proposition 2).

Consider a hypothesis h implying a belief bel about $\mathcal{U}$ (with $h \notin \mathcal{U}$). If we have a belief $bel_{\mathcal{H}}$ about $\mathcal{H} = \{h\}$, we can combine these two beliefs in the following way. We first expand bel to the belief about $\mathcal{U}' = \mathcal{U} \cup \mathcal{H}$ which contains nothing more than the implication $h \Rightarrow bel$: let $(h \Rightarrow bel)$ be the minimal belief about $\mathcal{U}'$ assigning for all $\varphi \in \mathcal{L}_{\mathcal{U}}$ the value $bel(\varphi)$ to the formula $h \to \varphi$.[8] Then we can combine $(h \Rightarrow bel)$ with the vacuous extension of $bel_{\mathcal{H}}$ to $\mathcal{L}_{\mathcal{U}'}$, obtaining

$$bel_{\mathcal{H}} \upharpoonleft^{\mathcal{U}'} \oplus (h \Rightarrow bel)\,.$$

The use of Dempster's rule is justified in the sense that this is only a formal construction to apply a "metabelief" $bel_{\mathcal{H}}$ about $\mathcal{H}$ to the consequence bel of the hypothesis h (in particular, there can be no conflict). The resulting belief about $\mathcal{U}$ is

$$\left(bel_{\mathcal{H}} \upharpoonleft^{\mathcal{U}'} \oplus (h \Rightarrow bel)\right) \downharpoonleft_{\mathcal{U}} = bel_{\mathcal{H}}(h)\, bel + (1 - bel_{\mathcal{H}}(h))\, bel_{\mathcal{U}}^{\top};$$

that is, the discounting of bel with discount rate $1 - bel_{\mathcal{H}}(h)$. This is sensible, since $pl_{\mathcal{H}}(\neg h) = 1 - bel_{\mathcal{H}}(h)$ measures the amount of our uncertainty about the hypothesis h.

If we get some information in the form of a belief bel' about $\mathcal{U}$, we can combine its vacuous extension to $\mathcal{L}_{\mathcal{U}'}$ with our belief about $\mathcal{U}'$, obtaining in particular a new belief $bel'_{\mathcal{H}}$ about $\mathcal{H}$:

$$bel'_{\mathcal{H}} = \left(\left(bel_{\mathcal{H}} \upharpoonleft^{\mathcal{U}'} \oplus (h \Rightarrow bel)\right) \otimes_{\underline{m}} bel' \upharpoonleft^{\mathcal{U}'}\right) \downharpoonleft_{\mathcal{H}}\,.$$

Thus in order to get $bel'_{\mathcal{H}}$, we must choose a jba $\underline{m}$. If we reason on the form of the marginal bbas, we can see that $\underline{m}$ is sensible only if it is "naturally" based on a jba $\underline{m}_h$ for the combination of bel and bel'.[9] Then $c(\underline{m}) = bel_{\mathcal{H}}(h)\, c(\underline{m}_h)$, so the combination is possible unless we are sure of the hypothesis and this totally conflicts with the new information (i.e. $bel_{\mathcal{H}}(h) = 1$ and $c(\underline{m}_h) = 1$). The changes in the belief about $\mathcal{H}$ are entirely determined by the conflict $c(\underline{m}_h)$:

$$bel'_{\mathcal{H}}(h) = \frac{bel_{\mathcal{H}}(h) - c(\underline{m})}{1 - c(\underline{m})} \leq bel_{\mathcal{H}}(h) \text{ and}$$

$$bel'_{\mathcal{H}}(\neg h) = \frac{bel_{\mathcal{H}}(\neg h)}{1 - c(\underline{m})} \geq bel_{\mathcal{H}}(\neg h)\,.$$

[8] If m is the bba associated with bel, then the bba associated with $(h \Rightarrow bel)$ is the function m' on $2^{V_{\mathcal{U}'}}$ defined by $m'(T_{\mathcal{U}'}(h \to \varphi)) = m(T_{\mathcal{U}}(\varphi))$ for all $\varphi \in \mathcal{L}_{\mathcal{U}}$, and $m'(A) = 0$ if $A \notin T_{\mathcal{U}'}(\{h \to \varphi : \varphi \in \mathcal{L}_{\mathcal{U}}\})$.

[9] If $m_{\mathcal{H}}$ and m' are the bbas associated with $bel_{\mathcal{H}}$ and bel', respectively, then $\underline{m}$ is the jba which satisfies (for all $\varphi, \psi \in \mathcal{L}_{\mathcal{U}}$)

$$\begin{aligned}
\underline{m}(T_{\mathcal{U}'}(h \wedge \varphi), T_{\mathcal{U}'}(\psi)) &= m_{\mathcal{H}}(T_{\mathcal{H}}(h))\, \underline{m}_h(T_{\mathcal{U}}(\varphi), T_{\mathcal{U}}(\psi)),\\
\underline{m}(T_{\mathcal{U}'}(h \to \varphi), T_{\mathcal{U}'}(\psi)) &= m_{\mathcal{H}}(V_{\mathcal{H}})\, \underline{m}_h(T_{\mathcal{U}}(\varphi), T_{\mathcal{U}}(\psi)) \text{ and}\\
\underline{m}(T_{\mathcal{U}'}(\neg h), T_{\mathcal{U}'}(\psi)) &= m_{\mathcal{H}}(T_{\mathcal{H}}(\neg h))\, m'(T_{\mathcal{U}}(\psi))\,.
\end{aligned}$$

If $0 < bel_{\mathcal{H}}(h) < 1$, then $bel'_{\mathcal{H}}(h)$ is a strictly decreasing function of $c(\underline{m}_h)$, and in particular we maintain our belief in h only if $c(\underline{m}_h) = 0$. Thus $c(\underline{m}_h)$ (that is, the conflict between the implications of the hypothesis h and the new information) is clearly a measure of disagreement. Therefore it is especially sensible to choose a jba $\underline{m}_h$ minimizing the conflict (and if we are only interested in the new belief about $\mathcal{H}$, then knowing the minimal conflict suffices). With such a choice we obtain in particular that if bel and bel' are compatible, then we maintain our belief in h (this is in general not true if $\underline{m}_h = \underline{m}_D$, even if $bel = bel'$).

Consider now the general case with n hypotheses $h_1, \ldots, h_n$ implying, respectively, the beliefs $bel_1, \ldots, bel_n$ about $\mathcal{U}$ (with $h_1, \ldots, h_n \notin \mathcal{U}$). Given an "a priori" belief $bel_{\mathcal{H}}$ about $\mathcal{H} = \{h_1, \ldots, h_n\}$ and an "observation" belief bel' about $\mathcal{U}$, we can combine these beliefs to obtain an "a posteriori" belief $bel'_{\mathcal{H}}$ about $\mathcal{H}$:

$$bel'_{\mathcal{H}} = \left(\left(bel_{\mathcal{H}} \uparrow^{\mathcal{U}'} \oplus \bigotimes_{i=1}^{n} (h_i \Rightarrow bel_i) \uparrow^{\mathcal{U}'} \right) \otimes_{\underline{m}} bel' \uparrow^{\mathcal{U}'} \right) \downarrow_{\mathcal{H}} .$$

As before, $\mathcal{U}' = \mathcal{U} \cup \mathcal{H}$ and the use of Dempster's rule in the first combination can be justified as a formal construction. The new element is $\bigotimes_{i=1}^{n} (h_i \Rightarrow bel_i) \uparrow^{\mathcal{U}'}$, which is any combination of the n beliefs $(h_i \Rightarrow bel_i) \uparrow^{\mathcal{U}'}$ using the general rule $\otimes$ (we can obtain it by $n-1$ applications of the binary rule or with a n-dimensional jba). This allows the hypotheses to be dependent (for instance if two hypotheses differ only by a detail and the two implied beliefs are almost the same), and it is important to notice that anyway there can be no conflict among the $n+1$ beliefs $bel_{\mathcal{H}} \uparrow^{\mathcal{U}'}$ and $(h_i \Rightarrow bel_i) \uparrow^{\mathcal{U}'}$.

This way to update a belief about $\mathcal{H}$ is a broad generalization of Bayes' theorem for epistemic probabilities and of Smets' generalized Bayesian theorem (gBt) for normalized beliefs (see for instance Smets [11]). The construction of $\bigotimes_{i=1}^{n} (h_i \Rightarrow bel_i) \uparrow^{\mathcal{U}'}$ allows a lot of freedom, which of course can be limited by some additional assumptions. Before introducing two such assumptions, I consider a simple special case.

Let $bel_{\mathcal{H}}$ be a belief about $\mathcal{H}$ satisfying

$$\sum_{i=0}^{n} bel_{\mathcal{H}}(\varphi_i) = 1, \text{ where}$$
$$\varphi_0 = \neg h_1 \wedge \ldots \wedge \neg h_n \text{ and}$$
$$\varphi_i = \neg h_1 \wedge \ldots \wedge \neg h_{i-1} \wedge h_i \wedge \neg h_{i+1} \ldots \wedge \neg h_n \text{ if } i \in \{1, \ldots, n\};$$

that is, $bel_{\mathcal{H}}$ is an epistemic probability on $\varphi_0, \ldots, \varphi_n$. Then

$$bel_{\mathcal{H}} \uparrow^{\mathcal{U}'} \oplus \bigotimes_{i=1}^{n} (h_i \Rightarrow bel_i) \uparrow^{\mathcal{U}'}$$

is independent of the choice of $\bigotimes_{i=1}^{n}(h_i \Rightarrow bel_i)\upharpoonleft^{\mathcal{U}'}$, and its restriction to $\mathcal{L}_{\mathcal{U}}$ is

$$\sum_{i=1}^{n} bel_{\mathcal{H}}(h_i)\, bel_i + bel_{\mathcal{H}}(\varphi_0)\, bel_{\mathcal{U}}^{\top}.$$

This shows that φ_0 can be considered as an additional hypothesis h_0 implying the vacuous belief $bel_0 = bel_{\mathcal{U}}^{\top}$, and $bel_{\mathcal{H}}$ can be seen as an epistemic probability on the mutually exclusive and exhaustive hypotheses $h_0, \ldots, h_n$. As before, in order to get $bel'_{\mathcal{H}}$, we must choose a jba $\underline{m}$. And as before, if we reason on the form of the marginal bbas, we can see that $\underline{m}$ is sensible only if it is "naturally" based on the jbas $\underline{m}_i$ of the combinations of bel_i and bel' (for each $i \in \{0, \ldots, n\}$).[10] Then $c(\underline{m}) = \sum_{i=1}^{n} bel_{\mathcal{H}}(h_i)\, c(\underline{m}_i)$ (notice that $c(\underline{m}_0) = 0$), so the combination is possible unless we are sure that the truth is among some hypotheses and these totally conflict with the new information. The belief about $\mathcal{H}$ remains an epistemic probability on $\varphi_0, \ldots, \varphi_n$, and as before, the changes are entirely determined by the conflicts $c(\underline{m}_1), \ldots, c(\underline{m}_n)$:

$$bel'_{\mathcal{H}}(h_i) = \frac{1 - c(\underline{m}_i)}{1 - c(\underline{m})} bel_{\mathcal{H}}(h_i) \text{ for each } i \in \{0, \ldots, n\}.$$

Thus the belief in a hypothesis h_i increases if and only if the respective conflict $c(\underline{m}_i)$ is less than $c(\underline{m})$, which is a weighted average of the conflicts of the $n+1$ hypotheses (h_0 included). Therefore the conflicts $c(\underline{m}_1), \ldots, c(\underline{m}_n)$ measure the disagreement between the respective hypotheses and the new information, and thus it is especially sensible to choose jbas $\underline{m}_1, \ldots, \underline{m}_n$ minimizing the conflict. With such a choice we obtain in particular that if bel_i and bel' are compatible, then the belief in h_i does not decrease (and it increases if $c(\underline{m}) > 0$); this is in general not true if $\underline{m}_i = \underline{m}_D$, even if $bel_i = bel'$.

I now consider the two announced assumptions which limit the freedom in the construction of $\bigotimes_{i=1}^{n}(h_i \Rightarrow bel_i)\upharpoonleft^{\mathcal{U}'}$. The first one is that the hypotheses $h_1, \ldots, h_n$ are mutually exclusive (i.e. $bel_{\mathcal{H}}(\varphi_0 \vee \ldots \vee \varphi_n) = 1$), but not necessarily exhaustive (which would mean $bel_{\mathcal{H}}(\varphi_1 \vee \ldots \vee \varphi_n) = 1$). The second one is that the beliefs $bel_1, \ldots, bel_n$ are issued from independent sources of information (the sources can be identified with the respective hypotheses $h_1, \ldots, h_n$). Since the hypotheses are mutually exclusive, this simply means that the belief about $\mathcal{U}$ implied

[10] $\underline{m}$ is the jba which satisfies (for all $\varphi, \psi \in \mathcal{L}_{\mathcal{U}}$ and $i \in \{0, \ldots, n\}$)

$$\underline{m}\left(T_{\mathcal{U}'}(\varphi_i \wedge \varphi), T_{\mathcal{U}'}(\psi)\right) = bel_{\mathcal{H}}(\varphi_i)\, \underline{m}_i\left(T_{\mathcal{U}}(\varphi), T_{\mathcal{U}}(\psi)\right).$$

by a disjunction of hypotheses is the disjunctive combination of the respective beliefs (the disjunctive rule of combination is defined for instance in Smets [11]). With these two additional assumptions, we obtain

$$bel'_{\mathcal{H}} = \left(\left(bel_{\mathcal{H}} \upharpoonleft^{\mathcal{U}'} \oplus \bigoplus_{i=1}^{n} (h_i \Rightarrow bel_i) \upharpoonleft^{\mathcal{U}'} \right) \otimes_{\underline{m}} bel' \upharpoonleft^{\mathcal{U}'} \right) \downharpoonleft_{\mathcal{H}} .$$

This is a generalization of Smets' gBt for normalized beliefs (which corresponds to the special case with $\underline{m} = \underline{m}_D$), and thus also of Bayes' theorem for epistemic probabilities. If *bel'* has the form $bel^{\varphi}_{\mathcal{U}}$ (in the literature the gBt is usually restricted to this case), the jba $\underline{m}$ is unique and the updated belief $bel'_{\mathcal{H}}$ is the one that we would obtain by applying the gBt to the $n+1$ hypotheses $h_0, \ldots, h_n$ (with $bel_0 = bel^{\top}_{\mathcal{U}}$). But if *bel'* has not the form $bel^{\varphi}_{\mathcal{U}'}$, we must choose a jba $\underline{m}$; and as before, $\underline{m}$ can be sensible only if it is "naturally" based on the jbas of the combinations of the new information *bel'* with the beliefs implied by the hypotheses or by any disjunction of hypotheses. Since also in this more general case the conflicts measure the disagreement between the respective hypotheses (or disjunctions of hypotheses) and the new information, it is especially sensible to choose jbas minimizing the conflict. With such a choice we obtain in particular that if the beliefs implied by some hypotheses are compatible with the new information, then the values of the belief in these hypotheses and in their disjunctions do not decrease (and they increase if $c(\underline{m}) > 0$). If instead we use Dempster's rule (that is, we use the gBt), we can get very bad results, since the conflict between the new information *bel'* and a hypothesis *h* implying the belief *bel* can be very high, even if $bel' = bel$ (i.e. the prevision of *h* is perfect). In fact, if a hypothesis is correct, can we assume that the belief which is a theoretical consequence of the hypothesis and the belief which is a practical consequence of the correctness of the hypothesis are independent?

6 Conclusion

In this paper a rule has been proposed to combine two belief functions issued from sources of information whose independence is doubtful. This rule increases the monotonicity of the reasoning, assuring in particular complete monotonicity if this is admissible. The proposed combination rule is commutative and idempotent. It is not associative, but it can be easily extended to a rule for the simultaneous combination of any number of belief functions.

The proposed combination rule leads to sensible results in a generalization of Bayes' theorem for epistemic probabilities and of Smets' generalized Bayesian theorem. This generalization allows the new information to be any belief function: in this situation the use of Dempster's rule (that is, the independence assumption) leads to questionable results.

References

[1] B. Ben Yaghlane, P. Smets and K. Mellouli. Belief Function Independence: I. The Marginal Case. *International Journal of Approximate Reasoning*, 29:47–70, 2002.

[2] A. P. Dempster. Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics*, 38:325–339, 1967.

[3] D. Dubois and H. Prade. On the Unicity of Dempster Rule of Combination. *International Journal of Intelligent Systems*, 1:133–142, 1986.

[4] R. L. Epstein. *Propositional Logics*. The Semantic Foundations of Logic, 2nd edition, Wadsworth, 2001.

[5] F. Klawonn and E. Schwecke. On the Axiomatic Justification of Dempster's Rule of Combination. *International Journal of Intelligent Systems*, 7:469–478, 1992.

[6] F. Klawonn and P. Smets. The Dynamic of Belief in the Transferable Belief Model and Specialization-Generalization Matrices. In Dubois, Wellman, D'Ambrosio and Smets, editors, *Proceedings of the 8th Conference on Uncertainty in Artificial Intelligence*, 130–137, Morgan Kaufmann, 1992.

[7] G. J. Klir and M. J. Wierman. *Uncertainty-Based Information: Elements of Generalized Information Theory*. Studies in Fuzziness and Soft Computing, 15, 2nd edition, Physica-Verlag, 1999.

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

[9] P. Smets. The Combination of Evidence in the Transferable Belief Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:447–458, 1990.

[10] P. Smets. The Concept of Distinct Evidence. In Bouchon-Meunier, Valverde and Yager, editors, *Proceedings of the 4th Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 789–794, Lecture Notes in Computer Science, 682, Springer, 1993.

[11] P. Smets. Belief Functions: The Disjunctive Rule of Combination and the Generalized Bayesian Theorem. *International Journal of Approximate Reasoning*, 9:1–35, 1993.

[12] R. R. Yager. Nonmonotonic Reasoning with Belief Structures. In Yager, Kacprzyk and Fedrizzi, editors, *Advances in the Dempster-Shafer Theory of Evidence*, Chapter 24, 533–554, Wiley, 1994.

Marco E.G.V. Cattaneo is with the Seminar for Statistics, Department of Mathematics, ETH Zürich, Switzerland. E-mail: cattaneo@stat.math.ethz.ch

Nonparametric Predictive Comparison of Two Groups of Lifetime Data

F.P.A. COOLEN
University of Durham, United Kingdom

K.J. YAN
University of Durham, United Kingdom

Abstract

We present the application of a recently introduced nonparametric predictive inferential method to compare two groups of data, consisting of observed event times and right-censoring times. Comparison is based on imprecise probabilities concerning one future observation per group.

Keywords

censored data, exchangeability, nonparametrics, prediction, survival analysis

1 Introduction

We apply a recently introduced method for statistical inference, called 'nonparametric predictive inference' (NPI) [1, 6], to the problem of comparing two groups of data, or, if one prefers to use such terminology, two underlying populations, where the data include right-censored observations. This generalizes the results presented by Coolen [3], who did not allow censoring. Right-censoring typically occurs in study of event times, e.g. survival times of patients in medical applications, or periods without failures of technical systems in reliability engineering, where a right-censoring at a time t just implies that the event of interest has not yet happened before or at time t. Throughout, we assume that no further information is available about the random quantities corresponding to right-censored observations, an assumption often called 'noninformative censoring' [6, 11, 13]. We also assume that the two populations compared are independent, in the sense that any information about the random quantities from one population does not influence our inferences on random quantities from the other population.

The method of statistical inference used here is based on quite minimal modelling assumptions, and is directly in terms of random quantities representing future observations. We assume that either a well-specified event happens, at a particular time, to each item for which, or individual for who, we have an observation, or that a time is reported at which such an event has not yet occurred. All data are referred to as 'observation (time)', if it is a time at which the event of interest actually occurred we call it 'event (time)', else '(right-)censoring (time)'. Speaking in terms of 'time', we restrict attention to non-negative random quantities, so to random quantities and observations on the time-axis $[0,\infty)$. However, the method presented is more widely applicable, as only a finite partition of (part of) the real line is required.

In Section 2, the basics of nonparametric predictive inference are briefly summarized. Section 3 presents the main result on predictive comparison of two groups of lifetime data, which is illustrated, and briefly compared with an alternative nonparametric method, via two examples in Section 4. For ease of notation, we assume that there are no ties of any kind in the data, so no two observations are equal. In Section 5, we briefly discuss how the method can be adapted for dealing with tied observations, and we add a few concluding remarks about the presented method and results, including some attention to when this method might be used.

2 Nonparametric predictive inference

In this section, we summarize NPI for data including right-censored observations, as recently presented by Coolen and Yan [6], to which we refer for the theoretical justification and further detailed discussion of this method.

Let a single group of data consist of n observations, of which u are event times, $0 < t_1 < \ldots < t_u$, and $v = n - u$ right-censoring times, $0 < c_1 < \ldots < c_v$. Let $t_0 = 0$ and $t_{u+1} = \infty$, and let the right-censoring times in (t_i, t_{i+1}) be $c_1^i < \ldots < c_{l_i}^i$. We assume that there are no ties among the data, the method is easily adapted for ties [6]. Let $\tilde{n}_t$ be the number of items with observation time greater than or equal to t. We call this the number of items 'at risk just prior to time t', at an observation time the corresponding item is included in $\tilde{n}_t$.

Based on such data, Coolen and Yan [6] introduce, and justify, the assumption 'right-censoring $A_{(n)}$' (rc-$A_{(n)}$) for NPI, for the random quantity X_{n+1} representing the lifetime of a future item, or the survival time of a future individual. Right-censoring $A_{(n)}$ generalizes Hill's $A_{(n)}$ [7], which underlies NPI if the data do not include right-censored observations [1, 3]. Description of rc-$A_{(n)}$ requires notation for partial specification of probability distributions, called 'M-function'.

Definition 1 (M-function) *[6]*
A partial specification of a probability distribution for a real-valued random quantity X can be provided via probability masses assigned to intervals, without any further restriction on the spread of the probability mass within each interval. A probability mass assigned, in such a way, to an interval (a,b), is denoted by $M_X(a,b)$, and referred to as M-function value for X on (a,b).

Clearly, all M-function values for X on all intervals should sum up to one, and each M-function value should be in $[0,1]$.

Definition 2 (rc-$A_{(n)}$) *[6]*
The assumption 'right-censoring $A_{(n)}$' (rc-$A_{(n)}$) is that the probability distribution for a nonnegative random quantity X_{n+1}, based on u event times and v right-censoring times, as described above, is partially specified by ($i = 0,\ldots,u$; $k = 1,\ldots,l_i$)

$$\begin{aligned} M_{X_{n+1}}(t_i,\ t_{i+1}) &= \frac{1}{n+1}\prod_{\{r:\,c_r<t_i\}}\frac{\tilde{n}_{c_r}+1}{\tilde{n}_{c_r}}, \\ M_{X_{n+1}}(c_k^i,\ t_{i+1}) &= \frac{1}{(n+1)\tilde{n}_{c_k^i}}\prod_{\{r:\,c_r<c_k^i\}}\frac{\tilde{n}_{c_r}+1}{\tilde{n}_{c_r}}. \end{aligned}$$

The product terms are defined as one if the product is taken over an empty set. The M-function values for X_{n+1} on other intervals are zero. This implicitly assumes non-informative censoring, as a post-data assumption related to exchangeability of all items known to be at risk at any time t, see Coolen and Yan [6], who also justify rc-$A_{(n)}$. We illustrate the M-function values in rc-$A_{(n)}$ via an example, followed by a brief explanation of the key ideas behind rc-$A_{(n)}$.

Example 1
Table 1 gives the data for group A which are part of Example 2 in Section 4, where the data are introduced in more detail. For this group, there are 10 observed event times and 6 right-censoring times. Table 1 also presents the M-function values, with corresponding intervals, according to rc-$A_{(n)}$ for these data.

These M-function values sum up to one (subject to a minor rounding effect), and illustrate the effects of right-censoring. Notice, for example, that there is some probability mass defined on each interval from a right-censoring time to the next observed event time, and that a right-censored observation also leads to larger M-function values between two later observed event times.

This assumption rc-$A_{(n)}$ is generalizing Hill's assumption $A_{(n)}$ [7], the idea is roughly as follows. If $n+1$ real-valued random quantities are exchangeable, and we assume that ties occur with probability zero, then the $n+1$-st of these random quantities has equal probability $1/(n+1)$ to fall in each of the intervals that form the partition created by the values of the other n random quantities, *before* any of these random quantities are actually observed. Hill [7] proposed this same property as a *posterior* predictive distribution, calling it $A_{(n)}$, and later he [8, 9] discussed further properties of this assumption and its use as an inferential procedure, and presented a prior process that leads to $A_{(n)}$ in the Bayesian framework (under finite additivity). Generally speaking, use of $A_{(n)}$ makes sense in case of very vague prior information, or indeed if one explicitly wishes not to use any

Table 1: Cervical cancer example ($> t$: right-censoring at t)

data		value
	$M(0,90)$	0.05882
90	$M(90,142)$	0.05882
142	$M(142,150)$	0.05882
150	$M(150,269)$	0.05882
269	$M(269,291)$	0.05882
291	$M(291,680)$	0.05882
>468	$M(468,680)$	0.00535
680	$M(680,837)$	0.06417
837	$M(837,1037)$	0.06417
>890	$M(890,1037)$	0.00802
1037	$M(1037,1297)$	0.07219
>1090	$M(1090,1297)$	0.01203
>1113	$M(1113,1297)$	0.01684
>1153	$M(1153,1297)$	0.02527
1297	$M(1297,1429)$	0.12634
1429	$M(1429,\infty)$	0.12634
>1577	$M(1577,\infty)$	0.12634

such prior information. Our generalization adopts the same idea for the situation of right-censored data, using the extra assumption that a right-censored item, at the moment the censoring takes place, had an exchangeable residual time till event with all those items for which the event had not yet taken place, and which had not been censored previously. This exchangeability at time of censoring is indeed a proper form of 'noninformative censoring', and the probabilities as specified by rc-$A_{(n)}$, via M-function values, for a single future observation are then derived via conditioning on possible values for the right-censored items. Further details of the derivation and justification of rc-$A_{(n)}$ are given by Yan [16] and Coolen and Yan [6].

Berliner and Hill [2] also presented the use of $A_{(n)}$ for right-censored data, but instead of adding an assumption to deal with the exact censoring information, they replaced each censored observation by just survival past the largest observed event time smaller than the censoring time, in which case no assumptions need to be added to $A_{(n)}$. This implies that at observed event times, our method coincides with the Berliner-Hill method, but these two methods differ in between event times if there are censoring times. In addition, Berliner and Hill assumed that the probability mass per interval is uniformly distributed (except for the last interval if there is no finite right-end point), whereas we use imprecise probabilities, as we discuss next.

It should be mentioned that, of course, imprecise probabilities have been used before for situations where not all data are complete, in the sense that not each event of interest has actually been observed. For example, Manski [12] considers the logical bounds on conditional probabilities based on censored samples alone. This would relate to our approach if we had not added any further assumption about the right-censored data, the novelty of rc-$A_{(n)}$ is the extra exchangeability-related assumption about the residual time till event for each censored observation, which has the effect of keeping imprecision relatively small, which is particularly useful if there are relatively many censored observations in the data set.

The partial specification of the probability distribution of X_{n+1}, via M-function values as specified by rc-$A_{(n)}$, enables NPI if the problems considered can be formulated in terms of a future observation X_{n+1}. However, for many problems of interest, the M-function values only imply bounds for predictive probabilities, where optimal bounds are imprecise probabilities [15].

As a consequence of the M-function values defined in rc-$A_{(n)}$, the events $\{X_{n+1} \in (t_i,\, t_{i+1})\}$, for $i = 0,\ldots,u$, have precise probabilities [6]

$$P(X_{n+1} \in (t_i,\, t_{i+1})) \;=\; M_{X_{n+1}}(t_i,\, t_{i+1}) + \sum_{k=1}^{l_i} M_{X_{n+1}}(c_k^i,\, t_{i+1}).$$

3 Comparing two groups of lifetime data

For the comparison of two groups of lifetime data we use the notation as introduced above, but consistently add an index a or b, corresponding to the groups which we call A and B. For example, for group A we have n_a observations, consisting of the event times $0 < t_{a,1} < \ldots < t_{a,u_a}$ and right-censoring times $0 < c_{a,1} < \ldots < c_{a,v_a}$, and the right-censoring times in the interval $(t_{a,i}, t_{a,i+1})$ are denoted by $c_{a,1}^i < \ldots < c_{a,l_{a,i}}^i$, et cetera. Throughout we assume that there are no ties at all among the observations (see Section 5), and that information on one group does not have any effect on probabilities of random quantities corresponding to the other group, so that X_{a,n_a+1} and X_{b,n_b+1} are independent and that data from group A does not influence our probabilities for X_{b,n_b+1}, and vice versa. We summarize this by stating that the groups are independent.

We require some additional notation, effectively counting the number of observed event times from group B to the left of observations from group A:

$$\begin{aligned} s_b(t_{a,i}) &= \#\{t_{b,j} \,|\, t_{b,j} < t_{a,i},\ j = 1,\ldots,u_b\}, \\ s_b(c_{a,k}^i) &= \#\{t_{b,j} \,|\, t_{b,j} < c_{a,k}^i,\ j = 1,\ldots,u_b\}, \end{aligned}$$

for $i = 1,\ldots,u_a$ and $k = 1,\ldots,l_{a,i}$. Similarly, we need notation for the number of right-censoring times from group B in the interval $(t_{b,s_b(t_{a,i})}, t_{a,i})$:

$$s_b^c(t_{a,i}) = \#\{c_{b,j} \,|\, c_{b,j} \in (t_{b,s_b(t_{a,i})}, t_{a,i}),\ j = 1,\ldots,u_b\},$$

for $i = 1, \dots, u_a + 1$.

The main results of this paper, namely the lower and upper probabilities for events $X_{a,n_a+1} > X_{b,n_b+1}$, based on the assumptions rc-$A_{(n_a)}$ and rc-$A_{(n_b)}$, are presented as a theorem below. The proof of the theorem is simplified via a lemma, which we present first, and which justifies the use of a variety of the theorem of total probability with conditioning on nested intervals, with probability distributions partially specified via M-function values.

Lemma 1 *For $s \geq 2$, let $J_l = (j_l, r)$, with $j_1 < j_2 < \ldots < j_s < r$, so we have nested intervals $J_1 \supset J_2 \supset \ldots \supset J_s$ with the same right end-point r (which may be infinity). We consider two independent real-valued random quantities, say X and Y. Let the probability distribution for X be partically specified via M-function values, with all probability mass $P(X \in J_1)$ described by the s M-function values $M_X(J_l)$, so $\sum_{l=1}^{s} M_X(J_l) = P(X \in J_1)$. Then, without additional assumptions, we have*

$$\sum_{l=1}^{s} P(Y < j_l) M_X(J_l) \leq P(Y < X, X \in J_1) \leq P(Y < r) P(X \in J_1),$$

and these bounds are optimal, so they are the maximum lower and minimum upper bounds that generally hold.

Proof. For any number s of nested intervals, the proof follows the same principle, so for ease of notation we present it for $s = 3$. We use the theorem of total probability to condition further on the partition $\{J_3, J_2 \setminus J_3, J_1 \setminus J_2\}$ of J_1 for the random quantity X. The probability distribution of X on J_1 is partially specified via M-function values for X defined on J_1, J_2, J_3. Let $M_X^l(J)$ denote the (unknown) part of the M-function value $M_X(J_l)$ that is actually in $J \subset J_l$, so we have

$$\begin{aligned}
P(X \in J_3) &= M_X^3(J_3) + M_X^2(J_3) + M_X^1(J_3), \\
P(X \in J_2 \setminus J_3) &= M_X^2(J_2 \setminus J_3) + M_X^1(J_2 \setminus J_3), \\
P(X \in J_1 \setminus J_2) &= M_X^1(J_1 \setminus J_2), \\
M_X(J_1) &= M_X^1(J_1 \setminus J_2) + M_X^1(J_2 \setminus J_3) + M_X^1(J_3), \\
M_X(J_2) &= M_X^2(J_2 \setminus J_3) + M_X^2(J_3), \\
M_X(J_3) &= M_X^3(J_3).
\end{aligned}$$

These M-function values are not further specified, but we can now use the theorem of total probability, and then derive bounds by solving the constrained optimization problems. The lower bound follows from (with $J_4 = \emptyset$ for ease of notation)

$$\begin{aligned}
P(Y < X, X \in J_1) &= \sum_{l=1}^{3} P(Y < X, X \in J_l \setminus J_{l+1}) \\
&= \sum_{l=1}^{3} P(Y < X \,|\, X \in J_l \setminus J_{l+1}) P(X \in J_l \setminus J_{l+1}) \\
&= P(Y < X \,|\, X \in J_1 \setminus J_2) M_X^1(J_1 \setminus J_2) + \\
&\quad P(Y < X \,|\, X \in J_2 \setminus J_3)[M_X^2(J_2 \setminus J_3) + M_X^1(J_2 \setminus J_3)] + \\
&\quad P(Y < X \,|\, X \in J_3)[M_X^3(J_3) + M_X^2(J_3) + M_X^1(J_3)].
\end{aligned}$$

With the constraints on these M-function values as given above, the lower bound is achieved by effectively putting the probability masses for X at the infimums of the intervals on which they are defined, so setting

$$M_X^1(J_2 \setminus J_3) = M_X^1(J_3) = M_X^2(J_3) = 0,$$

and taking the lower bounds for the conditional probabilities for $Y < X$, given $X \in I$, for the relevant I above, by replacing $X \in I$ by $X = \inf(I)$, leading to the terms $Y < j_l$ in the lower bound. The upper bound can be derived simultaneously, but is rather trivial as these nested intervals have the same right end-point. The fact that these bounds are optimal, without additional assumptions, follows easily from this construction. □

Bounds for the probability of $X_{a,n_a+1} > X_{b,n_b+1}$, based on rc-$A_{(n_a)}$ and rc-$A_{(n_b)}$, are presented in the following theorem. As these bounds are optimal, without any additional assumptions, they are lower and upper probabilities [15], which we denote by $\underline{P}(X_{a,n_a+1} > X_{b,n_b+1})$ and $\overline{P}(X_{a,n_a+1} > X_{b,n_b+1})$, respectively.

Theorem 1 *Assume that data are available from two independent groups, A and B, following the notation presented above. Based on the assumptions rc-$A_{(n_a)}$ and rc-$A_{(n_b)}$, predictive comparison of these two groups can be based on the following lower and upper probabilities for $X_{a,n_a+1} > X_{b,n_b+1}$,*

$$\begin{aligned}
&\underline{P}(X_{a,n_a+1} > X_{n_b+1}) \\
= \quad & \sum_{i=0}^{u_a} \left\{ \left[\sum_{j=0}^{s_b(t_{a,i})-1} P(X_{b,n_b+1} \in (t_{b,j}, t_{b,j+1})) \right] M_{X_{a,n_a+1}}(t_{a,i}, t_{a,i+1}) \right. \\
& \quad \left. + \sum_{k=1}^{l_{a,i}} \left(\left[\sum_{j=0}^{s_b(c^i_{a,k})} P(X_{b,n_b+1} \in (t_{b,j}, t_{b,j+1})) \right] M_{X_{a,n_a+1}}(c^i_{a,k}, t_{a,i+1}) \right) \right\}, \\
&\overline{P}(X_{a,n_a+1} > X_{b,n_b+1}) \\
= \quad & \sum_{i=0}^{u_a} \left\{ \left[\sum_{j=0}^{s_b(t_{a,i+1})-1} P(X_{b,n_b+1} \in (t_{b,j}, t_{b,j+1})) \right. \right. \\
& \quad + P(X_{b,n_b+1} \in (t_{b,s_b(t_{a,i+1})-1}, t_{b,s_b(t_{a,i+1})})) \\
& \quad \left. \left. + \sum_{l=1}^{s^c_b(t_{a,i+1})} M_{X_{b,n_b+1}}(c^{s^c_b(t_{a,i+1})}, t_{b,s_b(t_{a,i+1})+1}) \right] P(X_{a,n_a+1} \in (t_{a,i}, t_{a,i+1})) \right\}.
\end{aligned}$$

Proof. These lower and upper probabilities are derived by first writing

$$P(X_{a,n_a+1} > X_{b,n_b+1}) = \sum_{i=0}^{u_a} P(X_{b,n_b+1} < X_{a,n_a+1}, X_{a,n_a+1} \in (t_{a,i}, t_{a,i+1})),$$

and then applying the above lemma for each of the terms within this sum, and using the intervals on which the M-function values for X_{a,n_a+1} are defined according to rc-$A_{(n_a)}$. Then, bounds for the resulting probabilities (compare the lemma above) for X_{b,n_b+1} are determined, based on the corresponding M-function values according to rc-$A_{(n_b)}$, where a lower bound is derived by including only the M-function values on intervals that are fully included in the interval in the event of interest, and the upper bound is derived by including all M-function values on intervals that have non-empty intersection with the interval in the event of interest. Further details are relatively straightforward (see Yan [16] for a complete proof). □

These lower and upper probabilities are not available in a nice closed form. However, calculation is relatively easy as the individual terms are all product forms following from the definition of rc-$A_{(n)}$. If the data do not include any right-censorings, these lower and upper probabilities are identical to those presented by Coolen [3]. Although these formulae become fairly complex, the underlying idea for these optimal bounds is straightforward. The lower probability for $X_{a,n_a+1} > X_{b,n_b+1}$, based on the rc-$A_{(n)}$ assumptions per group, puts the probability masses as specified by the M-function values for X_{a,n_a+1} at the infimums of the intervals on which corresponding M-function values are specified, and for X_{b,n_b+1} at the supremums of the intervals, so at this bound the probability masses are effectively least supportive for this event, given the partial specifications via M-function values. Of course, the upper probability just relates to these probability masses being put at the other end-points per interval.

We have presented the lower and upper probabilities for $X_{a,n_a+1} > X_{b,n_b+1}$. Similar results are available for the complementary event $X_{b,n_b+1} > X_{a,n_a+1}$, which can be derived by interchanging the indices for the groups above. However, it is not necessary to calculate lower and upper probabilities for both these events, because the well-known conjugacy property for imprecise probabilities [15], $\underline{P}(E) = 1 - \overline{P}(E^c)$, holds, where E^c is the complementary event of E. Informally, this holds because our bounds are optimal, and correspond to the same assessments based on the rc-$A_{(n)}$ assumptions per group. Alternatively, one could only compute either the lower or upper probabilities for both these events, requiring only a single algorithm, and using this relation to derive the other imprecise probabilities of interest.

Implicit in our results is that the probability of $X_{a,n_a+1} = X_{b,n_b+1}$ is zero, which is reasonable for our method as long as there are no ties among the event times of different groups (it would become a problem if a particular event time had been observed twice or more in each group, we discuss ties briefly in Section 5), and which is a consequence of our method of comparison, where effectively we always put probability masses at end-points of different intervals. It should be remarked, however, that a positive upper probability for $X_{a,n_a+1} = X_{b,n_b+1}$ could also be justified on the basis of these rc-$A_{(n)}$ assumptions, but doing so consistently would have made the analysis presented here more awkward, with little relevance for most practical situations.

4 Examples

We illustrate our nonparametric predictive method for comparison of two groups of lifetime data via two examples. We also compare our method with Mantel’s two-sample test for censored data (see Section 11.7 of Hollander and Wolfe [10] for details), an established nonparametric method for such comparison, and discuss the important difference between our predictive approach and Mantel’s hypothesis test.

Example 2

The data for this example are given in Table 2, and were also used by Parmar and Machin [14] to illustrate nonparametric methods for survival data. It is a subset of data obtained from 183 patients entered into a randomised Phase III trial conducted by the Medical Research Council Working Party on Advanced Carcinoma of the Cervix.

Table 2: Cervical cancer survival data ($> t$: right-censoring at t).

Control (A)	New (B)
90	272
142	362
150	373
269	>383
291	>519
>468	>563
680	>650
837	827
>890	>919
1037	>978
>1090	>1100
>1113	1307
>1153	>1360
1297	>1476
1429	
>1577	

The data are on survival of 30 patients with cervical cancer, recruited to a randomised trial aimed at analysing the effect of addition of a radiosensitiser to radiotherapy (‘*new* treatment’, B), via comparison to the use of radiotherapy alone (‘*control* treatment’, A). Of these 30 patients, $n_a = 16$ received the control treatment A, and $n_b = 14$ received the new treatment B. The data are in days since start of the study, the event of interest is death of the patient caused by this cancer. Further variables recorded for patients in the original study are not taken into account (see Parmar and Machin [14] for further references to the original study), we only use this subset of all the data to illustrate our new method for comparison of two such groups of data.

Using the method presented in Section 3, we compare these two groups of data predictively, by focussing on future observations $X_{a,17}$, assuming rc-$A_{(16)}$, and $X_{b,15}$, assuming rc-$A_{(14)}$. The corresponding lower and upper probabilities are

$$\underline{P}(X_{a,17} > X_{b,15}) = 0.226 \text{ and } \overline{P}(X_{a,17} > X_{b,15}) = 0.473,$$

which, by the conjugacy property for imprecise probability, imply

$$\underline{P}(X_{b,15} > X_{a,17}) = 0.527 \text{ and } \overline{P}(X_{b,15} > X_{a,17}) = 0.774.$$

These imprecise probabilities indicate that a preference for the new treatment B over the control treatment A would be reasonable, if no further information (e.g. on side-effects) is taken into account, and if one aims at surviving longer. In particular from an individual's perspective, this seems to be a natural inference if choice between two treatments is possible.

Although we do not discuss it explicitly here, such a choice could also take further aspects into account via our general rc-$A_{(n)}$-based inferential method. For example, a patient may prefer the treatment with maximum lower probability of surviving a particular length of time, it is fairly straightforward to calculate such lower probabilities per treatment in our approach [6].

From a classical nonparametric point of view, inference on the difference between survival chances for the two treatments could, for example, be based on application of Mantel's two-sample test for censored data, which is a rank-based test of a null-hypothesis of two equal survival functions, using asymptotic normality of the relevant test statistic. Applying this test for these cervical cancer survival data leads to a one-sided p-value of 0.1020, which may not be regarded as strong enough evidence against the null-hypothesis.

Example 3

The data for this example are given in Table 3, and were also used by Hollander and Wolfe [10] to illustrate Mantel's test. These data are from a clinical trial on Hodgkin's disease, a cancer of the lymph system. Two treatments were considered, a radiation treatment of the affected node (Treatment A; 25 patients), and a radiation treatment of the affected node plus all nodes in the trunk of the body (Treatment B; 24 patients). The data represent the relapse-free survival times in days. If a relapse had not occurred before the end of the study, then the observation for that patient is right-censored.

Our method, as presented in Section 3, applied to these data, leads to predictive imprecise probabilities

$$\underline{P}(X_{b,25} > X_{a,26}) = 0.557 \text{ and } \overline{P}(X_{b,25} > X_{a,26}) = 0.893.$$

These values indicate that the data suggest pretty strongly that $T_{b,25} > T_{a,26}$, hence it seems to be in a patient's best interest to opt for Treatment B. Applying Mantel's test to these data leads to an approximate one-sided p-value of 0.0006, which suggests very strongly that the survival functions corresponding to these two treatments are not equal.

Table 3: Hodgkin's disease survival data ($> t$: right-censoring at t).

Treatment A		Treatment B	
86	822	173	>1726
107	836	498	>1763
141	>1309	615	>1807
296	1375	950	>1879
312	>1378	>1190	>1889
330	>1446	>1242	>1897
346	>1540	1408	>1968
364	>1645	>1493	>1972
401	>1818	>1572	>2022
419	>1910	>1576	>2070
505	>1953	>1585	>2177
570	>2052	>1684	
688		>1699	

Clearly, testing equality of survival functions is quite a different inference than our predictive comparison, and it is not unreasonable to consider the outcome of both when trying to get more insight into the different survival chances per treatment. In Example 2, our method suggests that the new treatment would be better for a future patient than the control treatment, although Mantel's test does not strongly reject the hypothesis that both survival functions could be equal. In Example 3, the conclusions from both methods seem to agree more.

In general, it could also happen that Mantel's test would reject the null hypothesis, while we would end up with lower and upper probabilities both close to 0.5, so care should be taken on interpretation of the results of our method and Mantel's test. In situations where the real problem of interest is naturally in terms of comparison of next observations, we believe that our new method should be preferred.

The imprecision in our upper and lower probabilities in Examples 2 and 3 is not unreasonably large, in particular when considering the relatively large number of right-censored observations. This is explicitly due to our assumption rc-$A_{(n)}$, without this exchangeability-related assumption for the residual times till event for the right-censored items, logical bounds on the relevant conditional probabilities would be much wider.

5 Concluding remarks

We suggest that our new method for comparison of two groups of survival data is particularly useful in situations where such comparison takes place from a single individual’s perspective, e.g. when a person has a choice between the two treatments. If one has more relevant information, e.g. covariates or prior knowledge, some established statistical methods will be more appropriate. Our method can then still serve as a sort of base method, which can provide insight into the effect of further information or model assumptions, used with those alternative methods, by comparing the ultimate inferences. Extending our approach to possible inclusion of covariates is an interesting and relevant topic for future research.

Generalization of this approach to more than two groups of data is feasible, in a way similar to Coolen and van der Laan [5], who considered this problem without censored observations. It is also possible to extend attention to multiple future observations per group, but this would lead to rather complex computations due to dependence of such future observations for the same group [4, 7].

Throughout, we have assumed that there are no ties in the data. If there are ties, these can relatively easily be taken into account by breaking the ties, so assuming that tied values are only nearly identical, applying our method, and then letting the differences decrease to zero. For ties between the groups, one should break them into all possible orderings among the groups, calculate lower (upper) probabilities for each such ordering, and then take the minimum (maximum) of all these lower (upper) probabilities as the actual lower (upper) probability to be used for the comparison.

Acknowledgements

We are grateful to the referees for thoughtful comments and suggestions that have led to better presentation.

References

[1] T. Augustin, and F.P.A. Coolen. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, to appear, 2003.

[2] L.M. Berliner, and B.M. Hill. Bayesian nonparametric survival analysis (with discussion). *Journal of the American Statistical Association,* 83:772-784, 1988.

[3] F.P.A. Coolen. Comparing two populations based on low stochastic structure assumptions. *Statistics & Probability Letters*, 29:297-305, 1996.

[4] F.P.A. Coolen. Low structure imprecise predictive inference for Bayes’ problem. *Statistics & Probability Letters*, 36:349-357, 1998.

[5] F.P.A. Coolen, and P. van der Laan. Imprecise predictive selection based on low structure assumptions. *Journal of Statistical Planning and Inference*, 98:185-203, 2001.

[6] F.P.A. Coolen, and K.J. Yan. Nonparametric predictive inference with right-censored data. *Journal of Statistical Planning and Inference*, to appear, 2003.

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

[8] B.M. Hill. De Finetti's theorem, induction, and A_n, or Bayesian nonparametric predictive inference (with discussion). In *Bayesian Statistics 3*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, Eds. Oxford University Press, 211-241, 1988.

[9] B.M. Hill. Parametric models for $A_{(n)}$: splitting processes and mixtures. *Journal of the Royal Statistical Society B*, 55:423-433, 1993.

[10] M. Hollander, and D.A. Wolfe. *Nonparametric Statistical Methods* (2nd ed.). Wiley, New York, 1999.

[11] J.F. Lawless. *Statistical Models and Methods for Lifetime Data*. Wiley, New York, 1982.

[12] C.F. Manski. *Identification Problems in the Social Sciences*. Harvard University Press, Cambridge, Massachusetts, 1995.

[13] W.Q. Meeker, and L.A. Escobar. *Statistical Methods for Reliability Data*. Wiley, New York, 1998.

[14] M.K.B. Parmar, and D. Machin. *Survival Analysis: a Practical Approach*. Wiley, Chichester, 1995.

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

[16] K.J. Yan. *Nonparametric predictive inference with right-censored data*. PhD-thesis, University of Durham (available from the corresponding author), 2002.

Frank Coolen is Reader in Statistics at the Department of Mathematical Sciences, University of Durham, Durham 3LE, United Kingdom. E-mail: Frank.Coolen@durham.ac.uk

Ke-Jian Yan finished his PhD studies, under supervision of Frank Coolen, at the Department of Mathematical Sciences, University of Durham, in August 2002.

Dynamic Programming for Discrete-Time Systems with Uncertain Gain[*]

G. DE COOMAN
Ghent University, Belgium

M. C. M. TROFFAES
Ghent University, Belgium

Abstract

We generalise the optimisation technique of dynamic programming for discrete-time systems with an uncertain gain function. We assume that uncertainty about the gain function is described by an imprecise probability model, which generalises the well-known Bayesian, or precise, models. We compare various optimality criteria that can be associated with such a model, and which coincide in the precise case: maximality, robust optimality and maximinity. We show that (only) for the first two an optimal feedback can be constructed by solving a Bellman-like equation.

Keywords

optimal control, dynamic programming, uncertainty, imprecise probabilities

1 Introduction to the Problem

The main objective in optimal control is to find out how a system can be influenced, or controlled, in such a way that its behaviour satisfies certain requirements, while at the same time maximising a given gain function. A very effective method for solving optimal control problems for discrete-time systems is the recursive *dynamic programming* method, introduced by Richard Bellman [1].

To explain the ideas behind this method, we refer to Figures 1 and 2. In Figure 1 we depict a situation where a system can go from state a to state c through state b in three ways: following the paths $\alpha\beta$, $\alpha\gamma$ and $\alpha\delta$. We denote the associated gains by $J_{\alpha\beta}$, $J_{\alpha\gamma}$ and $J_{\alpha\delta}$ respectively. Assume that path $\alpha\gamma$ is optimal: $J_{\alpha\gamma} > J_{\alpha\beta}$ and $J_{\alpha\gamma} > J_{\alpha\delta}$. Then it follows that path γ is the optimal way to go from b to c. To

[*] This paper presents research results of project G.0139.01 of the Fund for Scientific Research, Flanders (Belgium), and of the Belgian Programme on Interuniversity Poles of Attraction initiated by the Belgian state, Prime Minister's Office for Science, Technology and Culture.

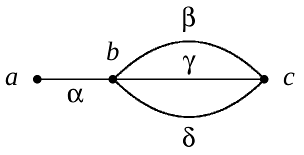


Figure 1: Principle of Optimality

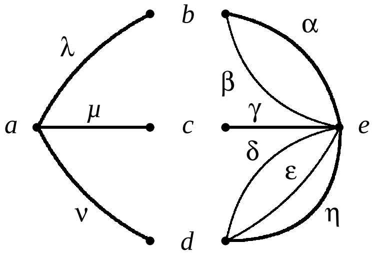


Figure 2: Dynamic Programming

see this, observe that $J_{\alpha\nu} = J_\alpha + J_\nu$ for $\nu \in \{\beta,\gamma,\delta\}$ (gains are assumed to be additive) and derive from the inequalities above that $J_\gamma > J_\beta$ and $J_\gamma > J_\delta$. This simple observation, which Bellman called the *principle of optimality*, forms the basis for the recursive technique of dynamic programming for solving an optimal control problem. To see how this is done in principle, consider the situation depicted in Figure 2. Suppose we want to find the optimal way to go from state a to state e. After one time step, we can reach the states b, c and d from state a, and the optimal paths from these states to the final state e are known to be α, γ and η, respectively. To find the optimal path from a to e, we only need to compare the costs $J_\lambda + J_\alpha$, $J_\mu + J_\gamma$ and $J_\nu + J_\eta$ of the respective candidate optimal paths $\lambda\alpha$, $\mu\gamma$ and $\nu\eta$, since the principle of optimality tells us that the paths $\lambda\beta$, $\nu\delta$ and $\nu\varepsilon$ cannot be optimal: if they were, then so would be the paths β, δ and ε. This, written down in a more formal language, is what is essentially known as *Bellman's equation*. It allows us to solve an optimal control problem very efficiently through a recursive procedure, by calculating optimal paths backwards from the final state.

In applications, it may happen that the gain function, which associates a gain with every control action and the resulting behaviour of the system, is not well known. This problem is most often treated by modelling the uncertainty about the gain by means of a probability measure, and by maximising the *expected gain* under this probability measure. Due to the linearity of the expectation operator, this approach does not change the nature of the optimisation problem in any essential way, and the usual dynamic programming method can therefore still be applied.

It has however been argued by various scholars (see [11, Chapter 5] for a detailed discussion with many references) that uncertainty cannot always be modelled adequately by (precise) probability measures, because, roughly speaking, there may not be enough information to identify a single probability measure. In those cases, it is more appropriate to model the available information through an *imprecise* probability model, e.g., by a lower prevision, or by a set of probability measures. For applications of this approach, see for instance [4, 10].

Two questions now arise naturally. First, how should we formulate the optimal control problem: what does it mean for a control to be optimal with respect to an *uncertain gain function*, where the uncertainty is represented through an imprecise probability model? In Section 2 we identify three different optimality criteria, each with a different interpretation (although they coincide for precise probability models), and we study the relations between them. Secondly, is it still possible to solve the corresponding optimal control problems using the ideas underlying Bellman’s dynamic programming method? We show in Section 3 that this is the case for only two of the three optimality criteria we study: only for these a generalised principle of optimality holds, and the optimal controls are solutions of suitably generalised Bellman-like equations. To arrive at this, we study the properties that an abstract notion of optimality should satisfy for the Bellman approach to work.

We recognise that other authors (see for instance [8]) have extended the dynamic programming algorithm to systems with imprecise gain and/or imprecise dynamics. However in doing so, none of them seems to have questioned in what sense their generalised dynamic programming method leads to optimal paths. In this article we approach the problem from the opposite, and in our opinion, more logical side: one should *first* define a notion optimality and investigate whether the dynamic programming argument holds for this notion of optimality, instead of blindly “generalising” Bellman’s algorithm. In the remainder of this section, we introduce the basic terminology and notation that will allow us to give a precise formulation of the problems under study. We have omitted proofs of technical results that do not contribute to a better understanding of the main ideas.

1.1 Preliminaries

1.1.1 The System

For a and b in $\mathbb{N}$, the set of natural numbers c that satisfy $a \leq c \leq b$ is denoted by $[a,b]$. Let $x_{k+1} = f(x_k,u_k,k)$ describe a discrete-time dynamical system with $k \in \mathbb{N}$, $x_k \in \mathcal{X}$ and $u_k \in \mathcal{U}$. The set $\mathcal{X}$ is the state space (e.g., $\mathbb{R}^n$, $n \in \mathbb{N} \setminus \{0\}$), and the set $\mathcal{U}$ is the control space (e.g., $\mathbb{R}^m$, $m \in \mathbb{N} \setminus \{0\}$). The map $f\colon \mathcal{X} \times \mathcal{U} \times \mathbb{N} \to \mathcal{X}$ describes the evolution of the state through time: given the state $x_k \in \mathcal{X}$ and the control $u_k \in \mathcal{U}$ at time $k \in \mathbb{N}$, it returns the next state x_{k+1} of the system. For practical reasons, we impose a final time N beyond which we are not interested in the dynamics of system. Moreover, it may happen that not all states and controls are allowed at all times: we demand that x_k should belong to a set of *admissible states* $\mathcal{X}_k$ at every instant $k \in [0,N]$, and that u_k should belong to a set of *admissible controls* $\mathcal{U}_k$ at every instant $k \in [0,N-1]$, where $\mathcal{X}_k \subseteq \mathcal{X}$ and $\mathcal{U}_k \subseteq \mathcal{U}$ are given. The set $\mathcal{X}_N$ may be thought of as the set we want the state to end up in at time N.

1.1.2 Paths

A *path* is a triple $(x,k,u.)$, where $x \in \mathcal{X}$ is a state, $k \in [0,N]$ a time instant, and $u.\colon [k,N-1] \to \mathcal{U}$ a sequence of controls. A path fixes a unique state trajectory $x.\colon [k,N] \to \mathcal{X}$, which is defined recursively through $x_k = x$ and $x_{i+1} = f(x_i,u_i,i)$ for every $i \in [k,N-1]$. It is said to be *admissible* if $x_\ell \in \mathcal{X}_\ell$ for every $\ell \in [k,N]$

and $u_\ell \in \mathcal{U}_\ell$ for every $\ell \in [k, N-1]$. We denote the unique map from $\emptyset$ to $\mathcal{U}$ by $u_\emptyset$. If $k = N$, the control $u_.$ does nothing: it is equal to $u_\emptyset$.

The set of admissible paths starting in the state $x \in \mathcal{X}_k$ at time $k \in [0, N]$ is denoted by $\mathcal{U}(x,k)$, i.e., $\mathcal{U}(x,k) = \{(x,k,u_.) \colon (x,k,u_.) \text{ admissible path}\}$. For example, $\mathcal{U}(x,N) = \{(x,N,u_\emptyset)\}$ whenever $x \in \mathcal{X}_N$ and $\mathcal{U}(x,N) = \emptyset$ otherwise.

If we consider a path with final time M different from N, then we write $(x,k,u_.)_M$ (assume $k \leq M \leq N$). Observe that $(x,k,u_.)_k$ can be identified with $(x,k,u_\emptyset)_k$; it is the unique path (of length zero) starting and ending at time k in x. Let $0 \leq k \leq \ell \leq m$. Two paths $(x,k,u_.)_\ell$ and $(y,\ell,v_.)_m$ can be concatenated if $y = x_\ell$. The concatenation is denoted by $(x,k,u_.,\ell,v_.)_m$ or $(x,k,u_.)_\ell \oplus (y,\ell,v_.)_m$, and represents the path that starts in state x at time k, and results from applying control u_i for times $i \in [k, \ell-1]$ and control v_i for times $i \in [\ell, m-1]$. In particular,

$$(x,k,u_.)_\ell = (x,k,u_.)_k \oplus (x,k,u_.)_\ell = (x,k,u_.)_\ell \oplus (x_\ell,\ell,u_.)_\ell.$$

The set of admissible paths starting in state $x \in \mathcal{X}_k$ at time $k \in [0,N]$ and ending at time $\ell \in [k,N]$ is denoted by $\mathcal{U}(x,k)_\ell$. In particular we have that $\mathcal{U}(x,k)_k = \{(x,k,u_\emptyset)_k\}$ if $x \in \mathcal{X}_k$, and $\mathcal{U}(x,k)_k = \emptyset$ otherwise. Moreover, for any $(x,k,u_.)_\ell \in \mathcal{U}(x,k)_\ell$ and any $\mathcal{V} \subseteq \mathcal{U}(x_\ell,\ell)$, we use the notation $(x,k,u_.)_\ell \oplus \mathcal{V}$ for the set

$$\{(x,k,u_.)_\ell \oplus (x_\ell,\ell,v_.) \colon (x_\ell,\ell,v_.) \in \mathcal{V}\}.$$

1.1.3 The Gain Function

Applying the control action $u \in \mathcal{U}$ to the system in state $x \in \mathcal{X}$ at time $k \in [0,N-1]$ yields a real-valued gain $g(x,u,k,\omega)$. Moreover, reaching the final state $x \in \mathcal{X}$ at time N also yields a gain $h(x,\omega)$. The parameter $\omega \in \Omega$ represents the (unknown) state of the world, used to model uncertainty of the gains. If we knew that the real state of the world was ω_o, we would know the gains to be $g(x,u,k,\omega_o)$ and $h(x,\omega_o)$. As it is, the real state of the world is uncertain, and so are the gains, which could be considered as random variables. It is important to note that the parameter ω only influences the gains; it has no effect on the system dynamics, which are assumed to be known perfectly well.

Assuming gain additivity, we can also associate a gain with a path $(x,k,u_.)$:

$$J(x,k,u_.,\omega) = \textstyle\sum_{i=k}^{N-1} g(x_i,u_i,i,\omega) + h(x_N,\omega),$$

for any $\omega \in \Omega$. If $M < N$, we also use the notation

$$J(x,k,u_.,\omega)_M = \textstyle\sum_{i=k}^{M-1} g(x_i,u_i,i,\omega).$$

It will be convenient to associate a zero gain with an empty control action: for $k \in [0,N]$ we let $J(x,k,u_.,\omega)_k = 0$.

The main objective of optimal control can now be formulated as follows: given that the system is in the initial state $x \in \mathcal{X}$ at time $k \in [0,N]$, find a control sequence $u_. \colon [k,N-1] \to \mathcal{U}$ resulting in an admissible path $(x,k,u_.)$ such that the corresponding gain $J(x,k,u.,\omega)$ is maximal. Moreover, we would like this control sequence $u.$ to be such that its value u_k at the time instant k is a function of x and k only, since in that case the control can be realised through state feedback.

If ω is known, then the problem reduces to the classical problem of dynamic programming, first studied and solved by Bellman [1]. We assume here that the available information about the true state of the world is modelled through a *coherent lower prevision* $\underline{P}$ defined on the set $\mathcal{L}(\Omega)$ of *gambles*, or bounded real-valued maps, on Ω. A special case of this obtains when $\underline{P}$ is a linear prevision P. Linear previsions are the precise probability models; they can be interpreted as expectation operators associated with (finitely additive) probability measures, and they are *previsions* or *fair prices* in the sense of de Finetti [6]. We assume that the reader is familiar with lower previsions and coherence (see [11] for more details).

For a given path $(x,k,u.)$, the corresponding gain $J(x,k,u.,\omega)$ can be seen as a real-valued map on Ω, which is denoted by $J(x,k,u.)$ and called the *gain gamble* associated with $(x,k,u.)$.[1] In the same way we define the gain gambles $g(x_k,u_k,k)$, $h(x_N)$ and $J(x,k,u.)_M$. There is gain additivity: $J(x,k,u.,\ell,v.)_m = J(x,k,u.)_\ell + J(x_\ell,\ell,v.)_m$ for $k \leq \ell \leq m \leq N$, and $J(x,k,u.)_k = 0$. We denote by $\mathcal{J}(x,k)$ the set of gain gambles for admissible paths from initial state $x \in \mathcal{X}_k$ at time $k \in [0,N]$:

$$\mathcal{J}(x,k) = \{J(x,k,u.)\colon (x,k,u.) \in \mathcal{U}(x,k)\}.$$

For fixed $k \in [0,N-1]$ and $x \in \mathcal{X}_k$, the gain $J(x,k,u.,\omega)$ can also be interpreted as a map from $\mathcal{U}(x,k)$ to $\mathcal{L}(\Omega)$; this map is denoted by $J(x,k)$.

2 Optimality Criteria

2.1 $\underline{P}$-Maximality

The lower prevision $\underline{P}(X)$ of a gamble X has a behavioural interpretation as a subject's supremum acceptable price for buying the gamble X: it is the highest value of μ such that the subject accepts the gamble $X - x$ (i.e., accepts to buy X for a price x) for all $x < \mu$. The conjugate upper prevision $\overline{P}(X) = -\underline{P}(-X)$ of X is then the subject's infimum acceptable price for selling X. This way of looking at a lower prevision $\underline{P}$ defined on the set $\mathcal{L}(\Omega)$ of all gambles allows us to define a strict partial order $>_{\underline{P}}$ on $\mathcal{L}(\Omega)$ whose interpretation is that of strict preference.

Definition 1 *For any gambles X and Y in $\mathcal{L}(\Omega)$ we say that X* strictly dominates *Y, or X* is strictly preferred to *Y (with respect to $\underline{P}$), and write $X >_{\underline{P}} Y$, if*

$$\underline{P}(X-Y) > 0 \textit{ or } (X \geq Y \textit{ and } X \neq Y).$$

Indeed, if $X \geq Y$ and $X \neq Y$, then the subject should be willing to exchange Y for X, since this transaction can only improve his gain. On the other hand,

[1] To simplify the discussion, we assume this map is bounded.

$\underline{P}(X-Y)>0$ expresses that the subject is willing to pay a strictly positive price to exchange Y for X, which again means that he strictly prefers X to Y.

It is clear that we can also use the lower prevision $\underline{P}$ to express a strict preference between any two *paths* $(x,k,u_\cdot)$ and $(x,k,v_\cdot)$, based on their gains: if $J(x,k,u_\cdot)>_{\underline{P}}J(x,k,v_\cdot)$ this means that the uncertain gain $J(x,k,u_\cdot)$ is strictly preferred to the uncertain gain $J(x,k,v_\cdot)$. We then say that the path $(x,k,u_\cdot)$ is strictly preferred to $(x,k,v_\cdot)$, and we use the notation $(x,k,u_\cdot)>_{\underline{P}}(x,k,v_\cdot)$.

$>_{\underline{P}}$ is anti-reflexive and transitive, and therefore a strict partial order on $\mathcal{L}(\Omega)$, and in particular also on $\mathcal{J}(x,k)$ and on $\mathcal{U}(x,k)$. But it is generally not linear: any two paths need not be comparable with respect to this order, and it does not always make sense to look for greatest elements, i.e., for paths that strictly dominate all the others. Rather, we should look for maximal, or undominated, elements: paths that are not dominated by any other path. Observe that a maximal gamble X in a set $\mathcal{K}$ with respect to $>_{\underline{P}}$ is a maximal element of $\mathcal{K}$ with respect to $\geq$ (i.e., it is point-wise undominated) such that $\overline{P}(X-Y)\geq 0$ for all $Y\in\mathcal{K}$. In case $\underline{P}$ is a linear prevision P, maximal gambles with respect to $>_P$ are just the point-wise undominated gambles whose prevision is maximal; they maximise expected gain.

Definition 2 *Let* $k\in[0,N]$, $x\in\mathcal{X}_k$ *and* $\mathcal{V}\subseteq\mathcal{U}(x,k)$. *A path* $(x,k,u^*_\cdot)$ *in* $\mathcal{V}$ *is called* $\underline{P}$-maximal, *or* $>_{\underline{P}}$-optimal, *in* $\mathcal{V}$ *if no path in* $\mathcal{V}$ *is strictly preferred to* $(x,k,u^*_\cdot)$, *i.e.,* $(x,k,u_\cdot)\not>_{\underline{P}}(x,k,u^*_\cdot)$ *for all* $(x,k,u_\cdot)\in\mathcal{V}$. *We denote the set of the* $\underline{P}$*-maximal paths in* $\mathcal{V}$ *by* $\operatorname{opt}_{>_{\underline{P}}}(\mathcal{V})$. *The operator* $\operatorname{opt}_{>_{\underline{P}}}$ *is called the* optimality operator induced by $>_{\underline{P}}$, *associated with* $\mathcal{U}(x,k)$.

The $\underline{P}$-maximal paths in $\mathcal{U}(x,k)$ are just those admissible paths starting at time k in state x for which the associated gain gamble is a maximal element of $\mathcal{J}(x,k)$ with respect to the strict partial order $>_{\underline{P}}$. If we denote the set of these $>_{\underline{P}}$-maximal gain gambles in $\mathcal{J}(x,k)$ by $\operatorname{opt}_{>_{\underline{P}}}(\mathcal{J}(x,k))$, then for all $(x,k,u_\cdot)\in\mathcal{U}(x,k)$:

$$(x,k,u_\cdot)\in\operatorname{opt}_{>_{\underline{P}}}(\mathcal{U}(x,k))\iff J(x,k,u_\cdot)\in\operatorname{opt}_{>_{\underline{P}}}(\mathcal{J}(x,k))\,.$$

$\underline{P}$-maximal paths do not always exist: not every partially ordered set has maximal elements. A fairly general sufficient condition for the existence of $\underline{P}$-maximal elements in $\mathcal{J}(x,k)$ (and hence in $\mathcal{U}(x,k)$) is that $\mathcal{J}(x,k)$ should be compact[2] (and of course non-empty). This follows from a general result mentioned in [11, Section 3.9.2]. In fact, Theorem 1 is a stronger result, whose Corollary 1 turns out to be very important in proving that the dynamic programming approach works for $\underline{P}$-maximality (see Section 3.2). Its proof is based on Zorn's lemma.

Theorem 1 *For every element* X *of a compact subset* $\mathcal{K}$ *of* $\mathcal{L}(\Omega)$ *that is not a maximal element of* $\mathcal{K}$ *with respect to* $>_{\underline{P}}$ *there is some maximal element* Y *of* $\mathcal{K}$ *with respect to* $>_{\underline{P}}$ *such that* $Y>_{\underline{P}}X$.

[2] In this paper, we always assume that $\mathcal{L}(\Omega)$ is provided with the supremum-norm topology.

Corollary 1 *Let $k \in [0,N]$ and let $x \in \mathcal{X}_k$. If $\mathcal{J}(x,k)$ is compact then for every admissible, non-$\underline{P}$-maximal path $(x,k,u.)$ in $\mathcal{U}(x,k)$ there is a $\underline{P}$-maximal path $(x,k,u^*_.)$ in $\mathcal{U}(x,k)$ that is strictly preferred to it.*

2.2 $\underline{P}$-Maximinity

We now turn to another optimality criterion that can be associated with a lower prevision $\underline{P}$. We can use $\underline{P}$ to define another strict order on $\mathcal{L}(\Omega)$:

Definition 3 *For any gambles X and Y in $\mathcal{L}(\Omega)$ we write $X \sqsupset_{\underline{P}} Y$ if*

$$\underline{P}(X) > \underline{P}(Y) \text{ or } (X \geq Y \text{ and } X \neq Y).$$

$\sqsupset_{\underline{P}}$ induces a strict partial order on $\mathcal{U}(x,k)$, since it is anti-reflexive and transitive on $\mathcal{L}(\Omega)$. A maximal element X of a subset $\mathcal{K}$ of $\mathcal{L}(\Omega)$ with respect to $\sqsupset_{\underline{P}}$ is easily seen to be a point-wise undominated element of $\mathcal{K}$ that maximises the lower prevision: $\underline{P}(X) \geq \underline{P}(Y)$ for all $Y \in \mathcal{K}$.

We can consider as optimal in $\mathcal{U}(x,k)$ those admissible paths $(x,k,u.)$ for which the associated gain gamble $J(x,k,u.)$ is a maximal element of $\mathcal{J}(x,k)$ with respect to $\sqsupset_{\underline{P}}$; they are the paths $(x,k,u.)$ that maximise the 'lower expected gain' $\underline{P}(J(x,k,u.))$ and whose gain gambles $J(x,k,u.)$ are point-wise undominated.

Definition 4 *Let $k \in [0,N]$, $x \in \mathcal{X}_k$ and $\mathcal{V} \subseteq \mathcal{U}(x,k)$. A path $(x,k,u^*_.)$ in $\mathcal{V}$ is called* $\underline{P}$-maximin, *or* $\sqsupset_{\underline{P}}$-optimal, *in $\mathcal{V}$ if no path in $\mathcal{V}$ is strictly preferred to $(x,k,u^*_.)$, i.e., $(x,k,u.) \not\sqsupset_{\underline{P}} (x,k,u^*_.)$ for all $(x,k,u.) \in \mathcal{V}$. We denote the set of the $\underline{P}$-maximin paths in $\mathcal{V}$ by $\operatorname{opt}_{\sqsupset_{\underline{P}}}(\mathcal{V})$. The operator $\operatorname{opt}_{\sqsupset_{\underline{P}}}$ is called the* optimality operator induced by $\sqsupset_{\underline{P}}$, *associated with $\mathcal{U}(x,k)$.*

Proposition 1 *$\underline{P}$-maximinity implies $\underline{P}$-maximality. For a linear prevision P, P-maximinity is equivalent to P-maximality.*

The existence of maximal elements with respect to $\sqsupset_{\underline{P}}$ in an arbitrary set of gambles $\mathcal{K}$ is obviously not guaranteed. But if $\mathcal{K}$ is compact, then we may easily infer from the continuity of any coherent lower prevision $\underline{P}$, that the counterparts of Theorem 1 and Corollary 1 hold for $\sqsupset_{\underline{P}}$.

2.3 $\mathcal{M}$-Maximality

There is a tendency, especially among robust Bayesians, to consider an imprecise probability model as a compact convex set of linear previsions $\mathcal{M} \subseteq \mathcal{P}(\Omega)$, where $\mathcal{P}(\Omega)$ is the set of all linear previsions on $\mathcal{L}(\Omega)$. $\mathcal{M}$ is assumed to contain the true, but unknown, linear prevision P_T that models the available information [2, 7].

A gamble X is then certain to be strictly preferred to a gamble Y under the true linear prevision P_T if and only if it is strictly preferred under all candidate models $P \in \mathcal{M}$. This leads to a 'robustified' strict partial order $>_{\mathcal{M}}$ on $\mathcal{L}(\Omega)$.

Definition 5 *$X >_{\mathcal{M}} Y$ if $X >_P Y$ for all $P \in \mathcal{M}$.*

Since $\mathcal{M}$ is assumed to be compact and convex, it is not difficult to show that the strict partial orders $>_{\mathcal{M}}$ and $>_{\underline{P}}$ are one and the same, where the coherent lower prevision $\underline{P}$ is the so-called lower envelope of $\mathcal{M}$, defined by $\underline{P}(X) = \inf\left\{P(X)\colon P \in \mathcal{M}\right\}$ for all $X \in \mathcal{L}(\Omega)$.[3] Conversely, given a coherent lower prevision $\underline{P}$, the strict partial orders $>_{\mathcal{M}(\underline{P})}$ and $>_{\underline{P}}$ are identical, where

$$\mathcal{M}(\underline{P}) = \{P \in \mathcal{P}(\Omega)\colon (\forall X \in \mathcal{L}(\Omega))(P(X) \geq \underline{P}(X))\}$$

is the set of linear previsions that dominate $\underline{P}$. These strict partial orders therefore have the same maximal elements, and lead to the same notion of optimality.

But there is in the literature yet another notion of optimality that can be associated with a compact convex set of linear previsions $\mathcal{M}$: a gamble X is considered optimal in a set of gambles $\mathcal{K}$ if it is a maximal element of $\mathcal{K}$ with respect to the strict partial order $>_P$ for *some* $P \in \mathcal{M}$. This notion of optimality is called 'E-admissibility' by Levi [9, Section 4.8]. It does not generally coincide with the ones associated with the strict partial orders $>_{\mathcal{M}}$ and $>_{\underline{P}}$, unless the set $\mathcal{K}$ is convex [11, Section 3.9]. We are therefore led to consider a third notion of optimality:

Definition 6 *Let $x \in \mathcal{X}$, $k \in [0,N]$ and $\mathcal{V} \subseteq \mathcal{U}(x,k)$. A path $(x,k,u_\cdot^*) \in \mathcal{V}$ is said to be $\mathcal{M}$-*maximal *in $\mathcal{V}$ if it is P-maximal in $\mathcal{V}$ for some P in $\mathcal{M}$, or in other words if it is $\geq$-maximal in $\mathcal{V}$ and maximises $P(J(x,k,u_\cdot))$ over $\mathcal{V}$ for some $P \in \mathcal{M}$. The set of all $\mathcal{M}$-maximal elements of $\mathcal{V}$ is denoted by* $\mathrm{opt}_{\mathcal{M}}(\mathcal{V})$.

Interestingly, for any set of paths $\mathcal{V} \subseteq \mathcal{U}(x,k)$:

$$\mathrm{opt}_{\mathcal{M}}(\mathcal{V}) = \bigcup_{P \in \mathcal{M}} \mathrm{opt}_{>_P}(\mathcal{V}). \tag{1}$$

3 Dynamic Programming

3.1 A General Notion of Optimality

We have discussed three different ways of associating optimal paths with a lower prevision $\underline{P}$, all of which occur in the literature. We now propose to find out whether, for these different types of optimality, we can use the ideas behind the dynamic programming method to solve the corresponding optimal control problems. To do this, we take a closer look at Bellman's analysis as described in Section 1, and we investigate which properties a generic notion of optimality must satisfy for his method to work. Let us therefore assume that there is some property, called $*$-*optimality*, which a path in a given set of paths $\mathcal{P}$ either has or does not have. If a path in $\mathcal{P}$ has this property, we say that it is $*$-optimal in $\mathcal{P}$. We

[3] Since $\mathcal{M}$ is compact, this infimum is actually achieved.

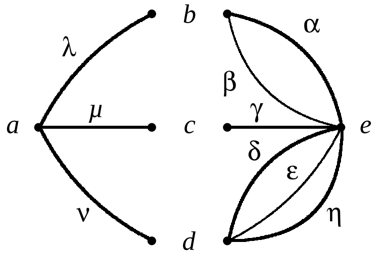


Figure 3: A More General Type of Dynamic Programming

shall denote the set of the $*$-optimal elements of $\mathcal{P}$ by $\operatorname{opt}_*(\mathcal{P})$. By definition, $\operatorname{opt}_*(\mathcal{P}) \subseteq \mathcal{P}$. Further on, we shall apply our findings to the various instances of $*$-optimality described above.

Consider Figure 3, where we want to find the $*$-optimal paths from state a to state e. Suppose that after one time step, we can reach the states b, c and d from state a. The $*$-optimal paths from these states to the final state e are known to be α, γ, and δ and η, respectively. For the dynamic programming approach to work, we need to be able to infer from this a generalised form of the Bellman equation, stating essentially that the $*$-optimal paths from a to e, *a priori* given by $\operatorname{opt}_*(\{\lambda\alpha,\lambda\beta,\mu\gamma,\nu\delta,\nu\varepsilon,\nu\eta\})$, are actually also given by $\operatorname{opt}_*(\{\lambda\alpha,\mu\gamma,\nu\delta,\nu\eta\})$, i.e., the $*$-optimal paths in the set of concatenations of λ, μ and ν with the respective $*$-optimal paths α, γ, and δ and η. It is therefore necessary to exclude that the concatenations $\lambda\beta$ and $\nu\varepsilon$ with the non-$*$-optimal paths β and ν can be $*$-optimal. This amounts to requiring that the operator opt_* should satisfy some appropriate generalisation of Bellman’s *principle of optimality* that will allow us to conclude that $\lambda\beta$ and $\nu\varepsilon$ cannot be $*$-optimal because then β and ε would be $*$-optimal as well. Definition 8 below provides a precise general formulation.

But, perhaps surprisingly for someone familiar with the traditional form of dynamic programming, opt_* should satisfy an *additional* property: the omission of the non-$*$-optimal paths $\lambda\beta$ and $\nu\varepsilon$ from the set of candidate $*$-optimal paths should not have any effect on the actual $*$-optimal paths: we need that

$$\operatorname{opt}_*(\{\lambda\alpha,\lambda\beta,\mu\gamma,\nu\delta,\nu\varepsilon,\nu\eta\}) = \operatorname{opt}_*(\{\lambda\alpha,\mu\gamma,\nu\delta,\nu\eta\}).$$

This is obviously true for the simple type of optimality that we have looked at in Section 1, but it need not be true for the more abstract types that we want to consider here. Equality will be guaranteed if opt_* is insensitive to the omission of non-$*$-optimal elements from $\{\lambda\alpha,\lambda\beta,\mu\gamma,\nu\delta,\nu\varepsilon,\nu\eta\}$, in the following sense.

Definition 7 *Consider a set $S \neq \emptyset$ and an* optimality operator opt_* *defined on the set $\wp(S)$ of subsets of S such that* $\operatorname{opt}_*(T) \subseteq T$ *for all $T \subseteq S$. Elements of* $\operatorname{opt}_*(T)$ *are called* $*$-optimal in T. opt_* *is called* insensitive to the omission of non-$*$-optimal elements from S *if* $\operatorname{opt}_*(S) = \operatorname{opt}_*(T)$ *for all T such that* $\operatorname{opt}_*(S) \subseteq T \subseteq S$.

The following proposition gives an interesting sufficient condition for this insensitivity in case optimality is associated with a (family of) strict partial order(s): it suffices that every non-optimal path is strictly dominated by an optimal path.

Proposition 2 *Let S be a non-empty set provided with a family of strict partial orders $>_j$, $j \in J$. Define for $T \subseteq S$,* $\operatorname{opt}_{>_j}(T) = \left\{a \in T\colon (\forall b \in T)(b \not>_j a)\right\}$ *as the set of maximal elements of T with respect to $>_j$, and let* $\operatorname{opt}_J(T) = \bigcup_{j \in J} \operatorname{opt}_{>_j}(T)$. *Then* $\operatorname{opt}_{>_j}$, $j \in J$ *and* opt_J *are optimality operators. If for some $j \in J$,*

$$(\forall a \in S \setminus \operatorname{opt}_{>_j}(S))(\exists b \in \operatorname{opt}_{>_j}(S))(b >_j a), \tag{2}$$

then $\operatorname{opt}_{>_j}$ *is insensitive to omission of non-$>_j$-optimal elements from S. If* (2) *holds for all $j \in J$, then* opt_J *is insensitive to omission of non-J-optimal elements from S.*

Proof. Consider j in J, and assume that (2) holds for this j. Let $\operatorname{opt}_{>_j}(S) \subseteq T \subseteq S$, then we must prove that $\operatorname{opt}_{>_j}(S) = \operatorname{opt}_{>_j}(T)$. First of all, if $a \in \operatorname{opt}_{>_j}(S)$ then $b \not>_j a$ for all b in S, and *a fortiori* for all b in T, so that $a \in \operatorname{opt}_{>_j}(T)$. Consequently, $\operatorname{opt}_{>_j}(S) \subseteq \operatorname{opt}_{>_j}(T)$. Conversely, let $a \in \operatorname{opt}_{>_j}(T)$ and assume *ex absurdo* that $a \notin \operatorname{opt}_{>_j}(S)$. It then follows from (2) that there is some c in $\operatorname{opt}_{>_j}(S)$ and therefore in T such that $c >_j a$, which contradicts $a \in \operatorname{opt}_{>_j}(T)$.

Next, assume that (2) holds for all $j \in J$. Let $\operatorname{opt}_J(S) \subseteq T \subseteq S$, then we must prove that $\operatorname{opt}_J(S) = \operatorname{opt}_J(T)$. Consider any $j \in J$, then $\operatorname{opt}_{>_j}(S) \subseteq \operatorname{opt}_J(S) \subseteq T \subseteq S$, so we may infer from the first part of the proof that $\operatorname{opt}_{>_j}(S) = \operatorname{opt}_{>_j}(T)$. By taking the union over all $j \in J$, we find that indeed $\operatorname{opt}_J(S) = \operatorname{opt}_J(T)$. □

We are now ready for a precise formulation of the dynamic programming approach for solving optimal control problems associated with general types of optimality. We assume that we have some type of optimality, called $*$-optimality, that allows us to associate with the set of admissible paths $\mathcal{U}(x,k)$ starting at time k in initial state x, an optimality operator opt_* defined on the set $\wp(\mathcal{U}(x,k))$ of subsets of $\mathcal{U}(x,k)$. For each such subset $\mathcal{V}$, $\operatorname{opt}_*(\mathcal{V})$ is then the set of admissible paths that are $*$-optimal in $\mathcal{V}$. The principle of optimality states that the optimality operators associated with the various $\mathcal{U}(x,k)$ should be related in a special way.

Definition 8 (Principle of Optimality) *$*$-optimality satisfies the* principle of optimality *if it holds for all $k \in [0,N]$, $x \in \mathcal{X}_k$, $\ell \in [k,N]$ and $(x,k,u_\cdot)$ in $\mathcal{U}(x,k)$ that if $(x,k,u_\cdot)$ is $*$-optimal in $\mathcal{U}(x,k)$, then $(x_\ell,\ell,u_\cdot)$ is $*$-optimal in $\mathcal{U}(x_\ell,\ell)$.*

This may also be expressed as:

$$\operatorname{opt}_*\left(\mathcal{U}(x,k)\right) \subseteq \bigcup_{(x,k,u_\cdot)_\ell \in \mathcal{U}(x,k)_\ell} (x,k,u_\cdot)_\ell \oplus \operatorname{opt}_*\left(\mathcal{U}(x_\ell,\ell)\right).$$

The Bellman equation now states that applying the optimality operator to the right hand side suffices to achieve equality. (Usually this is stated with $\ell = k+1$.)

Theorem 2 (Bellman Equation) *Let $k \in [0,N]$ and $x \in \mathcal{X}_k$. Assume that $*$-optimality satisfies the principle of optimality, and that the optimality operator* opt_* *for $\mathcal{U}(x,k)$ is insensitive to the omission of non-$*$-optimal elements from $\mathcal{U}(x,k)$. Then for all $\ell \in [k,N]$:*

$$\mathrm{opt}_*\left(\mathcal{U}(x,k)\right) = \mathrm{opt}_* \bigcup_{(x,k,u)_\ell \in \mathcal{U}(x,k)_\ell} (x,k,u)_\ell \oplus \mathrm{opt}_*\left(\mathcal{U}(x_\ell,\ell)\right),$$

that is, a path is $*$-optimal if and only if it is a $*$-optimal concatenation of an admissible path $(x,k,u.)_\ell$ and a $*$-optimal path of $\mathcal{U}(x_\ell,\ell)$.

Proof. Fix k in $[0,N]$, $\ell \in [k,N]$ and $x \in \mathcal{X}_k$. Define

$$\mathcal{V}_1 = \bigcup_{(x,k,u)_\ell \in \mathcal{U}(x,k)_\ell} (x,k,u)_\ell \oplus \mathrm{opt}_*\left(\mathcal{U}(x_\ell,\ell)\right), \text{ and,}$$

$$\mathcal{V}_2 = \bigcup_{(x,k,u)_\ell \in \mathcal{U}(x,k)_\ell} (x,k,u)_\ell \oplus \left(\mathcal{U}(x_\ell,\ell) \setminus \mathrm{opt}_*\left(\mathcal{U}(x_\ell,\ell)\right)\right).$$

Obviously, $\mathcal{U}(x,k) = \mathcal{V}_1 \cup \mathcal{V}_2$ and $\mathcal{V}_1 \cap \mathcal{V}_2 = \emptyset$. We have to prove that $\mathrm{opt}_*\left(\mathcal{U}(x,k)\right) = \mathrm{opt}_*\left(\mathcal{V}_1\right)$. By the principle of optimality, no path in $\mathcal{V}_2$ is $*$-optimal in $\mathcal{U}(x,k)$, so $\mathcal{V}_2 \cap \mathrm{opt}_*\left(\mathcal{U}(x,k)\right) = \emptyset$. This implies that $\mathrm{opt}_*\left(\mathcal{U}(x,k)\right) \subseteq \mathcal{V}_1 \subseteq \mathcal{U}(x,k)$, and since opt_* is assumed to be insensitive to the omission of non-$*$-optimal elements from $\mathcal{U}(x,k)$, it follows that $\mathrm{opt}_*\left(\mathcal{U}(x,k)\right) = \mathrm{opt}_*\left(\mathcal{V}_1\right)$. □

3.2 $\underline{P}$-Maximality

Let us now apply these general results to the specific types of optimality introduced before. We first consider the optimality operator $\mathrm{opt}_{>_{\underline{P}}}$ that selects from a set of gambles (or paths) S those gambles (or paths) that are the maximal elements of S with respect to the strict partial order $>_{\underline{P}}$. The following lemma roughly states that the preference amongst paths with respect to $>_{\underline{P}}$ is preserved under concatenation and truncation. It yields a sufficient condition for the principle of optimality with respect to $\underline{P}$-maximality to hold. Moreover, the lemma, and the principle of optimality, do not necessarily hold for preference with respect to $\underline{P}$-maximinity.

Lemma 1 *Let $k \in [0,N]$ and $\ell \in [k,N]$. Consider the paths $(x,k,u.)_\ell$ in $\mathcal{U}(x,k)_\ell$ and $(x_\ell,\ell,v.)$, $(x_\ell,\ell,w.)$ in $\mathcal{U}(x_\ell,\ell)$. Then $(x_\ell,\ell,v.) >_{\underline{P}} (x_\ell,\ell,w.)$ if and only if $(x,k,u.)_\ell \oplus (x_\ell,\ell,v.) >_{\underline{P}} (x,k,u.)_\ell \oplus (x_\ell,\ell,w.)$.*

Proof. Let X, Y and Z be gambles on Ω. The statement is proven if we can show that $Y >_{\underline{P}} Z$ implies $X+Y >_{\underline{P}} X+Z$. Assume that $Y >_{\underline{P}} Z$. If $\underline{P}(Y-Z) > 0$, then $\underline{P}((X+Y)-(X+Z)) = \underline{P}(Y-Z) > 0$. If $Y \geq Z$, then $X+Y \geq X+Z$, and finally, if $Y \neq Z$, then $X+Y \neq X+Z$. It follows that $X+Y >_{\underline{P}} X+Z$. □

Proposition 3 (Principle of Optimality) *Let $k \in [0,N]$, $x \in \mathcal{X}_k$ and $(x,k,u_\cdot^*) \in \mathcal{U}(x,k)$. If $(x,k,u_\cdot^*)$ is $\underline{P}$-maximal in $\mathcal{U}(x,k)$ then $(x_\ell,\ell,u_\cdot^*)$ is $\underline{P}$-maximal in $\mathcal{U}(x_\ell,\ell)$ for all $\ell \in [k,N]$.*

Proof. If $(x_\ell,\ell,u_\cdot^*)$ is not $\underline{P}$-maximal, there is a path $(x_\ell,\ell,u_\cdot)$ such that $(x_\ell,\ell,u_\cdot) >_{\underline{P}} (x_\ell,\ell,u_\cdot^*)$. By Lemma 1 we find that

$$(x,k,u_\cdot^*)_\ell \oplus (x_\ell,\ell,u_\cdot) >_{\underline{P}} (x,k,u_\cdot^*)_\ell \oplus (x_\ell,\ell,u_\cdot^*) = (x,k,u_\cdot^*).$$

This means that $(x,k,u_\cdot^*)_\ell \oplus (x_\ell,\ell,u_\cdot)$ is preferred to $(x,k,u_\cdot^*)$, and therefore $(x,k,u_\cdot^*)$ cannot be $\underline{P}$-maximal, a contradiction. □

As a direct consequence of Corollary 1 and Proposition 2, we see that if $\mathcal{J}(x,k)$ is compact, then the optimality operator $\mathrm{opt}_{>_{\underline{P}}}$ associated with $\mathcal{U}(x,k)$ is insensitive to the omission of non-$>_{\underline{P}}$-optimal elements. Together with Proposition 3 and Theorem 2, this allows us to infer a Bellman equation for $\underline{P}$-maximality.

Corollary 2 *Let $k \in [0,N]$ and $x \in \mathcal{X}_k$. If $\mathcal{J}(x,k)$ is compact, then for all $\ell \in [k,N]$*

$$\mathrm{opt}_{>_{\underline{P}}}\left(\mathcal{U}(x,k)\right) = \mathrm{opt}_{>_{\underline{P}}} \bigcup_{(x,k,u)_\ell \in \mathcal{U}(x,k)_\ell} (x,k,u)_\ell \oplus \mathrm{opt}_{>_{\underline{P}}}\left(\mathcal{U}(x_\ell,\ell)\right), \tag{3}$$

that is, a path is $\underline{P}$-maximal if and only if it is a $\underline{P}$-maximal concatenation of an admissible path $(x,k,u_\cdot)_\ell$ and a $\underline{P}$-maximal path of $\mathcal{U}(x_\ell,\ell)$.

Corollary 2 results in a procedure to calculate all $\underline{P}$-maximal paths. Indeed, $\mathrm{opt}_{>_{\underline{P}}}(\mathcal{U}(x,N)) = \{u_\emptyset\}$ for every $x \in \mathcal{X}_N$, and $\mathrm{opt}_{>_{\underline{P}}}(\mathcal{U}(x,k))$ can be calculated recursively through Eq. (3). It also provides a method for constructing a $\underline{P}$-maximal feedback: for every $x \in \mathcal{X}_k$, choose any $(x,k,u_\cdot^*(x,k)) \in \mathrm{opt}_{>_{\underline{P}}}(\mathcal{U}(x,k))$. Then $\phi(x,k) = u_k^*(x,k)$ realises a $\underline{P}$-maximal feedback.

3.3 $\mathcal{M}$-Maximality

We now turn to the optimality operator $\mathrm{opt}_{\mathcal{M}}$, satisfying (1). By Proposition 2 and (1), it follows that $\mathrm{opt}_{\mathcal{M}}$ is insensitive to the omission of non-$\mathcal{M}$-maximal elements of $\mathcal{U}(x,k)$ whenever $\mathcal{J}(x,k)$ is compact. By Proposition 3, $\mathrm{opt}_{\mathcal{M}}$ satisfies the principle of optimality (indeed, if a path is $\mathcal{M}$-maximal, then it must be P-maximal for some $P \in \mathcal{M}$, and by the proposition any truncation of it is also P-maximal, hence also $\mathcal{M}$-maximal). This means that the Bellman equation also holds for $\mathcal{M}$-maximality under similar conditions as for $\underline{P}$-maximality. As already mentioned in Section 2.3, both types of optimality coincide if $\mathcal{J}(x,k)$ is convex.

3.4 $\underline{P}$-Maximinity

Finally, we come to the type of optimality associated with the strict partial order $\sqsupset_{\underline{P}}$. It follows from Proposition 2 and the discussion at the end of Section 2.2

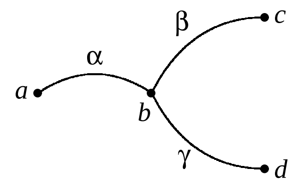


Figure 4: A Counterexample

that if $\mathcal{J}(x,k)$ is compact, the optimality operator $\mathrm{opt}_{\sqsupset_{\underline{P}}}$ for $\mathcal{U}(x,k)$ is insensitive to the omission of non-$\sqsupset_{\underline{P}}$-optimal paths from $\mathcal{U}(x,k)$. But, as the following counterexample shows, we cannot guarantee that the principle of optimality holds for $\sqsupset_{\underline{P}}$-optimality, and therefore dynamic programming may not work here—not even with a vacuous uncertainty model. Essentially, this is because the partial order $\sqsupset_{\underline{P}}$ is not a vector ordering on $\mathcal{L}(\Omega)$—it is not compatible with gain additivity: contrary to expected gain, lower expected gains are not additive.

Example 1 *Consider the dynamical system depicted in Figure 4. Let* $\Omega = \{\sharp, \flat\}$, *let* $\underline{P}$ *be the vacuous lower prevision on* Ω, *and denote the gamble* $\sharp \mapsto x$, $\flat \mapsto y$ *by* $\langle x, y \rangle$. *Assume that* $J(\alpha) = \langle 2, 0 \rangle$, $J(\beta) = \langle 0, -1 \rangle$ *and* $J(\gamma) = \langle -2, 0 \rangle$ *(there is zero gain associated with the final state). Then* $\alpha\beta \not\sqsupset_{\underline{P}} \alpha\gamma$: *indeed,* $\langle 2, -1 \rangle$ *does not dominate* $\langle 0, 0 \rangle$ *point-wise, and* $\inf \langle 2, -1 \rangle \not> \inf \langle 0, 0 \rangle$ *or equivalently* $\langle 0, 0 \rangle$ *maximises the worst expected gain. Hence, we find that* $\alpha\gamma$ *is* $\underline{P}$*-maximin. But* $\beta \sqsupset_{\underline{P}} \gamma$: *indeed,* $\inf \langle 0, -1 \rangle > \inf \langle 0, -2 \rangle$ *which means that* γ *is not* $\underline{P}$*-maximin. Thus the “principle of* $\underline{P}$*-maximin optimality” does not hold here.*

3.5 Yet Another Type of Optimality

We end this discussion with another type of optimality associated with a strict partial order, introduced by Harmanec in [8, Definition 3.4]. In our setting (precisely known system dynamics), its definition basically reduces to

$$X >^{\star}_{\underline{P}} Y \qquad \text{if} \qquad \underline{P}(X) > \overline{P}(Y) \text{ or } (X \geq Y \text{ and } X \neq Y).$$

It can be shown easily that if $\mathcal{J}(x,k)$ is compact, the optimality operator induced by $>^{\star}_{\underline{P}}$ for $\mathcal{U}(x,k)$ is insensitive to the omission of non-$>^{\star}_{\underline{P}}$-optimal paths from $\mathcal{U}(x,k)$. But, as the following counterexample shows, we cannot guarantee that the principle of optimality holds for $>^{\star}_{\underline{P}}$-optimality, and therefore the dynamic programming approach may not work here—not even with a vacuous uncertainty model. Again, this is because the partial order $\sqsupset_{\underline{P}}$ is not compatible with gain additivity. It also indicates that the solution of the Bellman-type equation advocated in [8] will not necessarily lead to optimal paths, in the sense we described above.

Example 2 *Consider the dynamical system depicted in Figure 4. Let* $\Omega = \{\sharp, \flat\}$, *let* $\underline{P}$ *be the vacuous lower prevision on* Ω, *and denote the gamble* $\sharp \mapsto x$, $\flat \mapsto y$

by $\langle x,y\rangle$. Assume that $J(\alpha)=\langle 2,0\rangle$, $J(\beta)=\langle 0,0\rangle$ and $J(\gamma)=\langle -1,-1\rangle$ (there is zero gain associated with the final state). Then $\alpha\beta \not>^{\star}_{\underline{P}} \alpha\gamma$: indeed, $\langle 2,0\rangle$ does not dominate $\langle 1,-1\rangle$ point-wise, and, $\inf\langle 2,0\rangle \not> \sup\langle 1,-1\rangle$. Hence, we find that $\alpha\gamma$ is $>^{\star}_{\underline{P}}$-maximal. But $\beta >^{\star}_{\underline{P}} \gamma$: indeed, $\langle 0,0\rangle$ dominates $\langle -1,-1\rangle$ point-wise, which means that γ is not $>^{\star}_{\underline{P}}$-maximal. Thus the "principle of $>^{\star}_{\underline{P}}$-maximal optimality" does not hold for this example.

4 Conclusion

The main conclusion of our work is that the method of dynamic programming can be extended to systems with imprecise gain. Our general study of what conditions a generalised notion of optimality should satisfy for the Bellman approach to work is of some interest in itself too. In particular, besides an obvious extension of the well-known principle of optimality, another condition emerges that relates to the nature of the optimality operators *per se*: the optimality of a path should be invariant under the omission of non-optimal paths from the set of paths under consideration. If optimality is induced by a strict partial ordering of paths, then this second condition is satisfied whenever the existence of dominating optimal paths for non-optimal ones is guaranteed.

Another important observation is that, in contradistinction to $\underline{P}$-maximality and $\mathcal{M}$-maximality, the dynamic programming method cannot be used to solve optimisation problems corresponding to $\underline{P}$-maximinity: for this notion the principle of optimality does not hold in general.

Throughout the paper we assumed the system dynamics to be deterministic, that is, independent of ω. This greatly simplifies the discussion, still encompasses a large number of interesting applications, and does not suffer from the computational issues often encountered when dealing with non-deterministic dynamical systems—simply because in general the number of possible (random) paths tends to grow exponentially with the size of the state space $\mathcal{X}$. However, we should note that dropping this assumption still leads to a Bellman-type equation, connecting operators of optimality associated with *random* states $x\colon \Omega\to\mathcal{X}$. A discussion of these matters has been omitted from the present paper due to limitations of space.

Acknowledgements

This paper presents research results of project G.0139.01 of the Fund for Scientific Research, Flanders (Belgium), and of the Belgian Programme on Interuniversity Poles of Attraction initiated by the Belgian state, Prime Minister's Office for Science, Technology & Culture. The scientific responsibility rests with the authors.

References

[1] BELLMAN, R. *Dynamic Programming*. Princeton University Press, Princeton, 1957.

[2] BERGER, J. O. The robust Bayesian viewpoint. In *Robustness of Bayesian Analyses*, J. B. Kadane, Ed. Elsevier Science, Amsterdam, 1984.

[3] BERNARDO, J. M., H., D. J., V., L. D., AND SMITH, A. F. M., Eds. *Bayesian Statistics*. University Press, Valencia, 1980.

[4] CHEVÉ, M., AND CONGAR, R. Optimal pollution control under imprecise environmental risk and irreversibility. *Risk Decision and Policy 5* (2000), 151–164.

[5] DE COOMAN, G., COZMAN, F. G., MORAL, S., AND WALLEY, P., Eds. *ISIPTA '99 – Proceedings of the First International Symposium on Imprecise Probabilities and Their Applications* (Ghent, 1999), Imprecise Probabilities Project.

[6] DE FINETTI, B. *Theory of Probability: a Critical Introductory Treatment*. Wiley, London, 1975.

[7] GIRON, F. J., AND RIOS, S. Quasi-Bayesian behaviour: A more realistic approach to decision making? In Bernardo et al. [3], pp. 17–38.

[8] HARMANEC, D. Generalizing Markov decision processes to imprecise probabilities. *Journal of Statistical Planning and Inference 105*, 1 (June 2002), 199–213.

[9] LEVI, I. *The Enterprise of Knowledge. An Essay on Knowledge, Credal Probability, and Chance*. MIT Press, Cambridge, 1983.

[10] UTKIN, L. V., AND GUROV, S. V. Imprecise reliability models for the general lifetime distribution classes. In de Cooman et al. [5], pp. 333–342.

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

Gert de Cooman is a member of the SYSTeMS research group of Ghent University.
Address: Technologiepark – Zwijnaarde 914, B-9052, Zwijnaarde, Belgium.
E-mail: gert.decooman@ugent.be

Matthias Troffaes is a member of the SYSTeMS research group of Ghent University.
Address: Technologiepark – Zwijnaarde 914, B-9052, Zwijnaarde, Belgium.
E-mail: matthias.troffaes@ugent.be

Computing Lower Expectations with Kuznetsov's Independence Condition[*]

FABIO GAGLIARDI COZMAN
Escola Politécnica, University of São Paulo, Brazil

Abstract

Kuznetsov's condition says that variables X and Y are independent when any product of bounded functions $f(X)$ and $g(Y)$ behaves in a certain way: the interval of expected values $\mathbb{E}[f(X)g(Y)]$ must be equal to the interval product $\mathbb{E}[f(X)] \times \mathbb{E}[g(Y)]$. The main result of this paper shows how to compute lower expectations using Kuznetsov's condition. We also generalize Kuznetsov's condition to conditional expectation intervals, and study the relationship between Kuznetsov's conditional condition and the semi-graphoid properties.

Keywords

sets of probability distributions, lower expectations, probability intervals, expectation intervals, independence concepts

1 Introduction

Kuznetsov's condition says that two variables X and Y are independent if, for any two bounded functions $f(X)$ and $g(Y)$, we have

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)] \times \mathbb{E}[g(Y)], \tag{1}$$

where $\mathbb{E}[\cdot]$ denotes an interval of expected values and the product is understood as interval multiplication [8].

Kuznetsov's condition is geared towards models that represent uncertainty through sets of probability measures and expectation intervals. In those models, Kuznetsov's condition is seen to be more general than the standard definition of stochastic independence. The condition can be viewed as a definition of independence, and also as a constraint to be used when building models that involve imprecise beliefs. The relationship between Kuznetsov's condition and other concepts of independence was analyzed in a previous paper [5]; several results from that publication are used in this paper.

[*] This work has been supported in part by CNPq through grant 300183/98-4, and in part by HP Labs through "Convênio Aprendizado de Redes Bayesianas".

This paper shows how to compute minima and maxima of expected values using Kuznetsov's condition. The main result is a characterization of the largest credal set that complies with Kuznetsov's condition — the "Kuznetsov's extension" of marginal sets. We discuss the computation of lower expectations from Kuznetsov's extensions, and investigate the connection between Kuznetsov's extensions and other extensions used in the literature. Section 4 contains these developments.

We then generalize Kuznetsov's condition to conditional beliefs (Section 5). To clarify the behavior of the resulting condition, we investigate its compliance to the semi-graphoid properties. We show that Kuznetsov's conditional condition satisfies symmetry, redundancy, decomposition and weak union, but fails the contraction property.

Kuznetsov's condition is an interesting tool for modeling independence with imprecise beliefs. This paper provides the basic machinery to manipulate the condition in practice. Section 6 presents our conclusions.

2 Credal sets, lower expectations, extensions

In this section we review the basic concepts necessary for later developments. Consider two random variables X and Y. In this paper all variables have finitely many values. The probability density for X is denoted by $p(X)$, and $E_p[f(X)]$ denotes the expectation of function $f(X)$ with respect to $p(X)$. A non-empty set of probability measures is called a *credal set* [9]; a credal set consisting of densities $p(X)$ is denoted by $K(X)$. A credal set $K(X,Y)$ consisting of joint densities $p(X,Y)$ is called a *joint credal set*. The *lower* and *upper* expectations of function $f(X)$ are respectively $\underline{E}[f(X)] = \min_{p(X)\in K(X)} E_p[f(X)]$ and $\overline{E}[f(X)] = \max_{p(X)\in K(X)} E_p[f(X)]$. The *lower probability* and the *upper probability* of event A are defined similarly. A credal set produces an expectation interval for any bounded function $h(X)$: $\mathbb{E}[h(X)] = \left[\underline{E}[h(X)], \overline{E}[h(X)]\right]$.

There are several concepts of independence that can be applied to credal sets [2, 7]; here we focus on *epistemic independence* and *strong independence*. Variable Y is *epistemically irrelevant* to X if $K(X|y)$ and $K(X)$ have the same convex hull for all possible values of Y (equivalently, $\underline{E}[f(X)|y] = \underline{E}[f(X)]$ for any bounded function $f(X)$ and all possible values of Y). Variables X and Y are *epistemically independent* if X is irrelevant to Y and Y is irrelevant to X. Strong independence focuses instead on decomposition of probability measures [1, 2, 4]: Variables X and Y are *strongly independent* when every extreme point of $K(X,Y)$ satisfies standard stochastic independence of X and Y.

Given marginal credal sets $K(X)$ and $K(Y)$, there may be several credal sets $K(X,Y)$ for which X and Y are independent. Each one of these sets is called an *extension* of $K(X)$ and $K(Y)$. Given marginal sets $K(X)$ and $K(Y)$, their *epistemic extension* (called the *independent natural extension* by Walley) is the largest joint credal set that satisfies epistemic independence with marginals $K(X)$ and $K(Y)$ [13]. Their *strong extension* is the largest joint credal set that satisfies strong independence with marginals $K(X)$ and $K(Y)$ [2, 4]. The term *natural extension* is used to indicate the largest possible extension given whatever constraints on probability and independence are adopted [13].

A credal set $K(X,Y)$ is *finitely generated* when it is a polytope in the space of probability measures — the convex hull of a finite number of probability distributions. Such a set is defined by a finite collection of linear inequalities such as $\sum_{X,Y} h(X,Y)p(X,Y) \geq 0$. In the remainder of this paper, f indicates a function of X, g indicates a function of Y and h indicates a function of X and Y. Similarly, p indicates a density for X, q indicates a density for Y; other densities, such as $p(X,Y)$, are indicated explicitly. We can view functions and probability densities as vectors, so we can write $(fg)\cdot(pq) \geq 0$ instead of $\sum_{X,Y} f(X)g(Y)p(X)q(Y) \geq 0$, using the dot product to produce summation.

Note that any hyperplane $h \cdot p(X,Y) = 0$ goes through the origin. The function/vector h is the normal vector of the hyperplane. If $\underline{E}[h] = 0$, then h defines a *supporting hyperplane* for the credal set. If $\overline{E}[h] = 0$, then $-h$ is a supporting hyperplane. A *face* of a polytope is the intersection of the polytope with a supporting hyperplane; a *facet* is a maximal face distinct of the polytope [11].

To simplify notation, we use the same letter (f, for instance) for a function, a vector (containing the values of a function), a normal vector (orthogonal to an hyperplane), an hyperplane (with the normal vector), or a facet (contained in the hyperplane with the normal vector), depending on the circumstances.

Any function/vector h can be written as $h' + \underline{E}[h]$ or as $-h'' + \overline{E}[h]$, where h' and h'' are supporting hyperplanes that are parallel to h. Consider any supporting hyperplane h' that goes through a vertex V. Take the facets intersecting at V, and the normal vectors to these facets. Then it must be possible to write h' as a linear of these normal vectors.

3 Kuznetsov's condition and Kuznetsov's extension

Kuznetsov's condition is a condition for independence operating on expectations of independent variables [8]. The condition can be expressed either in terms of expectation intervals (Expression (1)), or as

$$\underline{E}[f(X)g(Y)] = \min\left(\begin{array}{c} \underline{E}[f(X)]\,\underline{E}[g(Y)]\,,\underline{E}[f(X)]\,\overline{E}[g(Y)]\,, \\ \overline{E}[f(X)]\,\underline{E}[g(Y)]\,,\overline{E}[f(X)]\,\overline{E}[g(Y)] \end{array} \right). \tag{2}$$

To obtain (2) from (1), we recall that the interval product $[a,b] \times [c,d]$ is equal to

$$[\min(ac,ad,bc,bd), \max(ac,ad,bc,bd)]\,.$$

The following result is used later:

Theorem 1 *For any bounded functions $f(X)$ and $g(Y)$, any extension that satisfies Kuznetsov's condition must contain densities that attain $\underline{E}[f]\,\underline{E}[g]$, $\underline{E}[f]\,\overline{E}[g]$, $\overline{E}[f]\,\underline{E}[g]$, and $\overline{E}[f]\,\overline{E}[g]$.*

Proof. Suppose we have a credal set that satisfies Kuznetsov's condition. Consider a function $h_1 = (f - \underline{E}[f] + \alpha)(g - \underline{E}[g] + \beta)$, where $\alpha, \beta > 0$; then $\underline{E}[h_1] = (\underline{E}[f - \underline{E}[f]] + \alpha)(\underline{E}[g - \underline{E}[g]] + \beta) = \alpha\beta$ for any α, β. But for this to happen, we must have a density $p_1(X,Y)$ such that $E_{p_1}[f] = \underline{E}[f]$ and $E_{p_1}[g] = \underline{E}[g]$ at the same time. The proof can be completed by taking functions $h_2 = (f - \overline{E}[f] + \alpha)(g - \overline{E}[g] + \beta)$, $h_3 = -(f - \overline{E}[f] + \alpha)(g - \underline{E}[g] + \beta)$ and $h_4 = -(f - \underline{E}[f] + \alpha)(g - \overline{E}[g] + \beta)$. □

We can use Kuznetsov's condition to construct credal sets. Suppose we have $K(X)$ and $K(Y)$, and we obtain the information that X and Y satisfy Kuznetsov's condition, without further information on $K(X,Y)$. What can we say about the joint credal set $K(X,Y)$? A reasonable strategy is to focus on the largest joint credal set that satisfies Kuznetsov's condition and has the marginals $K(X)$ and $K(Y)$. This set is referred to as *Kuznetsov's extension* of $K(X)$ and $K(Y)$. It should be noted that a Kuznetsov's extension always exists [5].

Kuznetsov's extensions are smaller than epistemic extensions when all events have positive probability, as in this case Kuznetsov's independence implies epistemic independence — and even when all lower probabilities are larger than zero Kuznetsov's extensions can be strictly smaller than epistemic extensions [5]. A strong extension always satisfies Kuznetsov's condition and is contained in the corresponding Kuznetsov's extension (however, the Kuznetsov's extension can be strictly larger than the strong extension; also, it is possible that a credal set satisfies strong independence but does not satisfy Kuznetsov's condition) [5].

4 Characterizing Kuznetsov's extensions

Suppose we have two binary variables X and Y, and we construct the strong extension of $K(X)$ and $K(Y)$. In this case, it is known that the strong extension and the Kuznetsov's extension of $K(X)$ and $K(Y)$ are identical [5]. A more general result can actually be proved:

Theorem 2 *Consider a binary variable X with credal set $K(X)$, and a variable Y with N values and credal set $K(Y)$ with M vertices; the strong extension and Kuznetsov's extension of $K(X)$ and $K(Y)$ are identical.*

Proof. The strong extension is composed of vertices of the form $p_i(X)q_j(Y)$, where p_i indicates a vertex of $K(X)$ and q_j indicates a vertex of $K(Y)$. If $K(X)$ contains a single point, the result is immediate; suppose that $K(X)$ has two vertices p_1 and p_2 (so there is a function $f_1(X)$ such that $f_1 \cdot p_1 = 0$, and a function $f_2(X)$ such that $f_2 \cdot p_2 = 0$). The strong extension can have at most $2M$ vertices,

all of them with $2N$ components (thus the strong extension lives in $(2N-1)$ dimensional space). Any facet of the strong extension is contained in an hyperplane that is defined by selecting $(2N-1)$ vertices of the strong extension plus the origin. Take a facet and divide its vertices (other than the origin) in two sets:
(i) the set C_1 containing points of the form p_1q_j,
(ii) the set C_2 containing points of the form p_2q_k,
where q_j, q_k are vertices of $K(Y)$. Suppose that C_1 contains more points than C_2. Then we have at most $N-1$ points in C_2; we can always find an hyperplane defined by a function $g(Y)$ that goes through all these points. Thus we can construct a function $f_1(X)g(Y)$ such that

$$\sum_{X,Y} f_1(X)g(Y)p_1(X)q_j(Y) = \left(\sum_X f_1(X)p_1(X)\right)\left(\sum_Y g(Y)q_j(Y)\right) = 0$$

for every point in C_1 and every point in C_2. So the facet is represented by a decomposable function f_1g. The same construction can be followed if C_2 has more elements than C_1, in which case we will arrive at a decomposable function of the form f_2g' for some $g'(Y)$. Thus, any facet of the strong extension is defined by a decomposable hyperplane and consequently is a valid constraint for Kuznetsov's extension. The strong extension must then contain Kuznetsov's extension, and so both are equal. □

The facets generated in the proof of Theorem 2 are of the form $f(X)g(Y)$. A little reflection shows that this function $g(Y)$ must define a supporting hyperplane of $K(Y)$: If g were not a supporting hyperplane of $K(Y)$, there should be a point q_c such that $\sum_Y gq_c \geq 0$ and a point q_d such that $\sum_Y gq_d \leq 0$. But $g \cdot q_c \geq 0$ would imply $(fg)\cdot(p_1q_c) \geq 0$ and $g \cdot q_d \leq 0$ would imply $(fg)\cdot(p_1q_d) \leq 0$, contradicting the fact that fg is a supporting hyperplane for the strong extension. Consequently, the facets of the strong extension in Theorem 2 are defined by decomposable functions that factorize into facets of $K(X)$ and $K(Y)$.

Consider now a more general situation where we have categorical variables X and Y and finitely generated marginal credal sets $K(X)$ and $K(Y)$. Suppose that, instead of trying to compute Kuznetsov's extensions, someone simply constructed the following inequalities:

$$\begin{aligned} \sum_{X,Y} \tilde{f}_i(X)p(X,Y) &\geq 0, \\ \sum_{X,Y} \tilde{g}_j(Y)p(X,Y) &\geq 0, \\ \sum_{X,Y} (\tilde{f}_i(X)\tilde{g}_j(Y))p(X,Y) &\geq 0, \end{aligned} \tag{3}$$

which can be written as

$$\tilde{f}_i \cdot p(X,Y) \geq 0, \quad \tilde{g}_j \cdot p(X,Y) \geq 0, \quad (\tilde{f}_i\tilde{g}_j)\cdot p(X,Y) \geq 0, \tag{4}$$

for all combinations of i and j, where $\tilde{f}_i$ is a facet of $K(X)$ and $\tilde{g}_j$ is a facet of $K(Y)$. Note that any set of densities that satisfies these inequalities will also satisfy $(f'g') \cdot p(X,Y) \geq 0$, where f' and g' are supporting hyperplanes of $K(X)$ and $K(Y)$ respectively.

The next theorem is the main result: it shows how to explicitly construct Kuznetsov's extensions. The proof essentially consists of showing that any inequality required by Kuznetsov's condition is already implied by inequalities (4).

Theorem 3 *Consider a variable X with finitely generated credal set $K(X)$, defined by facets $\tilde{f}_i$, and a variable Y with finitely generated credal set $K(Y)$, defined by facets $\tilde{g}_j$. The Kuznetsov's extension is entirely defined by the facets $\tilde{f}_i$, $\tilde{g}_j$, and $(\tilde{f}_i\tilde{g}_j)$, for all combinations of i and j.*

Proof. Denote by $K_k(X,Y)$ the credal set constructed in the theorem. Every vertex of the strong extension is of the form $p(X)q(Y)$ and consequently satisfies $(\tilde{f}_i\tilde{g}_j) \cdot (pq) \geq 0$. We conclude that the strong extension is contained in $K_k(X,Y)$, thus the expectation intervals generated by the strong extension are contained in the expectation intervals generated by $K_k(X,Y)$. Furthermore, for every decomposable function $f(X)g(Y)$, there is a density in $K_k(X,Y)$ that attains the value prescribed by Kuznetsov's condition, as the strong extension is contained in $K_k(X,Y)$.
Now take two arbitrary bounded functions $f(X)$ and $g(Y)$. There are seven different situations to consider:

1. $\underline{E}[f] \geq 0$, $\underline{E}[g] \geq 0$: Kuznetsov's condition requires that $\underline{E}[fg] = \underline{E}[f]\,\underline{E}[g]$. Write f as $f' + \underline{E}[f]$ (f' is a supporting hyperplane of $K(X)$) and write g as $g' + \underline{E}[g]$ (g' is a supporting hyperplane of $K(Y)$). Then we have: $fg \cdot p(X,Y) = (f' + \underline{E}[f])(g' + \underline{E}[g]) \cdot p(X,Y) = f'g' \cdot p(X,Y) + \underline{E}[f]\,g' \cdot p(X,Y) + \underline{E}[g]\,f' \cdot p(X,Y) + \underline{E}[f]\,\underline{E}[g]$, an expression that is equal to or larger than $\underline{E}[f]\,\underline{E}[g]$ given that $p(X,Y)$ satisfies inequalities (4). This implies that $E_p[fg] \geq \underline{E}[f]\,\underline{E}[g]$ for every $p(X,Y)$ and we obtain $\underline{E}[fg] = \underline{E}[f]\,\underline{E}[g]$ (because the inclusion of the strong extension in $K_k(X,Y)$ guarantees that the equality obtains).

2. $\overline{E}[f] \leq 0$, $\overline{E}[g] \leq 0$: Kuznetsov's condition requires $\underline{E}[fg] = \overline{E}[f]\,\overline{E}[g]$. To show that $E_p[fg] \geq \overline{E}[f]\,\overline{E}[g]$ for every $p(X,Y)$, write f as $-f' + \overline{E}[f]$ and g as $-g' + \overline{E}[g]$ (where f' and g' are appropriate supporting hyperplanes), and then: $fg \cdot p(X,Y) = (-f' + \overline{E}[f])(-g' + \overline{E}[g]) \cdot p(X,Y)$, a quantity that is equal to or larger than $\overline{E}[f]\,\overline{E}[g]$ given inequalities (4).

3. $\underline{E}[f] \geq 0, \underline{E}[g] \leq 0$: Kuznetsov's condition requires $\underline{E}[fg] = \overline{E}[f]\,\underline{E}[g]$. Write $f = f' + \underline{E}[f]$, $f = -f'' + \overline{E}[f]$, and $g = -g' + \overline{E}[g]$ (where f', f'' and g' are appropriate supporting hyperplanes; note that f is written in two different ways) and then $fg \cdot p(X,Y) = f(g' + \underline{E}[g]) \cdot p(X,Y) = ((f' + \underline{E}[f])g' + (-f'' + \overline{E}[f])\underline{E}[g]) \cdot p(X,Y)$, which that is equal to or larger than $\overline{E}[f]\,\underline{E}[g]$.

4. $\underline{E}[f] \leq 0$, $\overline{E}[f] \geq 0$, $\overline{E}[g] \leq 0$: Kuznetsov's condition requires $\underline{E}[fg] = \overline{E}[f]\,\underline{E}[g]$. Write $f = -f' + \overline{E}[f]$, $g = g' + \underline{E}[g]$, and $g = -g'' + \overline{E}[g]$, and then $fg \cdot p(X,Y) = (-f'(-g'' + \overline{E}[g]) + \overline{E}[f]\,(g' + \underline{E}[g])) \cdot p(X,Y)$, which is equal to or larger than $\overline{E}[f]\,\underline{E}[g]$.

5. $\underline{E}[f] \leq 0, \underline{E}[g] \geq 0$: Kuznetsov's condition requires $\underline{E}[fg] = \underline{E}[f]\,\overline{E}[g]$. Write $f = f' + \underline{E}[f]$, $g = g' + \underline{E}[g]$, and $g = -g'' + \overline{E}[g]$, and then $fg \cdot p(X,Y) = (f'(g' + \underline{E}[g]) + \underline{E}[f]\,(-g'' + \overline{E}[g])) \cdot p(X,Y)$, which is equal to or larger than $\underline{E}[f]\,\overline{E}[g]$.

6. $\overline{E}[f] \leq 0$, $\underline{E}[g] \leq 0$, $\overline{E}[g] \geq 0$: Kuznetsov's condition requires that $\overline{E}[fg] = \underline{E}[f]\,\overline{E}[g]$. Write $f = f' + \underline{E}[f]$, $f = -f'' + \overline{E}[f]$, and $g = -g' + \overline{E}[g]$, and $fg \cdot p(X,Y) = (-g'(-f'' + \overline{E}[f]) + \overline{E}[g]\,(f' + \underline{E}[f])) \cdot p(X,Y) = (f''g' + \overline{E}[g]\,f' - \overline{E}[f]\,g') \cdot p(X,Y) + \underline{E}[f]\,\overline{E}[g]$, equal to or larger than $\underline{E}[f]\,\overline{E}[g]$.

7. $\underline{E}[f] \leq 0$, $\overline{E}[f] \geq 0$, $\underline{E}[g] \leq 0$, $\overline{E}[g] \geq 0$: Kuznetsov's condition requires $\underline{E}[fg] = \min(\underline{E}[f]\,\overline{E}[g], \overline{E}[f]\,\underline{E}[g])$. Divide $K_k(X,Y)$ into two sets. Define $K_1(X,Y)$ to contain the distributions in $K_k(X,Y)$ such that $f \cdot p(X,Y) \geq 0$, and $K_2(X,Y)$ to contain the distributions in $K_k(X,Y)$ such that $f \cdot p(X,Y) \leq 0$. The value of $\underline{E}[fg]$ with respect to $K_k(X,Y)$ is the minimum of $\underline{E}[fg]$ with respect to $K_1(X,Y)$ and $K_2(X,Y)$. Following the previous cases, we obtain $\overline{E}[f]\,\underline{E}[g]$ as $\underline{E}[fg]$ with respect to $K_1(X,Y)$, and $\underline{E}[f]\,\overline{E}[g]$ as $\underline{E}[fg]$ with respect to $K_2(X,Y)$. We finally obtain $\underline{E}[fg] = \min(\underline{E}[f]\,\overline{E}[g], \overline{E}[f]\,\underline{E}[g])$.

Thus $K_k(X,Y)$ satisfies Kuznetsov's condition, and Kuznetsov's extension must contain $K_k(X,Y)$ — however Kuznetsov's extension cannot be larger than the set $K_k(X,Y)$, as every inequality (4) is directly required by Kuznetsov's condition. □

Once we know how to construct Kuznetsov's extensions, we can compute $\underline{E}[h(X,Y)]$ for a non-decomposable function $h(X,Y)$:

$$\underline{E}[h(X,Y)] = \min\left(h(X,Y) \cdot p(X,Y)\right), \tag{5}$$

subject to $p(X,Y) \geq 0$, $\sum_{X,Y} p(X,Y) = 1$, and inequalities (4).

The linear program (5) provides the solution to the question, Which (decomposable) constraints to use when computing a lower expectation for Kuznetsov's extension? Theorem 3 proves that inequalities (4) contain all the relevant constraints. Kuznetsov himself seems to have obtained different results, using his condition and additional factorization conditions to define extensions — a framework that led him to prescribe linear programs with infinitely many constraints [8].

Finally, note that we can also use linear programming if we need to compute a conditional lower expectation such as $\underline{E}[h|A]$ for some event A where $\underline{P}(A) > 0$. The computation of $\underline{E}[h|A]$ requires the solution of a fractional linear program that can be performed using the Charnes-Cooper transformation and linear programming [3], using inequalities (4) as the starting point.

5 Kuznetsov’s conditional condition and the semi-graphoid properties

Kuznetsov’s condition does not deal with the concept of conditional independence, but it can certainly be extended to do so. Say that two variables X and Y are independent conditional on Z if, for bounded functions $f(X)$ and $g(Y)$,

$$\mathbb{E}[fg|z] = \mathbb{E}[f|z] \times \mathbb{E}[g|z],$$

for any value of Z (we assume that conditioning events have positive lower probability).

How appropriate is Kuznetsov’s conditional condition as a concept of conditional independence? One way to study concepts of independence is to verify the *semi-graphoid* properties satisfied by the concept [6, 10, 12]. A relation $(X \perp\!\!\!\perp Y \,|\, Z)$ is called a *semi-graphoid* when it satisfies the following axioms:
Symmetry: $(X \perp\!\!\!\perp Y \,|\, Z) \Rightarrow (Y \perp\!\!\!\perp X \,|\, Z)$
Redundancy: $(X \perp\!\!\!\perp Y \,|\, X)$
Decomposition: $(X \perp\!\!\!\perp (W,Y) \,|\, Z) \Rightarrow (X \perp\!\!\!\perp Y \,|\, Z)$
Weak union: $(X \perp\!\!\!\perp (W,Y) \,|\, Z) \Rightarrow (X \perp\!\!\!\perp Y \,|\, (W,Z))$
Contraction: $(X \perp\!\!\!\perp Y \,|\, Z)$ & $(X \perp\!\!\!\perp W \,|\, (Y,Z)) \Rightarrow (X \perp\!\!\!\perp (W,Y) \,|\, Z)$.

Denote by $(X \perp\!\!\!\perp_K Y \,|\, Z)$ the fact that X and Y satisfy Kuznetsov’s condition conditional on Z. The notation $\underline{\overline{E}}[f]$ is used to indicate either $\underline{E}[f]$ or $\overline{E}[f]$, whatever value is required by Kuznetsov’s condition. We have:

Theorem 4 *Kuznetsov’s conditional condition satisfies symmetry, redundancy, weak union and decomposition when applied to credal sets where no event has zero lower probability.*

Proof. Symmetry is immediate, and redundancy follows from $\mathbb{E}[f(X)g(Y)|x_0] = f(x_0)\mathbb{E}[g(Y)|x_0] = \mathbb{E}[f(X)|x_0] \times \mathbb{E}[g(Y)|x_0]$ for any $f(X)$, $g(Y)$, and any x_0. Decomposition follows from the fact that any function of Y is also a function of Y and W, so we have $\mathbb{E}[f(X)g(Y)|z] = \mathbb{E}[f(X)|z] \times \mathbb{E}[g(Y)|z]$ when $(X \perp\!\!\!\perp_K (W,Y) \,|\, Z)$. To simplify the proof of the weak union property, the conditioning variable Z is suppressed. What must be shown is that $\underline{E}[fg|w] = \underline{\overline{E}}[f]\underline{\overline{E}}[g|w]$ follows from $\underline{E}[fh] = \underline{\overline{E}}[f]\underline{\overline{E}}[h]$, where h is any function of W and Y (note that $\underline{\overline{E}}[f] = \underline{\overline{E}}[f|w]$ by hypothesis, as events have positive lower probability). Theorem 1 can be easily modified to prove that any credal set satisfying Kuznetsov’s condition must contain densities that attain $\underline{E}[f]\,\underline{E}[g|w]$, $\underline{E}[f]\,\overline{E}[g|w]$, $\overline{E}[f]\,\underline{E}[g|w]$ and $\overline{E}[f]\,\overline{E}[g|w]$; thus, there is always a density p in a set that satisfies Kuznetsov’s condition such that $E_p[fg|w] = \underline{E}[fg|w]$, where $\underline{E}[fg|w]$ follows Kuznetsov’s condition. Take Kuznetsov’s extension of $K(X)$ and $K(W,Y)$, denoted by $K_k(W,X,Y)$. This extension must be equal to or larger than any set satisfying $X \perp\!\!\!\perp_K (W,Y)$. If we determine that $\underline{E}[fg|w] \geq \underline{\overline{E}}[f]\underline{\overline{E}}[g|w]$ for $K_k(W,X,Y)$, then automatically we obtain $\underline{E}[fg|w] = \underline{\overline{E}}[f]\underline{\overline{E}}[g|w]$ for any set satisfying $(X \perp\!\!\!\perp_K (W,Y))$, and weak union follows. The Kuznetsov's extension $K_k(W,X,Y)$ satisfies any inequality $h(W,X,Y) \cdot p(W,X,Y) \geq 0$, and so it satisfies $(f(X)g(Y)I_w(W)) \cdot p(W,X,Y) \geq 0$ for any $f(X)$, $g(Y)$ and w. If we consider the conditional distributions $p(X,Y|w)$ obtained from $K_k(W,X,Y)$, they must satisfy $(f(X)g(Y)) \cdot p(X,Y|w) \geq 0$ as this last inequality is obtained by normalizing the previous one. If we were to construct the Kuznetsov's extension of $K(X)$ and $K(Y|w)$, where $K(Y|w)$ is obtained from $K(W,X)$ by conditioning, then this Kuznetsov's extension would also satisfy any inequality $(f(X)g(Y)) \cdot p(X,Y|w) \geq 0$. So, every inequality constraining the Kuznetsov's extension of $K(X)$ and $K(Y|w)$ is also a constraint for the conditional set obtained from $K_k(W,X,Y)$. Thus the former set is equal to or larger than the latter set. Now notice that, for the Kuznetsov's extension of $K(X)$ and $K(Y|w)$, $\underline{E}[fg|w] = \overline{\underline{E}}[f]\overline{\underline{E}}[g|w]$, and so we must have $\underline{E}[fg|w] \geq \overline{\underline{E}}[f]\overline{\underline{E}}[g|w]$ for $K_k(W,X,Y)$. □

Kuznetsov's condition does not imply the contraction property, as the next example shows.

Example 1 *Consider binary variables W, X, and Y, and a credal set $K(W,X,Y)$ with eight vertices such that each vertex decomposes as $p(W|Y)\,p(X)\,p(Y)$. Values of $p(w_0|y_0)$, $p(w_0|y_1)$, $p(x_0)$ and $p(y_0)$ are:*

Vertex	$[p(w_0\|y_0)\,,p(w_0\|y_1)\,,$ $p(x_0)\,,p(y_0)]$	*Vertex*	$[p(w_0\|y_0)\,,p(w_0\|y_1)\,,$ $p(x_0)\,,p(y_0)]$
1	*[0.7,0.4,0.3,0.2]*	*5*	*[0.7,0.4,0.3,0.3]*
2	*[0.7,0.5,0.2,0.2]*	*6*	*[0.7,0.5,0.3,0.3]*
3	*[0.8,0.4,0.2,0.2]*	*7*	*[0.8,0.4,0.3,0.3]*
4	*[0.8,0.5,0.2,0.2]*	*8*	*[0.8,0.5,0.2,0.3]*

It can be verified that the set of marginal densities $K(X,Y)$ contains every combination of $p(x_0)$ and $p(y_0)$, so $K(X,Y)$ is the Kuznetsov's extension for X and Y (Theorem 2). Likewise, $K(W,X|y_0)$ is the Kuznetsov's extension of W and X conditional on y_0, and $K(W,X|y_1)$ is the Kuznetsov's extension of W and X conditional on y_1. Thus the credal set $K(W,X,Y)$ satisfies $(X \perp\!\!\!\perp_K Y)$ and $(X \perp\!\!\!\perp_K W\,|\,Y)$, but it is not true that $X \perp\!\!\!\perp_K (W,Y)$. Take the function $f(X) = [1,2]$ and the function $h(W,Y) = [2,1,1,2]$. Then $\underline{E}[fh] = 2.652$ for $K(W,X,Y)$, but $\underline{E}[f]\,\underline{E}[h] = 1.7 \times 1.54 = 2.61$ — violating Kuznetsov's condition. □

Despite the failure of contraction for generic credal sets, there is an important situation where contraction holds with Kuznetsov's condition.

Theorem 5 *Kuznetsov's conditional condition satisfies the contraction property when applied to credal sets where no events have zero lower probability, and such that the sets $K(X)$, $K(Y)$, and $K(W|Y)$ are separately specified.*

Proof. As the relevant sets are separately specified, minimization can occur separately within each set, so $\underline{E}[f(X)h(W,Y)] = \min E_p[E_p[fh|Y]] = \min E_p[\underline{E}[fh|Y]]$.

As we have $(X \perp\!\!\!\perp_K W \,|\, Y)$, $\underline{E}[f(X)h(W,Y)] = \min E_p\left[\overline{\underline{E}}[f|Y]\overline{\underline{E}}[h|Y]\right]$, and because $X \perp\!\!\!\perp_K Y$, $\underline{E}[f(X)h(W,Y)] = \min \overline{\underline{E}}[f] E_p\left[\overline{\underline{E}}[h|Y]\right] = \overline{\underline{E}}[f]\overline{\underline{E}}[h]$. □

6 Conclusion

A Kuznetsov’s extension can be viewed as a set that “wraps” a strong extension using decomposable hyperplanes. In fact, there is an interesting duality between these two extensions; while the former is constructed with decomposable hyperplanes, the latter is constructed with decomposable measures.

Kuznetsov’s extensions can have complex structures, except when binary variables are present. The fact that the conditional version of Kuznetsov’s condition fails the contraction property is troubling. This failure suggests that it may be hard to simplify multivariate models using only judgements of conditional independence (according to Kuznetsov’s condition), as these judgements are coupled with the contraction property in traditional multivariate probabilistic models [10].

The challenges for the future are to determine when Kuznetsov’s extensions (and derived concepts) are applicable in practice and how to manipulate them efficiently.

References

[1] J. Cano, M. Delgado, and S. Moral. An axiomatic framework for propagating uncertainty in directed acyclic networks. *Int. Journal of Approximate Reasoning*, 8:253–280, 1993.

[2] I. Couso, S. Moral, and P. Walley. Examples of independence for imprecise probabilities. *First Int. Symp. on Imprecise Probabilities and Their Applications*, pages 121–130, Ghent, Belgium, 1999.

[3] F. G. Cozman. Calculation of posterior bounds given convex sets of prior probability measures and likelihood functions. *Journal of Computational and Graphical Statistics*, 8(4):824–838, 1999.

[4] F. G. Cozman. Separation properties of sets of probabilities. *XVI Conf. on Uncertainty in Artificial Intelligence*, pages 107–115, San Francisco, July 2000. Morgan Kaufmann.

[5] F. G. Cozman. Constructing sets of probability measures through Kuznetsov’s independence condition. In *Second Int. Symp. on Imprecise Probabilities and Their Applications*, pages 104–111, Ithaca, New York, 2001.

[6] A. P. Dawid. Conditional independence. *Encyclopedia of Statistical Sciences, Update Volume 2,* pages 146–153. Wiley, New York, 1999.

[7] L. de Campos and S. Moral. Independence concepts for convex sets of probabilities. *XI Conf. on Uncertainty in Artificial Intelligence*, pages 108–115, San Francisco, California, United States, 1995. Morgan Kaufmann.

[8] V. P. Kuznetsov. *Interval Statistical Methods*. Radio i Svyaz Publ., (in Russian), 1991.

[9] I. Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, Massachusetts, 1980.

[10] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, California, 1988.

[11] A. Schrijver. Theory of Linear and Integer Programming. John Wiley and Sons Ltd., New York, 1986.

[12] M. Studeny. Semigraphoids and structures of probabilistic conditional independence. *Annals of Mathematics and Artificial Intelligence*, 21(1):71–98, 1997.

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

Fabio Gagliardi Cozman is with the Engineering School (Escola Politécnica), University of São Paulo, Av. Prof. Mello Moraes, 2231, Cidade Universitária, São Paulo, SP, Brazil, CEP 05508-900. E-mail: fgcozman@usp.br

Geometry of Upper Probabilities[*]

F. CUZZOLIN
Università di Padova, Italy
Politecnico di Milano, Italy

Abstract

In this paper we adopt the geometric approach to the theory of evidence to study the geometric counterparts of the plausibility functions, or upper probabilities. The computation of the coordinate change between the two natural reference frames in the belief space allows us to introduce the dual notion of basic plausibility assignment and understand its relation with the classical basic probability assignment. The convex shape of the plausibility space Π is recovered in analogy to what was done for the belief space, and the pointwise geometric relation between a belief function and the corresponding plausibility vector is discussed. The orthogonal projection of an arbitrary belief function s onto the probabilistic subspace is computed and compared with other significant entities, such as the relative plausibility and mean probability vectors.

Keywords

theory of evidence, belief space, basic plausibility assignment, plausibility space, orthogonal projection

1 Introduction

Uncertainty measures are assuming a mayor role in fields like artificial intelligence and computer vision, where problems requiring formalized reasoning are common. However, during the last decades a number of different descriptions of uncertain state of knowledge have been proposed, as alternatives or extensions of the classical probability theory. The theory of evidence is one of the most popular formalisms, thanks perhaps to its nature of quite natural extension of the classical Bayesian methodology.

In a series of recent works ([7], [6]) we have proposed a geometric interpretation of the theory of evidence based on the notion of *belief space*, the set of all

[*] This work has been supported by the Autonomous Navigation and Computer Vision Lab, Department of Information Engineering, led by professor R. Frezza.

the b.f.s defined on a fixed domain. It is well known that upper and lower probabilities, belief functions, possibility measures, fuzzy sets can be all thought of as *fuzzy measures*. Hence, it would be highly desirable to find a common environment where to discuss and compare all these uncertainty descriptions in an unified fashion.

In this perspective, this paper proposes a geometric picture of the connections between upper and lower probabilities in the belief space framework. After recalling the basic notions of the theory of evidence, we will briefly introduce the geometric approach to the ToE. After computing the change of coordinates between the orthogonal and oblique reference frames in the belief space, the notion of basic plausibility assignment will be defined and its analytic relation with the basic probability assignment unveiled (Section 3). This will allow us to describe the space of all the plausibility vectors as a simplex, called *plausibility space*, and give a natural interpretation of its vertices in terms of degrees of belief.

Next (Section 4) we will try and understand the pointwise geometry of upper probabilities by noticing that the line connecting a belief function s and the corresponding plausibility function P_s^* is *orthogonal* to the Bayesian subspace $\mathcal{P}$. This will allow us to compute the *orthogonal projection* $s_{\perp\mathcal{P}}$ of s onto $\mathcal{P}$ and prove that it is a probability distribution. We will also find the position of the mean probability vector $\frac{s+P_s^*}{2}$ and the condition under which P_s^* is the reflection of s through the probabilistic subspace.

Finally, we will express the credal set of the probabilities consistent with s as a simplex, noticing that its center of mass is the geometric counterpart of the so called *pignistic* transformation, and discuss the geometry of these points in the perspective of the probabilistic approximation problem. To improve the readability of the paper the proofs of the major results have been moved to an appendix.

1.1 Previous work

The geometric approach to the theory of evidence and generalized probabilities is due to the author, even if close references can be the works of Ha and Haddawy [9] and Wang *et al.* [17]. Anyway, some interesting papers have been recently published on the geometry of lower probabilities and plausibilities of singletons. P. Black, in particular, has dedicated its doctoral thesis to the study of belief functions [2]. An abstract of his results on the geometry of belief functions and other monotone capacities can be found in [3], where he uses shapes of geometric loci to give a direct visualization of the distinct classes of monotone capacities. In particular a number of results about lengths of edges of convex sets representing monotone capacities are given, together with their *size* meant as the sum of those lengths.

A number of papers, on the other side, have been published on the approximation of belief functions (see [1] for a review), mainly in order to find efficient implementations of the rule of combination aiming to reduce the number of focal elements (see for instance the works of Tessem [16] and Lowrance *et al.* [11]).

2 Geometric approach to the Theory of Evidence

The *theory of evidence* [13] has been introduced in the late Seventies by Glenn Shafer as a way of representing epistemic knowledge, starting from a sequence of seminal works of Arthur Dempster [8]. In this formalism the best representation of chance is a *belief function* (b.f.) rather than a Bayesian mass distribution. Following Shafer [13] let us call the finite set of possible outcomes for a decision problem *frame of discernment* or simply *frame*. In the following we will denote by A^c the complement of an arbitrary set A, by $A \setminus B \doteq A \cap B^c$ the difference of two sets A and B, and by $|A|$ the cardinality (number of elements) of A.

A *basic probability assignment* (b.p.a.) over a frame Θ is a function $m : 2^\Theta \to [0,1]$ on its power set $2^\Theta = \{A \subset \Theta\}$ such that

$$m(\emptyset) = 0, \ \sum_{A \subset \Theta} m(A) = 1, \ \ m(A) \geq 0 \ \ \forall A \subset \Theta.$$

The subsets of Θ associated with non-zero values of m are called *focal elements* and their union $\mathcal{C}$ *core*.
The *belief function* $s : 2^\Theta \to [0,1]$ associated with a basic probability assignment m is defined as $s(A) = \sum_{B \subset A} m(B)$, while m can be uniquely recovered from s by means of the *Moebius formula*

$$m(A) = \sum_{B \subset A} (-1)^{|A \setminus B|} s(B). \tag{1}$$

In particular, a *Bayesian* belief function s is a belief function such that $m_s(A) = 0$ for all A s.t. $|A| > 1$. Hence, finite probabilities are nothing more than special b.f.s.

Belief functions representing distinct bodies of evidence can be combined by means of the *Dempster’s rule of combination* [8]. The *orthogonal sum* $s_1 \oplus s_2$ of two belief functions is a new belief function whose focal elements are all the possible intersections between the combining focal elements and whose b.p.a. is given by

$$m(C) = \frac{\sum_{i,j: A_i \cap B_j = C} m_1(A_i) m_2(B_j)}{1 - \sum_{i,j: A_i \cap B_j = \emptyset} m_1(A_i) m_2(B_j)}. \tag{2}$$

where $\{A_i\}$ and $\{B_j\}$ are the focal elements of s_1, s_2 respectively.
When all the intersections between focal elements of the two functions are empty, the denominator of Equation (2) goes to zero and we say that s_1 and s_2 are *not combinable*.

A *dual* representation of the evidence encoded by a belief function s is called *upper probability*[1], and expresses the amount of evidence *not against* a proposi-

[1] The name comes from the fact that belief values and upper probability values are respectively lower and upper bounds for the probabilities of the events.

tion A

$$P^*(A) \doteq 1 - s(A^c) = 1 - \sum_{B \subset A^c} m(B) = \sum_{B \cap A \neq \emptyset} m(B) \geq s(A). \tag{3}$$

Now, consider a frame of discernment Θ and introduce in the Euclidean space $\mathbb{R}^{|2^\Theta|-1}$ an orthonormal reference frame $\{X_A\}_{A \subset \Theta, A \neq \emptyset}$ such that each coordinate function x_A measures the belief value associated with the i-th subset of Θ.

Definition 1 *The* belief space *associated with Θ is the set of points $\mathcal{S}_\Theta$ of $\mathbb{R}^{|2^\Theta|-1}$ corresponding to a belief function.*

We usually assume the domain Θ fixed, and denote the belief space by $\mathcal{S}$. Let us call A-th *basis belief function*

$$P_A \doteq s \in \mathcal{S} \; s.t. \; m_s(A) = 1, \; m_s(B) = 0 \; B \neq A$$

the unique belief function assigning all the mass to a single subset A of Θ. It can be proved that (see [7], [6]), calling $\mathcal{E}_s$ the list of focal elements of s,

Theorem 1 *The set of all the belief functions with focal elements in a given collection X is closed and convex in $\mathcal{S}$: $\{s : \mathcal{E}_s \subset X\} = Cl(\{P_A : A \in X\})$.*

The shape of $\mathcal{S}$ follows immediately from Theorem 1.

Corollary 1 *The belief space $\mathcal{S}$ coincides with the convex closure of all the basis belief functions, $\mathcal{S} = Cl(P_A, \; A \subset \Theta, \; A \neq \emptyset)$.*

Moreover, any belief function $s \in \mathcal{S}$ can be written as a convex sum as follows:

$$s = \sum_{A \subset \Theta, \, A \neq \emptyset} m_s(A) \cdot P_A. \tag{4}$$

Clearly, since a probability is a belief function assigning non zero masses to singletons only, Theorem 1 implies that the set $\mathcal{P}$ of all the Bayesian belief functions is a subset of the border of $\mathcal{S}$, precisely $\mathcal{P} = Cl(P_{\{\theta_i\}}, \; i = 1, ..., |\Theta|)$.

3 Geometry of Plausibility Functions

Analogously to what done for the vectors of $\mathbb{R}^N$ ($N \doteq |2^\Theta| - 1$) representing belief functions, we would like to understand the geometric properties of the plausibility vectors $[P^*_s(A), A \subset \Theta]'$. A plausibility vector can indeed be expressed as

$$P^*_s = \sum_{A \subset \Theta} P^*_s(A) \cdot X_A \tag{5}$$

where $\{X_A, A \subset \Theta\}$ is the orthogonal reference frame of the belief space.
The basis belief functions P_A form a set of independent vectors in $\mathbb{R}^N$, so that the

collections $\{X_A\}$ and $\{P_A\}$ form two distinct coordinate frames in the belief space. To understand the place a plausibility vector takes in the belief reference frame $\{P_A\}$ we then need to compute the coordinate change between these frames. We first notice that basis b.f.s can be expressed as $P_A = \sum_{E\supset A} X_E$.

Proposition 1 *The coordinate change between the two coordinate frames $\{X_A\}$ and $\{P_A\}$ is given by*

$$X_A = \sum_{B\supset A} P_B \cdot (-1)^{|B\setminus A|}. \tag{6}$$

3.1 Basic Plausibility Assignment

Let us now replace expression (6) in Equation (5), obtaining for P_s^*[2]

$$\sum_{A\subset\Theta} P_s^*(A)\cdot X_A = \sum_{A\subset\Theta} P_s^*(A)\cdot \sum_{B\supset A} P_B \cdot (-1)^{|B\setminus A|} = \sum_{B\subset\Theta} P_B \cdot \sum_{A\subset B} (-1)^{|B-A|} P_s^*(A)$$

and after introducing the quantity

$$\mu(A) \doteq \sum_{B\subset A} (-1)^{|A-B|} P_s^*(B) \tag{7}$$

we can write

$$P_s^* = \sum_{A\subset\Theta} \mu(A)\cdot P_A. \tag{8}$$

We call the function $\mu : 2^\Theta \to \mathbb{R}$ defined by expression (7) *basic plausibility assignment*. It is easy to recognize the Moebius equation for plausibilities, which implies $P_s^*(A) = \sum_{B\subset A}\mu(B)$. A few calculations allow us to understand the relation between basic probabilities and plausibilities.

Theorem 2

$$\mu(A) = \begin{cases} (-1)^{|A|+1}\sum_{E\supset A} m(E) & A \neq \emptyset \\ 0 & A = \emptyset. \end{cases} \tag{9}$$

It is easy to see that basic plausibility assignments *meet the normalization constraint*. In fact

$$\sum_{A\subset\Theta} \mu(A) = -\sum_{A\subset\Theta, A\neq\emptyset} (-1)^{|A|} \sum_{E\supset A} m(E) = -\sum_{E\subset\Theta} m(E)\cdot \sum_{A\subset E, A\neq\emptyset} (-1)^{|A|} = 1$$

since $-\sum_{A\subset E, A\neq\emptyset}(-1)^{|A|} = -(0-(-1)^0) = 1$ for the expression of Newton's binomial $\sum_{k=0}^n \binom{n}{k} p^k q^{n-k} = (p+q)^n$, where in this case $k = |A|$, $p = -1$, $q = 1$. However, $\mu(A)$ is not always positive, so we can just infer that any plausibility vector lies on the affine subspace generated by the basis belief functions $\{P_A\}$.

[2] Note that $P_s^*(\emptyset) = 0$ so the expression is correct even if $X_\emptyset$ does not exist.

3.2 Plausibility Space

Analogously to what done for belief functions, let us call *plausibility space* the region Π of $\mathbb{R}^N$ whose points correspond to admissible plausibility functions. It is not difficult to prove that

Theorem 3 Π *is a simplex*

$$\Pi = Cl(\Pi_A, A \subset \Theta, A \neq \emptyset), \quad \Pi_A = -\sum_{B\subset A}(-1)^{|B|}P_B. \tag{10}$$

Proof. We just need to re-assemble expression (8) as a convex combination of points, getting (through Equation (9))

$$\begin{aligned} P^*_s &= \sum_{A\subset\Theta}\mu(A)\cdot P_A = \sum_{A\subset\Theta, A\neq\emptyset}(-1)^{|A|+1}\cdot\sum_{E\supset A}m(E)\cdot P_A = \\ &= \sum_{A\subset\Theta,A\neq\emptyset}\sum_{E\supset A}(-1)^{|A|+1}m(E)\cdot P_A = \sum_{E\subset\Theta,E\neq\emptyset}m(E)\cdot\sum_{A\subset E,A\neq\emptyset}(-1)^{|A|+1}P_A \end{aligned}$$

$= \sum_{E\neq\emptyset}m(E)\Pi_E$, that is a convex combination since basic probability assignments have unitary sum. □

It is easy to notice that $\Pi_{\{\theta\}} = -(-1)^{|\{\theta\}|}\cdot P_{\{\theta\}} = P_{\{\theta\}}\forall\theta\in\Theta$, so that $\mathcal{P}\subset\mathcal{S}\cap\Pi$. The inverse relation between basis belief functions and basis plausibilities has the same form of Equation (10):

Theorem 4

$$P_A = -\sum_{B\subset A}(-1)^{|B|}\cdot\Pi_B. \tag{11}$$

Proof. The proof follows the sketch of Proposition 1. Replacing expression (11) in Equation (10) yields for Π_A

$$-\sum_{B\subset A}(-1)^{|B|}P_B = \sum_{B\subset A}(-1)^{|B|}\cdot\sum_{E\subset B}(-1)^{|E|}\Pi_E = \sum_{E\subset A}(-1)^{|E|}\Pi_E\cdot\sum_{E\subset B\subset A}(-1)^{|B|}$$

but then, analogously to what previously done (see the Appendix),

$$\sum_{E\subset B\subset A}(-1)^{|B|} = \begin{cases}(-1)^{|A|} & E = A\\ 0 & E\neq A\end{cases}$$

and the thesis easily follows. □

The vertices of the plausibility space have a natural interpretation.

Theorem 5 *The vertex* Π_A *of the plausibility space is the plausibility vector associated with the basis belief function* P_A, $\Pi_A = P^*_{P_A}$.

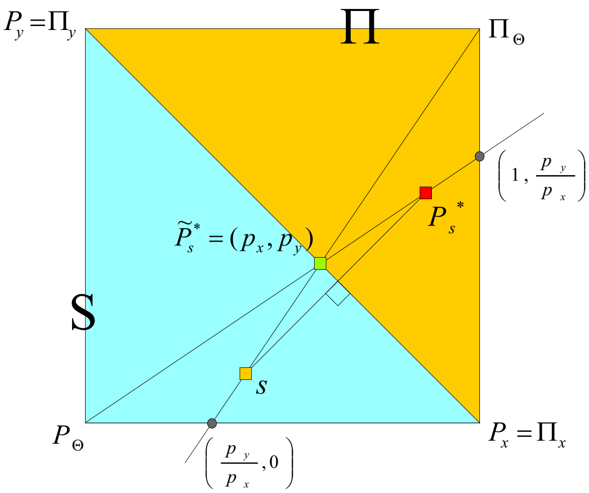

$P_y = \Pi_y$
Π
Π_Θ
$\left(1, \frac{p_y}{p_x}\right)$
P_s^*
$\tilde{P}_s^* = (p_x, p_y)$
S
s
P_Θ
$P_x = \Pi_x$
$\left(\frac{p_y}{p_x}, 0\right)$

Figure 1: Geometric relations between upper and lower probabilities in the belief space for a binary frame $\Theta = \{x, y\}$. The belief space $\mathcal{S}$ and the plausibility space Π are both simplices with vertices $\{P_\Theta = (0,0), P_x = (1,0), P_y = (0,1)\}$ and $\{\Pi_\Theta = (1,1), \Pi_x = P_x, \Pi_y = P_y\}$ respectively. In the picture a belief function s and the corresponding plausibility function P_s^* are indicated, showing that they are in symmetric positions with respect to the common subspace $\mathcal{P}$. The location of the relative plausibility of singletons $\tilde{P}_s^*$ is also shown, as intersection of the probabilistic subspace with the line joining P_s^* and $P_\Theta = (0,0)$. A dual line joining s and Π_Θ also appears.

Figure 1 shows the relation between belief and plausibility space for a the binary frame $\Theta = \{x, y\}$. Without reporting the calculations, we may notice another few interesting facts. The two simplices are perfectly symmetric with respect to the probabilistic subspace. Furthermore, upper and lower probability vectors determine a line that is orthogonal to $\mathcal{P}$, and they also lie on symmetric positions with respect to the Bayesian region. Notice that the relative plausibility vector $\tilde{P}_s^*$ (normalized version of P_s^*) does not coincide at all with the orthogonal projection of s (or P_s^*) onto $\mathcal{P}$. In the following we will try and understand what of those features retain their validity in the general case.

4 Upper and lower probability vectors

It is in fact natural to wonder what is the pointwise relation between vectors representing upper and lower probability functions generated by the same evidence.

Luckily enough, orthogonality turns out to be an actual property of those uncertainty descriptions.

4.1 Orthogonal projection

Let us first denote with P_x the basis belief function for $A = \{x\}$. Being $\mathcal{P} = Cl(P_x, x \in \Theta)$ an affine subspace, it can be written as the translated version of a vector space as $\mathcal{P} = P_x + span(P_y - P_x, \forall y \in \Theta, y \neq x)$, where the $n-1$ vectors $P_y - P_x$ form a basis of this vector space. They show a peculiar symmetry

$$P_y - P_x(A) = \begin{cases} 1 & A \supset \{y\}, A \not\supset \{x\} \\ 0 & A \supset \{x\}, \{y\} \; or \; A \not\supset \{x\}, \{y\} \\ -1 & A \not\supset \{y\}, A \supset \{x\}. \end{cases}$$

that can be usefully exploited for our goals. In particular, we can appreciate that

$$(P_y - P_x)(A) = 1 \Rightarrow A \supset \{y\}, A \not\supset \{x\} \Rightarrow A^c \supset \{x\}, A^c \not\supset \{y\} \Rightarrow (P_y - P_x)(A^c) = -1$$

and vice-versa, while $(P_y - P_x)(A) = 0 \Rightarrow A \supset \{y\}, A \supset \{x\}$ or $A \not\supset \{y\}, A \not\supset \{x\}$ so that in the first case $A^c \not\supset \{x\}, \{y\}$, in the second one $A^c \supset \{x\}, \{y\}$ but in both situations $(P_y - P_x)(A^c) = 0$. Summarizing we can write

$$(P_y - P_x)(A^c) = -(P_y - P_x)(A) \quad \forall A \subset \Theta$$

which directly implies that

Theorem 6 *The line connecting P_s^* and s is orthogonal to the probabilistic subspace, i.e.*

$$s - P_s^* \perp \mathcal{P}.$$

It is then clear that the orthogonal projection of s onto $\mathcal{P}$ is simply the intersection of this line with the probabilistic subspace,

$$s_{\perp\mathcal{P}} = \overrightarrow{sP_s^*} \cap \mathcal{P}.$$

We just have to find the value of α such that $s + \alpha(P_s^* - s) \in \mathcal{P}$.

Theorem 7 *The coordinates of the orthogonal projection of s onto $\mathcal{P}$ with respect to the basis $\{P_A\}$ can be expressed in terms of the basic probability assignment m of s as follows:*

$$m_{s_{\perp\mathcal{P}}}(\{x\}) = m(\{x\}) + \sum_{A \supseteq \{x\}} m(A) \cdot \frac{\sum_{|A|>1} m(A)}{\sum_{|A|>1} m(A)|A|}. \tag{12}$$

Equation (12) ensures that $m_{s_{\perp\mathcal{P}}}(\{x\})$ is always positive for each $x \in \Theta$, so that

Corollary 2 *The orthogonal projection $s_{\perp\mathcal{P}}$ of any arbitrary belief function s onto the probabilistic subspace $\mathcal{P}$ is a Bayesian belief function.*

This fact is not just a trivial consequence of its definition, since the probability simplex is a small region of $span(\mathcal{P})$ in general. A symmetric version of the formula can be obtained after realizing that $\frac{\sum_{|A|=1} m(A)}{\sum_{|A|=1} m(A)|A|} = 1$, so that we can write

$$m_{s_{\perp\mathcal{P}}}(\{x\}) = s(\{x\}) \cdot \frac{\sum_{|A|=1} m(A)}{\sum_{|A|=1} m(A)|A|} + [P^*_s - s](\{x\}) \cdot \frac{\sum_{|A|>1} m(A)}{\sum_{|A|>1} m(A)|A|}. \qquad (13)$$

It is natural to wonder whether the upper probability vector is actually the reflection of the lower probability vector through the probabilistic subspace as in the binary case, i.e. if $s_{\perp\mathcal{P}} = \dfrac{s+P^*_s}{2}$. In [5] we will show that

Proposition 2 *Orthogonal projection and mean probability coincide iff*

$$\sum_{|A|>1} m(A)|A| = 2 \sum_{|A|>1} m(A).$$

This apparently arid result is strictly related to the duality isuue concerning the geometric counterparts of upper and lower probabilities. Is this duality associated with some kind of symmetry through the probabilistic subspace? Further analysis [5] seem to hint that the situation is a bit more complex.

4.2 Simplex of Consistent Probabilities

It is well known, on the other side, that belief functions can be formally interpreted in terms of classes of unknown probabilities. Given the nature of basic probability assignments, it is natural to conjecture that the set of probabilities $P(s)$ consistent with a given belief function s has also the shape of a simplex. Is there any relation between the orthogonal projection of s onto $\mathcal{P}$ and this simplex?

Following Shafer [13] we can think of $m(A)$ as a probability free to move inside A. If we assign the mass of each focal element A_i to one of its elements a_i, intuitively we should get an extremum of the region of consistent probabilities. More formally, to each focal element A corresponds a mass $m(A)$ distributed among its elements, $m(A) \cdot Cl(P_a,\, a \in A)$, so that $P(s)$ can be expressed as

$$P(s) = \sum_{A \subset \Theta} m(A) \cdot Cl(P_a,\, a \in A).$$

Then, given an arbitrary belief function s with focal elements $A_1, ..., A_m$, we can define for each choice of m representatives $\{a_1, ..., a_m\}$, $a_i \in A_i \; \forall i$,

$$P_{a_1 \dots a_m} \doteq \sum_{i=1}^{m} m(A_i) \cdot P_{a_i}. \qquad (14)$$

It can be proved that [5] (as suggested by our intuition)

Proposition 3

$$P(s) = Cl(P_{a_1...a_m}, \{a_1,...,a_m\} \in A_1 \times ... \times A_m).$$

Accordingly, the center of mass $\bar{P}(s)$ of $P(s)$ gets the form

$$\frac{1}{\prod_i |A_i|} \cdot \sum_{\{a_1,...,a_m\}\in A_1\times...\times A_m} P_{a_1...a_m} = \frac{1}{\prod_i |A_i|} \cdot \sum_{\{a_1,...,a_m\}\in A_1\times...\times A_m} \sum_{i=1}^{m} m(A_i) P_{a_i} = \\ \frac{1}{\prod_i |A_i|} \sum_{a\in\mathcal{C}_s} P_a \sum_{A_j\supset\{a\}} m(A_j) \frac{\prod_i |A_i|}{|A_j|} = \sum_{a\in\mathcal{C}_s} P_a \sum_{A_j\supset\{a\}} \frac{m(A_j)}{|A_j|} = \sum_{x\in\Theta} P_x \sum_{A\supset\{x\}} \frac{m(A)}{|A|} \tag{15}$$

since no focal elements include points outside the core. Equation (15) possesses several interesting interpretations.

4.2.1 Center of mass and pignistic transformation

In his popular *transferable belief model* [15] Philippe Smets has proposed an approach to the theory of evidence in which beliefs are represented at credal level (as convex sets of probabilities or belief functions), while decisions are made by resorting to a probabilistic approximation of belief function called *pignistic transformation* (see for instance [4]). Smets justifies his transformation by means of a so-called “rationality” requirement, which mathematically translates into a linearity constraint (see Theorem 3 of [14]).

It is pretty surprising to see that the pignistic transformation $Pign[s]$ of a belief function s is exactly expressed by Equation (15)

$$Pign[s](x) = \sum_{A\supset\{x\}} \frac{m(A)}{|A|},$$

making clear that the geometric counterpart of the pignistic transformation coincides with the center of mass of the simplex $P(s)$ of consistent probabilities. The full implications of this fact are still unclear, and deserve further investigations.

4.2.2 Consistency and Epsilon Contamination

The geometric analysis of the convex region of the consistent probabilities can be also related to a popular technique in robust statistics, the Epsilon Contamination Model. For a fixed $0 < \varepsilon < 1$ and a probability distribution P^*, the associated ε-contamination model is a convex class of distributions of the form $\{(1-\varepsilon)P^* + \varepsilon Q\}$ where Q is arbitrary.

Teddy Seidenfeld has proved that (for discrete domains) any ε-contamination model is equivalent to a belief function, whose corresponding consistent probabilities form the largest convex set induced by the collection of coherent lower probabilities the model specifies for the elements of the domain (see [12], Theorem 2.10). It is worth noticing that in this special case P^* has the meaning of barycenter of the convex set, providing then another interesting interpretation of Equation (15).

5 Comments

What we have learned about the pointwise geometry of upper and lower probabilities can then be eventually depicted as in Figure 2. Each belief function s is associated with a simplex of consistent probabilities (the shaded triangle) $P(s)$ in the probabilistic subspace $\mathcal{P}$ (the larger triangle), whose center of mass $\overline{P}(s)$ (representing the pignistic transformation of s) is in general different from the orthogonal projection of s onto $\mathcal{P}$. The line $\overline{sP_s^*}$ is orthogonal to $\mathcal{P}$ but s and P_s^* are not on symmetric positions in general.

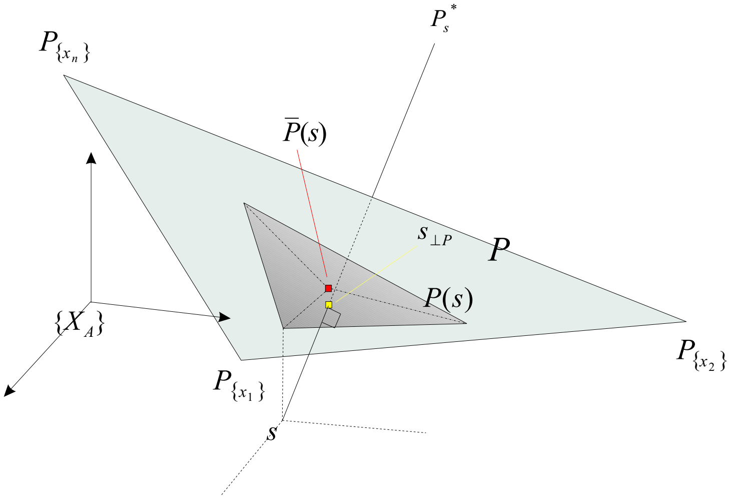


Figure 2: Geometric relation between upper and lower probability vectors.

The binary case turns out to be rather peculiar, since, recalling the definition of basic plausibility assignment (Section 3.1),

$$\begin{aligned}
&\bar{P}(s) = \textstyle\sum_{x\in\Theta_2} P_x \sum_{A\supset x} \frac{m(A)}{|A|} = P_x \cdot (m(x) + \frac{m(\Theta)}{2}) + P_y \cdot (m(y) + \frac{m(\Theta)}{2}),\\
&\tfrac{s+P_s^*}{2} = P_x \cdot \tfrac{m(x)+m(x)+m(\Theta)}{2} + P_y \cdot \tfrac{m(y)+m(y)+m(\Theta)}{2} +\\
&\qquad +P_\Theta \cdot \tfrac{m(\Theta)-m(\Theta)}{2} = P_x \cdot (m(x) + \tfrac{m(\Theta)}{2}) + P_y \cdot (m(y) + \tfrac{m(\Theta)}{2}),\\
&s_{\perp\mathcal{P}} = P_x \cdot [m(x) + (1 - m(y) - m(x)) \cdot \tfrac{m(\Theta)}{2m(\Theta)}] + P_y \cdot [m(y) + \tfrac{1-m(x)-m(y)}{2}]\\
&\qquad = P_x \cdot (m(x) + \tfrac{m(\Theta)}{2}) + P_y \cdot (m(y) + \tfrac{m(\Theta)}{2})
\end{aligned}$$

and these three quantities coincide.

In our vision this knowledge could represent a step towards a more comprehensive understanding of the various uncertainty measures that can be introduced on finite domains: classical probabilities, upper and lower probabilities, belief functions, possibility measures, fuzzy sets. A number of papers have been recently published, for instance, on the connection between fuzzy measures and belief functions ([10] among the others). The belief space framework could provide a unifying environment where those connections may emerge more clearly and lead to a better comprehension of the field.

In this paper, in particular, we have seen how the dual concept of plausibility function or upper probability transfer into a dual convex geometry. The analogous of basis belief functions and probability assignments have been developed and their geometric interpretation exposed. We concentrated our efforts on understanding the pointwise relation between lower and upper probability vectors, proving their orthogonality with respect to the probabilistic subspace.

We also analyzed the comparative geometry of relative plausibility, orthogonal projection and center of mass of the set of consistent probabilities. This can be seen as a preliminary work in the perspective of a geometric solution to the probabilistic approximation problem. Coherently, we are also working on the geometry of finite fuzzy sets and possibility measures, to investigate more closely the idea of *duality* between probabilistic and possibilistic measures and discuss possible alternative consonant approximations of belief functions.

From a purely technical viewpoint, it is not clear yet what is the exact position in the belief space of a generic plausibility vector, and its geometric relation with other significant points like the relative plausibility of singletons $\tilde{P}_s^*$. In the next future [5] we will show how this quantity turns out to be the best Bayesian approximation of a belief function in the framework of Dempster's combination rule, and "perfectly" represents (in a very precise way) the original belief function in probabilistic subspace. It will be interesting to compare these findings with the results of a recent working paper Cobb and Shenoy [4], where they describe some properties of the relative plausibility of singletons and discuss its nature of probability function that is equivalent to the original belief function.

The study of consistent probabilities could play as well an important role in the search for an alternative to Dempster's rule of combination, for their description in terms of convex sets opens the way to the application of our commutativity results [6]. Understanding their behavior in an inference process could give us a hint of the properties a combination rule should possess to guarantee coherency in terms of the corresponding credal sets.

Appendix: Mathematical Proofs

Proof. **(Proposition 1)** If the thesis is true we have, by replacing X_A with expression (6),

$$P_A = \sum_{E \supset A} X_E = \sum_{E \supset A} \sum_{B \supset E} P_B \cdot (-1)^{|B-E|} = \sum_{B \supset A} P_B \cdot \sum_{B \supset E \supset A} (-1)^{|B-E|}.$$

Let us consider the factor $\sum_{A \subset E \subset B} (-1)^{|B-E|}$. When $A = B$ then $E = A = B$ and the coefficient becomes 1. On the other side, when $B \neq A$ we have

$$\sum_{A \subset E \subset B} (-1)^{|B-E|} = \sum_{F \subset B \setminus A} (-1)^{|B \setminus A \setminus F|} = 0$$

for Newton’s binomial. Hence $P_A = P_A$. □

Proof. **(Theorem 2)** The definition (3) of upper probability yields

$$\begin{aligned} \mu(A) &= \sum_{B \subset A} (-1)^{|A-B|} P^*_s(B) = \sum_{B \subset A} (-1)^{|A-B|} (1 - s(B^c)) = \\ &= \sum_{B \subset A} (-1)^{|A-B|} - \sum_{B \subset A} (-1)^{|A-B|} s(B^c) \end{aligned} \tag{16}$$

where for Newton’s binomial $\sum_{B \subset A} (-1)^{|A \setminus B|} = 0$ if $A \neq \emptyset$, $(-1)^{|A|}$ otherwise. If $B \subset A$ then $B^c \supset A^c$, so that the second addendum becomes

$$\begin{aligned} &- \sum_{B \subset A, B \neq \emptyset} (-1)^{|A-B|} \sum_{E \subset B^c} m(E) = - \sum_{E \subset \Theta} m(E) \cdot \sum_{B: B \subset A, B^c \supset E} (-1)^{|A-B|} = \\ &= - \sum_{E \subset \Theta} m(E) \cdot \sum_{B \subset A \cap E^c} (-1)^{|A-B|} \end{aligned} \tag{17}$$

for $B^c \supset E, B \subset A$ is equivalent to $B \subset E^c, B \subset A \equiv B \subset (A \cap E^c)$.
Let us now analyze the function of E

$$f(E) \doteq \sum_{B \subset A \cap E^c} (-1)^{|A-B|}.$$

If $A \cap E^c = \emptyset$ then $B = \emptyset$ and the sum is $(-1)^{|A|}$. If $A \cap E^c \neq \emptyset$, instead, we can write $F \doteq E^c \cap A$ and obtain (since $B \subset F \subset A$ and $|A - B| = |A - F| + |F - B|$)

$$f(E) = \sum_{B \subset F} (-1)^{|A-B|} = \sum_{B \subset F} (-1)^{|A-F|+|F-B|} = (-1)^{|A-F|} \cdot \sum_{B \subset F} (-1)^{|F-B|} = 0$$

given that $\sum_{B\subset F}(-1)^{|F-B|} = 0$ for Newton's binomial again. Eventually

$$f(E) = \begin{cases} 0 & E^c \cap A \neq \emptyset \\ (-1)^{|A|} & E^c \cap A = \emptyset. \end{cases}$$

We can then rewrite expression (17) as follows

$$\begin{aligned} -\sum_{E\subset\Theta} m(E)f(E) &= -\sum_{E:E^c\cap A\neq\emptyset} m(E)\cdot 0 - \sum_{E:E^c\cap A=\emptyset} m(E)\cdot(-1)^{|A|} = \\ &= (-1)^{|A|+1}\sum_{E:E^c\cap A=\emptyset} m(E) = (-1)^{|A|+1}\sum_{E\supset A} m(E) \end{aligned}$$

and replacing it in Equation (16) yields Equation (9), after distinguishing the two cases $A = \emptyset$, $A \neq \emptyset$. □

Proof. **(Theorem 5)** Expression (10) is equivalent to $\Pi_A(X) = -\sum\limits_{B\subset A, B\neq\emptyset}(-1)^{|B|}P_B(X)$ $\forall X \subset \Theta$. But since $P_B(X) = 1$ if $X \supset B$ and 0 otherwise we have that

$$\Pi_A(X) = -\sum_{B\subset A, B\subset X, B\neq\emptyset}(-1)^{|B|} = -\sum_{B\subset A\cap X, B\neq\emptyset}(-1)^{|B|}.$$

Now, if $A\cap X = \emptyset$ there is no addenda in the above sum, that goes to zero. Otherwise, for Newton's binomial, we have $\Pi_A(X) = -\{[1+(-1)]^{|A\cap X|} - (-1)^0\} = 1$. But then the definition of upper probability yields exactly

$$P^*_{P_A}(X) = \sum_{B\cap X\neq\emptyset} m_{P_A}(B) = \begin{cases} 1 & A\cap X \neq \emptyset \\ 0 & A\cap X = \emptyset. \end{cases}$$

□

Proof. **(Theorem 6)** Clearly $P^*_s - s = \sum_{A\subset\Theta} X_A \cdot [P^*_s(A) - s(A)]$, where $[P^*_s - s](A^c) = P^*_s(A^c) - s(A^c) = 1 - s(A) - s(A^c) = 1 - s(A^c) - s(A) = P^*_s(A) - s(A) = [P^*_s - s](A)$. Hence,

$$\langle P^*_s - s, P_y - P_x\rangle = \sum_{A\subset\Theta}[P^*_s - s](A)\cdot[P_y - P_x](A) =$$

$$= \sum_{|A|\leq\lfloor|\Theta/2|\rfloor}[P^*_s - s](A)\cdot[(P_y - P_x)(A) - (P_y - P_x)(A^c)] = 0$$

since $(P_y - P_x)(A) = -(P_y - P_x)(A^c)$. □

Proof. **(Theorem 7)** The desired condition implies that, for any subset $A \subset \Theta$, $s(A) + \alpha\cdot[P^*_s(A) - s(A)] = s(A) + \alpha\cdot[1 - s(A^c) - s(A)] \in \mathcal{P}$. In particular, when $A = \{x\}$ is a singleton,

$$s(\{x\}) + \alpha\cdot[1 - s(\{x\}^c) - s(\{x\})] \in \mathcal{P}. \tag{18}$$

This point belongs to $\mathcal{P}$ iff the normalization criterion for singletons is met, i.e.

$$\sum_{x\in\Theta} s(\{x\}) + \alpha \cdot \sum_{x\in\Theta} (1 - s(\{x\}^c) - s(\{x\})) = 1 \Rightarrow \alpha = \frac{1 - \sum_{x\in\Theta} s(\{x\})}{\sum_{x\in\Theta}(1 - s(\{x\}^c) - s(\{x\}))}$$

and after replacing this value of α into Equation (18) we get

$$s_{\perp\mathcal{P}}(\{x\}) = s(\{x\}) + \frac{1 - \sum_{y\in\Theta} s(\{y\})}{\sum_{y\in\Theta}(1 - s(\{y\}^c) - s(\{y\}))} \cdot (1 - s(\{x\}^c) - s(\{x\})) =$$

$$= \frac{s(\{x\}) \cdot [\sum_{y\in\Theta}(1 - s(\{y\}^c) - s(\{y\})) - (1 - \sum_{y\in\Theta} s(\{y\}))]}{\sum_{y\in\Theta}(1 - s(\{y\}^c) - s(\{y\}))} +$$

$$+ \frac{(1 - s(\{x\}^c)) \cdot (1 - \sum_{y\in\Theta} s(\{y\}))}{\sum_{y\in\Theta}(1 - s(\{y\}^c) - s(\{y\}))} =$$

$$= \frac{s(\{x\}) \cdot [\sum_{y\in\Theta}(1 - s(\{y\}^c)) - 1] + (1 - s(\{x\}^c)) \cdot (1 - \sum_{y\in\Theta} s(\{y\}))}{\sum_{y\in\Theta}(1 - s(\{y\}^c) - s(\{y\}))}$$

that using the definition of plausibility function can be rewritten as

$$s_{\perp\mathcal{P}}(\{x\}) = \frac{s(\{x\}) \cdot (\sum_{y\neq x} P_s^*(\{y\}) - 1) + P_s^*(\{x\}) \cdot (1 - \sum_{y\neq x} s(\{y\}))}{\sum_{y\in\Theta}[P_s^*(\{y\}) - s(\{y\})]}. \qquad (19)$$

Equation (19) determines the coordinate of the orthogonal projection of a belief function s onto $\mathcal{P}$. The expression for the basic probability assignment associated with this projection (Equation (12)) can be found after a few passages, extensively reported in [5]. □

References

[1] BAUER, M. Approximation algorithms and decision making in the Dempster-Shafer theory of evidence–an empirical study. *International Journal of Approximate Reasoning 17* (1997), 217–237.

[2] BLACK, P. *An examination of belief functions and other monotone capacities*. PhD dissertation, Department of Statistics, Carnegie Mellon University, 1996. Pgh. PA 15213.

[3] BLACK, P. Geometric structure of lower probabilities. In *Random Sets: Theory and Applications*, Goutsias, Malher, and Nguyen, Eds. Springer, 1997, pp. 361–383.

[4] COBB, B., AND SHENOY, P. On transforming belief function models to probability models. Tech. rep., University of Kansas, School of Business, Working Paper No. 293, February 2003.

[5] CUZZOLIN, F. Probabilistic approximations of belief functions. *in preparation*.

[6] CUZZOLIN, F. Geometrical structure of belief space and conditional subspaces. *submitted to the IEEE Transactions on Systems, Man and Cybernetics part C* (November 2002).

[7] CUZZOLIN, F., AND FREZZA, R. Geometric analysis of belief space and conditional subspaces. In *Proceedings of the* 2^{nd} *International Symposium on Imprecise Probabilities and their Applications (ISIPTA2001)* (26-29 June 2001).

[8] DEMPSTER, A. Upper and lower probabilities generated by a random closed interval. *Annals of Mathematical Statistics 39* (1968), 957–966.

[9] HA, V., AND HADDAWY, P. Theoretical foundations for abstraction-based probabilistic planning. In *Proc. of the* 12^{th} *Conference on Uncertainty in Artificial Intelligence* (August 1996), pp. 291–298.

[10] HEILPERN, S. Representation and application of fuzzy numbers. *Fuzzy Sets and Systems 91* (1997), 259–268.

[11] LOWRANCE, J. D., GARVEY, T. D., AND STRAT, T. M. A framework for evidential-reasoning systems. In *Proceedings of the National Conference on Artificial Intelligence* (1986), A. A. for Artificial Intelligence, Ed., pp. 896–903.

[12] SEIDENFELD, T. Some static and dynamic aspects of rubust Bayesian theory. In *Random Sets: Theory and Applications*, Goutsias, Malher, and Nguyen, Eds. Springer, 1997, pp. 385–406.

[13] SHAFER, G. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[14] SMETS, P. Constructing the pignistic probability function in a context of uncertainty. In *Uncertainty in Artificial Intelligence, 5*, M. Henrion, R. Shachter, L. Kanal, and J. Lemmer, Eds. Elsevier Science Publishers, 1990, pp. 29–39.

[15] SMETS, P., AND KENNES, R. The transferable belief model. *Artificial Intelligence 66* (1994), 191–234.

[16] TESSEM, B. Approximations for efficient computation in the theory of evidence. *Artificial Intelligence 61:2* (1993), 315–329.

[17] WANG, C.-C., AND DON, H.-S. A geometrical approach to evidential reasoning. In *Proceedings of IEEE* (1991), pp. 1847–1852.

Fabio Cuzzolin is with the Department of Information Engineering, University of Padova, 35131 Padova Italy. E-mail: cuzzolin@dei.unipd.it

The *Decide*IT Decision Tool

M. DANIELSON
Örebro University, Sweden

L. EKENBERG
Stockholm University and Mid Sweden University, Sweden

J. JOHANSSON
Mid Sweden University, Sweden

A. LARSSON
Mid Sweden University, Sweden

Abstract

The nature of much information available to decision makers is vague and imprecise, be it information for human managers in organisations or for process agents in a distributed computer environment. Several models for handling vague and imprecise information in decision situations have been suggested. In particular, various interval methods have prevailed, i.e. methods based on interval estimates of probabilities and, in some cases, interval utility estimates. Even if these approaches in general are well founded, little has been done to take into consideration the evaluation perspective and, in particular, computational aspects and implementation issues. The purpose of this paper is to demonstrate a tool for handling imprecise information in decision situations. The tool is an implementation of our earlier research focussing on finding fast algorithms for solving bilinear systems of equations together with a graphical user interface supporting the interpretation of evaluations of imprecise data.

Keywords

decision analysis, interval probabilities, utility theory, decision tools

1 Introduction

The idea of using computers to support decision making has been around almost as long as computers have been available for humans in usable form. The past decades have witnessed a tremendous development in the graphical user interface, which facilitates the use of more advanced computational techniques to a wider group of users. As a consequence, several decision analytic tools have emerged during the last decade. Decision software based on classical decision theory, such as Standard & Poor's DPL (www.dpl.adainc.com), Palisades' PrecisionTree (www.palisade.com), and TreeAge's DATA (www.treeage.com), have successfully been commercialised and are used by various professional decision analysts and decision makers to aid them in their work.

However, most classical decision models and software based on them consist of some straightforward set of rules applied to precise numerical estimates of probabilities and values. Matrix, tree, and influence diagram models have proliferated, but since they mostly handle precise numeric figures, sensitivity analysis is often not easy to carry out in more than a few dimensions at a time. The requirement to provide numerically precise information in such models has often been considered unrealistic in real-life decision situations, and a number of models with representations allowing imprecise statements have been suggested. Some of them use standard probability theory while others contain some specialised formalism. Most of them focus more on representation and probabilistic inference, and less on evaluation [15], [21], [22], [23], [24].

The purpose of this paper is to present a new decision tool currently being developed, called *Decide*IT. It allows the decision maker to be as deliberately imprecise as he feels is natural and provides him with the means of expressing varying degrees of imprecision in the input sentences, facilitating both the use of decision trees and influence diagrams as decision models. The application takes advantage of a set of algorithms defined as the DELTA method [4], [5], [8], [9], combined with a user-friendly interface which provides an intuitive graphical representation of evaluation results.

Pre-release versions of *Decide*IT have been used in a number of various areas and situations, such as contract formulations [1], investment decisions [7], and insurance policies and flood management [10]. *Decide*IT is currently in a beta-stage of the development phase and will be distributed by Doctor Decide (www.doctordecide.com). Academic licenses will be available for a symbolic fee.

2 The DELTA Method

The main concern of the DELTA method is evaluation of decision problems, with probability and utility intervals to express numerically imprecise information. The method originates from research on handling decision problems involving a finite number of alternatives and consequences [16].

Interval sentences are of the form: "The probability of c_{ij} lies between the numbers a_k and b_k" and are translated into $p_{ij} \in [a_k, b_k]$. Comparative sentences are of the form: "The probability of c_{ij} is greater than the probability of c_{kl}". Such a sentence is translated into an inequality $p_{ij} \geq p_{kl}$. The conjunction of constraints of the types above together with $\sum_j p_{ij} = 1$ for each alternative A_i

involved is called *the probability base* (P). The *value base* (V) consists of similar translations of vague and numerically imprecise value estimates.

A collection of interval constraints concerning the same set of variables is called a *constraint set*. For such a set of constraints to be meaningful, there must exist some vector of variable assignments that simultaneously satisfies each inequality, i.e., the system must be *consistent*. The *orthogonal hull* is a concept that in each dimension signals which parts are incompatible with the constraint set, thus it consists of consistent value assignments for each variable.

Definition 1: Given a consistent constraint set X in $\{x_i\}_{i \in I}, I = \{1,\ldots,n\}$, and a function f, ${}^X\max(f(x)) =_{def} \sup(a|\{f(x) > a\} \cup X$ is consistent).
Similarly, ${}^X\min(f(x)) =_{def} \inf(a|\{f(x) < a\} \cup X$ is consistent).

Definition 2: Given a consistent constraint set X in $\{x_i\}_{i \in I}, I = \{1,\ldots,n\}$, the set of pairs $\{\langle {}^X\min(x_i), {}^X\max(x_i)\rangle\}$ is the *orthogonal hull* of the set and is denoted $\langle {}^X\min(x_i), {}^X\max(x_i)\rangle_n$.

The orthogonal hull greatly simplifies the computational effort and can be pictured as the result of wrapping the smallest orthogonal hyper-cube around the constraint set. For the probability base P, such a wrapping of a consistent system yields *feasible* interval probabilities, in the sense that none of the lower and upper bounds of the probability assignments are inconsistent [24].

2.1 Strength of Alternatives

An *information frame* contains the probability and value bases. In an information frame, an alternative A_i is represented by its consequence set $C_i = \{c_{i1},\ldots,c_{ih_i}\}$.

Definition 3: Given an information frame $\langle\{C_1,\ldots,C_n\},P,V\rangle$ the *strength*, δ_{ij}, denotes the expression $E(C_i) - E(C_j)$, i.e., $\sum_k p_{ik} \cdot v_{ik} - \sum_k p_{jk} \cdot v_{jk}$, over all consequences in the consequence sets C_i and C_j.

To analyse the strength of the alternatives, ${}^{PV}\max(\delta_{ij})$ is calculated. This means that we choose the feasible solutions to the constraints in P and V that are most favourable to $E(C_i)$ and demeaning to $E(C_j)$. This means that if there are no dependencies[1] between the alternatives, ${}^{PV}\max(\delta_{ij}) = {}^{PV}\max(E(C_i)) - {}^{PV}\min(E(C_j))$ and ${}^{PV}\min(\delta_{ij}) = {}^{PV}\min(E(C_i)) - {}^{PV}\max(E(C_j))$. The concept of strength expresses the maximum differences between the alternatives under consideration. It is however used in a comparative way so that formally the maximum and minimum is calculated. In this way, we get a measure about the proportions of the information frame, where the respective alternatives are dominant. When applying the hull

[1] cf. [4] for details when there are various dependencies between the alternatives.

cut operation (see section 2.2), we also receive a measure of the stability of these differences.

This is, however, not enough. Sometimes, the decision maker wants to put more emphasis on the maximal difference (displaying a difference-prone behaviour). At other times, the minimal difference is of more importance. This is captured in the medium difference.

Definition 4: Given an information frame $\langle\{C_1,\ldots,C_n\},P,V\rangle$, let $\alpha \in [0,1]$ be a real number. The α-*medium difference* of δ_{ij} in the frame is ${}^{PV}[\alpha]\mathrm{mid}(\delta_{ij}) = \alpha \cdot {}^{PV}\mathrm{max}(\delta_{ij}) + (1-\alpha) \cdot {}^{PV}\mathrm{min}(\delta_{ij})$.

The α can be considered a precedence parameter that indicates if one boundary should be given more weight than the other. It is, consequently, a measure of difference in strength between the consequence sets. This view duality is a key to understanding the selection process. This is further discussed in [6].

For the pairwise evaluation of our alternatives, [4] suggests the two algorithms ${}^{P}\!BOpt$ and ${}^{V}\!BOpt$. The first algorithm (*probability bilinear optimisation*) can handle any statement except comparisons between value variables from different C_i's, and is described as follows.

Definition 5: Given an information frame $\langle\{C_1,\ldots,C_n\},P,V\rangle$, let C_i be the set $\{c_{i1},\ldots,c_{ih_i}\}$. Then ${}^{V}E_i^{max}$ is $p_{i1}\cdot a_{i1}+\ldots+p_{ih_i}\cdot a_{ih_i}$, where a_{in}, $1 \le n \le h_i$, is $\sup(b|\{b = v_{in}\} \cup \{a_{i(n-1)} = v_{i(n-1)}\} \cup \ldots \cup \{a_{i1} = v_{i1}\}$ is consistent with V).

Further, ${}^{V}E_i^{min}$ is $p_{i1}\cdot a_{i1}+\ldots+p_{ih_i}\cdot a_{ih_i}$, where $a_{in}, 1 \le n \le h_i$, is $\inf(b|\{b \ge v_{in}\} \cup \{a_{i(n-1)} = v_{i(n-1)}\} \cup \ldots \cup \{a_{i1} = v_{i1}\}$ is consistent with V).
Let C_j be the set $\{c_{j1},\ldots,c_{jh_j}\}$. Then ${}^{V}\delta_{ij}$ is ${}^{V}E_i^{max} - {}^{V}E_j^{min}$.

The idea behind this is to transform a bilinear expression into a linear expression with the property of having the same extremal value under specific conditions. Under conditions satisfied by a majority of information frames, $\max\delta_{ij} = \max^{V}\delta_{ij}$ and $\min\delta_{ij} = \min^{V}\delta_{ij}$. When comparisons between value variables from different C_i's are important, the ${}^{V}\!BOpt$ algorithm should be considered instead. ${}^{V}\!BOpt$ is a twin algorithm to ${}^{P}\!BOpt$, working essentially in the same way, but for other preconditions [4].

2.2 Cutting the Orthogonal Hull

A problem with evaluating interval statements is that the results could be overlapping, i.e., an alternative might not be dominating[2] for all instances of the feasible values in the probability and value bases. A suggested solution to this is to further investigate in which regions of the bases the respective alternatives are dominating. For this purpose, the *hull cut* is introduced in the framework. The hull cut

[2] Alternative i dominates alternative j iff ${}^{PV}\mathrm{min}(\delta_{ij}) > 0$.

can be seen as generalised sensitivity analyses to be carried out to determine the stability of the relation between the consequence sets under consideration. The hull cut avoids the complexity in combinatorial analyses, but it is still possible to study the stability of a result by gaining a better understanding of how important the interval boundary points are.

If dominance is evaluated on a sequence of ever-smaller sub-bases, a good appreciation of the strength’s dependency on boundary values can be obtained. This is taken into account by cutting off the dominated regions indirectly using the hull cut operation. This is denoted cutting the bases, and the amount of cutting is indicated as a percentage p, which can range from 0 % to 100 %. For a 100 % cut, the bases are transformed into single points, and the evaluation becomes the calculation of the ordinary expected value.

Definition 6: X is a base with the variables $x_1,\ldots,x_n$, $\pi \in [0,1]$ is a variable referred to as the *cut level*. $\langle a_i,b_i \rangle_n$ is the orthogonal hull, and $\overline{k} = (k_1,\ldots,k_n)$ is a consistent point in X. A π-*cut* of X is to add the interval statements $\{x_i \in [a_i + \pi \cdot (k_i - a_i), b_i - \pi \cdot (b_i - k_i)] : i = 1,\ldots,n\}$ to the base X. $\overline{k}$ is called the *contraction point*.

If no consistent contraction point is given explicitly by the decision maker, *Decide*IT suggests one by minimising the distance to the orthogonal hull midpoints. The choice of the calculated contraction point is motivated by being the centroid in the (non-explicit) second-order belief distributions over the intervals [12]. Intuitively, the hull cuts in *Decide*IT are based on values closer to the centre of the interval being more reliable, i.e., there is an underlying assumption that the second-order distributions have a mass concentrated to the centre. Since the belief in peripheral values is somewhat less, the interpretation of the cut is to zoom in on more believable values that are more centrally located. The centroid of a distribution is exactly this point where this geometrical property of the distribution can be regarded as concentrated. Furthermore, it has very attractive properties from computational as well as intuitive view-points [12].

By co-varying the cutting of an arbitrary set of intervals, it is possible to gain much better insight into the influence of the structure of the information frame on the solutions. Contrary to volume estimates, hull cuts are not measures of the sizes of the solution sets but rather of the strength of statements when the original solution sets are modified in controlled ways. Both the set of intervals under investigation and the scale of individual hull cuts can be controlled.

2.3 Risk Constraints and Security Levels

It is reasonable to extend the framework based on the principle of maximising the expected utility with other decision rules. A number of rules have been suggested, see, e.g., [14], [18] and [20], but these are mostly applicable to decisions under strict uncertainty.

A more general approach is to introduce risk constraints that provide thresholds beyond which a strategy is undesirable. However, when the information is numerically imprecise, the meaning of such thresholds is not obvious. In [11] it is suggested that the interval limits together with stability analyses should be considered in such cases. In *Decide*IT, such thresholds are referred to as *security levels*, and the exclusion of undesirable consequence sets takes the following form,

$$S(C_i, r, s) = (\sum_{v_{ij} \leq r} p_{ij} \leq s)$$

where r denotes the lowest acceptable value and s the highest acceptable probability of ending up with a lower value than r. This means that the sum of the probabilities, where the consequences violate the security level r, must not exceed s. When dealing with interval statements it is not obvious what r and s represents, but one approach is to study the worst and best case by using lower and upper bounds. The contraction points can be used to study the normal case. The concept of security levels is of general use when implementing risk constraints, as suggested in [8].

3 The Tool

The decision tools currently available on the market (e.g., DPL, PrecisionTree, DATA etc.) have set a useful de facto standard for how users may interact with the software, and construct models of their decision problems. Therefore, *Decide*IT has about the same look-and-feel as these tools.

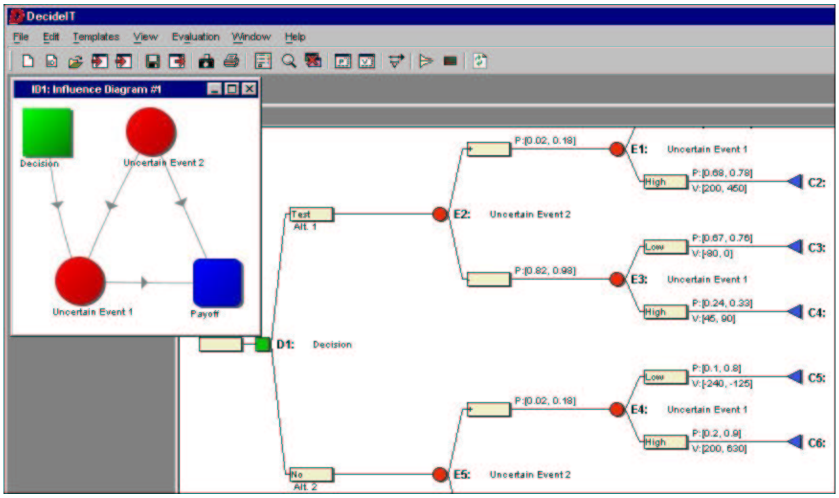
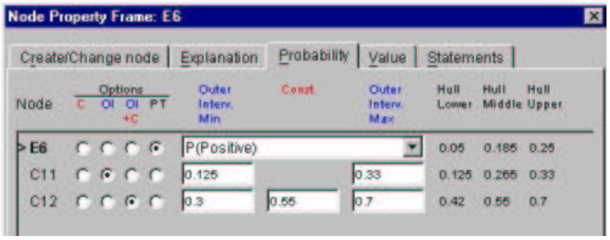


Figure 1: Screenshot of *Decide*IT holding an influence diagram that has been converted to a decision tree.

Currently, three types of nodes may be used in the application: decision nodes, chance nodes, and consequence nodes. Work is carried out on deterministic nodes for influence diagrams.

3.1 Decision Trees

A decision tree is graphically illustrated on the screen, showing explicitly the probabilities and values for all nodes. Interaction with the model is performed through the GUI. Editing probabilities, values, and other properties of a certain node is performed through a node property frame.



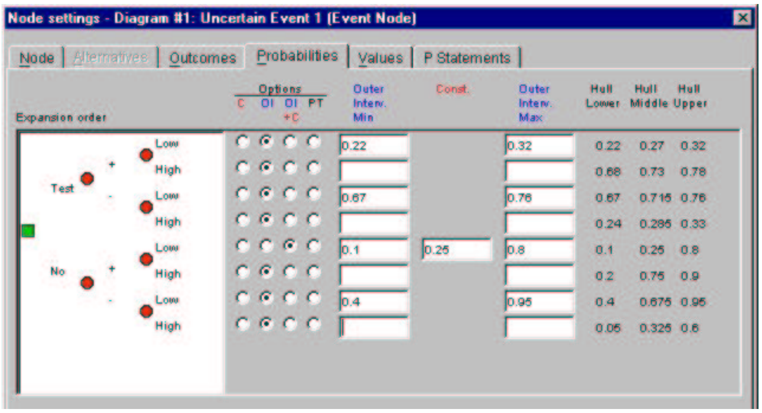
Node	Options C	 OI	 OI +C	 PT	Outer Interv. Min	Const.	Outer Interv. Max	Hull Lower	Hull Middle	Hull Upper
E6	○	○	○	●	P(Positive)			0.05	0.185	0.25
C11	○	●	○	○	0.125		0.33	0.125	0.265	0.33
C12	○	○	●	○	0.3	0.55	0.7	0.42	0.55	0.7

Figure 2: Entering imprecise probabilities, using a probability template for the outcome leading to E_6. For the outcome C_{12}, we explicitly set the contraction point to 0.55.

3.2 Influence Diagrams

Influence diagrams are, when evaluated, transformed into a corresponding symmetric decision tree using a conversion algorithm that creates a total ordering of all connected nodes in the diagram, barren nodes discarded. This conversion algorithm traverses along the directed arcs, and orders the nodes according to a set of rules. In some cases, when only the topology of the graph is not enough to order the nodes, a node placed to the left is converted before a node to the right. It is also possible to convert an influence diagram into an instance of a decision tree, and continue the modelling work on this tree.

Editing the properties of a node in an influence diagram is analogous to the same procedure for a decision tree. There is, however, some differences between the node property frames of the two models. In an influence diagram, the user gets an overview of the conditional expansion order when editing properties of a conditionally dependent chance node.



Expansion order	Options C	Options OI	Options OI +C	Options PT	Outer Interv. Min	Const.	Outer Interv. Max	Hull Lower	Hull Middle	Hull Upper
Low	○	●	○	○	0.22		0.32	0.22	0.27	0.32
High	○	●	○	○				0.68	0.73	0.78
Low	○	●	○	○	0.67		0.76	0.67	0.715	0.76
High	○	●	○	○				0.24	0.285	0.33
Low	○	○	●	○	0.1	0.25	0.8	0.1	0.25	0.8
High	○	●	○	○				0.2	0.75	0.9
Low	○	●	○	○	0.4		0.95	0.4	0.675	0.95
High	○	●	○	○				0.05	0.325	0.6

Figure 3: Entering conditional probabilities for a conditionally dependent chance node in an influence diagram.

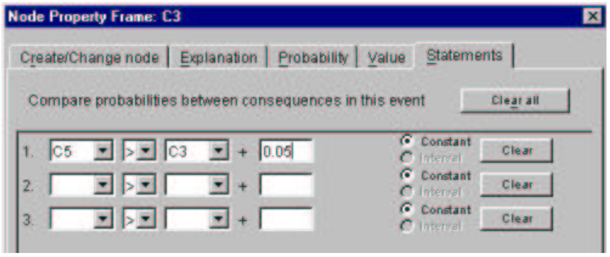
Reversal of arcs is possible between two chance nodes in an influence diagram, who shares a common information state and have no other directed path between them. Thus, according to Shachter, the two chance nodes must inherit each other's conditional predecessors before reversal of an arc between them [19]. Bayes' theorem is invoked, and to determine the lower bound we maximise the denominator and minimise the numerator, and vice versa for the upper bound. This means that as of today reversal of arcs in *Decide*IT simply employ the *intuitive concept* of conditional probability, and a re-flip of the arc will not restore the values for interval probabilities as they do in the precise case. One solution is to implement the Fertig and Breese algorithm [13], but since we do not wish to lose the upper bounds this solution seems less interesting. There does not exist one superior algorithm for this problem taking both lower and upper bounds in account [2],

[3].

Because of this drawback, development of *Decide*IT will focus on employing the *canonical concept* of conditional probabilities [24], but this is a matter of further research regarding the computational aspects. The user of *Decide*IT may however choose not to let the software automatically suggest any new conditional probabilities when flipping an arc.

3.3 Probability and Value Statements

In a chance node in a tree or influence diagram, it is possible to set comparative statements between the probabilities of different outcomes. These statements are then added to the constraint sets. Value statements are set in an analogous fashion.



Node Property Frame: C3

Compare probabilities between consequences in this event

1. C5 > C3 + 0.05 ● Constant ○ Interval

2. > + ● Constant ○ Interval

3. > + ● Constant ○ Interval

Figure 4: Setting a comparative probability statement, that the probability of the outcome leading to C_5 is at least 0.05 higher than the probability of ending up with C_3.

Note that by using this feature, it is possible to handle qualitative probabilities and utilities in a common framework together with the interval approach. Such statements let both decision trees and influence diagrams handle both quantitative and qualitative information, as a step towards evaluation of more qualitative models defined in [17].

3.4 Presentation of Evaluation Results

Results are presented as a graph. Along the x-axis we have the cut in per cent ranging from 0% to 100%, and along the y-axis the possible differences of the expected values between a pair of alternatives. It is also possible to compare one alternative against an average of a set of alternatives. In Figure 5, the upper line is $\max(\delta_{13})$, the middle is $^{PV}[0.5]\text{mid}(\delta_{13})$, and the lower is $\min(\delta_{13})$. The shrinking area depicts the expected value under different degrees of cutting. As can be seen, the higher cut level that is used, the more equal the alternatives seem to be, according to the principle of maximising the expected utility. For a 100% cut, where the results from the algorithms coincide with the ordinary expected value, the result implies that A_3 is the better alternative. However, taking impreciseness in account, it may not be that simple.

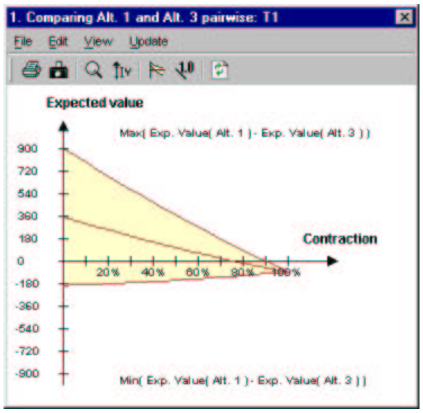
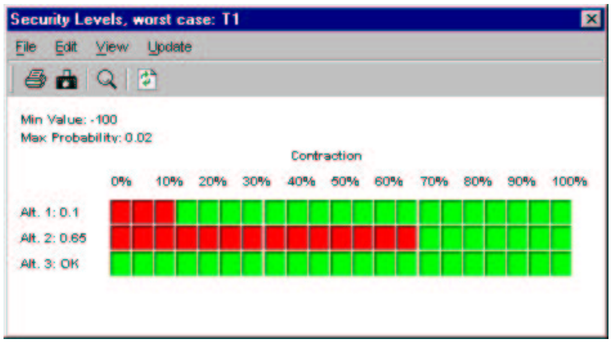


Figure 5: Pairwise comparison of two alternatives, using the DELTA method. After about 75% cut, we see that ${}^{PV}[0.5]\mathrm{mid}(\delta_{13}) < 0$.

3.5 Security Levels

In Figure 6, we investigate at which cut level a given security level will hold in the worst case[3]. An all-green (light grey) alternative can then from this perspective be considered as completely safe.



Min Value: -100
Max Probability: 0.02

Contraction

	0%		10%		20%		30%		40%		50%		60%		70%		80%		90%		100%
Alt. 1: 0.1	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Alt. 2: 0.65	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Alt. 3: OK	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

Figure 6: A security analysis with a security level of -100 as the lowest acceptable value and 0.02 as the highest acceptable probability.

A_3 does not violate the security levels for any cut level and seems to be the desired course of action for a risk avoidant decision maker. This is represented by green (brighter) in the figure above. After a 70% cut level, A_2 does not violate the given security level. If the decision maker is eager for choosing A_1 or A_3, the security analysis imply that A_1 is more risky than A_3, leaving the decision maker to seriously consider choosing A_3 over A_1.

[3] It is possible to investigate best and normal cases as well.

3.6 Preference Ordering Among Consequences

In complex decision situations with large sets of consequences, it might be time-consuming to identify the preference ordering of consequences, and *Decide*IT offers a graphical overview of such a relation on a set of consequences. The ordering is easily determined by checking whether $v_{ij} - v_{kl} > 0$ is consistent with the value base. If not, v_{ij} is before v_{kl} in the partial ordering. Thereafter, obvious transitive relationships are removed.

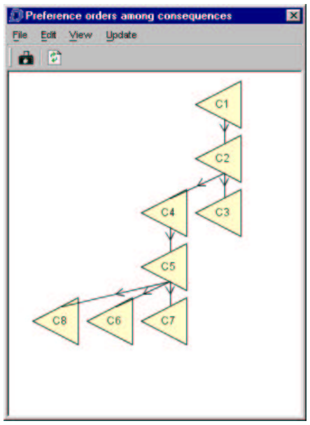
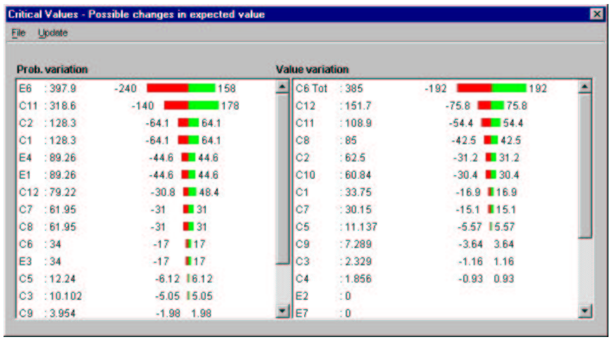


Figure 7: Preference order among consequences, where C_1 is the most preferred consequence.

3.7 Critical Values

Even though the concept of hull cut is a general form of sensitivity analysis, a model may be further investigated through identifying the most critical elements of a decision problem. By varying each event’s probability and utility values within their intervals, it is possible to identify the elements with highest impact on the expected value. This feature lets a decision maker identify where to put his efforts in the information gathering procedure in order to make more safe decisions.



Prob. variation

E6	: 397.9	-240	158
C11	: 318.6	-140	178
C2	: 128.3	-64.1	64.1
C1	: 128.3	-64.1	64.1
E4	: 89.26	-44.6	44.6
E1	: 89.26	-44.6	44.6
C12	: 79.22	-30.8	48.4
C7	: 61.95	-31	31
C8	: 61.95	-31	31
C6	: 34	-17	17
E3	: 34	-17	17
C5	: 12.24	-6.12	6.12
C3	: 10.102	-5.05	5.05
C9	: 3.954	-1.98	1.98

Value variation

C6 Tot	: 385	-192	192
C12	: 151.7	-75.8	75.8
C11	: 108.9	-54.4	54.4
C8	: 85	-42.5	42.5
C2	: 62.5	-31.2	31.2
C10	: 60.84	-30.4	30.4
C1	: 33.75	-16.9	16.9
C7	: 30.15	-15.1	15.1
C5	: 11.137	-5.57	5.57
C9	: 7.289	-3.64	3.64
C3	: 2.329	-1.16	1.16
C4	: 1.856	-0.93	0.93
E2	: 0		
E7	: 0		

Figure 8: Identifying the critical elements of a decision problem, illustrated as a tornado diagram.

For probability variation, the event E_6 has the highest impact on the expected value. By varying the probabilities for this uncertain event, the expected value may differ 397.9 value units. For value variation, the impreciseness in the value of consequence C_6 affects the expected value the most.

4 Concluding Remarks

Based on our earlier research on fast algorithms for solving bilinear problems, we have presented a tool integrating various procedures for handling vague and numerically imprecise probabilities and utilities. The tool has been tested in several real-life applications, and provides means for evaluating decision situations using alternative evaluation principles beside the conventional pointwise maximisation of the expected utility. The latter has turned out to be too limited in many situations. Thus, we also suggest that the alternatives should be further investigated with respect to their relative strengths and also to the number of values consistent with the given domain. Furthermore, the alternatives can also be evaluated relative to a set of security parameters considering how risky they are. To refine the evaluations, we have also shown how hull cut procedures can be introduced in the model. These indicate the effects of choosing different degrees of reliability of the input data. In this way, it is possible to investigate critical variables and the stability of the evaluations. The result of such an analysis often point out reasonable strategies, but also what aspects are crucial to consider for a reliable and stable result.

References

[1] N-P. Andersson, L. Ekenberg, and A Larsson. A Bayesian Approach to Operational Decisions in Transportation Businesses. In *Proc. of 15th International FLAIRS Conference*, 514–518, AAAI Press, 2002.

[2] A. Cano and S. Moral. A Review of Propagation Algorithms for Imprecise Probabilities. In *Proc. of 1st International Symposium on Imprecise Probabilities and their Applications*, 1999.

[3] F.G. Cozman. Computing Posterior Upper Expectations. In *Proc. of 1st International Symposium on Imprecise Probabilities and their Applications*, 1999.

[4] M. Danielson and L. Ekenberg. A Framework for Analysing Decisions Under Risk. In *European Journal of Operational Research*, vol. 104/3, 474–484, 1998.

[5] M. Danielson. Generalized Evaluation in Decision Analysis. To appear in *European Journal of Operational Research*, 2003.

[6] M. Danielson and L. Ekenberg. Symmetry in Decision Evaluation. In *Proc. of 14th International FLAIRS Conference*, 575–579, AAAI Press, 2001.

[7] M. Danielson, L. Ekenberg, J. Johansson and A. Larsson. Investment Decision Analysis – A Case Study at SCA Transforest. To appear in *Proc. of The 2003 International Conference on Information and Knowledge Engineering*, CSREA Press, 2003.

[8] L. Ekenberg. Risk Constraints in Agent Based Decisions. In A. Kent & J.G. Williams (Eds) *Encyclopaedia of Computer Science and Technology*, vol. 23/48, 263–280, Marcel Dekker nc., 2000.

[9] L. Ekenberg, M. Boman, and J. Linneroth-Bayer. General Risk Constraints. In *Journal of Risk Research* 4/1, 31–47, 2001.

[10] L. Ekenberg, L. Brouwers, M. Danielson, K. Hansson, A. Riabacke, J. Johansson, and A. Vari. Simulations and Analysis of Three Flood Management Strategies. International Institute for Applied Systems Analysis (IIASA), 2003.

[11] L. Ekenberg, M. Danielson, and M. Boman. Imposing Risk Constraints on Agent-Based Decision Support. In *Decision Support Systems*, vol. 20/1, 3–15, 1997.

[12] L. Ekenberg and J. Thorbiörnson. Second-Order Decision Analysis. In *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 9/1, 13–38, 2001.

[13] K.W. Fertig and J.S. Breese. Probability Intervals Over Influence Diagrams. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15, 280–286, 1993.

[14] L. Hurwicz. Optimality Criteria for Decision Making under Ignorance. *Cowles Commission Discussion Paper*, vol. 370, 1951.

[15] H.E. Kyburg. Semantics for Interval Probabilities. In *Proc. of 15th International FLAIRS Conference*, 253–257, AAAI Press, 2002.

[16] P-E. Malmnäs. Towards a Mechanization of Real Life Decisions. In *Logic and Philosophy of Science in Uppsala*, 231–243, 1994.

[17] S. Renooij and L. van der Gaag. Decision Making in Qualitative Influence Diagrams. In *Proc. of 11th International FLAIRS Conference*, 410–414, AAAI Press, 1998.

[18] L. Savage. The Theory of Statistical Decision. In *Journal of the American Statistical Association*, vol. 46, 55–67, 1951.

[19] R.D. Shachter. Evaluating Influence Diagrams. In *Operations Research* 34, 871–882, 1986.

[20] A. Wald. *Statistical Decision Functions*, John Wiley and Sons, 1950.

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

[22] P. Walley. Imprecise Probabilities. From *The Imprecise Probabilities Project*, http://ippserv.rug.ac.be, 1997.

[23] K. Weichselberger and S. Pöhlman. *A Methodology for Uncertainty in Knowledge-Based Systems*, Springer-Verlag, 1990.

[24] K. Weichselberger. The Theory of Interval-Probability as a Unifying Concept for Uncertainty. In *Proc. of 1st International Symposium on Imprecise Probabilities and their Applications*, 1999.

Mats Danielson is with the Department of Informatics / ESI at Örebro University, Örebro, Sweden. E-mail: mad@dsv.su.se

Love Ekenberg is with the Department of Computer and Systems Sciences at Stockholm University and KTH, Stockholm, Sweden, and the Department of Information Technology and Media at Mid Sweden University, Sundsvall, Sweden. E-mail: lovek@dsv.su.se

Jim Johansson is with the Department of Information Technology and Media at Mid Sweden University, Sundsvall, Sweden. E-mail: jim.johansson@mh.se

Aron Larsson is with the Department of Information Technology and Media at Mid Sweden University, Sundsvall, Sweden. E-mail: aron.larsson@mhs.studit.com

Convenient Interactive Computing for Coherent Imprecise Prevision Assessments

JAMES DICKEY
University of Minnesota, USA

Abstract

A generalization of deFinetti's Fundamental Theorem of Probability facilitates coherent assessment, by iterated natural extension, of imprecise probabilities or expectations, conditional and unconditional. Point values are generalized to assessed bounds, accepted under weak coherence, that is, allowing the input of redundant loose bounds. The method is realized in a convenient interactive computer program, which is demonstrated here, and made available as open source code. This work suggests that a consulting expert's fees should not be paid unless his/her assessed probabilities cohere.

Keywords

assessment, imprecise probabilities, previsions, coherence, natural extension, interactive computing

1 Introduction

We consider previsions of random quantities, loosely, expectations of random variables, a probability being the prevision of an event, or 0-1 random quantity. Prevision assessments can either be intended as estimates of frequencies, more generally averages, or they can be intended as mere quantitative expressions of human uncertainty. In either case, they should be coherent, that is, extendible to at least one full probability distribution. For estimates of frequencies or averages to be taken seriously, this says that their values must not be impossible when interpreted together as limiting frequencies or limiting averages in an experiment. They can describe a conceivable, possibly infinite, population. For previsions intended as expressions of uncertainty, coherence is a kind of rationality, a direct generalization of non-contradiction for statements of fact, a self-consistency in the sense that, if taken as a person's betting prices, the person could not be made a sure-loser merely by combining a finite number of bets at such prices.

2 Coherent Assessment by Iterated Natural Extension

It is becoming more widely known that deFinetti's Fundamental Theorem of Probability [12, 13] provides a dynamic for interactive computational assessment of coherent previsions. For a sequence of mathematically related random quantities (including logically related events), if coherent prevision values are given for an initial segment of the sequence, the available cohering values for the prevision of the next quantity comprise an interval whose endpoints can be computed by linear programming (first noted by Boole [2], Hailperin [14], and Bruno and Gilio [6]). Walley [22] calls this interval the "natural extension" of the given coherent previsions.

The linear-programming variables are interpretable as the probabilities of the "constituent" events, the events of the joint-range points of the random quantities. Coherence restricts the prevision vector of the quantities to the convex hull of the joint- range set, that is, the prevision point must be some weighted average of the join-range points. The assessed previsions impose additional linear constraints.

In textbook-type problems, where a probability is determined by given probabilities, the extension interval reduces to a single value. If the given values, themselves, are not coherent, the linear programming calculation will so indicate by reporting that there are "no feasible solutions," which implies an empty extension interval. Coherent previsions are always capable of being extended coherently with the value for any further random quantity assignable in an extend-assess cycle. If supplementary calculations are made of the extension interval for a random quantity of special interest, the interval will be seen to shrink to a subinterval whenever a further coherent prevision is assessed.

The method generalizes to include conditional previsions, as inputs and/or outputs. In addition, since prevision is a linear operator, a linear combination of previsions can be assessed directly as the prevision of a linear combination of random quantities. For example, if the assessor defines the difference of two events as a random quantity, then the difference of their probabilities can be assessed as a prevision, and so included in the analysis. The convenience of the method suggests that any consulting expert should not be paid unless her/his probability assessments cohere.

3 Coherence for Imprecise Assessments

An interval, or even a single bound, generalizes a point value, and experts may only be willing to report such imprecise previsions. So how do the coherence concept and the iterated extend-assess algorithm generalize to handle imprecise previsions?

If mere bounds are input, instead of precise values, the output extension interval consists of all the available values for the further prevision for which there exists at least one mutually coherent list of precise values satisfying the input bounds. And, of course, for each such precise list, the corresponding cohering values for the further random quantity would form a subinterval of the output interval. This was defined as the problem of probability logic by Hailperin [15] (following Boole [2]), included as “natural extension” by Walley [22], and presented in a generalization of deFinetti’s Fundamental Theorem by Lad, Dickey, and Rahman [18, 19]. The latter two papers are the basis for the algorithm coded in the present program. A prototype program written in Mathematica in 1991 has had limited distribution.

So, what assessed further bounds should one say “cohere” with the output extension interval?

Definition 1 *(Weak Coherence) Assessed bounds that do not contradict the output bounds will be said to cohere weakly with the given input bounds. An assessed lower (upper) bound must not lie above (below) the output upper (lower) bound, that is, the assessed interval must overlap the extension interval. Also, of course, an assessed lower (upper) bound must not be higher (lower) than the corresponding assessed upper (lower) bound. Weak coherence is directly equivalent to the prevention of sure-loss combined bets.*

Definition 2 *(Strong Coherence) Assessed bounds that neither contradict, in the weak-coherence sense, nor relax the output bounds will be said to cohere strongly with the given input bounds. So, in addition, an assessed lower (upper) bound must not lie below (above) the output lower (upper) bound, that is, the assessed interval must be a subinterval of the extension interval.*

P. Walley [22] uses the term “coherence” to refer to strong coherence, with the interpretation that an assessed lower (upper) value is asserted as the highest (lowest) agreeable relative purchase (selling) price for the random quantity scaled in monetary units, an interpretation under which dynamic refinement of assessed previsions would seem less than natural. Whereas, weakly coherent buying (selling) prices can be interpreted as conservative purchase offers (offers to sell) that can be refined upward (downward). The weak version of coherence was termed “g-coherence” by Biazzo and Gilio [3]. Weak coherence is relevant to our program, for if a user chooses a bound that is a relaxation of the latest extension interval, it has no effect on any subsequent computed interval. Being subject to later refinement, it need not be the tightest bound, now.

In a trivial mathematical sense, the order in which assessments are made does not matter. If an expert asserts the same coherent bounds in a different order, then the same coherent joint bounds will result. (The tightest implied bounds prevail, of course.) In a practical sense, however, ones psychological reaction to encountering different computed intervals for a different order can make a substantial difference in ones assessed values or bounds.

Relevant further references on coherence and coherence methods for imprecise unconditional and conditional previsions, as suggested by referees, include [1], [7], [8], [9], [10], [11], [16].

4 Implementation

This is to introduce an interactive computer program for coherent assessment of imprecise previsions by iterated coherent extension, in which the user communicates with the program through a combined input-output text file. The interaction proceeds as a series of steps, each in the form of an extend-assess cycle:

1. Based on all the prevision bounds assessed so far, the program computes natural extensions, the implied extension interval(s), for the previsions of one or more user-selected quantities.

2. The user assesses a lower and/or upper bound (or a point value) for a prevision, cohering with its computed extension interval.

4.1 Algorithm

To calculate the extension interval for the unspecified prevision of a quantity, say $p_n = P(X_n)$, the program must determine the convex hull of the joint range set of the considered quantities, and then impose the linear constraints of the assessed prevision values and bounds. Denote by $\mathbf{X}\,(n \times 1)$ the vector of n quantities, $R\,(n \times N)$ the matrix of N joint-range points, and $\mathbf{C}\,(N \times 1)$ the vector of N "constituent" events (joint point-value events). Then $\mathbf{C}$ is a partition, and $\mathbf{X} = R\mathbf{C}$. The convex hull of the set of columns of R is the set of all convex combinations,

$$\mathbf{p} = R\mathbf{q}, \tag{1}$$

where $\mathbf{q} \geq \mathbf{0}$ and $\mathbf{1^T q} = \mathbf{1}$. Now, suppose our assessments impose the further constraints,

$$A\mathbf{p} \leq \mathbf{b},$$

some of the inequalities of which may be equalities. The prevision variable to be optimized is $p_n = \mathbf{r_n^T q}$, from Eq. (1). This fully defines the relevant linear-programming calculations.

The steps to achieve this construction and calculation are:

1. Define the product quantities needed for any conditional previsions considered.

2. Define subroutines to reject the potential columns of R that do not satisfy the logical and mathematical constraints on $\mathbf{X}$.

3. Border R for any new random quantities, or start over to reconstruct R from scratch if any old quantities are omitted or redefined.

4. For each prevision to be optimized, form a linear-programming input file and run the routine lp-solve. (Perform a change-of-variables if a conditional prevision is to be optimized.)

4.2 Zero Probabilties

A coherent prevision conditional on an event of zero probability is not determined by the usual unconditional previsions: if $P(A) = 0$, then $P(XA) = 0$ and $P(X|A) = 0/0$, which is indeterminate. Nor can such a conditional prevision have any coherent effect on unconditional prvisions: if $P(A) = 0$, then $P(X) = P(X|A)P(A) + P(X|nA)P(nA) = P(X|nA)$. So, although the program can accept, as input, prevision assessments that are conditional on an event of probability zero, as presently coded, it will not respond to a request to calculate extension bounds on such a prevision. The practical reason for this is that the program solves the fractional-programming problem for a bound on conditional prevision by a change-of-variable that divides by $P(A)$. Improvements in this aspect of the program are contemplated.

4.3 Input/Output

The combined input/output file is organized as a sequence of records, or lines, separated by carriage returns. The following two types of records represent utterances about previsions.

1. Assessed lower and/or upper bound(s) (or point value) on the prevision of a quantity. (Input.)

2. A computed extension interval for the prevision of a quantity. (Output.)

In each type of utterance about a prevision, the case of equal lower and upper bounds, a single point value, is handled by special notation. (A pair of equal bounds are optional on input.)

In order to keep track of what assessed bounds are assumed as the bases for computed intervals, and to promote the stepwise coherence-preserving use of the method, a step number is assigned to a new assessment the first time it is imposed in the calculation of an extension interval. That step number is also assigned to all extension intervals that are subsequently calculated before any further assessments are introduced.

It should be noted that a computed interval will only guarantee coherence of an assessment one new quantity at a time. If more than one quantity's new assessment is uttered in the same step, the linear programming routine could find that they are not coherent, even though each new assessment would be coherent if added singly. The program will issue a warning, yet it will not prevent the user from introducing multiple new assessments in a single step. The user may happen to know that coherence will be preserved, or may just wish to take a chance.

A third type of record provides the framework for prevision utterances:

3. A definition of a random quantity, stated with identifying name, description, range set, and relation(s) (if any) to preceding random quantities. An event is a quantity with the range set $\{0, 1\}$.

The records that define random quantities are spaced out in the file in the order they are introduced, and each is immediately followed by its corresponding prevision utterances, with step numbers. This format seems an important contribution, lending great convenience to the use of the program. The program actually allows the prevision utterances to be placed arbitrarily, but arranging them by quantity seems helpful. What the program requires for quantities is that they be defined and listed in a logical order that facilitates the computation of the joint range set.

4.4 Relations and the Joint Range

Hailperin [14, 15] seems not to have noticed that logical and other mathematical relations among random quantities can substantially reduce the size of their joint range set and, hence, diminish computing costs. It is not necessary, first, to define a full product space and then discard all the points made impossible by the relations. The program brings in only the possible points during the formation of the joint range set. Each definition of a quantity, imposing constraints relating it to previously defined quantities, enables the program to construct only those points that are possible as each quantity is introduced to the joint range. Any reference to a quantity that has not yet been introduced will raise an exception. Of course, the user can wait until the very last quantity defined in the file to impose all the relations, but this can be very inefficient, hence even nonfeasible.

Consider, for example, a partition, $A_1, \ldots, A_n$. The relation $A_1 + \ldots + A_n = 1$, meaning mutually exclusive and exhaustive (for $0-1$ quantities), can be more efficiently imposed piecemeal, as $A_1 + \ldots + A_k \leq 1$ at each definition of A_k, $k = 1, \ldots, n-1$, and then $= 1$ at $k = n$. However, a more convenient approach, also efficient, is to define the A_k's as the value events of an artificial random quantity X with the arbitrary range $\{1, \ldots, n\}$. After first defining X with that range, let $A_k : X = k$, for $k = 1, \ldots, n$. Then $A_1, \ldots, A_n$ will automatically comprise a partition.

4.5 Availability

The program, a moderately large Perl script wrapper on a publicly available open-source linear programming routine, lp-solve, currently runs under unix/linux. User control is through a program command line and the vi editor. The menu for the

MENU			
FILE:		ACTION:	
n	New	a	Assess
o	Open	au	Undo Assessment
s	Save	e	Extend
p	Print	eu	Undo Extension
q	Quit	t	Option

Figure 1: Program menu.

command line is shown in Fig. 1. To obtain the program via e-mail or ftp transfer, contact the author at dickey@stat.umn.edu. A tutorial file is also available.

5 Example: A Medical Screening Test

We demonstrate the use of the program with a simplified example of medical diagnosis. The assessed probability values here will help introduce the program, but they are not necessarily appropriate to the real problem, nor is the problem claimed to be a typical use of the program. Interaction with the program in the example will be described by showing the progressive states of the input/output file.

Suppose a person from the general population receives a positive test result, event S, in a screening skin test for tuberculosis. What is the conditional probability of the event T that she/he has tuberculosis, $P(T\,|\,S)$? This is a classic Bayes' Theorem problem, but the program does not see it as such, treating it more directly as a problem of implied bounds on conditional probability.

Assuming, first, the bounds on the prior probability, $5\times 10^{-5} \leq P(T) \leq 10^{-4}$, and the test-performance probabilities, $P(S|T) = 1$, $1/20 \leq P(S|nT) \leq 1/10$, we will obtain the implied bounds on the marginal symptom probability, $.05005 \leq P(S) \leq .10001$, and the posterior-probability bounds, $.0004998 \leq P(T|S) \leq .001996$. This posterior probability, following a positive symptom, is small; but of course, it lies between about ten and twenty times the prior probability.

We will then use the computed extension interval of the marginal probability, $0.05005 \leq P(S) \leq 0.1001$, as a coherent guide for a further, precise assessment, $P(S) = 0.07$. It could be known, for example, that the relevant empirical frequency of positive test readings is equal to this value. The program then outputs the corresponding step-2 extension intervals. The interval for the conditional false-positive probability will shrink almost to a single point, $.06991 \leq P(S\,|\,nT) \leq .06995$, and the posterior probability of having the disease will be confined to the subinterval, $0.0007143 \leq P(T\,|\,S) \leq 0.001429$. Again, this is small; but it's about 15 times the prior probability.

```
* TITLE/DESCR:        -> SCREENING TEST FOR TB <-
*  Separate fields by “; ” (semicolon space(s)). Records (lines) by <Enter>.
* EVENT/QUANT DEFN FIELDS: xname; descr; rangeSet(or”fun”); relation(or
expr)
* PROB/EXPEC FIELDS: (Indent)P(xname); bdL; bdU; “a”(assess) or
“e”(extend)stepN

T;        Patient has TB;                (0, 1);          none
   P(T);     5.00e-05;     1.00e-04;     a

nT;       Doesn’t have TB;               fun;             not $T

S;        Pos skin test;                 (0, 1);          none
   P(S|T);      eq;        1;         a
   P(S|nT);     1/20;      1/10;      a
   P(S);        ;          ;          e
```

Figure 2: Initial input file.

6 Using the Program

The initial input file is given in Fig. 2. Four automatic header lines, each starting with a star “*”, consist of: title/description of the problem (as entered by user), a line giving formats, and two lines defining the fields of the quantity-definition lines and the fields of the prevision-utterance lines. (We drop these header lines in the subsequent figures.) The user-input lines of the three types follow.

The three left-justified lines here define the events, T, nT (for not$T$), and S. The ranges of T and S are given as the Perl list “$(0,1)$”, followed by “none” (for no relation). The event nT is defined as the function (“fun”), *not* T, of the 0-1 event quantity T. (The dollar sign in the expression “*not* $\$T$” signifies a variable in Perl.) Alternatively, nT could be defined as an event subject to a relation, with the fields, “$(0,1)$; $(\$T\ or\ \$nT)\ ==\ 1$”.

The remaining indented lines are prevision utterances. The second and third fields are for lower and upper bounds, respectively, or for a point value when “*eq*” is entered in the second field followed in the third field by a single number. In the fourth field of a prevision utterance, the user indicates whether an assessment is being asserted (“*a*”) or an extension requested (“*e*”). Perpetual calculation, at each step, of the current extension interval for a quantity is the default action triggered by “*e*”. (To prevent the automatic later extensions, enter “*e*!”.) For a check on the effectiveness of assessed bounds, “*a*” also, by default, triggers the perpetual calculation of extension intervals. (Use “*a*!” to prevent it.) After an interval is calculated, the current step number is automatically appended to the fourth field.

Fig. 3 shows the file updated to report the calculation of three extension inter-

```
T;        Patient has TB;                 (0, 1);             none
    P(T);    5.00e-05;    1.00e-04;    a1
    ;          [5e-05;       0.0001];     e1

nT;       Doesn’t have TB;               fun;                not $T

S;         Pos skin test;                  (0, 1);             none
S_T;    S and T;                            fun;                $S and $T
S_nT;   S and nT;                           fun;                $S and $nT
    P(S|T);       eq;              1;              a1
    P(S|nT);     1/20;           1/10;          a1
    ;                [0.05;          0.1];          e1
    P(S);          [0.05005;    0.1001];     e1
```

Figure 3: First output

```
    P(T|S);    [0.0004998;    0.001996];    e1
```

Figure 4: Second output (fragment)

vals: for $P(T)$; for $P(S\,|\,nT)$, both as checks; and for $P(S)$, the marginal probability of a positive skin test result. Note the new events, $(S\,and\,T)$ and $(S\,and\,nT)$, automatically defined by the program as needed to work with bounds on the conditional probabilities, $P(S\,|\,T) = P(S\,and\,T)\,/\,P(T)$ and $P(S\,|\,nT) = P(S\,and\,nT)\,/\,P(nT)$.

After including a new request for the conditional (“posterior”) probability $P(T\,|\,S)$, we obtain the output as given in Fig. 4, differing only in this one line from the output in Fig. 3. Note that, because no additional assessments were input, the assigned step number remains at 1.

Finally, we use the computed extension interval of the marginal probability $0.05005 \leq P(S) \leq 0.1001$, as a coherent guide for a further, precise assessment, $P(S) = 0.07$. Then the program outputs the corresponding step-2 extension intervals as given in Fig. 5.

7 Relevance of the Method

It seems to the author that these interactive methods could potentially be employed to advantage by real-time decision makers, such as physicians or military commanders. In personal discussion, Glen Meeden has suggested simultaneous cooperative use by a group of experts as an aid to achieving a jointly agreeable coherent assessment.

```
T;        Patient has TB;                (0, 1);          none
    P(T);       5.00e-05;       1.00e-04;     a1
    ;           [5e-05;         0.0001];      e1
    ;           [5e-05;         0.0001];      e2
    P(T|S);     [0.0004998;     0.001996];    e1
    ;           [0.0007143;     0.001429];    e2

nT;       Doesn't have TB;               fun;             not $T

S;        Pos skin test;                 (0, 1);          none
S_T;    S and T;                         fun;             $S and $T
S_nT;   S and nT;                        fun;             $S and $nT
    P(S|T);       eq;           1;              a1
    P(S|nT);      1/20;         1/10;           a1
    ;             [0.05;        0.1];           e1
    ;             [0.06991;     0.06995];       e2
    P(S);         [0.05005;     0.1001];        e1
    ;             eq;           0.07;           a2
    ;             EQ;           [0.07];         e2
```

Figure 5: Third output

Because of the convenience of this coherent assessment algorithm and its interactive implementation, and the flexibility afforded by imprecise assessments, the method would seem destined for heavy use. However, the need for and advantage of such a method hinges on the recognition of logical and other mathematical relations among the quantities whose previsions are subject to assessment. It seems still an open question whether such relations are rare or common in practice. Early Wittgenstein, in what has been called his Logical Independence Thesis, might be interpreted as claiming that such relations tend not to be basic in an analyses. Quoting from the Tractatus [23]:

> The world divides into facts [Prop. 1.2]
>
> Each can be the case or not the case, while the others remain the same [Prop. 1.21]
>
> (See also [Props. 2.061, 2.062].)

In the opinion of a referee, this is no longer an open question, "We simply have applications where logical independence holds and other cases where the random quantities are not logically independent."

8 Further Developments

1. Events or quantities having special properties are amenable to special coding:

   (a) Exchangeable events. Logical independence is usually assumed. Direct definition of variables representing the common invariant joint probabilities seems preferable to imposing the equality constraints for exchangeability on probability variables for a large number of events: $n+1$ variables with 1 constraint, versus 2^n variables with $2^n - n$ constraints.

   (b) Interval events on a random quantity. These can usefully accommodate envelope and other statements regarding the c.d.f.

2. Various upper and lower probability systems (C.A.B. Smith, Dempster-Shafer, etc.) can be incorporated as special program modes. Comparisons can be made in such applications as the use of multiple messages with specifiable reliabilities.

3. Reconciliation of incoherent previsions, by minimum distance under weighted least squares, or other, metric. See, for example, Nau [20, 21].

4. A graphical user interface (Perl/Tk) is being put onto the current functionality, for unix/linux and win32.

5. Charles Geyer has suggestied integration of the program into the emacs editor environment with separate simultaneous displays for the menu and input/output file.

6. Charles Geyer suggested that the program be recast as an ad hoc computing language, for possible inclusion in rweb, or other general system.

9 A Plea

The author would like to hear from conference participants and others interested in using or improving the program. Advice is welcome on what to do or how to do it better, and collaborative and coding help is especially welcome.

Acknowledgement

The author is grateful to anonymous referees for valuable suggestions that improved the paper.

References

[1] M.Baioletti, A.Capotorti, S.Tulipani, B.Vantaggi. Simplification rules for the coherent probability assessment problem, Annals of Mathematics and Artificial Intelligence, 35 (2002), 11-28.

[2] G. Boole. An Investigation of the Laws of Thought. Walton and Maberly, London, 1854.

[3] 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.

[4] V. Biazzo, A. Gilio. "On the linear structure of betting criterion and the checking of coherence", Annals of Mathematics and Artificial Intelligence 35: 83-106, 2002.

[5] V. Biazzo, A. Gilio, G. Sanfilippo. "Coherence checking and propagation of lower probability bounds", Soft Computing 7 (2003), 310-320.

[6] G. Bruno and A. Gilio. Applicazione del metodo del simplesso al teorema fondamentale per le probabilita nella concezione soggettivistica. Statistica, Vol. 40, 1980, No. 3, pp 337-344.

[7] A.Capotorti, B.Vantaggi. A simplified algorithm for inference by lower conditional probabilities, Proc. ?2nd ISIPTA? (2001) Ithaca (New York), 68-76;

[8] A.Capotorti, B.Vantaggi. Locally strong coherence in inferential processes, Annals of Mathematics and Artificial Intelligence, 35 (2002), 125-149.

[9] G.Coletti. Coherent numerical and ordinal probabilistic assessments, IEEE Trans. on Systems, Man, and Cybernetics, 24(12) (1994), 1747-1754.

[10] G.Coletti, R.Scozzafava. The Role of Coherence in Eliciting and Handling Imprecise Probabilities and its Application to Medical Diagnosis. Information Science, 130 (2000), 41-65.

[11] G.Coletti, R.Scozzafava. Probabilistic logic in a coherent setting. Trends in logic n.15, Kluwer, Dordrecht/Boston/London 2002.

[12] B. deFinetti. La prévision, ses lois logiques, ses sources subjectives. Annales de L'Institute Henri Poincaré, Vol. 7, 1937, pp 1-68. H. Kyburg (tr.) Foresight, its logical laws, its subjective sources, in Kyburg, H., and Smokler, H. (eds.) Studies in Subjective Probability, Wiley, New York, 1964; 2nd ed., Krieger, NewYork, 1980.

[13] B. deFinetti. Theory of Probability, Vol. 1. English translation of Teoria delle Probabilità (1970). Wiley, London, 1974.

[14] T. Hailperin. Best possible inequalities for the probability of a logical function of events. American Mathematical Monthly, Vol. 72, 1965, pp. 343-359.

[15] T. Hailperin. Sentential Probability Logic. Lehigh University Press, Bethlehem, PA, 1996.

[16] B. Jaumard; P. Hansen; M. Poggi de Aragao. Column generation methods for probabilistic logic, ORSA J. Comput. 3 (1991) 135-148.

[17] F. Lad. Operational Subjective Statistical Methods. Wiley, New York, 1996.

[18] F. Lad; J. Dickey; M. Rahman. The fundamental theorem of prevision. Statistica, Vol. 50, 1990, No. 1, pp 15-38.

[19] F. Lad; J. Dickey; M. Rahman. Numerical application of the fundamental theorem of prevision. J. of Statistical Computation and Simulation, Vol 40, 1992, pp 135-151.

[20] R. Nau. Decision Analysis with Indeterminate or Incoherent Probabilities. Annals of Operations Research, 1989.

[21] R. Nau. Indeterminate Probabilities on Finite Sets. Annals of Statistics, 1992.

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

[23] L. Wittgenstein. Tractatus Logico-Philosophicus. English translation by D. F. Pears and B. F. McGuiness from the German, in Annalen der Naturpholosophie (1921). Routledge & Kegan Paul Ltd, London, 1961.

James Dickey is with the School of Statistics of the University of Minnesota.

Independence with Respect to Upper and Lower Conditional Probabilities Assigned by Hausdorff Outer and Inner Measures

SERENA DORIA
Università "G.d'Annunzio," Italy

Abstract

Upper and lower conditional probabilities assigned by Hausdorff outer and inner measures are given; they are *natural extensions* to the class of all subsets of Ω=[0,1] of finitely additive conditional probabilities, in the sense of Dubins, assigned by a class of Hausdorff measures. A *weak disintegration property* is introduced when conditional probability is defined by a class of Hausdorff dimensional measures. Moreover the definition of *s-independence* and *s-irrelevance* are given to assure that logical indepedence is a necessary condition of independence. The interpretation of *commensurable* events in the sense of de Finetti as sets with finite and positive Hausdorff measure and with the same Hausdorff dimension is proposed.

Keywords

upper and lower conditional probabilities, Hausdorff measures, disintegration property, independence

1 Introduction

The necessity to introduce a new tool to assess conditional probabilities is due to some problems related to the aximatic definition of regular conditional probability (or regular conditional distribution) Q(A,ω) on a σ-field **F** given a sub σ-field **G**. A regular conditional probability can not exist [7]; moreover even if it exists, if **F** is a σ-field countably generated and **G** is sub σ-field of **F** not countably generated, than there exists no regular, *proper* conditional probability Q(A,ω) on **F** given **G**, that is Q(H,ω) =1 for ω ∈H∈**G** ([2], [3]). In a recent paper of Seidenfeld, Schervish and Kadane [16] *improper regular conditional distributions* are studied. The authors established that when regular conditional probability exists and the sub σ-field **G** is countably generated almost surely it is proper, but when the sub σ-field **G** is not countable generated the regular conditional probability can be *maximally improper*, that is Q(H,ω) =0 for ω ∈H∈**G**, almost surely. Alternative probabilistic approaches that always assure the existence of a proper conditional probability are those proposed by de Finetti [5, 6], Dubins [9] and Walley [17]. In [8] finitely additive conditional probabilities in the sense of Dubins are given by a class of Hausdorff dimensional measures. Their *natural extensions* are given in Section 2 of this paper by outer and inner Hausdorff measures. In particular the case where the σ-field of the conditioning events is not countable generated is analysed. In fact we consider **G** equal to the σ-field of countable or co-countable sets, to the tail σ-field and equal to the σ-field of symmetric events. A problem related to the theory of finitely additive conditional probability is that it does not always satisfy the disintegration property. In section 3 we analyse the meaning of the disintegration property when conditional probability is assigned by a class of Hausdorff dimensional measures. In particular a *weak disintegration property* is introduced and it is proved that this property is verified by conditional probability assigned by a class of Hausdorff measures. There is an other reason to investigate coherent conditional probabilities: it is that, some paradoxical situations about stochastic independence, can be solved if a stronger definition of independence, tested with respect upper and lower conditional probabilities assigned by outer and inner Hausdorff measure, is given. To this aim in section 4 we introduce the definitions of *s-independence* and *s-irrelevance* that are based on the fact that epistemic independence and irrelevance, introduce by Walley, must be tested for events A and B such that the intersection A∩B and the events A and B have the same Hausdorff dimension. With this further condition we prove that s-independence implies logical independence. The results proposed in this paper are based on the idea that *commensurable* events in the sense of de Finetti [4], are subsets of Ω with the same Hausdorff dimension when conditional probability is assigned by a class of Hausdorff measures. At the end of this paper we put in evidence the possibility to use conditional probabilities, assigned by Hausdorff dimensional measures, to deal uncertainty in complex natural phenomena and to give hazard assessments. In fact in different fields of science (geology, biology, architecture) many data sets are fractal sets, i.e. are sets with non-integer Hausdorff dimension. So conditional probabilities, assigned by a Hausdorff dimensional measures, can be used as tool to make inference given fractal sets of data.

2 Upper and Lower Conditional Probabilities Assigned by Hausdorff Outer and Inner Measures

In Walley [17] (Chap. 6) coherent conditional probabilities are considered as a special case of coherent conditional previsions, that are characterized in the case where conditioning events form a partition **B** of Ω.The real number $\overline{P}(X|B)$ are specified for B in **B** and all gambles X in some domain H(B). Conditional previsions $\overline{P}(X|B)$, defined for B in **B** and all gambles X in H(B), are s*eparately* *coherent* when for every conditioning event B, $\overline{P}(\cdot|B)$ is a coherent upper prevision on the domain H(B) and $\overline{P}(B|B) = 1$.

When the domain H(B) is a class of events, that can be regarded as a class of 0-1 valued gambles, $\overline{P}(X|B)$ is a coherent upper conditional probability. In particular when $P(\cdot|B)$ is a countably additive probability defined on a σ-field, its *natural extensions* to the class of all subsets of Ω, called coherent upper and lower probabilities are the outer and inner measures generated by it (see Theorem 3.1.5 of [17]).

In the standard theory, conditional previsions P(| **G**) are defined with respect to a σ-field of events **G**, rather then a partition **B.** The two approches are closely related when **G** is the σ-field made up of all unions of sets in **B.**

In this section coherent upper and lower conditional probabilities are given by the inner and outer measures generated by the Hausdorff dimensional measures. They are *natural extensions* to the class of all subsets of Ω=[0,1] of finitely additive conditional probabilities, in the sense of Dubins [9] assigned by a class of Hausdorff measures.

Let **F** and **G** be two fields of subsets of Ω, with **G**⊆**F**, P* is a *finitely additive conditional probability* [9] on (**F**,**G**) if it is a real function defined on **F**×$\mathbf{G}^0$, where $\mathbf{G}^0$= **G**-{∅} such that the following conditions hold:

I) given any H∈G^0 and A_1,...,A_n ∈**F** with $A_i \cap A_j = \emptyset$ for i≠j, the function P*(·|H) defined on **F** is such that

$$P*(A|H) \geq 0, P*(\bigcup_{k=1}^{n} A_k|H) = \sum_{k=1}^{n} P^*(A_k|H), P*(|H) = 1$$

II) P*(H|H)=1 if H∈**F**∩$\mathbf{G}^0$

III) given E∈**F**, H∈**F**, EH∈**F** with A∈$\mathbf{G}^0$ and EA∈$\mathbf{G}^0$ then P*(EH|A)=P*(E|A)P*(H|EA).

From conditions I) and II) we have

II') P*(A|H)=1 if A∈**F**, H∈$\mathbf{G}^0$and H⊂A.

These conditional probabilities are coherent in the sense of de Finetti, since conditions I), II), III) are sufficient [14] for the coherence of P* on **C**=**F**×$\mathbf{G}^0$ when **F** and **G** are fields of subsets of Ω with **G**⊆ **F** or when **G** is an additive subclass of **F**; otherwise if **F** and **G** are two arbitrary families of subsets of Ω, such that Ω ∈**F** the previous conditions are necessary for the coherence [11, 14], but not sufficient.

Now we recall some definitions about Hausdorff dimensional outer measures that we use as tool to give upper conditional probabilities (for more details about Hausdorff measures see for example [10]).

Let (Ω,d) be the Euclidean metric space with Ω=[0,1]. The diameter of a nonempty set U of Ω is defined as |U|=sup{|x-y|: x,y∈U} and if a subset A of Ω is such that A⊂ $\bigcup_i U_i$ and 0< $|U_i| < \delta$ for each *i*, the class {U_i} is called a δ-cover of

A. Let s be a non-negative number. For $\delta > 0$ we define $h^s(A)=\inf\sum_{i=1}^{\infty}|U_i|^s$, where the infimum is over all (countable) δ-covers $\{U_i\}$. The Hausdorff s-dimensional outer measure of A, denoted by $h^s(A)$, is defined as $h^s(A)=\lim_{\delta\to 0} h^s_\delta(A)$. This limit exists, but may be infinite, since $h^s_\delta(A)$ increases as δ decreases.

The Hausdorff dimension of a set A, $\dim_H(A)$, is defined as the unique value, such that $h^s(A)=\infty$ if $0\leq s < \dim_H A$ and $h^s(A)=0$ if $\dim_H A < s < \infty$. We can observe that if $0< h_s(A) < \infty$ then $\dim_H(A)=s$, but the converse is not true. We assume the Hausdorff dimension of the empty set equal to -1. So no event has Hausdorff dimension equal to the empty set.

Remark: It is important to note the link between the Hausdorff dimension of an event and the Hausdorff dimension of its complement. In fact, denoted by $\dim_H(A)$ the Hausdorff dimension of A we have [10] that

$$dim_H(A\cup B) = max\{dim_H(A), dim_H(B)\};$$

in particular if $A=B^c$we obtain that $1=\dim_H(\Omega)=\max\{\dim_H(B), \dim_H(B^c)\}$; so if $\dim_H(B)$ ¡ $\dim_H(B^c)$ then $\dim(B^c)=1$.

Upper conditional probabilities are given by outer Hausdorff dimensional measures, firstly in the case where conditioning events have finite and positive Hausdorff outer measure.

Theorem 1 *Let Ω=[0,1] and let **F** be the σ-field of all subsets of [0,1] and let **G** be an additive subclass of **F** of sets such that for every H in **G** we have $0<h^s(H)<\infty$, where s is the Hausdorff dimension of H and h_s is the Hausdorff s-dimensional outer measure. Then for each H in **G** the real function P(·|H) defined on **F**, such that*

$$\overline{P}(A|H) = \frac{h^s(AH)}{h^s(H)}$$

verifies the following properties:

a) $0\leq \overline{P}(A|H)\leq 1$;

b) $\overline{P}(A\cup B|H)\leq \overline{P}(A|H)+\overline{P}(B|H)$ and $\overline{P}(A\cup B|H) = \overline{P}(A|H)+\overline{P}(B|H)$ whenever A and B are positively separated, that is $d(A,B)=\inf\{d(x,y): x\in A, y\in B\}>0$;

c) for each $H\in G$ $\overline{P}(\cdot|H)$ is a coherent upper probability.

Proof. For each H belonginig to **G** we have, for the monotony of the Hausdorff outer measures, that

$$0 \leq \overline{P}(A|H) = \frac{\mathrm{h}^s(\mathrm{AH})}{\mathrm{h}^s(\mathrm{H})} \leq \frac{\mathrm{h}^s(\mathrm{H})}{\mathrm{h}^s(\mathrm{H})} = 1;$$

Moreover, since h^s is an outer measure for every s then it is subadditive. For every s the Hausdorff outer measure h^s is a *metric outer measure* that is $h^s(A\cup B)=h^s(A)+h^s(B)$ whenever A and B are positively separated.

Property c) follows from Theorem 3.1.5. of [17]. □

In the general case, when conditioning events can have infinite or zero Hausdorff measure, conditional probability is defined by a 0-1 valued finitely additive (but not countable additive) probability measure m; this assures condition III) of a finitely conditional probability in the sense of Dubins, is verified.

Theorem 2 *Let Ω=[0,1], let* ***F*** *be the σ-field of all subsets of [0,1] and let* **G** *be an additive sub-class of* ***F***. *Let us denoted by h^s the Hausdorff s-dimensional outer measure, by s the Hausdorff dimension of H and by t Hausdorff dimension of AH; let m be a 0-1 valued finitely additive (but not countable additive) probability measure. Then the function $\overline{P}$ defined on* **C**=***F***×$\mathbf{G}^0$ *such that*

$$\overline{P}(A|H) = \begin{cases} \frac{h^s(AH)}{h^s(H)} & if \quad 0 < h^s(H) < \infty \\ m(AH) & if \quad h^s(H) = 0, \infty \end{cases}$$

is an upper conditional probability.

Proof. Firstly we prove that the restriction of $\overline{P}$ to the Catersian product of **B**×$\mathbf{G}^0$, where **B** is the Borel σ-field of [0,1] is a coherent conditional probability. The restriction of the Hausdorff s-dimensional outer measure to the σ-field of the borelian sets of [0,1] is a measure for every *s* so, by definition, we have, that $\overline{P}$ (·|H) verifies condition I) and II).

To prove condition III), that is $\overline{P}$(EH|A)=$\overline{P}$(E|A)$\overline{P}$(H|EA), for E∈**B**, H∈**B** EH∈**B** with A∈$\mathbf{G}^0$ and EA∈$\mathbf{G}^0$, we distinguish the following cases:

a) conditioning events A and EA have positive and finite Hausdorff measures, then condition III) can be written as

$$\frac{\mathrm{h}^s(\mathrm{EAH})}{\mathrm{h}^s(\mathrm{A})} = \frac{\mathrm{h}^s(\mathrm{EA})}{\mathrm{h}^s(\mathrm{A})} \cdot \frac{\mathrm{h}^t(\mathrm{EAH})}{\mathrm{h}^t(\mathrm{EA})} \tag{1}$$

Two cases are possible: i) $s = t$ or ii) $s > t$.

If i) holds than (1) is obviously satisfied. If ii) holds than h^s(EA)=0 and also, by the monotony of h^s, h^s(EAH)=0; so equation (1) is satisfied.

b) conditioning events A and EA have both infinite or zero Hausdorff measures then condition III) becomes m(EAH)=m(EAH)m(EA) and it is always satisfied because m is monotone;

c) conditioning event A has infinite Hausdorff measure and conditioning event EA has positive and finite Hausdorff measure then from the definition of m it follows that condition III) becomes 0=0 , and it is obviously satisfied.

Then from Theorem 3.1.5 of [16] we have that if 0<h^s(H)< ∞ then $\overline{P}$ is the natural extension to **C**=**F**×$\mathbf{G}^0$ moreover if h^s(H)=0 or ∞ then m can be extended to **C**=**F**×$\mathbf{G}^0$ since m is finitely additive, but not countable additive. □

The upper conditional probability defined in the previous Theorem 2 can be used to assess conditional upper probabilities when the class of conditioning events is not a countably generated σ-field. In particular if **G** is equal to the σ-field of countable or co-countable sets, to the tail σ-field or to the σ-field of symmetric events. In all these cases conditioning events have Lebesgue measure equal to one or zero. So upper conditional probability can be defined as in Theorem 2.

Example 1 *Let (Ω,**F**,P) be a probability space where Ω=[0,1], **F** is the σ-field of Borel of Ω and P is the Lebesgue measure on **F**. Let **G** be the sub σ-field of **F** of sets that are either countable or co-countable. Since the probability of the events of the σ-field **G** is either 0 or 1, we have that the probability of A given **G** is equal to P(A), with probability 1, if conditional probability is defined by the Radon-Nikodym derivative. That is*

$$P[A||\mathbf{G}]_{\omega} = P(A)$$

except on a P zero subset of [0,1].

Given A=[a,b] with $0 < a < b < 1$ let P^{*} be the real function defined on $\mathbf{C}$=$\mathbf{F}\times\mathbf{G}^{0}$ such that the restriction P^{*}_{r} to $\mathbf{E}$=$\{(\mathrm{A},\{\omega\})$:$\omega \in [0,1]\}$ is equal, with probability 1, to the Radon-Nikodym derivative P[A||**G**] $_{\omega}$. We have that P* is not coherent on **C**, since it does not satisfy the property that P*(A,{ω}) is equal to 1 or 0 according to whether ω belongs to A or not.

A finitely additive conditional probability on $\mathbf{C}$=$\mathbf{F}\times\mathbf{G}^{0}$ can be defined by

$$\overline{P}(A|H) = \begin{cases} \frac{h^{1}(AH)}{h^{1}(H)} & H \quad \text{co-countable} \\ \frac{h^{0}(AH)}{h^{0}(H)} & H \quad \text{finite} \\ m(AH) & H \quad \text{countable} \end{cases}$$

where m is a 0-1 valued finitely additive (but not countably additive) probability measure.

The function $\overline{P}$ is a coherent conditional probability since it verifies the axioms of a finitely additive probability in the sense of Dubins as proved in Theorem 2.

The lower conditional probability $\underline{\mathrm{P}}$(A|H) can be define as in the previous theorems if h^{s} denotes the Hausdorff s-dimensional inner measure.

3 The Disintegration Property

In this section we analyse the meaning of the disintegration property when conditional probability is assigned by a class of Hausdorff dimensional measures. In particular a *weak disintegration property* is introduced. If conditional probability is defined by the Radon-Nikodym derivative $\mathrm{P[A||G]}_{\omega}$, it verifies the disintegration property, that is the functional equation $\mathrm{P(A\cap H)}=\int_{H}\mathrm{P[A||G]}_{\omega}\mathrm{dP}$ with H∈**G**.

This property is not always satisfied in the theory of finitely additive probability of Dubins. In fact with a finitely additive probability P it is not assured that P(A)=

$\int_{\Omega}$P[A||G]$_{\omega}$dP for A in **F**. In the paper of Schervish, Seidenfeld and Kadane [15] has been shown that each finitely but not countably additive probability P will fail to be disintegrable on some denumerable partition of Ω.

Let Ω=[0,1], let **F** be the σ-field of the Borel subsets of [0,1], **G** a sub σ-field of **F** and let P be equal to h^1, that is the Lebesgue measure. We recall that since the class of subsets of Ω measurable with respect to h^s, for every s, is the class of Borel subsets of [0,1], than each h^s is a measure (σ-additive) on **F**. We denote by P* the restriction to **F**×**G**^{0}, of the upper conditional probability assigned in Theorem 2. For each H in **G**^{0}P$^*(A|H)$ is a function on H.

The starting point is that when the conditioning event H has Hausdorff dimension s less then 1, the equation P(A∩H)=$\int_H$P*(A|H)dP is obviously verified since dim(A∩H)≤dim(H)<1 then P(A∩H)=0=P(H) and $\int_H$P*(A|H)dP=0. So it can be interesting to investigate if an analogous equation holds with respect to the measure h^s. We observe that, if h^s(H)=∞ then the functions P*(A|H), defined in the previous section, are not integrable since no constant different from zero is integrable on H with respect to h^s; so we introduce the following definition

Definition 1. Let Ω=[0,1], let **F** be the σ-field of the Borel subsets of [0,1] and let P be equal to h^1, that is the Lebesgue measure. Let **G** be a sub-σ−field of **F**. Denoted by h^s the Hausdorff s-dimensional measure where s is the Hausdorff dimension of H. A coherent conditional probability P* verifies the *weak disintegration property* if the following functional equation h^s(A∩H)=$\int_H$P*(A|H)dh^s is verified for every H in **G**^{0} with h^s(H)< ∞.

Remark: If dim(A∩H)<dim(H)=s and h^s(H)< ∞ then the equation

$$h^s(A\cap H) = \int_H P*(A|H)dh^s$$

is satisfied since both members are equal to zero. So to verify that a given coherent conditional probability satisfied the weak disintegration property we have to prove that the equation is verified for every pair of event A, H with dim(AH)=dim(H).

Theorem 3 *Let Ω=[0,1], let* **F** *be the σ-field of the Borel subsets of [0,1] and let P be equal to h^1, that is the Lebesgue measure. Let* **G** *be a sub-σ-field of* **F**. *Having fixed A in* **F**, *let us denoted by h^s the Hausdorff s-dimensional measure, by s the Hausdorff dimension of H; let m be a 0-1 valued finitely additive (but not countable additive) probability measure. The coherent conditional probability P* defined on* **C**=**F**×**G**^{0} *such that*

$$P^*(A|H) = \begin{cases} \frac{h^s(AH)}{h^s(H)} & if \quad 0 < h^s(H) < \infty \\ m(AH) & if \quad h^s(H) = 0, \infty \end{cases}$$

verifies the weak disintegration property.

Proof. We have to prove that the equation

$$h^s(A\cap H) = \int_H P^*(A|H)dh^s \tag{2}$$

is verified for every H in $\mathbf{G}^0$ with h^s(H)$<\infty$.

Firstly we suppose h^s(H) positive and finite; for each A and H, the function P*(A|H) is nonnegative and less or equal to 1, so it is integrable with respect to h^s; then we observe that the equation (2) is always satisfied since

$$\int_H P^*(A|H)dh^s = \int_H \frac{h^s(A\cap H)}{h^s(H)}dh^s = h^s(A\cap H).$$

Moreover if h^s(H) is equal to zero, then equation (2) vanishes to 0=0. □

4 Independence

In this section we introduce a new definition of independence for events, called *s-independence*, based on the fact that the relative events and their intersection must have the same Hausdorff dimension. This notion does not require any assumption of positivity for the probability of the conditioning event. This is one of the difference with the concepts of *confirmational irrelevance* and *strong confirmational irrelevance*, proposed by Levi [12].

We prove that s-independence between events implies their logical independence when both events have Hausdorff dimension less than 1. Moreover also when the events have Hausdorff dimension equal to 1 and positive and finite Lebesgue outer measure then logical dependence is a necessary condition for the s-independence. Firstly we analyse the concept of *epistemic independence* for events proposed by Walley [17] with respect to conditional upper and lower probabilities defined by Hausdorff dimensional outer and inner mesures. The concept of epistemic independence is based on the notion of *irrelevence*; given two events A and B, we say that B is *irrelevant* to A when $\underline{\text{P}}$(A|B)=$\underline{\text{P}}$(A|B^c)=$\underline{\text{P}}$(A) and $\overline{P}$(A|B)=$\overline{P}$(A|B^c)=$\overline{P}$(A).

A and B are *epistemic independent* when B is irrelevant to A and A is irrelevant to B. As a consequence of this definition we can obtain the factorisation property P(A∩B)=P(A)P(B) that constitutes the standard definition of independence for events. Let Ω=[0,1] and let $\overline{P}$ and $\underline{\text{P}}$ be the upper and lower conditional probabilities defined by the outer and inner Hausdorff measures. The unconditional upper and lower probabilities can be obtained from the conditional ones by the equalities $\overline{P}$(A)=$\overline{P}$(A|Ω)= and $\underline{\text{P}}$(A)=$\underline{\text{P}}$(A|Ω).

When the events A and B or their complements have not upper probability equal to zero, epistemic independence implies *logical independence*, (i.e. each of four sets A∩B, A∩B^c, A^c∩B, A^c∩B^c are non-empty). Otherwise logically dependent events can be epistemically independent.

Example 2 *Let Ω=[0,1], let **F** be the σ-field of all subsets of [0,1] and let **G** be the additive sub-class of **F** of sets that are finite and co-finite. Let A and B two finite subsets of [0,1] such that A∩B=∅. If conditional probability is defined as in Theorem 2 we have that*

$$\underline{P}(A|B) = \overline{P}(A|B) = \frac{h^0(AB)}{h^0(B)} = 0$$

$$\underline{P}(A|B^c) = \overline{P}(A|B^c) = \frac{h^1(AB^c)}{h^1(B^c)} = 0$$

$$\underline{P}(A) = \overline{P}(A) = \overline{P}(A|\Omega) = \frac{h^{\,1}(A)}{h^1(\Omega)} = 0$$

So A and B are logical dependent but epistemically independent.

The previous example puts in evidence the necessity to introduce the following definition.

Definition 2. Let Ω=[0,1], let **F** be the σ-field of all subsets of [0,1] and let **G**=**F**. Denoted by $\overline{P}$ and $\underline{P}$ be the upper and lower conditional probabilities defined by the outer and inner Hausdorff measures and given A and B in $\mathbf{G}^0$, then they are *s-independent* if the following conditions hold:

1) dim$_H$(AB)=dim$_H$(B)=dim$_H$(A)

2) $\underline{P}$(A|B)=$\underline{P}$(A|B^c)=$\underline{P}$(A) and $\overline{P}$(A|B)=$\overline{P}$(A|B^c)=$\overline{P}$(A).

3) $\underline{P}$(B|A)=$\underline{P}$(A|A^c)=$\underline{P}$(B) and $\overline{P}$(B|A)=$\overline{P}$(B|A^c)=$\overline{P}$(B).

Remark: Two disjoint events A and B are s-dependent since the Hausdorff dimension of the empty set can not be equal to that one of any other set so condition 1) is never satisfied. In particular the events A and B of Example 1, that are logical dependent but epistemically independent, are not s-independent.

We prove that logical independence between two events A and B is a necessary condition for s-independence when dim$_H$(A) and dim$_H$(B) are both less then 1.

Theorem 4 *Let Ω=[0,1], let **F** be the σ-field of all subsets of [0,1], let **G**=**F** and let us denoted by $\overline{P}$ and $\underline{P}$ be the upper and lower conditional probabilities defined by the outer and inner Hausdorff as in Theorem 2. Then two events A and B of $\mathbf{G}^0$, s-independent and with Hausdorff dimension less then 1, are logical independent.*

Proof. Since dim$_H$(A) and dim$_H$(B) are both less then 1 if A and B are s-independent then the following conditions hold:

1) dim$_H$(AB)=dim$_H$(B)=dim$_H$(A)

2) $\underline{P}$(A|B)=$\underline{P}$(A|B^c)=$\underline{P}$(A)=$\underline{h}^1$(A)=0 and $\underline{P}$(B|A)=$\underline{P}$(A|A^c)=$\underline{P}$(B)=$\underline{h}^1$(B)=0

3) $\overline{P}$(A|B)=$\overline{P}$(A|B^c)=$\overline{P}$(A)=$\overline{h}^1$(A)=0 and $\overline{P}$(B|A)=$\overline{P}$(B|A^c)=$\overline{P}$(B)=$\overline{h}^1$(B)= 0.

From 1) we have that A∩B≠ ∅ since the Hausdorff dimension of the empty set can not be equal to that one of any other set, from 3) we have $\overline{P}$(A|B)=0 then B is not contained in A and $\overline{P}$(B|A)=0 then A is not contained in B. Moreover since dim$_H$A and dim$_H$B are both less then 1 then h^1(A∪B)=0 while h^1(Ω)=1 so Ω ≠A∪B. □

We prove that logical independence is a necessary condition for the s-independence when the events have Hausdorff dimension equal to 1 and positive and finite Lebesgue outer measure.

Theorem 5 *Let Ω=[0,1], let* ***F*** *be the σ-field of all subsets of [0,1], let* ***G***=***F*** *and let us denoted by* $\overline{P}$ *and* $\underline{P}$ *be the upper and lower conditional probabilities defined by the outer and inner Hausdorff as in Theorem 2. Two events A and B of* **G**^{0}*, s-independent, with Hausdorff dimension equal to 1 and such that 0<* $\overline{h}^1$*(A)<1 and 0<* $\overline{h}^1$*(B)<1, are logically independent.*

Proof. Since A and B are s-independent, from condition 1) we have dim$_H$A∩B=1, that implies A∩B≠ ∅; from condition 3) we have $\overline{P}$(A|B)=$\overline{P}$(A|B^c)=$\overline{P}$(A)=$\overline{h}^1$(A)≠1 so B is not contained in A and B^c is not contained in A; moreover $\overline{P}$(B|A)=$\overline{P}$(B|A^c)=$\overline{P}$(B)= $\overline{h}^1$(B) ≠1 so A is not contained in B and A^c is not contained in B.Then A and B are logically independent. □

We can observe that the converse of Theorems 3) and 5) is not true; in fact logical independence is not a sufficient condition for the s-independence.

Example 3 *Let Ω=[0,1], let* ***F*** *be the σ-field of all subsets of [0,1], let* ***G***=***F*** *and let us denoted by* $\overline{P}$ *and* $\underline{P}$ *be the upper and lower conditional probabilities defined by the outer and inner Hausdorff measures as in Theorem 2. Let A and B two finite subsets of [0,1] such that each of four sets A∩B, A∩B^c, A^c∩B, A^c∩B^c is non-empty, that is A and B are logical independent. We have that A and B are not s-independent since conditions 2) and 3) of Definition 1 is never satisfied.*

If **G** is properly contained in **F** and A belong to **F**-**G**, for any H in **G**^{0} we cannot test the s-independence between A and H because epistemic independence is symmetric, so it requires that also A belongs to **G**^{0}; in this case we introduce the following definition.

Definition 2. Let Ω=[0,1], let **F** be the σ-field of all subsets of [0,1] and let **G** a sub field of **F**. Denoted by $\overline{P}$ and $\underline{P}$ be the upper and lower conditional probabilities defined by the outer and inner Hausdorff measures and given A in F and B in **G**^{0}, then B is *s-irrelevant* to A if the following conditions hold:

1) dim$_H$(AB)=dim$_H$(B)=dim$_H$(A)

2) $\underline{P}$(A|B)=$\underline{P}$(A|B^c)=$\underline{P}$(A) and $\overline{P}$(A|B)=$\overline{P}$(A|B^c)=$\overline{P}$(A).

Proposition 1 *Let Ω=[0,1], let **F** be the σ-field of all subsets of [0,1] and **G** a sub field properly contained in **F**. Given A in **F** and B in $\mathbf{G}^0$ such that dim_H(A)<1, dim_H(B)<1 and B is s-irrelevant to A then the following conditions hold:*

1a) A∩B≠ ∅;

2a) B is not contained in A and B^c is not contained in A;

3a) Ω ≠A∪B;

Proof. The result follows from Theorem 5. □

Definition 3. Let Ω=[0,1], let **F** be the σ-field of all subsets of [0,1] and let **G** an additive subclass contained in **F**. Given A in **F** we say that **G** is *s-irrelevant* to A if any event H of **G** such that $\dim_H$(A)=$\dim_H$(H) is irrelevant to A.

The previous results can be used to solve paradoxical situations proposed in literature that show that the interpretation of conditional probability in terms of partial knowledge breaks down in certain cases. A conditional probability can be used to represent partial information as proposed by Billingsley [1]. A probability space (Ω,**F**,P) can be use to represent a random phenomenon or an experiment whose outcome is drown from Ω according to the probability given by P. Partial information about the experiment can be represented by a sub σ-field **G** of **F** in the following way: an observer does not know which ω has been drawn but he knows for each H in **G**, if ω belongs to H or if ω belongs to H^c.

A sub σ-field **G** of **F** can be identified as partial information about the random experiment, and, fixed A in **F**, conditional probability can be used to represent partial knowledge about A given the information on **G**. By standard definition, an event A is independent from the σ-field **G** if it is independent from each H in **G**, that is, if conditional probability is defined by the Radon-Nikodym derivative, P[A||**G**]$_\omega$=P(A) with probability 1. Example 3 shows that the interpretation of conditional probability in terms of partial knowledge breaks down in certain cases. In fact the event A is independent from the information represented by **G** and this is a contradiction according to the fact that the information represented by **G** is complete since **G** contains all the singletons of Ω. The contradiction can be dissolved if s-irrelevance is tested with respect to conditional probabilities assigned by a class of Hausdorff dimensional measures.

Example 4 *Let (Ω,**F**,P) be a probability space where Ω=[0,1], **F** is the σ-field of Borel of Ω and P is the Lebesgue measure on **F**. Let **G** be the sub σ-field of **F** of sets that are either countable or co-countable. Let $\overline{P}$ be the finitely additive conditional probability defined on $\mathbf{C}=\mathbf{F}\times\mathbf{G}^0$ by*

$$\overline{P}(A|H) = \begin{cases} \frac{h^1(AH)}{h^1(H)} & H \quad \textit{co-countable} \\ \frac{h^0(AH)}{h^0(H)} & H \quad \textit{finite} \\ m(AH) & H \quad \textit{countable} \end{cases} \tag{3}$$

where m is a 0-1 valued finitely additive (but not countably additive) probability measure.

Given A=[a,b] with 0<a<b<1, we have that **G** is not s-irrelevant to A, since condition 2) of the definition of s-irrelevance is not satisfied.

In fact for every H= [0,1]-{ω} we have that $\overline{P}$(A)=$\overline{P}$(A|Ω)=h^1(A) is different from 0 and 1, while P*(A|H^c)=P*(A|{ω}) must be, for the coherence, equal to 1 or 0 according to the fact that ω belongs to A or not.

5 Conclusions and Applications

The results proposed in this paper would be an attempt to show that Hausdorff dimensional measures can be used as a tool to define coherent conditional probabilities. This approach is based on the idea that *commensurable* events [4] with respect to the given coherent conditional probability, are subsets of Ω with the same Hausdorff dimension. Given a coherent conditional probabilities P* defined on **C** =**F**×$\mathbf{G}^0$, any pair of events A and B of $\mathbf{G}^0$ can be compare as proposed by de Finetti. In fact

$$P^*(A|A\cup B)+P^*(B|A\cup A)\geq 1$$

so the above conditional probabilities cannot be both zero and their ratio can be used to introduce an ordering between A and B. In fact this ratio is finite if either P^*(A|A∪B) and P^*(B|A∪B) are finite and in this case A and B are called *commensurable*. Otherwise if one of the conditional probability is zero the corresponding event has a probability infinitely less then the other and the two events A and B belong to different layers [5]. We can observe that when conditional probability P^* is countably additive there can be only finitely many layers above a given layer, but not so when P is only finitely additive.

Two events A and B of $\mathbf{G}^0$, commensurable with respect to the coherent conditional probability defined by (3) of Example 4, are subsets of Ω with the same Hausdorff dimension. The converse is not true, in fact if A is countable and B finite then the two events have Hausdorff dimension equal to 0, but they are not commensurable with respect to the previous conditional probability, since coherence requires that P^*(B|A∪B)=0. Two events are commensurable in the sense of de Finetti if and only if they have both finite and positive Hausdorff measure and the same Hausdorff dimension.

Also from a practical point of view there are some advantages to assess coherent conditional probabilities, by a class of Hausdorff dimensional measures. In fact they can be used as a tool to assess probability to an event given a data set coming from a real problem. In different fields of science (geology, biology, architecture, economics) many data sets are fractal sets, i.e. are sets with non-integer Hausdorff dimension; for example the hypocentre distribution of earthquakes is a fractal set so if we want to assess the probability that a given place will be the hypocentre of a future earthquake knowing the set of the previous ones, we need to have a tool able to handling fractal sets. Moreover the classification of several soils can be done by their Hausdorff dimensions. A future aim of this research is to implement these results to dealing uncertainty in natural hazard and risk assessment.

6 Acknowledgements.

I wish to thank Prof. Teddy Seidenfeld of Carnegie Mellon University for his valuable remarks and insightful comments.

References

[1] P.Billingsley. (1985), *Probability and measure*, John Wiley, New York.

[2] D. Blackwell and L.Dubins, (1975), On existence and non-existence of proper, regular, conditional distributions, *The Annals of Probability*, Vol 3, No5, 741-752.

[3] D. Blackwell and C. Ryll-Nardzewski, (1963), Non-existence of everywhere proper conditional distributions, *Ann.Math.Statist.*, 34, 223-225.

[4] B.de Finetti, (1936), Les probabilits nulles, in *Bulletin de Sciences Matematiques*, Paris, pp. 275-288.

[5] B.de Finetti, (1970), *Teoria della Probabilita'*, Einaudi Editore, Torino.

[6] B.de Finetti. (1972), *Probability, Induction, Statistics*, Wiley, London.

[7] J.L. Doob *Stochastic Processes*, John Wiley, 1953.

[8] S.Doria (2001), Conditional Upper Probabilities Assigned by a Class of Hausdorff Outer Measures, in ISIPTA 01, *Proceedings of the Second International Symposium on Imprecise Probabilites and Their Applications*, Edited by G.de Cooman, T.L.Fine, T.Seidenfeld, Ithaca, NY,USA, pp.147-151.

[9] L. Dubins. (1975), Finitely additive conditional probabilities, conglomerability and disintegrations, *The Annals of Probability*, VOL. 3, No.1,89-99.

[10] K.J.Falconer. (1986), *The geometry of the fractal sets*, Cambridge University Press.

[11] S. Holzer. (1985), On coherence and conditional prevision, *Bollettino U.M.I.* Serie VI, vol IV-C-N.1, 441-460.

[12] Levi, I (1980), *The Enterprise of knowledge.* MIT Press.

[13] E. Regazzini. (1987), de Finetti coherence and statistical inference, *The Annals of Statistics*, vol 15, No.2, 845-864.

[14] E. Regazzini. (1985), Finitely additive conditional probabilities, *Rend. Sem. Mat. Fis.* Milano, 55, 69-89.

[15] M. Schervish, T. Seidenfeld and J.B. Kadane. (1984), The extent of non-conglomerability of finitely additive probability. *Z.War.* 66 205-226.

[16] T. Seidenfeld, M. Schervish and J.B. Kadane. (2001), Improper regular conditional distributions, *The Annals of Probability*, Vol. 29, No 4,1612-1624.

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

Serena Doria is with the Facoltà di Scienze Matematiche Fisiche e Naturali, Università "G.d'Annunzio," Via dei Vestini, 31 66013 Chieti, Italy. E-mail: s.doria@dst.unich.it

Towards a Chaotic Probability Model for Frequentist Probability: The Univariate Case[*]

P.I. FIERENS
Cornell University, U.S.A.

T.L. FINE
Cornell University, U.S.A.

Abstract

We adopt the same mathematical model of a set **M** of probability measures as is central to the theory of coherent imprecise probability. However, we endow this model with an **objective, frequentist interpretation** in place of a behavioral subjective one. We seek to use **M** to model **stable physical sources of time series data that have highly irregular behavior** and not to model states of belief or knowledge that are assuredly imprecise. The approach we present in this paper is to understand a set of measures model **M** not as a traditional compound hypothesis, in which one of the measures in **M** is a true description, but rather as one in which none of the individual measures in **M** provides an adequate description of the potential behavior of the physical source as actualized in the form of a long time series.

We provide an **instrumental interpretation** of random process measures consistent with **M** and the highly irregular physical phenomena we intend to model by **M**. This construction provides us with the basic tools for simulation of our models.

We present a method to estimate **M** from data which studies any given data sequence by analyzing it into subsequences selected by a set of computable rules. We prove results that help us to choose an adequate set of rules and evaluate the performance of the estimator.

Keywords

imprecise probability, sets of measures, objective, frequentist interpretation

[*] Research of both authors partially supported by AFOSR Grant F30602-00-2-0558

1 Introduction

1.1 Orientation

We adopt the same mathematical model of a set $\mathbf{M} = \{\nu\}$ of probability measures as is central to the theory of coherent imprecise probability (e.g., see Walley [18]). However, we endow this model with an **objective, frequentist interpretation** in place of a behavioral subjective one, and ask completely different questions of this model. While the mathematical model **M** is the same in the two theories of probability (as it is in a variety of interpretations that have been offered for conventional probability), on our account there is no focus on imprecision as is appropriate in the behavioral account. In order to signal the distinction between the two theories sharing the same mathematical model, we do not use the descriptor “imprecise” and instead use **“chaotic”**. Although we remain interested in alternatives to this term, it does connote a highly irregular sequence of physical (typically mechanical) origin. We seek to use **M** to model **stable** (although not stationary in the traditional stochastic sense) **physical sources of time series data that have highly irregular behavior** and not to model states of belief or knowledge that are assuredly imprecise. Support for the existence of such chaotic sources is lent by the following quotation from Kolmogorov [9]:

> *In everyday language we call random those phenomena where we cannot find a regularity allowing us to predict precisely their results. Generally speaking, there is no ground to believe that random phenomena should possess any definite probability. Therefore, we should distinguish between randomness proper (as absence of any regularity) and stochastic randomness (which is the subject of probability theory). There emerges the problem of finding reasons for the applicability of the mathematical theory of probability to the real world.*

1.2 Previous Work

Previous work focused on asymptotics or laws of large numbers for interval-valued probability models can be found in Fine et al. [10][12][7][15]. Cozman and Chrisman [1] estimate credal sets by looking at the limiting relative frequencies along several subsequences of a time series. Our current work focussed on modelling finite length time series.

Our previous attempt at supplying an objective frequentist interpretation for a set of measures **M**, reported at ISIPTA ’01 in Fierens and Fine [3], was based upon the use of Kolmogorov complexity to enable us to simulate highly complex time series data from the model and then to estimate the model from such data through the sequence of alternating minima and maxima of relative frequencies calculated along a given sequence. The underlying motivation was an attempt at an analog of the $i.i.d.$ standard probability model; the model **M** gave us the marginal or univariate description and the high complexity was meant to ensure that there was no further exploitable structure in the time evolution. We subsequently judged this approach to be inadequate, in part after considering the performance of martingale betting systems on such time series as advocated by the then newly-published Shafer and Vovk [16].

1.3 Overview

As in our previous work, we focus on a description of univariate or marginal events and not on descriptions of k-tuples of outcomes. This restriction is intended only to simplify our search for a meaningful interpretation and not because we deny the importance of an extension to k-tuples. In Section 2.1, we provide an **instrumental interpretation** of random process measures consistent with $\mathbf{M}$ and the highly irregular physical phenomena we intend to model by $\mathbf{M}$. Although we do not offer this description as an explanation for real world data, we develop it because it helps us to better understand chaotic probability models by reference to well-known standard stochastic processes, and, at the same time, this description provides us with the basic tools for simulation of our models (see Section 2.2). Essentially, our instrumental interpretation consists of a decision mechanism that at each time instant chooses a probability measure $\nu \in \mathbf{M}$ from which the next outcome of a sequence will be generated. This measure selection function has both properties of being highly complex so that it is difficult to discover it from any given data sequence, and having enough simple structure to allow for the estimation of $\mathbf{M}$ (see Theorems 1-3). The approach we present in this paper is to understand a set of measures model $\mathbf{M}$ not as a traditional compound hypothesis, in which one of the measures in $\mathbf{M}$ is a true description, but rather as one in which none of the individual measures in $\mathbf{M}$ provides an adequate description of the potential behavior of the physical source as actualized in the form of a long time series. Instead, it is the whole set $\mathbf{M}$ that describes the potential behavior, and this distinction has operational significance in terms of the time series data that is anticipated from the physical source.

As explained in Section 3, **we estimate $\mathbf{M}$ from a data sequence by computing the relative frequencies along some of its subsequences**. Subsequence selection is a well-entrenched method of exposing behavioral patterns in time series. It formed the basis of Richard von Mises' pioneering definition of randomness ([17],[4],[11],[13]) for infinitely long sequences and the seminal work of A.N. Kolmogorov on randomness of finite strings ([8]). Cozman and Chrisman [1] estimate credal sets by looking at the relative frequencies along several subsequences. In a similar way, we also study a given sequence by analyzing it into subsequences selected by rules in some set Ψ. Technically, we use **causal subsequence selection rules**, also known as **Church place selection rules** (see Definition 1 and also Li and Vitányi [11]). For any given model $\mathbf{M}$, we expect to find some set of rules Ψ_V for which $\mathbf{M}$ becomes **"visible"**, that is, a set of rules such that all measures in $\mathbf{M}$ can be estimated by the relative frequencies along the selected subsequences (see Definition 2 and Theorems 2 and 3). Although such a set Ψ_V may exist, identifying it will not be easy. Furthermore, there are sets of rules Ψ_T for which a chaotic source may appear to be **“temporally homogeneous”**, that is, for a certain set Ψ_T there may exist a chaotic source generating sequences such that the relative frequencies along subsequences selected by rules in Ψ_T cannot expose more than a small neighborhood of a single measure contained in the convex hull of **M** (see Definition 3, Lemma 1 and Theorem 4).

Proofs have been omitted in what follows. However, they are available in the appendices of Fierens [2].

2 From the Model to Data

2.1 An Instrumental Interpretation of the Model

Let $\mathbf{X} = \{z_1, z_2, \cdots, z_\xi\}$ be a finite sample space. We denote by $\mathbf{X}^*$ the set of all finite sequences of elements taken in **X**. A particular sequence of n samples from **X** is denoted by $x^n = \{x_1, x_2, \cdots, x_n\}$. **P** denotes the set of all measures on the power set of **X**. A chaotic probability model **M** is a subset of **P** and models the “marginals” of some process generating sequences in $\mathbf{X}^*$. In this section, we present an instrumental (that is, without commitment to reality) interpretation of such a process.

Consider the generation of a sequence x^n by the following pseudo-algorithm:

```
FOR k = 1 TO k = n
  1. Choose $\nu \in \mathbf{M}$.
  2. Generate $x_k$ according to $\nu$.
```

If the decision mechanism in `1` is very complex[1], say, random, with decisions made in an $i.i.d.$ manner according to some distribution on **M**, we would not be able to distinguish whether x^n was produced by an $i.i.d.$ process according to some measure in $ch(\mathbf{M})$, the convex hull of **M**, or by the algorithm in question. On the other hand, if the decision rule were very simple and deterministic, we would possibly be able to make such a distinction. For example, consider the simple choice mechanism that alternates between two measures $\nu_1, \nu_2 \in \mathbf{M}$. In this case, for sufficiently large n, we expect to discover the alternating-measure rule and to be able to estimate ν_1 and ν_2. However, if the choice mechanism in `1` were neither too complex (as in the first example) nor too simple (as in the second example), we may still be able to estimate **M** (or part of it), but we would probably

[1] Although Kolmogorov complexity captures part of the complexity to which we make reference here, it seems not to suffice. Thus, the discussion in this paragraph follows at a more intuitive level.

find it difficult (if not impossible given our computational resources) to discover the choice mechanism itself. It is in this case that we believe chaotic probability models to be useful: when dealing with chaotic sources, the measure selection function F has both properties of being highly complex so that it is difficult to discover it from any given data sequence, and having enough simple structure to allow for the estimation of $\mathbf{M}$.

We formalize the decision in `1` of the previous algorithm by means of a function $F : \mathbf{X}^* \to \mathbf{M}$. Furthermore, we restrict ourselves to *causally* made decisions, ones dependent only upon the past:

```
FOR k = 1 TO k = n
  1. Choose $\nu = F(x^{k-1}) \in \mathbf{M}$.
  2. Generate $x_k$ according to $\nu$.
```

Let $\nu_k = F(x^{k-1})$. For any $k \leq n$, F determines the probability distribution of the *potential* kth outcome X_k of the sequence,

$$(\forall \mathbf{A} \subseteq \mathbf{X}) \;\; P(X_k \in \mathbf{A} | X^{k-1} = x^{k-1}) = \nu_k(X_k \in \mathbf{A}).$$

An actual data sequence x^n is assessed by the graded potential of the realization of a sequence of random variables X^n described by

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{k=1}^{n} \nu_k(X_k = x_k).$$

We denote by $\mathbf{M}^*$ the family of all such process measures P. From the analysis of data, we do not expect in general to be able to pinpoint a single $P \in \mathbf{M}^*$ or even a small subset of $\mathbf{M}^*$, what we call a **fine-grained picture** of the source. On the contrary, we expect our knowable **operational quantities to be (large) subsets of $\mathbf{M}^*$** which provide an appropriate **coarse-grained** description of the source. These ideas are related to those of **coarse grainedness** and **fine grainedness** in physics. For example, in classical physics we commonly have situations, say, kinetic theory, in which a coarse description suffices even though we have access in principle to a more detailed quantum mechanical one. Unlike the case of classical physics, there need be no more than instrumental reality in the fine details of our model $\mathbf{M}^*$. A similar situation may be found in quantum mechanics where there are fine-grained pictures that have no empirical reality (see Gell-Mann [6], Chapter 11, especially pp. 143-147).

2.2 Simulation

Simulation of sequences coming from a source modelled by a set of measures $\mathbf{M}$ can be achieved by simply choosing an appropriate function F and applying the algorithm presented above. Since we expect not to know F in general, the choice of the measure selection functions used for simulation depends on our judgment, based on our knowledge of the physical phenomenon being modelled, the intended use of the simulated sequences, etc.

In the typical case where $\mathbf{M}$ has infinite cardinality, we need a notion of approximation to the measures in $\mathbf{M}$ by finitely many other measures in (or close to) $\mathbf{M}$. Given a distance or metric d on $\mathbf{P}$, a particular form of approximation is provided by an **ε-covering** of $\mathbf{M}$, that is, by a covering of the set $\mathbf{M}$ by open balls of radius ε (according to d) and centers in some set $\mathbf{M}_\varepsilon \subset \mathbf{P}$ (perhaps a subset of $\mathbf{M}$). Note that, if $\mathbf{P}$ is compact with respect to d, we can find a **finite** ε-covering of $\mathbf{M}$. Choose a minimal set $\mathbf{M}_\varepsilon$ so that each ball has a non-empty intersection with $\mathbf{M}$ and call $B(\varepsilon, \nu)$ the ball with center $\nu \in \mathbf{M}_\varepsilon$ and radius ε. Then, given an appropriate measure selection function $F : \mathbf{X}^* \to \mathbf{M}_\varepsilon$, the following algorithm can be used for simulation.

```
FOR k = 1 TO k = n
  1. Choose $\nu = F(x^{k-1}) \in \mathbf{M}_\varepsilon$.
  2. Choose any $\nu' \in B(\varepsilon, \nu) \cap \mathbf{M}$.
  3. Use a pseudo-random number generator to
     generate $x_k$ according to $\nu'$.
```

Since we want to expose all of $\mathbf{M}$ in a single, but sufficiently long, simulated sequence, we require F to visit, many times, each measure in $\mathbf{M}_\varepsilon$. Theorems 1-2 in Section 3 can help us choose the minimum number of times that each measure should be visited. Examples of simulation algorithms based on the basic strategy presented above are available in the appendices of Fierens [2] (see, e.g., the proof of Theorem 4) and in Section 3.5

3 From Data to the Model

3.1 Subsequence Analysis

We begin the study of a sequence $x^n \in \mathbf{X}^*$ by analyzing it into several subsequences. These subsequences are selected by rules that satisfy the following

Definition 1 (Causal Subsequence Selection Rule)

An effectively computable function ψ *is a* **causal subsequence selection rule** *(also known as a Church place selection rule) if*

$$\psi : \mathbf{X}^* \to \{0, 1\},$$

and, for any $x^n \in \mathbf{X}^*$, x_k is the j-th term in the generated subsequence $x^{\psi,n}$, of length $\lambda_{\psi,n}$, if

$$\psi(x^{k-1}) = 1, \quad \sum_{i=1}^{k} \psi(x^{i-1}) = j, \quad \lambda_{\psi,n} = \sum_{k=1}^{n} \psi(x^{k-1}).$$

Let $\Psi = \{\psi_\alpha\}$ be a set of causal subsequence selection rules. For each $\psi \in \Psi$, we study the behavior of the relative frequency of (only) marginal events along the chosen subsequence. That is, given x^n and a selection rule $\psi \in \Psi$ we determine the **frequentist empirical (relative frequency) measure** $\overline{\mu}_{\psi,n}$ along the subsequence $x^{\psi,n}$ through

$$(\forall \mathbf{A} \subset \mathbf{X})\ \overline{\mu}_{\psi,n}(\mathbf{A}) = \frac{1}{\lambda_{\psi,n}} \sum_{k=1}^{n} I_{\mathbf{A}}(x_k)\psi(x^{k-1}),$$

where $I_{\mathbf{A}}(\cdot)$ is the $\{0,1\}$-valued indicator function of the event $\mathbf{A}$. In a similar manner, for any such rule ψ, we may compute the **time average conditional measure** $\overline{\nu}_{\psi,n}$ defined by

$$(\forall \mathbf{A} \subset \mathbf{X})\ \overline{\nu}_{\psi,n}(\mathbf{A}) = \frac{1}{\lambda_{\psi,n}} \sum_{k=1}^{n} \mathrm{E}\left[I_{\mathbf{A}}(X_k) \,\middle|\, X^{k-1} = x^{k-1}\right] \psi(x^{k-1}).$$

Rewritten in terms of our instrumental understanding of the measure selection function F,

$$\overline{\nu}_{\psi,n}(\mathbf{A}) = \frac{1}{\lambda_{\psi,n}} \sum_{k=1}^{n} \nu_k(\mathbf{A})\psi(x^{k-1}),$$

where $\nu_k = F(x^{k-1})$.

Since we want to expose some of the structure of the chaotic probability model $\mathbf{M}$ by means of the rules in Ψ, we are interested in how good an estimator of $\overline{\nu}_{\psi,n}$ is $\overline{\mu}_{\psi,n}$. Introduce the norm-based metric

$$(\forall \mu, \mu' \in \mathbf{P})\ d(\mu, \mu') = \max_{z \in \mathbf{X}} \left|\mu(z) - \mu'(z)\right|,$$

which quantifies the "closeness" between two probability measures on $\mathbf{X}$. We call a rule ψ applied to x^n **causally faithful** if the resulting subsequence yields a small value of $d(\overline{\nu}_{\psi,n}, \overline{\mu}_{\psi,n})$. The existence of such rules is guaranteed by

Theorem 1 *Let ξ be the cardinality of $\mathbf{X}$ and denote the cardinality of Ψ by $||\Psi||$. Let $m \leq n$. If $||\Psi|| \leq t_n$, then for any process measure $P \in \mathbf{M}^*$*

$$P\left(\max_{\psi \in \Psi} \left\{ d(\overline{\mu}_{\psi,n}, \overline{\nu}_{\psi,n}) :\ \lambda_{\psi,n} \geq m \right\} \geq \varepsilon\right) \leq 2\xi t_n \mathrm{e}^{-\frac{\varepsilon^2 m^2}{2n}}.$$

Hence, so long as we restrict to a family of causal selection rules of size t_n and examine discrepancies of size ε only over subsequences of length at least m, with m large, we can with high probability avoid uncontrollably imposing our own patterns through some of the selected subsequences and instead exhibit only the patterns that have inductive validity. If, to the contrary, we allow the set of subsequence selection rules to be too large, we will observe with non-negligible probability measures that are outside the convex hull of **M**. For example, if we enlarge the set of subsequence selection rules by including all possible subsequences, then we will observe measures that concentrate all the mass on a single atom (outcome in **X**).

3.2 Visibility and Estimation

The possibility of exposing all of **M** by means of the rules in Ψ is expressed in the following

Definition 2 (Visibility)
(a) **M** *is* **made visible** (Ψ,θ,δ,m,n) *by* $P \in \mathbf{M}^*$ *if*

$$P\left(\bigcap_{\mu\in\mathbf{M}}\bigcup_{\psi\in\Psi}\{X^n : \lambda_{\psi,n}(X^n) \geq m, d(\bar{\mu}_{\psi,n},\mu) \leq \theta\}\right) \geq 1-\delta.$$

(b) *A subset of* $\mathbf{M}^*$ **renders M uniformly visible** (Ψ,θ,δ,m,n) *if* **M** *is made visible* (Ψ,θ,δ,m,n) *by each of its elements. The maximal such subset is denoted* $\mathbf{M}_V(\Psi)$ *and* $\mathbf{M}_V(\Psi)$ *may be empty.*

The non-triviality of Definition 2**(a)**, and, hence, of Definition 2**(b)**, is asserted in

Theorem 2 *Let* $0 < 2\varepsilon < \theta$ *and* $\mathbf{M}_\varepsilon \subseteq \mathbf{M}$ *be the centers of a minimal covering of* **M** *by* N_ε *balls of radius* ε *(according to the metric d as defined above*[2]*). Then, for large n, there exists a process measure P and a family* Ψ *of size* N_ε *such that* **M** *is made visible* (Ψ,θ,δ,m,n) *with*

$$\delta = 2(\xi+1)N_\varepsilon \mathrm{e}^{-\frac{(\theta-2\varepsilon)^2 m^2}{2n}}.$$

Theorem 2 asserts the existence of a set of rules Ψ such that $\mathbf{M}_V(\Psi)$ is not empty.

The following theorem shows how it is possible to estimate **M** by means of an appropriate set of rules Ψ.

[2] According to our choice of d, although **M** is not necessarily compact, **P** certainly is. Therefore, as a subset of compact **P**, **M** will always have a finite open covering.

Theorem 3 *Let* $\mathbf{M}_\alpha$ *be a subset of the non-empty maximal* $\mathbf{M}_V(\Psi) \subseteq \mathbf{M}^*$ *that renders* $\mathbf{M}$ *uniformly visible* $(\Psi, \theta, \delta, m, n)$. *Let* $[\mathbf{A}]^\varepsilon$ *denote the* ε**-enlargement** *of a set* $\mathbf{A}$ *defined by*

$$(\forall \mathbf{A} \subseteq \mathbf{P})\ (\forall \varepsilon > 0)\ [\mathbf{A}]^\varepsilon = \{\mu : (\exists \mu' \in \mathbf{A}) d(\mu, \mu') < \varepsilon\}.$$

Let $\hat{\mathbf{M}}_{\theta,\Psi}$ *be an estimator of* $\mathbf{M}$ *defined by*

$$(\forall x^n \in \mathbf{X}^*)\ \hat{\mathbf{M}}_{\theta,\Psi}(x^n) = \bigcup_{\{\psi : \psi \in \Psi,\ \lambda_{\psi,n}(x^n) \geq m\}} B(\theta, \bar{\mu}_{\psi,n}).$$

Then the estimator $\hat{\mathbf{M}}_{\theta,\Psi}$ *satisfies*

$$(\forall P \in \mathbf{M}_\alpha) \quad P\left([ch(\mathbf{M})]^{\theta+\varepsilon} \supset \hat{\mathbf{M}}_{\theta,\Psi} \supset \mathbf{M}\right) \geq 1 - \delta - \tau_n,$$

where $ch(\mathbf{M})$ *is the convex hull of* $\mathbf{M}$ *and*

$$\tau_n = 2\xi ||\Psi|| e^{-\frac{\varepsilon^2 m^2}{2n}}.$$

3.3 Temporal Homogeneity

Not every set of rules Ψ can expose all of $\mathbf{M}$. The following definition deals with some sets of rules that can only expose a small neighborhood of a single probability measure in $ch(\mathbf{M})$.

Definition 3 (Temporal Homogeneity)
(a) $P \in \mathbf{M}^*$ *is* **temporally homogeneous** $(\Psi, \theta, \delta, m, n)$ *if*

$$P\left(\max_{\psi_1, \psi_2 \in \Psi} \{d(\bar{\mu}_{\psi_1,n}, \bar{\mu}_{\psi_2,n}) : \lambda_{\psi_1,n}(X^n), \lambda_{\psi_2,n}(X^n) \geq m\} \leq \theta\right) \geq 1 - \delta.$$

(b) *A subset of* $\mathbf{M}^*$ *is* **uniformly temporally homogeneous** $(\Psi, \theta, \delta, m, n)$ *if each of its elements is temporally homogeneous* $(\Psi, \theta, \delta, m, n)$. *The maximal such subset is denoted* $\mathbf{M}_T(\Psi)$.

The non-triviality of Definition 3**(a)**, and, hence, of Definition 3**(b)**, is established by

Lemma 1 *Choose* $\mu_0 \in \mathbf{P}$ *and* $0 < 2\varepsilon < \theta$ *and constrain the measure selection mechanism* F *so that*

$$(\forall x^* \in \mathbf{X}^*)\ F(x^*) \in B(\varepsilon, \mu_0),$$

where $B(\varepsilon, \mu_0)$ is a ball with center μ_0 and radius ε; that is, every P induced by F is approximately i.i.d. μ_0. Then each such P is temporally homogeneous $(\Psi, \theta, \delta, m, n)$ provided that δ satisfies

$$\delta = 2\xi t_n \mathrm{e}^{-\frac{[(\theta-2\varepsilon)m]^2}{8n}},$$

where $t_n = ||\Psi||$.

3.4 Consistency between Visibility and Temporal Homogeneity

We can better appreciate the difficulty of choosing an appropriate set of rules for estimation of $\mathbf{M}$ by means of the next theorem, which in some sense complements Theorem 2 and Lemma 1.

Theorem 4 *Let $\varepsilon > \frac{1}{m}$. Assume that there is an ε-cover of $\mathbf{M}$ by N_ε open balls with centers in a set $\mathbf{M}_\varepsilon = \{\mu_1, \mu_2, \cdots, \mu_{N_\varepsilon}\}$ such that, for each μ_i, there is a* **recursive probability measure** *$\nu \in B(\varepsilon, \mu_i) \cap \mathbf{M}$. Let Ψ_0 be a set of (causal deterministic) place selection rules. Then, there are a process measure P and a family Ψ_1 such that, for large enough n, P will both render $\mathbf{M}$ visible $(\Psi_1, 3\varepsilon, \delta, m, n)$ and ensure temporal homogeneity $(\Psi_0, 6\varepsilon, \delta, m, n)$ with*

$$\delta = 2\xi t_n \mathrm{e}^{-\frac{\varepsilon^2 m^2}{2n}},$$

where

$$t_n = \max\{||\Psi_0||, ||\Psi_1||\}.$$

A more transparent version of Theorem 4, given in terms of an analyzing set Ψ_0 formed by **finite history rules** defined as follows

Definition 4 Finite History Rules

We say that ψ is a **finite history rule** *if there is a positive integer L, called* **the history length of** *ψ, and a function $\Gamma : \mathbf{X}^L \to \{0,1\}$ such that for all $x^n \in \mathbf{X}$ we have*

$$\psi(x^{k-1}) = \begin{cases} \Gamma(x^{k-L}, x^{k-L+1}, \cdots, x^{k-1}) & \text{if } k > L, \\ 0 & \text{otherwise.} \end{cases}$$

The next theorem is similar to Theorem 4:

Theorem 5 *Assume that $\mathbf{M}$ makes all atoms possible, i.e., there is $\phi > 0$ such that*

$$\inf_{\mu \in \mathbf{M}} \min_{z \in \mathbf{X}} \mu(\{z\}) \geq \phi.$$

Let Ψ_0 consist of finite history rules with length smaller than a given L. Then, for $\varepsilon > 0$*, there are a process measure P and a family Ψ_1 such that P will both render* **M** *visible* $(\Psi_1, 2\varepsilon, \delta, m, n)$ *and ensure temporal homogeneity* $(\Psi_0, 4\varepsilon, \delta, m, n)$ *with*

$$\delta = 4\xi t_n e^{-\frac{\varepsilon^2 m^2}{2n}},$$

where

$$t_n = \max\{||\Psi_0||, ||\Psi_1||\}.$$

Although we do not present a complete proof here (it can be found in Fierens [2]), we give the basic idea behind the construction of P in Section 3.5 because it provides a simple example of several ideas in this paper.

Put picturesquely, the results in this section show that Ψ determines the **resolving power of the analytical microscope** with which we examine **M**. When one prepares a sample to be put under the lenses of the microscope, little or nothing is seen of the structure of the sample, e.g., it may just look like some watery solution. Similarly, in the case of a chaotic probability model, the temporal homogeneity property tells us that **M** looks just like the traditional single measure. As we explore **M** with a large numbers of more complex selection rules, say, under the more powerful lenses of the microscope, we begin to see or isolate different relative frequency measures and begin to see **M** as a set of measures. However, we do not know in advance the final scale at which **M** exhibits all of its structure and do not know in advance how to choose Ψ to render all of **M** visible. Our abilities at progressive exploration are, of course, limited both by the increasing computational burden, and by considerations of extracting faithful subsequences. Preserving the faithful subsequence property requires a relation between $||\Psi||$ and the resulting confidence level $1 - \delta$. As $||\Psi||$ increases, maintaining confidence levels requires longer subsequences (larger m) and in turn more data (larger n). These considerations make good traditional statistical sense.

3.5 Simulation Example

Let $\mathbf{M}_\varepsilon = \{\mu_1, \mu_2, \cdots, \mu_{N_\varepsilon}\} \subseteq \mathbf{M}$ be the centers of a finite ε-cover of **M** by N_ε open balls. Let γ be defined as

$$\gamma = \left\lceil \log_\xi N_\varepsilon \right\rceil.$$

Let $\mathbf{B}_1, \cdots, \mathbf{B}_{N_\varepsilon}$ be a partition of $\mathbf{X}^\gamma$, the histories of length γ, into N_ε subsets and consider the memory-γ Markov process defined by the following transition probabilities:

$$(\forall \mathbf{A} \subseteq \mathbf{X})\ P\left(X_k \in \mathbf{A} | X_{k-1} = x_{k-1}, \cdots, X_{k-\gamma} = x_{k-\gamma}\right) = \mu_i\left(X_k \in \mathbf{A}\right), \qquad (1)$$

iff $(x_{k-\gamma}, \cdots, x_{k-1}) \in \mathbf{B}_i$. It can be proved that this Markov process has a unique stationary probability measure μ_S (see Fierens [2]).

Let R be an integer greater than a given L and consider the construction of a process measure $P \in \mathbf{M}^*$ by an algorithm that: a) initializes R *i.i.d.* Markov processes (as described by Eqn. 1) at the stationary measure; b) generates the sequence x^n by choosing outcomes from the R Markov processes in a round-robin fashion. A more detailed description of this algorithm follows.

```
FOR l = 1 TO l = $R$
  1. Generate $(x_{l,1}, x_{l,2}, \cdots, x_{l,\gamma})$ according to $\mu_S$.
  2. FOR k = $\gamma+1$ TO k = $\lceil n/R \rceil$
     (a) Find the set $\mathbf{B}_i$ such that $(x_{k-\gamma}, \cdots, x_{k-1}) \in \mathbf{B}_i$.
     (b) Generate $x_k$ according to $\mu_i$.
Set $R$ counters $i_1, i_2, \cdots, i_R$ to 1.
FOR k = 1 TO k = n
  1. Let $l = [(k-1) \bmod R] + 1$.
  2. Let $x_k = x_{l,i_l}$.
  3. Let $i_l = i_l + 1$.
```

We now sketch the proof that the previous algorithm succeeds in constructing a process measure $P \in \mathbf{M}^*$ satisfying the conditions stated in Theorem 5. By the previous algorithm, for $k > R\gamma$, the outcome X_k depends on $X_{k-R\gamma}$, $X_{k-R(\gamma-1)}$, $\cdots$, X_{k-R}, but it does not depend on X_{k-R+1}, X_{k-R+2}, $\cdots$, X_{k-1}. Let ψ be any rule in Ψ_0. Since ψ has a limited time horizon L which is strictly smaller than R, we have for all $\mathbf{A} \subset \mathbf{X}$

$$\mathrm{E} \sum_{k=1}^{n} \psi(X^{k-1}) \left[\mathrm{E} \left[I_{\mathbf{A}}(X_k) \,\middle|\, X^{k-1} \right] - \mu_S(\mathbf{A}) \right] =$$

$$= \sum_{k=1}^{n} P\left(\psi(X^{k-1}) = 1 \right) \mathrm{E} \left[I_{\mathbf{A}}(X_k) - \mu_S(\mathbf{A}) \,\middle|\, \psi(X^{k-1}) = 1 \right] = \text{ (by memory } L)$$

$$= \sum_{k=1}^{n} P\left(\psi(X^{k-1}) = 1 \right) \mathrm{E} \left[I_{\mathbf{A}}(X_k) - \mu_S(\mathbf{A}) \,\middle|\, \psi(X^{k-L:k-1}) = 1 \right] = \text{ (by indep.)}$$

$$= \sum_{k=1}^{n} P\left(\psi(X^{k-1}) = 1 \right) \mathrm{E} \left[I_{\mathbf{A}}(X_k) - \mu_S(\mathbf{A}) \right] = 0.$$

It can be shown, by means of the same techniques used in the proof of Theorem 1, that this fact implies that

$$P\left(\max_{\psi \in \Psi_0} \left\{ d(\bar{\nu}_{\psi,n}, \mu_S) : \lambda_{\psi,n} \geq m \right\} \geq \varepsilon \right) \leq 2\xi \|\Psi_0\| \mathrm{e}^{-\frac{\varepsilon^2 m^2}{2n}}.$$

Finally, this statement together with Theorem 1 imply that **M** is Ψ_0-temporal homogeneous.

Let $\Psi_1 = \{\psi_1, \cdots, \psi_{N_\varepsilon}\}$ be a set of rules such that

$$\psi_i(x^{k-1}) = \begin{cases} 1 & \text{if } k > R\gamma \text{ and } (x_{k-R\gamma}, x_{k-R(\gamma-1)}, \cdots, x_{k-R}) \in \mathbf{B}_i, \\ 0 & \text{otherwise.} \end{cases}$$

Then, it is easy to see that

$$(\forall \mathbf{A} \subseteq \mathbf{X})\ \bar{\nu}_{\psi_i,n}(\mathbf{A}) = \frac{1}{\lambda_{\psi_i,n}} \sum_{k=1}^{n} \mathrm{E}\left[I_{\mathbf{A}}(X_k) \,\middle|\, X^{k-1} = x^{k-1} \right] \psi_i(x^{k-1}) = \mu_i(\mathbf{A}).$$

This fact together with Theorem 1 ensure Ψ_1-visibility.

4 Conclusions and Future Work

As is well known in cognitive psychology (see, e.g., [5]), perception is intimately related to expectation: in many cases, we see what we expect to see. In a similar manner, our capacity to recognize new phenomena is conditioned by our existing mathematical constructs (see Fierens and Fine [3]). In the words of Meno to Socrates[3]:

> *And how will you enquire, Socrates, into that which you do not know? What will you put forth as the subject of enquiry? And if you find what you want, how will you ever know that this is the thing which you did not know?* (From [14]).

We have presented here a new way of "seeing" time series by introducing chaotic probability models. Although we have not shown real-world data supporting our models, we have provided the basic tools needed to recognize and study such data. We have developed a basic understanding of chaotic sources by means of the instrumental interpretation in Section 2 and we have presented methods to estimate the model from data and to simulate it given the model in Section 3.

Bridge-building provides a metaphor for our approach to the development of an objective theory based on sets of probability models **M**. The two piers of the bridge are: the model **M** as a set of probability measures on all subsets of **X** (see Section 2) representing potential; time series data in the form of a sequence x^n of finite length n with terms in the sequence all drawn from a finite sample space **X** (see Section 3) representing the actualization of potential. Our models need to show consistent descriptions of both piers and methods to traverse this bridge in both directions. In estimation we have many ways to proceed from a unique data sequence to an approximate model. In simulation we have many ways to proceed from a model to multiple data sequences that are typical of the model.

[3] We owe this quote to an anonymous referee.

Structural soundness of the bridge amounts to self-consistency of estimation and simulation, in the sense that a model $\hat{\mathbf{M}}$ estimated from a simulated sequence $\hat{x}^n$ must be similar to the original model $\mathbf{M}$ being simulated:

$$\mathbf{M} \xrightarrow[\text{source gen.}]{} x^n \xrightarrow[\text{estimation}]{} \hat{\mathbf{M}}(x^n)(\approx \mathbf{M}) \xrightarrow[\text{simulation}]{} \hat{x}^n \xrightarrow[\text{estimation}]{} \hat{\mathbf{M}}(\hat{x}^n) \approx \mathbf{M}.$$

More work remains to be done on estimation and simulation before being able to evaluate fairly this kind of consistency in our chaotic probability models. Also, we need to find a way of quantifying such consistency. Do the models obtained from simulated sequences look similar to the models used for simulation? How do we quantify these similarities? These questions need an answer if we want the framework of chaotic probability models to be consistent.

In view of the instrumental interpretation in Section 2.1, it may be argued that a set of probability measures $\mathbf{M}$ provides only an unfinished picture of a chaotic source, the description of F being needed for a complete model. However, we believe that $\mathbf{M}$ provides, not an incomplete picture of the source, but a coarse grained one. As thermodynamics in physics provides good (complete) enough models of gases for many practical purposes, we believe sets of measures $\mathbf{M}$ may be good (complete) enough models of chaotic sources in many cases. Although examples of the successful use of chaotic models in applied probability have yet to be provided, the main elements needed for the application of our models have been given in this paper.

References

[1] COZMAN, F., AND CHRISMAN, L. Learning convex sets of probability from data. Tech. Rep. CMU-RI-TR-97-25, Robotics Institute, Carnegie Mellon University, 1997.

[2] FIERENS, P. I. *Towards a Chaotic Probability Model for Frequentist Probability*. PhD thesis, Cornell University, August 2003. Available at http://www.people.cornell.edu/pages/pif1/thesis.pdf.

[3] FIERENS, P. I., AND FINE, T. L. Towards a frequentist interpretation of sets of measures. In *Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications* (The Netherlands, 2001), G. de Cooman, T. L. Fine, and T. Seidenfeld, Eds., Shaker Publishing.

[4] FINE, T. L. *Theories of Probability: An Examination of Foundations*. Academic Press, 1973.

[5] GALOTTI, K. M. *Cognitive Science In and Out of the Laboratory*. Brooks/Cole, 1994.

[6] GELL-MANN, M. *The Quark and The Jaguar*. W. H. Freeman and Company, 1994.

[7] GRIZE, Y.-L., AND FINE, T. L. Continuous lower probability-based models for stationary processes with bounded and divergent time averages. *Annals of Probability 15* (1987), 783–803.

[8] KOLMOGOROV, A. N. On tables of random numbers. *Sankhya: The Indian Journal of Statistics, Series A* (1963), 369–376.

[9] KOLMOGOROV, A. N. On logical foundations of probability theory. In *Probability Theory and Mathematical Statistics*, K. Ito and J. Prokhorov, Eds., vol. 1021 of *Lecture Notes in Mathematics*. Springer-Verlag, 1983.

[10] KUMAR, A., AND FINE, T. L. Stationary lower probabilities and unstable averages. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 69* (1985), 1–17.

[11] LI, M., AND VITÁNYI, P. *An Introduction to Kolmogorov Complexity and Its Applications*, second ed. Graduate Texts in Computer Science. Springer, 1997.

[12] PAPAMARCOU, A., AND FINE, T. L. A note on undominated lower probabilities. *Annals of Probability 14* (1986), 710–723.

[13] PAPAMARCOU, A., AND FINE, T. L. Unstable collectives and envelopes of probability measures. *Annals of Probability* (1991), 893–906.

[14] PLATO. Meno. Translated by Bengamin Jowett. Available at http://textkit.com.

[15] SADROLHEFAZI, A., AND FINE, T. L. Finite-dimensional distribution and tail behavior in stationary interval-valued probability models. *Annals of Statistics 22* (1994), 1840–70.

[16] SHAFER, G., AND VOVK, V. *Probability and Finance. It's Only a Game!* Wiley Series in Probability and Statistics. John Wiley & Sons, 2001.

[17] VON MISES, R. *Probability, Statistics and Truth*. Dover, 1981.

[18] WALLEY, P. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

Bounding Analysis of Lung Cancer Risks Using Imprecise Probabilities[*]

MINH HA-DUONG
Centre National de la Recherche Scientifique, France

ELIZABETH CASMAN
Carnegie Mellon University, USA

M. GRANGER MORGAN
Carnegie Mellon University, USA

Abstract

For cancers with more than one risk factor, the sum of probabilistic estimates of the number of cancers attributable to each individual factor may exceed the total number of cases observed when uncertainties about exposure and dose-response for some factors is high. In this study we outline a method to bound the fraction of lung cancer fatalities not attributed to specific well-studied causes in which available data and expert judgment are used to attribute portions of the observed lung cancer mortality to known causes such as smoking, residential radon, and asbestos fibers. An upper bound on the residual risk due to other causes is then inferred using a coherence constraint on the total number of deaths, a maximum uncertainty principle, and imprecise probabilities.

Keywords

bounding analysis, lung cancer, belief functions, assessment methods, medicine

1 Introduction

Usually, the health risk of exposure to an environmental contaminant is calculated using a “front-to-back” procedure, which involves estimating toxic releases, modeling environmental and physiological transformations, and then employing exposure models and dose-response functions, see for example [6]. That methodology works best when the relevant science is well developed; however, when

[*] This work has been supported by CNRS of France, and by the National Science Foundation under grants SES-0216897 and SRB-9521914. A similar paper has been submitted for publication in the journal *Risk Analysis*.

Well characterized factors	Less well characterized factors
Cigarette smoking	Occupational exposures:
Passive smoking	Asbestos
Indoor radon	Arsenic
	Chromates
	Chloromethyl ethers
	Diesel exhaust
	Nickel
	Polycyclic aromatic hydrocarbons (PAHs)
	Ambient air pollution

Table 1: Examples of environmental risk factors for lung cancer

there are several *risk factors* (as the expression is used in the epidemiology literature), and uncertainty about some of the science is large, such a procedure can lead to estimates for the numbers of cancers attributable to the various factors that, summed, exceed the total number of cases actually observed.

Morgan [12] argued that methods of bounding analysis could be used for environmental risk analysis. For health risks with multiple external causes, the available knowledge constrains the magnitude of the poorly characterized risks. If most risks were known with precision, this would be a simple subtraction problem. However disease risks from environmental causes are often estimated from models or inferred from studies involving limited numbers of subjects and inconsistent notions of controls or have other methodological problems that contribute to the uncertainty of the results. It is common to see the central tendencies of such risk estimates expressed as ranges, especially when there are competing plausible models. Sometimes the sum of the individual risks exceeds the total risk. How to quantify and bound the residual "unclaimed" risk is the subject of this paper.

Using lung cancer mortality from environmental factors as an illustrative example, this paper presents a method for bounding the remaining uncertainty when only some of the risk factors are well characterized. The result is an upper bound on the mortality that can be attributed to all other, less well-characterized factors. Some of the major environmental risk factors for lung cancer are shown in Table 1. "Well characterized" here means that population-wide longitudinal attributional studies exist.

In the method presented, expert judgment is used to attribute a portion of the observed cancers to known causes such as smoking, radon and asbestos. Information about the risks from unspecified causes is inferred using a coherence constraint on the total number of deaths, and a principle we term maximum uncertainty.

Our method builds upon the work of Walley [23, chapter 4]. Mathematically, this is an application of Smets' Transferable Belief Model [20], which was developed to solve some paradoxes in combining expert opinion in the theory of evidence [19]. We elicit information about a finite set of variables (risk factors for cancer) and represent this information as constraints on a linear programming problem involving a convex family of probabilities. We invoke the maximum unspecificity criterion in order to estimate the upper bound for the less well-studied members of the set.

Ours is not the first combination of linear programming, expert elicitation, and imprecise probabilities. Lins combined these elements [10] to assess prior probabilities for a single continuous parameter.

The paper is organized as follows. Section 2 presents the conceptual model, which is an application of the mathematical Transferable Beliefs Model to risk assessment. Based on this, Section 3 discusses our method to elicit and validate expert opinion using a maximum unspecificity criterion. From our reading of the literature, we then provide a tentative attribution among the causes (because the expert elicitation phase of this project is currently incomplete), and in Section 4 illustrate the method with a numerical application.

2 Model

2.1 Multiple pollutants may cause lung cancer

Let N denote the magnitude of the health end-point, in this case, the total annual number of lung cancer deaths. Let Ω denote the set of all possible causes of lung cancer deaths. For example, $\Omega = \{C, \qquad R, A, X\}$ where C means tobacco smoke primarily from cigarettes, R means indoor exposure to radon, A means asbestos and X is the group of all other more poorly understood environmental factors of interest.

The model assumes that N is readily observable and therefore known with precision. While this is not strictly true in the case of lung cancer [3, 2] the assumption is not limiting, since the results of the method can be stated in percentage terms and then applied to a range of possible numerical values of N. We also assume exposure to be binary, which is of course not true, but the assumption is consistent with the exposure definitions used in the supporting epidemiological studies. With these two assumptions, each death can be linked to zero or more possible causes in Ω. Most lung cancer deaths are caused by smoking alone, but there are synergistic cases in which more than one cause is involved, such as smoking and radon.

Figure 1 shows one way to subdivide N by causes that includes synergistic effects. We denote the number of deaths linked to cause s as $n(s)$, where s is any subset of Ω. In our example we consider four possible causes in Ω, so there could be sixteen ($= 2^4$) possible s, but to simplify the analysis and to be consistent with the cancer literature, we will consider only the two-factor interactions involving

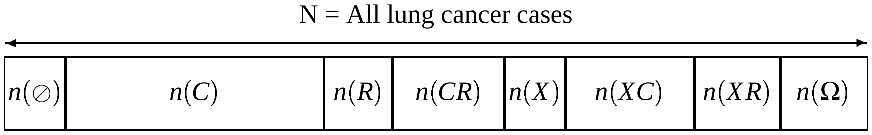


Figure 1: The basic statistic n, simplified to include only the risk factors cigarettes (C), radon (R), and all other causes (X). N is the total number of lung cancer fatalities. n is the number of fatalities attributable to each risk factor, or combination of factors. $n(\oslash)$ is the background number of lung cancer deaths that would occur absent all the various risk factors. $n(\Omega) = n(CRX)$, those cases for which no risk factor can be excluded.

cigarette smoke.

To adopt a more precise and cautious definition, $n(s)$ is the number of cases not caused by pollutants not in s. This implies that causes not in s are known to be non-contributing to that lung cancer. For deaths in $n(s)$, any cause in s may have caused the lung cancer, but which one is uncertain and there may have been synergies.

Our intuitive interpretation for this definition of "ambiguous causality" is that $n(s)$ represents the number of cases that were exposed to the possibly multiple risk factors in s.

The number of lung cancer deaths where all causes of Ω have been positively excluded is $n(\oslash)$ shown to the left of the bar in Figure 1. Cases that could not be linked to any pollutant in Ω are considered spontaneous lung cancer. It is important to underline that $n(\oslash)$ does not have the same status as $n(X)$, which will be deduced as a residual. It corresponds to the background rate of lung cancer that occurs in a population without exposure to any environmental, dietary, occupational or other carcinogen.

The function n does not come from real data. Direct measurement of the basic statistic n is impossible, since exposure to a pollutant does not necessarily result in a cancer fatality and because retrospectively, lifetime exposures to the various carcinogens can only be roughly estimated. It is only a mathematical tool used in to support expert elicitation of consistent bounds, as discussed next.

2.2 Bounding the risk attributable to single and joint pollutants

The basic statistic n can be used to bound the number of cases attributable to smoking C as follows, where $\overline{n}(C)$ and $\underline{n}(C)$ denote the upper and lower bounds on $n(C)$, respectively:

- The lower bound is the number of cases attributed only to smoking (we

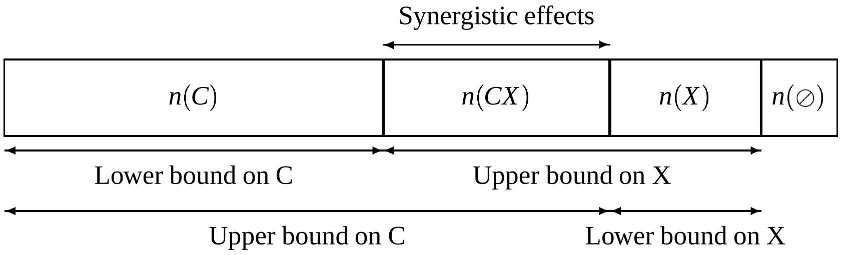


Figure 2: Upper and lower bounds on the number of lung cancer deaths attributable to C and X

lump both passive and active smoking together). That is $\underline{n}(C)$=$n(C)$.

- The upper bound is the number of cases exposed to smoke and possibly other factors. That is $\overline{n}(C) = n(C) + n(XC) + n(CR) + n(XCR) + n(CA) + n(XCA) + n(CRA) + n(XCRA)$ or:

$$\overline{n}(C) = \sum_E n(E) \text{ for all subsets } E \text{ of } \Omega \text{ containing } C$$

Figure 2 illustrates this definition of the upper and lower bounds of the number of lung cancer deaths attributable to X and C. For clarity the figure is drawn showing only two causes, with $\Omega = \{C, X\}$.

In epidemiologists’ terms, the *attributable fraction* of pollutant C is the proportion of all cases that could be avoided if this pollutant were eliminated, denoted $af(C)$. The model suggests the following bounds for smoking attributable fraction:

$$\frac{\underline{n}(C)}{N}(1 - r_0) \leq af(C) \leq \frac{\overline{n}(C)}{N}(1 - r_0) \tag{1}$$

The lower bound accounts for the $1 - r_0$ share of spontaneous lung cancer cases in those cases exposed to cigarettes. The upper bound attributes all cigarette-exposed deaths to this factor.

For this paper we will assume that the background rate of lung cancer mortality in the U.S., r_0, is 3 deaths per 100 000 people. This background rate is the number of lung cancer deaths in the unexposed population divided by the unexposed population. Denoting p_C, p_R and p_A as the exposure probabilities of C, R, A; and T as the total population; assuming independence (meaning that people who smoke are no more or less likely to be exposed to radon or to asbestos):

$$r_0 = \frac{n(\emptyset)}{(1 - p_C)(1 - p_R)(1 - p_A)T} \tag{2}$$

Consider now the bounds on deaths attributed to multiple synergistic causes. Denote these causes s, a subset of Ω, for example $s = CR$. For the lower bound on the number of deaths attributable to these causes acting jointly, we continue to adopt the number of cases exposed only to these causes, that is:

$$\underline{n}(s) = n(s) \tag{3}$$

And as the upper bound, we continue to adopt the number of cases exposed to s and possibly other factors, that is:

$$\overline{n}(s) = \sum_E n(E) \text{ for all subsets } E \text{ of } \Omega \text{ containing } s \tag{4}$$

This $\overline{n}$ corresponds to the commonality function in the Transferable Belief Model [20]. Bounds on the attributable fraction can be computed as in equation 1.

2.3 Unspecificity, a measure of uncertainty

Structurally, the only uncertainty in this bounding analysis model comes from the synergistic causes, because it is not possible to attribute the cancer to any one of these causes. Consider these two (of the three) extreme cases:

- If each death were attributed to exactly one cause, then there would be no uncertainty, and all lower bounds would coincide with their upper counterpart. We would have $n(C) + n(R) + n(A) + n(X) = N - n(\emptyset)$. Note that since n is a positive function that sums up to N, this implies that $n(s) = 0$ for all other subsets.
- If no information were available, each death would be attributed to the synergy of all factors. We would have all the lower bounds at 0 and all upper bounds at N. Mathematically, this is $n(\Omega) = N$. Note that this constitutes a proper uninformative distribution: it is not the Bayesian uniform prior probability distribution on Ω. It represents the family of all probability distributions that can be defined on Ω.

Unspecificity is an numeric indicator that equals one in the first case, and in the second case equals the number of elements of Ω. It is the expected value of the number of elements of s with respect to the probability distribution $m(s) = \frac{n(s)}{N}$, that is:

$$U = \frac{n(C) + n(R) + n(A) + n(X) + 2(n(CR) + n(RA) + \ldots) + 3\ldots + 4n(\Omega)}{N} \tag{5}$$

In this paper unspecificity is a kind of generalized cardinality, that specifies the number of alternatives. The reason for using this word is that when a death is attributed to the synergy of k factors, it can be said that the unspecificity of this information is k. See [16] for an extensive discussion of this concept.

A lower unspecificity measure corresponds to better information, so a third extreme case needs discussion: unspecificity is zero when and only when $n(\oslash) = N$. This is the case when for all deaths, all non-spontaneous causes of Ω have been positively excluded. It means that all the substances in Ω are actually safe (with respect to lung cancer). This is the highest level of information achievable, to the point that it makes Ω irrelevant.

Regarding unspecificity as a measure of information allows to implement numerically the general principle of maximum uncertainty, also known as Laplace's principle of "raison insuffisante". The principle states that one should select the statistic that is the most unspecific, compatible with existing information. This is the principle that we use in the next section to estimate the bounds on the unknown cause, given information about all others.

3 Expert elicitation

3.1 Procedure

When we apply this procedure, we will elicit a set of judgments regarding $n(s)$ from a number of leading health scientists using methods previously developed for expert elicitation in domains in which there is considerable scientific evidence [13, 14, 15]

The results from an elicitation will be interpreted as linear constraints on n. These constraints determine a set $\mathcal{B}$ of basic statistics, that is a set of n that are all compatible with the expert's judgments. The most unspecific n in $\mathcal{B}$ is chosen to represent the expert judgment, according to the maximum uncertainty principle. This amounts to solving a linear program in a space with $2^{|\Omega|}$ dimensions.

Other ways of translating judgments into constraints are possible, for example using relative risk, but are not used in this introductory paper. Note that both quantitative and comparative judgments are possible, which may ultimately be important because some of the pollutants have been well studied, but we are interested in the less well-known pollutants.

In addition to elicited information, we impose these constraints:

- It is understood that all $n(s)$ are non-negative, summing up to N.
- Three-way interactions and higher are not allowed. That is, $n(s) = 0$ if s has 3 or more elements.

The constraint on three-way interaction is a zero-order approximation. We assume that that the number of deaths caused by multiple interactions are a very small number that can be neglected. In a more sophisticated approach, this assumption could be replaced by explicit considerations about causes interactivity and independence. But there is little scientific empirical knowledge about these interactions.

3.2 Ensuring consistency

Maximizing unspecificity is possible only if $\mathcal{B}$ is not empty. This means that the different items of information given by the expert should be coherent with each other. For example, one could not allow the expert to say that the lower bound for C is 90 percent, and the lower bound for R is 20 percent at the same time, because that would exceed 100 percent. Walley has shown [23] that the coherence condition is:

$$\overline{af}(s_i) + \sum_{j \neq i} \underline{af}(s_j) \leq 1 \leq \underline{af}(s_i) + \sum_{j \neq i} \overline{af}(s_j)$$

The double inequality should hold for all causes i in $\{1, \dots, |\Omega|\}$.

Besides mathematical consistency, it is also important to provide safeguards so that the expert can check that formal implications of the elicited n are consistent with its informal understanding of the problem. We propose two checks.

The first check on n is to make sure that the results in terms of bounds on relative risks and on interactions between pollutants make sense. The definition of relative risk for smoking cigarettes $rr(C)$, for example, is the lung cancer rate associated with exposure to tobacco smoke divided by the background lung cancer rate. Given exposure probabilities in the general population, we will assess the bounds on the relative risk for the various pollutants using the formula in [6, appendix C p. 229].

The second check on n is to make sure that the risk-ranking it implies makes sense. We will ask experts to rank risks during the elicitation process. The consistency of results will be assessed by comparing the partial order derived from n with the expert's *a priori* risk ranking.

Informally, this partial order says that the lung cancer risk related to R is not larger than the risk related to C when we know with certainty that R causes fewer lung cancer deaths than C. For example, one sufficient condition for this is that the lower bound on C is greater than the upper bound on R. But the mathematical definition of the natural partial order relation associated with a basic statistic n requires more explanations.

Let P denote a function such that $P(C)+P(R)+P(A)+P(X)=N$. It is a basic statistic with unspecificity one, describing an hypothetical world where each lung cancer death is attributed to one and only one cause. For such a P, the number of deaths caused by any set of causes s is $\sum_x P(x)$, for all causes x in s. We say that P is compatible with the basic statistic n if and only if for all s, that number respects the bound determined by n in the following way:

$$\forall s \subset \Omega, \sum_{x \in s} P(x) \leq \sum_{y, s \cap y \neq \emptyset} n(y) \tag{6}$$

The right hand side of Equation 6 can be interpreted in the present model as the upper bound on the number of deaths related to the causes in s acting either jointly *or separately*. This function of s corresponds to the belief function in the Transferable Belief Model.

The heart of the problem is that P is hypothetical. Because there are interactions, more than one P is compatible with n. Denote $\mathcal{P}$ the family of all P compatible with n. The natural partial order is mathematically defined by:

$$R \preceq C \Leftrightarrow \forall P \in \mathcal{P},\, P(R) \leq P(C) \tag{7}$$

Numerically, this is determined by checking the sign of the minimum of $P(C) - P(R)$ under constraint 6. It is tractable to work with the full partial order, since there is at most $|\Omega|(|\Omega| - 1)/2$ comparisons. Assuming $|\Omega| = 7$ for example, there are no more than 21 information items, which can be presented naturally in the diagonal half of a table. Moreover, practically there will be fewer than 21 items, since not all risks can be compared. It is to be expected, for example, that some experts may prefer to find that some of the less-known risks are not comparable, because of missing scientific information.

4 Application

Our numerical simulations were performed using a *Mathematica* notebook[1]. The code directly implements matrix calculus for belief functions as outlined in [21]. This is the most straightforward method given that Ω remains small, but it would not scale well to tens of pollutants, since it involves square matrices with $2^{2|\Omega|}$ elements. For example, 10 pollutants implies storing in memory arrays with 1M numbers.

In our illustration Ω, the set of possible causes of lung cancer, consists of:

C	Smoking
R	Radon
A	Asbestos, glass wool, ceramic fibers
X	All other environmental risk factors

Based on our own review of the literature [6, 7, 4, 18, 22, and others] we have constructed a set of judgments attributing lung cancer deaths among the major causes, as the expert elicitations have not at this time been performed. We offer the following breakdown: Cigarette smoking combined with passive smoking accounts for 70 to 95 percent of lung cancer mortality; indoor radon exposures for 02 to 21 percent; asbestos, 1 to 5 percent.

[1] Available on the web at `http://www.andrew.cmu.edu/user/mduong`, or upon request, under the GNU General Public License.

Bounds	C	R	A	X
$\overline{af}$	95%	21%	5%	3.2%
$\underline{af}$	70%	02%	1%	0%
Exposure probability	45%	50%	5%	5%
$\overline{rr}$	43.2	1.53	2.05	1.66
$\underline{rr}$	6.19	1.04	1.20	1.

Table 2: Results of optimization: Upper and lower bounds on attributable fractions and relative risks

We used a 3% background rate [1, 5, 11, 9, 17]. With our assumptions on exposure probabilities, equation 2 implies that $n(\oslash) = 0.013N$.

The next table shows the implications for bounds of af and rr of the most unspecific imprecise probability distribution compatible with these constraints. The exposure probabilities needed to compute rr are exogenous: radon exposure is defined as living in a home with radon concentration at or above 25 Bqm^{-1}, and exposure to X is our estimate. The effect of this calculation on the bounds of rr would serve as a calibration/validation reference for the expert who may be more familiar with small sample studies than population effects, and might adjust his or her initial responses in light of seeing their mathematical implications.

This result attributes between 0 and 3.2 percent of lung cancer deaths to X, the group of unknown environmental pollutants. For the group of known and suspected lung carcinogens other than C, A and R, the risk analyst concludes that, *if one is confident in the bounds assigned to the well understood risk factors*, the sum of the effects of the other factors accounts for no more than 3.2% of total lung cancer mortality.

The implication for judging future risk assessments of members of X is that, if the assessment projects the lung cancer risk in the U. S. population from these pollutants to be in excess of 3.2% of the annual lung cancer mortality, then the assumptions of the model should be re-examined and the upper bound on the resulting estimate constrained.

5 Concluding remarks

5.1 Discussion

With less than ten pollutants, computing time is not a problem. Expert elicitation could be done interactively, solving for n after each expert's reply. This would allow the interviewer to point out and resolve inconsistency when there is no solution. But assuming that experts were willing to form judgments on a wider range of pollutants, the curse of dimensionality can be addressed along the following lines. Rather than using matrix calculus, it is possible to use faster algorithms (namely the Fast Möbius transform) for belief function computations. If this is not enough, further simplifications can be made if additional assumptions on n, for example disallowing 3-way or higher interactions, are accepted.

The proposed method takes all information items provided by the expert with equal force. A potential advance of this research could be to ask experts to rank the reliability of each information item, or even to give an estimate of confidence for them.

Further research could deal with inter-expert validation, a question linked with the unresolved issue of judgment fusion. The Transferable Belief Model underlying this work offers a measure of contradiction between different sources of information: it reinterprets $n(\emptyset)$, the number of spontaneous lung cancer deaths found when one combines the opinion of all experts. The problem is how to combine the experts.

Each expert's judgment determines a set $\mathcal{B}$ of coherent basic statistics. If the intersection of all these sets is non-empty, then experts agree on this intersection. The principle of maximum unspecificity can be used to form a group judgment.

If the intersection is empty, the experts contradict each other. Studying which information items cause the contradiction (which constraints make the LP infeasible) can identify the substantive sources of disagreement, and in that way inform both future research priorities as well as the decision-making process. How (or if) to fuse the judgments and quantify the degree of contradiction is still an active research question, see [8] for example.

5.2 Conclusion

This paper has proposed an application of the Transferable Belief Model [20] to estimate an upper bound on the number of lung cancers caused annually by the group of causes for which comprehensive longitudinal studies are lacking. Such a result is interesting from a risk management perspective, as it gives an indication of the level of effort control of these pollutants deserve.

This was done by attributing a portion of the observed cancers to known causes such as smoking, radon and asbestos, and then deducing information about the residual using maximum unspecificity. The critical aspects of this procedure are:

1. Uncertainty in the known causes is explicitly stated, using statements on upper and lower bounds.

2. Synergistic effects in the known causes are part of the framework.

3. Consistency between known causes and poorly understood agents is required. (As Figure 2 illustrates, it is the lower bound on smoking that mostly

constrains the upper bound on the residual.)

This paper presents the methodology. The results revealed by future expert elicitation will be the subject of another paper.

References

[1] ALAVANJA, M. C. R., BROWNSON, R. C., BENICHOU, J., SWANSON, C., AND BOICE, J. D. J. Attributable risk of lung cancer in lifetime nonsmokers and long-term ex-smokers (missouri, united states). *Cancer Causes and Control 6* (1995), 209–216.

[2] AMERICAN CANCER INSTITUTE. Cancer facts & figures. Tech. rep., American Cancer Society, Inc. Surveillance Research, 2003.

[3] ARCHER, V. E., AND LYON, J. L. Errors and biases in the diagnosis of cancer of the lung and their influence on risk estimates. *Medical hypotheses 54(3)* (2000), 400–407.

[4] AXELSON, O. Alternative for estimating the burden of lung cancer from occupational exposures – some calculations based on data from swedish men. *Scandinavian Journal of Work, Environment, and Health 28*, 1 (2002), 58–63.

[5] AXELSON, O., DAVIS, D. L., FORESTIERE, F., SCHNEIDERMAN, M., AND WAGENER, D. Lung cancer not attributable to smoking. *Annals of the New York Academy of Sciences 609*, 78 (1990), 165–178.

[6] BEIR VI, C. *Health effects of exposure to radon*. National Academy Press, 1999.

[7] CENTERS FOR DISEASE CONTROL AND PREVENTION. Smoking-attributable mortality, morbidity, and economic costs (SAMMEC): adult SAMMEC and maternal and child health (MCH) SAMMEC software, 2002.

[8] ISIF (INTERNATIONAL SOCIETY ON INFORMATION FUSION). *Fusion 2000 Conference* (Paris, 10–13 July 2000).

[9] KOO, L. C., AND HO, J. H. Worldwide epidemiological patterns of lung cancer in nonsmokers. *International Journal of Epidemiology 19,* Suppl. 1 (1990), s14–s23.

[10] LINS, G. C. N., AND DE SOUZA, F. M. C. A protocol for the elicitation of prior distributions. In *Second International Symposium on Imprecise Probability and their Applications* (New York, USA, 26–29 June 2001), Cornell University.

[11] Lyon, J. L., Gardner, J. W., and West, D. W. Cancer in utah: risk by religion and place of residence. *J Natl Cancer Inst 65*, 5 (1980), 1063–71.

[12] Morgan, M. G. The neglected art of bounding analysis. *Environmental Science and Technology 35*, 7 (Apr.1 2001), 162A–164A.

[13] Morgan, M. G., and Henrion, M. *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge University Press, 1990.

[14] Morgan, M. G., and Keith, D. Subjective judgments by climate experts. *Environmental Science and Technology 29*, 10 (Oct. 1995), 468A–476A.

[15] Morgan, M. G., Pitelka, L. F., and Shevliakova, E. Elicitation of expser judgments of climate change impact on forest ecosystems. *Climatic Change 49* (2001), 279–307.

[16] Rocha, L. M. Relative uncertainty and evidence sets: a constructivist framework. *International Journal of General Systems 26*, 1–2 (1997), 35–61.

[17] Samet, J. M., Brenner, D., Brooks, A. L., Ellett, W. H., Gilbert, E. S., Goodhead, D. T., Hall, E. J., Hopke, P. K., Krewski, D., Lubin, J. H., McClellan, R. O., and Ziemer, P. L. Health effects of exposure to radon, 1999.

[18] Schoenberg, J. B., Stemhagen, A., Mason, T. J., Patterson, J., Bill, J., and Altman, R. Occupation and lung cancer risk among new jersey white males. *Journal of the National Cancer Institute 79*, 1 (1987), 13–21.

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

[20] Smets, P. Belief functions and the transferable belief model, 2000.

[21] Smets, P. Matrix calculus for belief functions, 2001.

[22] Vineis, P., and Simonato, L. Proportion of lung and bladder cancers in males resulting from occupation: a systematic approach. *Archives of Environmental Health 46*, 1 (1991), 6–15.

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

Minh Ha-Duong is Visiting Professor at the Engineering and Public Policy Department, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213. E-mail: minh.ha.duong@cmu.edu

Elizabeth Casman is Research Engineer at the Engineering and Public Policy Department, Carnegie Mellon University.

M. Granger Morgan is Lord Chair Professor in Engineering; Professor and Department Head, Engineering and Public Policy; Professor, Electrical and Computer Engineering, and The H. John Heinz III School of Public Policy and Management, Carnegie Mellon University.

Robust Estimators under the Imprecise Dirichlet Model

MARCUS HUTTER
IDSIA, Switzerland

Abstract

Walley's Imprecise Dirichlet Model (IDM) for categorical data overcomes several fundamental problems which other approaches to uncertainty suffer from. Yet, to be useful in practice, one needs efficient ways for computing the imprecise=robust sets or intervals. The main objective of this work is to derive exact, conservative, and approximate, robust and credible interval estimates under the IDM for a large class of statistical estimators, including the entropy and mutual information.

1 Introduction

This work derives interval estimates under the Imprecise Dirichlet Model (IDM) [Wal96] for a large class of statistical estimators. In the IDM one considers an i.i.d. process with unknown chances[1] π_i for outcome i. The prior uncertainty about $\boldsymbol{\pi}$ [2] is modeled by a set of Dirichlet priors[3] $\{p(\boldsymbol{\pi}) \propto \prod_i \pi_i^{st_i-1} : \boldsymbol{t} \in \Delta\}$, where[4] $\Delta := \{\boldsymbol{t} : t_i \geq 0, \sum_i t_i = 1\}$, and s is a hyper-parameter, typically chosen between 1 and 2. Sets of probability distributions are often called Imprecise probabilities, hence the name IDM for this model. We avoid the term *imprecise* and use *robust* instead, or capitalize *Imprecise*. IDM overcomes several fundamental problems which other approaches to uncertainty suffer from [Wal96]. For instance, IDM satisfies the representation invariance principle and the symmetry principle, which are mutually exclusive in a pure Bayesian treatment with proper prior [Wal96]. The counts n_i for i form a minimal sufficient statistic of the data of size $n = \sum_i n_i$. Statistical estimators $F(\boldsymbol{n})$ usually also depend on the chosen

[1] Also called *objective* or *aleatory* probabilities.

[2] We denote vectors by $\boldsymbol{x} := (x_1, ..., x_d)$ for $\boldsymbol{x} \in \{\boldsymbol{n}, \boldsymbol{t}, \boldsymbol{u}, \boldsymbol{\pi}, ...\}$.

[3] Also called *second order* or *subjective* or *belief* or *epistemic* probabilities.

[4] Strictly speaking, Δ should be the open simplex [Wal96], since $p(\boldsymbol{\pi})$ is improper for $\boldsymbol{t}$ on the boundary of Δ. For simplicity we assume that, if necessary, considered functions of $\boldsymbol{t}$ can and are continuously extended to the boundary of Δ, so that, for instance, minima and maxima exist. All considerations can straightforwardly, but cumbersomely, be rewritten in terms of an open simplex. Note that open/closed Δ result in open/closed robust intervals, the difference being numerically/practically irrelevant.

prior: so a set of priors leads to a set of estimators $\{F_{\boldsymbol{t}}(\boldsymbol{n}) : \boldsymbol{t} \in \Delta\}$. For instance, the expected chances $E_{\boldsymbol{t}}[\pi_i] = \frac{n_i+st_i}{n+s} =: u_i(\boldsymbol{t})$ lead to a robust interval estimate $[\frac{n_i}{n+s}, \frac{n_i+s}{n+s}] \ni E_{\boldsymbol{t}}[\pi_i]$. Robust intervals for the variance $\text{Var}[\pi_i]$ [Wal96] and for the mean and variance of linear-combinations $\sum_i \alpha_i \pi_i$ have also been derived [Ber01]. Bayesian estimators (like expectations) depend on $\boldsymbol{t}$ and $\boldsymbol{n}$ only through $\boldsymbol{u}$ (and $n+s$ which we suppress), i.e. $F_{\boldsymbol{t}}(\boldsymbol{n}) = F(\boldsymbol{u})$. The main objective of this work is to derive approximate, conservative, and exact intervals $[\min_{\boldsymbol{t}\in\Delta} F(\boldsymbol{u}), \max_{\boldsymbol{t}\in\Delta} F(\boldsymbol{u})]$ for general $F(\boldsymbol{u})$, and for the expected (also called predictive) entropy and the expected mutual information in particular. These results are key building blocks for applying IDM. Walley suggests, for instance, to use $\min_{\boldsymbol{t}} P_{\boldsymbol{t}}[\mathcal{F} \geq c] \geq \alpha$ for inference problems and $\min_{\boldsymbol{t}} E_{\boldsymbol{t}}[\mathcal{F}] \geq c$ for decision problems [Wal96], where $\mathcal{F}$ is some function of $\boldsymbol{\pi}$. One application is the inference of robust tree-dependency structures [Zaf01, ZH03], in which edges are partially ordered based on Imprecise mutual information.

Section 2 gives a brief introduction to IDM and describes our problem setup. In Section 3 we derive exact robust intervals for concave functions F, such as the entropy. Section 4 derives approximate robust intervals for arbitrary F. In Section 5 we show how bounds of elementary functions can be used to get bounds for composite function, especially for sums and products of functions. The results are used in Section 6 for deriving robust intervals for the mutual information. The issue of how to set up IDM models on product spaces is discussed in Section 7. Section 8 addresses the problem of how to combine Bayesian credible intervals with the robust intervals of the IDM. Conclusions are given in Section 9.

2 The Imprecise Dirichlet Model

Random i.i.d. processes. We consider discrete random variables $\imath \in \{1,...,d\}$ and an i.i.d. random process with outcome $i \in \{1,...,d\}$ having probability π_i. The chances $\boldsymbol{\pi}$ form a probability distribution, i.e. $\boldsymbol{\pi} \in \Delta := \{\boldsymbol{x} \in I\!R^d : x_i \geq 0 \,\forall i,\, x_+ = 1\}$, where we have used the abbreviation $\boldsymbol{x} = (x_1,...,x_d)$ and $x_+ := \sum_{i=1}^d x_i$. The likelihood of a specific data set $\boldsymbol{D}$ with n_i observations i and total sample size $n = n_+ = \sum_i n_i$ is $p(\boldsymbol{D}|\boldsymbol{\pi}) = \prod_i \pi_i^{n_i}$. The chances π_i are usually unknown and have to be estimated from the sample frequencies n_i. The frequency estimate $\frac{n_i}{n}$ for π_i is one possible point estimate.

Second order p(oste)rior. In the Bayesian approach one models the initial uncertainty in $\boldsymbol{\pi}$ by a (second order) prior “belief” distribution $p(\boldsymbol{\pi})$ with domain $\boldsymbol{\pi} \in \Delta$. The Dirichlet priors $p(\boldsymbol{\pi}) \propto \prod_i \pi_i^{n'_i-1}$, where n'_i comprises prior information, represent a large class of priors. n'_i may be interpreted as (possibly fractional) virtual number of “observation”. High prior belief in i can be modeled by large n'_i. It is convenient to write $n'_i = s \cdot t_i$ with $s := n'_+$, hence $\boldsymbol{t} \in \Delta$. Having no initial bias one should choose a prior in which all t_i are equal, i.e. $t_i = \frac{1}{d} \,\forall i$.

Examples for s are 0 for Haldane's prior [Hal48], 1 for Perks' prior [Per47], $\frac{d}{2}$ for Jeffreys' prior [Jef46], and d for Bayes-Laplace's uniform prior [GCSR95]. From the prior and the data likelihood one can determine the posterior $p(\boldsymbol{\pi}|\boldsymbol{D}) = p(\boldsymbol{\pi}|\boldsymbol{n}) \propto \prod_i \pi_i^{n_i+st_i-1}$.

The posterior $p(\boldsymbol{\pi}|\boldsymbol{D})$ summarizes all statistical information available in the data. In general, the posterior is a very complex object, so we are interested in summaries of this plethora of information. A possible summary is the expected value or mean $E_{\boldsymbol{t}}[\pi_i] = \frac{n_i+st_i}{n+s}$ which is often used for estimating π_i. The accuracy may be obtained from the covariance of $\boldsymbol{\pi}$.

Usually one is not only interested in an estimation of the whole vector $\boldsymbol{\pi}$, but also in an estimation of scalar functions $\mathcal{F} : \Delta \to \mathbb{R}$ of $\boldsymbol{\pi}$, such as the entropy $\mathcal{H}(\boldsymbol{\pi}) = -\sum_i \pi_i \log \pi_i$, where log denotes the natural logarithm. Since $\mathcal{F}$ is itself a random variable we could determine the posterior distribution $p(\mathcal{F}_0|\boldsymbol{n}) = \int_\Delta \delta(\mathcal{F}(\boldsymbol{\pi}) - \mathcal{F}_0) p(\boldsymbol{\pi}|\boldsymbol{n}) d\boldsymbol{\pi}$ of $\mathcal{F}$, which may further be summarized by the posterior mean $E_{\boldsymbol{t}}[\mathcal{F}] = \int_\Delta \mathcal{F}(\boldsymbol{\pi}) p(\boldsymbol{\pi}|\boldsymbol{n}) d\boldsymbol{\pi}$ and possibly the posterior variance $\text{Var}_{\boldsymbol{t}}[\mathcal{F}]$. A simple, but crude approximation for the mean can be obtained by exchanging E with $\mathcal{F}$ (exact only for linear functions): $E_{\boldsymbol{t}}[\mathcal{F}(\boldsymbol{\pi})] \approx \mathcal{F}(E_{\boldsymbol{t}}[\boldsymbol{\pi}])$. The approximation error is typically of the order $\frac{1}{n}$.

The Imprecise Dirichlet Model. The classical approach, which consists of selecting a single prior, suffers from a number of problems. Firstly, choosing for example a uniform prior $t_i = \frac{1}{d}$, the prior depends on the particular choice of the sampling space. Secondly, it assumes exact prior knowledge of $p(\boldsymbol{\pi})$. The solution to the second problem is to model our ignorance by considering sets of priors $p(\boldsymbol{\pi})$, often called Imprecise probabilities. The specific *Imprecise Dirichlet Model* (IDM) [Wal96] considers the set of *all* $\boldsymbol{t} \in \Delta$, i.e. $\{p(\boldsymbol{\pi}|\boldsymbol{n}) : \boldsymbol{t} \in \Delta\}$ which solves also the first problem. Walley suggests to fix the hyperparameter s somewhere in the interval $[1,2]$. A set of priors results in a set of posteriors, set of expected values, etc. For real-valued quantities like the expected entropy $E_{\boldsymbol{t}}[\mathcal{H}]$ the sets are typically intervals, which we call robust intervals

$$E_{\boldsymbol{t}}[\mathcal{F}] \in [\min_{\boldsymbol{t}\in\Delta} E_{\boldsymbol{t}}[\mathcal{F}], \max_{\boldsymbol{t}\in\Delta} E_{\boldsymbol{t}}[\mathcal{F}]].$$

Problem setup and notation. Consider any statistical estimator F. F is a function of the data $\boldsymbol{D}$ and the hyperparameters $\boldsymbol{t}$. We define the general correspondence

$$u_i^{\cdots} = \frac{n_i + st_i^{\cdots}}{n+s}, \quad \text{where } ^{\cdots} \text{ can be various superscripts.} \tag{1}$$

F can, hence, be rewritten as a function of $\boldsymbol{u}$ and $\boldsymbol{D}$. Since we regard $\boldsymbol{D}$ as fixed, we suppress this dependence and simply write $F = F(\boldsymbol{u})$. This is further motivated by the fact that all Bayesian estimators of functions $\mathcal{F}$ of $\boldsymbol{\pi}$ only depend on $\boldsymbol{u}$ and the sample size $n+s$. It is easy to see that this holds for the mean, i.e. $E_{\boldsymbol{t}}[\mathcal{F}] = F(\boldsymbol{u}; n+s)$, and similarly for the variance and all higher (central) moments. The main focus of this work is to derive exact and approximate expressions for upper and lower F values

$$\overline{F} := \max_{\boldsymbol{t}\in\Delta} F(\boldsymbol{u}) \quad \text{and} \quad \underline{F} := \min_{\boldsymbol{t}\in\Delta} F(\boldsymbol{u}), \qquad \overline{\underline{F}} := [\underline{F}, \overline{F}]$$

$\boldsymbol{t} \in \Delta \Leftrightarrow \boldsymbol{u} \in \Delta'$, where $\Delta' := \{\boldsymbol{u} : u_i \geq \frac{n_i}{n+s},\ u_+ = 1\}$. We define $\boldsymbol{u}^{\overline{F}}$ as the $\boldsymbol{u} \in \Delta'$ which maximizes F, i.e. $\overline{F} = F(\boldsymbol{u}^{\overline{F}})$, and similarly $\boldsymbol{t}^{\overline{F}}$ through relation (1). If the maximum of F is assumed in a corner of Δ' we denote the index of the corner by $i^{\overline{F}}$, i.e. $t_i^{\overline{F}} = \delta_{ii^{\overline{F}}}$, where δ_{ij} is Kronecker's delta function. Similarly $\boldsymbol{u}^{\underline{F}}$, $\boldsymbol{t}^{\underline{F}}$, $i^{\underline{F}}$.

3 Exact Robust Intervals for Concave Estimators

In this section we derive exact expressions for $\overline{\underline{F}}$ if $F : \Delta \to I\!R$ is of the form

$$F(\boldsymbol{u}) = \sum_{i=1}^{d} f(u_i) \quad \text{and concave} \quad f : [0,1] \to I\!R. \tag{2}$$

The expected entropy is such an example (discussed later). Convex f are treated similarly (or simply take $-f$).

The nature of the solution. The approach to a solution of this problem is motivated as follows: Due to symmetry and concavity of F, the global maximum is attained at the center $u_i = \frac{1}{d}$ of the probability simplex Δ, i.e. the more uniform $\boldsymbol{u}$ is, the larger $F(\boldsymbol{u})$. The nearer $\boldsymbol{u}$ is to a vertex of Δ, i.e. the more unbalanced $\boldsymbol{u}$ is, the smaller is $F(\boldsymbol{u})$. The constraints $t_i \geq 0$ restrict $\boldsymbol{u}$ to the smaller simplex

$$\Delta' = \{\boldsymbol{u} : u_i \geq u_i^0,\ u_+ = 1\} \quad \text{with} \quad u_i^0 := \frac{n_i}{n+s},$$

which prevents setting $u_i^{\overline{F}} = \frac{1}{d}$ and $u_i^{\underline{F}} = \delta_{i1}$. Nevertheless, the basic idea of choosing $\boldsymbol{u}$ as uniform / as unbalanced as possible still works, as we will see.

Greedy $F(\boldsymbol{u})$ minimization. Consider the following procedure for obtaining $\boldsymbol{u}^{\underline{F}}$. We start with $\boldsymbol{t} \equiv \boldsymbol{0}$ (outside the usual domain Δ of F, which can be extended to $[0,1]^d$ via (2)) and then gradually increase $\boldsymbol{t}$ in an axis-parallel way until $t_+ = 1$. With axis-parallel we mean that only one component of $\boldsymbol{t}$ is increased, which one possibly changes during the process. The total zigzag curve from $\boldsymbol{t}^{start} = \boldsymbol{0}$ to $\boldsymbol{t}^{end}$ has length $t_+^{end} = 1$. Since all possible curves have the same (Manhattan) length 1, $F(\boldsymbol{u}^{end})$ is minimized for the curve which has (on average) smallest F-gradient along its path. A greedy strategy is to follow the direction i of currently smallest F-gradient $\frac{\partial F}{\partial t_i} = f'(u_i)\frac{s}{n+s}$. Since f' is monotone decreasing ($f'' < 0$), $\frac{\partial F}{\partial t_i}$ is smallest for largest u_i. At $\boldsymbol{t}^{start} = \boldsymbol{0}$, $u_i = \frac{n_i}{n+s}$ is largest for $i = i^{min} := \arg\max_i n_i$. Once we start in direction i^{min}, $u_{i^{min}}$ increases even further whereas all other u_i ($i \neq i^{min}$) remain constant. So the moving direction is never changed and finally we reach a local minimum at $t_i^{end} = \delta_{ii^{min}}$. In [Hut03] we show that this is a global minimum, i.e.

$$t_i^{\underline{F}} = \delta_{ii^{\underline{F}}} \quad \text{with} \quad i^{\underline{F}} := \arg\max_i n_i. \tag{3}$$

Greedy $F(\boldsymbol{u})$ maximization. Similarly we maximize $F(\boldsymbol{u})$. Now we increase $\boldsymbol{t}$ in direction $i = i_1$ of maximal $\frac{\partial F}{\partial t_i}$, which is the direction of smallest $u_i \propto n_i + st_i$. Again, (only) u_{i_1} increases, but possibly reaches a value where it is no longer the smallest one. We stop if it becomes equal to the second smallest u_i, say $i = i_2$. We now have to increase u_{i_1} and u_{i_2} with same speed (or in an ε-zigzag fashion) until they become equal to u_{i_3}, etc or until $u_+ = 1 = t_+$ is reached. Assume the process stops with direction i_m and minimal u being $\tilde{u}$, i.e. finally $u_{i_k} = \tilde{u}$ for $k \leq m$ and $t_{i_k} = 0$ for $k > m$. From the constraint $1 = u_+ = \sum_{k\leq m} u_{i_k} + \sum_{k>m} u_{i_k} = m\tilde{u} + \sum_{k>m} \frac{n_{i_k}}{n+s}$ we obtain $\tilde{u}(m) = \frac{1}{m}[1 - \sum_{k>m} \frac{n_{i_k}}{n+s}] = [s + \sum_{k\leq m} n_{i_k}]/[m(n+s)]$. One can show that $\tilde{u}(m)$ has one global minimum (no local ones) and that the final m is the one which minimizes $\tilde{u}$, i.e.

$$\tilde{u} = \min_{m\in\{1...d\}} \frac{s + \sum_{k\leq m} n_{i_k}}{m(n+s)}, \text{ where } n_{i_1} \leq n_{i_2} \leq ... \leq n_{i_d}, \quad u_i^{\overline{F}} = \max\{u_i^0, \tilde{u}\}. \tag{4}$$

If there is a unique minimal n_{i_1} with gap $\geq s$ to the second smallest n_{i_2} (which is quite likely for not too small n), then $m = 1$ and the maximum is attained at a corner of Δ (Δ').

Theorem 1 (Exact extrema for concave functions on simplices) *Assume $F : \Delta' \to I\!R$ is a concave function of the form $F(\boldsymbol{u}) = \sum_{i=1}^d f(u_i)$. Then F attains the global maximum $\overline{F}$ at $\boldsymbol{u}^{\overline{F}}$ defined in (4) and the global minimum $\underline{F}$ at $\boldsymbol{u}^{\underline{F}}$ defined in (3).*

Proof. What remains to be shown is that the solutions obtained in the last paragraphs by greedy minimization/maximization of $F(\boldsymbol{u})$ are actually global minima/maxima. For this assume that $\boldsymbol{t}$ is a local minimum of $F(\boldsymbol{u})$. Let $j := \arg\max_i u_i$ (ties broken arbitrarily). Assume that there is a $k \neq j$ with non-zero t_k. Define $\boldsymbol{t}'$ as $t'_i = t_i$ for all $i \neq j, k$, and $t'_j = t_j + \varepsilon$, $t'_k = t_k - \varepsilon$, for some $0 < \varepsilon \leq t_k$. From $u_k \leq u_j$ and the concavity of f we get[5]

$$\begin{aligned} F(\boldsymbol{t}') - F(\boldsymbol{t}) &= [f(u'_j) + f(u'_k)] - [f(u_j) + f(u_k)] \\ &= [f(u_j{+}\sigma\varepsilon) - f(u_j)] - [f(u_k) - f(u_k{-}\sigma\varepsilon)] \;<\; 0 \end{aligned}$$

where $\sigma := \frac{s}{n+s}$. This contradicts the minimality assumption of $\boldsymbol{t}$. Hence, $t_i = 0$ for all i except one (namely j, where it must be 1). (Local) minima are attained in the vertices of Δ. Obviously the global minimum is for $t_i^{\underline{F}} = \delta_{ii^{\underline{F}}}$ with $i^{\underline{F}} := \arg\max_i n_i$. This solution coincides with the greedy solution. Note that the global minimum

[5] Slope $\frac{f(u+\varepsilon)-f(u)}{\varepsilon}$ is a decreasing function in u for any $\varepsilon > 0$, since f is concave.

may not be unique, but since we are only interest in the value of $F(\boldsymbol{u}^{\underline{F}})$ and not its argument this degeneracy is of no further significance.

Similarly for the maximum, assume that $\boldsymbol{t}$ is a (local) maximum of $F(\boldsymbol{u})$. Let $j := \arg\min_i u_i$ (ties broken arbitrarily). Assume that there is a $k \neq j$ with non-zero t_k *and* $u_k > u_j$. Define $\boldsymbol{t}'$ as above with $0 < \varepsilon < \min\{t_k, t_k - t_j\}$. Concavity of f implies

$$F(\boldsymbol{t}') - F(\boldsymbol{t}) = [f(u_j + \sigma\varepsilon) - f(u_j)] - [f(u_k) - f(u_k - \sigma\varepsilon)] > 0,$$

which contradicts the maximality assumption of $\boldsymbol{t}$. Hence $t_i = 0$ if u_i is not minimal ($\tilde{u}$). The previous paragraph constructed the unique solution $\boldsymbol{u}^{\overline{F}}$ satisfying this condition. Since this is the only local maximum it must be the unique global maximum (contrast this to the minimum case). □

Theorem 2 (Exact extrema of expected entropy) *Let $\mathcal{H}(\boldsymbol{\pi}) = -\sum_i \pi_i \log \pi_i$ be the entropy of $\boldsymbol{\pi}$ and the uncertainty of $\boldsymbol{\pi}$ be modeled by the Imprecise Dirichlet Model. The expected entropy $H(\boldsymbol{u}) := E_{\boldsymbol{t}}[\mathcal{H}]$ for given hyperparameter $\boldsymbol{t}$ and sample $\boldsymbol{n}$ is given by*

$$H(\boldsymbol{u}) = \sum_i h(u_i) \quad \textit{with} \quad h(u) = u \cdot [\psi(n+s+1) - \psi((n+s)u+1)] = u \cdot \sum_{k=(n+s)u+1}^{n+s} k^{-1} \qquad (5)$$

where $\psi(x) = d \log \Gamma(x)/dx$ is the logarithmic derivative of the Gamma function and the last expression is valid for integral s and $(n+s)u$. The lower $\underline{H}$ and upper $\overline{H}$ expected entropies are assumed at $\boldsymbol{u}^{\underline{H}}$ and $\boldsymbol{u}^{\overline{H}}$ given in (3) and (4) (with $F \rightsquigarrow H$, see also (1)).

A derivation of the exact expression (5) for the expected entropy can be found in [WW95, Hut02]. The only thing to be shown is that h is concave. This may be done by exploiting special properties of the digamma function ψ (see [AS74]).

There are fast implementations of ψ and its derivatives and exact expressions for integer and half-integer arguments

Example. For $d = 2$, $n_1 = 3$, $n_2 = 6$, $s = 1$ we have $n = 9$, $u_1 = \frac{3+t_1}{10}$, $u_2 = \frac{6+t_2}{10}$, $\boldsymbol{t}^0 = 0$, $\boldsymbol{u}^0 = \binom{.3}{.6}$, see (1). From (3), $i^{\underline{H}} = 2$, $\boldsymbol{t}^{\underline{H}} = \binom{0}{1}$, $\boldsymbol{u}^{\underline{H}} = \binom{.3}{.7}$. From (4), $i_1 = 1$, $i_2 = 2$, $\tilde{u} = \min\{\frac{1+3}{9+1}, \frac{1+3+6}{2\cdot(9+1)}\} = \frac{4}{10}$, $\boldsymbol{u}^{\overline{H}} = \max\{\boldsymbol{u}^0, \tilde{u}\} = \binom{.4}{.6} \Rightarrow \boldsymbol{t}^{\overline{H}} = \binom{1}{0}$ is in corner. From (5), $h(\frac{3}{10}) = \frac{2761}{8400}$, $h(\frac{4}{10}) = \frac{2131}{6300}$, $h(\frac{6}{10}) = \frac{1207}{4200}$, $h(\frac{7}{10}) = \frac{847}{3600}$, hence $\overline{\underline{H}} = [H(\boldsymbol{u}^{\underline{H}}), H(\boldsymbol{u}^{\overline{H}})] = [h(\frac{3}{10}) + h(\frac{7}{10}), h(\frac{4}{10}) + h(\frac{6}{10})] = [0.5639..., 0.6256...]$, so $\overline{H} - \underline{H} = O(\frac{1}{10})$.

4 Approximate Robust Intervals

In this section we derive approximations for $\overline{\underline{F}}$ suitable for arbitrary, twice differentiable functions $F(\boldsymbol{u})$. The derived approximations for $\overline{\underline{F}}$ will be robust in the sense of covering set $\overline{\underline{F}}$ (for any n), and the approximations will be “good” if n is not too small. In the following, we treat $\sigma := \frac{s}{n+s}$ as a (small) expansion parameter. For $\boldsymbol{u},\boldsymbol{u}^* \in \Delta'$ we have

$$u_i - u_i^* \;=\; \sigma\cdot(t_i - t_i^*) \quad \text{and} \quad |u_i - u_i^*| \;=\; \sigma|t_i - t_i^*| \;\leq\; \sigma \quad \text{with} \quad \sigma := \tfrac{s}{n+s}. \tag{6}$$

Hence we may Taylor-expand $F(\boldsymbol{u})$ around $\boldsymbol{u}^*$, which leads to a Taylor series in σ. This shows that F is approximately linear in $\boldsymbol{u}$ and hence in $\boldsymbol{t}$. A linear function on a simplex assumes its extreme values at the vertices of the simplex. This has already been encountered in Section 3. The consideration above is a simple explanation for this fact. This also shows that the robust interval $\overline{\underline{F}}$ is of size $\overline{F} - \underline{F} = O(\sigma)$.[6] Any approximation to $\overline{\underline{F}}$ should hence be at least $O(\sigma^2)$. The expansion of F to $O(\sigma)$ is

$$F(\boldsymbol{u}) \;=\; \overbrace{F(\boldsymbol{u}^*)}^{F_0=O(1)} + \overbrace{\sum_i [\partial_i F(\breve{\boldsymbol{u}})](u_i - u_i^*)}^{F_R=O(\sigma)} \tag{7}$$

where $\partial_i F(\breve{\boldsymbol{u}})$ is the partial derivative $\frac{\partial_i F(\breve{\boldsymbol{u}})}{\partial \breve{u}_i}$ of $F(\breve{\boldsymbol{u}})$ w.r.t. $\breve{u}_i$. For suitable $\breve{\boldsymbol{u}} = \breve{\boldsymbol{u}}(\boldsymbol{u},\boldsymbol{u}^*) \in \Delta'$ this expansion is exact (F_R is the exact remainder). Natural points for expansion are $t_i^* = \frac{1}{d}$ in the center of Δ, or possibly also $t_i^* = \frac{n_i}{n} = u_i^*$. See [Hut03] for such a general expansion. Here, we expand around the improper point $t_i^* := t_i^0 \equiv 0$, which is outside(!) Δ, since this makes expressions particularly simple.[7] (6) is still valid in this case, and F_R is exact for some $\breve{\boldsymbol{u}}$ in

$$\Delta'_e := \{\boldsymbol{u} : u_i \geq u_i^0 \forall i,\, u_+ \leq 1\}, \quad \text{where} \quad u_i^0 = \frac{n_i}{n+s}.$$

Note that we keep the exact condition $\boldsymbol{u} \in \Delta'$. F is usually already defined on Δ'_e or extends from Δ' to Δ'_e without effort in a natural way (analytical continuation). We introduce the notation

$$F \sqsubseteq G \quad :\Leftrightarrow \quad F \leq G \quad \text{and} \quad F = G + O(\sigma^2) \tag{8}$$

stating that G is a “good” upper bound on F. The following bounds hold for arbitrary differentiable functions. In order for the bounds to be “good,” F has to be Lipschitz differentiable in the sense that there exists a constant c such that

$$|\partial_i F(\boldsymbol{u})| \leq c \quad \text{and} \quad |\partial_i F(\boldsymbol{u}) - \partial_i F(\boldsymbol{u}')| \leq c|\boldsymbol{u} - \boldsymbol{u}'|$$

$$\forall \boldsymbol{u},\boldsymbol{u}' \in \Delta'_e \quad \text{and} \quad \forall\, 1 \leq i \leq d. \tag{9}$$

[6] $f(\boldsymbol{n},\boldsymbol{t},s) = O(\sigma^k) \;:\Leftrightarrow\; \exists c \forall \boldsymbol{n} \in I\!N_0^d, \boldsymbol{t} \in \Delta, s > 0 : |f(\boldsymbol{n},\boldsymbol{t},s)| \leq c\sigma^k$, where $\sigma = \frac{s}{n+s}$.

[7] The order of accuracy $O(\sigma^2)$ we will encounter is for all choices of $\boldsymbol{u}^*$ the same. The concrete numerical errors differ of course. The choice $\boldsymbol{t}^* = \boldsymbol{0}$ can lead to $O(d)$ smaller F_R than the natural center point $\boldsymbol{t}^* = \frac{1}{d}$, but is more likely a factor $O(1)$ larger. The exact numerical values depend on the structure of F.

If F depends also on $\boldsymbol{n}$, e.g. via σ or $\boldsymbol{u}^0$, then c shall be independent of them.

The Lipschitz condition is satisfied, for instance, if the curvature $\partial^2 F$ is uniformly bounded. This is satisfied for the expected entropy H (see (5)), but violated for the approximation $E_{\boldsymbol{t}}[\mathcal{H}] \approx \mathcal{H}(\boldsymbol{u})$ if $n_i = 0$ for some i.

Theorem 3 (Approximate robust intervals) *Assume $F : \Delta'_e \to I\!R$ is a Lipschitz differentiable function (9). Let $[\underline{F},\overline{F}]$ be the global [minimum,maximum] of F restricted to Δ'. Then*

$$F(\boldsymbol{u}^1) \;\sqsubseteq\; \overline{F} \;\sqsubseteq\; F_0 + F_R^{ub} \;\; \textit{where} \;\; F_R^{ub} = \max_i F_{iR}^{ub} \;\; \textit{and} \;\; F_{iR}^{ub} \;=\; \sigma \max_{\boldsymbol{u}\in\Delta'_e}[\partial_i F(\boldsymbol{u})]$$

$$F_0 + F_R^{lb} \;\sqsubseteq\; \underline{F} \;\sqsubseteq\; F(\boldsymbol{u}^2) \;\; \textit{where} \;\; F_R^{lb} = \min_i F_{iR}^{lb} \;\; \textit{and} \;\; F_{iR}^{lb} \;=\; \sigma \min_{\boldsymbol{u}\in\Delta'_e}[\partial_i F(\boldsymbol{u})]$$

$F_0 = F(\boldsymbol{u}^0)$, and $u_i^1 = \delta_{ii^1}$ with $i^1 = \arg\max_i F_{iR}^{ub}$, and $u_i^2 = \delta_{ii^2}$ with $i^2 = \arg\min_i F_{iR}^{lb}$, and $\sqsubseteq$ defined in (8) means $\leq$ and $=$ *$+O(\sigma^2)$, where $\sigma = 1 - u_+^0$.*

For conservative estimates, the lower bound on $\underline{F}$ and the upper bound on $\overline{F}$ are the interesting ones.

Proof. We start by giving an $O(\sigma^2)$ bound on $\overline{F}_R = \max_{\boldsymbol{u}\in\Delta'} F_R(\boldsymbol{u})$. We first insert (6) with $\boldsymbol{t}^* = \boldsymbol{t}^0 \equiv \boldsymbol{0}$ into (7) and treat $\breve{\boldsymbol{u}}$ and $\boldsymbol{t}$ as separate variables:

$$F_R(\breve{\boldsymbol{u}},\boldsymbol{t}) \;=\; \sigma \sum_i [\partial_i F(\breve{\boldsymbol{u}})] \cdot t_i \;\sqsubseteq\; \max_{\breve{\boldsymbol{u}}\in\Delta'_e} \left\{ \sigma \sum_i [\partial_i F(\breve{\boldsymbol{u}})] \cdot t_i \right\} \;\sqsubseteq\; \sum_i F_{iR}^{ub} \cdot t_i$$

$$\text{with} \quad F_{iR}^{ub} \;:=\; \sigma \max_{\breve{\boldsymbol{u}}\in\Delta'_e} [\partial_i F(\breve{\boldsymbol{u}})] \tag{10}$$

The first inequality is obvious, the second follows from the convexity of max. From assumption (9) we get $\partial_i F(\boldsymbol{u}) - \partial_i F(\boldsymbol{u}') = O(\sigma)$ for all $\boldsymbol{u},\boldsymbol{u}' \in \Delta'_e$, since Δ'_e has diameter $O(\sigma)$. Due to one additional σ in (10) the expressions in (10) change only by $O(\sigma^2)$ when introducing or dropping $\max_{\breve{\boldsymbol{u}}}$ anywhere. This shows that the inequalities are tight within $O(\sigma^2)$ and justifies $\sqsubseteq$. We now upper bound $F_R(\boldsymbol{u})$:

$$\overline{F}_R = \max_{\boldsymbol{u}\in\Delta'} F_R(\boldsymbol{u}) \sqsubseteq \max_{\boldsymbol{t}\in\Delta} \max_{\breve{\boldsymbol{u}}\in\Delta'_e} F_R(\breve{\boldsymbol{u}},\boldsymbol{t}) \sqsubseteq \max_{\boldsymbol{t}\in\Delta} \sum_i F_{iR}^{ub} \cdot t_i = \max_i F_{iR}^{ub} =: F_R^{ub} \tag{11}$$

A linear function on Δ is maximized by setting the t_i component with largest coefficient to 1. This shows the last equality. The maximization over $\breve{\boldsymbol{u}}$ in (10) can often be performed analytically, leaving an easy $O(d)$ time task for maximizing over i.

We have derived an upper bound F_R^{ub} on $\overline{F}_R$. Let us define the corner $t_i = \delta_{ii^1}$ of Δ with $i^1 := \arg\max_i F_{iR}^{ub}$. Since $\overline{F}_R \geq F_R(\boldsymbol{u})$ for all $\boldsymbol{u}$, $F_R(\boldsymbol{u}^1)$ in particular is a lower bound on $\overline{F}_R$. A similar line of reasoning as above shows that that $F_R(\boldsymbol{u}^1) = \overline{F}_R + O(\sigma^2)$. Using $\overline{F + const.} = \overline{F} + const.$ we get $O(\sigma^2)$ lower and upper bounds on $\overline{F}$, i.e. $F(\boldsymbol{u}^1) \sqsubseteq \overline{F} \sqsubseteq F_0 + F_R^{ub}$. $\underline{F}$ is bound similarly with all max's replaced by min's and inequalities reversed. Together this proves the Theorem 3.

□

5 Error Propagation

Approximation of $\underline{F}$ (special cases). For the special case $F(\boldsymbol{u}) = \sum_i f(u_i)$ we have $\partial_i F(\boldsymbol{u}) = f'(u_i)$. For concave f like in case of the entropy we get particularly simple bounds

$$F_{iR}^{ub} = \sigma \max_{\boldsymbol{u}\in\Delta'_e} f'(u_i) = \sigma f'(u_i^0), \qquad F_R^{ub} = \sigma \max_i f'(u_i^0) = \sigma f'(\tfrac{\min_i n_i}{n+s}),$$

$$F_{iR}^{lb} = \sigma \min_{\boldsymbol{u}\in\Delta'_e} f'(u_i) = \sigma f'(u_i^0+\sigma), \quad F_R^{lb} = \sigma \min_i f'(u_i^0+\sigma) = \sigma f'(\tfrac{\max_i n_i+s}{n+s}),$$

where we have used $\max_{\boldsymbol{u}\in\Delta'_e} f'(u_i) = \max_{u_i\in[u_i^0,u_i^0+\sigma]} f'(u_i) = f'(u_i^0)$, and similarly for min. Analogous results hold for convex functions. In case the maximum cannot be found exactly one is allowed to further increase Δ'_e as long as its diameter remains $O(\sigma)$. Often an increase to $\Box' := \{\boldsymbol{u} : u_i^0 \leq u_i \leq u_i^0+\sigma\} \supset \Delta'_e \supset \Delta'$ makes the problem easy. Note that if we were to perform these kind of crude enlargements on $\max_{\boldsymbol{u}} F(\boldsymbol{u})$ directly we would loose the bounds by $O(\sigma)$.

Example (continued). $\sigma = \frac{1}{10}$, $h'(\frac{3}{10}) = \frac{13051}{2520} - \frac{1}{2}\Pi^2$, $h'(\frac{7}{10}) = \frac{91717}{8400} - \frac{7}{6}\Pi^2$, $H_0 = H(\boldsymbol{u}^0) = h(\frac{3}{10}) + h(\frac{6}{10})$, $H_R^{ub} = \frac{1}{10}h'(\frac{3}{10})$, $H_R^{lb} = \frac{1}{10}h'(\frac{7}{10}) \Rightarrow [H_0+H_R^{lb}, H_0+H_R^{ub}] = [0.5564..., 0.6404...]$, hence $H_0 + H_R^{ub} - \overline{H} = 0.0148 = O(\frac{1}{10^2})$, $\underline{H} - H_0 - H_R^{lb} = 0.0074... = O(\frac{1}{10^2})$.

Error propagation. Assume we found bounds for estimators $G(\boldsymbol{u})$ and $H(\boldsymbol{u})$ and we want now to bound the sum $F(\boldsymbol{u}) := G(\boldsymbol{u}) + H(\boldsymbol{u})$. In the direct approach $\overline{F} \leq \overline{G} + \overline{H}$ we may lose $O(\sigma)$. A simple example is $G(\boldsymbol{u}) = u_i$ and $H(\boldsymbol{u}) = -u_i$ for which $F(\boldsymbol{u}) = 0$, hence $0 = \overline{F} \leq \overline{G} + \overline{H} = u_i^0 + \sigma - u_i^0 = \sigma$, i.e. $\overline{F} \not\sqsubseteq \overline{G} + \overline{H}$. We can exploit the techniques of the previous section to obtain $O(\sigma^2)$ approximations.

$$F_{iR}^{ub} \;=\; \sigma \max_{\boldsymbol{u}\in\Delta'_e} \partial_i F(\boldsymbol{u}) \;\sqsubseteq\; \sigma \max_{\boldsymbol{u}\in\Delta'_e} \partial_i G(\boldsymbol{u}) + \sigma \max_{\boldsymbol{u}\in\Delta'_e} \partial_i H(\boldsymbol{u}) \;=\; G_{iR}^{ub} + H_{iR}^{ub}$$

Theorem 4 (Error propagation: Sum) *Let $G(\boldsymbol{u})$ and $H(\boldsymbol{u})$ be Lipschitz differentiable and $F(\boldsymbol{u}) = \alpha G(\boldsymbol{u}) + \beta H(\boldsymbol{u})$, $\alpha, \beta \geq 0$, then $\overline{F} \sqsubseteq F_0 + F_R^{ub}$ and $\underline{F} \sqsupseteq F_0 + F_R^{lb}$, where $F_0 = \alpha G_0 + \beta H_0$, and $F_{iR}^{ub} \sqsubseteq \alpha G_{iR}^{ub} + \beta H_{iR}^{ub}$, and $F_{iR}^{lb} \sqsupseteq \alpha G_{iR}^{lb} + \beta H_{iR}^{lb}$.*

It is important to notice that $F_R^{ub} \not\sqsubseteq G_R^{ub} + H_R^{ub}$ (use previous example), i.e. $\max_i[G_{iR}^{ub} + H_{iR}^{ub}] \not\sqsubseteq \max_i G_{iR}^{ub} + \max_i H_{iR}^{ub}$. $\max_i$ can not be pulled in and it is important to propagate F_{iR}^{ub}, rather than F_R^{ub}.

Every function F with bounded curvature can be written as a sum of a concave function G and a convex function H. For convex and concave functions, determining bounds is particularly easy, as we have seen. Often F decomposes naturally into convex and concave parts as is the case for the mutual information, addressed later. Bounds can also be derived for products.

Theorem 5 (Error propagation: Product) *Let $G,H:\Delta'_e\to[0,\infty)$ be non-negative Lipschitz differentiable functions (9) with non-negative derivatives $\partial_i G,\partial_i H\geq 0\;\forall i$ and $F(\boldsymbol{u})=G(\boldsymbol{u})\cdot H(\boldsymbol{u})$, then $\overline{F}\sqsubseteq F_0+F_R^{ub}$, where $F_0=G_0\cdot H_0$, and $F_{iR}^{ub}\sqsubseteq G_{iR}^{ub}(H_0+H_R^{ub})+(G_0+G_R^{ub})H_{iR}^{ub}$, and similarly for $\underline{F}$.*

Proof. We have

$$F_{iR}^{ub}\;=\;\sigma\max\partial_i F\;=\;\sigma\max\partial_i(G{\cdot}H)\;=\;\sigma\max[(\partial_i G)H+G(\partial_i H)]\;\sqsubseteq$$

$$\sigma(\max\partial_i G)(\max H)+\sigma(\max G)(\max\partial_i H)\sqsubseteq G_{iR}^{ub}(H_0{+}H_R^{ub})+(G_0{+}G_R^{ub})H_{iR}^{ub}$$

where all functions depend on $\boldsymbol{u}$ and all max are over $\boldsymbol{u}\in\Delta'_e$. There is one subtlety in the last inequality: $\max G\neq\overline{G}\sqsubseteq G_0+G_R^{ub}$. The reason for the $\neq$ being that the maximization is taken over Δ'_e, not over Δ' as in the definition of $\overline{G}$. The correct line of reasoning is as follows:

$$\max_{\boldsymbol{u}\in\Delta'_e}G_R(\boldsymbol{u})\sqsubseteq\max_{\boldsymbol{t}\in\Delta_e}\sum_i G_{iR}^{ub}\cdot t_i=\max\{0,\max_i G_{iR}^{ub}\}=G_R^{ub}\;\Rightarrow\;\max G\sqsubseteq G_0+G_R^{ub}$$

The first inequality can be proven in the same way as (11). In the first equality we set the $t_i=1$ with maximal G_{iR}^{ub} *if* it is positive. If all G_{iR}^{ub} are negative we set $\boldsymbol{t}\equiv\boldsymbol{0}$. We assumed $G\geq 0$ and $\partial_i G\geq 0$, which implies $G_R\geq 0$. So, since $G_R\geq 0$ anyway, this subtlety is ineffective. Similarly for $\max H_R$. □

It is possible to remove the rather strong non-negativity assumptions. Propagation of errors for other combinations like ratios $F=G/H$ may also be obtained.

6 Robust Intervals for Mutual Information

Mutual Information. We illustrate the application of the previous results on the Mutual Information between two random variables $\imath\in\{1,...,d_1\}$ and $\jmath\in\{1,...,d_2\}$. Consider an i.i.d. random process with outcome $(i,j)\in\{1,...,d_1\}\times\{1,...,d_2\}$ having joint probability π_{ij}, where $\boldsymbol{\pi}\in\Delta:=\{\boldsymbol{x}\in\mathbb{R}^{d_1\times d_2}:x_{ij}\geq 0\forall ij,\;x_{++}=1\}$. An important measure of the stochastic dependence of $\imath$ and $\jmath$ is the mutual information

$$I(\boldsymbol{\pi})=\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\pi_{ij}\log\frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}=\sum_{ij}\pi_{ij}\log\pi_{ij}-\sum_i\pi_{i+}\log\pi_{i+}-\sum_j\pi_{+j}\log\pi_{+j}\tag{12}$$
$$=\;\mathcal{H}(\boldsymbol{\pi}_{\imath+})+\mathcal{H}(\boldsymbol{\pi}_{+\jmath})-\mathcal{H}(\boldsymbol{\pi}_{\imath\jmath})$$

$\pi_{i+}=\sum_j\pi_{ij}$ and $\pi_{+j}=\sum_i\pi_{ij}$ are row and column marginal chances. Again, we assume a Dirichlet prior over $\boldsymbol{\pi}_{\imath\jmath}$, which leads to a Dirichlet posterior $p(\boldsymbol{\pi}_{\imath\jmath}|\boldsymbol{n})\propto\prod_{ij}\pi_{ij}^{n_{ij}+st_{ij}-1}$ with $\boldsymbol{t}\in\Delta$. The expected value of π_{ij} is

$$E_{\boldsymbol{t}}[\pi_{ij}]=\frac{n_{ij}+st_{ij}}{n+s}=:u_{ij}$$

The marginals $\boldsymbol{\pi}_{i+}$ and $\boldsymbol{\pi}_{+j}$ are also Dirichlet with expectation u_{i+} and u_{+j}. The expected mutual information $I(\boldsymbol{u}) := E_{\boldsymbol{t}}[I]$ can, hence, be expressed in terms of the expectations of three entropies $H(\boldsymbol{u}) := E_{\boldsymbol{t}}[\mathcal{H}]$ (see (5))

$$I(\boldsymbol{u}) = H(\boldsymbol{u}_{\imath+}) + H(\boldsymbol{u}_{+\jmath}) - H(\boldsymbol{u}_{\imath\jmath}) \; = \; H_{row} + H_{col} - H_{joint}$$
$$= \; \sum_i h(u_{i+}) + \sum_j h(u_{+j}) - \sum_{ij} h(u_{ij})$$

where here and in the following we index quantities with *joint*, *row*, and *col* to denote to which distribution the quantity refers.

Crude bounds for $I(\boldsymbol{u})$. Estimates for the robust IDM interval $[\min_{\boldsymbol{t}\in\Delta} E_{\boldsymbol{t}}[I]$, $\max_{\boldsymbol{t}\in\Delta} E_{\boldsymbol{t}}[I]]$ can be obtained by [minimizing,maximizing] $I(\boldsymbol{u})$. A crude upper bound can be obtained as

$$\overline{I} \; := \; \max_{\boldsymbol{t}\in\Delta} I(\boldsymbol{u}) \; = \; \max[H_{row} + H_{col} - H_{joint}] \; \leq$$
$$\max H_{row} + \max H_{col} - \min H_{joint} \; = \; \overline{H}_{row} + \overline{H}_{col} - \underline{H}_{joint},$$

where exact solutions to $\overline{H}_{row}$, $\underline{H}_{row}$ and $\underline{H}_{joint}$ are available from Section 3. Similarly $\underline{I} \geq \underline{H}_{row} + \underline{H}_{col} - \overline{H}_{joint}$. The problem with these bounds is that, although good in some cases, they can become arbitrarily crude. The following $O(\sigma^2)$ bound can be derived by exploiting the error sum propagation Theorem 4.

Theorem 6 (Bound on lower and upper Mutual Information) *The following bounds on the expected mutual information $I(\boldsymbol{u}) = E_{\boldsymbol{t}}[I]$ are valid:*

$$\begin{aligned}
&I(\boldsymbol{u}^1) \sqsubseteq \overline{I} \sqsubseteq I_0 + I_R^{ub} \quad \textit{and} \quad I_0 + I_R^{lb} \sqsubseteq \underline{I} \sqsubseteq I(\boldsymbol{u}^2), \quad \textit{where} \\
&I_0 = I(\boldsymbol{u}^0) = H_{0row} + H_{0col} - H_{0joint} = h(u_{i+}^0) + h(u_{+j}^0) - h(u_{ij}^0), \\
&I_{ijR}^{ub} \sqsubseteq H_{iRrow}^{ub} + H_{jRcol}^{ub} - H_{ijRjoint}^{lb} = h'(u_{i+}^0) + h'(u_{+j}^0) - h'(u_{ij}^0 + \sigma), \\
&I_{ijR}^{lb} \sqsupseteq H_{iRrow}^{lb} + H_{jRcol}^{lb} - H_{ijRjoint}^{ub} = h'(u_{i+}^0 + \sigma) + h'(u_{+j}^0 + \sigma) - h'(u_{ij}^0),
\end{aligned}$$

with h defined in (5), and $t_{ij}^0 = 0$, and $t_{ij}^1 = \delta_{(ij)(ij)^1}$ with $(ij)^1 = \arg\max_{ij} I_{ijR}^{ub}$, and $t_{ij}^2 = \delta_{(ij)(ij)^2}$ with $(ij)^2 = \arg\min_{ij} I_{ijR}^{lb}$.

7 IDM for Product Spaces

Product spaces $\Omega = \Omega_1 \times ... \times \Omega_m$ with $\Omega_k = \{1, ...d_k\}$ occur frequently in practical problems, e.g. in the mutual information ($m = 2$), in robust trees ($m = 3$), or in Bayesian nets in general (m large). Without loss of generality we only discuss the $m = 2$ case in the following. Ignoring the underlying structure in Ω, a Dirichlet prior in case of unknown chances $\pi_{\imath\jmath}$ and an IDM as used in Section 6 with

$$\boldsymbol{t} \in \Delta := \{\boldsymbol{t} \in \mathbb{R}^{d_1 \times d_2} \equiv \mathbb{R}^{d_1} \otimes \mathbb{R}^{d_2} \, : \, t_{ij} \geq 0 \forall ij, \, t_{++} = 1\} \tag{13}$$

seems natural. On the other hand, if we take into account the structure of Ω and go back to the original motivation of IDM this choice is far less obvious. Recall that one of the major motivations of IDM was its reparametrization invariance in the sense that inferences are not affected when grouping or splitting events in Ω. For unstructured spaces like Ω_k this is a reasonable principle. For illustration, let us consider objects of various *shape* and *color*, i.e. $\Omega=\Omega_1\times\Omega_2$, $\Omega_1=\{ball,pen,die,...\}$, $\Omega_2=\{yellow,red,green,...\}$ in generalization to Walleys bag of marbles example. Assume we want to detect a potential dependency between *shape* and *color* by means of their mutual information I. If we have no prior idea on the possible kind of colors, a model which is independent of the choice of Ω_2 is welcome. Grouping red and green, for instance, corresponds to $(x_{i1},x_{i2},x_{i3},x_{i4},...)\rightsquigarrow(x_{i1},x_{i2}+x_{i3},x_{i4},...)$ *for all shapes i*, where $\boldsymbol{x}\in\{\boldsymbol{n},\boldsymbol{\pi},\boldsymbol{t},\boldsymbol{u}\}$. Similarly for the different shapes, for instance we could group all round or all angular objects. The "smallest IDM" which respects this invariance is the one which considers all

$$\boldsymbol{t}\in\mathbb{\Delta}:=\Delta_{d_1}\otimes\Delta_{d_2}\;\subsetneq\;\Delta. \tag{14}$$

The tensor or outer product $\otimes$ is defined as $(\boldsymbol{v}\otimes\boldsymbol{w})_{ij}:=v_iw_j$ and $V\otimes W:=\{\boldsymbol{v}\otimes\boldsymbol{w}:\boldsymbol{v}\in V,\boldsymbol{w}\in W\}$. It is a bilinear (not linear!) mapping. This "small tensor" IDM is invariant under arbitrary grouping of columns and rows of the chance matrix $(\boldsymbol{\pi}_{ij})_{1\leq i\leq d_1,1\leq j\leq d_2}$. In contrast to the larger Δ IDM model it is not invariant under arbitrary grouping of matrix cells, but there is anyway little motivation for the necessity of such a general invariance. General non-column/row cross groupings would destroy the product structure of Ω and with that the mere concepts of shape and color, and their correlation. For $m>2$ as in Bayes-nets cross groupings look even less natural. Whether the $\mathbb{\Delta}$ or the larger simplex Δ is the more appropriate IDM model depends on whether one regards the structure $\Omega_1\times\Omega_2$ of Ω as a natural prior knowledge or as an arbitrary a posteriori choice. The smaller IDM has the potential advantage of leading to more precise predictions (smaller robust sets).

Let us consider an estimator $F:\Delta\to I\!R$ and its restriction $F_\otimes:\mathbb{\Delta}\to I\!R$. Robust intervals $[\underline{F},\overline{F}]$ for Δ are generally wider than robust intervals $[\underline{F}_\otimes,\overline{F}_\otimes]$ for $\mathbb{\Delta}$. Fortunately not much. Although $\mathbb{\Delta}$ is a *lower-dimensional* subspace of Δ, it contains all vertices of Δ. This is possible since $\mathbb{\Delta}$ is a *nonlinear* subspace. The set of "vertices" in both cases is $\{\boldsymbol{t}:t_{ij}=\delta_{ii_0}\delta_{jj_0},\;i_0\in\Omega_1,\;j_0\in\Omega_2\}$. Hence, *if* the robust interval boundaries $\overline{\underline{F}}$ are assumed in the vertices of Δ *then* the interval for the $\mathbb{\Delta}$ IDM model is the same ($\overline{\underline{F}}=\overline{\underline{F}}_\otimes$). Since the condition is "approximately" true, the conclusion is "approximately" true. More precisely:

Theorem 7 (IDM bounds for product spaces) *The $O(\sigma^2)$ bounds of Theorem 3 on the robust interval $\overline{\underline{F}}$ in the full IDM model Δ (13), remain valid for $\overline{\underline{F}}_\otimes$ in the product IDM model $\mathbb{\Delta}$ (14).*

Proof.

$$F(\boldsymbol{u}^1) \leq \overline{F}_{\otimes} \leq \overline{F} \leq F_0 + F_R^{ub} = F(\boldsymbol{u}^1) + O(\sigma^2),$$

where $\overline{F}_{\otimes} := \max_{\boldsymbol{t}\in\mathbb{\Delta}} F(\boldsymbol{u})$ and $\boldsymbol{u}^1$ was the “F_R maximizing” vertex as defined in Theorem 6 ($F(\boldsymbol{u}^1) \sqsubseteq \overline{F}$). The first inequality follows from the fact that all Δ vertices also belong to $\mathbb{\Delta}$, i.e. $\boldsymbol{t}^1 \in \mathbb{\Delta}$. The second inequality follows from $\mathbb{\Delta} \subset \Delta$. The remaining (in)equalities follow from Theorem 3. This shows that $|\overline{F}_{\otimes} - \overline{F}| = O(\sigma^2)$, hence $F_0 + F_R^{ub}$ is also an $O(\sigma^2)$ upper bound to $\overline{F}_{\otimes}$. This implies that to the approximation accuracy we can achieve, the choice between Δ and $\mathbb{\Delta}$ is irrelevant. □

8 Robust Credible Intervals

Bayesian credible sets/intervals. For a probability distribution $p : I\!R^d \to [0,1]$, an α-credible region is a measurable set A for which $p(A) := \int p(x)\chi_A(x)d^dx \geq \alpha$, where $\chi_A(x) = 1$ if $x \in A$ and 0 otherwise, i.e. $x \in A$ with probability at least α. For given α, there are many choices for A. Often one is interested in “small” sets, where the size of A may be measured by its volume $\mathrm{Vol}(A) := \int \chi_A(x)d^dx$. Let us define a/the smallest α-credible set

$$A^{min} := \operatorname*{arg\,min}_{A:p(A)\geq\alpha} \mathrm{Vol}(A)$$

with ties broken arbitrarily. For unimodal p, A^{min} can be chosen as a connected set. For $d = 1$ this means that $A^{min} = [a,b]$ with $\int_a^b p(x)dx = \alpha$ is a minimal length α-credible interval. If, additionally p is symmetric around $E[x]$, then $A^{min} = [E[x] - a, E[x] + a]$ is also symmetric around $E[x]$.

Robust credible sets. If we have a set of probability distributions $\{p_t(x),\, t \in T\}$, we can choose for each t an α-credible set A_t with $p_t(A_t) \geq \alpha$, a minimal one being $A_t^{min} := \arg\min_{A:p_t(A)\geq\alpha} \mathrm{Vol}(A)$. A robust α-credible set is a set A which contains x with p_t-probability at least α for *all* t. A minimal size robust α-credible set is

$$A^{min} := \operatorname*{arg\,min}_{A=\bigcup_t A_t : p_t(A_t)\geq\alpha\forall t\in T} \mathrm{Vol}(A) \tag{15}$$

It is not easy to deal with this expression, since A^{min} is *not* a function of $\{A_t^{min} : t \in T\}$, and especially does not coincide with $\bigcup_t A_t^{min}$ as one might expect.

Robust credible intervals. This can most easily be seen for univariate symmetric unimodal distributions, where t is a translation, e.g. $p_t(x) = \mathrm{Normal}(E_t[x] = t, \sigma = 1)$ with 95% credible intervals $A_t^{min} = [t-2, t+2]$. For, e.g. $T = [-1,1]$ we get $\bigcup_t A_t^{min} = [-3,3]$. The credible intervals *move* with t. One can get a smaller union if we take the intervals $A_t^{sym} = [-a_t, a_t]$ symmetric around 0. Since A_t^{sym} is a non-central interval w.r.t. p_t for $t \neq 0$, we have $a_t > 2$, i.e. A_t^{sym} is larger than A_t^{min}, but one can show that the increase of a_t is smaller than the shift of A_t^{min} by t, hence we save something in the union. The optimal choice is neither A_t^{sym} nor A_t^{min}, but something in-between. In the extended version [Hut03] this is illustrated for the triangular distribution $p_t(x) = \max\{0\,,\,1-|x-t|\}$ with $t \in T := [-\gamma,\gamma]$, where closed form solutions can be given.

An interesting open question is under which general conditions we can expect $A^{min} \subseteq \bigcup_t A_t^{min}$. In any case, $\bigcup_t A_t$ can be used as a conservative estimate for a robust credible set, since $p_t(\bigcup_{t'} A_{t'}) \geq p_t(A_t) \geq \alpha$ for all t.

A special (but important) case which falls outside the above framework are one-sided credible intervals, where only A_t of the form $[a,\infty)$ are considered. In this case $A^{min} = \bigcup_t A_t^{min}$, i.e. $A^{min} = [a_{min},\infty)$ with $a_{min} = \max\{a : p_t([a,\infty]) \geq \alpha \forall t\}$.

Approximations. For complex distributions like for the mutual information we have to approximate (15) somehow. We use the following notation for shortest α-credible *intervals* w.r.t. a univariate distribution $p_t(x)$:

$$\underset{\sim}{\widetilde{x}}_t \;\equiv\; [\underset{\sim}{x}_t, \widetilde{x}_t] \;\equiv\; [E_t[x] - \Delta\underset{\sim}{x}_t\,,\, E_t[x] + \Delta\widetilde{x}_t] \;:=\; \mathop{\arg\min}_{[a,b]:p_t([a,b])\geq\alpha} (b-a),$$

where $\Delta\widetilde{x}_t := \widetilde{x}_t - E_t[x]$ ($\Delta\underset{\sim}{x}_t := E_t[x] - \underset{\sim}{x}_t$) is the distance from the right boundary $\widetilde{x}_t$ (left boundary $\underset{\sim}{x}_t$) of the shortest α-credible interval $\underset{\sim}{\widetilde{x}}_t$ to the mean $E_t[x]$ of distribution p_t. We can use $\underset{\simeq}{\overline{\widetilde{x}}} \equiv [\underset{\simeq}{x}, \overline{\widetilde{x}}] := \bigcup_t \underset{\sim}{\widetilde{x}}_t$ as a (conservative, but not shortest) robust credible interval, since $p_t(\underset{\simeq}{\overline{\widetilde{x}}}) \geq p_t(\underset{\sim}{\widetilde{x}}_t) \geq \alpha$ for all t. We can upper bound $\overline{\widetilde{x}}$ (and similarly lower bound $\underset{\simeq}{x}$) by

$$\overline{\widetilde{x}} \;=\; \max_t (E_t[x] + \Delta\widetilde{x}_t) \;\leq\; \max_t E_t[x] + \max_t \Delta\widetilde{x}_t \;=\; \overline{E[x]} + \overline{\Delta\widetilde{x}}. \qquad (16)$$

We have already intensively discussed how to compute upper and lower quantities, particularly for the upper mean $\overline{E[x]}$ for $x \in \{\mathcal{F}, \mathcal{H}, I, ...\}$, but the linearization technique introduced in Section 4 is general enough to deal with all in t differentiable quantities, including $\Delta\widetilde{x}_t$. For example for Gaussian p_t with variances σ_t we have $\Delta\widetilde{x}_t = \kappa\sigma_t$ with κ given by $\alpha = \text{erf}(\kappa/\sqrt{2})$, where erf is the error function (e.g. $\kappa = 2$ for $\alpha \approx 95\%$). We only need to estimate $\max_t \sigma_t$.

For non-Gaussian distributions, exact expression for $\Delta\widetilde{x}_t$ are often hard or impossible to obtain and to deal with. Non-Gaussian distributions depending on some sample size n are usually close to Gaussian for large n due to the central limit theorem. One may simply use $\kappa\sigma_t$ in place of $\Delta\widetilde{x}_t$ also in this case, keeping in mind that this could be a non-conservative approximation. More systematically, simple (and for large n good) upper bounds on $\Delta\widetilde{x}_t$ can often be obtained and should preferably be used.

Further, we have seen that the variation of sample depending differentiable functions (like $E_t[x] = E_t[x|\boldsymbol{n}]$) w.r.t. $t \in \Delta$ are of order $\frac{s}{n+s}$. Since in such cases the standard deviation $\sigma_t \sim n^{-1/2} \sim \Delta\widetilde{x}_t$ is itself suppressed, the variation of $\Delta\widetilde{x}_t$

with t is of order $n^{-3/2}$. If we regard this as negligibly small, we may simply fix some $t^* \in \Delta$:

$$\max_t \Delta\widetilde{x}_t = \kappa\sigma_{t^*} + O(n^{-3/2})$$

Since $\Delta\widetilde{x}_t$ is "nearly" constant, this also shows that we lose at most $O(n^{-3/2})$ precision in the bound (16) (equality holds for $\Delta\widetilde{x}_t$ independent of t). Expressions for the variance of I, for instance, have been derived in [WW95, Hut02].

9 Conclusions

This is the first work, providing a systematic approach for deriving closed form expressions for interval estimates in the Imprecise Dirichlet Model (IDM). We concentrated on exact and conservative *robust* interval ([lower,upper]) estimates for concave functions $F = \sum_i f_i$ on simplices, like the entropy. The conservative estimates widened the intervals by $O(n^{-2})$, where n is the sample size. Here is a dilemma, of course: For large n the approximations are good, whereas for small n the bounds are more interesting, so the approximations will be most useful for intermediate n. More precise expressions for small n would be highly interesting. We have also indicated how to propagate robust estimates from simple functions to composite functions, like the mutual information. We argued that a reduced IDM on product spaces, like Bayesian nets, is more natural and should be preferred in order to improve predictions. Although improvement is formally only $O(n^{-2})$, the difference may be significant in Bayes nets or for very small n. Finally, the basics of how to combine robust with credible intervals have been laid out. Under certain conditions $O(n^{-3/2})$ approximations can be derived, but the presented approximations are not conservative. All in all this work has shown that IDM has not only interesting theoretical properties, but that explicit (exact/conservative/approximate) expressions for robust (credible) intervals for various quantities can be derived. The computational complexity of the derived bounds on $F = \sum_i f_i$ is very small, typically one or two evaluations of F or related functions, like its derivative. First applications of these (or more precisely, very similar) results, especially the mutual information, to robust inference of trees look promising [ZH03].

Acknowledgements. I want to thank Peter Walley for introducing IDM to me and Marco Zaffalon for encouraging me to investigate this topic. This work was supported by SNF grant 2000-61847.00 to Jürgen Schmidhuber.

References

[AS74] M. Abramowitz and I. A. Stegun, editors. *Handbook of mathematical functions*. Dover publications, inc., 1974.

[Ber01] J.-M. Bernard. Non-parametric inference about an unknown mean using the Imprecise Dirichlet Model. In G. de Cooman, T. Fine, and T. Seidenfeld, editors, *Proceedings of the 2nd International Symposium on Imprecise Probabilities and Their Application (ISIPTA-2001)*, pages 40–50, The Netherlands, 2001. Shaker Publishing.

[GCSR95] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman, 1995.

[Hal48] J. B. S. Haldane. The precision of observed values of small frequencies. *Biometrika*, 35:297–300, 1948.

[Hut02] M. Hutter. Distribution of mutual information. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 399–406, Cambridge, MA, 2002. MIT Press.

[Hut03] M. Hutter. Robust estimators under the Imprecise Dirichlet Model (extended version). Technical report, IDSIA, Manno(Lugano), Switzerland, 2003. http://www.idsia.ch/∼marcus/ai/idmx.ps.

[Jef46] H. Jeffreys. An invariant form for the prior probability in estimation problems. In *Proc. Royal Soc. London (A)*, volume 186, pages 453–461, 1946.

[Per47] W. Perks. Some observations on inverse probability. *J. Inst. Actuar.*, 73:285–312, 1947.

[Wal96] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society B*, 58(1):3–57, 1996.

[WW95] D. H. Wolpert and D. R. Wolf. Estimating functions of distributions from a finite set of samples. *Physical Review E*, 52(6):6841–6854, 1995.

[Zaf01] M. Zaffalon. Robust discovery of tree-dependency structures. In G. de Cooman, T. Fine, and T. Seidenfeld, editors, *Proceedings of the 2nd International Symposium on Imprecise Probabilities and Their Application (ISIPTA-2001)*, pages 394–403, The Netherlands, 2001. Shaker Publishing.

[ZH03] M. Zaffalon and M. Hutter. Robust inference of trees. Technical Report IDSIA-11-03, IDSIA, Manno (Lugano), CH, 2003.

Marcus Hutter is with the AI research institute IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland. E-mail: marcus@idsia.ch, HP: http://www.idsia.ch/∼marcus/idsia

How to Deal with Partially Analyzed Acts? A Proposal

JEAN-YVES JAFFRAY
Université Paris 6, France

MEGLENA JELEVA
Université de Nantes et EUREQua, France

Abstract

In some situations, a decision is best represented by an incompletely analyzed act: conditionally to a certain event, the consequences of the decision on sub-events are perfectly known and uncertainty becomes expressable through probabilities, whereas the plausibility of this event itself remains vague and the decision outcome on the complementary event is imprecisely known. In this framework, we study an axiomatic decision model and prove a representation theorem. Decision criteria must aggregate partial evaluations consisting in: i) the conditional expected utility associated with the analyzed part of the decision and ii) the best and worst outcomes of its non-analyzed part.

Keywords

decision making under uncertainty, partially analyzed decision

1 Introduction

Consider the famous oil wildcatter problem of decision analysis textbooks. Its description only involves geophysical data and results of seismic tests, which makes it quite convincingly expressible in a Savagean setting where decisions are acts and events are endowed with subjective probabilities. However, it may well be that the relevance of that analysis is only contingent on local political stability. A complete description of the problem would require introducing this factor explicitly. The likelihood of political events being generally difficult to assess and their impact on the wildcatter profits difficult to evaluate, the standard Savagean approach reveals itself unsuitable for taking this aspect into account.

As another example, consider the question of the use of genetically modified organisms (GMO) in agriculture. Without GMO, farmers' income depend basically on climatic and market variables. Available data allow to estimate their probability distribution and their impact on income. With GMO, expected income remains assessable conditionally on the absence of cross-fertilization and contamination of other plants. However, neither the plausibility of the contamination, nor its consequences on the farmers' income, can be precisely evaluated. Here again, the standard approach appears unsatisfactory.

In these situations, and many others (introduction of new technologies, marketing of new medicines,...) decisions seem best represented by incompletely analyzed acts: conditionally to some events consequences of decisions on sub-events are perfectly known and uncertainty becomes expressable through probabilities, whereas the plausibility of these events themselves remains vague and the decision outcomes on complementary events are imprecisely known.

The axiomatic model proposed below is an attempt at formalizing such situations and at justifying adapted decision criteria.

2 The Model

Decisions.

Consider: Ω, set of states of nature; $\mathcal{E}$, $\sigma-$algebra of events; $\mathcal{C}$, a set of consequences; $\mathcal{G}$, $\sigma-$algebra of subsets of $\mathcal{C}$ containing singletons. A decision problem involves a particular set of decisions, $\mathcal{D}$, which are (measurable) acts in the sense of Savage[1], i.e., mappings $(\Omega, \mathcal{E}) \longrightarrow (\mathcal{C}, \mathcal{G})$. However, in the decision model below, these acts are not completely known by the decision maker. Specifically, the decisions are only partially analyzed, i.e., for any decision $a \in \mathcal{D}$ there is an event A such that the restriction of a to A - the analyzed part of a - denoted $a|_A$ is exactly known but the only information about $a|_{A^c}$ - the non-analyzed part of a - is its range $M_a = a(A^c)$. Thus, preferences will depend on pairs $(a|_A\,,\, M_a)$.

A specific feature of the model is that $\mathcal{D}$ is not assumed to contain all conceivable pairs $(a|_A\,,\, M_a)$. The reason is that decision makers cannot be expected to meaningfully evaluate unrealistic decisions. Thus the range M on an ”unfavorable” event (such as a natural catastrophe) should not include any blissful consequence. Similarly, in some situations, major ignorance about the relevant event will necessarily imply much uncertainty about outcomes i. e. a wide consequence range M on this event.

Completely analyzed decisions, denoted by $(a|_\Omega\,, \cdot)$, can exist. In particular, for evaluation purposes, we shall assume the existence of completely analyzed $\mathcal{R}$-measurable acts, where subalgebra $\mathcal{R}$ of $\mathcal{E}$ can be interpreted as events associated with sequences of heads and tails (see Savage [6, p. 38-39] and de Finetti [2, p.199-202]).

[1] More precisely, we use Savage's remark [6, § 3.4, p. 42] that the results in his model remain valid with events, consequences and acts defined in the present way.

A decision a analyzed on an event A is called an A - act. It generates a σ - algebra of subsets of A : $\left\{ a|_A^{-1}(G),\ G \in \mathcal{G} \right\}$, which we embed into a richer one, the σ - algebra $\mathcal{A}_a$ of subsets of A generated by $\left\{ a|_A^{-1}(G) \cap R,\ G \in \mathcal{G},\ R \in \mathcal{R} \right\}$.

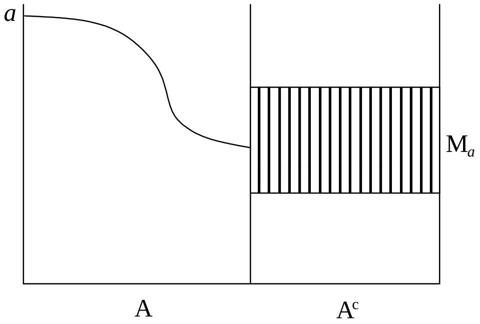


Figure 1: A partially analyzed act

We denote by $\mathcal{F}_a$ the set of all pairs $g = (g|_A,\ M_a)$ where $g|_A$ is any conceivable Savagean act $(A,\ \mathcal{A}_a) \longrightarrow (\mathcal{C},\ \mathcal{G})$. Thus, $g \in \mathcal{F}_a$ implies $M_g = M_a$. There is one such set corresponding to each $a \in \mathcal{D}$ and their union is denoted by $\mathcal{F}$. We denote by $\mathcal{A}_{\mathcal{F}}$ the set of all events A such that $\mathcal{F}$ contains at least one A-act.

Note that the fact that two acts a' and a'' are both A-acts, i.e., are analyzed on the same event A, does not imply the identity of $\mathcal{A}_{a'}$ and $\mathcal{A}_{a''}$, nor that of $\mathcal{F}_{a'}$ and $\mathcal{F}_{a''}$.

Example 1 *Acts a, a', a'' characterize various oil field management strategies in the same country. Political risk (event A^c) may imply partial or complete loss of the investment. Act a' involves the same investment level I as a but concerns the exploitation of a different oil field, whereas act a'' corresponds to a more intensive exploitation of the same field as a. Thus, it is likely that $M_{a'} = M_a = [0, -I]$ but $\mathcal{A}_{a'} \neq \mathcal{A}_a$ (oil yields depend on different events), whereas $M_{a''} = [0, -I''] \neq M_a$ and $\mathcal{A}_{a''} = \mathcal{A}_a$. Hence, although the three acts are analyzed on the same event A, $\mathcal{F}_a$, $\mathcal{F}_{a'}$ and $\mathcal{F}_{a''}$ all differ from one another.*

Preferences.

Preferences on $\mathcal{F}$ are expressed by a binary relation $\succsim$. We assume:

Axiom 1 *$\succsim$ is a weak order on $\mathcal{F}$.*

We want to endow $\succsim$ with standard properties and, moreover, to establish links between its restrictions $\succsim_a$ to the various $\mathcal{F}_a$. For this, we need in particular an appropriate version of Savage’s Sure Thing Principle.

Due to the partial information on the decisions, the common part $Com(a,b)$ of two acts a and b analyzed on events A and B, respectively, is defined as

$$Com(a,b) = \begin{cases} \{\omega \in A \cap B : a(\omega) = b(\omega)\} \text{ if } M_a \neq M_b \\ \{\omega \in A \cap B : a(\omega) = b(\omega)\} \cup (A^c \cap B^c) \text{ if } M_a = M_b \end{cases}$$

Axiom 2 *(Sure Thing Principle for partially analyzed decisions)*

Let $a, \widehat{a}, b, \widehat{b} \in \mathcal{F}$ where $\widehat{a}$ results from a and $\widehat{b}$ from b by a common modification in the sense that $Com(a,b) = Com(\widehat{a},\widehat{b})$.

Then $a \succsim b \Longleftrightarrow \widehat{a} \succsim \widehat{b}$.

Note that the feasible common modifications of a given pair of acts are strongly limited by the fact that the modified acts must still belong to $\mathcal{F}$.

Note also that $\mathcal{F}_a$, $\mathcal{F}_{\widehat{a}}$, $\mathcal{F}_b$, $\mathcal{F}_{\widehat{b}}$ may differ.

Example 2 *Suppose there are three countries: $\mathbb{A},\mathbb{B}$ and $\mathbb{C}$. Country $\mathbb{A}$ (resp. $\mathbb{B}$) may possibly face an economic crisis (event A^c (resp. B^c)) which however is unlikely in country $\mathbb{C}$. A firm has to take a decision concerning a productive investment of amount I. The decision a of investing I in country $\mathbb{A}$ will generate sales shared out among countries $\mathbb{A},\mathbb{B}$ and $\mathbb{C}$ in proportions 45% in country $\mathbb{A}$, 5% in country $\mathbb{B}$ and 50% in country $\mathbb{C}$, unless economic crisis (event A^c) happens in $\mathbb{A}$ in which case I may be partially or completely lost, independently of crisis occurring or not in country $\mathbb{B}$.*

On the other hand, consider a' with the same amount of investment in $\mathbb{A}$ as a but generating a different sales sharing, namely 70%, 30% and 0% respectively in countries $\mathbb{A},\mathbb{B}$ and $\mathbb{C}$ if there is no economic crisis. With this investment decision, the firm may loose up to I if crisis occurs only in $\mathbb{A}$, but is sure to loose the investment completely if the crisis takes place simultaneously in $\mathbb{A}$ and $\mathbb{B}$ (event $A^c \cap B^c$).

Decisions b and b' have similar characteristics with the roles of countries $\mathbb{A}$ and $\mathbb{B}$ exchanged. We assume moreover that the countries are "similar", in the sense that the return from sales is the same in $\mathbb{A}$ as in $\mathbb{B}$, that is $a|_A = c$ and $b|_B = c$ with $c \in \mathcal{C}$.

Thus, a and b are respectively an $A-$act and a $B-$act with

$$Com(a,b) = (A \cap B) \cup (A^c \cap B^c)$$

and $M_a = M_b = [0, -I]$.

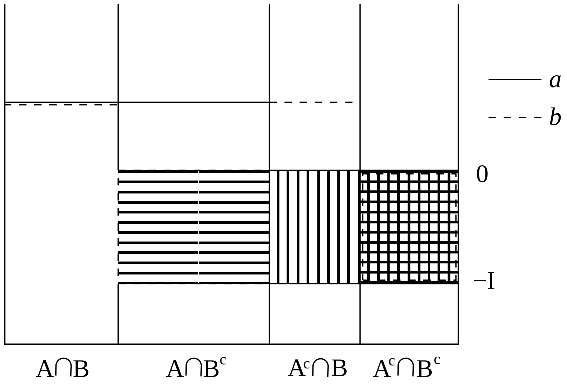


Figure 2: Original acts a and b.

a' and b' are $(A \cap B) \cup (A^c \cap B^c)$ – acts resulting from a and b by a modification of their common part. More precisely,

$$a|_{A\cap B} = b|_{A\cap B} = a'\big|_{A\cap B} = b'\big|_{A\cap B},$$
$$M_{a'} = M_{b'} = [0,-I] \text{ and } a'\big|_{A^c\cap B^c} = b'\big|_{A^c\cap B^c} = -I$$

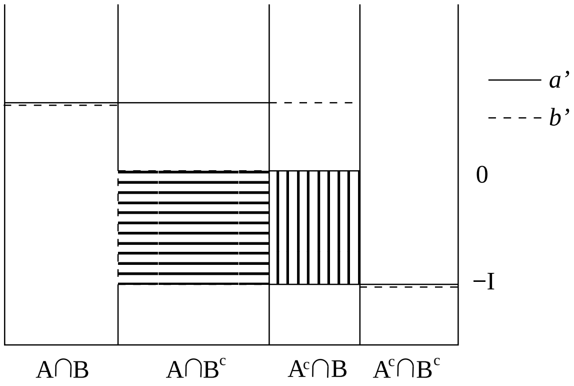


Figure 3: Modified acts a' and b'.

3 Preferences on Analyzed Events and SEU

From preferences $\succsim_a$ on $\mathcal{F}_a$, we can now derive, ”à la Savage”, $\succsim_a^E$, conditional preferences given event E, where $E \in \mathcal{A}_a$, by

$$g \succsim_a^E h \Leftrightarrow g' \succsim_a h' \text{ where } g'\big|_E = g|_E\,,\ h'\big|_E = h|_E \text{ and } g'\big|_{A\backslash E} = h'\big|_{A\backslash E}$$

Axiom 2 ensures that the ordering of g' and h' is independent from their common values on $A \backslash E$.

Note that $\succsim_a^A$ is the same as $\succsim_a$.

More generally, given an A-act, $a \in \mathcal{D}$ and a B-act, $b \in \mathcal{D}$ where $B \in \mathcal{A}_a$ (hence $B \subset A$), orderings $\succsim_a^B$ on $\mathcal{F}_a$ and $\succsim_b$ on $\mathcal{F}_b$ are related as shown by the following lemma.

Lemma 1 *Let* $a' = (\, a'|_A \,,\, M_a)\,,\ a'' = (\, a''|_A \,,\, M_a)$ *with* $a', a'' \in \mathcal{F}_a$*. Suppose that, for some* $B \in \mathcal{A}_a$*,* b *is a* $B-$*act and* $b' = (\, a'|_B \,,\, M_b)\,, b'' = (\, a''|_B \,,\, M_b) \in \mathcal{F}_b$*. Then*

$$a' \succsim_a^B a'' \Leftrightarrow b' \succsim_b b''.$$

Proof. Consider g' and g'' resulting from b' and b'' by the common modification consisting in giving them a constant common consequence $g'(\omega) = g''(\omega) = c$ for $\omega \in A \backslash B$ and the same range M_a on A^c. By Axiom 2, $g' \succsim g'' \Leftrightarrow b' \succsim_b b''$. Moreover, g' and g'' also belong to $\mathcal{F}_a$ and can be obtained by modifying a' and a'' on $A \backslash B$ and giving them the constant value c. By definition, $a' \succsim_a^B a'' \Leftrightarrow g' \succsim g''$. Hence $a' \succsim_a^B a'' \Leftrightarrow b' \succsim_b b''$. □

As a direct consequence of Lemma 1, conditional preferences given E are intrinsic in the sense that they do not depend on which $\mathcal{A}_a$ containing E (hence on which a in $\mathcal{F}$) is considered, and can be defined by $g \succsim^B h \Leftrightarrow$ there is a such that $g \succsim_a^E h$.

We also need slightly modified versions of the other Savage's definitions and axioms.

A constant A-act f_a^c in $\mathcal{F}_a$ is defined by: $f_a^c(A) = \{c\}$ with $c \in \mathcal{C}$ and $f_a^c(A^c) = M_a$.

Savage's P3 becomes:

Axiom 3 *For* $c', c'' \in \mathcal{C}$*, let* $f_a^{c'}, f_a^{c''}$ *be constant A-acts in* $\mathcal{F}_a$ *and* $f_b^{c'}, f_b^{c''}$ *be constant B-acts in* $\mathcal{F}_b$*.*

Then $f^{c'} \succsim f^{c''} \Longleftrightarrow f_b^{c'} \succsim f_b^{c''}$.

Preferences among consequences can now be defined by

$c' \succeq_{\mathcal{C}} c'' \Longleftrightarrow$ there exist $a \in \mathcal{D}$ and constant A-acts $f_a^{c'}, f_a^{c''}$ in $\mathcal{F}_a$ such that $f_a^{c'} \succsim f_a^{c''}$.

Since $\mathcal{C}$ can always be replaced by its quotient, we henceforth assume w.l.o.g. that $\succeq_{\mathcal{C}}$is an order (i.e. is antisymmetric) which justifies the use of symbol $\succeq_{\mathcal{C}}$.

We moreover assume the existence of $\underline{c}$, $\overline{c}$, respectively the worst and the best consequence in $\mathcal{C}$.

We now require Savage's P4 (irrelevance of the values of the prizes on the events) in every $\mathcal{F}_e$, where $e \in \mathcal{D}$ is an E-act.

Axiom 4 *Let $A, B \in \mathcal{A}_e$ where $e \in \mathcal{D}$ is an E-act; let $c_1, c_1', c_2, c_2' \in \mathcal{C}$ be such that $c_1 \succ_{\mathcal{C}} c_1'$ and $c_2 \succ_{\mathcal{C}} c_2'$. Define acts $f, f', g, g' \in \mathcal{F}_e$ by:*
i) $f(E^c) = f'(E^c) = g(E^c) = g'(E^c) = M_e$;
ii) $f(\omega) = c_1, \quad f'(\omega) = c_1'$, for $\omega \in A$;
$f(\omega) = c_2, \quad f'(\omega) = c_2'$, for $\omega \in E \backslash A$;
iii) $g(\omega) = c_1, \quad g'(\omega) = c_1'$, for $\omega \in B$;
$g(\omega) = c_2, \quad g'(\omega) = c_2'$, for $\omega \in E \backslash B$;
then $f \succsim g \Leftrightarrow f' \succsim g'$.

Whenever $f \succsim g$ holds for f, g defined as in Axiom 4, we can write $A \succsim_e^E B$. However, if $A, B \in \mathcal{A}_{e^*}$ for some other $e^* \in \mathcal{D}$ which is also an $E-$act, it results from Axiom 2 ($f(E^c) = f'(E^c) = g(E^c) = g'(E^c) = M_e$ above can be replaced by $f(E^c) = f'(E^c) = g(E^c) = g'(E^c) = M_{e^*}$) that: $A \succsim_e^E B \Leftrightarrow A \succsim_{e^*}^E B$. We can therefore drop the subscript e and simply write $A \succsim^E B$ and read "event A is qualitatively more probable than B conditionally to event E".

The next axiom is Savage's P5.

Axiom 5 *There exists a pair $c', c'' \in \mathcal{C}$ such that $c' \succ_{\mathcal{C}} c''$.*

We also introduce a version of Savage's P6. It makes it clear that one of the roles of the coin-toss related subalgebra of events $\mathcal{R}$ is to make all $(\mathcal{A}_a, \succsim_a)$ atomless.

Axiom 6 *Let $f, g \in \mathcal{F}_a$, where $a \in \mathcal{D}$ is an A-act, with $f \succ g$ and $c \in \mathcal{C}$. There exists a partition of A, consisting of events $R \cap A$, $R \in \mathcal{R}$, such that if f (resp. g) is modified on any element of the partition and given constant outcome c on this element, then the modified act f' (resp. g') also satisfies $f' \succ g$ (resp. $f \succ g'$).*

We also need Savage's P7 for each $\succsim_a$.

Axiom 7 *Let $f, g \in \mathcal{F}_a$, where $a \in \mathcal{D}$ is an A-act and let $E \in \mathcal{A}_a$. If $f \succsim_a^E$ (resp. $\precsim_a^E$) $g(\omega)$ for all $\omega \in E$, then $f \succsim_a^E$ (resp. $\precsim_a^E$) g.*

Axioms 1-7 imply the validity of Savage's P1-7 in every $\mathcal{F}_a$, where thus his main result holds: preferences in $\mathcal{F}_a$ can be represented by a subjective expected utility (SEU) criterion with respect to an atomless probability on $\mathcal{A}_a$.

Moreover, due to the explicit introduction of σ-algebra $\mathcal{R}(A) = \{A \cap R, R \in \mathcal{R}\}$ in the statement of Axiom 6, it is clear that this result still holds if $\mathcal{F}_a$ is replaced by its restriction to $\mathcal{R}(A)$ - measurable acts. We can thus state:

Proposition 1 *For every $a \in \mathcal{D}$ there exist a bounded utility u_a and an additive probability P_a such that*

$$f \succsim g \Leftrightarrow \int_A u_a \circ f \, dP_a \geq \int_A u_a \circ g \, dP_a, \quad \forall f, g \in \mathcal{F}_a$$

where

u_a is unique up to an affine transformation;

P_a is unique and for every $\rho \in [0,1]$ there exists $B \in \mathcal{A}_a$ such that $P_a(B) = \rho$.

Moreover, these existence and uniqueness statements are also valid when $\mathcal{F}_a$ is replaced by its restriction to $\mathcal{R}(A)$ - measurable acts and thus $\mathcal{A}_a$ by $\mathcal{R}(A)$.

4 Intrinsic Utility and Probability Consistency

It is well known that Savage's axioms do not imply the existence of certainty equivalents for the acts. However, this property is easily acceptable for sufficiently rich consequence sets (for instance when $\mathcal{C}$ is a real interval) and, although not necessary, will be technically helpful later in the paper. So, we assume:

Axiom 8 *For any $a \in \mathcal{F}$ there exist $c \in \mathcal{C}$ such that the constant A-act f_a^c satisfies $f_a^c \sim_a a$*

The next assumption and the lemma that follows assert that coin-toss related events are "qualitatively" independent and thus "quantitatively" independent from events in $\mathcal{E}$.

Axiom 9 *For every $A, B \in \mathcal{A}_{\mathcal{F}}$ conditional preferences on events $\succsim^A$ and $\succsim^B$ satisfy, for all $R', R'' \in \mathcal{R}$:*

$$A \cap R' \succsim^A A \cap R'' \Longleftrightarrow B \cap R' \succsim^B B \cap R''.$$

Lemma 2 *Let $a, b \in \mathcal{D}$. For every $R \in \mathcal{R}$, $P_a(A \cap R) = P_b(B \cap R)$.*

Proof. For any $R', R'' \in \mathcal{R}$, $P_a(A \cap R') \geq P_a(A \cap R'') \Leftrightarrow A \cap R' \succsim^A A \cap R'' \Longleftrightarrow B \cap R' \succsim^B B \cap R'' \Leftrightarrow P_b(B \cap R') \geq P_b(B \cap R'')$. Thus, the mapping $\mathcal{R}(A) \longmapsto [0,1]$ defined by $A \cap R \longmapsto P_b(B \cap R)$ is a probability measure representing $\succsim^A$ which however is uniquely represented by P_a. Therefore $P_a(A \cap R) = P_b(B \cap R)$ for every $R \in \mathcal{R}$. □

Whenever $A \cap R' \succsim^A A \cap R''$ holds for $R', R'' \in \mathcal{R}$ and some $A \in \mathcal{A}_{\mathcal{F}}$, we shall simply write $R' \succsim^{\mathcal{R}} R''$ and read "event R' is qualitatively more probable than R'' ". Qualitative probability $\succsim^{\mathcal{R}}$ is uniquely represented by probability $P_{\mathcal{R}}$ defined by $P_{\mathcal{R}}(R) = P_a(A \cap R)$ for some A.

Thus, Axiom 8 ensures the existence of an intrinsic probability $P_{\mathcal{R}}$ on $\mathcal{R}$.

We shall use this result to derive properties of utilities. That far, all we know about the $u_a, a \in \mathcal{D}$ is that they represent the same ordering $\succeq_C$ and are therefore increasing transforms from one another. We would like functions u_a to be identical (after calibration).

According to Proposition 1 for every triple $c' \succeq_C c \succeq_C c''$, with $c' \succ_C c''$, there is an event $R \in \mathcal{R}$ such that act $g \in \mathcal{F}_a$ with $g(\omega) = c'$, for $\omega \in A \cap R$, and $g(\omega) = c''$ for $\omega \in A \cap R^c$ is indifferent to the constant $A-$act f_a^c in $\mathcal{F}_a$. In other terms, there is $R \in \mathcal{R}$ such that $P_a(A \cap R)$ satisfies:

$$u_a(c) = P_a(A \cap R)u_a(c') + (1 - P_a(A \cap R))u_a(c''), \tag{1}$$

hence, according to the definition that follows Lemma 2

$$u_a(c) = P_{\mathcal{R}}(R)u_a(c') + (1 - P_{\mathcal{R}}(R))u_a(c'').$$

Thus, all we need is an axiom ensuring that the event R in (1) only depends on c.

Axiom 10 *For every triple $c' \succeq_C c \succeq_C c''$, with $c' \succ_C c''$, there exist an event $R \in \mathcal{R}$ such that for every $a \in \mathcal{D}$, act $g \in \mathcal{F}_a$ with $g(\omega) = c'$, for $\omega \in A \cap R$, and $g(\omega) = c''$ for $\omega \in A \cap R^c$ is indifferent to the constant $A-$act f_a^c in $\mathcal{F}_a$.*

If follows immediately that:

Proposition 2 *Utilities u_a ($a \in \mathcal{D}$) are affine transforms from one another.*

Thus, after calibration u_a's are identical and we will write from now on u instead of u_a. Note that u is a utility function representing $\succeq_C$.

Next proposition guarantees the existence of intrinsic conditional probabilities in the sense that they are independent from the context in which they are evaluated.

Proposition 3 *Let $a, b \in \mathcal{D}$ be analyzed on A and B, respectively, with $B \in \mathcal{A}_a$ and let moreover $E \in \mathcal{A}_b$ (hence $E \subset B \subset A$). Then $P_a(E/B) = P_b(E)$.*

Proof. By Proposition 2, there exists $R \in \mathcal{R}$ such that $R \cap B \sim_b E$, and thus, by Lemma 1, $R \cap B \sim_a^B E$, implying

$$P_b(R \cap B) = P_b(E) \text{ and } P_b(R \cap B/B) = P_a(E/B). \tag{2}$$

Moreover, by applying Lemma 1 to acts offering prizes on events $R' \cap B$ and $R'' \cap B$, where $R', R'' \in \mathcal{R}$, we get $R' \cap B \succsim_b R'' \cap B \Leftrightarrow R' \cap B \succsim_a^B R'' \cap B$. Thus, the same ordering (say $\succsim_b$) on set of events $\{R \cap B,\ R \in \mathcal{R}\}$ is representable by (restrictions of) probabilities P_b and $P_a(./B)$; by uniqueness of such a representation (see Proposition 2), $P_b(R \cap B) = P_a(R \cap B/B)$, for all $R \in \mathcal{R}$. Then according to (2) $P_b(E) = P_a(E/B)$. □

Thus, as for conditional preferences, intrinsic conditional probabilities can be defined by $P(E/B) = P_a(E/B)$ where $E,\ B \in \mathcal{A}_a$ and $E \subset B$.

5 Preferences on Non-analyzed Events

We now turn to the non-analyzed part of the decisions.

Let $\mathcal{M}$ denote the set of ranges corresponding to all the decisions:

$\mathcal{M} = \{M_a,\ \exists\, a \in \mathcal{D}$ such that $a = (a|_A\,,\, M_a)\}$.

We assume that every $M \in \mathcal{M}$ has a $\succeq_C$ - greatest and a $\succeq_C$ - lowest consequence, respectively denoted $g(M)$ and $l(M)$.

We define a partial preference relation over $\mathcal{M}$. For this, two axioms are needed: Axiom 11 ensures the existence of the relation and Axiom 12 its transitivity.

Axiom 11 *Let a', a'' be $A-$acts such that $a' = (a'|_A\,,\, M_{a'})$, $a'' = (a''|_A\,,\, M_{a''})$ with $a'|_A = a''|_A$and let b', b'' be $B-$acts such that $b' = (b'|_B\,,\, M_{b'})$, $b'' = (b''|_B\,,\, M_{b''})$ with $b'|_B = b''|_B\,,\, M_{b'} = M_{a'}$ and $M_{b''} = M_{a''}$. Then*

$$a' \succsim a'' \Leftrightarrow b' \succsim b''.$$

Preferences among ranges can now be defined by the transitive closure $\succsim_{\mathcal{M}}$ of the relation $\succsim^0_{\mathcal{M}}$ given by:

$M' \succsim^0_{\mathcal{M}} M'' \Longleftrightarrow$ there exist $A-$ acts $a', a'' \in \mathcal{D}$ such that $M_{a'} = M'$, $M_{a''} = M''$, $a'|_A = a''|_A$ and $a' \succsim a''$.

$\succsim_{\mathcal{M}}$ is automatically a partial order if:

Axiom 12 *$\succsim^0_{\mathcal{M}}$ is acyclic i.e. there is no sequence M^i, $i = 1..n$ in $\mathcal{M}$ such that $M^i \succsim^0_{\mathcal{M}} M^{i+1}$, $i = 1..n-1$ and $M^n \succ^0_{\mathcal{M}} M^1$.*

Let's now turn to the representation of the preference relation $\succsim_{\mathcal{M}}$.

The following requirement will allow us to extend a result of Barbera, Barrett and Pattanaik [1].

Axiom 13 *(1) $\forall M, c, \exists\, A$ and two $A-$acts a', a'' such that $M_{a'} = M$ and $M_{a''} = M \cup \{c\}$*

(2) Let $c_1, c_2 \in \mathcal{C}$ be such that $c_1 \succ_C c_2$. Then, for any $M_0 \in \mathcal{M}$ such that $c_1, c_2 \notin M_0$,

$$\{c_1\} \cup M_0 \succsim_{\mathcal{M}} \{c_1, c_2\} \cup M_0 \succsim_{\mathcal{M}} \{c_2\} \cup M_0.$$

Moreover, if $c \succ_C c_2$ for all $c \in M_0$, then:

$$\{c_1\} \cup M_0 \succ_{\mathcal{M}} \{c_1, c_2\} \cup M_0$$

and if $c_1 \succ_C c$ for all $c \in M_0$, then

$$\{c_1, c_2\} \cup M \succ_{\mathcal{M}} \{c_2\} \cup M_0.$$

Note that, if $M_0 = \emptyset$, we get

$$\{c_1\} \succ_{\mathcal{M}} \{c_1, c_2\} \succ_{\mathcal{M}} \{c_2\}.$$

Note that Axiom 16 makes both existence and comparability requirements.

Lemma 3 *(i) For all finite $M \in \mathcal{M}$ such that $g(M) \succ_C l(M)$, $M \sim_{\mathcal{M}} \{g(M), l(M)\}$.*
(ii) For finite $M', M'' \in M$:

$$\left.\begin{array}{r} g(M') \succeq_C g(M'') \\ l(M') \succeq_C l(M'') \end{array}\right\} \Rightarrow M' \succsim_{\mathcal{M}} M''. \tag{3}$$

Moreover,

$$\left.\begin{array}{r} g(M') \succ_C g(M'') \\ l(M') \succ_C l(M'') \end{array}\right\} \Rightarrow M' \succ_{\mathcal{M}} M''. \tag{4}$$

Proof. *(i)* For $c \in M \backslash \{g(M), l(M)\}$, by Axiom 13, $g(M) \succ_C c$ implies $M \backslash \{c\} \succsim_{\mathcal{M}} M$ (take $M_0 = M \backslash \{g(M), c\}$) and symmetrically $c \succ_C l(M)$ implies $M \succsim_{\mathcal{M}} M \backslash \{c\}$; hence $M \sim_{\mathcal{M}} M \backslash \{c\}$.

Let $M = \{g(M), c_1, c_2, ..., c_n, l(M)\}$ where $g(M) \succ_C c_1 \succ_C c_2 \succ_C ... \succ_C c_n \succ_C l(M)$. Then, by repeated application of last relation:

$M \sim_{\mathcal{M}} M \backslash \{c_1\} \sim_{\mathcal{M}} M \backslash \{c_1, c_2\} \sim_{\mathcal{M}} ...$
$\sim_{\mathcal{M}} M \backslash \{c_1, c_2, ..., c_n\} = \{g(M), l(M)\}$.

(ii) From (i) of the Lemma, we have $M' \sim_{\mathcal{M}} \{g(M'), l(M')\}$ and $M'' \sim_{\mathcal{M}} \{g(M''), l(M'')\}$. $\succsim_{\mathcal{M}}$ being transitive (Axiom 12), we just need to prove that $\{g(M'), l(M')\} \succsim_{\mathcal{M}} \{g(M''), l(M'')\}$. Assume that in the left side of (3) there is at least one strict preference, for instance $g(M') \succ_C g(M'')$ (if it is not the case, the result is straightforward). By Axiom 13 (point (2)) with $M_0 = \{l(M')\}$, we have $\{g(M'), l(M')\} \succsim_{\mathcal{M}} \{g(M''), l(M')\}$. If $l(M') \succ_C l(M'')$, by the same Axiom with $M_0 = \{g(M'')\}$, $\{g(M''), l(M')\} \succsim_{\mathcal{M}} \{g(M''), l(M'')\}$. Else $(l(M') \sim_C l(M''))$, from the proof of (i)

$$\left\{g(M''), l(M'), l(M'')\right\} \sim_{\mathcal{M}} \left\{g(M''), l(M')\right\} \sim_{\mathcal{M}} \left\{g(M''), l(M'')\right\}.$$

The proof of the second part of (ii) is similar and uses the second part of point (2) in Axiom 13 (strict inequalities). □

Lemma 3 directly implies that, for a finite sequence $(M_i)_{i=1}^n$ of finite M_i with $g(M_i)$ and $l(M_i)$ independent of i, $\cup_{j=1}^n M_j \sim_{\mathcal{M}} M_i$, $i = 1..n$. We extend this property to infinite unions in the following axiom.

Axiom 14 *For any family $(M_i)_{i \in I}$, of finite $M_i \in \mathcal{M}$ such that $g(M_i)$ and $l(M_i)$ are independent of i, $\cup_{j \in I} M_j \sim_{\mathcal{M}} M_i$, $i \in I$.*

We can then prove the following proposition:

Proposition 4 *For all $M \in \mathcal{M}$ such that $g(M) \succ_{\mathcal{C}} l(M)$, $M \sim_{\mathcal{M}} \{g(M), l(M)\}$.*

Proof. It is sufficient to note that any M in $\mathcal{M}$ is the infinite union of finite subsets of it also in $\mathcal{M}$ and with the same greatest and lowest elements. □

Proposition 5 *There exists a mapping $v : \mathcal{M} \to \mathbb{R}$ such that*

$$M' \succ_{\mathcal{M}} M'' \Rightarrow v(M') > v(M'')$$
$$M' \sim_{\mathcal{M}} M'' \Rightarrow v(M') = v(M'')$$

with $M \mapsto v(M) = \varphi(g(M), l(M))$ and

$$\left.\begin{array}{c} g(M') \succ_{\mathcal{C}} g(M''), l(M') \succeq_{\mathcal{C}} l(M'') \\ \textit{or} \\ g(M') \succeq_{\mathcal{C}} g(M''), l(M') \succ_{\mathcal{C}} l(M'') \end{array}\right\} \Rightarrow v(M') > v(M'').$$

Proof. Let the elements of $\mathcal{C}$ be indexed as $c_1 \succ_{\mathcal{C}} c_2 \succ_{\mathcal{C}} \ldots \succ_{\mathcal{C}} c_N$ and mapping φ defined:

$$\text{for } i < j \text{ by } \varphi(c_i, c_j) = \sum_{(r,s) \in E_{ij}} \frac{1}{2^{r+s}},$$
$$\text{where } E_{ij} = \left\{(r,s) : r < s \text{ and } \left\{c_i, c_j\right\} \succ_{\mathcal{M}} \{c_r, c_s\}\right\}$$
$$\text{for } i = j \text{ by } \varphi(c_i, c_i) = \sum_{(r,s) \in F_i} \frac{1}{2^{r+s}},$$
$$\text{where } F_i = \{(r,s) : r < s \text{ and } \{c_i\} \succ_{\mathcal{M}} \{c_r, c_s\}\}$$

Then, v defined by $v(M) = \varphi(g(M), l(M))$ has the required properties since if $g(M) \succ_{\mathcal{M}} l(M)$ then $M \sim_{\mathcal{M}} \left\{c_i, c_j\right\}$ for some $c_i = g(M)$ and $c_j = l(M)$ and if $g(M) = l(M)$ $M \sim_{\mathcal{M}} \{c_i\}$ for $c_i = g(M)$. □

6 Representation Theorem

We now want to construct a utility representation of preferences $\succsim$ in $\mathcal{F}$ that incorporates the results obtained so far concerning its restrictions $\succsim_a$ to the various $\mathcal{F}_a$ as well as those concerning $\succsim_{\mathcal{M}}$.

This construction will be based on the existence of certainty equivalent for the acts which is directly required by the following axiom, where f^k denotes the constant act: $f^k(\Omega) = \{k\}$.

Axiom 15 *For any act $a \in \mathcal{F}$ there exists $k \in \mathcal{C}$ such that $f^k \sim a$.*

Proposition 6 *The weak order $\succsim$ on $\mathcal{F}$ is representable by a utility function V :*

- *For an A-act a such that $A \neq \Omega$,*

$$a = (\, a|_A\,,\, M_a) \longmapsto V(a) = \Phi\left(A,\ \int_A u \circ a\, dP_a,\ g(M_a),\ l(M_a)\right)$$

where

P_a is a subjective conditional probability on the σ-algebra $\mathcal{A}_a$;
$g(M_a)$, $l(M_a)$ are the $\succeq_C$ - greatest and the $\succeq_C$ - lowest consequences in M_a;
and Φ is increasing in $\int_A u \circ a\, dP_a$, $g(M_a)$, $l(M_a)$.

- *Otherwise, for $A = \Omega$,*

$$a = (\, a|_\Omega\,,\, \cdot) \longmapsto V(a) = \Psi\left(\int_\Omega u \circ a\, dP_a\right)$$

with Ψ increasing in $\int_\Omega u \circ a\, dP_a$.

Proof. Any a in $\mathcal{F}$ has a certainty equivalent k in $\mathcal{C}$ (by Axiom 15) and $\succeq_C$ is representable by utility function u. A priori consequence k, hence number $u(k)$, depends on all the elements characterizing a namely A, $\mathcal{A}_a$, $a|_A$ and $\mathcal{A}_a$.

Since, by Axiom 8, there exist c in $\mathcal{C}$ such that $a \sim_a f^c_a$, then

$$a \sim (A,\ \mathcal{A}_a,\ f^c_a|_A\,,\ M_a). \tag{5}$$

The constant A-act f^c_a being measurable with respect to any σ-algebra $\mathcal{A}_a$ of subsets of A, we have, for any A-acts a', a'' such that $M_{a'} = M_{a''}$ and $f^c_{a'} = f^c_{a''}$, $a' \sim a''$. Thus, the preference between a' and a'' does not explicitly depend on $\mathcal{A}_{a'}$ and $\mathcal{A}_{a''}$ and (5) becomes:

$$a \sim (A,\ f^c_a|_A\,,\ M_a). \tag{6}$$

Moreover, the certainty equivalent k depends on $a|_A$ only through $\int_A u \circ a\, dP_a$ (by Proposition 1) and on M_a only through $g(M_a)$, $l(M_a)$ (by Proposition 4). □

Example 3 *A common practice in international borrowing consists in classifying countries into various groups according to their insolvency risk. The rating is generally based on a check-list of economic indicators through a multiple criteria decision model; probability evaluations are rarely involved (Cf: Saini and Bates [5]). A given country is then allowed to borrow money at an interest rate* *equal to the LIBOR, i plus a risk spread* Δi*, which depends on its group. Thus, the net expected present value of a one period investment* I *is*

$$EV = -I + \frac{ER}{(1+i+\Delta i)} = -I + \frac{ER}{(1+i)} - \frac{\Delta i \times ER}{(1+i+\Delta i)}; \tag{7}$$

the risk premium, given by the last term, is proportional to the expected return ER*. On the contrary, a particular, additive, instance of our model would evaluate the preceding investment according to formula:*

$$V^* = -I + \frac{ER}{(1+i)} - k \times I$$

i. e. require the risk premium to be proportional to the maximal possible loss, here I*, which seems to make more sense.*

7 Discussion

The family of criteria described by the representation theorem is still rather wide and various behavioural assumptions could be added and lead to more specific criteria. On the other hand, the building blocks of the model, SEU for the analyzed part and "(max, min)" for the non-analyzed one could easily be replaced by other theories for instance the analyzed part would still be be endowed with probabilities but Quiggin's Rank Dependent Utility [4] would replace EU or information on the non-analyzed part of the acts would not be quantified in terms of consequence sets but according to symbolic categories.

The model is consistent with various generalizations of SEU. For instance partially analyzed acts are a special case of multivalued acts; once restricted to this special class, the criteria of Ghirardato's model [3], become a subfamily of ours. Moreover, our model allows the expression of various types of beliefs concerning the relative plausibility of the analyzed and the non analyzed events ranging from probabilities $(P(A)+P(A^c)=1)$ to complete ignorance that include capacities $(v(A)+v(A^c)\neq 1)$, and in particular necessities (for instance $N(A)=\alpha, N(A^c)=0)$.

References

[1] S. Barbera, C. R. Barrett, P. K. Pattanaik. On Some Axioms for Ranking Sets of Alternatives. *Journal of Economic Theory,* 33:301–308, 1984.

[2] B. de Finetti. *Theory of Probability (vol. 1),* Wiley, 1974.

[3] P. Ghirardato. Coping with Ignorance: Unforeseen Contingencies and Non-Additive Uncertainty. *Economic Theory*, 17:247–276, 2001.

[4] J. Quiggin. A Theory of Anticipated Utility. *Journal of Economic Behavior and Organization*, 3:324–343, 1982.

[5] K. G. Saini, P. S. Bates. A survey of the quantitative approaches to country risk analysis. *Journal of Banking and Finance*, 8-2:341–356, 1984.

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

Jean-Yves Jaffray is a Professor at Université Paris 6. LIP6, pôle IA, 8, rue du capitaine Scott, F-75015 Paris, France. E-mail: Jean-Yves.Jaffray@lip6.fr

Meglena Jeleva is an Associate Professor at Université de Nantes. Member of EUREQua, Maison des Sciences Economiques, 106-112 boulevard de l’Hôpital, F-75647 Paris cedex 13, France. E-mail: Meglena.Jeleva@univ-paris1.fr

On Approximating Multidimensional Probability Distributions by Compositional Models*

R. JIROUŠEK
Academy of Science, Czech Republic

Abstract

Because of computational problems, multidimensional probability distributions must be approximated by distributions which can be defined by a reasonable number of parameters. As a rule, distributions with a special dependence structure (i.e., complying with a system of conditional independence relations) are considered; graphical Markov models and especially Bayesian networks are often used. This paper proposes application of compositional models for this puropose. In addition to a theoretical background, a heuristic algorithm solving one part of a model learning process is presented. Its basic idea, construction of an approximation exploiting informational content of given low-dimensional distributions in a maximal possible way, was proposed by Albert Perez as early as in 1977.

Keywords

multidimensional distributions, approximations, conditional independence, operator of composition

1 Introduction

Data-driven methods for probability model construction usually suffer from a lack of data. This is why one must always keep in mind that any probability estimate is imprecise and the more probabilities, the less precise their estimates. Moreover, it would be absurd to try to get estimates of (let us say) 2^{50} probabilities defining a 50-dimensional distribution (of binary variables) from a file whose size is only several Mbytes. Such an effort would also be in contradiction with the *Minimum Description Length* principle often employed in the field of AI. Therefore, application of probabilistic models to problems of practice, when the dimensionality

*This work has been supported in part by GA AV ČR, under grant A2075302.

of considered multidimensional probability distributions is expressed in hundreds rather than tens, quite naturally leads to the necessity of approximations.

The present paper proposes to look for an approximation of a probability distribution in a class of so-called *compositional models* (CM), which is an alternative apparatus to that usually called *Graphical Markov Modeling* (GMM). GMM is used as a general term describing any of the approaches representing multidimensional probability distributions by means of graphs and systems of quantitative parameters like Bayesian networks (BN), decomposable and graphical models, influence diagrams and chain graph models.

The main idea of CM is the same as that of GMM: not to strive for estimating multidimensional distribution but only its oligo-dimensional marginals, from which the multidimensional model is subsequently composed. In a way this model resembles a jigsaw puzzle that has a great number of parts, each bearing a local piece of a picture, and the goal is to find how to assemble them in a way that the global picture makes sense, reflecting all the individual small parts. Naturally, the whole task can be split into two subproblems: how to find which oligo-dimensional distributions are to be estimated and how to compose them into a multidimensional model. Though the present paper concentrates exclusively on the latter one, let us mention that, to be consistent with the apparatus employed in this paper, the problem of selection of oligodimensional distributions should be solved with the help of information theoretic quantities; distributions with the highest informational content (see section 5) should be selected.

Before introducing the apparatus of CM let us mention that both GMM and CM are based on the very idea published by Albert Perez as early as 1977 in his unfortunately neglected paper [10]. In this paper Perez calls these probability distributions *dependence structure simplification approximations* and studies increase of risk connected with statistical decision problem when, instead of Bayes optimal solution, ε-Bayes optimal solution (ie., Bayes optimal with respect to ε-approximation) is accepted.

2 Notation

In this text, we will deal with a finite system of finite-valued random variables. Let N be an arbitrary finite index set, $N \neq \emptyset$. Each variable from $\{X_i\}_{i \in N}$ is assumed to have a finite (non-empty) set of values $\mathbf{X}_i$. Distributions of these variables will be denoted by Greek letters (π, κ); thus for $K \subseteq N$, we can consider a distribution $\pi((X_i)_{i \in K})$. To make the formulae more lucid, the following simplified notation will be used. Symbol $\pi(x_K)$ will denote both a $|K|$-dimensional distribution and a value of a probability distribution π (when several distributions are considered, we shall distinguish between them by indices), which is defined for variables $(X_i)_{i \in K}$ at a combination of values x_K; x_K thus represents a $|K|$-dimensional vector of values of variables $\{X_i\}_{i \in K}$. Analogously, we shall also denote the set of all these vectors $\mathbf{X}_K$:

$$\mathbf{X}_K = \times_{i \in K} \mathbf{X}_i.$$

For a probabilistic distribution $\pi(x_K)$ and $J \subset K$ we will often consider a *marginal distribution* $\pi(x_J)$ of distribution $\pi(x_K)$, which can be computed by

$$\pi(x_J) = \sum_{x_{K\setminus J} \in \mathbf{X}_{K\setminus J}} \pi(x_K) = \sum_{x_{K\setminus J} \in \mathbf{X}_{K\setminus J}} \pi(x_{K\setminus J}, x_J).$$

In this simple formula we have introduced a notation used throughout this article: a vector x_K is composed of two subvectors $x_{K\setminus J}$ and x_J, where x_J is a *projection* of x_K into $\mathbf{X}_J$, and, analogously $x_{K\setminus J}$ is a projection of x_K into $\mathbf{X}_{K\setminus J}$. For computation of marginal distributions we need not exclude situations when $J = \emptyset$. In accordance with the above-introduced formula we get $\pi(x_\emptyset) = 1$.

In some situations we will want to stress that we are dealing with a marginal distribution of a distribution π; we will use symbol $\pi^{(J)}$ to denote the marginal distribution of π for variables $(X_i)_{i \in J}$. That is, for $J \subseteq K$ and a distribution $\pi(x_K)$,

$$\pi^{(J)} = \pi(x_J).$$

For a distribution $\pi(x_K)$ and two disjoint subsets $J, L \subseteq K$ we will also speak about a *conditional distribution* $\pi(x_J|x_L)$, which is, for each fixed $x_L \in \mathbf{X}_L$, a $|J|$-dimensional probability distribution, for which $\pi(x_J|x_L)\pi(x_L) = \pi(x_{J\cup L})$. (Notice that this definition is ambiguous if $\pi(x_L) = 0$ for some combination(s) of values $x_L \in \mathbf{X}_L$.) The reader can immediately see that if $J = \emptyset$ then $\pi(x_J|x_L) = 1$, and if $L = \emptyset$ then $\pi(x_J|x_L) = \pi(x_J)$.

Consider $K \subseteq L \subseteq N$ and a probability distribution $\pi(x_K)$. With $\Pi^{(L)}$ we shall denote the set of all probability distributions defined for variables X_L. Similarly, $\Pi^{(L)}(\pi)$ will denote the system of all *extensions* of the distribution π to L-dimensional distributions:

$$\Pi^{(L)}(\pi) = \left\{ \kappa \in \Pi^{(L)} : \kappa(x_K) = \pi(x_K) \right\},$$

(where $\kappa(x_K)$ naturally denotes the marginal distribution of κ for variables X_K). Having a system

$$\Xi = \{\pi_1(x_{K_1}), \pi_2(x_{K_2}), \ldots, \pi_n(x_{K_n})\},$$

of oligo-dimensional distributions ($K_1 \cup \ldots \cup K_n \subseteq L$), the symbol $\Pi^{(L)}(\Xi)$ denotes the system of distributions that are extensions of all the distributions from Ξ:

$$\Pi^{(L)}(\Xi) = \left\{ \kappa \in \Pi^{(L)} : \kappa^{(K_i)} = \pi_i \ \ \forall i = 1, \ldots, n \right\} = \bigcap_{i=1}^{n} \Pi^{(L)}(\pi_i).$$

3 Operator of composition

To be able to compose low-dimensional distributions to get a distribution of a higher dimension we will introduce an *operator of composition*.

To make this construction clear from the very beginning, let us stress that it is just a generalization of the idea of computing the three-dimensional distribution from two two-dimensional ones introducing the conditional independence:

$$\pi(x_1,x_2) \triangleright \kappa(x_2,x_3) = \frac{\pi(x_1,x_2)\kappa(x_2,x_3)}{\kappa(x_2)} = \pi(x_1,x_2)\kappa(x_3|x_2).$$

Consider two probability distributions $\pi(x_K)$ and $\kappa(x_L)$, such that $\kappa(x_{L\cap K})$ dominates[1] $\pi(x_{L\cap K})$; in symbol: $\pi(x_{L\cap K}) \ll \kappa(x_{L\cap K})$. The *composition* of these two distributions is defined by the formula

$$\pi(x_K) \triangleright \kappa(x_L) = \frac{\pi(x_K)\kappa(x_L)}{\kappa^{(L\cap K)}}.$$

Since we assume $\pi^{(L\cap K)} \ll \kappa^{(L\cap K)}$, if for any $x \in \mathbf{X}_{(L\cap K)}$ $\kappa^{(L\cap K)}(x) = 0$ then there is a product of two zeros in the nominator and we take $0.0/0 = 0$. If $L \cap K = \emptyset$ then $\kappa^{(L\cap K)} = 1$ and the formula degenerates to a simple product of π and κ.

Let us stress that in the case $\pi^{(L\cap K)} \not\ll \kappa^{(L\cap K)}$, the expression $\pi \triangleright \kappa$ remains undefined.

Thus, the formal definition of the operator $\triangleright$ is as follows.

Definition 1 *For two arbitrary distributions $\pi \in \Pi^{(K)}$ and $\kappa \in \Pi^{(L)}$ their* composition *is given by the following formula*

$$\pi(x_K) \triangleright \kappa(x_L) = \begin{cases} \frac{\pi(x_K)\kappa(x_L)}{\kappa(x_{K\cap L})} & \textit{if } \pi(x_{K\cap L}) \ll \kappa(x_{K\cap L}), \\ \textit{undefined} & \textit{otherwise}. \end{cases}$$

The following simple assertion proven in [5] answers the question: what is the result of the composition of two distributions?

Theorem 1 *If $\pi(x_{L\cap K}) \ll \kappa(x_{L\cap K})$ (i.e., if $\pi(x_K) \triangleright \kappa(x_L)$ is defined) then $\pi(x_K) \triangleright \kappa(x_L)$ is a probability distribution from $\Pi^{(L\cup K)}(\pi)$, i.e., it is a probability distribution of $X_{K\cup L}$ and its marginal distribution for variables X_K equals π: $(\pi \triangleright \kappa)(x_K) = \pi(x_K)$.*

An importance of this operator arises from the fact that, when applied iteratively, it defines a multidimensional distribution from a system of low-dimensional ones.

[1] The concept of dominance (or absolute continuity) $\pi \ll \kappa$ in finite case simplifies to

$$\forall x \in \mathbf{X} \;\; (\kappa(x) = 0 \Longrightarrow \pi(x) = 0).$$

4 Generating sequences

Let us now consider a system of n low-dimensional distributions $\pi_1(x_{K_1}), \pi_2(x_{K_2}), \ldots, \pi_n(x_{K_n})$, and start studying a distribution $\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n$, which (if defined) is a distribution of variables $X_{K_1 \cup K_2 \cup \ldots \cup K_n}$. Regarding the fact that the operator is neither commutative nor associative, let us stress that we always apply the operators from left to right:

$$\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n = (\ldots((\pi_1 \triangleright \pi_2) \triangleright \pi_3) \triangleright \ldots \triangleright \pi_n).$$

Therefore, in order to construct a multidimensional distribution it is sufficient to determine a sequence – we call it a *generating sequence* – of low-dimensional distributions.

Example 1 *In agreement with what has just been said, the generating sequence*

$$\pi_1(x_1,x_3), \pi_2(x_3,x_5), \pi_3(x_1,x_4,x_5,x_6), \pi_4(x_2,x_5,x_6)$$

defines distribution

$$\begin{aligned}
&(\pi_1 \triangleright \pi_2 \triangleright \pi_3 \triangleright \pi_4)(x_1,x_2,x_3,x_4,x_5,x_6)\\
&\quad = \big((\pi_1(x_1,x_3) \triangleright \pi_2(x_3,x_5)) \triangleright \pi_3(x_1,x_4,x_5,x_6)\big) \triangleright \pi_4(x_2,x_5,x_6)\\
&\quad = \pi_1(x_1,x_3)\pi_2(x_5|x_3)\pi_3(x_4,x_6|x_1,x_5)\pi_4(x_2|x_5,x_6). \qquad \diamond
\end{aligned}$$

Not all generating sequences are equally efficient in their representations of multidimensional distributions. Among them, the so-called perfect sequences hold an important position.

Definition 2 *A generating sequence of probability distributions* $\pi_1, \pi_2, \ldots, \pi_n$ *is called* perfect *if for all* $k = 2, \ldots, n$ *distributions* $\pi_1 \triangleright \ldots \triangleright \pi_k$ *are defined and*

$$\pi_1 \triangleright \ldots \triangleright \pi_k = \pi_k \triangleright (\pi_1 \triangleright \ldots \triangleright \pi_{k-1}).$$

This definition enables us to check whether a generating sequence is perfect[2] but one can hardly see from it the importance of perfect sequences. This importance becomes clearer from the following characterization theorem (Theorem 2 in [7]).

Theorem 2 *A sequence of distributions* $\pi_1, \pi_2, \ldots, \pi_n$ *is perfect* iff *all the distributions from this sequence are marginals of the distribution* $(\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n)$.

What is the main message conveyed by this characterization theorem? Considering that low-dimensional distributions π_k are carriers of local information, the constructed multidimensional distribution represents global information, faithfully reflecting all of the local input.

Let us briefly summarize the main properties of distributions represented by perfect sequences and their relation to the well-known concepts of GMM.

[2] A sequence is perfect *iff* for all $k = 2, \ldots, n$, $(\pi_1 \triangleright \ldots \triangleright \pi_{k-1})^{(K_n \cap (K_1 \cup \ldots \cup K_{i-1}))} = \pi_k^{(K_n \cap (K_1 \cup \ldots \cup K_{i-1}))}$.

(i) It was shown that perfect sequences are equivalent to BNs in the sense that any distribution representable by a perfect sequence can be represented by BN (and vice versa) and both of these strucures are defined with the same number of parameters – probabilities (for details see [7]) .

(ii) In analogy to BN, for each distribution represented by a perfect sequence a list of conditional independence relations holds true. For a BN, one can read all these relations from its graph by the famous d-separation criterion. How to determine them for CM was shown in [8].

(iii) Let us stress that whether a generating sequence is perfect does not depend only on structural properties (those corresponding to sets $K_1,\ldots,K_n$ and their ordering), but also on probabilities. To make this remark clearer notice the two extreme sufficient conditions, guaranteeing perfecness of a generating sequence:

(a) if distributions $\pi_1(x_{K_1}),\ldots,\pi_n(x_{K_n})$ are pairwise consistent ($\pi_i^{(K_i\cap K_j)} = \pi_j^{(K_1\cap K_j)}$) and the sequence $K_1,\ldots,K_n$ meets the *running intersection property*[3] then $\pi_1(x_{K_1}),\ldots,\pi_n(x_{K_n})$ is perfect;

(b) if all the distributions $\pi_k(x_{K_k})$ are uniform then $\pi_1(x_{K_1}),\ldots,\pi_n(x_{K_n})$ is always perfect.

(iv) Distributions represented by perfect sequences are unique in the following sense: if two permutations $\pi_{i_1},\ldots,\pi_{i_n}$ and $\pi_{j_1},\ldots,\pi_{j_n}$ of a system of oligodimensional distributions are perfect then $\pi_{i_1} \triangleright \ldots \triangleright \pi_{i_n} = \pi_{j_1} \triangleright \ldots \triangleright \pi_{j_n}$. This property, somehow resembling decomposable distributions, is especially important for designing computational procedures.

(v) Notice that we have not imposed any conditions on sets K_k. For example, considering a generating sequence where one distribution is defined for a subset of variables of another distribution (ie., $K_j \subset K_k$) is fully sensible and may enrich a system of considered multidimensional distributions (cf. Algorithm in Section 6.3).

5 Information-theoretic notions

In Section 6 several notions characterizing probability distributions and their relationship will be used. The first is the well-known *Shannon entropy* defined (for $\pi \in \Pi^{(N)}$)

$$H(\pi) = -\sum_{x\in\mathbf{X}_N} \pi(x)\log\pi(x).$$

[3] $\forall k = 2,\ldots,n \quad \exists j (1 \le j < k) \quad K_k \cap (K_1 \cup \ldots \cup K_{k-1}) \subset K_j.$

Recall that for two disjoint index sets $K, L \subset N$ one can also define a *conditional entropy* $H(\pi(x_K|x_L)$ using the expression:

$$H(\pi(x_K|x_L)) \quad = \quad -\sum_{x \in \mathbf{X}_{K \cup L}} \pi(x) \log \pi(x_K|x_L).$$

To compare two distributions defined for the same system of variables (i.e. $\pi, \kappa \in \Pi^N$) we will use *Kullback-Leibler divergence* (in literature sometimes called I-divergence, or cross-entropy). It is in fact a relative entropy of the first distribution with respect to the other:

$$Div(\pi||\kappa) = \begin{cases} \sum\limits_{x \in \mathbf{X}_N} \pi(x) \log \frac{\pi(x)}{\kappa(x)} & \text{if } \pi \ll \kappa, \\ +\infty & \text{otherwise.} \end{cases}$$

The reader can immediately see that if $\pi = \kappa$ then $Div(\pi||\kappa) = 0$. It is a well-known property of Kullback-Leibler divergence (and not too difficult to be proven) that its value is always non-negative and equals 0 if and only if $\pi = \kappa$. (Recall also that this divergence is not symmetric, i.e., generally $Div(\pi||\kappa) \neq Div(\kappa||\pi)$.)

One of the fundamental notions of information theory is a *mutual information*. Having a distribution $\pi(x_N)$ and two disjoint subsets $K, L \subset N$, it expresses how much one group of variables X_K influences the other one – X_L. It is defined

$$MI_\pi(X_K; X_L) = \sum_{x_{K \cup L} \in \mathbf{X}_{K \cup L}} \pi(x_{K \cup L}) \log \frac{\pi(x_{K \cup L})}{\pi(x_K)\pi(x_L)},$$

and equals 0 if and only if the groups of variables X_K and X_L are independent under the distribution π. Otherwise, it is always positive.

The last notion, which will be of great importance, but which is not as famous as Shannon entropy or mutual information, is an *informational content* of a distribution defined by the formula:

$$I(\pi) = \sum_{x \in \mathbf{X}_N} \pi(x) \log \frac{\pi(x)}{\prod\limits_{j \in N} \pi(x_j)}.$$

Notice that this formula is nothing but a Kullback-Leibler divergence of two distributions: $\pi(x_N)$ and $\prod_{j \in N} \pi(x_j)$. Therefore, it is always non-negative and equals 0 if and only if $\pi(x_N) = \prod_{j \in N} \pi(x_j)$. In fact, this value expresses how much individual variables are dependent under the distribution π. Therefore the higher this value, the more dependent the variables, and consequently, the greater amount of information carried by the distribution.

One can also immediately see that for a 2-dimensional distribution $\pi(x_1, x_2)$

$$I(\pi) = MI_\pi(X_1; X_2).$$

6 Approximations

Let us consider an arbitrary multidimensional distribution $\kappa \in \Pi^{(N)}$ and assume that for one reason or another we are looking for its approximation in the form of a compositional model. Such situations appear quite often in practical problems; κ can be, for example, a sample distribution of a large database, or it can be an unknown theoretical distribution, from which some data file has been generated. In any case, we need its approximation.

Criterion function.

For a candidate compositional distribution $\pi = \pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n \in \Pi^{(N)}$, the Kull- back-Leibler divergence $Div(\kappa \| \pi)$ will be used as a criterion function. Naturally, the smaller the value of the Kullback-Leibler divergence, the better approximation π.

For compositional models this divergence can be expressed in a special form, which enables us to analyze individual factors of the divergence. To make the formulae more transparent we will use the following notation: for each $i = 1, \ldots, n$ set K_i is split into two disjoint parts

$$R_i = K_i \setminus (K_1 \cup \ldots \cup K_{i-1}), \quad S_i = K_i \cap (K_1 \cup \ldots \cup K_{i-1}).$$

(Naturally, $R_1 = K_1$ and $S_1 = \emptyset$.) In the following computations we shall use a standard trick, according to which

$$\begin{aligned}\sum_{x \in \mathbf{X}_N} \kappa(x) \log \kappa(x_K) &= \sum_{x_K \in \mathbf{X}_K} \kappa(x_K) \log \kappa(x_K) \sum_{x_{N \setminus K} \in \mathbf{X}_{N \setminus K}} \kappa(x_{N \setminus K} | x_K) \\ &= \sum_{x_K \in \mathbf{X}_K} \kappa(x_K) \log \kappa(x_K)\end{aligned}$$

because $\sum_{x_{N \setminus K} \in \mathbf{X}_{N \setminus K}} \kappa(x_{N \setminus K} | x_K) = 1$. Thus, assuming $Div(\kappa \| \pi)$ is finite, we can compute

$$\begin{aligned}Div(\kappa \| \pi) &= \sum_{x \in \mathbf{X}_N} \kappa(x) \log \frac{\kappa(x)}{\pi_1(x_{K_1}) \triangleright \ldots \triangleright \pi_n(x_{K_n})} \\ &= \sum_{x \in \mathbf{X}_N} \kappa(x) \log \kappa(x) - \sum_{x \in \mathbf{X}_N} \kappa(x) \log \prod_{i=1}^{n} \pi_i(x_{R_i} | x_{S_i}) \\ &= -H(\kappa) - \sum_{i=1}^{n} \sum_{x \in \mathbf{X}_N} \kappa(x) \log \pi_i(x_{R_i} | x_{S_i}) \\ &= -H(\kappa) - \sum_{i=1}^{n} \sum_{x_{K_i} \in \mathbf{X}_{K_i}} \kappa(x_{K_i}) \log \pi_i(x_{R_i} | x_{S_i})\end{aligned}$$

$$
\begin{aligned}
&= -H(\kappa) + \sum_{i=1}^{n} \sum_{x_{K_i}\in \mathbf{X}_{K_i}} \kappa(x_{K_i}) \log \frac{\kappa(x_{R_i}|x_{S_i})}{\pi_i(x_{R_i}|x_{S_i})} - \sum_{i=1}^{n}\sum_{x_{K_i}\in\mathbf{X}_{K_i}} \kappa(x_{K_i})\log\kappa(x_{R_i}|x_{S_i}) \\
&= -H(\kappa) + \sum_{i=1}^{n} Div(\kappa(x_{R_i}|x_{S_i})\|\pi_i(x_{R_i}|x_{S_i})) + \sum_{i=1}^{n} H(\kappa(x_{R_i}|x_{S_i})).
\end{aligned}
$$

Now, let us have a look at the meaning of the expression

$$\sum_{i=1}^{n} H(\kappa_i(x_{R_i}|x_{S_i})) - H(\kappa).$$

First, for each $i = 1,\ldots,n$ we get

$$
\begin{aligned}
H(\kappa_i(x_{R_i}|x_{S_i})) &= -\sum_{x_{K_i}\in\mathbf{X}_{K_i}} \kappa(x_{K_i})\log\kappa(x_{R_i}|x_{S_i}) \\
&= -\sum_{x_{K_i}\in\mathbf{X}_{K_i}} \kappa(x_{K_i})\log\frac{\kappa(x_{K_i})}{\kappa(x_{S_i})}\frac{\prod_{j\in K_i}\kappa(x_j)}{\prod_{j\in K_i}\kappa(x_j)} \\
&= -I(\kappa(x_{K_i})) + I(\kappa(x_{S_i})) + \sum_{j\in R_i} H(\kappa(x_j)).
\end{aligned}
$$

Since all sets R_i are mutually disjoint and their union is the whole set N we are getting

$$
\begin{aligned}
\sum_{i=1}^{n} H(\kappa_i(x_{R_i}|x_{S_i})) - H(\kappa) &= \sum_{i=1}^{n} \left(I(\kappa(x_{S_i})) - I(\kappa(x_{K_i}))\right) + \sum_{j\in N} H(\kappa(x_j)) - H(\kappa) \\
&= \sum_{i=1}^{n} \left(I(\kappa(x_{S_i})) - I(\kappa(x_{K_i}))\right) + I(\kappa).
\end{aligned}
$$

In this way we have deduced that

$$
\begin{aligned}
&Div(\kappa\|\pi) \\
&\quad = \sum_{i=1}^{n} Div(\kappa(x_{R_i}|x_{S_i})\|\pi_i(x_{R_i}|x_{S_i})) + \sum_{i=1}^{n}\left(I(\kappa(x_{S_i})) - I(\kappa(x_{K_i}))\right) + I(\kappa), \qquad (1)
\end{aligned}
$$

which is a result that is worth being formulated as a theorem.

Theorem 3 *Let a distribution $\kappa \in \Pi^{(N)}$ and a sequence of distributions $\pi_1(x_{K_1})$, $\pi_2(x_{K_2}),\ldots,\pi_n(x_{K_n})$, for which $\bigcup_{i=1}^{n} K_i = N$, be such that $Div(\kappa\|\pi_1 \triangleright \ldots \triangleright \pi_n)$ is finite. Then, denoting $\pi = \pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n$, for the Kullback-Leibler divergence $Div(\kappa\|\pi)$ the equation (1) holds true.*

So, the divergence of distributions κ and π consists of two parts. The first one

$$\sum_{i=1}^{n} Div(\kappa(x_{R_i}|x_{S_i})||\pi_i(x_{R_i}|x_{S_i}))$$

describes the "local" difference between κ and π (more precisely it renders the difference between conditional distributions $\kappa(x_{R_i}|x_{S_i})$ and $\pi_i(x_{R_i}|x_{S_i})$), and the second part

$$I(\kappa) - \sum_{i=1}^{n} \left(I(\kappa(x_{K_i})) - I(\kappa(x_{S_i}))\right)$$

describes the difference resulting from the application of a compositional model. As it will be shown below, in the case that $\kappa(x_{K_1}), \kappa(x_{K_2}), \ldots, \kappa(x_{K_n})$ is a perfect sequence, it is exactly a difference between the informational content of the distributions κ and $\kappa(x_{K_1}) \triangleright \ldots \triangleright \kappa(x_{K_n})$.

Corollary 1 *If for a distribution κ a generating sequence of its marginals $\kappa(x_{K_1})$, $\kappa(x_{K_2}), \ldots, \kappa(x_{K_n})$ is perfect then*

$$I(\kappa(x_{K_1}) \triangleright \kappa(x_{K_2}) \triangleright \ldots \triangleright \kappa(x_{K_n})) = \sum_{i=1}^{n} \left(I(\kappa(x_{K_i})) - I(\kappa(x_{S_i}))\right),$$

and therefore also

$$Div(\kappa||\kappa(x_{K_1}) \triangleright \ldots \triangleright \kappa(x_{K_n})) = I(\kappa) - I(\kappa(x_{K_1}) \triangleright \ldots \triangleright \kappa(x_{K_n})).$$

Proof. The first equation can immediately be obtained by substituting $\kappa(x_{K_1}) \triangleright \kappa(x_{K_2}) \triangleright \ldots \triangleright \kappa(x_{K_n})$ for both κ and π in equation (1), because then the Kullback-Leibler divergence must equal 0. The second one is a direct consequence of the first equality following from (1). □

Perfect sequence approximations.

Problem of model learning in context of CM means that one wants to find a properly ordered system of oligodimensional distributions. It is evident from the expression (1) that the best approximations are defined by generating sequences consisting of distributions which are marginals[4] of the approximated distribution κ. In this case, namely, for all $i = 1, \ldots n$, $Div(\kappa(x_{R_i}|x_{S_i})||\pi_i(x_{R_i}|x_{S_i}))$ equal 0 and the formula (1) simplifies to

$$Div(\kappa||\pi) = I(\kappa) - \sum_{i=1}^{n} \left(I(\kappa(x_{K_i})) - I(\kappa(x_{S_i}))\right), \tag{2}$$

which does not depend on values of distributions π_i (quite naturally, because they are marginals of κ) but only on the system, or more precisely sequence, $K_1, K_2, \ldots, K_n$. In the following example we shall show that different orderings of the distributions in generating sequences can result in different values of the Kullback-Leibler divergence.

[4] In fact, it is enough when all $\pi_i(x_{R_i}|x_{S_i}) = \kappa(x_{R_i}|x_{S_i})$.

Example 2 *Consider a* 4*-dimensional distribution* $\kappa(x_1,x_2,x_3,x_4)$ *and its three marginal distributions denoted* π_1,π_2,π_3*:*

$$\pi_1(x_1,x_2) = \kappa(x_1,x_2), \quad \pi_2(x_2,x_3) = \kappa(x_2,x_3), \quad \pi_3(x_3,x_4) = \kappa(x_3,x_4).$$

Compute $Div(\kappa\|\pi)$ *and* $Div(\kappa\|\hat{\pi})$ *for* $\pi = \pi_1 \triangleright \pi_2 \triangleright \pi_3$ *and* $\hat{\pi} = \pi_1 \triangleright \pi_3 \triangleright \pi_2$*. For the first distribution it is*

$$\begin{aligned} Div(\kappa\|\pi) &= I(\kappa) - \big(I(\kappa(x_{\{1,2\}})) + I(\kappa(x_{\{2,3\}})) + I(\kappa(x_{\{3,4\}}))\big) \\ &\qquad + \big(I(\kappa(x_\emptyset)) + I(\kappa(x_{\{2\}})) + I(\kappa(x_{\{3\}}))\big) \\ &= I(\kappa) - I(\kappa(x_{\{1,2\}})) - I(\kappa(x_{\{2,3\}})) - I(\kappa(x_{\{3,4\}})), \end{aligned}$$

whereas for $\hat{\pi}$ *we get*

$$\begin{aligned} Div(\kappa\|\hat{\pi}) &= I(\kappa) - \big(I(\kappa(x_{\{1,2\}})) + I(\kappa(x_{\{3,4\}})) + I(\kappa(x_{\{2,3\}}))\big) \\ &\qquad + \big(I(\kappa(x_\emptyset)) + I(\kappa(x_\emptyset)) + I(\kappa(x_{\{2,3\}}))\big) \\ &= I(\kappa) - I(\kappa(x_{\{1,2\}})) - I(\kappa(x_{\{3,4\}})) - I(\kappa(x_{\{2,3\}})) + I(\kappa(x_{\{2,3\}})) \\ &= Div(\kappa\|\pi) + I(\kappa(x_{\{2,3\}})). \end{aligned}$$

The reader probably noticed that, for the sake of simplicity, we introduced a situation corresponding to a decomposable model. It is perhaps worth mentioning that even in this case it may happen that both of the sequences defining distributions π *and* $\hat{\pi}$ *are perfect. In correspondence with the assertion mentioned in Section 4 (item* (iv)*), it happens only when* $\pi = \hat{\pi}$ *and therefore also* $Div(\kappa\|\pi) = Div(\kappa\|\hat{\pi})$*, from which we get that* $I(\kappa(x_{\{2,3\}})) = 0$*. This means that variables* X_2 *and* X_3 *are independent.* ◇

In the example we have shown that a quality of a compositional approximation depends not only on the selected system of low-dimensional distributions (possibly marginals of the approximated distribution) but also on their ordering. We could see that κ was better approximated by perfect sequence π_1,π_2,π_3 than by π_1,π_3,π_2, in case that the latter one was not perfect. From the following assertion we will see that perfect sequences are always, in a sense, the best approximations.

Theorem 4 *If* $\pi_1,\pi_2,\ldots,\pi_n$ *is a perfect sequence of marginal distributions of* κ *(*$\kappa \in \Pi^{(K_1\cup\ldots\cup K_n)}$*) then*

$$Div(\kappa\|\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n) \le Div(\kappa\|\pi_{i_1} \triangleright \pi_{i_2} \triangleright \ldots \triangleright \pi_{i_n})$$

for any permutation $i_1,i_2,\ldots,i_n$ *of indices* $1,2,\ldots,n$*.*

Proof. Since $\pi_1,\pi_2,\ldots,\pi_n$ is a perfect sequence of marginals of κ, we get from Corollary 1

$$Div(\kappa\|\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n)) = I(\kappa) - I(\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n),$$

and, because the Kullback-Leibler divergence is always nonnegative,

$$I(\kappa) \geq I(\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n).$$

We assume that $\pi_1, \pi_2, \ldots, \pi_n$ are marginals of κ, and since they form a perfect sequence (due to Theorem 2) they are also marginals of $\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n$. Therefore, equation (2) can be applied to both $Div(\kappa \| \pi_{i_1} \triangleright \ldots \triangleright \pi_{i_n}))$ and $Div(\pi_1 \triangleright \ldots \triangleright \pi_n \| \pi_{i_1} \triangleright \ldots \triangleright \pi_{i_n}))$:

$$Div(\kappa \| \pi_{i_1} \triangleright \ldots \triangleright \pi_{i_n})) = I(\kappa) - \sum_{\ell=1}^{n} \Big(I(\kappa(x_{K_{i_\ell}})) - I(\kappa(x_{S_{i_\ell}})) \Big), \tag{3}$$

$$Div(\pi_1 \triangleright \ldots \triangleright \pi_n \| \pi_{i_1} \triangleright \ldots \triangleright \pi_{i_n})) = I(\pi_1 \triangleright \ldots \triangleright \pi_n) - \sum_{\ell=1}^{n} \Big(I(\kappa(x_{K_{i_\ell}})) - I(\kappa(x_{S_{i_\ell}})) \Big).$$

The latter equality gives (respecting again the fact that the Kullback-Leibler divergence value must be nonnegative)

$$I(\pi_1 \triangleright \ldots \triangleright \pi_n) \geq \sum_{\ell=1}^{n} \Big(I(\kappa(x_{K_{i_\ell}})) - I(\kappa(x_{S_{i_\ell}})) \Big).$$

Combining this with equality (3) we get

$$Div(\kappa \| \pi_{i_1} \triangleright \ldots \triangleright \pi_{i_n})) \geq I(\kappa) - I(\pi_1 \triangleright \ldots \triangleright \pi_n),$$

where the right-hand side part of the inequality equals, as mentioned at the very beginning of the proof, $Div(\kappa \| \pi_1 \triangleright \ldots \triangleright \pi_n)$. □

Heuristic algorithm.

Regarding the above-mentioned fact that perfect sequence models are equivalent to Bayesian networks, it is obvious that all the methods for Bayesian network learning can be adapted to CM construction (see eg. [1]). Another very simple and effective possibility, though far from being optimal, is the process discussed in the rest of the paper.

We split the model construction process into two steps. The first one, which is not discussed in this paper, is selection of oligodimensional distributions, from which the model will be constructed. In some situations one can be quite naturally relieved of necessity to perform this step. For example, when the data file is too small and only 2-dimensional distributions can be estimated, then all these 2-dimensional distributions can be considered. In other situations, an expert can select the distributions from which the model should be constructed. Otherwise, informational content of low-dimensional distributions should be taken as a criterion for selection of a system of oligodimensional distributions.

The second step of the model construction process is to find a proper ordering of the selected oligodimensional distributions. The properties presented in the above sections theoretically support a heuristic algorithm, which arranges low-dimensional distributions into a generating sequence in a manner that utilizes its informational content as much as possible. In this section its simplest version is presented that enables the reader to understand the basic principle of exploiting the informational content of individual input low-dimensional distributions.

The reader will see that the procedure considers not only the given system of distributions but also their marginals; this can, in some situations, improve exploitation of the informational content of distributions, since it considers a greater variety of conditional independence structures.

Algorithm

```
Input: System of low-dimensional distributions $\pi_1(x_{K_1}),\ldots\pi_n(x_{K_n})$.

Initialization: Select a variable $X_m$ and a distribution $\pi_j$ such that $m \in K_j$.
        Set $\kappa_1 := \pi_j(x_m)$, $L := \{m\}$ and $k := 1$.

Computational Cycle: While $K_1 \cup \ldots \cup K_n \setminus L \neq \emptyset$ perform the following
        3 steps:

    1. for all $j = 1,\ldots,n$ and all $m \in K_j \setminus L$ compute the mutual
       information
                   $MI_{\pi_j}(X_m;X_{K_j \cap L})$.

    2. Fix $j$ and $m$ for which $MI_{\pi_j}(X_m;X_{K_j \cap L})$ achieved its maximal
       value.

    3. Increase $k$ by 1. Set $\kappa_k := \pi_j(X_{(K_j \cap L)\cup\{m\}})$ and $L := L \cup \{m\}$.

Output: Generating sequence $\kappa_1,\kappa_2,\ldots,\kappa_k$.
```

What can be said about the resulting generating sequence $\kappa_1,\kappa_2,\ldots,\kappa_k$? Distribution $\kappa^* = \kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_k$ is a probability distribution of $X_{K_1 \cup K_2 \cup \ldots \cup K_n}$. The goal of the algorithm is to get a distribution with the highest possible informational content $I(\kappa^*)$ (we know that the higher informational content, the lower the criterion function – Kullback-Leibler divergence). Important questions concern the facts whether the resulting sequence $\kappa_1,\kappa_2,\ldots,\kappa_k$ is perfect and contains all the distributions from $\pi_1,\pi_2,\ldots,\pi_n$. Unfortunately, answers to both these questions are negative.

Though the heuristics employed in the algorithm do not guarantee that a perfect sequence will always be found when it does exist, the advantage is in its efficiency and in the fact that it always suggests a subset of distributions that may form a perfect sequence, exploiting the available information in a suboptimal way[5]. One should realize, however, that a distribution from such a perfect

[5] Any generating sequence can be converted into a perfect sequence according to the following

sequence, though defined for groups of variables for which some input distribution π_j is defined, can differ from this input distribution π_j. In such a case, we employ a *process of verification and refinement*.

The detailed description of this process is beyond the scope (and extent) of this paper. Briefly said, verification consists of computation of Kullback-Leibler divergence of model distributions and the respective input distributions (or their marginals). If we find that some of the distributions defining the perfect model are too far from the required marginals (assuming that input distributions are marginals of the approximated distribution), then refinement is applied. This is realized by substituting a group of input marginal distributions by one distribution defined for all of the variables which are arguments of the deleted distributions. Naturally, this must be applied carefully, to avoid too much increase in the dimension of input distributions. New, more-dimensional input distributions are either estimated from data, or often computed from the original input distributions by the well-known Iterative Proportional Fitting Procedure ([3]). Then, having a new group of input distributions, the process starts from the very beginning by application of Algorithm.

The same process of verification and refinement is also applied when some of the input distributions are not included in the model.

Let us illustrate this process by a simple example.

Example 3 *Let us consider the following* 10 3*-dimensional distributions (their values were estimated from a data file):*

$$\begin{array}{lll}
\pi_1(x_1,x_2,x_4), & \pi_2(x_1,x_2,x_6), & \pi_3(x_1,x_4,x_6),\\
\pi_4(x_3,x_6,x_{11}), & \pi_5(x_3,x_{10},x_{11}), & \pi_6(x_4,x_6,x_{11}),\\
\pi_7(x_5,x_6,x_8), & \pi_8(x_6,x_8,x_{11}), & \pi_9(x_7,x_{10},x_{11}),\\
 & \pi_{10}(x_9,x_{10},x_{11}). &
\end{array}$$

The algorithm (starting with variable X_1 and distribution π_1) produced the sequence

$$\pi_1(x_1),\pi_1(x_1,x_4),\pi_3(x_1,x_4,x_6),\pi_6(x_4,x_6,x_{11}),\pi_8(x_6,x_8,x_{11}),\pi_4(x_3,x_6,x_{11}),$$
$$\pi_7(x_5,x_6,x_8),\pi_5(x_3,x_{10},x_{11}),\pi_{10}(x_9,x_{10},x_{11}),\pi_9(x_7,x_{10},x_{11}),\pi_1(x_1,x_2,x_4).$$

There are two points that can be made about this sequence. First, since all the distributions were estimated from one data file (with no missing values), all the distributions were pairwise consistent, and thus both $\pi_1(x_1)=\pi_3(x_1)$ and $\pi_1(x_1,x_4)=\pi_3(x_1,x_4)$, and therefore also

$$\pi_1(x_1)\triangleright\pi_1(x_1,x_4)\triangleright\pi_3(x_1,x_4,x_6)=\pi_3(x_1,x_4,x_6).$$

assertion ([5, 6]).
Theorem 5 *Let $\pi_1\triangleright\ldots\triangleright\pi_n$ be defined and let the sequence $\kappa_1,\ldots,\kappa_n$ be: $\kappa_1=\pi_1$, $\kappa_2=\kappa_1^{(K_2\cap K_1)}\triangleright\pi_2$, and generally $\kappa_j=(\kappa_1\triangleright\ldots\triangleright\kappa_{j-1})^{(K_j\cap(K_1\cup\ldots\cup K_{j-1}))}\triangleright\pi_j$. Then $\kappa_1,\ldots,\kappa_n$ is perfect and $\pi_1\triangleright\ldots\triangleright\pi_n=\kappa_1\triangleright\ldots\triangleright\kappa_n$.*

Therefore, the result of the algotihm was, in fact, a generating sequence

$$\pi_3, \pi_6, \pi_8, \pi_4, \pi_7, \pi_5, \pi_{10}, \pi_9, \pi_1,$$

which was perfect (see assertion (iiia) *in Section 4).*

The negative property of this result was the fact that the algorithm finished before exploiting distribution $\pi_2(x_1,x_2,x_6)$. *Since we are looking for an approximation of a distribution from which the data file was generated, (following the verification and refinement process) we have to assess how much omitting* π_2 *influences the quality of the achieved result. This is done by considering the Kullback-Leibler divergence* $Div(\pi_2(x_1,x_2,x_6)||\pi_{appr}(x_1,x_2,x_6))$, *for*

$$\pi_{appr} = \pi_3 \triangleright \pi_6 \triangleright \pi_8 \triangleright \pi_4 \triangleright \pi_7 \triangleright \pi_5 \triangleright \pi_{10} \triangleright \pi_9 \triangleright \pi_1$$

(let us mention that in this case $\pi_{appr}(x_1,x_2,x_6) = (\pi_3 \triangleright \pi_1)(x_1,x_2,x_6)$*). If we are not satisfied, refinement results in getting a distribution* $\pi_{11}(x_1,x_2,x_4,x_6)$ *and substituting it for* π_1, π_2 *and* π_3. *Subsequent application of the algorithm to the set of distributions* $\pi_4, \pi_5, \pi_6, \pi_7, \pi_8, \pi_9, \pi_{10}, \pi_{11}$ *resulted in obtaining the perfect sequence*

$$\pi_{11}, \pi_6, \pi_8, \pi_4, \pi_7, \pi_5, \pi_{10}, \pi_9.$$

◇

7 Conclusions

We have presented theoretical results showing that if an approximation of a probability distribution is looked for in a family of compositional distributions then the Kullback-Leibler divergence representing a quality of the approximation can be expressed as a sum of two contributions. The first one, which can easily be suppressed by considering only marginals of the approximated distribution, describes “local” differences, while the other one corresponds to the loss of information resulting from the compositional model (from introducing the respective conditional independence relations). This knowledge was exploited for designing a heuristic algorithm based on an effort to maximize informational content of the constructed approximation.

Let us conclude the paper by a brief comment advocating CMs. Based on de Cooman approach to conditionning [2], J. Vejnarová introduced the operator of composition also in possibility theory [11], which made it possible to extend the whole approach beyond probabilistic framework.

References

[1] R.R. Boukaert, *Bayesian belief networks – from construction to inference*, PhD. thesis, University of Utrecht (Netherlands), 1995.

[2] G. de Cooman, Possibility theory I – III *International Journal of General Systems* **25** (1997), pp. 291–371.

[3] W.E. Deming and F.F. Stephan, On a least square adjustment of a sampled frequency table when the expected marginal totals are known, *Ann. Math. Stat.* **11** (1940), pp. 427-444.

[4] F. V. Jensen, *Introduction to Bayesian Network*, UCL Press 1996.

[5] R. Jiroušek, Composition of probability measures on finite spaces. In: *Proc. of the 13th Conf. Uncertainty in Artificial Intelligence UAI’97* (D. Geiger and P. P. Shenoy, eds.), Morgan Kaufmann Publ., San Francisco, California, 1997, pp. 274-281.

[6] Jiroušek, R. Graph Modelling without Graphs. In *Proc. of the 17th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-based Systems IPMU’98* (B. Bouchon-Meunier, R.R. Yager, eds.), Editions E.D.K. Paris, 1998, pp. 809-816.

[7] R. Jiroušek, Marginalization in composed probabilistic models. In: *Proc. of the 16th Conf. Uncertainty in Artificial Intelligence UAI’00* (C. Boutilier and M. Goldszmidt eds.), Morgan Kaufmann Publ., San Francisco, California, 2000, pp. 301-308.

[8] Jiroušek, R. Detection of independence relations from persegrams. In: *Proceedings of the 9th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge Based Systems* (Bernadette Bouchon-Meunier, Ronald R. Yager (ed)), ESIA, Annecy, France, 2002, pp. 1261-1267.

[9] S.L. Lauritzen, *Graphical Models*. Clarendon Press, Oxford, 1996.

[10] A. Perez, ε-admissible simplification of the dependence structure of a set of random variables, *Kybernetika* **13** (1977), pp. 439–450.

[11] J. Vejnarová, Possibilistic independence and operators of composition of possibility measures. In *Prague Stochastics’98* (M. Hušková, J. Á. Víšek and P. Lachout eds.), JČMF 1998, pp. 575–580.

Radim Jiroušek is with the Institute of Information Theory and Automation of the Academy of Sciences of the Czech Republic. He is also a visiting Professor at Faculty of Management, Jindřichův Hradec, Czech Republic. E-mail: radim@utia.cas.cz

An Update on Generalized Information Theory

GEORGE J. KLIR
Binghamton University (SUNY), USA

Abstract

The purpose of this paper is to survey recent developments and trends in the area of generalized information theory (GIT) and to discuss some of the issues of current interest in GIT regarding the measurement of uncertainty-based information for imprecise probabilities on finite crisp sets.

Keywords

uncertainty, uncertainty-based information, generalized information theory

1 Introduction

The term “Generalized Information Theory” (GIT) was introduced in the early 1990s to name a research program whose objective was to develop a broader treatment of uncertainty-based information, not restricted to the classical notions of uncertainty [6]. In GIT, the primary concept is uncertainty, and information is defined in terms of uncertainty reduction.

The basic tenet of GIT is that uncertainty can be formalized in many different ways, each based on some specific assumptions. To develop a fully operational theory for some conceived type of uncertainty, we need to address issues at four levels:

- LEVEL 1 – we need to find an appropriate mathematical representation of the conceived type of uncertainty
- LEVEL 2 – we need to develop a calculus by which this type of uncertainty can be properly manipulated
- LEVEL 3 – we need to find a meaningful way of measuring the amount of relevant uncertainty in any situation formalizable in the theory
- LEVEL 4 – we need to develop methodological aspects of the theory

GIT is an outgrowth of two classical uncertainty theories. The older one, which is also simpler and more fundamental, is based on the notion of possibility. The newer one, which has been considerably more visible, is based on the notion of probability. Proper ways of measuring uncertainty in these classical theories were established, respectively, by Hartley [5] and Shannon [12]. Basic features of the theories are outlined in [8].

The various nonclassical uncertainty theories in GIT are obtained by expanding the conceptual framework upon which the classical theories are based. At this time, the expansion is two-dimensional. In one dimension, the formalized language of the classical set theory is expanded to a more expressive language of *fuzzy set theory*, where further distinctions are based on various special types of fuzzy sets [10]. In the other dimension, the classical (additive) measures theory [4] is expanded to a less restrictive *fuzzy measure theory* [14], within which further distinctions are made by using fuzzy measures with various special properties. This expanded conceptual framework is a broad base for formulating and developing various theories of imprecise probabilities.

The subject of this paper is to discuss some of the issues of current interest regarding the measurement of uncertainty for imprecise probabilities on finite crisp sets. The various issues of possible fuzzifications of imprecise probabilities and of imprecise probablities on infinite sets are not addressed here. To facilitate the discussion, some common characteristics of imprecise probabilities on finite crisp sets are introduced in Section 2.

2 Imprecise Probabilities: Some Common Characteristics

One of the common characteristics of imprecise probabilities on finite crisp sets is that evidence within each theory is fully described by a *lower probability function (or measure)*, $\underline{g}$, or, alternatively, by an *upper probability function* (or measure) $\underline{g}$. These functions are always regular fuzzy measures that are superadditive and subadditive [14], respectively, and

$$\sum_{x\in X}\underline{g}(\{x\})\leq 1,\sum_{x\in X}\overline{g}(\{x\})\geq 1. \tag{1}$$

In the various special theories of uncertainty, they possess additional special properties.

When evidence is expressed (at the most general level) in terms of an arbitrary closed and convex set $\mathcal{D}$ of probability distribution functions p on a finite set X, functions $\underline{g}_{\mathcal{D}}$ and $\overline{g}_{\mathcal{D}}$ associated with $\mathcal{D}$ are determined for each $A\in\mathcal{P}(X)$ by the formulas

$$\underline{g}_{\mathcal{D}}(A)=\inf_{p\in\mathcal{D}}\sum_{x\in A}p(x) \text{ and } \overline{g}_{\mathcal{D}}(A)=\sup_{p\in\mathcal{D}}\sum_{x\in A}p(x).$$

Since

$$\sum_{x\in A} p(x) + \sum_{x\notin A} p(x) = 1,$$

for each $p \in \mathcal{D}$ and each $A \in \mathcal{P}(X)$, it follows that

$$\overline{g}_{\mathcal{D}}(A) = 1 - \underline{g}_{\mathcal{D}}(\overline{A}). \tag{2}$$

Due to this property, functions $\underline{g}_{\mathcal{D}}$ and $\overline{g}_{\mathcal{D}}$ are called *dual* (or *conjugate*). One of them is sufficient for capturing given evidence; the other one is uniquely determined by (2). It is common to use the lower probability function $\underline{g}_{\mathcal{D}}$ to capture the evidence.

As is well known [2, 3], any given lower probability function $\underline{g}_{\mathcal{D}}$ is uniquely represented by a set-valued function $m_{\mathcal{D}}$ for which $m_{\mathcal{D}}(\emptyset) = 0$ and

$$\sum_{A\in P(X)} m_{\mathcal{D}}(A) = 1. \tag{3}$$

Any set $A \in \mathcal{P}(X)$ for which $m_{\mathcal{D}}(A) \neq 0$ is often called a *focal set*, and the family of all focal sets, $\mathcal{F}$, with the values assigned to them by function $m_{\mathcal{D}}$ is called a *body of evidence*. Function $m_{\mathcal{D}}$ is called a *Möbius representation* of $\underline{g}_{\mathcal{D}}$ when it is obtained for all $A \in \mathcal{P}(X)$ via the *Möbius transform*

$$m_{\mathcal{D}}(A) = \sum_{B|B\subseteq A} (-1)^{|A-B|} \underline{g}_{\mathcal{D}}(B), \tag{4}$$

where $|A - B|$ denotes the cardinality of the finite set $A - B$. The inverse transform is defined for all $A \in \mathcal{P}(X)$ by the formula

$$\underline{g}_{\mathcal{D}}(A) = \sum_{B|B\subseteq A} m_{\mathcal{D}}(B). \tag{5}$$

It follows directly from (2) that

$$\overline{g}_{\mathcal{D}}(A) = \sum_{B|B\cap A\neq\emptyset} m_{\mathcal{D}}(B). \tag{6}$$

for all $A \in \mathcal{P}(X)$.

Assume now that evidence is expressed in terms of a given lower probability function $\underline{g}$. Then, the set of probability distribution functions that are consistent with $\underline{g}$, $\mathcal{D}(\underline{g})$, which is always closed and convex, is defined as follows:

$$D(\underline{g}) = \{p(x)|x \in X, p(x) \in [0,1], \sum_{x\in X} p(x) = 1, \text{ and}$$
$$\underline{g}(A) \leq \sum_{x\in A} p(x) \text{ for all } A \in P(X)\}. \tag{7}$$

That is, each given function g̲ is associated with a unique set $\mathcal{D}$ and vice-versa.

3 Measures of Uncertainty

A *measure of uncertainty* of some conceived type in a given theory of imprecise probabilities is a functional, U, that assigns to each lower probability function in the theory a nonnegative real number. This number is supposed to measure, in an intuitively meaningful way, the amount of uncertainty of the considered type that is embedded in the lower probability function. To be acceptable as a measure of the amount of uncertainty, the functional U must satisfy several intuitively essential axiomatic requirements. Considering the most general level, when evidence is represented in terms of an arbitrary closed and convex set $\mathcal{D}$ of probability distribution functions p on finite set $X \times Y$, function U must satisfy the following requirements:

1. *Subadditivity*: $U(\mathcal{D}) \leq U(\mathcal{D}_X) + U(\mathcal{D}_Y)$, where
$$\mathcal{D}_X = \{p_X | p_X(x) = \sum_{y \in Y} p(x,y) \text{ for some } p \in \mathcal{D}\},$$
$$\mathcal{D}_Y = \{p_Y | p_Y(y) = \sum_{x \in X} p(x,y) \text{ for some } p \in \mathcal{D}\}.$$

2. *Additivity*: $U(\mathcal{D}) = U(\mathcal{D}_X) + U(\mathcal{D}_Y)$ if and only if $\mathcal{D}_X$ and $\mathcal{D}_Y$ are not interactive, which means that for all $A \in \mathcal{P}(X)$ and all $B \in \mathcal{P}(X)$, $m_{\mathcal{D}}(A \times B) = m_{D_X}(A) \cdot m_{D_Y}(B)$ and $m_{\mathcal{D}}(R) = 0$ for all $R \neq A \times B$.

3. *Monotonicity*: if $\mathcal{D} \subseteq \mathcal{D}'$, then $U(\mathcal{D}) \subseteq U(\mathcal{D}')$; and similarly for $\mathcal{D}_X$ and $\mathcal{D}_Y$.

4. *Range*: if uncertainty is measured in bits, then $U(\mathcal{D}) \in [0, log_2|X \times Y|]$, and similarly for $\mathcal{D}_X$ and $\mathcal{D}_Y$.

The requirement of subadditivity and additivity, as stated here, are generalized counterparts of the classical requirements of subadditivity and additivity for probabilistic and possibilistic measures of uncertainty. The requirement of monotonicity (not applicable to classical probabilistic uncertainty) means that reducing the set of probability distributions consistent with a given lower (or upper) probability function cannot increase uncertainty. The requirement of range, which depends on the choice of measurement units, is defined by the two extreme cases: the full certainty and the total ignorance.

When distinct types of uncertainty coexist in a given uncertainty theory, it is not necessary that these requirements be satisfied by each uncertainty type. However, they must be satisfied by an overall uncertainty measure, which appropriately aggregates measures of the individual uncertainty types.

It is well established that two types of uncertainty coexist in all theories of imprecise probabilities [8, 9]. They are generalized counterparts of the classical possibilistic and probabilistic uncertainties. They are measured, respectively, by appropriate generalizations of the Hartley and Shannon measures.

4 Generalized Hartley Measures

An historical overview of efforts to generalize the classical Hartley measure of uncertainty can be found in [9]. Its full generalization (to arbitrary closed and convex sets of probability distributions) was completed fairly recently by Abellan and Moral [1]. They showed that the functional

$$GH(m_{\mathcal{D}}) = \sum_{A \in \mathcal{F}} m_{\mathcal{D}}(A) \log_2 |A|, \tag{8}$$

where $m_{\mathcal{D}}$ is the Möbius representation of the lower probability associated with a given closed and convex set $\mathcal{D}$ of probability distributions, satisfies all the essential axiomatic requirements defined in Sec. 3 (subadditivity, additivity, etc.). Moreover, this functional is also directly connected with the classical Hartley measure: it is the weighted average of the Hartley measure for each given body of evidence $(\mathcal{F}, m_{\mathcal{D}})$.

It is fairly obvious that the functional *GH* defined by (8) measures the lack of specificity in evidence. Large focal elements result in less specific predictions, diagnoses, etc., than their smaller counterparts. The type of uncertainty measured by *GH* is thus well characterized by the term nonspecificity.

Observe that $GH(m_{\mathcal{D}}) = 0$ for precise probabilities, where $\mathcal{D}$ consists of a single probability distribution function, which is expressed in (8) by function $m_{\mathcal{D}}$. All focal sets are in this case singletons. Evidence expressed by precise probabilities is thus fully specific.

Eq. (8) is clearly applicable only to functions $m_{\mathcal{D}}$ defined on finite sets. It must be properly modified when $m_{\mathcal{D}}$ is defined on the n-dimensional Euclidean space for some $n \geq 1$, as shown in [9]. However, this modification is not a subject of this paper.

5 Generalized Shannon Measures

There have been many promising, but eventually unsuccessful efforts to generalize the classical Shannon measure (usually referred to as the *Shannon entropy*). Virtually all these efforts were based on the recognition that the Shannon entropy measures the mean (expected) value of the conflict among evidential claims expressed by a single probability distribution function on a finite set of mutually exclusive alternatives [9]. An historical overview of most of these efforts is given in [9].

All the proposed generalizations of the Shannon entropy were intuitively promising as measures of conflict among evidential claims in general bodies of evidence, but each of them was eventually found to violate the essential requirement of subadditivity. In fact, no generalized Shannon entropy can be subadditive on its own, as is shown in [13]. The subadditivity may be obtained only in terms of the total uncertainty — an aggregate of the two coexisting types of uncertainty (nonspecifivity and conflict). However, when the total uncertainty is viewed as the sum of the generalized Hartley measure with the various candidates for the generalized Shannon entropy, none of these aggregated uncertainty measures is still subadditive, as demonstrated by relevant counterexamples in each case [13].

The latest promising candidate (not previously analyzed in terms of the requirement of subadditivity) is based on the so-called Shapley index, which plays an important role in game theory [11, 15]. For any given finite universal set X, this candidate for the generalized Shannon entropy, GS, is defined as the average Shannon entropy of differences in a given lower probability (or, alternatively, an upper probability) for all maximal chains in the lattice $(\mathcal{P}(X),\subseteq)$. Unfortunately, the sum $GH+GS$ does not satisfy in this case again the requirement of subadditivity. This can be demonstrated by the following counterexample.

Let $X=[x_1,x_2]$ and $Y=[y_1,y_2]$, and let us consider a body of evidence on $X\times Y$ whose Möbius representation is:

$$\begin{aligned} m(\{(x_1,y_1),(x_2,y_2),(x_2,y_1)\}) &= a,\\ m(X\times Y) &= 1-a, \end{aligned}$$

where $a\in[0,1]$. Then, $m_X(X)=m_Y(Y)=1$, and, hence, $GS_X(m_X)=GS_Y(m_Y)=0$ and $GH_X(m_X)+GH_Y(m_Y)=2$. Furthermore,

$$\begin{aligned} GS(m) &= [-a\log_2 a-(1-a)log_2(1-a)]/4,\\ GH(m) &= alog_2 3+2-2a \end{aligned}$$

For subadditivity of $GH+GS$, the difference

$$\begin{aligned} \Delta &= (GH_X+GH_Y+GS_X+GS_Y)-(GH+GS)\\ &= [a\log_2 a+(1-a)log_2(1-a)]/4+2a-alog_2 3 \end{aligned}$$

is required to be nonnegative for all values $a\in[0,1]$. However, Δ is negative in this case for any value $a\in(0,0.58)$ and it reaches its minimum, $\Delta=-0.1$, at $a=0.225$.

6 Total Uncertainty Measures

Generalized Shannon measure, GS, was eventually defined indirectly, via an *aggregated uncertainty*, AU, covering both nonspecificity and conflict, and the well established generalized Hartley measure of nonspecificity, GH, defined by (8). Since it must be that $GH+GS=AU$, the generalized Shannon measure can be defined as

$$GS=AU-GH \tag{9}$$

Using this definition, the unsuccessful effort to find *GS* directly is replaced with the effort to find *AU* and define *GS* indirectly via Eq. (9). The latter effort was successful in the mid 1990s, when a functional *AU* satisfying all essential requirements was established in evidence theory [9]. However, this functional is applicable to all the other theories of imprecise probabilities as well, which follows from the common properties shared by these theories (Sec. 2). Given any lower probability function $\underline{g}_{\mathcal{D}}$ associated with a closed convex set $\mathcal{D}$ of probability distributions (or vice versa), $AU(\underline{g}_{\mathcal{D}})$ is defined by the formula

$$AU(\underline{g}_{\mathcal{D}}) = \max_{p \in D}[-\sum_{x \in X} p(x) \log_2 p(x)]. \tag{10}$$

It is the maximum Shannon entropy within $\mathcal{D}$. An efficient algorithm for computing this maximum, which was proven correct for belief functions of evidence theory [9], is applicable without any change when belief functions are replaced with arbitrary lower probability functions of any other kind.

Given an arbitrary lower probability function $\underline{g}$ on $\mathcal{P}(x)$, the generalized version of this algorithm consists of the following seven steps:

Step 1. Find a non-empty set $A \subseteq X$, such that $\underline{g}(A)/|A|$ is maximal. If there are more such sets than one, take the one with the largest cardinality.

Step 2. For all $x \in A$, put $p(x) = \underline{g}(A)/|A|$.

Step 3. For each $B \subseteq X - A$, put $\underline{g}(B) = \underline{g}(B \cup A) - \underline{g}(A)$.

Step 4. Put $X = X - A$.

Step 5. If $X \neq \emptyset$ and $\underline{g}(X) > 0$, then go to Step 1.

Step 6. If $\underline{g}(X) = 0$ and $X \neq \emptyset$, then put $p(x) = 0$ for all $x \in X$ and $m(X) = 1 - na$.

Step 7. Calculate $AU = -\sum_{x \in X} p(x) \log_2 p(x)$.

Although functional *AU* is a well-justified measure of total uncertainty in the various theories of uncertainty, it is highly insensitive to changes in evidence due to its aggregated nature. It is an aggregate of the two coexisting types of uncertainty, nonspecificity and conflict. It is thus desirable to express the total uncertainty, TU, in a disaggregated form

$$TU = (GH, GS), \tag{11}$$

where *GH* is defined by (8) and *GS* is defined by (9) and (10). It is assumed here that the axiomatic requirements are defined in terms of the sum of the two functionals involved, which is always the well-justified aggregate measure *AU*. In this sense the measure satisfies trivially all the requirements. Its advantage is that measures of both types of uncertainty that coexist in uncertainty theory employed (nonspecificity and conflict) are expressed explicitly and, consequently, the measure is sensitive to changes in evidence.

To appreciate the difference between AU and TU, let us consider three simple examples of given evidence within a finite universal set X and let $|X| = n$ for convenience: (i) in the case of total ignorance (when $m(X) = 1$), we obtain $AU = \log_2 n$ and $TU = (\log_2 n, 0)$; (ii) when evidence is expressed by the uniform probability distribution on X, then again we have $AU = \log_2 n$, but $TU = (0, \log_2 n)$; (iii) when evidence is expressed by $m(\{x\}) = a$ for all $x \in X$ and $m(X) = 1 - na$, then again $AU = \log_2 n$ for all values $a \leq 1/n$, while

$$TU = ((1 - na)\log_2 n, na\log_2 n).$$

It is clear that TU defined by (11) possesses all the required properties in terms of the sum of its components, since $GH + GS = AU$. Moreover, as proven by Smith [13], $GS \geq 0$ for all bodies of evidence. Additional properties of GS defined by (9) can be determined by employing the algorithm for computing AU, as shown for some properties in Section 7.

It is also reasonable to express the generalized Shannon entropy by the interval $[\underline{S}, \overline{S}]$, where $\underline{S}$ and $\overline{S}$ are, respectively, the minimum and maximum values of the Shannon entropy within the set of all probability distributions that are consistent with a given lower probability function. Clearly $\overline{S} = AU$ and $\underline{S}$ is defined by replacing max with min in Eq. (10). Then, the total uncertainty, TU', has the form

$$TU' = (GH, [\underline{S}, \overline{S}]). \tag{12}$$

Let us define a partial ordering of these total uncertainties as follows:

$$TU'_1 \leq TU'_2 \text{ iff } GH_1 \leq GH_2 \text{ and } [\underline{S}_1, \overline{S}_1] \subseteq [\underline{S}_2, \overline{S}_2].$$

Then, due to subadditivity of $\overline{S}$, subadditivity of TU' is guaranteed. Indeed,

$$[\underline{S}_X + \underline{S}_Y, \overline{S}_X + \overline{S}_Y] \not\subset [\underline{S}, \overline{S}]$$

for any joint and associated marginal bodies of evidence. However, no algorithm for computing $\underline{S}$ that has been proven correct is available as yet.

7 Some Properties of Generalized Shannon Entropy

The purpose of this section is to examine the generalized Shannon entropy defined by (9). To facilitate this examination, let

$$\mathcal{F} = \{A_i | A_i \in \mathcal{P}(X), i \in \mathbb{N}_q\}$$

denote the family of all focal sets of a given body of evidence, where $\mathbb{N}_q = \{1, 2, \ldots, q\}$ for some integer q, and let $m_i = m(A_i)$ for convenience. Moreover, let

$$E = \bigcup_{i \in \mathbb{N}_q} A_i.$$

The algorithm for computing $\overline{S}(= AU)$ produces a partition,

$$\mathcal{E} = \{E_k | k \in \mathbb{N}_r, r \leq q\}$$

of E. For convenience, assume that block E_k of this partition was produced in k-th iteration of the algorithm and let $e_k = |E_k|$. Then

$$\overline{S}(m) = -\sum_{k \in \mathbb{N}_r} \underline{g}_k \log_2(\underline{g}_k / e_k)$$

where $\underline{g}_k$ denotes the lower probability of E_k in k-th iteration of the algorithm, where $a_k = |A_k|$. This equation can be rewritten as

$$\overline{S}(m) = -\sum_{k \in \mathbb{N}_r} \underline{g}_k \log_2 \underline{g}_k + \sum_{k \in \mathbb{N}_r} \underline{g}_k \log_2 e_k.$$

It follows from this equation and from Eq. (9) that

$$GS(m) = S(\underline{g}_k | k \in \mathbb{N}_r) + GH(\underline{g}_k | k \in \mathbb{N}_r) - GH(m), \tag{13}$$

where S denotes the Shannon entropy.

Assume now that $\mathcal{F}$ consists of pair-wise disjoint focal sets. Then, the Möbius representation, m, is a positive function since any negative value m_i for some $A_i \in \mathcal{F}$ would clearly violate in this case the requirement that values of the associated lower probability function must be in $[0, 1]$. When applying the algorithm for computing $\overline{S}$ to our case, it turns out that the values m_i for all $A_i \in \mathcal{F}$ are uniformly distributed among elements of each focal set A_i. This only requires to prove that

$$\sum_{i \in I} m_i / \sum_{i \in I} a_i \leq m_k / a_k$$

for each $k \in I$ and all nonempty sets $I \subseteq \mathbb{N}_q$, where $a_k = |A_k|$. The proof of this inequality, which is omitted here due to limited space, can be obtained by the method of contradiction. The maximum entropy probability distribution function, p, for the given body of evidence is thus defined for all $x_{i_k} \in A_i (k \in \mathbb{N}_{|A_i|})$ and all

$A_i \in \mathcal{F}$. by the formula $p(x_{i_k}) = m_i/a_i$ where $a_i = |A_i|$. Hence,

$$\begin{aligned}\overline{S}(m) &= -\sum_{i=1}^{n}\sum_{k=1}^{a_i} p(x_{i_k})\log_2 p(x_{i_k})\\ &= -\sum_{i=1}^{n} m_i \log_2(m_i/a_i)\\ &= -\sum_{i=1}^{n} m_i \log_2 m_i + \sum_{i=1}^{n} m_i \log_2 a_i\\ &= -\sum_{i=1}^{n} m_i \log_2 m_i + GH(m).\end{aligned}$$

Consequently,

$$GS(m) = -\sum_{i=1}^{n} m_i \log_2 m_i.$$

This is clearly a property that we would expect, on intuitive grounds, the generalized Shannon entropy to satisfy.

To examine some properties of the generalized Shannon entropy for nested bodies of evidence, let $X = \{x_i | i \in \mathbb{N}_n\}$ and assume that elements of X are ordered in such a way that the family

$$\mathcal{A} = \{\mathcal{A}_i = \{x_1, x_2, \ldots, x_i\} | i \in \mathbb{N}_n\}$$

contains all focal sets. That is, $\mathcal{F} \subseteq \mathcal{A}$. For convenience, let $m_i = m(A_i)$ for all $i \in \mathbb{N}_n$.

To express $GS(m)$, we need to express $GH(m)$ and $\overline{S}(m)$. Clearly,

$$GH(m) = -\sum_{i=1}^{n} m_i \log_2 i \tag{14}$$

To express $\overline{S}(m)$, three cases must be distinguished in terms of values m_i:

(a) $m_i \geq m_{i+1}$ for all $i \in \mathbb{N}_{n-1}$;

(b) $m_i \leq m_{i+1}$ for all $i \in \mathbb{N}_{n-1}$;

(c) neither (a) nor (b).

Following the algorithm for computing $\overline{S}$, we obtain the formula

$$GS_a(m) = -\sum_{i=1}^{n} m_i \log_2(m_i i) \tag{15}$$

for any function m that conforms to Case (a). By applying the method of Lagrange multipliers, we can readily find out that the maximum, $GS_a^*(n)$, of this functional for some $n \in \mathbb{N}$ is obtained for

$$m_i = (1/i)2^{(-1/\ln 2+\alpha)}(i \in \mathbb{N}_n), \tag{16}$$

where the value of α is determined by solving the equation

$$2^{-(1/\ln 2+\alpha)} \sum_{i=1}^{n} (1/i) = 1.$$

Let $s_n = \sum_{i=1}^{n} (1/i)$. Then,

$$\alpha = -\log_2(1/s_n) - (1/\ln 2)$$

and, hence,

$$\begin{aligned} m_i &= (1/i)2^{\log_2(1/s_n)} \\ &= 1/(is_n). \end{aligned}$$

Substituting this expression for m_i in (15), we obtain

$$\begin{aligned} GS_a^*(n) &= \sum_{i=1}^{n} (1/i)(1/s_n)\log_2 s_n \\ &= [(1/s_n)\log_2 s_n] \sum_{i=1}^{n} (1/i). \end{aligned}$$

Consequently,

$$GS_a^*(n) = \log_2 s_n. \tag{17}$$

In Case (b), $\overline{S} = \log_2 n$ and GH is given by (8). Hence,

$$GS_b(m) = \log_2 n - \sum_{i=1}^{n} m_i \log_2 i.$$

The maximum, $GS_b^*(n)$, of this functional for some $n \in \mathbb{N}$ subject to the inequalities that are assumed in Case (b), is obtained for $m_i = 1/n$. Hence,

$$GS_b^*(n) = \log_2 \frac{n}{n!^{1/n}}. \tag{18}$$

Employing Stirling's formula for approximating $n!$, it can be shown that

$$\begin{aligned} \lim_{n\to\infty} \log_2 \tfrac{n}{n!^{1/n}} &= \log_2 e \\ &= 1.442695. \end{aligned}$$

GS_b^* is thus bounded, contrary to $GS_a^*(n)$. Moreover, $GS_b^*(n) < GS_a^*(n)$, for all $n \in \mathbb{N}$.

Case (c) is more complicated for a general analytic treatment since it covers a greater variety of bodies of evidence with respect to the computation of GS. This follows from the algorithm for computing $\overline{S}$. For each given body of evidence, the algorithm partitions the universal set in some way, and distributes the value of the lower probability in each block of the partition uniformly. For nested bodies of evidence, the partitions preserve the induced order of elements of X. There are 2^{n-1} order preserving partitions. The most refined partition and the least refined one are represented by Cases (a) and (b), respectively. All the remaining $2^{n-1}-2$ partitions are represented by Case (c). A conjecture, based on a complete analysis for $n=3$ and extensive simulation experiments for $n>3$, is that the maxima of GS for all these partitions are for all $n \in \mathbb{N}$ smaller than the maximum GS_a^* for Case (a). According to this plausible conjecture, whose proof is an open problem, the difference between the maximum nonspecificity, $GH^*(n)$, and maximum conflict, $GS_a^*(n)$, grows rapidly with n. For example, $GH^*(2) = 1$ and $GS_a^*(2) = 0.585$, while $GH^*(10^4) = 13.29$ and $GS_a^*(10^4) = 3.29$. Similarly, the maximum value of conflict is 36.9% of the maximum value of total uncertainty for $n = 2$, but it reduces to 19.8% for $n = 10^4$. For nested (consonant) bodies of evidence, this feature makes intuitively a good sense.

8 Conclusions

For the last two decades or so, research in GIT has been focusing on developing justifiable ways of measuring uncertainty and the associated uncertainty-based information in the various emerging uncertainty theories. This objective is now, by and large, achieved. However, some research in this direction is still needed to improve our understanding of the generalized Shannon entropy, defined either by (9) or by the interval $[\underline{S}, \overline{S}]$. Results presented in this paper are intended to contribute a little to this understanding.

In the years ahead, the focus of GIT will likely divide into two branches of research. One of them will focus on developing methodological tools based on our capability to measure uncertainty in the various established theories of uncertainty. Methodological tools for making the principles of uncertainty maximization, minimization, and invariance operational will in particular be sought due to the broad utility of these principles [7, 9]. The other branch of research will pursue the development of additional uncertainty theories. One direction in this research area will undoubtedly include a comprehensive investigation of the various ways of fuzzifying existing uncertainty theories.

References

[1] Abellán, J. and S. Moral, "A non-specificity measure for convex sets of probability distributions." *Intern. J. of Uncertainty, Fuzziness, and Knowledge-Based Systems*, **8**(3): 357–367, 1999.

[2] Chateauneuf, A. and Jaffray, J. Y., "Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion." *Mathematical Social Sciences*, **17**: 263–283, 1989.

[3] Grabisch, M. "The interaction and Möbius representations of fuzzy measures on finite spaces, k-additive measures: a survey." Grabish, M. et al., eds., *Fuzzy Measures and Integrals: Theory and Applications*. Springer-Verlag, New York, pp. 70–93.

[4] Halmos, P. R., *Measure Theory*. D. Van Nostrand, Princeton, NJ, 1950.

[5] Hartley, R. V. L., "Transmission of information." *The Bell System Technical J.*, **7**(3): 535–563, 1928.

[6] Klir, G. J., "Generalized information theory." *Fuzzy Sets and Systems*, **40**(1): 127–142, 1991.

[7] Klir, G. J., "Principles of uncertainty: What are they? Why do we need them?" *Fuzzy Sets and Systems*, **74**(1): 15–31, 1995.

[8] Klir, G. J., "Uncertainty-based information." Teodorescu, H. and P. Melo, eds, *Systemic Organization of Information in Fuzzy Systems*. IOS Press, Amsterdam, pp. 21–52, 2003.

[9] Klir, G. J. and Wierman M. J., *Uncertainty-Based Information: Elements of Generalized Information Theory* (Second Edition). Physica-Verlag/Springer-Verlag, Heidelberg and New York, 1999.

[10] Klir, G. J. and Yuan, B., *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice Hall, PTR, Upper Saddle River, NJ, 1995.

[11] Marichal, J. and Roubens, M., "Entropy of discrete fuzzy measures." *Intern. J. of Uncertainty, Fuzziness, and Knowledge-Based Systems*, **8**(6): 625–640, 2000.

[12] Shannon, C. E., "The mathematical theory of communication." *The Bell System Technical J.*, **27**(3&4): 379–423, 623–656, 1948.

[13] Smith, R. M., *Generalized Information Theory: Resolving Some Old Questions and Opening Some New Ones*. PhD. Dissertation, Binghamton University (SUNY), Binghamton, NY, 2000.

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

[15] Yager, R. R., “On the entropy of fuzzy measures.” *IEEE Trans. on Fuzzy Systems*, **8**(4): 453–461, 2000.

George J. Klir is with the Center for Intelligent Systems, Binghamton University (SUNY), Binghamton, New York, 13902 USA

Reducing Uncertainty by Imprecise Judgements on Probability Distributions: Application to System Reliability*

I.KOZINE
Risø National Laboratory, Denmark

V.KRYMSKY
Ufa State Aviation Technical University, Russia

Abstract

In this paper the judgement consisting in choosing a function that is believed to dominate the true probability distribution of a continuous random variable is explored. This kind of judgement can significantly increase precision in constructed imprecise previsions of interest, which of great importance for applications. New formulae for computing system reliability are derived on the basis of the technique developed.

Keywords

imprecise probabilities, probability density function, reliability

1 Introduction

Natural extension, a tool to extend statistical knowledge to other domains and to make a set of available statistical partial evidence coherent, can appear and be used in different forms. In [1] four equivalent forms of the natural extension were reported. They are all nothing other than properly stated optimisation problems for obtaining lower and upper coherent bounds of probability characteristics of interest. The primal form suggests seeking coherent bounds defined by a set of feasible probability distributions, and this set, in turn, is formed by the available evidence expressed as constraints in the optimisation task. If no evidence is available (the state of complete ignorance), then the solution is sought over the set of all possible probability distributions, which brings us to the vacuous probability of the event of interest A, i.e., $P(A) \in [0,1]$. The crux of such optimisation problems

*Participation of V. Krymsky in this work has been supported by NATO grant #26-02-0001.

is that their solutions are defined on the family of degenerate probability distributions[1], which are included on equal footing in the set of all possible probability distributions. As proven in [1], solving these optimisation problems, on the set of all possible probability distributions, gives the same solution as that obtained on only the set of degenerate distributions. This issue is closely related to the central theorems and methods of Chebychev systems as described in [2]. All this would simply be mathematical subtlety, that is, far from practitioners’ interest, if this did not give us a clue for deriving more precise previsions of interest for continuous random variables. For these 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. The existence of solutions on degenerate distributions often results in high imprecision, negating the pragmatic value of the assessments of interest. For example, in reliability applications the time to failure of a system/component can not admit (except for very special cases) the concentration of probability masses in a very few points of the positive real line. Not being able to utilise such evidence leads to the fact that imprecision in the reliability of a system grows rapidly as the number of components in the system increases, making the results rather practically useless [3].

This feature of the natural extension was found disturbing and precluded wider implementation of imprecise statistical reasoning into reliability analysis. An attempt to mitigate the influence of degenerate probability distributions on the solutions was undertaken in [4]. No significant effect was attained 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 stays the same.

In this paper we explore a more drastic and, as it will be demonstrated, effective way to exclude the family of degenerate distributions from the set of probability distributions, which, as was expected, results in more precise previsions of interest. This is attained through judgements on a value (or a function, in general) that dominates the probability density function $\rho(x)$ of a continuous random variable X. That is, we introduce judgements of the form $\rho(x) \leq \Psi(x)$, where $\Psi(x)$ is a real-valued positive function satisfying the inequalities $1 \leq \int\limits_{R_+} \Psi(x)dx < \infty$, and demonstrate a way of their utilisation. In particular, $\Psi(x)$ can be set as $\Psi(x) = K \cdot I_{[a,b]}(x)$ where $a, b \in R_+$ and $a \leq b$, $I_{[a,b]}(x)$ is the indicator function such that $I_{[a,b]}(x) = 1$ if $x \in [a,b]$, and $I_{[a,b]}(x) = 0$ otherwise, and $K \geq (b-a)^{-1}$ is a constant.

Similar ideas of utilising bounds on density functions were explored in [5]. The tool of their utilisation was dynamic programming which gives us numerical

[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

solutions of the stated problems, while we suggest an approach to solving the problems analytically.

Breaking down a multidimensional case, $\boldsymbol{X}$=$(X_1,\dots,X_n)$, provides a theoretical basis for system reliability computations, which is a subject of the second part of the paper.

2 Relevant basics of the approach

Comprehensive coverage of the foundation of the theory of imprecise previsions can be found in the books [6] and [7]. In this section we briefly describe only those concepts that are necessary to understand the approach developed.

Consider a system consisting of n components. Let $f_{ij}(x_i)$ be $j-$th function of the i-th component lifetime x_i, $i=1,\dots,n$, and $j=1,\dots,m_i$. Suppose that reliability characteristics of the components are not known precisely and represented as a set of lower and upper previsions $\underline{a}_{ij}=\underline{M}(f_{ij}(x_i)),\overline{a}_{ij}=\overline{M}(f_{ij}(x_i)),i=1,\dots,n$, and $j=1,\dots,m_i$, which means that there exist m_i interval-valued judgements for the i-th component formally represented as expected values. The functions $f_{ij}(x_i)$ can be regarded as gambles, where a gamble is a real-valued function on a possibility space whose value is uncertain [6]. If, for instance, $f_{ij}(x_i)=x$, then the lower prevision $\underline{a}_{ij}$ is the lower bound of the mean time to failure of the i-th component; or if $f_{ij}(x_i)=I_{[t,\infty)}(x_i)$, where $I_{[t,\infty)}(x_i)$=1 if $x_i\in[t,\infty)$ and $I_{[t,\infty)}(x_i)$=0 otherwise, then the lower prevision $\underline{a}_{ij}$ is the lower bound of the probability of a failure occurrence within $[t,\infty)$.

Denote $\boldsymbol{X}$=$(X_1,\dots,X_n)$ a random vector and $x=(x_1,\dots,x_n)$ is the vector of numerical values for X_1,...,X_n. Then, there exists a function $g(\boldsymbol{X})$ of the component lifetimes that characterises the system's reliability. The function $g(\boldsymbol{X})$ is also a gamble.

In order to compute the coherent lower and upper previsions $\underline{M}(g)$ and $\overline{M}(g)$ of interest characterising the system reliability, a proper optimisation problem (also referred to as the natural extension in its primal form) can be posed

$$\underline{M}(g)\langle\overline{M}(gt)\rangle=\inf_{\Re^n}\left\langle\sup_{\Re^n}\right\rangle\int\limits_{R_+^n}g(x)\rho(x)dx \tag{1}$$

subject to

$$\left.\begin{array}{l} 0\le\rho(x),\int\limits_{R_+^n}\rho(x)dx=1,\\ \underline{a}_{ij}\le\int\limits_{R_+^n}f_{ij}(x_i)\rho(x)dx\le\overline{a}_{ij},i=1,...,n,j=1,...,m_i. \end{array}\right\} \tag{2}$$

Here the minimum and maximum are taken over the set $\Re^n$ of all possible n-dimensional density functions $\{\rho(\boldsymbol{x})\}$ satisfying conditions (2). That is, each constraint in (2) is associated with a subset of $\Re^n$ and the intersection of those subsets, if not empty, defines the solutions of the above optimisation problems. If the initial interval-valued data, forming the constraints, are not consistent, then some of the subsets of $\Re^n$ associated with the constraints are disjoint and the solution does not exist. The requirement of the existence of a non-empty set of probability distributions associated with the set of constraints is the only consistency principle imposed on the initial interval-valued data. This requirement is equivalent to the principle of avoiding sure loss [6] and is easily subject to technical checks.

If the components of a system are independent, then $\rho(\boldsymbol{x}) = \rho_1(x_1),\ldots,\rho_n(x_n)$.

In some cases the duals of optimisation problems (1)-(2) can be stated, which makes it technically easy to solve them [1]. The duals of (1)-(2) are

$$\underline{M}(g) = \sup_{c_0,c_{ij},d_{ij}} \left(c_0 + \sum_{i=1}^{n} \sum_{j=1}^{m_i} (c_{ij}\underline{a}_{ij} - d_{ij}\overline{a}_{ij}) \right), \tag{3}$$

subject to $c_0 \in R,\ c_{ij}, d_{ij} \in R_+$ and for any $x_i \geq 0$, $i = 1,2,...,n, \quad j = 1,2,...,m_i$,

$$c_0 + \sum_{i=1}^{n} \sum_{j=1}^{m_i} (c_{ij} - d_{ij}) I_{[t,\infty)}(x_i) \leq g(x). \tag{4}$$

And

$$\overline{M}(g) = \inf_{c_0,c_{ij},d_{ij}} \left(c_0 + \sum_{i=1}^{n} \sum_{j=1}^{m_i} (c_{ij}\overline{a}_{ij} - d_{ij}\underline{a}_{ij}) \right), \tag{5}$$

subject to $c_0 \in R, c_{ij}, d_{ij} \in R_+$ and for any $x_i \geq 0$, $i = 1,2,...,n$, $j = 1,2,...,m_i$,

$$c_0 + \sum_{i=1}^{n} \sum_{j=1}^{m_i} (c_{ij} - d_{ij}) I_{[t,\infty)}(x_i) \geq g(x). \tag{6}$$

The validity of the transition from a primal form similar to (1)-(2) to the dual form is explained in [1], [8].

Problems (3)-(4) and (5)-(6) are linear optimisation problems and and have technically straightforward solutions.

In some cases dual problems do not exist. This takes place if a primal optimisation problem is not linear. For example, the judgement of independence among system components, which is equivalent to the introduction of $\rho(\boldsymbol{x}) = \rho_1(x_1),\ldots,\rho_n(x_n)$, makes the problem non-linear, and, as a consequence, it leads to the non-existence of the dual optimisation problem.

3 Extending knowledge: one-dimensional case

Let us consider first a one-dimensional case of extending partial statistical information to probability characteristics of interest. That is, we will be focusing in this section on the construction of new imprecise characteristics provided some other imprecise statistical characteristics are known on the same possibility set, and, more important, we will demonstrate how "soft" judgements on the probability density function of a random variable can be modelled and utilised in the framework of the theory of imprecise probabilities.

Assume that there are m interval-valued judgements on probability characteristics on a specific possibility set, i.e. $M\left(f_i(X)\right) \in [\underline{a}_i, \overline{a}_i]$, and there is an additional judgement of $\rho(x) \leq \Psi(x)$, $1 \leq \int\limits_{R_+} \Psi(x)dx < \infty$. The objective is to extend this evidence to the prevision of interest $M\left(g(X)\right)$ that cannot be found precisely, as the initial data are partial.

Write the primal form of natural extension

$$\underline{M}(g)\left\langle \overline{M}(g) \right\rangle = \inf_{\Re}\left\langle \sup_{\Re} \right\rangle \int\limits_{R_+} g(x)\rho(x)dx \tag{7}$$

subject to

$$\left.\begin{array}{l} 0 \leq \rho(x) \leq \Psi(x), \int\limits_{R_+} \Psi(x)dx = H < \infty, \int\limits_{R_+} \rho(x)dx = 1 \text{ and} \\ \underline{a}_i \leq \int\limits_{R_+} f_i(x)\rho(x)dx \leq \overline{a}_i, i = 1, 2, ..., m. \end{array}\right\} \tag{8}$$

The dual of the above optimisation problem cannot be straightforwardly written. First, introduce a new variable $z(x)$ instead of $\rho(x)$

$$z(x) = \frac{\Psi(x) - \rho(x)}{H - 1},$$

and denote $\Gamma = \int\limits_{R_+} g(x)\Psi(x)dx$; $\Phi_i = \int\limits_{R_+} f_i(x)\Psi(x)dx$, $i = 1, 2, ..., m$.

It is clear that $\int\limits_{R_+} z(x)dx = 1$. Then, optimisation problem (7)-(8) can be rewritten

$$\begin{aligned} \underline{M}(g)\left\langle \overline{M}(g) \right\rangle &= \inf_{\Re}\left\langle \sup_{\Re} \right\rangle \int\limits_{R_+} g(x)\rho(x)dx = \\ &= \Gamma - (H-1)\sup_{Z}\left\langle \inf_{Z} \right\rangle \left\{ \int\limits_{R_+} g(x)z(x)dx \right\} \end{aligned} \tag{9}$$

subject to

$$\left.\begin{array}{c} 0 \leq z(x), \int\limits_{R_+} z(x)dx = 1, \frac{\Phi_i - \overline{a}_i}{H-1} \leq \int\limits_{R_+} f_i(x)z(x)dx \leq \frac{\Phi_i - \underline{a}_i}{H-1}, \\ i = 1, ..., m. \end{array}\right\} \tag{10}$$

And finally, the challenge is to solve the following problems

$$\underline{s}(g)\left\langle \overline{s}(g) \right\rangle = \inf_{Z}\left\langle \sup_{Z} \right\rangle \int\limits_{R_+} g(x)z(x)dx, \tag{11}$$

subject to (10).

Before we go on to the duals, one consistency condition must be fulfilled. It is of avoiding sure loss [6] and is transparent from the stand of common sense and can be written as $\inf(f(x)) \leq \underline{M}(f(x)) \leq \overline{M}(f(x)) \leq \sup(f(x))$. Applied to objective functions (9), it appears as

$$\inf(g) \leq \Gamma - (H-1)\sup_Z \left\{ \int\limits_{R_+} g(x)z(x)dx \right\} \leq$$
$$\Gamma - (H-1)\inf_Z \left\{ \int\limits_{R_+} g(x)z(x)dx \right\} \leq \sup(g)$$

Optimisation problems (11) subject to (10) have their duals

$$\underline{s}(g) = \sup_{c_0,c_i,d_i} \left\{ c_0 + \sum_{i=1}^{m} \left[c_i \left(\frac{\Phi_i - \overline{a}_i}{H-1} \right) - d_i \left(\frac{\Phi_i - \underline{a}_i}{H-1} \right) \right] \right\} \tag{12}$$

subject to $c_0 \in R,\ c_i, d_i \in R_+$ and for any $x \geq 0\ c_0 + \sum\limits_{i=1}^{m} (c_i - d_i) f_i(x) \leq g(x)$. And

$$\overline{s}(g) = \inf_{c_0,c_i,d_i} \left\{ c_0 + \sum_{i=1}^{m} \left[c_i \left(\frac{\Phi_i - \underline{a}_i}{H-1} \right) - d_i \left(\frac{\Phi_i - \overline{a}_i}{H-1} \right) \right] \right\} \tag{13}$$

subject to $c_0 \in R,\ c_i, d_i \in R_+$ and for any $x \geq 0\ c_0 + \sum\limits_{i=1}^{m} (c_i - d_i) f_i(x) \geq g(x)$.

Thus, having derived the dual optimisation problems (12) and (13), we have got a tool for utilising “soft” judgements concerning probability density functions and extending them to other probability characteristics of interest defined on a one-dimensional possibility set.

Example 1. The information concerning a continuous random variable X is that of $\rho(x) \leq \Psi(x) = K \cdot I_{[0,T]}(x) < \infty$, where T, K are fixed positive numbers. What are the bounds for the expectation $M(X)$?

The above approach brings us to the following results

$$\underline{M}(X) = \frac{KT^2}{2} - (KT-1)T = T\left(1 - \frac{KT}{2}\right) \text{ and } \overline{M}(X) = \frac{KT^2}{2}.$$

Example 2. Assume now that besides the information stated in example 1 we know precisely the probability $P\{\underline{a} \leq X \leq \overline{a}\} = p$, where $0 \leq \underline{a} < \overline{a} \leq T$. How would the given information change the bounds for the expectation $M(X)$?

The result is

$$\underline{M}(X) = T\left(1 - \tfrac{KT}{2}\right) + (T - \overline{a})\left[K(\overline{a} - \underline{a}) - p\right],$$
$$\overline{M}(X) = \tfrac{KT^2}{2} - \underline{a}\left[K(\overline{a} - \underline{a}) - p\right].$$

4 Computation of system reliability

Extending knowledge on multidimensional possibility sets, taking into account imprecise judgements on probability density functions, is undertaken in a similar way to the one-dimensional case described above. The multidimensional case is broken down in detail in [9]. In this section we represent the results concerning system reliability computations that follow from this case.

As it has been found earlier (see elsewhere [3], [10], [11]), the reliability of a system, P_{Series}, the components of which are connected in series given the lower and upper bounds of the components' reliabilities and the state of complete ignorance concerning their dependence, is calculated according to the formulae

$$\underline{P}_{Series} = \underline{M}\left(I_{[t,\infty)}(\min_i x_i)\right) = \max\left(0; \sum_{i=1}^{n} \underline{p}_i - (n-1)\right),$$

and

$$\overline{P}_{Series} = \overline{M}\left(I_{[t,\infty)}(\min_i x_i)\right) = \min_i \overline{p}_i,$$

where $\underline{P}_{Series} \leq P_{Series} \leq \overline{P}_{Series}$, and $\underline{p}_i$ and $\overline{p}_i, i = 1, ..., n$ are the lower and upper reliabilities of the components.

By applying the above described approach, the formulas for the calculation of the reliability of series systems become updated in the light of the evidence concerning the probability density function of time to failure

$$\underline{P}_{Series} = \Gamma - (H-1)\min_i\left(\frac{\Phi_i - \underline{a}_i}{H-1}\right) = \Gamma - \min_i(\Phi_i - \underline{a}_i),$$

$$\begin{aligned}\overline{P}_{Series} &= \Gamma - (H-1)\max\left(0; \sum_{i=1}^{n}\left(\tfrac{\Phi_i - \overline{a}_i}{H-1}\right) - (n-1)\right) = \\ &= \Gamma - \max\left(0; \sum_{i=1}^{n}(\Phi_i - \overline{a}_i) - (H-1)\cdot(n-1)\right).\end{aligned}$$

The reliability of a system, $P_{Parallel}$, the components of which are connected in parallel given the lower and upper bounds of the components' reliabilities and the state of ignorance concerning their independence, is calculated according to the formulas (see elsewhere [3], [10], [11])

$$\underline{P}_{Parallel} = \underline{M}\left(I_{[t,\infty)}(\max_i x_i)\right) = \max_i \underline{p}_i,$$

$$\overline{P}_{Parallel} = \overline{M}\left(I_{[t,\infty)}(\max_i x_i)\right) = \min\left(1; \sum_{i=1}^{n} \overline{p}_i\right)$$

Their update in the light of the new evidence appears as follows

$$\underline{P}_{Parallel} = \Gamma - (H-1)\min\left(1; \sum_{i=1}^{n}\left(\tfrac{\Phi_i - \underline{a}_i}{H-1}\right)\right) =$$
$$= \Gamma - \min\left((H-1); \sum_{i=1}^{n}\left(\Phi_i - \underline{a}_i\right)\right),$$

and

$$\overline{P}_{Parallel} = \Gamma - (H-1)\max_i\left(\frac{\Phi_i - \overline{a}_i}{H-1}\right) = \Gamma - \max_i(\Phi_i - \overline{a}_i).$$

For a system of an arbitrary structure the reliability bounds satisfy the inequalities [3], [10], [11]:

$$\underline{P}_{ArbStruct} \geq \max_{1\leq j\leq r}\max(0, L_j),$$

where r is a number of system minimal paths $\pi_1, \pi_2, ..., \pi_r$, $L_j = \sum\limits_{i\in\pi_j}\underline{p}_i - (\mu_j - 1)$, and μ_j is the number of components belonging to path π_j, and

$$\overline{P}_{ArbStruct} \leq \min_{1\leq j\leq s}\min\left(\sum_{i\in \mathrm{K}_j}\overline{p}_i; 1\right),$$

where s is a number of system minimal cut sets denoted by $\mathrm{K}_1, \mathrm{K}_2, ..., \mathrm{K}_s$.

Now by applying the approach developed and using the substitutions $\overline{p}_i = \frac{\Phi_i - \underline{a}_i}{H-1}$, $\underline{p}_i = \frac{\Phi_i - \overline{a}_i}{H-1}$, we obtain

$$\underline{P}_{ArbStruct} \geq \Gamma - \min_{1\leq j\leq s}\min\left(\sum_{i\in \mathrm{K}_j}(\Phi_i - \underline{a}_i); (H-1)\right),$$
$$\overline{P}_{ArbStruct} \leq \Gamma - \max_{1\leq j\leq r}\max(0, L_j^*),$$

where $L_j^* = \sum\limits_{i\in\pi_j}(\Phi_i - \overline{a}_i) - (H-1)\cdot(\mu_j - 1)$.

Example 3. A system consists of two components (n=2) connected in series, and the reliability of the first component is $p_1 \in [\underline{a}_1, \overline{a}_1]$ and the second is $p_2 \in [\underline{a}_2, \overline{a}_2]$. One more judgement is of the form $\rho(x_1, x_2) \leq \Psi(x_1, x_2) = K \cdot I_{\{[0,T];[0,T]\}}(x_1, x_2)$, where K and$T$ are constants and $I_{\{[0,T];[0,T]\}}(x_1, x_2)$ is a two-dimensional indicator function. What is system reliability?

The reliabilities of the components for an arbitrary time $t$are to be written in the form (2)

$$\underline{a}_1 \leq \int_0^T I_{[t,T]}(x_1)\int_0^T \rho(x_1, x_2)dx_1dx_2 \leq \overline{a}_1,$$
$$\underline{a}_2 \leq \int_0^T I_{[t,T]}(x_2)\int_0^T \rho(x_1, x_2)dx_1dx_2 \leq \overline{a}_2.$$

Note that in this case $H = KT^2$, $\Gamma = K(T-t)^2$, hence

$$\begin{aligned}
&\underline{P}_{Series} = \Gamma - \min_i(\Phi_i - \underline{a}_i) = \Gamma - \int_t^T\int_0^T \Psi(x_1,x_2)dx_1dx_2 + \max_i(\underline{a}_i) = \\
&= \max_i(\underline{a}_i) - Kt(T-t); \\
&\overline{P}_{Series} = \Gamma - \max\left(0; \sum_{i=1}^{n}(\Phi_i - \overline{a}_i) - (H-1)\cdot(n-1)\right) = \\
&= \min\left(K(T-t)^2; (Kt^2 + \sum_{i=1}^{2}\overline{a}_i - 1)\right).
\end{aligned}$$

5 Concluding remarks

Judgements concerning the function $\Psi(x)$, which is believed to dominate the true probability distribution of a continuous variable, are practically elicitable and may be unambiguously understood by those inexperienced in probabilistic reasoning. So, a sample probability density function is defined by the totality of the values $\rho_i = {}^{n_i}\!/_{(N\Delta x)}$, $i = 1, 2, ...$, where n_i is the number of observed realisations of a continuous random variable X falling in the $i-$th bin with a width of Δx, and N is the size of the sample. For example, in reliability analysis the continuous random variable is time to failure or time between failures, and usually reliability characteristics are counted for a time period of 1 year. That is, the width of the bins is equal to 1 year for any i except for the last bin which is an open interval [x_k, ∞). As a matter of fact, any reliability calculation and failure reporting systems are scaled to one-year assessments so that the experts in the field are used to think of reliability characteristics as values scaled to a year. A question of "what would be the maximum percentage of failures per year for a specified component over its lifetime?" or alike would be quite easy to answer for an expert or to assess based on even scarce failure evidence.

References

[1] Utkin L. and Kozine I. Different faces of the natural extension. In: *Proc. of the Second International Symposium on Imprecise Probabilities and Their Applications, ISIPTA '01*, 316-323, 2001.

[2] Karlin S. and Studden W.J. Tchebycheff systems: With applications in analysis and statistics. *Interscience Publishers. A division of John Willey & Sons*, (1966).

[3] Kozine I. Prior reliability assessments based on coherent imprecise probabilities. *Int. J. General Systems*, 30(3):283-307, 2001.

[4] Utkin L. Imprecise calculation with the qualitative information about probability distributions. In *Proc. of the conference on Soft Methods in Probability and Statistics. Eds. P.Grzegorzewski, O.Hryniewicz and M.A.Gil*, 164-169, 2002.

[5] Smith E.J. Generalized Chebychev inequalities: theory and applications in decision analysis. *Operations Research*, Vol. 43(5) (1995), pp. 807-825.

[6] Walley P. Statistical reasoning with imprecise probabilities. *Chapman and Hall*, 1991.

[7] Kuznetsov V. Interval statistical models. *Radio and Sviaz*, 1991(in Russian).

[8] Kuznetsov V. Interval methods for processing statistical characteristics. In: *Proceedings of International Workshop on Applications of Interval Computations*, 116-122, 1995.

[9] Kozine I. and Krymsky V. Widening the scope of statistical judgements in reliability analysis. In: *Proceedings of 25 Symposium i Anvendt Statistik*, Frederiksberg, Denmark, 27-28 January, 137-148, 2003.

[10] Kozine I. and Filimonov Y. Imprecise reliabilities: experiences and advances. *Reliability Engineering & System Safety*, 67:75-83, 2000.

[11] Utkin L. and Gurov S. Imprecise reliability of general structures. *Knowledge and Information Systems*, 1:459-480, 1999.

Igor O.Kozine is with the System Analysis Department, Risø National Laboratory, P.O.Box 49 DK-4000 Roskilde, Denmark. E-mail: igor.kozine@risoe.dk

Victor G.Krymsky is with the Ufa State Aviation Technical University, Industrial Electronics Department, 12 K.Marx Street, Ufa, 450000, Russia. E-mail: kvg@mail.rb.ru

Climate Projections for the 21st Century Using Random Sets*

E. Kriegler
Potsdam Institute of Climate Impact Research, Germany

H. Held
Potsdam Institute of Climate Impact Research, Germany

Abstract

We apply random set theory to an analysis of future climate change. Bounds on cumulative probability are used to quantify uncertainties in natural and socio-economic factors that influence estimates of global mean temperature. We explore the link of random sets to lower envelopes of probability families bounded by cumulative probability intervals. By exploiting this link, a random set for a simple climate change model is constructed, and projected onto an estimate of global mean warming in the 21st century. Results show that warming estimates on this basis can generate very imprecise uncertainty models.

Keywords

climate change, climate sensitivity, imprecise probability, random sets, belief functions

1 Introduction

It is widely acceped by now that a discernible influence of anthropogenic emissions of greenhouse gases (GHGs) on the earth's climate exists. Greenhouse gas concentrations in the atmosphere have risen by, to name just a few, 30% (carbon dioxide), 250% (methane) and 15% (nitrous oxide) in the industrial era since 1750, mainly due to human activity. Empirical evidence for a growing climate change signal is mounting, and nearly all climate models need the increased radiative forcing due to growing GHG concentrations to reproduce this signal. Still, uncertainty abounds. How sensitive is the climate to growing GHG concentrations? What amount of greenhouse gases will humankind put into the atmosphere in the 21st century?

*This work has been supported in part by the Deutsche Bundesstiftung Umwelt (German Federal Foundation of the Environment).

We believe that the application of imprecise probability concepts carries the potential to greatly improve the situation in climate change forecasting and integrated assessment of climate change policies. However, an obstacle might be the dynamical nature of climate change models, and the large number of uncertain variables which mostly range over continuous possibility spaces. In this paper, we present an application of random set methods to the estimation of global mean temperature (GMT) change in the 21st century. We interprete the corresponding belief functions as a lower envelope of a set of probability measures, and try to respect this interpretation throughout the reasoning process. The uncertain model parameters are initially quantified by lower and upper cumulative probability distribution functions (CDF) on the real line. In section 2, we discuss how this information can be converted into a random set, combined for independent model parameters, and projected onto the model output. In section 3, we present the simple temperature change model, and construct a random set for its uncertain parameters. In section 4, the uncertainty in the input values is projected onto an estimate of global mean temperature change.

2 Methods

Random Sets of Imprecise CDF Models.

Consider an uncertain quantity X that enters a model of some causal relationship, e.g. of the link between GHG emissions and GMT. The imprecise uncertainty about X shall be described by a lower bound $\underline{F}_X : \mathbb{R} \to [0,1]$ and an upper bound $\overline{F}_X : \mathbb{R} \to [0,1]$ for a set of CDFs $F_X(x) := P(X \leq x)$ on the real line $\mathbb{R}$. In the following, such an uncertainty assessment will be called an *imprecise CDF model*

$$\mathcal{M}_X(\underline{F},\overline{F}) := \{\, P \,|\, \forall\, x \in \mathbb{R} \;\; \underline{F}(x) \leq P(-\infty,x] \leq \overline{F}(x) \,\} \tag{1}$$

A monotone set function $\underline{P} : \mathcal{R} \to [0,1]$, $\underline{P}(\emptyset) = 0$, $\underline{P}(\mathbb{R}) = 1$ is a *lower envelope* or *coherent lower probability* on the Borel algebra $\mathcal{R}$ of the real line, if it defines a non-empty set of countably additive probability measures $\mathcal{M}(\underline{P}) := \{\, P \,|\, \forall\, A \in \mathcal{R} \;\; \underline{P}(A) \leq P(A) \}$, and $\forall\, A \in \mathcal{R} \;\; \underline{P}(A) = \inf_{P \in \mathcal{M}(\underline{P})} P(A)$ [13, theorem 3.3.3]. An ∞-monotone lower envelope is a *belief function* Bel.

In the theory of Dempster [4], belief functions are generated by a multi-valued mapping from an underlying space $\Psi = \{\psi_1, ..., \psi_n\}$ onto a field of sets, in our case the Borel algebra $\mathcal{R}$. By means of the multi-valued mapping, a *probability mass assignment m* on Ψ can be transferred to $\mathcal{R}$, i.e. there exists $m : \mathcal{R} \to [0,1]$, with $m(A) > 0$ for only a finite number of sets $\mathcal{F} = \{E_1, ..., E_n\} \subset \mathcal{R}$ and $\sum_{A \in \mathcal{R}} m(A) = 1$. The pair $(\mathcal{F}, m)$ is called a (finite support) *random set*, and the sets $E_i \in \mathcal{F}$ *focal elements*. A belief function *Bel* and its conjugate *plausibility* *function Pl* are connected to a random set by [4, 11]

$$Bel(A) = \sum_{B \subseteq A} m(B) \;=\; \sum_{i\,|\,E_i \subseteq A} m_i\,, \qquad Pl(A) = \sum_{B \cap A \neq \emptyset} m(A) \;=\; \sum_{i\,|\,E_i \cap A \neq \emptyset} m_i$$

Thus, knowledge of the random set $(\mathcal{F}, m)$ suffices to determine *Bel* and *Pl* on $\mathcal{R}$.

We explore the relationship between the lower envelope of an imprecise CDF model and a belief function that can be represented by a finite support random set (In the following, the reference to the finiteness of the random set will be omitted). The goal is to capture the information content of an imprecise CDF model with a random set.

Proposition 1 *Let $\mathcal{M}_X(\underline{F}, \overline{F})$ be an imprecise CDF model as defined in (1). Let $\mathcal{A}$ be the algebra generated by the set of half-closed intervals $(a, b]$, $a < b$ of the real line $\mathbb{R}$. Let $(\mathcal{F}, m)$ be a random set, and $Bel_{\mathcal{F}}$, $Pl_{\mathcal{F}}$ the corresponding belief and plausibility functions, respectively.*

If (I) $(\mathcal{F}, m)$ contains only closed intervals $E_i = [\underline{x}_i, \overline{x}_i]$,
(II) $(\mathcal{F}, m)$ includes no pair of focal elements E_i, E_j with $\underline{x}_i < \underline{x}_j < \overline{x}_j < \overline{x}_i$, and
(III) $\forall\, x \in \mathbb{R}$ $Bel_{\mathcal{F}}(-\infty, x] = \underline{F}(x)$, $Pl_{\mathcal{F}}(-\infty, x] = \overline{F}(x)$,

$$\textit{then} \qquad \forall\, A \in \mathcal{A} \quad Bel_{\mathcal{F}}(A) = \underline{P}_X(A) := \inf_{P \in \mathcal{M}_X(\underline{F}, \overline{F})} P(A)$$

Proof. **Step 1**: Consider an arbitrary $(a, b] \in \mathcal{A}$, $a < b$. We have to show $\underline{P}_X(a, b] = Bel_{\mathcal{F}}(a, b]$ and $\overline{P}_X(a, b] = Pl_{\mathcal{F}}(a, b]$. Since $\underline{P}_X(A) = 1 - \overline{P}_X(A^c)$ and $Bel_{\mathcal{F}}(A) = 1 - Pl_{\mathcal{F}}(A^c)$, this implies that the equalities hold for the complement $(a, b]^c$ as well.

$$\begin{aligned} 1a) \quad \overline{P}_X(a, b] &= \overline{F}(b) - \underline{F}(a) \;=\; \sum_{i\,|\,E_i \cap (-\infty, b] \neq \emptyset} m_i - \sum_{j\,|\,E_j \subseteq (-\infty, a]} m_j \\ &= \sum_{s(i)\,|\,E_{s(i)} \subseteq (-\infty, a]} m_{s(i)} + \sum_{t(i)\,|\,E_{t(i)} \cap (a, b] \neq \emptyset} m_{t(i)} - \sum_{j\,|\,E_j \subseteq (-\infty, a]} m_j \\ &= Pl_{\mathcal{F}}(a, b] \end{aligned}$$

$$1b) \quad \underline{P}_X(a, b] = \max[0, \underline{F}(b) - \overline{F}(a)] = \max[0, \sum_{i\,|\,E_i \subseteq (-\infty, b]} m_i - \sum_{j\,|\,E_j \cap (-\infty, a] \neq \emptyset} m_j\,]$$

If $\underline{F}(b) < \overline{F}(a)$, there exists $E_* = [\underline{x}_*, \overline{x}_*] \in \mathcal{F}$ with $E_* \cap (-\infty, a] \neq \emptyset$ and $E_* \not\subseteq (-\infty, b]$. Assume an arbitrary $E' = [\underline{x}', \overline{x}'] \in \mathcal{F}$ with $\underline{x}' > a \geq \underline{x}_*$. By condition (II), $\overline{x}' \geq \overline{x}_* > b$. Thus, $E' \not\subseteq (a, b]$, and $Bel_{\mathcal{F}}(a, b] = 0$.

Assume there exists $E_* \in \mathcal{F}$ with $E_* \cap (-\infty, a] \neq \emptyset$ and $E^* \not\subseteq -(\infty, b]$. By condition (I)+(II), all $E_i \subseteq (-\infty, b] \in \mathcal{F}$ intersect $(-\infty, a]$, and $\underline{F}(b) < \overline{F}(a)$.

Thus, if $\underline{F}(b) \geq \overline{F}(a)$, there is no such focal element $E_* \in \mathcal{F}$. In other words, $\forall E_i \in \mathcal{F}\;\; E_i \nsubseteq (-\infty,b] \Rightarrow E_i \cap (-\infty,a] = \emptyset$.

$$\begin{aligned}
\Rightarrow \underline{P}_X(a,b] &= \sum_{s(i)\,|\,E_{s(i)}\subseteq(a,b]} m_{s(i)} + \sum_{t(i)\,|\,E_{t(i)}\cap(-\infty,a]} m_{t(i)} - \sum_{j\,|\,E_j\cap(-\infty,a]\neq\emptyset} m_j \\
&= Bel_{\mathcal{F}}(a,b]
\end{aligned}$$

Step 2: Consider an arbitrary union of k disjoint half-closed intervals $A_k = (a_1,b_1] \cup \ldots \cup (a_k,b_k]$, $a_1 < b_1 < \ldots < a_k < b_k$.

Choose a CDF $F^* : \mathbb{R} \to [0,1]$ with $F^*(a_1) = \min[\overline{F}(a_1), \underline{F}(b_1)]$, $F^*(b_1) = \underline{F}(b_1)$, ..., $F^*(a_k) = \min[\overline{F}(a_k), \underline{F}(b_k)]$, $F^*(b_k) = \underline{F}(b_k)$. Since $F^*(a_1) \leq F^*(b_1) \leq \ldots \leq F^*(a_k) \leq F^*(b_k)$, such a CDF does exist, and is contained in $\mathcal{M}_X(\underline{F},\overline{F})$.

$$\begin{aligned}
P^*(A_k) &= F^*(b_k) - F^*(a_k) + \ldots + F^*(b_1) - F^*(a_1) \\
&= \max[0, \underline{F}(b_k) - \overline{F}(a_k)] + \ldots + \max[0, \underline{F}(b_1) - \overline{F}(a_1)] \\
&= \underline{P}_X(a_k,b_k] + \ldots + \underline{P}_X(a_1,b_1]
\end{aligned}$$

Since the lower envelope $\underline{P}_X$ is super-additive on a union of disjoint sets [13, Ch. 2.7.4], $\underline{P}_X(A_k) = P^*(A_k)$. Thus, $\underline{P}_X(\bigcup_{l=1}^{k} (a_l,b_l]) = \sum_{l=1}^{k} \underline{P}_X(a_l,b_l]$. Since $\underline{P}_X(a_l,b_l] = Bel(a_l,b_l]$ as shown in step 1:

$$2a) \quad \underline{P}_X(A_k) = \sum_{l=1}^{k} \sum_{i|E_i\subseteq(a_l,b_l]} m_i = \sum_{i|E_i\subseteq \bigcup_{l=1}^{k}(a_l,b_l]} m_i = Bel_{\mathcal{F}}(A_k)$$

$$\begin{aligned}
2b) \quad \overline{P}_X(A_k) &= \overline{P}_X(-\infty,b_k] - \underline{P}_X((-\infty,a_1] \cup (b_1,a_2] \cup \ldots \cup (b_{k-1},a_k]) \\
&= \overline{F}(b_k) - \underline{F}(a_1] - \underline{P}_X(b_1,a_2] - \ldots - \underline{P}_X(b_{k-1},a_k] \\
&= \sum_{i|E_i\cap(a_1,b_k]\neq\emptyset} m_i - \sum_{j|E_j\subseteq\bigcup_{l=1}^{k-1}(b_l,a_{l+1}]} m_j = Pl_{\mathcal{F}}(A_k)
\end{aligned}$$

Every element of $\mathcal{A}$ is either $\emptyset$, $\mathbb{R}$, a union of $k \in \mathbb{N}$ disjoint half-closed intervals, or its complement. For the latter, $\underline{P}_X(A) = Bel_{\mathcal{F}}(A)$ has been shown in step 1 and 2. For $\emptyset$, $\mathbb{R}$, $\underline{P}_X(\emptyset) = Bel_{\mathcal{F}}(\emptyset) = 0$ and $\underline{P}_X(\mathbb{R}) = Bel_{\mathcal{F}}(\mathbb{R}) = 1$. □

Since the random set $(\mathcal{F}, m)$ contains only a finite number of focal elements, its corresponding belief and plausibility function cannot fulfil condition (III) of proposition 1 for continuous $\underline{F}$ and/or $\overline{F}$. For application purposes, however, this defect is not disturbing. Every imprecise CDF model with continuous lower and upper bound can be approximated by two step functions approaching the lower bound from below and the upper bound from above [7, 12]. Consider two step functions $SF_*, SF^* : \mathbb{R} \to [0,1]$ of the form $0 = SF(x_1) < ... < SF(x_k) = 1$,

$$SF_*(x) = \begin{cases} SF_*(x_{*i}) & x_{*i} \leq x < x_{*i+1} \\ 0 & x < x_{*1} \\ SF_*(x_{*k}) & x_{*k} \leq x \end{cases} \qquad SF^*(x) = \begin{cases} SF^*(x^*_{j+1}) & x^*_j < x \leq x^*_{j+1} \\ 0 & x \leq x^*_1 \\ SF^*(x^*_k) & x^*_{k'} < x \end{cases}$$

If $\forall\, x \in \mathbb{R} \quad SF^*(x) \geq SF_*(x)$, the two step functions define an imprecise CDF model $\mathcal{M}(SF_*, SF^*) := \{P \,|\, \forall\, x \in \mathbb{R} \quad SF_*(x) \leq P(-\infty, x] \leq SF^*(x)\}$. The following algorithm can be used to construct a random set $(\mathcal{F}, m)$, which fulfils the requirements of proposition 1, from two arbitrary $SF_* \leq SF^*$. Let the lower bound have cumulative probability $SF_*(x_{*i})$ at the “step” points $x_{*1} < ... < x_{*n}$, and the upper bound have cumulative probability $SF^*(x^*_j)$ at $x^*_1 < ... < x^*_m$.

Algorithm 1 *1. Initialize indices $k = 1$ (running over the focal elements of the random set to be constructed), $i = 1$ (running over x_{*i}), $j = 1$ (running over x^*_j). Let p_k denote the cumulative probability already accounted for in step k. Assign $p_0 = 0$.*

2. Construct random set $E_k = [x^*_j, x_{*i}]$.

3. (a) $SF_*(x_{*i}) < SF^*(x^*_j)$: $m_k = SF_*(x_{*i}) - p_{k-1}$, $p_k = SF_*(x_{*i})$. Raise indices $k \to k+1$, $i \to i+1$. Return to step 2.

(b) $SF_*(x_{*i}) > SF^*(x^*_j)$: $m_k = SF^*(x^*_j) - p_{k-1}$, $p_k = SF^*(x^*_j)$. Raise indices $k \to k+1$, $j \to j+1$. Return to step 2.

(c) $SF_*(x_{*i}) = SF^*(x^*_j)$: $m_k = SF^*(x^*_j) - p_{k-1}$. If $SF_*(x_{*i}) = SF^*(x^*_j) = 1$ abort the algorithm.
If $SF_*(x_{*i}) = SF^*(x^*_j) < 1$, set $p_k = SF^*(x^*_j)$. Raise indices $k \to k+1$, $i \to i+1$, $j \to j+1$. Return to step 2.

Algorithm 1 is well defined. For each step k, $x^*_j \leq x_{*i}$, $m_k > 0$, and the algorithm will always reach the points x_{*n}, x^*_m with $SF_*(x_{*n}) = SF^*(x^*_m) = 1$ and abort. It constructs a random set $(\mathcal{F}, m)$ with $k \leq n + m$ focal elements. The E_k are either closed intervals $[a_k, b_k]$ or singletons $\{a\} = [a_k, a_k]$. The algorithm is also applicable to the case of a precise probability, where $SF_* = SF^* = SF$.

Combining and Extending Random Sets.

In almost all assessments of climate change, uncertainty accumulates from different sources. In general, we need to consider a multivariate uncertainty model that arises from a vector of uncertain quantities $X = \{X_1, ..., X_n\}$, each of which is described by an imprecise CDF model $\mathcal{M}_{X_i}(\underline{F}, \overline{F})$ on the real line $\mathbb{R}$. There are different ways to construct a joint lower envelope $\underline{P}_X$ from the lower envelopes of independent marginals $\underline{P}_{X_i}$. They depend on the concept of independence that is employed to generate the joint envelope [2, 13]. In general, the resulting envelopes agree only on product sets $A_1 \times ... \times A_n$, $A_i \subseteq \mathbb{R}$.

In our case, the lower envelopes $\underline{P}_{X_i}$ of the independent marginals are represented by belief functions $Bel_{\mathcal{F}_i}$ with corresponding random sets $(\mathcal{F}_i, m_i) = \{(E_{1_i}, m_{1_i}), \ldots, (E_{k_i}, m_{k_i})\}$. The concept of *random set independence* [4] leads to joint belief functions by applying Dempster's rule of combination to logically independent "marginal" random sets $(\mathcal{F}_i, m_i)$, $1 \leq i \leq n$.

$$(\mathcal{F}, m) = \{\, (E_{l_1 \ldots l_n} = E_{l_1} \times \ldots \times E_{l_n}, m_{l_1 \ldots l_n} = m_{l_1} \cdot \ldots \cdot m_{l_n}),\ 1 \leq l_i \leq k_i \,\} \qquad (2)$$

It can be easily checked that $(\mathcal{F}, m)$ generates indeed a belief and plausibility function $Bel_{\mathcal{F}}$ and $Pl_{\mathcal{F}}$ that agree with the joint lower and upper envelopes $\underline{P}_X$ and $\overline{P}_X$ on product sets, no matter under which concept of independence they were generated. However, it is less clear, how $Bel_{\mathcal{F}}$ relates to the different types of the joint lower envelope on sets $A \in \mathcal{R}^n$ that are not product sets. Comparisons of different independence concepts on finite possibility spaces indicate that random set independence yields a lower envelope that is dominated by the envelopes emanating from epistemic or strong independence [2]. It needs to be further investigated how far these findings translate to the special case presented here. For the time being, we use random set independence to construct the joint lower envelope $Bel_{\mathcal{F}}$ from the independent marginals $Bel_{\mathcal{F}_i}$.

Consider a model of some causal relationship, which generates a transfer function $f : \mathbb{R}^n \to \mathbb{R}^m$, $y = f(x)$. Let the uncertainty in the input variables x be described by $\mathcal{M}_X(Bel_{\mathcal{F}}) := \{\, P_X \,|\, \forall\, A \in \mathcal{R}^n \;\; Bel_{\mathcal{F}}(A) \leq P_X(A) \,\}$. The corresponding random set $(\mathcal{F}, m) = \{(E_1, m_1), \ldots, (E_k, m_k)\}$ can be transferred to the model output y by applying the extension principle for random set-valued variables [5]:

$$f(E_i) := \{\, y \,|\, \exists\, x \in E_i \;\; y = f(x) \,\} \,, \qquad m_f(B) := \sum_{f(E_i) = B} m_i \quad B \in \mathbb{R}^m \qquad (3)$$

Let $(f(\mathcal{F}), m_f)$ denote the transferred random set. It corresponds to a belief function $Bel_{f(\mathcal{F})}$ that is the lower envelope of a set of probabilities $\mathcal{M}_Y(Bel_{f(\mathcal{F})})$. Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be Borel measurable, i.e. $\forall\, B \in \mathcal{R}^m \quad f^{-1}(B) = \{x \in \mathbb{R}^m : f(x) \in B\} \in \mathcal{R}^n$. Then, every probability measure P on $(\mathbb{R}^n, \mathcal{R}^n)$ is transformed by the mapping f into a probability measure P_f on $(\mathbb{R}^m, \mathcal{R}^m)$ defined by $\forall\, B \in \mathcal{R}^m \;\; P_f(B) := P(f^{-1}(B))$. Using this definition, we can transform each element of $\mathcal{M}_X(Bel_{\mathcal{F}})$ to a probability measure on $(\mathbb{R}^m, \mathcal{R}^m)$, thus generating:

$$f(\mathcal{M}_X(Bel_{\mathcal{F}})) := \{\, P_Y \,|\, \exists\, P_X \in \mathcal{M}_X(Bel_{\mathcal{F}}) \;\; \forall\, B \in \mathcal{R}^m \quad P_Y(B) = P_X(f^{-1}(B)) \,\}$$

Proposition 2 *Let $\mathcal{R}^n$, $\mathcal{R}^m$ be Borel algebras, $f : \mathbb{R}^n \to \mathbb{R}^m$ a Borel measurable transfer function. Let $(\mathcal{F}, m)$, $Bel_{\mathcal{F}}$ describe the set of probabilities $\mathcal{M}_X(Bel_{\mathcal{F}})$. Let $f(\mathcal{M}_X(Bel_{\mathcal{F}}))$ be the f-tranformed set of probabilities as defined above.*

Similarly, let $(f(\mathcal{F}), m_f)$ be the f-extension of $(\mathcal{F}, m)$ calculated from equation (3), and $Bel_{f(\mathcal{F})}$ the corresponding belief function. Then

$$f(\mathcal{M}_X(Bel_{\mathcal{F}})) \;\; \subseteq \;\; \mathcal{M}_Y(Bel_{f(\mathcal{F})}) := \{\, P_Y \,|\, \forall\, B \in \mathcal{R}^m \quad Bel_{f(\mathcal{F})}(B) \leq P_Y(B) \,\}$$

Proof. Consider an arbitrary $P_Y \in f(\mathcal{M}_X(Bel_{\mathcal{F}}))$. There exists a $P_X \in \mathcal{M}_X(Bel_{\mathcal{F}})$ with $\forall\, B \in \mathcal{R}^m \;\; P_Y(B) = P_X(f^{-1}(B))$. For a particular, yet arbitrary $B \in \mathcal{R}^m$

$$\begin{aligned} P_Y(B) \;=\; P_X(f^{-1}(B)) \;\;\geq\;\; Bel_{\mathcal{F}}(f^{-1}(B)) \;&=\; \sum_{E_i \subseteq f^{-1}(B)} m_i \\ &=\; \sum_{f(E_i) \subseteq B} m_i \;=\; Bel_{f(\mathcal{F})}(B) \end{aligned}$$

□

3 A Random Set for a Simple Climate Model

Global Mean Temperature Model.

We use a simple dynamical model to link radiative forcing $F(t)$ to a change ΔT in global mean temperature (GMT) since preindustrial times [14].

$$C_e \cdot \Delta T'(t) \;=\; F(t) - F_{2x} \cdot \frac{\Delta T(t)}{T_{2x}} \tag{4}$$

C_e effective ocean heat capacity

F_{2x} radiative forcing for a doubling of atmospheric CO_2

T_{2x} climate sensitivity

Differential equation (4) is the simplest type of energy balance model. It equates the net radiative flux into the system at the top of the atmosphere to oceanic heat uptake $C_e \Delta T'$. If the radiative forcing was kept constant at a value $F(t) = F_{2x}$, the system would undergo an equilibrium temperature change of $\Delta T = T_{2x}$. Climate sensitivity T_{2x} is a crucial parameter to characterize the response of the climate system to an increase in GHG concentrations.

The Intergovernmental Panel on Climate Change (IPCC) gives an estimate of climate sensitivity $T_{2x} = [1.5\text{ K}, 4.5\text{ K}]$ [3]. The panel explicitely refrains from specifying probabilistic information. Recently, models of intermediate complexity (EMICs) were used to establish probability distributions from a comparison of model results with historical atmosphere, surface and deep ocean temperature data [1, 6, 8]. Efforts are hampered by the presence of natural variability, the lack of long-term data and the multitude of forcings.

In this analysis, we use the probability distributions of [1, 6] to generate an imprecise CDF model for T_{2x} (fig. 1). The estimates of [1] are shifted to considerably higher values of climate sensitivity compared to [6], ranging up to values of $T_{2x} = 22$ K. One reason could be that [1] does not compare their results with deep ocean temperature data. [6] requires the ocean record to restrict T_{2x} from above. However, [8] considers ocean heat uptake, and fails to discriminate between climate sensitivity in the range $T_{2x} = [1\text{ K}, 10\text{ K}]$. In this situation, we simply cut of

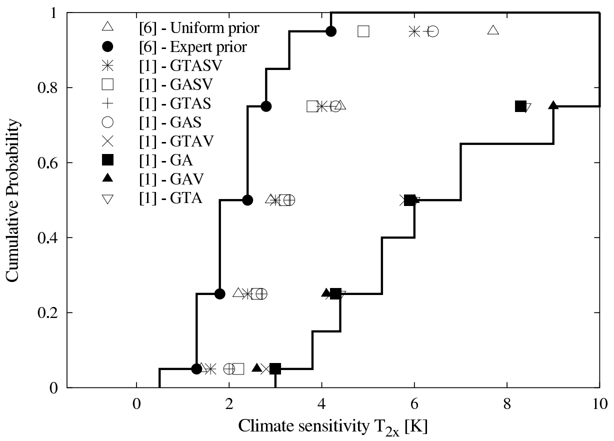


Figure 1: Imprecise CDF model for T_{2x}: Shown are 5%, 25%, 50%, 75% and 95% quantiles of probability distributions from [1, 6]. Estimates of [6] depend on a prior probability for T_{2x}. Estimates of [1] depend on whether solar forcing (S), volcanic aerosol forcing (V) and tropospheric ozone (T) was added to greenhouse gas (G) and aerosol forcing (A). The capital letters G, A, T, S, V in the figure key specify the radiative forcing components that were considered for the particular estimate of [1].

the probability distributions of [1] at $T_{2x} = 10$ K, and allocate their total probability mass $P(T_{2\times} \geq 10\text{ K})$ to this value.

Fig. 1 depicts the resulting ranges for 5%, 25%, 50%, 75% and 95% quantile estimates in [1, 6]. We interpolate the extreme values of the ranges to generate a lower and upper CDF, and approximate the resulting imprecise CDF model with two step functions SF_* and SF^* (Fig. 1). There is some arbitrariness here. It could be resolved by fixing the number of "step" points $T_{2x,i*}$ and $T^*_{2x,j}$, and calculating the optimal approximation according to some accuracy measure [7, 12]. Algorithm 1 is applied to construct a random set $(\mathcal{F}_{T_{2x}}, m_{T_{2x}})$ that corresponds to $\mathcal{M}_{T_{2x}}(SF_*,SF^*) := \{\,P \,|\, \forall\, T_{2x} \in \mathbb{R} \quad SF_*(T_{2x}) \leq P(-\infty, T_{2x}] \leq SF^*(T_{2x})\,\}$ (in the sense of proposition 1). $\mathcal{M}_{T_{2x}}(SF_*,SF^*)$ can be compared with the IPCC estimate [1.5 K, 4.5 K] for climate sensitivity. The probability for $T_{2x} \in$ [1.5 K, 4.5 K] lies in the interval $[0,1]$, for $T_{2x} < 1.5K$ in $[0,0.25]$, and for $T_{2x} > 4.5$ K in $[0,0.75]$. The numbers show that $\mathcal{M}_{T_{2x}}(SF_*,SF^*)$ does not support the IPCC estimate, especially for high climate sensitivities $T_{2x} > 4.5$ K. This reflects the fact that the upper bound of the IPCC estimate is not supported by [1, 6, 8].

Effective ocean heat capacity C_e is an artificial quantity that arises from the simple form of the energy balance model (4). It depends on ocean characteristics, but also on climate sensitivity [3]. A comparison of model (4) with emulations of different AOGCMs suggest a functional dependence of C_e on T_{2x} of the form $C_e \sim T_{2x}^{\gamma_c}$ with $0 < \gamma_c \leq 1$. We specify an interval uncertainty for the parameters $\overline{C} = C_e(T_{2x} = 3\text{ K})$ and γ_c, which is an adequate choice in the light of the large uncertainty surrounding ocean characteristics like vertical diffusivity [6]. Interval uncertainty is the simplest form of an imprecise CDF model. Lower and upper CDF are either 0 or 1. The model can be immediately captured by a random set $(\mathcal{F}_{\bar{C},\gamma_c}, m_{\bar{C},\gamma_c})$ containing just one focal element $E = [40$ Wa/m^{2}K, 50 Wa/m^{2}K$] \times [0.6,1]$ with probability mass assignment $m(E) = 1$.

An additional uncertainty concerns the present day global mean warming ΔT_o since 1860, which enters model (4) as initial value. Estimates of ΔT_o lie in the range 0.6 ± 0.2 K. We adopt the interval uncertainty [0.4 K, 0.8 K] for ΔT_o, since its small influence on future GMT projections does not justify a more complicated imprecise CDF model.

Radiative Forcing Model.

We group the anthropogenic sources of radiative forcing $F(t)$ into carbon dioxide, which is the most important GHG, the "other" greenhouse gases (OGHG) including both the remaining direct as well as indirect GHGs, and aerosols. Solar and volcanic sources are neglected since we are interested in estimating the anthropogenic climate change signal.

$$F(t) \quad = \quad F_{2x} \ln\left(\frac{C_{CO_2}(t)}{C_{CO_2}(1750)}\right) / \ln 2 + F_{Aer}\, g(E_{Aer}(t)) + F_{OGHG}\, h(t) \qquad (5)$$

C_{CO_2} atmospheric CO_2 concentration

E_{Aer} anthropogenic sulfate aerosol emissions

F_{Aer} Total aerosol forcing in the period 1990-2000

F_{OGHG} Total OGHG forcing in the period 1990-2000

The radiative properties of aerosol particles are most uncertain. Aerosols influence the radiation balance not only directly, but also indirectly by altering cloud formation processes. The IPCC estimates that the negative forcing of aerosols has been in the range [-0.8 W/m^2, -0.2 W/m^2] (direct effect) and [-2 W/m^2, 0 W/m^2] (indirect effect) for the period 1990-2000 [10]. [1, 6, 8] have investigated F_{Aer} in their comparison of model results with historical data. Fig. 2 shows the ranges for the 5%, 25%, 50%, 75% and 95% quantile estimates from [1, 6]. [8] presents a histogram probability which can be converted into two step functions for the lower and upper bound on the CDFs that are supported by the probability masses allocated to the bins of the histogram. Analogous to the case of climate sensitivity, we construct a lower CDF SF_* and upper CDF SF^* (solid lines in fig. 2). Algorithm 1 is used to generate the random set $(\mathcal{M}_{F_{Aer}}, m_{F_{Aer}})$ that corresponds to the imprecise CDF model $\mathcal{M}_{F_{Aer}}(SF_*, SF^*)$.

The probability that F_{Aer} is contained in the IPCC estimate [-2.8 K, -0.2 K] (direct and indirect effect combined) lies in the range $[0.95, 1]$. In contrast to climate sensitivity, the IPCC range includes $\mathcal{M}_{F_{Aer}}(SF_*, SF^*)$ almost entirely. The results in [1, 6, 8] support a more narrow range, where in particular the potential of a very strong negative aerosol forcing contribution is discarded.

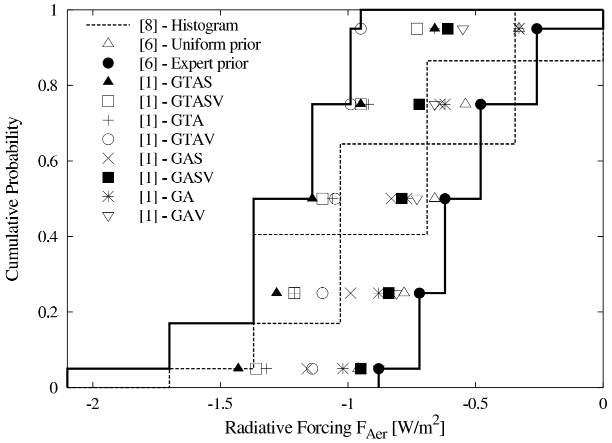


Figure 2: Imprecise CDF model for F_{Aer}: Shown are 5%, 25%, 50%, 75% and 95% quantiles of probability distributions from [1, 6], and a histogram probability from [8]. See Fig. 1 for additional explanation of the figure key.

Estimates for the radiative forcing contributions of indirect GHGs, in particular troposheric and stratospheric ozone, exhibit relative errors between 40%-70%. The indirect GHGs have contributed around 30-40% to F_{OGHG} in the last decade. We capture the uncertainty by the interval $F_{OGHG} \in [0.8 \text{ W/m}^2, 1.2 \text{ W/m}^2]$.

We link the uncertainty in the time-dependent paths of atmospheric CO_2 concentration $C_{CO_2}(t)$, future changes in the radiative forcing of the OGHG $h(t)$, and anthropogenic aerosol emissions $E_{Aer}(t)$ directly to the socio-economic sphere. Thereby, we neglect any uncertainty about the response of the biogeochemical cycles to anthropogenic emissions. In a special report on emissions scenarios (SRES) [9], the IPCC has formulated a range of scenarios describing future pathways of society and economy on a global scale. The major branching points of these scenarios are globalization vs. regionalization and sustainability orientation vs. growth orientation. In this analysis, we specify just two parameters G (“Growth”) and S (“Shift”), with $C_{CO_2}(t)$, $h(t) \sim e^{Gt-St^2}$. We restrict $S \leq G/200$, so that the growth in atmospheric CO_2 concentration and radiative forcing of OGHGs can be dampened, but not reversed by a “shift” S in the 21st century.

As the future socio-economic development is entirely uncertain, it is appropriate to specify interval uncertainties for $G \in [0.004/a, 0.012/a]$ and $S \in [0, G/200]$. Growth rates from 0.4% to 1.2% per year lead to atmospheric CO_2 concentrations from 480 ppmv to 1230 ppmv in 2100 (present day: 370 ppmv), and to a forcing contribution of the OGHG from 1 W/m^2 to 4 W/m^2. This covers the full range of the SRES scenarios including uncertainty in the biogeochemical cycles [3].

Combining the Random Set Information.

Most parameter pairs are physically and epistemically independent. Present day warming T_o depends physically on climate sensitivity and ocean heat capacity, but knowledge of T_o alone does not constrain the assessment of T_{2x} and C_e. A more critical issue is the epistemic dependence of F_{Aer} and T_{2x}. Although physically independent, comparisons of model results with historical data will have a tendency to produce high estimates of T_{2x} for a large negative radiative forcing F_{Aer} of aerosols, and vice versa [6]. Neglecting this dependence will yield a more imprecise estimate of future GMT change, since the probability weight of combinations with large negative F_{Aer} and low T_{2x} leading to a weak GMT increase, and with small negative F_{Aer} and high T_{2x} leading to a strong rise of GMT, will be overestimated. This issue needs to be investigated in further studies. For the time being, we use equation (2) based on random set independence to combine the random sets for all eight parameters $par := (\Delta T_0,\, T_{2x},\, \bar{C},\, \gamma_c,\, F_{Aer},\, F_{OGHG},\, G,\, S)$ to a joint random set $(\mathcal{F}_{par}, m)$.

4 Estimation of Global Mean Temperature Change

Differential equation (4) and radiative forcing model (5) generate a continuous transfer function that maps the uncertain model parameters to an increase ΔT in GMT since 1860. The extension principle for random sets ([5], equation 3) transfers the random set $(\mathcal{F}_{par}, m)$ for the uncertain parameters to a random set $(\mathcal{F}_{\Delta T}, m)$ for GMT increase. In our specific case, the images $f(E_{i,par}) = [\underline{\Delta T}_i(t), \overline{\Delta T}_i(t)]$ can be calculated with standard gradient-based optimization methods. After discretizing time in sufficiently small time steps Δt, the boundaries of the range at time $t_k = k\Delta t + t_o$ are found by solving

$$\underline{\Delta T}_i(t_k) \quad = \quad \min_{(\Delta T_0, T_{2x}, \bar{C}, \gamma_c, F_{Aer}, F_{OGHG}, G, S) \in E_{i,par}} \Delta T(t_k) \tag{6}$$

$$\text{subject to} \qquad \Delta T(t_l) = \Delta T(t_{l-1}) + \Delta t \cdot \left(\frac{F(t_{l-1})}{C_e} - \frac{F_{2x}}{C_e} \cdot \frac{\Delta T(t_{l-1})}{T_{2x}} \right) \qquad 1 \leq l \leq k$$

$$\overline{\Delta T}_i(t_k) \quad = \quad \max_{(\Delta T_0, T_{2x}, \bar{C}, \gamma_c, F_{Aer}, F_{OGHG}, G, S) \in E_{i,par}} \Delta T(t_k) \tag{7}$$

$$\text{subject to} \qquad \Delta T(t_l) = \Delta T(t_{l-1}) + \Delta t \cdot \left(\frac{F(t_{l-1})}{C_e} - \frac{F_{2x}}{C_e} \cdot \frac{\Delta T(t_{l-1})}{T_{2x}} \right) \qquad 1 \leq l \leq k$$

It can be checked that $\Delta T(t)$ is monotone in $\Delta T_o, \bar{C}, \gamma_c, F_{Aer}, F_{OGHG}, G, S$ and convex in T_{2x}. The latter is due to the fact that T_{2x} influences ΔT both directly and indirectly through its connection to effective ocean heat capacity. Thus, program (7) is a well-defined convex optimization problem. Care has to be taken with program (6). The solution will be a boundary point of the focal element $E_{i,par}$, and we have to check both for the lower and upper bound of T_{2x}.

Fig. 3 shows the image $[\underline{\Delta T}(t), \overline{\Delta T}(t)]$ of a single focal element. The range

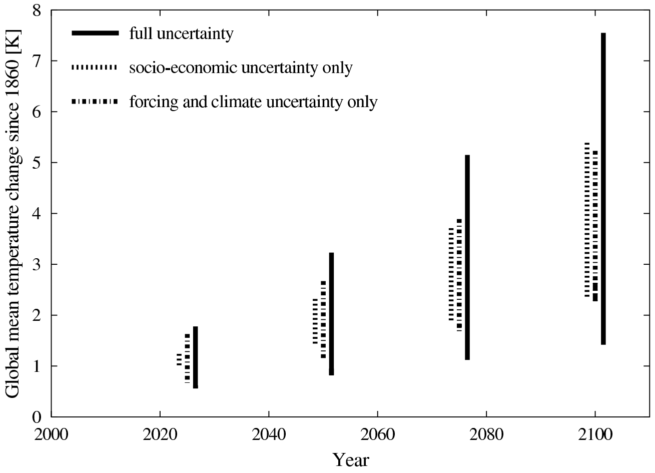


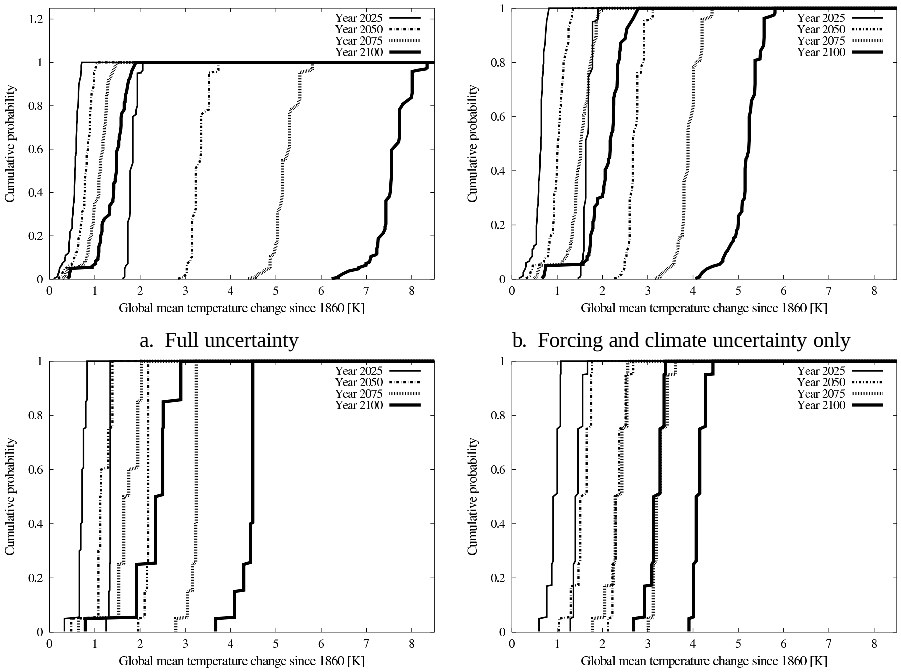
Figure 3: Image $[\underline{\Delta T}(t), \overline{\Delta T}(t)]$ of a single focal element $E^* =$ [1.8 K, 6.0 K] × [-1.37 W/m^2,-0.62 W/m^2] $\times (\Delta T_o, C, \gamma_c, F_{OGHG}, G, S) \in \mathcal{F}_{par}$ for the years 2025, 2050, 2075 and 2100. Shown are also the cases with solely socio-economic or solely forcing and climate uncertainty.

of the image grows considerably in time. We performed a sensitivity analysis with partly resolved uncertainty. Uncertainty in the radiative forcing and climate parameters dominates the overall uncertainty in the first half of the 21st century, but socio-economic uncertainty becomes equally important in the second half of the 21st century. Most strikingly, the uncertainties on the subspaces combine in a nonlinear way. A much larger overall uncertainty is found in particular for cases where the natural systems and socio-economic uncertainties are of similar size.

The projected random set $(\mathcal{F}_{\Delta T}, m)$ for GMT increase can be used to construct the lower CDF $\underline{F}_{\Delta T}$ and upper CDF $\overline{F}_{\Delta T}$. It is important to note that the corresponding imprecise CDF model $\mathcal{M}_{\Delta T}(\underline{F}, \overline{F}) := \{\, P \,|\, \forall\, x \in \mathbb{R} \;\; \underline{F}_{\Delta T}(x) \leq P(-\infty, x] \leq \overline{F}_{\Delta T}(x) \,\}$ can be more imprecise than $\mathcal{M}_{\Delta T}(Bel_{\mathcal{F}_{\Delta T}}) := \{\, P \,|\, \forall\, A \in \mathcal{A} \;\; Bel_{\mathcal{F}_{\Delta T}}(A) \leq P(A) \,\}$, i.e. $\mathcal{M}_{\Delta T}(\underline{F}, \overline{F}) \supseteq \mathcal{M}_{\Delta T}(Bel_{\mathcal{F}_{\Delta T}})$. This is due to the fact, that after applying the extension principle, the focal elements $E_{i,\Delta T} = [\underline{\Delta T}_i(t), \overline{\Delta T}_i(t)] \in \mathcal{F}_{\Delta T}$ might violate condition (II) of proposition 1. In this case, the lower envelope $\underline{P}_{\Delta T}$ of $\mathcal{M}_{\Delta T}(\underline{F}, \overline{F})$ is strictly smaller than $Bel_{\mathcal{F}_{\Delta T}}$ for some $A \in \mathcal{A}$. Recalling proposition 2, it can be seen that $\mathcal{M}_{\Delta T}(\underline{F}, \overline{F})$ does not contain more information than $(\mathcal{F}_{par}, m)$, which captures the uncertainty in the model parameters, would allow.

$$\mathcal{M}_{\Delta T}(\underline{F}, \overline{F}) \supseteq \mathcal{M}_{\Delta T}(Bel_{\mathcal{F}_{\Delta T}}) \supseteq \mathcal{M}_{par}(Bel_{\mathcal{F}_{par}})$$

Fig. 4 shows the lower and upper CDFs that are generated by the random set $(\mathcal{F}_{\Delta T}, m)$. We consider the area between lower and upper CDF as an indicator for the *imprecision* in the uncertainty. It can be seen that the imprecision in the GMT



a. Full uncertainty

b. Forcing and climate uncertainty only

c. Climate uncertainty only

d. Forcing uncertainty only

Figure 4: Lower and upper CDFs for GMT increase ΔT in the years 2025, 2050, 2075, 2100

estimate for the case of full uncertainty in the model parameters is enormous. This is partly due to the large number of uncertain parameters, as a comparison with the other cases shows. However, the cases (4.b) and (4.c) also exhibit large imprecision. This reflects the fact that the underlying imprecise CDF models for the climate parameters are already very imprecise. Certainly, they are conservative estimates, as the results of different studies were not weighed against each other. Some imprecision is also induced by the combination of the uncertainty for single parameters using random set independence (sec. 3).

The results can be compared with the IPCC estimate [1.8 K, 6.6 K] for GMT increase in 2100 relative to 1860 [3]. The probability for $\Delta T \in$ [1.8 K, 6.6 K] lies in the interval $[0, 1]$, for $\Delta T <$ 1.8 K in $[0, 0.95]$, and for $\Delta T >$ 6.6 K in $[0, 0.965]$. Despite the large range of the IPCC estimate, the uncertainty in GMT increase is too imprecise to discriminate against values outside this range. The probability mass allocated to values smaller than 1.8 K stems from random sets allowing for climate sensitivity values that are below the IPCC estimate for climate sensitivity. Similarly, the probability mass allocated to GMT increases higher than 6.6 K is due to climate sensitivity values above the IPCC estimate. As a comparison of (4.c) and (4.d) underlines, the uncertainty in climate parameters is the most influential factor on the uncertainty in GMT increase.

5 Conclusion

Imprecise probability concepts carry the potential to consistently capture the different types of uncertainties and different degrees of knowledge that are encountered in climate change analysis. However, they need to be applicable to dynamical problems with a large number of continuous uncertain variables. We suggest that imprecise CDF models are conceptually flexible and mathematically tractable enough to fulfil these competing requirements to some extent. When the imprecise CDF model is bounded by lower and upper step functions on the real line, the information about the encompassed set of additive probabilities can be condensed in a random set $(\mathcal{F},m)$. The corresponding belief function $Bel_{\mathcal{F}}$ is the lower envelope of the imprecise CDF model on the algebra generated by the half-closed intervals of the real line. Moreover, if the random set extension principle is used to project a random set onto the range of a measurable function, no information is added in the sense that every additive probability dominating $Bel_{\mathcal{F}}$ is transferred into a probability dominating the “extended” belief function.

We have constructed a random set for a simple climate model, and projected it onto an estimate for global mean temperature increase. The resulting estimate is very imprecise, with uncertainties about socio-economic development, radiative forcing and climate characteristics combining in a nonlinear way. The large imprecision of the estimate has different reasons and implications. Firstly, we incorporated a very broad range of factors in the analysis. Imprecision will be reduced if the range of factors is limited by formulating more specific questions. Secondly, we combined the random sets of single uncertain factors by assuming random set independence. This has increased the imprecision in the overall estimate, since aerosol forcing and climate sensitivity are not epistemically independent, when estimated from the present day climate change signal. Thirdly, the CDF models for the single parameters should be considered conservative estimates, which can be improved upon, when more comparisons of model results with historical data become available. Imprecision can be reduced in particular, if it is discriminated between the reliability of different models and methods.

Nevertheless, the results show that uncertainty is a key issue in the integrated assessment of climate change. Random set methods provide new insights into the structure of the uncertainty, particularly into its imprecision. The link to imprecise CDF models seems to be an important yardstick for assessing information losses when combining random sets, and applying the extension principle. More theoretical work is needed here to enhance the applicability of random sets to climate change analysis. In addition, methods need to be developed to determine imprecise CDF models directly from a comparison of model results with historical data.

References

[1] Andronova, N. G., and Schlesinger, M. E. Objective estimation of the probability density function for climate sensitivity. *Journal of Geophysical Research 106* (2001), 22605–22611.

[2] Couso, I., Moral, S., and Walley, P. Examples of independence for imprecise probabilities. In *Proceedings of the First International Symposium on Imprecise Probabilities and Their Applications* (1999), pp. 121–130.

[3] Cubasch, U., and Meehl, G. Projections of future climate change. In *Climate Change 2001: The Scientific Basis*, J. Houghton and Y. Ding, Eds. Cambridge University Press, Cambridge, 2001, pp. 525–582.

[4] Dempster, A. P. Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist. 38* (1967), 325–339.

[5] Dubois, D., and Prade, H. Random sets and fuzzy interval analysis. *Fuzzy Sets and Systems 42* (1991), 87–101.

[6] Forest, C. E., Stone, P. H., Sokolow, A. P., Allen, M. R., and Webster, M. D. Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science 295* (2002), 113–117.

[7] Hall, J., and Lawry, J. Generation, combination and extension of random set approximations to coherent lower and upper probabilities. *Reliability Engineering and System Safety* (2003), in press.

[8] Knutti, R., Stocker, T. F., Joos, F., and Plattner, G. K. Constraints on radiative forcing and future climate change from observations and climate model ensembles. *Nature 416* (2002), 719–723.

[9] Nakićenović, N., and Swart, R. *Emissions Scenarios. Special Report of the IPCC*. Cambridge University Press, Cambridge, 2000.

[10] Ramaswamy, V. Radiative forcing of climate change. In *Climate Change 2001: The Scientific Basis*, J. Houghton and Y. Ding, Eds. Cambridge University Press, Cambridge, 2001, pp. 289–348.

[11] Shafer, G. *A Mathematical Theory of Evidence*. Princeton U. Press, Princeton, 1976.

[12] Tonon, F. Using random set theory to propagate uncertainty through a mechanical system. *Reliability Engineering and System Safety* (2003), in press.

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

[14] Watterson, I. G. Interpretation of simulated global warming using a simple model. *Journal of Climate 13* (2000), 202–215.

H. Held is with the Potsdam Institute of Climate Impact Research, Germany. E-mail: kriegler@pik-potsdam.de

E. Kriegler is with the Potsdam Institute of Climate Impact Research, Germany. E-mail: kriegler@pik-potsdam.de

Exploring Imprecise Probability Assessments Based on Linear Constraints

RADU LAZAR
University of Minnesota, USA

GLEN MEEDEN
University of Minnesota, USA

Abstract

For many problems there is only sufficient prior information for a Bayesian decision maker to identify a class of possible prior distributions. In such cases it is of interest to find the range of possible values for the prior expectation for some real valued function of the parameter of interest. Here we show how this can be done when the imprecise prior assessment is based on linear constraints. In particular we find the joint range of possible values for a pair of such functions. We also study the joint range of the posterior expectation for a pair of functions.

Keywords

linear constraints, probability assessment, Bayesian inference, Metropolis-Hastings algorithm

1 Introduction

Consider the usual statistical inference problem where a subjective Bayesian must select a prior probability distribution which reflects their prior knowledge and beliefs about the unknown state of nature. Often one is unable to actually choose a single prior even though some prior information is present. When this occurs the Bayesian often selects a family of possible prior distributions. In such cases one could be interested in the range of possible values for some function defined on the the family of possible priors. More generally one could be interested in the set of possible values for a pair of functions. By judiciously selecting different pairs of functions such a graphical representation could help the Bayesian assess how sensible their initial choice for the family of priors really is. These graphical representations could help a pair of experts resolve possible conflicting prior beliefs. It could be used to check for areas of disagreement and to see how adjustments of some of their beliefs could lead to a merging of opinions.

In section two we assume that the parameter space contains only a finite number of points, say k. If $p = (p_1, \ldots, p_k)$ denotes a typical prior distribution we assume that the prior information can be expressed through linear equalities and inequalities involving the p_i's. This restricts the class of possible priors to a convex subset, $\mathcal{C}$ say, of the $k-1$ dimensional simplex of possible probability vectors of length k. Note $\mathcal{C}$ must be a convex polytope generated by a finite number of extreme points or vertices. Let ϕ denote a real valued function defined on the parameter space. Then its prior expectation is a linear function on $\mathcal{C}$. Dickey (see Dickey (2003)) has developed an interactive computing environment which computes the minimum and maximum of its expectation over $\mathcal{C}$. This programs allows the statistician to incorporated their prior information in stages and see how the range of ϕ changes. Here we are interested in finding the range of a pair of such functions. Because the prior expectations of the two functions are linear functions of p the range of possible values must be a convex set. Moreover, its extreme points must be contained in set of points which are images of the extreme points of $\mathcal{C}$. Hence, this problem is easily solved if one knows the extreme points of $\mathcal{C}$. But these can be found using a program that developed by Fuduka. See Fuduka (2003).

When considering the posterior expectation of such functions our problem becomes much more difficult analytically since the posterior expectation is no longer a linear function over $\mathcal{C}$. However, knowing the extreme points of $\mathcal{C}$ lets one find an approximate solution quite easily. Using these points one can generate random values in $\mathcal{C}$ by assigning random weights to them. Then one finds the values of the two functions at each of the realizations and plots these pairs of values.

In section three we consider the situation where the parameter is a r dimensional vector. We assume that it belongs to a convex polytope in r-dimensional Euclidean space defined by some known linear equalities and inequalities which reflect some of prior information of the statistician. A Bayesian needs to select a prior or possibly a family of prior distributions over this set to further reflect their uncertainty. For a given prior one is interested in computing prior and posterior expectations of some function of the parameter. In practice, it is usually not possible to make independent draws from a probability density defined on such a set. In such cases statisticians often employ Markov chain Monte Carlo methods to generate dependent samples from the posterior from which expectations can be computed approximately. Here, we use the Metropolis-Hastings algorithm to construct dependent samples drawn from a prior of interest. If the statistician selects as their family of priors all possible convex combinations of some finite collection of priors defined on the parameter space then for any pair of functions defined on the parameter space one can find the range of all possible values of their prior or posterior expectations.

As far as we know statisticians have not really addressed problems where imprecise knowledge is expressed through linear constraints. In section four we will note examples of somewhat similar problems that have been studied in the operations research literature. Finally, we will point out some of the difficulties when either the dimension of $\mathcal{C}$ or the parameter space gets too large.

2 The parameter space is finite

We begin by assuming that the parameter space contains only finitely many points, say k. In k-dimensional Euclidean space let Λ denote the $k-1$ dimensional simplex of p vectors with $p_i \geq 0$ and $\sum_{i=1}^{k} p_i = 1$. We assume that the known relations among the p_i's can be expressed by

$$Ap = a \tag{1}$$

where A is a known $r \times k$ matrix and a is a known vector of length r and

$$Bp \leq b \tag{2}$$

$$p \geq 0 \tag{3}$$

where B is a known $s \times k$ matrix and b is a known vector of length s. The set of $p \in \Lambda$ which satisfy the above equations form a closed convex subset of Λ which we will denote by $\mathcal{C}$.

Note that the interior of $\mathcal{C}$ is empty but we assume that properly considered $\mathcal{C}$ will have a nonempty interior in some smaller dimensional Euclidean space with a dimension of at least two. If ϕ_1 and ϕ_2 are two functions defined on the parameter space we let $\mathcal{C}(\phi_1, \phi_2)$ denote their range of possible values over $\mathcal{C}$. Our problem is to find this set. To see what could happen in practice we considered the following simple example.

Example 1 *We let $k = 10$ and imposed two equality constraints and two inequality constraints. The equality constraints were $p_5 = p_6$ and $\sum_{i=1}^{10} ip_i = 5.5$ while the inequality constraints were $p_1 \leq p_2$ and $\sum_{i=1}^{4} p_i \leq 0.5$. When doing the posterior calculations we assumed that the probabilities of seeing the observed data under the 10 possible parameter values were 0.1, 0.15, 0.09, 0.2, 0.3, 0.2, 0.1, 0.05, 0.07 and 0.02.*

As we noted in the introduction $\mathcal{C}(\phi_1, \phi_2)$ is a convex set whose extreme points are contained in the image of all the extreme points of $\mathcal{C}$. Hence, this becomes an easy problem once we know the extreme points of $\mathcal{C}$.

Fortunately for many problems the extreme points of $\mathcal{C}$ can be found easily using a program that is available over the Internet. See Fuduka (2003). It turns out for our example that $\mathcal{C}$ has 28 extreme points.

For definiteness we let $\phi_1(i) = (i - 5.5)^2$ for $i = 1, \ldots, 10$ and ϕ_2 be the indicator function of the set $\{2, 3, 4, 5\}$.

Plotting ϕ_1 and ϕ_2 at the extreme points of $\mathcal{C}$ we found the seven extreme points of $\mathcal{C}(\phi_1,\phi_2)$.

Since we know the extreme points of $\mathcal{C}$ we can use the Dirichlet family of distributions to generate a fairly flexible class of distributions which take values in $\mathcal{C}$ and from which one can simulate directly. If $q^1,\ldots,q^m$ are the extreme points of $\mathcal{C}$ and W is Dirichlet(α) where $\alpha = (\alpha_1,\ldots,\alpha_m)$ then $\sum_{i=1}^M W_i q^i$ defines a random distribution on $\mathcal{C}$. In some cases one may be able to make a judicious choice of α if some partial information about the ϕ's are available. For our example we generated 1,000 random values using the Dirichlet distribution with each of the 10 parameter values set equal to 0.1.

In the upper plot of Figure 1 we plotted the posterior expectations of the ϕ's for these 1,000 pairs of values along with the seven extreme points for the prior expectations. The prior expectations are darker and four of these points are clearly visible. They form the lower boundary of $\mathcal{C}(\phi_1,\phi_2)$. Another, (14.03,0.56), is clearly visible as well but maybe hard to identify because the plot is quite small. The other two, (3.82,0.64) and (4.58,0.67), are totally obscured by the posterior expectations. As to be expected the posterior expectations form a smaller set than the prior expectations.

Being able to find the extreme points of $\mathcal{C}$ is a powerful tool. In our somewhat limited experience the program we used seems quite good. In one constrained problem we considered in a different context it found over 28,000 extreme points. Using the Dirichlet distribution on the set of weights associated with the extreme points is a convenient distribution to sample from. These distributions are known as multivariate B-splines and are well studied. See for example Dahmen and Micchelli (1983). If one had a closed form expression for their densities then one could use importance sampling to approximate expectations under other densities. Unfortunately this can only be done in practice for very small problems. See for example Choudhuri (2003).

In practice one would use a much larger sample that 1,000 when studying the posterior expectations. But if the dimension of $\mathcal{C}$ gets to large one may not be able to take a large enough sample to get a reasonable approximation. In such cases one can find the minimum or maximum of the posterior expectation of a particular ϕ using a random search. The basic idea underlying random searches is well know and is quite simple. See for example Swann (1974).

3 The parameter space is a convex polytope

Here we will consider cases where the parameter space is no longer finite. We will assume that the parameter is a m dimensional vector, θ and that any possible choice for θ must satisfy

$$A\theta = a \tag{4}$$

where A is a known $r \times m$ matrix and a is a known vector of length r and

$$B\theta \leq b \tag{5}$$

where B is a known $s \times m$ matrix and b is a known vector of length s.

These constraints represent some of the statistician's prior information about θ. The parameter space, Θ, is the set of all θ which satisfy the above two equations. It is a convex polytope in m dimensional Euclidean space with an empty interior.

We begin by assuming that the statistician can select a prior density f over the parameter space to reflect the rest of their prior information. After discussing this case we will consider the situation where the statistician can only determine a family of possible prior densities.

Let ϕ denote some function defined on Θ. Then we are interested in finding

$$\mu = \int_{\Theta} \phi(\theta) h(\theta)\, d\theta \tag{6}$$

approximately. Interesting choices of h include f and the posterior density of θ under f given the data.

For most cases of interest this means employing Markov chain Monte Carlo methods. Here we will use the Metropolis-Hastings algorithm. With the Metropolis-Hastings algorithm one generates dependent observations, $Y_1, Y_2, \ldots$ from a suitable chosen Markov chain with values in Θ and then calculates

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i)$$

In nice situations $\hat{\mu}_n \to \mu$ almost surely.

If the current value of the chain is θ^* then one selects a proposal value for the next possible value in the sequence, θ', according to some probability distribution. Whether or not this new value is used depends on this probability distribution and the value of h at the two points. If $q(u,v)$ denotes the probability of selecting v as the proposal point when u is the current value we let

$$R = \frac{h(v)q(v,u)}{h(u)q(u,v)} \tag{7}$$

and accept the proposal with probability $\min(1,R)$. From this equation we see that the density h needs to be know only up to a constant since the value of R does not depend on this value.

We now explain how this can be done for our problem. Suppose θ^* is our current state which is assumed to lie in the relative interior of Θ. In particular, this means that it yields a strict inequality in all the equations in 5. There is an intuitive two step process by which we can choose the proposal. First, we select a random direction, d, in Θ. Remember that even though d is a vector of length m

it must remain in Θ which is essentially a lower dimensional set. The distribution for choosing d will not depend on the value θ^* Next, we find the set of points in Θ which lie either in the direction d or $-d$ from θ^* and choose the proposal at random from this set.

It is important to note that for this scheme if θ' is the proposal then

$$q(\theta^*, \theta') = q(\theta', \theta^*)$$

This follows because if d is the direction from θ^* to θ' then the only way to move from θ^* to θ' is if the directions d or $-d$ were selected in the first step. Of course the same is true if we are moving from θ' to θ^*. Clearly, the second step has the same distribution no matter which of the two points was chosen as the initial point.

To implement this scheme we proceed as follows. To get our direction d we choose at random a vector from $\mathcal{S}$, the unit sphere in m dimensional Euclidean space and then project it onto the null space of A. Next, we normalize this vector so that its length is one and denote it by d. Remember that $d \in \mathcal{S}$. Let $\mathcal{S}(A)$ denote the the subset $\mathcal{S}$ consisting of all vectors which can be generated in this way. Since we use the uniform distribution on the surface of $\mathcal{S}$ to choose the first vector, the distribution of d must be uniform on $\mathcal{S}(A)$. This follows by symmetry. The probability of d falling in any region of some fixed shape is independent of the location of the region in $\mathcal{S}(A)$.

Let θ^* denote a point which lies in the relative interior of Θ and consider vectors of the form

$$\theta^* + \alpha d \tag{8}$$

where α is a real number. Note if d does not belong to the null space of A this point cannot belong to Θ whenever $\alpha \neq 0$. On the other hand if $Ad = 0$ and α is sufficiently close to zero then this point will belong to Θ. Hence, α should be selected from the set of possible values that satisfy all the constraints of equation 5. This is a total of s constraints. Each constraint will result in either an upper or lower bound for α. Consider the interval formed by the maximum of these lower bounds to the minimum of these upper bounds. This is the range of possible values for α for which the vector in equation 8 will belong to Θ. Given a θ^* and a d in this equation then one just selects a value for α from the uniform distribution on its interval of possible values resulting in the proposal

$$\theta' = \theta^* + \alpha d$$

To recap, this is essentially a very simple procedure. Given a current value in the interior of Θ we first pick a random direction in Θ and then find how far we can move either in this or the opposite direction and still remain in Θ. Note that this method of selecting a proposal point does not depend in any way on the function h although from equation 7 the probability of it being accepted does depend on h. When h is the uniform distribution then the proposal is always accepted.

Suppose now the statistician does not have enough prior information to select a single prior density over Θ but can only specify a family of possible priors on Θ. A convenient and often sensible choice for such a family is all possible convex combinations of some finite set of densities. Let $f_1, \ldots, f_n$ denote such a finite set of densities and Π the be the family of all possible convex combinations of the f_i's. For a function ϕ define on Θ let $\Pi(\phi)$ be the range of possible values of the prior expectation of ϕ as the prior ranges over all members of Π. Since for each f_i we can find its ϕ expectation approximately we can find $\Pi(\phi)$ approximately. It is just the interval from the minimum to the maximum of these n prior expectations. Suppose instead we wished to find the joint range of the prior expectations of two functions as the prior ranged over Π, say $\Pi(\phi_1, \phi_2)$. This is easily done since it is just the convex hull of all points of the form $(\int \phi_1 f_i, \int \phi_2 f_i)$ for $i = 1, \ldots, n$. Posterior calculations are handled in exactly the same way since for any prior in Π the posterior is just a convex combination of the n posteriors. Finally, we emphasize the importance of only needing to know any probability density up to a constant since for most of these kind of problems the normalizing constant will be unknown.

Example 2 *We let $m = 5$ and supposed θ is the parameter for the multinomial distribution. We imposed the constraints $\theta_1 \leq \theta_2$, $(\theta_1 + \theta_2)/2 \leq (\theta_3 + \theta_4 + \theta_5)/3$ and $\theta_4 \leq \theta_5$ to get the parameter space Θ. We assumed the class of possible priors is all possible convex combinations of three Dirichlet distributions restricted to Θ. These were taken to be Dirichlet(2,2,2,2,2), Dirichlet(0.5,0.5,0.5,0.5,0.5) and Dirichlet(0.5,1.0,1.5,2,2.5). The two functions of interest were $\phi(\theta) = \theta_4 - \theta_3$ and $\phi_2(\theta) = \theta_5 - \theta_3$. To compare the prior and posterior expectations we assumed that a random sample of size 20 had been observed with the observed counts of states 1, 2, 3, 4 and 5 being 2, 3, 4, 3 and 8 respectively.*

Using our methods we found approximately the three extreme points of $\Pi(\phi_1, \phi_2)$ and the corresponding extreme points for the posterior problem. As we noted before it is crucial that the densities on Θ need only be known up to a constant to find these integrals approximately. The results are given in the lower plot of Figure 1. The prior expectations are marked by 0 and the posterior expectations are marked by x. As to be expected the posterior range is much smaller than the prior range. Using our approach and considering various sets of constraints, families of priors, choices of ϕ_1 and ϕ_2 and hypothetical samples can be useful in helping one select a sensible representation of their prior beliefs for a particular problem.

We end this section by noting that our method of picking a proposal point can be adapted to the random search method we mentioned in section 2. Formally, the two spaces $\mathcal{C}$ and Θ are essentially the same. In the random search algorithm given a point in the interior of $\mathcal{C}$ one needs to be able to choose a point at random from a small neighborhood that contains it. Hence after a direction has been selected at random rather than allowing a move that is as far as possible in either direction

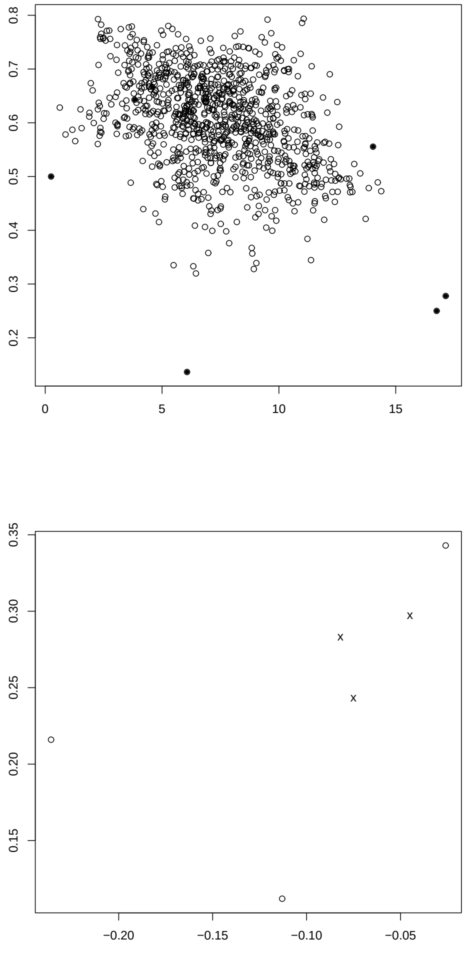


Figure 1: The upper plot contains the seven extreme points for the prior expectations and the plot of 1,000 posterior expectations for the functions ϕ_1 and ϕ_2 in Example 1. The lower plot gives the extreme points for prior (marked by 0) and posterior (marked by *x*) expectations of (ϕ_1, ϕ_2) for Example 2. In both cases the horizonal and vertical axes are the ϕ_1 and ϕ_2 expectations respectively.

one restricts the new point to be no more than ε away in either direction from the current point where $\varepsilon > 0$ is fixed.

4 Discussion

Although constraints seem a natural way to incorporate prior information into an inference problem they have not been widely considered in the statistical literature. The main reason seems to be that they are difficult to deal with both theoretically and practically. Betrò and Guglielmi (2000) considered robust Bayesian analysis under moment constraints in a fairly abstract setting and concluded that none of the current algorithms were good enough to be adopted for routine use. Generating random samples from distributions defined over bounded subsets of m dimensional Euclidean has been considered in a variety of contexts. Smith (1984) considers the problem of generating independent uniform observations from a bounded region while Belisle, Romeijin and Smith (1993) considers algorithms for generating observations from a general multivariate distribution. They assume that the region of interest is open with a nonempty interior which will not work here. Boender et al. (1991) and Chen and Schmeiser (1993) consider somewhat related problems.

At the present time Markov chain Monte Carlo methods seem to be the best way to handle the types of problems considered in section 3. They come with no guarantees however. If run long enough they will converge to the correct answer but in a given example it can be very difficult to know when to stop. When m is large it is impossible to visualize Θ. From our experience, it seems one should select a starting value for θ that is somewhere in the “center” of Θ. In some cases it seems that it is possible for the chain to spend long periods trapped in a corner near the boundary of Θ. If you start in the center then any region of Θ you eventually reach you will also eventually leave. When trying to compute equation 6 approximately it is not necessary to visit every niche and corner of Θ especially those where h puts little weight. But in examples we have studied we have seen that it can take a very long time to reach certain regions very near the boundary. As be noted by many authors when m increases we must deal with the “curse of dimensionality”. For a helpful discussion on the convergence of Markov chain Monte Carlo simulations see Geyer (1992) and Gelman and Rubin (1992).

The methods discussed here not only can help a single statistician evaluate the consequences of their prior assessments but could help a pair of experts resolve possibly conflicting prior beliefs. It can be used to check how areas of disagreement will effect their inferences. It could help them study how adjustments of some of their beliefs could lead to a merging of opinions. Clearly this is not a theory of how to readjust one beliefs when face with new information but a way to explore the consequences of readjustments of linear constraints. Here we have emphasized exploring jointly the prior or posterior expectations for a pair of functions. The methods work for jointly exploring three and even more functions however convenient graphical representations are no longer possible.

Our simulations were done using *R*. We are preparing a small package that would make it easy for others to implement these methods. Once it is finished we will submit it to the *R* archives for public distribution. We hope to complete this sometime this year.

Finally, the authors wish to thank Charles Geyer for many helpful discussions.

References

[1] C. J. P. Belisle, H. E. Romeijn and R. L. Smith. Hit-and-Run Algorithms for Generating Multivariate Distributions. *Mathematics of Operations Research*, 18:255-266, 1993.

[2] B. Betrò and A. Gugliemi. Methods for Global Prior Robustness under Generalized Moment Conditions. In *Robust Bayesian Analysis* (D. R. Insua and F. Ruggeri eds.) 273-293, Springer 2000.

[3] C. G. E. Boender, R. J. Caron, J. F. McDonald, A. H. G Rinnooy Kan, H. E. Romejin and R. L. Smith. Shake-and-Bake Algorithms for Generating Uniform Points on the Boundary of Bounded Polyhedra. *Operations Research* 39: 945-954, 1991.

[4] M. Chen and B. Schmeiser. Performance of the Gibbs, Hit-and-Run and Metropolis Samplers. *Journal of Computational and Graphical Statistics*, 2, 251-272, 1993.

[5] N. Choudhuri. Computing Multivariate B-splines: A Simulation Based Approach. Technical Report, Case Western Reserve University, 2003.

[6] W. Dahmen and C. A. Micchelli. Recent Progress in Multivariate Splines. In *Approximation Theory IV* (C. K. Chui, L. L. Schumaker and J. D. Ward, eds.) 17-121. Academic Press, New York 1983.

[7] J. M. Dickey. Convenient Interactive Computing for Coherent Imprecise Prevision Assessment. submitted to ISIPTA '03, 2003.

[8] K. Fuduka. cdd and cddplus Homepage for locating vertices. http://www.cs.mcgill.ca/∼fukuda/soft/cdd_home/cdd.html, 2003.

[9] A. Gelman and D. B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science,* 7, 457:472, 1992.

[10] C. J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7, 473-483, 1992.

[11] W. H. Swann. Constrained Optimization by Direct Search. In *Numerical Methods for Constrained Optimization* (P. E. Gill and W. Murray, eds.) 191-217. Academic Press, New York 1974.

[12] R. L. Smith. Efficient Monte Carlo Procedures for Generating Points Uniformly Distributed over Bounded Regions. *Operations Research* 32:1296-1308, 1984.

Radu Lazar is a graduate student in the School of Statistics at the University of Minnesota, Minneapolis, MN 55455, USA. E-mail: lazar@stat.umn.edu

Glen Meeden is with the School of Statistics at the University of Minnesota, Minneapolis, MN 55455, USA. E-mail: glen@stat.umn.edu

Continuous Linear Representation of Coherent Lower Previsions

SEBASTIAN MAASS
Universität Bremen, Germany

Abstract

This paper studies the possibility of representing lower previsions by continuous linear functionals. We prove the existence of a linear isomorphism between the linear space spanned by the coherent lower previsions and that of an appropriate space of continuous linear functionals. Moreover, we show that a lower prevision is coherent if and only if its transform is monotone. We also discuss the interpretation of these results and the new light they shed on the theory of imprecise probabilities.

Keywords

coherent lower previsions, Möbius transform, Choquet's theorem, Bishop-de Leeuw theorem, Dempster-Shafer-Shapley representation theorem

1 Introduction

The theory of imprecise probabilities especially that of coherent lower previsions has been designed to mathematically cope with subjective behavior in decision situations (cf. Walley [11]). It has evolved so extensively that coherent lower previsions have been repeatedly reinvented under different names like e.g. “coherent risk measures” (cf. Delbaen [4]) or “maxmin expected utility” (cf. Gilboa and Schmeidler [7]).

From an applicational and often also mathematical point of view nonlinear functionals like coherent lower previsions cannot as nice be handled as (monotone) continuous linear functionals. So, in this paper, we are interested in representing the former functionals by the latter. For nonadditive set functions such a representation is well-known as Dempster-Shafer-Shapley Representation Theorem in the discrete case or as Möbius transform in the general case (cf. Denneberg [5], Gilboa and Schmeidler [6] and Marinacci [10]).

The main steps of constructing such a transformed set function run as follows. First, given a totally monotone set function ν on an algebra $\mathcal{A}$, to every set $A \in \mathcal{A}$ is assigned a function $\tilde{A}$ on the extreme points of the convex set of normalized totally monotone set functions and defined by $\tilde{A}(\eta) := \eta(A)$. Since the extreme points are the filter games and therefore $\{0,1\}$-valued (cf. Choquet [3] p. 260 f.) all $\tilde{A}$ can be interpreted as characteristic functions of the sets $\{\eta \mid \eta(A) = 1\}$. Then, by different methods, it can be shown that there exists a bijective mapping from the set of totally monotone set functions to the set of (positive) measures on the σ-algebra generated by the $\tilde{A}$, $A \in \mathcal{A}$. Finally, this bijective mapping can be extended to the linear spaces each spanned by the respective class of set functions.

In this paper we will show that the main results of these theorems do not presuppose the functions being totally monotone set functions. Even a structured domain like an algebra is not necessary to obtain analogous results for coherent lower previsions. In our main theorem (Theorem 2) we provide a representation theorem for coherent lower previsions which contains results analogously to those sketched in the preceding paragraph for totally monotone set functions.

2 Preliminaries

Let Ω be a nonempty set, $B(2^\Omega)$ the linear space of bounded (w.r.t. the supremum norm) real-valued functions on Ω and $K \subset B(2^\Omega)$ be nonempty. To avoid laborious considerations of special cases, we will assume that there is at least one nonzero function in K. A *lower prevision* on K is a real-valued functional $\underline{P} : K \to \mathbb{R}$. A lower prevision $\underline{P}$ is called *coherent*, if $\underline{P}(f) \geq \sum_{i=1}^{n} \lambda_i \underline{P}(f_i) + \lambda_0$ whenever $f \geq \sum_{i=1}^{n} \lambda_i f_i + \lambda_0$ with $f, f_i \in K$, $\lambda_i > 0$, $\lambda_0 \in \mathbb{R}$, $n \in \mathbb{N}$. This definition is not the usual one (cf. Walley [11, Definition 2.5.1]) but it follows immediately from Proposition 3.1.2 (d) and Lemma 3.1.3 (b) in Walley's book and it will be of use to prove a functional being *not* coherent. Furthermore, this characterization of coherence can nicely be interpreted in the following way. As usual, Ω denotes a possibility space, K a set of *gambles*, i.e. positive or negative rewards depending on the uncertain state $\omega \in \Omega$. A lower prevision $\underline{P}$ of a gamble f is then the supremum buying price for f one is willing to pay. Since the system of buying prices have to fulfill some justified consistency properties, Walley introduced the notion of coherence which, using the characterization given above, means that, whenever a gamble f is dominating a portfolio of other gambles (possibly including a sure gain or loss λ_0) independently of the state ω, one should be willing to pay at least as much for f as one is willing to pay for the individual gambles included in the portfolio (not for the portfolio as whole - this would be considered as one gamble).

If K consists of characteristic functions then $\underline{P}$ can be interpreted as a set function and then is called a *coherent lower probability*. We have shown in [8] that the normalized exact games in cooperative game theory are the coherent lower probabilities. Simple examples of coherent lower probabilities are unanimity games, i.e. set functions u_A on an algebra $\mathcal{A}$ with $A \in \mathcal{A}$ and $u_A(B) = 1$ if $B \supset A$ and 0 else. If K is a linear space containing constant functions then $\underline{P}$ is a coherent lower prevision if and only if it is monotone, positively homogeneous, superadditive, normalized (i.e. $\underline{P}(1) = 1$) and constant additive (i.e. $\underline{P}(f + c) = \underline{P}(f) + \underline{P}(c)$). This characterization is almost equivalent to that of “coherent risk measures” (cf. Delbaen [4] and Maaß [9]). It is well-known (cf. Walley [11], Chapter 3) that every coherent lower prevision can be extended coherently to the linear space of all bounded real-valued functions. The minimum of all such extensions exists and is called the *natural extension*. Since coherence implies that every function $f \in K$ is mapped into the bounded interval $\inf f, \sup f$, $CLP(K)$ is contained in $BLP(K)$. Denote by $BLP(K)$ the linear space of all lower previsions on K which are bounded w.r.t. the operator norm $||\cdot||$, $||\underline{P}|| := \sup_{f\in K, f\neq 0} \frac{|\underline{P}(f)|}{||f||_\infty}$, and by $CLP(K)$ the convex set of all coherent lower previsions on K.

The linear space $BLP(K)$ will additionally be considered as a topological space endowed with the topology $\mathcal{T}$ having as subbase the sets $B(\underline{P}, f, \varepsilon) := \{\underline{P}' \in BLP(K) \mid |\underline{P}'(f) - \underline{P}(f)| < \varepsilon\}$, with $\underline{P} \in BLP(K)$, $f \in K$, $\varepsilon > 0$. The definition of $\mathcal{T}$ is similar to that of the weak* topology and it is the smallest making all functions

$$\tilde{f} : BLP(K) \to \mathbb{R}, \quad \tilde{f}(\underline{P}) := \underline{P}(f)$$

continuous for all $f \in K$. The set of all such $\tilde{f}$ will be denoted by $\tilde{K}$, the linear space spanned by $\tilde{K}$ will be denoted by $\mathrm{span}(\tilde{K})$. The topology $\mathcal{T}$ is also known as the topology of pointwise convergence and, by definition of the product topology, $\mathcal{T}$ is identical with the relative topology of $BLP(K)$ as a subset of the product space $\Pi_{f\in K}\mathbb{R}_f$, $\mathbb{R}_f := \mathbb{R}$ for all $f \in K$.

We start with some topological results that will serve as technical basis for the following analysis.

Proposition 1 *Under the topology $\mathcal{T}$ the linear space $BLP(K)$ is a locally convex and Hausdorff topological linear space.*

Proof. We have to show that $\mathcal{T}$ possesses a base consisting of convex sets. Since convexity is preserved under forming intersections it suffices to show that the given subbase of $\mathcal{T}$ consists of convex sets. Therefore, suppose $\underline{P}_1, \underline{P}_2 \in B(\underline{P}, f, \varepsilon)$ with $\underline{P} \in BLP(K)$, $f \in K$ and $\varepsilon > 0$ and let $\lambda \in [0,1]$. Then

$$|\lambda\underline{P}_1(f) + (1-\lambda)\underline{P}_2(f) - \underline{P}(f)| \leq \lambda|\underline{P}_1(f) - \underline{P}(f)| + (1-\lambda)|\underline{P}_2(f) - \underline{P}(f)| < \varepsilon,$$

i.e. $B(\underline{P}, f, \varepsilon)$ is convex since $\underline{P}_1, \underline{P}_2$ and λ were chosen arbitrarily. Hence, all elements of the subbase are convex since $\underline{P}$, f and ε were chosen arbitrarily. □

Proposition 2 *The unit ball in $(BLP(K), ||\cdot||)$, $B := \{\underline{P} \in BLP(K) \mid ||\underline{P}|| \leq 1\}$, is $\mathcal{T}$-compact.*

Proof. Let $I := \Pi_{f\in K}[-1,1]$. By Tychonoff’s Theorem, I is compact w.r.t. the product topology. Let $\tau : B \to I$ be the injective mapping $\tau(\underline{P}) := \Pi_{f\in K} \frac{\underline{P}(f)}{||f||_\infty}$. Since the sets $B(\underline{P}, f, \varepsilon) := \{\underline{P}' \in B \mid |\underline{P}'(f) - \underline{P}(f)| < \varepsilon\}$ with $\underline{P} \in B$, $f \in K$, $\varepsilon > 0$ form a subbase for the relative topology $\mathcal{T}_B$ of B generated by $\mathcal{T}$ and since $\{\Pi_{f\in K} U_f \mid U_f = \mathbb{R}\ \forall f \in K \setminus \{f'\}, U_{f'} =]x-\varepsilon, x+\varepsilon[, f' \in K, x \in \mathbb{R}, \varepsilon > 0[\}$ is a subbase of the product topology in $\mathbb{R}^K$, the images of the $\mathcal{T}_B$-subbase elements of $\mathcal{T}_B$ form a subbase of the relative product topology in $\tau(B)$. Thus τ is a homeomorphism between B endowed with the relative $\mathcal{T}$-topology, and $\tau(B)$ endowed with the relative product topology. Therefore, to prove that B is $\mathcal{T}$-compact, it suffices to show that B is $\mathcal{T}$-closed. This is easily done since for any $\underline{P} \in BLP(K)$ with $\|\underline{P}\| > 1$ there exist a $f \in K$ and a $\varepsilon > 0$ with $|\underline{P}(f)| > \|f\|_\infty + \varepsilon$ such that $B(\underline{P}, f, \varepsilon)$ is an open neighborhood of $\underline{P}$ disjoint from B, i.e. B is $\mathcal{T}$-closed. □

Proposition 3 *The set $CLP(K)$ is $\mathcal{T}$-compact in $BLP(K)$.*

Proof. Obviously, $CLP(K)$ is a subset of the $\mathcal{T}$-compact set B. So, it remains to prove that $CLP(K)$ is $\mathcal{T}$-closed. Suppose $\underline{P}$ is a noncoherent lower prevision. Then there exist $f, f_i \in K$, $\lambda_i > 0$, $\lambda_0 \in \mathbb{R}$, $i \in \{1, \ldots, n\}$ and $\varepsilon > 0$ with $f \geq \sum_{i=1}^n \lambda_i f_i + \lambda_0$ and $\underline{P}(f) + \varepsilon < \sum_{i=1}^n \lambda_i \underline{P}(f_i) + \lambda_0$. Setting $\varepsilon_i := \varepsilon / (2 \sum_{k=1}^n \lambda_k)$, the set $B(\underline{P}, f, \frac{1}{2}\varepsilon) \cap \bigcap_{i=1}^n B(\underline{P}, f_i, \varepsilon_i)$ is an open neighborhood of $\underline{P}$ which is disjoint from $CLP(K)$. Hence, $CLP(K)$ is $\mathcal{T}$-compact. □

The main result of this paper will heavily base on the Bishop-de Leeuw Theorem (cf. Alfsen [1, Theorem I.4.14]) which, like Choquet’s Theorem, belongs to a group of results generalizing the famous Krein-Milman Theorem. We recall that the Baire σ-algebra is the smallest σ-algebra for which all continuous real-valued functions are measurable, with, as usual, the Borel σ-algebra on the range space $\mathbb{R}$. Furthermore, denote by $\mathrm{ex}(X)$ the set of extreme points of X.

Theorem 1 (Bishop-de Leeuw) *Suppose E is a locally convex Hausdorff space over $\mathbb{R}$ and X a nonempty compact convex subset of E. Denote by $A(X)$ the linear space of continuous real-valued functions $a : X \to \mathbb{R}$ which are affine, i.e. $a(\lambda x + (1-\lambda)y) = \lambda a(x) + (1-\lambda)a(y)$ for $x, y \in X$, $0 \leq \lambda \leq 1$ and by $\mathcal{B}_0$ the Baire σ-algebra on X. Then for every $x \in X$ there exists a probability measure μ_x on the σ-algebra $\mathrm{ex}(X) \cap \mathcal{B}_0$, such that*

$$a(x) = \int a \, d\mu_x \quad \textit{for all } a \in A(X). \tag{1}$$

Generally, it is not possible to replace the Baire σ-algebra by the more usual Borel σ-algebra (cf. Alfsen [1, p. 39 f.]).

3 Main Results

In this section, we present the announced isomorphism between the linear space spanned by $CLP(K)$ and a linear space of continuous linear functionals and characterize the previsions in $CLP(K)$ by monotonicity of their transform. As a preparation, we start with a simple application of the Bishop-de Leeuw Theorem.

Lemma 1 *For every coherent lower prevision $\underline{P}$ on K there exists a probability measure $\mu_{\underline{P}}$ on the σ-algebra $\mathrm{ex}(CLP(K)) \cap \mathcal{B}_0$, such that*

$$\underline{P}(f) = \int \tilde{f}\, d\mu_{\underline{P}} \quad \textit{for all } f \in K. \tag{2}$$

Proof. The assertion made in the lemma follows directly from Theorem 1 using Proposition 1 and 3 and from $\tilde{f} \in A(CLP(K))$ for all $f \in K$. □

We obviously have found that the continuous linear functional $\int \cdot\, d\mu_{\underline{P}}$ represents the coherent lower prevision $\underline{P}$ via the nonlinear application $f \mapsto \tilde{f}$. Unfortunately, the representing measure $\mu_{\underline{P}}$ needs not to be unique as the following example shows.

Example 1 *Let $\Omega = \{1,2,3\}$ and $\nu : 2^\Omega \to \mathbb{R}$ be the coherent lower probability defined by $\nu(A) := \frac{1}{2}$ iff $|A| = 2$ and $\nu(A) := 0$ iff $|A| < 2$. Then ν is an extreme point of the set of coherent lower probabilities on 2^Ω, $CLP(2^\Omega)^1$. Suppose ν is a convex combination of two coherent lower probabilities ν_1 and ν_2 Obviously, $\nu_1(A) = \nu_2(A) = \nu(A)$ for all A with $\nu(A) \in \{0,1\}$, i.e. $|A| \neq 2$. Therefore, suppose $\nu_1(\{1,2\}) > \nu(\{1,2\}) = \frac{1}{2}$. By coherence of ν_1, $1_{\{1\}} \geq 1_{\{1,2\}} + 1_{\{1,3\}} - 1$ implies $\nu_1(\{1\}) \geq \nu_1(\{1,2\}) + \nu_1(\{1,3\}) - 1$ such that $\nu_1(\{1,3\}) < \frac{1}{2}$. Analogously, we conclude $\nu_1(\{2,3\}) < \frac{1}{2}$. The same argument applied to ν_2 implies that both ν_1 and ν_2 are at least for two of three sets A with $|A| = 2$ smaller than or equal to $\nu(A)$. Hence, $\nu_1 = \nu_2 = \nu$.*
Further on, it is easy to see that all unanimity games on 2^Ω are extreme points of $CLP(2^\Omega)$.
The coherent lower probability $\nu' : 2^\Omega \to \mathbb{R}$ defined by $\nu'(A) := \frac{1}{3}$ iff $|A| = 2$ and $\nu'(A) := 0$ iff $|A| < 2$ can be obtained by two different convex combinations of extreme points of $CLP(2^\Omega)$, $\nu' = \frac{1}{3}u_{\{1,2\}} + \frac{1}{3}u_{\{1,3\}} + \frac{1}{3}u_{\{2,3\}}$ and $\nu' = \frac{2}{3}\nu + \frac{1}{3}u_\Omega$. Since the coefficients of the extreme points used in the convex combinations are the masses of the transform $\mu_{\nu'}$ of ν', we obtain that uniqueness of the representing measure cannot be guaranteed.

To obtain uniqueness, we have to draw our attention to the integrals because for two representing measures $\mu_{\underline{P}}$ and $\mu'_{\underline{P}}$ of $\underline{P}$ we have, by Lemma 1,

$$\int \tilde{f}\, d\mu_{\underline{P}} = \int \tilde{f}\, d\mu'_{\underline{P}} \quad \text{for all } f \in K. \tag{3}$$

So, if we just restrict the continuous linear functional $\int \cdot\, d\mu_{\underline{P}}$ to the linear space $\mathrm{span}(\tilde{K})$ we get the desired uniqueness.

[1] It can be shown that ν is the only non-unanimity game in the set of extreme points of $CLP(2^\Omega)$.

The subsequent lemma (cf. Maaß [9, Proposition 6]) is mainly for technical use in the proof of the following theorem. As will be discussed after Lemma 3, it can be of practical use.

Lemma 2 *Let $\{\underline{P}_i\}_{i\in I}$ be a nonempty indexed set of coherent lower previsions on $K \subset B(2^\Omega)$ and the lower prevision $\underline{P} : B(2^I) \to \mathbb{R}$ be coherent. Then the functional*

$$K \to \mathbb{R}, \quad f \mapsto \underline{P}(i \mapsto \underline{P}_i(f)) \tag{4}$$

is a coherent lower prevision.

Proof. The functional defined in (4) is well defined since coherence of the $\underline{P}_i$ implies $-\infty < \inf f \leq \underline{P}_i(f) \leq \sup f < \infty$ such that the function $i \mapsto \underline{P}_i(f)$ is bounded for every $f \in K$. By considering the natural extensions $\underline{E}_i$ of $\underline{P}_i$, coherence is easily verified for the functional $B(2^\Omega) \to \mathbb{R}$, $f \mapsto \underline{P}(i \mapsto \underline{E}_i(f))$ by using the characterization of coherence on linear spaces, and therefore for its restriction to K as defined in (4). □

This rather abstract lemma can be used to prove results which were formulated as individual theorems in Walley's book (cf. Walley [11, 2.6.3 - 2.6.7] and Maaß [8, Corollary 4.2]). The following lemma generalizes one of these results, namely that convex combinations of coherent lower previsions are again coherent.

Lemma 3 *Let $X \subset CLP(K)$, $\mathcal{A}$ be a σ-algebra over X making all $\tilde{f}$ measurable and μ be a probability measure on $\mathcal{A}$. Then the lower prevision*

$$\underline{P} : K \to \mathbb{R}, \quad \underline{P}(f) := \int \tilde{f}\, d\mu \tag{5}$$

is coherent.

Proof. The integral $\int \cdot\, d\mu$ is of course coherent and applies to functions $X \to \mathbb{R}$, $\underline{P}' \mapsto \tilde{f}(\underline{P}') = \underline{P}'(f)$. Applying Lemma 2 yields the desired result. □

Before proceeding with the main issue of this paper, a possible application of Lemma 2 should be sketched. Suppose I is a nonempty set of persons assigning values in a coherent way to all gambles $f \in K$, i.e. $\{\underline{P}_i\}_{i\in I}$ is an indexed set of coherent lower previsions. Furthermore, suppose we also want to assign values coherently to all $f \in K$ just by incorporating the $\underline{P}_i$. Using the already cited well-known theorems (cf. Walley [11, 2.6.3 - 2.6.7]), we could take the lower envelope of all $\underline{P}_i$, $\inf_{i\in I} \underline{P}_i$, as our coherent lower prevision if we were very cautious. If we had certain opinions on the coherent lower previsions of all persons we also could assign weights λ_i to every $\underline{P}_i$ and take $\sum_{i\in I} \lambda_i \underline{P}_i$ as our coherent lower prevision (cf. Lemma 3). But using Lemma 2 we can go even further. We can assign weights $\mu(J)$ to "coalitions" $J \subset I$ in order to express that if certain persons agree on the evaluation of some gamble f this should count more than the evaluations of other persons. If this set function μ is supermodular then the Choquet integral $\int \cdot\, d\mu$ is coherent and, by Lemma 2, so is the lower prevision $f \mapsto \int (i \mapsto \underline{P}_i)\, d\mu$

By merely collecting the results from Lemma 1, the remarks following Example 1 (especially Equation (3)) and Lemma 3, we obtain the subsequent proposition which contains the essential mathematical part of the main theorem of this paper (Theorem 2).

Proposition 4 *The mapping*

$$\begin{aligned} CLP(K) \;&\to\; \Big\{ \big(\int \cdot\, d\mu\big)_{|\mathrm{span}(\tilde{K})} \mid \mu : \mathrm{ex}(CLP(K)) \cap \mathcal{B}_0 \to \mathbb{R} \textit{ probability measure} \Big\}, \\ \underline{P} \;&\mapsto\; \big(\int \cdot\, d\mu_{\underline{P}}\big)_{|\mathrm{span}(\tilde{K})} \end{aligned} \tag{6}$$

with $\underline{P}(f) = \int \tilde{f}\, d\mu_{\underline{P}}$ *for all* $f \in K$ *is bijective.*

We now expand this first result to the linear spaces spanned by the respective sets used in Proposition 4. Thus, denote by

$$V_1 := \Big\{ \lambda_1 \underline{P}_1 - \lambda_2 \underline{P}_2 \mid \lambda_1, \lambda_2 \geq 0, \underline{P}_1, \underline{P}_2 \in CLP(K) \Big\} \tag{7}$$

the linear space of functionals spanned by $CLP(K)$ and by

$$V_2 := \Big\{ \big(\int \cdot\, d\mu\big)_{|\mathrm{span}(\tilde{K})} \mid \mu : \mathrm{ex}(CLP(K)) \cap \mathcal{B}_0 \to \mathbb{R} \text{ of bounded variation} \Big\}. \tag{8}$$

the linear space of restricted integrals w.r.t. signed measures on $\mathrm{ex}(CLP(K)) \cap \mathcal{B}_0$ of bounded variation. Let V_1 be endowed with $\mathcal{T}_{V_1}$, the relative topology of V_1 generated by $\mathcal{T}$, i.e. the smallest topology making all $\tilde{f}$ restricted to V_1 continuous and let V_2 be endowed with $\mathcal{T}_{V_2}$, the weak* topology, i.e. the smallest topology making all natural embeddings $\tilde{\tilde{f}} : V_2 \to \mathbb{R}$, $\tilde{\tilde{f}}\big((\int \cdot\, d\mu)_{|\mathrm{span}(\tilde{K})}\big) := \int \tilde{f}\, d\mu$ continuous. Further on, let the norm $||\cdot||_{V_1}$ be defined by

$$||\underline{P}||_{V_1} := \inf\{\lambda_1 + \lambda_2 \mid \underline{P} = \lambda_1 \underline{P}_1 - \lambda_2 \underline{P}_2, \lambda_1, \lambda_2 \geq 0, \underline{P}_1, \underline{P}_2 \in CLP(K)\} \tag{9}$$

and the norm $||\cdot||_{V_2}$ be analogously to $||\cdot||_{V_1}$ defined by

$$\begin{aligned} &||\big(\int \cdot\, d\mu\big)_{|\mathrm{span}(\tilde{K})}||_{V_2} \\ &:= \inf\Big\{\lambda_1 + \lambda_2 \mid \big(\int \cdot\, d\mu\big)_{|\mathrm{span}(\tilde{K})} = \lambda_1 \big(\int \cdot\, d\mu_1\big)_{|\mathrm{span}(\tilde{K})} - \lambda_2 \big(\int \cdot\, d\mu_2\big)_{|\mathrm{span}(\tilde{K})}, \\ &\qquad\qquad \lambda_1, \lambda_2 \geq 0, \mu_1, \mu_2 \text{ probability measures}\Big\}. \end{aligned} \tag{10}$$

We defer the easy but technical proof of $||\cdot||_{V_1}$ and $||\cdot||_{V_2}$ really being norms to the end of this section and just proceed with the main result.

Theorem 2 *There is a linear isomorphism J^* between the linear spaces V_1 and V_2. The isomorphism is determined by the identity*

$$\underline{P}(f) = \int \tilde{f}\, d\mu \quad \textit{for all } f \in K. \tag{11}$$

The isomorphism J^* is topological, i.e. a homeomorphism, between the topological spaces $(V_1, \mathcal{T}_{V_1})$ and $(V_2, \mathcal{T}_{V_2})$. The isomorphism J^* is isometric between the normed spaces $(V_1, \|\cdot\|_{V_1})$ and $(V_2, \|\cdot\|_{V_2})$. Moreover, $\underline{P}$ is coherent if and only if its transformed is monotone.

Proof. To prove that J^* is well defined, it suffices to show that for every $\underline{P} \in V_1$ there is a measure $\mu : \mathrm{ex}(CLP(K)) \cap \mathcal{B}_0 \to \mathbb{R}$ of bounded variation with $\underline{P}(f) = \int \tilde{f}\, d\mu$ for all $f \in K$ because uniqueness of the image is guaranteed by Equation (11). Suppose $\underline{P} = \lambda_1 \underline{P}_1 - \lambda_2 \underline{P}_2$ with $\lambda_1, \lambda_2 \geq 0$ and $\underline{P}_1, \underline{P}_2 \in CLP(K)$. Then, by Proposition 4, there exist probability measures μ_1, μ_2 on $\mathrm{ex}(CLP(K)) \cap \mathcal{B}_0$ satisfying $\underline{P}_1(f) = \int \tilde{f}\, d\mu_1$ and $\underline{P}_2(f) = \int \tilde{f}\, d\mu_2$ for all $f \in K$. Thus,

$$\underline{P}(f) = \lambda_1 \underline{P}_1(f) - \lambda_2 \underline{P}_2(f) = \lambda_1 \int \tilde{f}\, d\mu_1 - \lambda_2 \int \tilde{f}\, d\mu_2 = \int \tilde{f}\, d(\lambda_1\mu_1 - \lambda_2\mu_2) \tag{12}$$

for all $f \in K$, i.e. J^* is well defined. Injectivity of J^* directly follows from Equation (11) since $\underline{P}_1 \neq \underline{P}_2$, $\underline{P}_1, \underline{P}_2 \in V_1$, implies $\int \tilde{f}\, d\mu_1 \neq \int \tilde{f}\, d\mu_2$ for all $f \in K$ with $\underline{P}_1(f) \neq \underline{P}_2(f)$ and μ_1 resp. μ_2 satisfying Equation (11) for $\underline{P}_1$ resp. $\underline{P}_1$. Since, by the Hahn-Jordan Decomposition Theorem, every measure μ of bounded variation can be decomposed into a difference $\mu = \lambda_1\mu_1 - \lambda_2\mu_2$, $\lambda_i \geq 0$, μ_i probability measures, $i \in \{1,2\}$, we obtain surjectivity of J^* simply by reading Equation (12) from right to left, again using Proposition 4. Linearity of J^* is rather obvious. So, we have shown that J^* is a linear isomorphism between the linear spaces V_1 and V_2.

By setting $X := K$ and $V := V_1$ in the subsequent Proposition 5, it follows immediately that J^* also is a homeomorphism between the topological spaces $(V_1, \mathcal{T}_{V_1})$ and $(V_2, \mathcal{T}_{V_2})$.

For proving isometry of J^*, we observe that any decomposition of $J^*(\underline{P})$, $J^*(\underline{P}) = \lambda_1 \big(\int \cdot\, d\mu_1\big)_{|\mathrm{span}(\tilde{K})} - \lambda_2 \big(\int \cdot\, d\mu_2\big)_{|\mathrm{span}(\tilde{K})}$, with $\lambda_1, \lambda_2 \geq 0$, μ_1, μ_2 probability measures, directly corresponds to a decomposition of $\underline{P}$ by Proposition 4, $\underline{P} = \lambda_1 \underline{P}_{\mu_1} - \lambda_2 \underline{P}_{\mu_2}$. Therefore, the infima in the respective definitions of $\|\cdot\|_{V_1}$ and $\|\cdot\|_{V_2}$ are taken over the same sets, i.e. $\|J^*(\underline{P})\|_{V_2} = \|\underline{P}\|_{V_1}$ for all $\underline{P} \in V_1$. □

We now provide the deferred, fairly general proposition used in Theorem 2.[2]

[2] This proposition can also be used to prove that the isomorphism between the linear spaces respectively spanned by the totally monotone set functions and the signed bounded Borel measures (cf. Marinacci [10, Theorem 3]) is a homeomorphism. Marinacci proved homeomorphy only for the respective unit balls (w.r.t. the norm which is not compatible to the topology) instead of the whole spaces.

Proposition 5 *Let X be a nonempty set and V a linear space of real-valued functions on X. Define*

$$\begin{array}{rcll} \tilde{X} & := & \{\tilde{x} : V \to \mathbb{R} \mid \tilde{x}(v) := v(x), x \in X\}, & (13)\\ \tilde{V} & := & \{\tilde{v} : \tilde{X} \to \mathbb{R} \mid \tilde{v}(\tilde{x}) := \tilde{x}(v), v \in V\}, & (14)\\ \tilde{\tilde{X}} & := & \{\tilde{\tilde{x}} : \tilde{V} \to \mathbb{R} \mid \tilde{\tilde{x}}(\tilde{v}) := \tilde{v}(\tilde{x}), \tilde{x} \in \tilde{X}.\} & (15) \end{array}$$

Endow $\mathcal{T}_V$ with the smallest topology on V making all $\tilde{x} \in \tilde{X}$ continuous and endow $\mathcal{T}_{\tilde{V}}$ with the smallest topology on $\tilde{V}$ making all $\tilde{\tilde{x}} \in \tilde{\tilde{X}}$ continuous.
Then $J : V \to \tilde{V}$, $v \mapsto \tilde{v}$ is a linear topological isomorphism.

Proof. Linearity and injectivity of J is easily verified by successively applying the definitions of $\tilde{V}$ and $\tilde{X}$. Additionally, by definition of $\tilde{V}$, J is surjective. For proving J being a homeomorphism it suffices to show that the elements of the respective subbase of $\mathcal{T}_V$, $\{\tilde{x}^{-1}(O) \mid O \subset \mathbb{R} \text{ open}\}$ and $\mathcal{T}_{\tilde{V}}$, $\{\tilde{\tilde{x}}^{-1}(O) \mid O \subset \mathbb{R} \text{ open}\}$, are mapped onto each other as preimages under J and J^{-1}. This follows almost directly from the above definitions since

$$\begin{aligned} J^{-1}(\tilde{\tilde{x}}^{-1}(O)) &= J^{-1}(\{\tilde{v} \mid \tilde{\tilde{x}}(\tilde{v}) \in O\})\\ &= J^{-1}(\{\tilde{v} \mid \tilde{x}(v) \in O\})\\ &= J^{-1}(\{\tilde{v} \mid v \in \tilde{x}^{-1}(O)\})\\ &= \tilde{x}^{-1}(O) \end{aligned}$$

and analogously $J(\tilde{x}^{-1}(O)) = \tilde{\tilde{x}}^{-1}(O)$. □

We end this section with a lemma proving the function $||\cdot||_{V_1}$ and $||\cdot||_{V_2}$ in fact being norms.

Lemma 4 *The functions $||\cdot||_{V_1}$ and $||\cdot||_{V_2}$ are norms on the respective spaces.*

Proof. Obviously, $||0||_{V_1} = 0$. Now suppose $||\underline{P}|| = 0$ for a $\underline{P} \in V_1$. Since all $f \in K$ are bounded and since every coherent lower prevision maps f into the bounded interval $[\inf f, \sup f]$, we obtain $|\underline{P}(f)| < \varepsilon$ for every $\varepsilon > 0$, i.e. $\underline{P} = 0$. Further on, for all $\underline{P} \in V_1$ and $c \in \mathbb{R}$, $c \neq 0$,

$$\begin{aligned} ||c\underline{P}|| &= \inf\left\{\lambda_1 + \lambda_2 \mid c\underline{P} = \lambda_1\underline{P}_1 - \lambda_2\underline{P}_2, \lambda_1, \lambda_2 \geq 0, \underline{P}_1, \underline{P}_2 \in CLP(K)\right\}\\ &= \inf\left\{|c|\tfrac{\lambda_1+\lambda_2}{|c|} \mid \underline{P} = \tfrac{\lambda_1}{c}\underline{P}_1 - \tfrac{\lambda_2}{c}\underline{P}_2, \lambda_1, \lambda_2 \geq 0, \underline{P}_1, \underline{P}_2 \in CLP(K)\right\}\\ &= |c| \inf\left\{\tfrac{\lambda_1}{|c|} + \tfrac{\lambda_2}{|c|} \mid \underline{P} = \tfrac{\lambda_1}{|c|}\underline{P}_1 - \tfrac{\lambda_2}{|c|}\underline{P}_2, \lambda_1, \lambda_2 \geq 0, \underline{P}_1, \underline{P}_2 \in CLP(K)\right\}\\ &= |c| \cdot ||\underline{P}||. \end{aligned}$$

Finally, the triangle inequality holds because whenever $\underline{P} = \underline{P}_1 + \underline{P}_2$ with $\underline{P}, \underline{P}_1, \underline{P}_2 \in V_1$, $\underline{P}_1 = \lambda_{1,1}\underline{P}_{1,1} - \lambda_{1,2}\underline{P}_{1,2}$, $\underline{P}_2 = \lambda_{2,1}\underline{P}_{2,1} - \lambda_{2,2}\underline{P}_{2,2}$, $\underline{P}_{i,j} \in CLP(K)$

and $\lambda_{i,j} \geq 0$ with $i,j \in \{1,2\}$ then $\underline{P} = (\lambda_{1,1}\underline{P}_{1,1} + \lambda_{2,1}\underline{P}_{2,1}) - (\lambda_{1,2}\underline{P}_{1,2} + \lambda_{2,2}\underline{P}_{2,2})$ holds whereat $(\lambda_{1,1}\underline{P}_{1,1} + \lambda_{2,1}\underline{P}_{2,1}), (\lambda_{1,2}\underline{P}_{1,2} + \lambda_{2,2}\underline{P}_{2,2}) \in CLP(K)$. Therefore, $\|\underline{P}_1 + \underline{P}_2\| \leq \|\underline{P}'\| + \|\underline{P}''\|$. □

4 Summary, Outlook and Open Problems

In this paper we have presented a linear isomorphism between the linear space V_1, spanned by the coherent lower previsions on an arbitrary nonempty set K and an appropriate linear space V_2 of continuous linear functionals. Thereby, we have shown that the famous representation theorems for totally monotone set functions do not depend on this special class, not even on the structure of the domain.

For applications, we are heavily interested in transformations of coherent lower previsions that can practically be handled. It is well-known that the set of extreme points of the set of normalized totally monotone set functions on a finite algebra is finite and consists of all unanimity games which is a finite set. Therefore, every totally monotone set function on a finite domain can be represented as a convex combination of unanimity games. It remains as an open problem to determine the set of extreme points of $CLP(K)$ for a given K. Additionally, for possible application of Theorem 2, it remains as an open problem what condition K has to meet in order to make $\mathrm{ex}(CLP(K))$ finite.

Theorem 2 can be used to construct coherent lower previsions in the following way. After determining the extreme points of the convex set of coherent lower previsions any coherent lower prevision can be obtained by assigning weights to all extreme points. There is an analogous situation in Dempster-Shafer Theory where these weights are called "basic probability assignments". So, by working on the set of extreme points of $CLP(K)$ with linear functionals, things are getting easier and often more applicable.

Finally, we will outline why the transform given in Theorem 2 should not be called "Möbius transform" like in the case of totally monotone set functions. On an algebra $\mathcal{A}$ the zeta function can be expressed in terms of unanimity games, $\zeta(A,B) := u_A(B)$. In the case of considering totally monotone set functions on a finite algebra instead of coherent lower previsions (cf. Denneberg [5], Gilboa and Schmeidler [6] and Marinacci [10]), the integrand of Equation (11) is always a zeta function because the set of extreme points of the set of normalized totally monotone set functions consists of all unanimity games. This gives rise to call the two set functions appearing in the transformation equation the zeta transform resp., since the zeta function and the Möbius function are mutually inverse, the Möbius transform of the respective other set function. Since we have seen in Example 1 that the set of extreme points of $CLP(\mathcal{A})$ contains more than unanimity games the interpretation of using zeta functions can not be preserved such that the term "Möbius transform" can not be justified in our case.

Acknowledgements

The author is indebted to D. Denneberg for valuable discussions and to the anonymous referees for their helpful suggestions and comments.

References

[1] E. M. Alfsen. *Compact Convex Sets and Boundary Integrals*. Springer, Berlin, 1971.

[2] E. Bishop and K. de Leeuw. The Representation of Linear Functionals by Measures on Sets of Extreme Points. *Annales de l'Institut Fourier*, 9:305–331, 1959.

[3] G. Choquet. Theory of Capacities. *Annales de l'Institut Fourier*, 5:131–295, 1953/54.

[4] F. Delbaen. Coherent Risk Measures on General Probability Spaces. 1–37 in K. Sandmann and P. J. Schönbucher. *Advances in Finance and Stochastics*, Springer, Berlin, 2002.

[5] D. Denneberg. Representation of the Choquet Integral with the σ-additive Möbius Transform. *Fuzzy Sets and Systems*, 92:139–156, 1997.

[6] I. Gilboa and D. Schmeidler. Canonical Representation of Set Functions. *Mathematics of Operations Research*, 20:197–212, 1995.

[7] I. Gilboa and D. Schmeidler. Maxmin Expected Utility with non-unique Prior. *Journal of Mathematical Economics*, 18:141–153, 1989.

[8] S. Maaß. Coherent Lower Previsions as Exact Functionals and their (Sigma-)Core. In *ISIPTA'01: Proceedings of the Second International Symposium on Imprecise Probabilitics and their Applications*, Shaker, Maastricht, 230–236, 2001.

[9] S. Maaß. Exact Functionals and their Core. *Statistical Papers*, 43:75–93, 2002.

[10] M. Marinacci. Decomposition and Representation of Coalitional Games. *Mathematics of Operations Research*, 21:1000–1015, 1996.

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

Sebastian Maaß is with the Department of Mathematics, Universität Bremen, Germany.
E-mail: Sebastian.Maass@web.de

Study of the Probabilistic Information of a Random Set[*]

ENRIQUE MIRANDA
University of Oviedo, Spain

INÉS COUSO
University of Oviedo, Spain

PEDRO GIL
University of Oviedo, Spain

Abstract

Given a random set coming from the imprecise observation of a random variable, we study how to model the information about the distribution of this random variable. Specifically, we investigate whether the information given by the upper and lower probabilities induced by the random set is equivalent to the one given by the class of the distributions of the measurable selections; together with sufficient conditions for this, we also give examples showing that they are not equivalent in all cases.

Keywords

random sets, upper and lower probabilities, measurable selections, Choquet integral

1 Introduction

Random sets have been successfully applied in such different fields as economy ([11]) or stochastic geometry ([14]), and they have been studied under different interpretations, like the behavioral ([19]) or the evidential one ([7]). In this paper, we will interpret a random set as the result of the imprecise observation of a random variable ([13]). Under this interpretation, our information about the probability distribution of the random variable is given by the class of distributions of the *measurable selections* of the random set. This class of distributions is a subset of the class of probability measures bounded between the *upper* and *lower* probabilities ([7]) of the random set. These functions satisfy Walley's axioms of

[*] This paper has been partially funded by FEDER-MCYT, grant number BFM2001-3515, and Fundación Banco Herrero.

coherence ([21]), and are moreover ∞-alternating and ∞-monotone, respectively ([20]).

Although working with the upper and lower probabilities leads to a number of mathematical simplifications ([20, 21]), the information they provide is in general more imprecise than the one given by the set of distributions of the measurable selections ([16, 18]). In this paper, we will investigate under which conditions these two models are equivalent. The results we obtain will show when it is advisable to model our information through the upper and lower probabilities and when this produces a loss of precision.

In Section 2, we introduce some concepts and notations that we will use in the rest of the paper, and recall some previous works on the subject. In Section 3, we investigate the information that the upper and lower probabilities give about the distribution of the original random variable, and about the value of this distribution on an arbitrary set. Finally, in Section 4 we give some additional comments and remarks.

2 Preliminary concepts

We will consider a probability space $(\Omega, \mathcal{A}, P)$, a measurable space $(X, \mathcal{A}')$ and a multi-valued mapping $\Gamma : \Omega \to \mathcal{P}(X)$. If X is a topological space, we will denote β_X its Borel σ-field. A topological space is said to be *Polish* when it is separable and complete for some compatible metric d, and it is called *Souslin* if it is the bijective image of a Polish space. The multi-valued mapping will be called open (resp. complete, closed, compact) if $\Gamma(\omega)$ is an open (resp. complete, closed, compact) subset of X for every $\omega \in \Omega$.

Formally, a random set is a multi-valued mapping satisfying some measurability condition. There are different conditions, such as the weak, the strong, or the graph-measurability ([12]). Most of them are based on the notion of upper and lower inverse:

Definition 1 *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $(X, \mathcal{A}')$ be a measurable space and $\Gamma : \Omega \to \mathcal{P}(X)$ a multi-valued mapping. Given $A \in \mathcal{A}'$, its* **upper inverse** *is $\Gamma^*(A) = \{\omega \in \Omega \mid \Gamma(\omega) \cap A \neq \emptyset\}$, and its* **lower inverse** *is $\Gamma_*(A) = \{\omega \in \Omega \mid \emptyset \neq \Gamma(\omega) \subseteq A\}$.*

When there is no possible confusion about the multi-valued mapping we are working with, we will use the notation $A^* := \Gamma^*(A)$ and $A_* := \Gamma_*(A)$. By a **random set** we will mean throughout a strongly measurable multi-valued mapping. The strong measurability is necessary for the upper and lower probabilities of the random set to be defined on $\mathcal{A}'$.

Definition 2 *A multi-valued mapping is called* **strongly measurable** *if $A^* \in \mathcal{A}$ $\forall A \in \mathcal{A}'$.*

Note that $A_* = X^* \cap ((A^c)^*)^c\ \forall A \in \mathcal{A}'$, whence if Γ is strongly measurable, we also have $A_* \in \mathcal{A}\ \forall A \in \mathcal{A}'$. The concepts of upper and lower probabilities induced by a random set were introduced by Dempster in [7]:

Definition 3 *Given a random set $\Gamma : \Omega \to \mathcal{P}(X)$, the* **upper probability** *of $A \in \mathcal{A}'$ is $P^*_\Gamma(A) = \frac{P(A^*)}{P(X^*)}$, and its* **lower probability** *is $P_{*\Gamma}(A) = \frac{P(A_*)}{P(X^*)}$.*

When there is no ambiguity about which random set is inducing the upper and lower probability, we will denote $P^* := P^*_\Gamma$ and $P_* := P_{*\Gamma}$.

As we said in the introduction, we will regard a random set as the result of the imprecise observation of a random variable $U_0 : \Omega \to X$ (which we will call *original* random variable), in the sense that for every ω in the initial space all we know about $U_0(\omega)$ is that it belongs to the set $\Gamma(\omega)$. As a consequence, $\Gamma(\omega)$ will be assumed to be non-empty for every ω, and hence $P^*(A) = P(A^*)$ and $P_*(A) = P(A_*)$ for all $A \in \mathcal{A}'$. The upper and lower probabilities induced by a random set are conjugate functions, and they are moreover ∞-alternating and ∞-monotone capacities, respectively ([20]). This means in particular that they satisfy Walley’s axioms of coherence ([21]).

If Γ is the imprecise observation of U_0, all we know about this variable is that it belongs to the class of **measurable selections** (or **selectors**) of Γ,

$$S(\Gamma) := \{U : \Omega \to X \text{ measurable} \mid U(\omega) \in \Gamma(\omega)\ \forall \omega\}.$$

The probability distribution of U_0 belongs to

$$P(\Gamma) := \{P_U \mid U \in S(\Gamma)\},$$

and our information about $P_{U_0}(A)$ is given by the set of values

$$P(\Gamma)(A) := \{P_U(A) \mid U \in S(\Gamma)\}.$$

There are two other classes of probabilities that may be useful in some situations. The first one is

$$\Delta(\Gamma) := \{Q \text{ probability} \mid Q(A) \in P(\Gamma)(A)\ \forall A \in \mathcal{A}'\}.$$

This is the set of distributions whose values are compatible with the information given by the random set. It is clear that $P(\Gamma) \subseteq \Delta(\Gamma)$. If they coincide, the information about the distribution of the original random variable is equivalent to the information about the values it takes. On the other hand, we can also consider the class

$$M(P^*) := \{Q \text{ probability} \mid Q(A) \leq P^*(A)\ \forall A \in \mathcal{A}'\}$$

of distributions dominated by P^*, or **credal set** generated by P^*. This class is convex and easier to handle in practice than $P(\Gamma)$. Using the inequalities $P_*(A) \leq$

$P_U(A) \leq P^*(A)$, valid for any $U \in S(\Gamma),\ A \in \mathcal{A}'$, we deduce that $\Delta(\Gamma) \subseteq M(P^*)$. We see then that $P(\Gamma) \subseteq \Delta(\Gamma) \subseteq M(P^*)$. As we showed in [16], both inclusions can be strict, and in some cases the use of the upper and lower probabilities can produce a loss of precision, which in turn can cause some misjudgements. It is therefore interesting to see in which cases it is reasonable to use P^* and P_*.

Although the class of the distributions of the selectors of a random set ([1, 9]) and the upper probability it induces ([14, 20]) have been thoroughly studied in the literature, the connection between them has not received much attention. It was investigated for the case of X finite in [16], and for some particular infinite spaces in [3, 6, 10, 15, 18]. Our goal in this paper is to somewhat fill this gap. Specifically, we will study two different problems:

- First, we will investigate the relationship between $\Delta(\Gamma)$ and $M(P^*)$, which tells us if the upper and the lower probabilities are informative enough about the value $P_{U_0}(A)$ for some arbitrary $A \in \mathcal{A}'$.
- Then, we will study when $P(\Gamma) = M(P^*)$, i.e., under which conditions the upper probability keeps all the information about P_{U_0}.

3 Study of the probabilistic models for P_{U_0}

$P^*(A), P_*(A)$ as a model for $P_{U_0}(A)$. Let us start investigating the relationship between $\Delta(\Gamma)$ and $M(P^*)$. As we mentioned before, $\Delta(\Gamma)$ models the information that Γ gives about the probability values of the elements in $\mathcal{A}'$. Therefore, by investigating its equality with $M(P^*)$ we will see whether P^* and P_* are informative enough about the 'true' probability of an arbitrary set A. This is formally stated in the following proposition.

Proposition 1 *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $(X, \mathcal{A}')$ a measurable space and $\Gamma : \Omega \to \mathcal{P}(X)$ a random set. Then,*

$$\Delta(\Gamma) = M(P^*) \Leftrightarrow P(\Gamma)(A) = [P_*(A), P^*(A)]\ \forall A \in \mathcal{A}'.$$

Let us consider then some arbitrary $A \in \mathcal{A}'$, and let us study the relationship between $P(\Gamma)(A)$ and $[P_*(A), P^*(A)]$. It is clear that the latter is a superset of the former. In order to give conditions for the equality, we must see if the maximum and minimum values of $P(\Gamma)(A)$ coincide with $P^*(A)$ and $P_*(A)$, and also if $P(\Gamma)(A)$ is convex.

This problem was studied in [18]. We showed there that $P(\Gamma)(A)$ has a maximum and a minimum value (it is indeed a closed subset of $[0,1]$), and that these values do not coincide in all cases with $P^*(A), P_*(A)$, even in the non-trivial case of $S(\Gamma) \neq \emptyset$. Moreover, $P(\Gamma)(A)$ is not convex in general. The following theorem gives sufficient conditions for the equalities $P^*(A) = \max P(\Gamma)(A)$ and $P_*(A) = \min P(\Gamma)(A)$. It generalizes previous results from [6].

Theorem 1 *[18] Consider $(\Omega,\mathcal{A},P)$ a probability space, (X,τ) a topological space and $\Gamma:\Omega\to\mathcal{P}(X)$ a random set. Under any of the following conditions:*

1. *Ω is complete, X is Souslin and $Gr(\Gamma)\in\mathcal{A}\otimes\beta_X$,*
2. *X is a separable metric space and Γ is compact,*
3. *X is a separable metric space and Γ is open,*
4. *X is a Polish space and Γ is closed,*
5. *X is a σ-compact metric space and Γ is closed,*

we have $P^*(A)=\max P(\Gamma)(A)$ and $P_*(A)=\min P(\Gamma)(A)\ \forall A\in\beta_X$. Moreover, if

6. *X is a separable metric space and Γ is complete,*

then $P^*(A)=\max P(\Gamma)(A), P_*(A)=\min P(\Gamma)(A)\ \forall A\in Q(\{B_n\}_n)$, where $\{B_n\}_n=\{B(x_i;q_j)\mid i\in\mathbb{N},q_j\in\mathbb{Q}\}$ is a countable basis of $\tau(d)$ associated to a countable dense set $\{x_n\}_n$ and $Q(\{B_n\}_n)$ is the field generated by $\{B_n\}_n$.

This theorem gives sufficient conditions for the equalities $P^*=\max P(\Gamma)$ and $P_*=\min P(\Gamma)$. The coherence of P^* implies ([21]) that it is the upper envelope of the set of the finitely additive probabilities it dominates. We have proven that, under conditions (1) to (5) from Theorem 1, it is indeed the upper envelope of the class of *countably* additive probabilities induced by the selectors. A similar (symmetrical) remark can be made for P_*.

Let us remark in passing that results established in Theorem 1 guarantee the existence of a selector of Γ whose distribution coincides with P^* on a finite chain. Indeed, in [5] Couso showed that the equality $P^*(A)=\sup P(\Gamma)(A)\ \forall A\in\mathcal{A}'$ implies the equality between the Choquet integral of a bounded random variable respect to the upper probability of a random set ([8]) and the supremum of class of the integrals respect to the distributions of the measurable selections. This allows us to deduce the following result, which generalizes theorem 1 from [3].

Theorem 2 *Let $\Gamma:\Omega\to\mathcal{P}(X)$ be a random set and $V:X\to\mathbb{R}$ a bounded random variable. Under any of the conditions (1) to (5) from the previous theorem, $(C)\int V dP^*=\sup\{\int V dP_U\mid U\in S(\Gamma)\}$ and $(C)\int V dP_*=\inf\{\int V dP_U\mid U\in S(\Gamma)\}$.*

On the other hand, we have already remarked that the equality between $\Delta(\Gamma)$ and $M(P^*)$ relies on the equalities $P^*(A)=\max P(\Gamma)(A)$ and on the convexity of $P(\Gamma)(A)$ for every $A\in\mathcal{A}'$. Concerning the latter, we have proven the following:

Proposition 2 *[18] Let $\Gamma:\Omega\to\mathcal{P}(X)$ be a random set, and consider $A\in\mathcal{A}'$. Let $U_1,U_2\in S(\Gamma)$ satisfy $P_{U_1}(A)=\max P(\Gamma)(A)$, $P_{U_2}(A)=\min P(\Gamma)(A)$. Then,*

$$P(\Gamma)(A) \textit{ is convex} \Leftrightarrow U_1^{-1}(A) \setminus U_2^{-1}(A) \textit{ is not an atom } ^1.$$

In particular, $P(\Gamma)(A)$ is convex $\forall A \in \mathcal{A}'$ if the initial space is non-atomic; this condition holds for instance if we have some additional information stating that P_{U_0} is continuous. Nevertheless, the non-atomicity of $(\Omega, \mathcal{A}, P)$ is not necessary for $P(\Gamma)(A)$ to be convex, as we showed in [16]. If we join Theorem 1 and Proposition 2, we derive the following corollary:

Corollary 1 *Let $\Gamma : \Omega \to \mathcal{P}(X)$ be a random set satisfying any of the conditions (1) to (5) from Theorem 1. If $A^* \setminus A_*$ is not an atom for any $A \in \beta_X$, $\Delta(\Gamma) = M(P^*)$.*

P^*, P_* as a model for P_{U_0}. Let us study now the equality between $P(\Gamma)$ and $M(P^*)$, which tells whether the upper probability keeps all the available information about the distribution of the original random variable, P_{U_0}. The class of the distributions of the selectors has been studied for some types of random sets (see for instance [1, 9, 10]). However, its relationship with the credal set generated by the upper probability has not been investigated in detail. In [16], we studied this problem for the case of X finite, and in [15] the attention was focused on random intervals. On the other hand, Castaldo and Marinacci proved in [3] a result for compact random sets on Polish spaces.

The equality between $\Delta(\Gamma)$ and $M(P^*)$ does not guarantee that $P(\Gamma) = M(P^*)$, and neither does the equality between $P(\Gamma)$ and $\Delta(\Gamma)$ ([16]). Then, a possible approach for our problem would be determining sufficient conditions for $P(\Gamma) = \Delta(\Gamma)$, and join them with the ones stated in Corollary 1. Unfortunately, it does not seem easy (except in trivial situations) to characterize this last equality. We are going to show that a reasoning based on the extreme points of $M(P^*)$ will be more fruitful in our context: it allows us to easily characterize the equality between $P(\Gamma)$ and $M(P^*)$ in the finite case, and we can use this to derive some results for the case of X separable metric. When X is finite, the extreme points of $M(P^*)$ are in correspondence with the permutations on X, in the following manner[2]:

Theorem 3 *[4] Consider $X = \{x_1, \dots, x_n\}$ finite and μ a 2-alternating capacity on $\mathcal{P}(X)$. For any $\pi \in S^n$, define a probability Q_π on $\mathcal{P}(X)$ satisfying*

$$Q_\pi(\{x_{\pi(1)}, \dots, x_{\pi(j)}\}) = \mu(\{x_{\pi(1)}, \dots, x_{\pi(j)}\}) \forall j = 1, \dots, n.$$

Then, $Ext(M(\mu)) = \{Q_\pi \mid \pi \in S^n\}$ and $M(\mu) = Conv(\{Q_\pi \mid \pi \in S^n\})$.

We can see ([16]) that given X finite and $\Gamma : \Omega \to \mathcal{P}(X)$ a random set, it is $Ext(M(P^*)) \subseteq P(\Gamma)$, and as a consequence $P(\Gamma) = M(P^*) \Leftrightarrow P(\Gamma)$ is convex.

[1] By this we mean that for every $\alpha \in (0,1)$ there is some measurable $B \subseteq U_1^{-1}(A) \setminus U_2^{-1}(A)$ with $P(B) = \alpha P(U_1^{-1}(A) \setminus U_2^{-1}(A))$.

[2] This theorem is an extension, for 2-alternating capacities, of a result given by Dempster ([7]) for random sets on finite spaces.

This equivalence does not hold for the case of X infinite, as the following example shows:

Example 1 (sketch) *Let $\Gamma : (0,1) \to \mathcal{P}([0,1])$ be defined between the probability space $((0,1), \beta_{(0,1)}, \lambda_{(0,1)})$ and the measurable space $([0,1], \beta_{[0,1]})$ by $\Gamma(\omega) = (0,\omega)\ \forall \omega \in (0,1)$. It is easy to see that this mapping is strongly measurable.*

- *Given $U \in S(\Gamma)$, it can be checked that $P_U(\{0\}) = 0, P_U([0,x]) \geq x\ \forall x$, and $\lambda_{(0,1)}(\{x \in (0,1) \mid P_U([0,x]) = x\}) = 0$.*

- *Conversely, consider a probability measure $Q : \beta_{[0,1]} \to [0,1]$ satisfying the three previous properties. This implies that it also satisfies $Q([0,x)) \geq x$ and $Q([0,x)) > x$ for all but a null subset of $(0,1)$, that we will denote N_Q. The quantile function U of Q is a measurable mapping satisfying $P_U = Q, U(\omega) \in \Gamma(\omega)\ \forall \omega \notin N_Q$. We can modify U on N_Q without affecting its measurability so that all its values are included in those of Γ, whence we deduce that $Q \in P(\Gamma)$.*

- *We deduce that $P(\Gamma)$ is the class of probability measures with $Q(\{0\}) = 0, Q([0,x]) \geq x\ \forall x$ and $Q([0,x]) > x$ for all but a null subset of $[0,1]$, and we can easily check that this class is convex.*

- *The Lebesgue measure $\lambda_{[0,1]}$ on $\beta_{[0,1]}$ satisfies $\lambda_{[0,1]}(A) \leq P^*(A)\ \forall A \in \beta_{[0,1]}$; hence, it belongs to $M(P^*)$, and clearly it does not satisfy $\lambda_{[0,1]}([0,x]) > x$ with probability 1. As a consequence, $P(\Gamma) \subsetneq M(P^*)$.*

In [17], we investigated the form of the extreme points of $M(\mu)$ for the case of μ 2-alternating and upper continuous, and for (X,d) a separable metric space. The idea in that paper was to approximate a distribution $Q : \beta_X \to [0,1]$ by distributions coinciding with Q on some finite fields. We will use a similar reasoning in our next theorem, where we consider the upper probability P^* induced by a random set (and hence not necessarily upper continuous). We will work in this paper with the topology of the weak convergence, whose main properties can be found in [2]. Together with the well-known Portmanteau's theorem, we will also use the following result:

Proposition 3 *[2] Let (X,d) be a separable metric space, and consider a class $\mathcal{U} \subseteq \beta_X$ such that (i) it is closed under finite intersections and (ii) every open set is a finite or countable union of elements from $\mathcal{U}$. Let $\{P_n\}_n, P$ be a family of probability measures on β_X such that $P_n(A) \to P(A)\ \forall A \in \mathcal{U}$. Then, the sequence $\{P_n\}_n$ converges weakly to P.*

Let $\{x_n\}_n$ be a countable set dense on (X,d), and define $\{B_n\}_n := \{B(x_i;q_j) \mid i \in \mathbb{N}, q_j \in \mathbb{Q}\}$ a countable basis of $\tau(d)$. Let us denote $\mathcal{Q}(\{B_n\}_n)$ the field generated by $\{B_n\}_n$, $\mathcal{Q}_n$ the field generated by $\{B_1,\ldots,B_n\}$. Then, $\mathcal{Q}(\{B_n\}_n) = \cup_n \mathcal{Q}_n$, and it can easily be checked that $\mathcal{Q}(\{B_n\}_n)$ satisfies the hypotheses (i) and (ii) stated in the previous proposition. Any element of $\mathcal{Q}_n$ is a (finite and disjoint) union of elements from $\mathcal{D}_n := \{C_1 \cap C_2 \cap \cdots \cap C_n \mid C_i \in \{B_i, B_i^c\} \forall i : 1, \ldots, n\}$. Let us denote this class $\mathcal{D}_n := \{E_1^n, \ldots, E_{k_n}^n\}$.

Theorem 4 *Let $(\Omega, \mathcal{A}, P)$ be a probability space, (X, d) a separable metric space and $\Gamma : \Omega \to \mathcal{P}(X)$ a random set such that $P^*(A) = \max P(\Gamma)(A)$ $\forall A \in \mathcal{Q}(\{B_n\}_n)$. Then,*

1. $\overline{M(P^*)} = \overline{Conv(P(\Gamma))}$.
2. $\overline{P(\Gamma)} = \overline{M(P^*)} \Leftrightarrow \overline{P(\Gamma)}$ *is convex.*

Proof.

1. It is clear that $\overline{Conv(P(\Gamma))} \subseteq \overline{M(P^*)}$. Conversely, consider $Q_1 \in M(P^*)$, and fix $n \in \mathbb{N}$. Consider the finite measurable space $(\mathcal{D}_n, \mathcal{P}(\mathcal{D}_n))$, and let us define the multi-valued mapping

$$\begin{array}{rcl} \Gamma_n : \Omega & \to & \mathcal{P}(\mathcal{D}_n) \\ \omega & \hookrightarrow & \{E_i^n \mid \Gamma(\omega) \cap E_i^n \neq \emptyset\}. \end{array}$$

- Given $I \subseteq \{1, \ldots, k_n\}, \Gamma_n^*(\{E_i^n\}_{i \in I}) = \{\omega \mid \exists i \in I, E_i^n \in \Gamma_n(\omega)\} = \{\omega \mid \exists i \in I, \Gamma(\omega) \cap E_i^n \neq \emptyset\} = \Gamma^*(\cup_{i \in I} E_i^n) \in \mathcal{A} \Rightarrow \Gamma_n$ is strongly measurable.
- Define a probability measure $Q : \mathcal{P}(\mathcal{D}_n) \to [0,1]$ by $Q(\{E_i^n\}) = Q_1(E_i^n)$ $\forall i$. Then, given $I \subseteq \{1, \ldots, k_n\}$,

$$Q(\{E_i^n\}_{i \in I}) = Q_1(\cup_{i \in I} E_i^n) \leq P_\Gamma^*(\cup_{i \in I} E_i^n) = P_{\Gamma_n}^*(\{E_i^n\}_{i \in I}),$$

whence $Q \in M(P_{\Gamma_n}^*)$.

Now, from Theorem 3 $M(P_{\Gamma_n}^*) = Conv(\{Q_\pi \mid \pi \in S^{k_n}\})$, where the probability measure $Q_\pi : \mathcal{P}(\mathcal{D}_n) \to [0,1]$ is defined by $Q_\pi(\{E_{\pi(1)}^n, \ldots, E_{\pi(j)}^n\}) = P_{\Gamma_n}^*(\{E_{\pi(1)}^n, \ldots, E_{\pi(j)}^n\}) = P_\Gamma^*(\cup_{i=1}^j E_{\pi(j)}^n)$ $\forall j = 1, \ldots, k_n$.

For any of these extreme points, there is some $P_\pi \in P(\Gamma)$ with $P_\pi(E_j^n) = Q_\pi(\{E_j^n\})$ $\forall j = 1, \ldots, k_n$: it suffices to take into account that, as we have seen in Theorem 2, we can approximate P_Γ^* on a finite chain. As a consequence, for the probability $Q \in Conv(\{Q_\pi \mid \pi \in S^n\})$ defined above, there is some $P_n \in Conv(P(\Gamma))$ such that $P_n(E_j^n) = Q(\{E_j^n\}) = Q_1(E_j^n)$ $\forall j = 1, \ldots, k_n$. The sequence $\{P_n\}_n \subseteq Conv(P(\Gamma))$ satisfies $P_n(A) \to Q_1(A)$ for all $A \in \mathcal{Q}(\{B_n\}_n)$. Applying Proposition 3, we conclude that $\{P_n\}_n$ converges weakly to Q_1, whence $M(P^*) \subseteq \overline{Conv(P(\Gamma))}$ and we deduce the desired equality.

2. For the direct implication, it suffices to see that $\overline{M(P^*)}$ is convex. Consider $P_1, P_2 \in \overline{M(P^*)}, \alpha \in (0,1)$; then, there are $\{P_n^1\}_n, \{P_n^2\}_n \subset M(P^*)$ converging weakly to P_1, P_2, respectively. Let $A \in \beta_X$ be a $(\alpha P_1 + (1-\alpha)P_2)$-continuity set. It is $0 = (\alpha P_1 + (1-\alpha)P_2)(\delta(A))$[3] $= \alpha P_1(\delta(A)) + (1-\alpha)P_2(\delta(A))$, and therefore A is also a P_1, P_2-continuity set. From Portmanteau's theorem (see for instance [2]), $P_n^1(A) \to P_1(A)$ and $P_n^2(A) \to P_2(A)$, whence $(\alpha P_n^1 + (1-\alpha)P_n^2)(A) \to (\alpha P_1 + (1-\alpha)P_2)(A)$, and again using Portmanteau's theorem we deduce that the sequence $\{\alpha P_n^1 + (1-\alpha)P_n^2\}_n \subset M(P^*)$ converges weakly to $\alpha P_1 + (1-\alpha)P_2$. Hence, this probability belongs to $\overline{M(P^*)}$.

   For the converse implication, assume that $\overline{P(\Gamma)}$ is convex. Then, applying the first point of this theorem, it is

$$\overline{M(P^*)} = \overline{Conv(P(\Gamma))} \subseteq \overline{Conv(\overline{P(\Gamma)})} = \overline{\overline{P(\Gamma)}} = \overline{P(\Gamma)} \Rightarrow \overline{P(\Gamma)} = \overline{M(P^*)}.$$

□

The second part of this theorem extends a result mentioned before for the finite case (it can be checked that in that case both $P(\Gamma)$ and $M(P^*)$ are closed). We deduce that a way to determine conditions for the equality $\overline{M(P^*)} = \overline{P(\Gamma)}$ is to give sufficient conditions for the convexity of $\overline{P(\Gamma)}$. We have done this in our next theorem. It uses the following supporting result.

Lemma 1 *Let $(\Omega, \mathcal{A}, P)$ be a non-atomic probability space, (X,d) a separable metric space and $\Gamma : \Omega \to \mathcal{P}(X)$ a random set. Then, the class of probabilities $\mathcal{H}_n := \{Q : \mathcal{P}(\mathcal{D}_n) \to [0,1]$ probability $\mid \exists Q' \in P(\Gamma)$ such that $Q(\{E_i^n\}) = Q'(E_i^n)\ \forall E_i^n \in \mathcal{D}_n\}$ is convex for every n.*

Proof. Fix $n \in \mathbb{N}$, and consider $P_1, P_2 \in \mathcal{H}_n, \alpha \in (0,1)$. Then, there exist $U_1, U_2 \in S(\Gamma)$ with $P_{U_1}(E_i^n) = P_1(\{E_i^n\}), P_{U_2}(E_i^n) = P_2(\{E_i^n\})\ \forall i = 1, \ldots, k_n$. Let us consider the measurable partition of Ω given by $\{C_{ij} \mid i,j = 1, \ldots, k_n\}$ with $C_{ij} = U_1^{-1}(E_i^n) \cap U_2^{-1}(E_j^n)$; from the non-atomicity of $(\Omega, \mathcal{A}, P)$, there is, for every i,j, some measurable $D_{ij} \subseteq C_{ij}$ such that $P(D_{ij}) = \alpha P(C_{ij})$. Define $C = \cup_{i,j} C_{ij}$, and

$$U := U_1 I_C + U_2 I_{C^c}.$$

Then, U is a measurable combination of selectors of Γ, whence $U \in S(\Gamma)$. Moreover,

[3] $\delta(A)$ denotes here the boundary of the set A.

$$\begin{aligned} P_U(E_l^n) &= P(U_1^{-1}(E_l^n)\cap C) + P(U_2^{-1}(E_l^n)\cap C^c) \\ &= \sum_{i=1}^{k_n} P(D_{li}) + \sum_{j=1}^{k_n} (P(C_{jl}) - P(D_{jl})) \\ &= \sum_{i=1}^{k_n} \alpha P(C_{li}) + \sum_{j=1}^{k_n} (1-\alpha)P(C_{jl}) \\ &= \alpha P_{U_1}(E_l^n) + (1-\alpha)P_{U_2}(E_l^n)\ \forall l = 1,\dots,k_n, \end{aligned}$$

and we deduce that $\alpha P_1 + (1-\alpha)P_2 \in \mathcal{H}_n$. □

Theorem 5 *Let $(\Omega,\mathcal{A},P)$ be a non-atomic probability space, (X,d) a separable metric space and $\Gamma:\Omega\to\mathcal{P}(X)$ a random set. Then, $\overline{P(\Gamma)}$ is convex.*

Proof. Let us show first that $Conv(P(\Gamma)) \subseteq \overline{P(\Gamma)}$. Consider $P_1,P_2 \in P(\Gamma), \alpha \in (0,1)$. Applying the previous lemma, for every n there is $Q_n \in P(\Gamma)$ with $Q_n(A) = (\alpha P_1 + (1-\alpha)P_2)(A)\ \forall A \in \mathcal{Q}_n$, where $\mathcal{Q}_n$ is the field generated by $\{B_1,\dots,B_n\}$. Now, applying Proposition 3 we deduce that $\{Q_n\}_n$ converges weakly to $\alpha P_1 + (1-\alpha)P_2$ and this probability belongs to $\overline{P(\Gamma)}$. From this, we deduce in particular the equality $\overline{Conv(P(\Gamma))} = \overline{P(\Gamma)}$. The first set in this equality is the closure of a convex set of probabilities defined on a separable metric space. Following the course of reasoning from the proof of point 2 from Theorem 4, we can deduce that $\overline{Conv(P(\Gamma))}$ (and hence $\overline{P(\Gamma)}$) is convex. □

A similar proof would allow us to deduce that $\overline{\Delta(\Gamma)}$ is convex when $(\Omega,\mathcal{A},P)$ is non-atomic and (X,d) separable. Now, using Theorems 1, 4 and 5, we derive the following result:

Corollary 2 *Let $(\Omega,\mathcal{A},P)$ be a probability space, (X,d) be a separable metric space, and $\Gamma:\Omega\to\mathcal{P}(X)$ a random set. Under any of the following conditions:*

1. *Γ is open,*
2. *Γ is complete,*
3. *X is σ-compact and Γ is closed,*

$\overline{M(P^*)} = \overline{Conv(P(\Gamma))}$. If in addition $(\Omega,\mathcal{A},P)$ is non-atomic, then $\overline{M(P^*)} = \overline{P(\Gamma)}$.

Proof. The first part follows from Theorem 1 and the first point of Theorem 4. For the second part, it suffices to apply the second point of Theorem 4 and Theorem 5. □

This corollary extends results from [3, 16], and tells us that under fairly general conditions, the upper probability can be used to model the available information without producing a (big) loss of precision. It also extends some results from [10]: it is proven there that given two closed random sets Γ_1, Γ_2 on a separable Banach space, the equality between $P^*_{\Gamma_1}$ and $P^*_{\Gamma_2}$ implies that $\overline{Conv(P(\Gamma_1))}$ is equal to $\overline{Conv(P(\Gamma_2))}$. Similar results can be seen in [1, 9]. We have proven that it is indeed $P^*_{\Gamma_1} = P^*_{\Gamma_2} \Rightarrow \overline{Conv(P(\Gamma_1))} = \overline{Conv(P(\Gamma_2))} = \overline{M(P^*_{\Gamma_1})} = \overline{M(P^*_{\Gamma_2})}$, and only requiring Γ_i to be complete on a separable metric space $\forall i = 1,2$. On the other hand, we deduce that under the hypotheses of the second part of Corollary 2, if $P(\Gamma)$ is weakly closed, it is also convex, and $M(P^*)$ is closed. The converses are not true in general. The following example shows that $P(\Gamma)$ is not necessarily closed when $M(P^*)$ is closed:

Example 2 *[15] Consider* $\Gamma : [0,1] \to \mathcal{P}([0,1])$ *defined by* $\Gamma(\omega) = [-\omega, \omega]\ \forall \omega \in [0,1]$. *Then, it can be proven that* $M(P^*)$ *is closed (indeed, this holds for any compact random set on a Polish space). However, given the selectors* $A, B \in S(\Gamma)$ *defined by* $A(\omega) = -\omega, B(\omega) = \omega\ \forall \omega$*, it can be checked that* $\frac{P_A + P_B}{2} \notin P(\Gamma)$*. This shows that* $P(\Gamma)$ *is not convex. As a consequence, it is not closed either: if it were, it would be* $P(\Gamma) = \overline{P(\Gamma)} = \overline{M(P^*)} = M(P^*)$ *convex, a contradiction.*

On the other hand, Example 1 shows that $P(\Gamma)$ is not closed either if it is convex. Indeed, that implication does not hold even if $P(\Gamma) = M(P^*)$:

Example 3 *Consider* $(\Omega, \mathcal{A}, P) = ((0,1), \beta_{(0,1)}, \lambda_{(0,1)})$ *a non-atomic probability space, and let* $\Gamma : \Omega \to \mathcal{P}(\mathbb{R})$ *be constant on* $(0,1)$*. Then,* $M(P^*) = \{Q : \beta_{\mathbb{R}} \to [0,1] \text{ probability} \mid Q((0,1)) = 1\}$*. Given a probability measure* $Q \in M(P^*)$*, its quantile function* $U : (0,1) \to \mathbb{R}$ *is a selector of* Γ *and satisfies* $P_U = Q$*, whence* $P(\Gamma) = M(P^*)$ *convex. However, the sequence of degenerate probability measures on* $\frac{1}{n}$, $\{\delta_{\frac{1}{n}}\}_n \subseteq P(\Gamma)$*, converges weakly to* $\delta_0 \notin P(\Gamma)$*. Hence, this set is not closed.*

4 Conclusions and open problems

In this paper, we have compared the different models of the probabilistic information given by a random set, when this random set is the imprecise observation of a random variable. We have considered three different sets of probability measures, and through them we have investigated whether an imprecise probability model in terms of the upper and lower probabilities is useful in this context.

The results we have established allow us to conclude that under fairly general conditions, the upper and lower probabilities induced by a random set can be used to summarize the information on the distribution of the original random variable without a substantial loss of precision. Nevertheless, there are still a number of particular cases of random sets which are worth a detailed study. We would also like to study the topological structure of $P(\Gamma)$ and $M(P^*)$ under other than the topology of the weak convergence, and derive other sufficient conditions for the equalities $\Delta(\Gamma) = M(P^*)$ and $\overline{P(\Gamma)} = \overline{M(P^*)}$. Finally, it would also be interesting (though we are not very optimistic in this respect) to obtain sufficient conditions for the equality $P(\Gamma) = M(P^*)$ in terms of the images of the random set, as it was done in [15] for the particular case of random intervals.

References

[1] Z. Arstein and S. Hart. Law of large numbers for random sets and allocation processes. *Mathematics of Operations Research*, 6:485–492, 1981.

[2] P. Billingsley. *Convergence of probability measures*. Wiley, New York, 1968.

[3] A. Castaldo and M. Marinacci. Random correspondences as bundles of random variables. In *Proceedings of the 2nd ISIPTA Conference*, Ithaca (New York), 2001.

[4] A. Chateauneuf and J.-Y. Jaffray. Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. *Mathematical Social Sciences*, 17:263–283, 1989.

[5] I. Couso. *Teoría de la Probabilidad para datos imprecisos. Algunos aspectos*. Ph.D. Thesis, University of Oviedo, 1999. In Spanish.

[6] I. Couso, S. Montes and P. Gil. Second order possibility measure induced by a fuzzy random variable. In C. Bertoluzza, M. A. Gil and D. A. Ralescu, editors, *Statistical modeling, analysis and management of fuzzy data*. Springer, 2002.

[7] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[8] D. Denneberg. *Non-additive measure and integral*. Kluwer Academic Publishers, 1994.

[9] S. Hart and E. Kohlberg. Equally distributed correspondences. *Journal of Mathematical Economics*, 1:167–674, 1974.

[10] C. Hess. The distribution of unbounded random sets and the multivalued strong law of large numbers in nonreflexive banach spaces. *Journal of Convex Analysis*, 6:163–182, 1999.

[11] W. Hildenbrand. *Core and Equilibria of a Large Economy*. Princeton University Press, Princeton, 1974.

[12] C. J. Himmelberg. Measurable relations. *Fundamenta Mathematicae*, 87:53–72, 1975.

[13] R. Kruse and K. D. Meyer. *Statistics with vague data*. D. Reidel Publishing Company, Dordretch, 1987.

[14] G. Matheron. *Random sets and integral geometry*. Wiley, New York, 1975.

[15] E. Miranda. Estudio de la información probabilística de los intervalos aleatorios. In *Proceedings of the 27th SEIO conference*, Lérida (Spain), 2003. In Spanish.

[16] E. Miranda, I. Couso and P. Gil. Upper probabilities and selectors of random sets. In P. Grzegorzewski, O. Hryniewicz and M. A. Gil, editors, *Advances in soft computing*. Physica-Verlag, 2002.

[17] E. Miranda, I. Couso and P. Gil. Extreme points of credal sets generated by 2-alternating capacities. *International Journal of Approximate Reasoning*, 33:95-115, 2003.

[18] E. Miranda, I. Couso and P. Gil. Upper probabilities attainable by distributions of measurable selections, 2002, submitted.

[19] E. Miranda, G. de Cooman and I. Couso. Imprecise probabilities induced by multi-valued mappings. In *Proceedings of the 9th IPMU Conference*, Annecy (France), 2002.

[20] H. T. Nguyen. On random sets and belief functions. *Journal of Mathematical Analysis and Applications*, 63:531–542, 1978.

[21] P. Walley. *Statistical reasoning with imprecise probabilities*. Chapman and Hall, London, 1991.

Enrique Miranda is with the Department of Statistics and O. R. of the University of Oviedo, C-Calvo Sotelo s/n, 33007, Oviedo. E-mail:emiranda@correo.uniovi.es

Inés Couso is with the Department of Statistics and O. R. of the University of Oviedo, E. S. de Marina Civil, Ctra. de Villaviciosa s/n 33203, Gijón. E-mail:couso@pinon.ccu.uniovi.es

Pedro Gil is with the Department of Statistics and O. R. of the University of Oviedo, C-Calvo Sotelo s/n, 33007, Oviedo. E-mail:pedro@pinon.ccu.uniovi.es

An Extended Set-valued Kalman Filter

D. R. MORRELL
Arizona State University, USA

W. C. STIRLING
Brigham Young University, USA

Abstract

Set-valued estimation offers a way to account for imprecise knowledge of the prior distribution of a Bayesian statistical inference problem. The set-valued Kalman filter, which propagates a set of conditional means corresponding to a convex set of conditional probability distributions of the state of a linear dynamic system, is a general solution for linear Gaussian dynamic systems. In this paper, the set-valued Kalman filter is extended to the non-linear case by approximating the non-linear model with a linear model that is chosen to minimize the error between the non-linear dynamics and observation models and the linear approximation. An application is presented to illustrate and interpret the estimator results.

Keywords

imprecise probabilities, statistical inference, dynamic systems, convex sets of probability measures, set-valued estimation

1 Introduction

In this paper we address the statistical inference problem of estimating a set of time-varying parameters of a discrete-time dynamical system that is monitored with discrete-time observations of its behavior. Such a real-time estimator is called a *filter*. For example, consider an aircraft flight for which radar data are collected as functions of its kinematic parameters (position and velocity). The filtering problem is to obtain instantaneous estimates of its trajectory.

A reasonable model structure for this class of problems is for the system dynamics to be modeled as a finite-dimensional Markov process that is characterized by a stochastic difference equation of the form

$$\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k) + \mathbf{w}_k \tag{1}$$

for $k = 0, 1, \ldots$, where the p-dimensional vector $\mathbf{x}_k$ is the *state* of the system at time k, and is the time-varying parameter set to be estimated. The p-dimensional vector function $\mathbf{f}$ is the dynamical model for the system, and the p-dimensional vector $\mathbf{w}_k$ is an uncorrelated process termed the *process noise*, with covariance matrix $\mathbf{Q}_k$. The process noise represents random disturbances to the system.

The observation model for this system is of the form

$$\mathbf{y}_k = \mathbf{h}(\mathbf{x}_k) + \mathbf{v}_k \tag{2}$$

for $k = 1, 2, \ldots$, where the q-dimensional vector function $\mathbf{h}$ models the observations as a function of the state. The q-dimensional vector $\mathbf{v}_k$ is an uncorrelated process, termed the *observation noise*, with covariance matrix $\mathbf{R}$. The observation noise represents random measurement errors.

The general filtering problem for this class of systems is to determine the conditional distribution of $\{\mathbf{x}_k, k > 0\}$, given $\{\mathbf{y}_j,\ j \leq k\}$. This problem is easily solved formally: densities are propagated forward via the Chapman-Kolmogorov equation and observations are incorporated using Bayes theorem. However, there are very few system models that lead to closed form solutions. An important special case for which the solution is well known is the linear Gaussian system with precise probability distributions. According to this model, the dynamical and observational equations are linear functions of the state, *i.e.* ,

$$\mathbf{x}_{k+1} = \mathbf{F}_k \mathbf{x}_k + \mathbf{w}_k \tag{3}$$

and

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k, \tag{4}$$

where the processes $\mathbf{w}_k$ and $\mathbf{v}_k$ and the initial state $\mathbf{x}_0$ are all assumed to be jointly normally distributed and mutually uncorrelated. For this special case, the subsequent states $\mathbf{x}_k$, being linear combinations of normally distributed random variables, are also normally distributed, and the problem is solved by directly computing the conditional expectation and covariance of the state. The stationary linear filtering problem (that is, when $\mathbf{F}_k$ and $\mathbf{H}_k$ are constant matrices) was solved by Wiener [14, 4], and the nonstationary case was solved by Kalman [5], Kalman and Bucy [7], and Kalman [6], resulting in the well-known Kalman filter.

Since the normal distribution is not preserved under non-linear transformations, it is not straightforward to compute the conditional mean and variance for non-linear systems. The set-valued estimation problem was addressed for the non-linear case by Kenney and Stirling [8], who provide an approximate solution for the propagation for a set of conditional densities of the state based upon Galerkin approximations to Kolmogorov's equations. Unfortunately, however, this latter approach, although theoretically elegant, is very computationally intensive and has not yet proven to be a practical solution. Practical non-linear estimation techniques include linearization approaches such as the extended Kalman filter [3], Monte Carlo particle filters [2], and interacting multiple-model filtering [1]. Although our approach is essentially Kalman filter based, alternative approaches to set-valued filtering are possible topics of future research. Our preliminary assessment, however, is that extending a particle filter to the imprecise case would be computationally very demanding.

Kalman filter-based approaches (both linear and extended) typically employ a precise prior distribution for the initial state. This is a strict Bayesian approach that is often assumed out of convenience. If this assumption is unwarranted, the precision attributed to the resulting state estimates will not be a realistic indication of their accuracy. Of course, if the system is observable, the influence of the initial conditions will become asymptotically negligible as more and more data are processed. But, for systems with limited data, or if accuracy assessments after a few observations are of interest, then the effect of the initial conditions will be critical.

Imprecise probability theory [13] has emerged as a way to account for ignorance as well as uncertainty in decision making. For the problem here considered, we are concerned with situations where we are unable to specify with confidence the prior distribution of $\mathbf{x}_0$. One way to approach this problem is to characterize the prior as a convex set of distributions, rather than a singleton. This convex Bayesian approach is advocated by Levi [9, 10] as a way of suspending judgment between choices when there is insufficient information to choose a single distribution. Thus, if $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ are possible prior distributions for $\mathbf{x}_0$, then so is every convex combination $\alpha p_1(\mathbf{x}) + (1-\alpha) p_2(\mathbf{x})$, where $\alpha \in [0,1]$. The filtering problem is then to propagate and update this convex set of distributions. This problem was solved for the linear case by Morrell and Stirling [12], resulting in the *set-valued Kalman filter*.

This paper presents an alternative approach to set-valued non-linear filtering. In Section 2 we review linear set-valued Kalman filtering, which we then extend to deal with non-linear systems in Section 3. Finally, we provide an example in Section 4, and we finish with a discussion in Section 5.

2 Linear Set-Valued Filtering

Consider the system dynamics and observation equations presented in (3) and (4). The set-valued Kalman filter computes a sequence of estimate sets and a corresponding sequence of estimate covariances [12]. An estimate set is denoted $X_{k|j}$, the set of estimates of the system state at time k given the observations $\mathbf{y}_1$ through $\mathbf{y}_j$, and is represented in terms of the p-dimensional vector $\mathbf{c}_{k|j}$ and the $p \times p$ matrix $\mathbf{K}_{k|j}$ as

$$X_{k|j} = \left\{ \mathbf{x}\colon \left(\mathbf{x} - \mathbf{c}_{k|j}\right)^T \mathbf{S}_{k|j}^{-1} \left(\mathbf{x} - \mathbf{c}_{k|j}\right) \leq 1 \right\}, \tag{5}$$

where $\mathbf{S}_{k|j} = \mathbf{K}_{k|j} \mathbf{K}_{k|j}{}^T$. The set-valued Kalman filtering equations provide a two-stage recursion for computing $\mathbf{c}_{k|j}$, $\mathbf{K}_{k|j}$, and the estimation error covariance $\mathbf{P}_{k|j}$

for $j = k-1$ (prediction between observations) and $j = k$ (updating new observations, or filtering).

Initialization: We assume that the initial state of the dynamic system is characterized by a distribution that lies in the set

$$\mathbf{X} = \left\{\mathbf{x} \sim \mathcal{N}\big(\mathbf{m}, \mathbf{P}_{0|0}\big) \colon \mathbf{m} \in X_{0|0}\right\},$$

where $\mathcal{N}\big(\mathbf{m}, \mathbf{P}_{0|0}\big)$ denotes the normal distribution with mean $\mathbf{m}$ and positive-definite covariance matrix $\mathbf{P}_{0|0}$, and $X_{0|0}$ denotes a hyper-ellipsoid defined by

$$X_{0|0} = \left\{\mathbf{x} \colon \left(\mathbf{x} - \mathbf{c}_{0|0}\right)^T \mathbf{S}_{0|0}^{-1} \left(\mathbf{x} - \mathbf{c}_{0|0}\right) \leq 1\right\}, \tag{6}$$

where $\mathbf{S}_{0|0} = \mathbf{K}_{0|0}\mathbf{K}_{0|0}{}^T$.

Prediction Step:

$$\mathbf{c}_{k|k-1} = \mathbf{F}_{k-1}\mathbf{c}_{k-1|k-1} \tag{7}$$

$$\mathbf{P}_{k|k-1} = \mathbf{F}_{k-1}\mathbf{P}_{k-1|k-1}\mathbf{F}_{k-1}{}^T + \mathbf{Q}_{k-1} \tag{8}$$

$$\mathbf{K}_{k|k-1} = \mathbf{F}_{k-1}\mathbf{K}_{k-1|k-1} \tag{9}$$

The predicted set-valued state estimate is given by

$$X_{k|k-1} = \left\{\mathbf{x} \colon \left(\mathbf{x} - \mathbf{c}_{k|k-1}\right)^T \mathbf{S}_{k|k-1}^{-1} \left(\mathbf{x} - \mathbf{c}_{k|k-1}\right) \leq 1\right\}, \tag{10}$$

with $\mathbf{S}_{k|k-1} = \mathbf{K}_{k|k-1}\mathbf{K}_{k|k-1}{}^T$.

Filter Step:

$$\mathbf{c}_{k|k} = \mathbf{c}_{k|k-1} + \mathbf{W}_k \left[\mathbf{y}_k - \mathbf{H}_k\mathbf{c}_{k|k-1}\right] \tag{11}$$

$$\mathbf{P}_{k|k} = \left[\mathbf{I} - \mathbf{W}_k\mathbf{H}_k\right] \mathbf{P}_{k|k-1} \tag{12}$$

$$\mathbf{K}_{k|k} = \left[\mathbf{I} - \mathbf{W}_k\mathbf{H}_k\right] \mathbf{K}_{k|k-1}, \tag{13}$$

where $\mathbf{W}_k$ is the Kalman gain:

$$\mathbf{W}_k = \mathbf{P}_{k|k-1}\mathbf{H}_k{}^T \left[\mathbf{H}_k\mathbf{P}_{k|k-1}\mathbf{H}_k{}^T + \mathbf{R}\right]^{-1}. \tag{14}$$

The filtered set-valued estimate is then given by

$$X_{k|k} = \left\{\mathbf{x} \colon \left(\mathbf{x} - \mathbf{c}_{k|k}\right)^T \mathbf{S}_{k|k}^{-1} \left(\mathbf{x} - \mathbf{c}_{k|k}\right) \leq 1\right\} \tag{15}$$

It is shown in [12] that, if the linear system defined by (3) and (4) is uniformly observable and controllable (*e.g.* , see [3]), then $\mathbf{K}_{k|k} \to \mathbf{0}$ as $k \to \infty$. Thus, in the limit, the set-valued estimates converge to a point, and the imprecise probability distributions converge to a precise distribution. For systems that are not uniformly observable and controllable, or if the time sequence is not infinite, then imprecision cannot be eliminated. Observability and controllability guarantee only that $\mathbf{P}_{k|k}$ will be bounded [3]. This does *not* mean, however, that the estimation error covariance $\mathbf{P}_{k|k}$ tends to zero as $k \to \infty$.

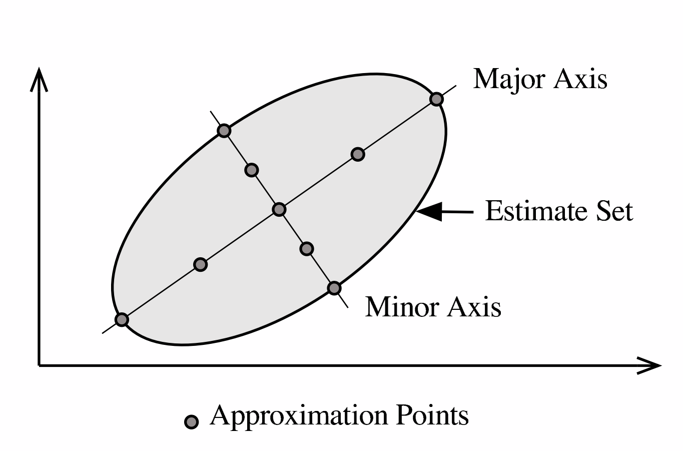


Figure 1: Approximation points for the extended set-valued Kalman filter.

3 Extension to Non-linear System Models

We desire to apply the set-valued Kalman filter to non-linear systems using a linear approximation to the system model. In an extended Kalman filter, such an approximation is made by computing a first-order Taylor series expansion of the non-linear functions about a point-valued state estimate. Unfortunately, because we have a set of state estimates, the set-valued Kalman filter cannot be extended in the same way, and we instead choose approximations that best fit the non-linear functions over the estimate set.

We propose the following approach to finding approximations of the system dynamic and observation functions over the entire estimate set. A set of *approximation points* is chosen; the parameters of affine approximations to the dynamics and observation functions are computed to minimize the weighted squared errors between the function values and approximation values evaluated at the approximation points. Our method of choosing approximation points relies on the hyper-ellipsoidal shape of the estimation sets. Figure 1 illustrates our method for a two-dimensional estimate set (i.e., $p = 2$). Specifically, we form the set of approximation points from the centroid of the estimate set, each point where an axis of the hyper-ellipse intersects the ellipse, and all points equidistant from the centroid and boundary points. Since the estimate set is a p-dimensional hyper-ellipse, the set of approximation points will require $4p+1$ elements. The set-valued Kalman filter requires (approximate) linear dynamics and observation models in which the approximations are good over the entire estimate set.

Approximating the Dynamics and Observation Functions. We choose approximations of the following form:

$$\mathbf{f}(\mathbf{x}_k) \approx \mathbf{F}_k \mathbf{x}_k + \mathbf{f}_k^0 \tag{16}$$

and

$$\mathbf{h}(\mathbf{x}_k) \approx \mathbf{H}_k \mathbf{x}_k + \mathbf{h}_k^0. \tag{17}$$

The linearizations are obtained by solving the following problems for $\mathbf{F}_k$ and $\mathbf{f}_k^0$ and for $\mathbf{H}_k$ and $\mathbf{h}_k^0$. Let $\mathbf{x}_{k|k}^{(0)}$ through $\mathbf{x}_{k|k}^{(N-1)}$ be values in $X_{k|k}$, denoted the *filtered approximation points,* and let $\mathbf{x}_{k|k-1}^{(0)}$ through $\mathbf{x}_{k|k-1}^{(N-1)}$ be values in $X_{k|k-1}$, denoted the *predicted approximation points*. Let $\mathbf{d}_k^{(n)}$ be the error between the actual dynamics function and the linear approximation evaluated at $\mathbf{x}_{k|k}^{(n)}$:

$$\mathbf{d}_k^{(n)} = \mathbf{f}\left(\mathbf{x}_{k|k}^{(n)}\right) - \mathbf{F}_k \mathbf{x}_{k|k}^{(n)} - \mathbf{f}_k^0.$$

Also, let $\mathbf{e}_k^{(n)}$ be the error between the actual observation function and the linear approximation evaluated at $\mathbf{x}_{k|k-1}^{(n)}$:

$$\mathbf{e}_k^{(n)} = \mathbf{h}\left(\mathbf{x}_{k|k-1}^{(n)}\right) - \mathbf{H}_k \mathbf{x}_{k|k-1}^{(n)} - \mathbf{h}_k^0.$$

We choose $\mathbf{F}_k$, $\mathbf{f}_k^0$ and $\mathbf{H}_k$, $\mathbf{h}_k^0$ to minimize the sums, respectively, of weighted squared dynamics and observation errors evaluated at the approximation points:

$$\mathbf{F}_k, \mathbf{f}_k^0 = \arg\min_{\mathbf{F},\mathbf{f}^0} \sum_{n=0}^{N-1} \ell_k^{(n)} \left[\mathbf{d}_k^{(n)}\right]^T \left[\mathbf{d}_k^{(n)}\right]$$

and

$$\mathbf{H}_k, \mathbf{h}_k^0 = \arg\min_{\mathbf{H},\mathbf{h}^0} \sum_{n=0}^{N-1} \ell_k^{(n)} \left[\mathbf{e}_k^{(n)}\right]^T \left[\mathbf{e}_k^{(n)}\right],$$

where $\ell_k^{(n)}$ is a weight associated with the nth approximation point.

This is a simple weighted least squares problem [11]. We define the following matrices:

$$\mathbf{L}_k = \operatorname{diag}\left(\ell_k^{(0)}, \ldots, \ell_k^{(N-1)}\right)$$

$$\mathbf{A}_k = \begin{bmatrix} \mathbf{x}_{k|k}^{(0)} & \cdots & \mathbf{x}_{k|k}^{(N-1)} \\ 1 & \cdots & 1 \end{bmatrix}, \qquad \mathbf{B}_k = \begin{bmatrix} \mathbf{x}_{k|k-1}^{(0)} & \cdots & \mathbf{x}_{k|k-1}^{(N-1)} \\ 1 & \cdots & 1 \end{bmatrix}$$

$$\mathbf{C}_k = \begin{bmatrix} \mathbf{f}^T\left(\mathbf{x}_{k|k}^{(0)}\right) \\ \vdots \\ \mathbf{f}^T\left(\mathbf{x}_{k|k}^{(N-1)}\right) \end{bmatrix}, \qquad \mathbf{D}_k = \begin{bmatrix} \mathbf{h}^T\left(\mathbf{x}_{k|k-1}^{(0)}\right) \\ \vdots \\ \mathbf{h}^T\left(\mathbf{x}_{k|k-1}^{(N-1)}\right) \end{bmatrix}.$$

The solution to the weighted least squares problem is

$$\begin{bmatrix} \mathbf{F}_k{}^T \\ \mathbf{f}_k^{0T} \end{bmatrix} = \left(\mathbf{A}_k \mathbf{L}_k \mathbf{A}_k{}^T\right)^{-1} \mathbf{A}_k \mathbf{L}_k \mathbf{C}_k$$

and

$$\begin{bmatrix} \mathbf{H}_k{}^T \\ \mathbf{h}_k^{0T} \end{bmatrix} = \left(\mathbf{B}_k \mathbf{L}_k \mathbf{B}_k{}^T\right)^{-1} \mathbf{B}_k \mathbf{L}_k \mathbf{D}_k.$$

An example of choosing approximation points is given in Section 4 in the context of a target tracking problem. Once these quantities are defined, the set-valued Kalman filter is applied to the equations

$$\mathbf{x}_{k+1} = \mathbf{F}_k \mathbf{x}_k + \mathbf{f}_k^0 + \mathbf{w}_k \tag{18}$$

and

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{h}_k^0 + \mathbf{v}_k. \tag{19}$$

Initialization: The extended set-valued Kalman filter is initialized in the same way the set-valued Kalman filter is initialized.

Prediction Step:

$$\mathbf{c}_{k|k-1} = \mathbf{f}(\mathbf{c}_{k-1|k-1}) \tag{20}$$

$$\mathbf{P}_{k|k-1} = \mathbf{F}_{k-1} \mathbf{P}_{k-1|k-1} \mathbf{F}_{k-1}{}^T + \mathbf{Q}_{k-1} \tag{21}$$

$$\mathbf{K}_{k|k-1} = \mathbf{F}_{k-1} \mathbf{K}_{k-1|k-1} \tag{22}$$

Filter Step:

$$\mathbf{c}_{k|k} = \mathbf{c}_{k|k-1} + \mathbf{W}_k \left[\mathbf{y}_k - \mathbf{h}(\mathbf{c}_{k|k-1})\right] \tag{23}$$

$$\mathbf{P}_{k|k} = [\mathbf{I} - \mathbf{W}_k \mathbf{H}_k] \mathbf{P}_{k|k-1} \tag{24}$$

$$\mathbf{K}_{k|k} = [\mathbf{I} - \mathbf{W}_k \mathbf{H}_k] \mathbf{K}_{k|k-1}, \tag{25}$$

where $\mathbf{W}_k$ is the Kalman gain as given by (14). The filtered set estimate is then given by (15).

4 Example: Target Tracking using Range Measurements

In this section, we present an example of this linearization technique for the set-valued filter. We track a moving target using measured range from one or two fixed sensors; one or both sensors may operate at any point in time. The target moves in a two-dimensional Cartesian coordinate system. Figure 2 illustrates the target motion, sensor locations, and range measurements. The set-valued filter estimates

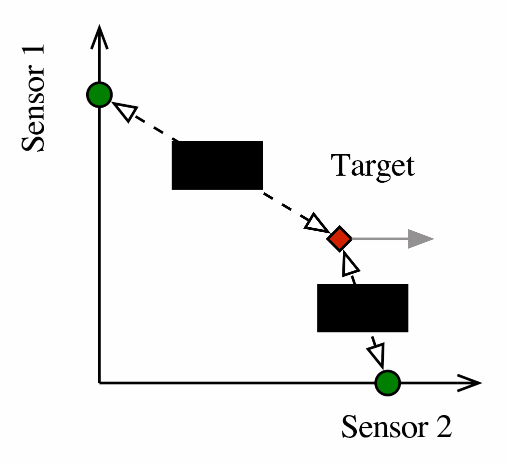


Figure 2: The tracking scenario for application of the set-valued Kalman filter. The target moves in a straight line from left to right. Sensors 1 and 2 measure their range to the target at each time.

the target position and velocity in both dimensions as a function of time from the range measurements.

We use a linear model of the form (3) for the target dynamics. The target state $\mathbf{x}_k$ consists of four elements: the target position in the x and y directions, denoted $x_k(0)$ and $x_k(1)$, and the target velocity in the x and y directions, denoted $x_k(2)$ and $x_k(3)$. The system dynamics matrix is the following:

$$\mathbf{F}=\begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where Δt is the time between observations. The process noise covariance matrix is

$$\mathbf{Q}_k = \sigma^2 \begin{bmatrix} \frac{\Delta t^3}{3} & 0 & \frac{\Delta t^2}{2} & 0 \\ 0 & \frac{\Delta t^3}{3} & 0 & \frac{\Delta t^2}{2} \\ \frac{\Delta t^2}{2} & 0 & \Delta t & 0 \\ 0 & \frac{\Delta t^2}{2} & 0 & \Delta t \end{bmatrix},$$

where σ^2 is the intensity of a white continuous-time Gaussian noise process modeling the target acceleration.

For this example, we locate Sensor 1 at coordinates $(0,20)$ and Sensor 2 at coordinates $(20,0)$. The ranges from the sensors to the target at time k are denoted as $r_k(1)$ and $r_k(2)$. These ranges are computed as

$$r_k(1) = \sqrt{(x_k(0) - 0)^2 + (x_k(1) - 20)^2}$$

$$r_k(2) = \sqrt{(x_k(0) - 20)^2 + (x_k(1) - 0)^2}.$$

Since one or both sensors may be in use at any given k, the observation function $\mathbf{h}(\mathbf{x}_k)$ will be either a one- or two- dimensional vector function of the state $\mathbf{x}_k$. When only Sensor 1 is in use, $\mathbf{h}(\mathbf{x}_k) = r_k(1)$.When both sensors are in use, $\mathbf{h}(\mathbf{x}_k) = \begin{bmatrix} r_k(1) \\ r_k(2) \end{bmatrix}$.

In this problem, the system dynamics are linear; thus, we need to approximate only the observation function. We select the predicted approximation points $\mathbf{x}_{k|k-1}^{(n)}$ as follows. The observations depend only on the target position and not on its velocity, so we select approximation points to cover the range of position values in the estimate set $X_{k|k-1}$. These position values lie in an ellipse defined by the upper left sub-matrix of $\mathbf{S}_{k|k-1}$. We use the centroid of the ellipse, the four points at the intersection of the boundary of the ellipse with its axes, and the four points equidistant between the centroid and boundary points. We use weights of 1.0 for the centroid point, 0.5 for the midpoints, and 0.1 for the boundary points.

In the example scenario, the target starts at the point (10,10) with a velocity of one unit/second to the right. The time Δt between observations is 2 seconds. Only Sensor 1 provides range measurements from time $k = 1$ to $k = 4$; after $k = 4$, both sensors provide range measurements. Figure 3 shows the set estimates of the target position for this scenario. The initial estimate set is circular. The range observations from Sensor 1 quickly reduce the size of the estimate set in the direction of the target from the sensor, but do not provide information about the target location along the perpendicular direction. When range information from Sensor 2 becomes available at time $k = 5$, the set of estimates becomes much smaller, since now there is enough information in the observations to locate the target. In other words, during the first 4 time units, when the system is not fully observable, the set of estimates does not shrink in the unobservable direction, but thereafter, the system is fully observable and the set of means shrinks in both directions.

It must be emphasized that the ellipses in Figure 3 do *not* correspond to likelihood contours (contours of constant value of probability density); rather, they define a set of position estimates, each of which has a legitimate claim to being a valid assessment of the true state of the system. If time were to increase without bound with both sets of observations available, the system would be observable and, in the limit, it would converge to a singleton representing the mean value of a unique limiting distribution (a precise probability). The covariance of this distribution, however, would converge to a steady-state, but non-zero, level, such

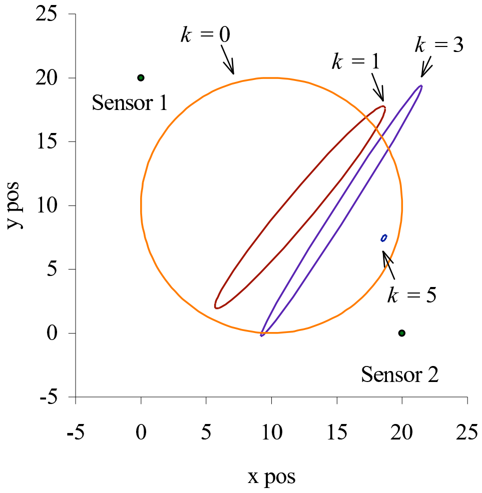


Figure 3: Estimate sets.

that no increase in the accuracy of the (now point-valued) state estimates can be achieved.

5 Discussion

For time-varying estimation scenarios that are either not uniformly observable and controllable or, even if they are observable and controllable, are of such short duration that transients in the estimator dynamics do not have time to damp out, set-valued estimation provides a realistic means of accounting for imprecise knowledge of the mean of the prior distribution.

Non-linear filtering requires the propagation of the entire distribution, in contrast to the need to propagate only the first two moments with linear filtering. This accounts for the difficulty associated with non-linear estimation. The conventional extended Kalman filter is a well-accepted and practical solution for point-valued estimates, but it does not apply to the set-valued case. The extended set-valued Kalman filter provides an approximate solution to the non-linear set-valued dynamic state estimation problem that is computationally feasible. As with the conventional extended Kalman filter, however, it is not possible to prove global convergence of the extended set-valued Kalman filter.

References

[1] Blom, H. A. P., and Bar-Shalom, Y. The interacting multiple model algorithm for systems with markovian switching coefficients. *IEEE Trans. on Automatic Control AC-33*, 8 (1988), 780–783.

[2] Doucet, A. N., de Freitas, N., and Gordon, N. J., Eds. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York, 2001.

[3] Jazwinski, A. H. *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970.

[4] Kailath, T. *Lectures on Wiener and Kalman Filtering*. Springer-Verlag, New York, 1981.

[5] Kalman, R. E. A new approach to linear filtering and prediction problems. *Trans. ASME, Ser. D: J. Basic Eng 82* (1960), 35–45.

[6] Kalman, R. E. New methods in Wiener filtering theory. In *Proc. Symp. Appl. Random Function Theory and Probability* (New York, 1963), J. L. Bogdanoff and F. Kozin, Eds., Wiley.

[7] Kalman, R. E., and Bucy, R. S. New results in linear filtering and prediction theory. *Trans. ASME, Ser. D: J. Basic Eng 83* (1961), 95–108.

[8] Kenney, J. D., and Stirling, W. C. Nonlinear filtering of convex sets of probability distributions. *J. Stat. Plann. Inference 105* (2002), 123–137.

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

[10] Levi, I. Imprecision and indeterminacy in probability judgement. *Philosophy of Science 52*, 3 (1985), 390–409.

[11] Moon, T. K., and Stirling, W. C. *Mathematical Methods and Algorithms in Signal Processing*. Prentice-Hall, Upper Saddle River, NJ, 2000.

[12] Morrell, D. R., and Stirling, W. C. Set-valued filtering and smoothing. *IEEE Trans. Systems, Man, Cybernet. 21*, 1 (January/February 1991), 184–193.

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

[14] Wiener, N. *The extrapolation, Interpolation and Smoothing of Stationary Time Series*. Wiley, New York, 1949.

Darryl R. Morrell is with the Department of Electrical Engineering, Arizona State University, Tempe, Arizona, USA. E-mail Morrell@asu.edu

Wynn C. Stirling is with the Department of Electrical and Computer Engineering, Brigham Young University, Provo, Utah, USA. E-mail wynn@ee.byu.edu

The Shape of Incomplete Preferences[*]

ROBERT NAU
Duke University, USA

Abstract

The emergence of robustness as an important consideration in Bayesian statistical models has led to a renewed interest in normative models of incomplete preferences represented by imprecise (set-valued) probabilities and utilities. This paper presents a simple axiomatization of incomplete preferences and characterizes the shape of their representing sets of probabilities and utilities. Deletion of the completeness assumption from the axiom system of Anscombe and Aumann yields preferences represented by a convex set of state-dependent expected utilities, of which at least one must be a probability/utility pair. A strengthening of the state-independence axiom is needed to obtain a representation purely in terms of a set of probability/utility pairs.

Keywords

axioms of decision theory, Bayesian robustness, state-dependent utility, coherence, partial order, imprecise probabilities and utilities

1 Introduction

In the Bayesian theory of choice under uncertainty, a decision maker holds rational preferences among acts, which are mappings from states of nature $\{s\}$ to consequences $\{c\}$. It is typically assumed that rational preferences are *complete*, meaning that for any two acts **X** and **Y**, either $\mathbf{X} \succsim \mathbf{Y}$ ("**X** is weakly preferred to **Y**) or else $\mathbf{Y} \succsim \mathbf{X}$, or both. This assumption, together with other rationality axioms such as transitivity and independence, leads to a representation of preferences by a unique subjective probability distribution on states $p(s)$ and a unique utility function $u(c)$ on consequences, such that $\mathbf{X} \succsim \mathbf{Y}$ if and only if the subjective expected utility of **X** is greater than or equal to that of **Y** (Savage 1954, Anscombe and Aumann 1963, Fishburn 1982). However, the completeness assumption may be inappropriate if we have only partial information about the decision maker's preferences, or if realistic limits on her powers of discrimination are assumed, or if there are actually many decision makers whose preferences may disagree.

[*] This research was supported by the Fuqua School of Business and by the National Science Foundation. The author is grateful to David Rios Insua, Teddy Seidenfeld, Jim Smith, and Peter Wakker for comments on earlier drafts.

Incomplete preferences are generally represented by indeterminate (i.e., set-valued) probabilities and/or utilities. Varying degrees of such indeterminacy have been modeled previously in the literature of statistical decision theory and rational choice:

i. If probabilities alone are considered to be indeterminate, then preferences can be represented by a set of probability distributions $\{p(s)\}$ and a unique (perhaps linear) utility function $u(c)$. The set of probability distributions is typically convex, so the representation can be derived by separating hyperplane arguments (e.g., Smith (1961), Suppes (1974), Williams (1976), Giron and Rios (1980), Nau (1992).) Representations of this kind are are widely used in robust Bayesian statistics; an extensive treatment is given by Walley (1991).

ii. If utilities alone are considered to be indeterminate, preferences can be represented by a set of utility functions $\{u(c)\}$ and a unique (perhaps objectively specified) probability distribution $p(s)$, a representation that has been axiomatized and applied to economic models by Aumann (1962). The set of utility functions in this case is also typically convex, so that separating hyperplane arguments are again applicable.

iii. If both probabilities and utilities are allowed to be indeterminate, they can be represented by separate sets of probability distributions $\{p(s)\}$ and utility functions $\{u(c)\}$ whose elements are paired up arbitrarily. This representation of preferences preserves the traditional separation of information about *beliefs* from information about *values* when both are indeterminate (Rios Insua 1990, 1992), but lacks a natural axiomatic basis. Rather, it arises only as a special case of more general representations when probability and utility assessments are carried out independently.

iv. More generally, we can represent incomplete preferences by sets of probability distributions paired with state-independent utility functions $\{(p(s), u(c))\}$, a.k.a. “probability/utility pairs.” This representation has an appealing multi-Bayesian interpretation and provides a normative basis for techniques of robust decision analysis (Moskowitz, Preckel and Yang, 1993) and asset pricing in incomplete financial markets (Staum 2002). It has been axiomatized by Seidenfeld, Schervish, and Kadane (1995, henceforth SSK), starting from the “horse lottery” formalization of decision theory introduced by Anscombe and Aumann (1963). However, as pointed out by SSK, the set of probability/utility pairs is typically nonconvex and may even be unconnected, so that separating hyperplane arguments are not directly applicable. Instead, SSK rely on methods of transfinite induction and indirect reasoning to obtain their results.

The objective of this paper is to derive a simple representation of incomplete preferences for the elementary case of finite state and reward spaces, and to characterize the shape of the resulting sets of probabilities and utilities. We begin by deleting both completeness and state-independence from the horse-lottery axiom system of Anscombe and Aumann, showing that this leads immediately to a representation of preferences by a set of probabilities paired with state-*dependent* utility functions $\{(p(s), u(s,c))\}$. Such pairs will be called *state-dependent expected utility* (s.d.e.u.) functions. State-dependent utilities have been used in economic models by Karni (1985) and Drèze (1987) and are also discussed by Schervish et al. (1990). A set of s.d.e.u. functions is typically convex—unlike a set of probability/utility pairs—so that separating-hyperplane methods are still applicable at this stage. We then re-introduce Anscombe and Aumann's state-independence assumption and show that it imposes (only) the further requirement that the representing set should contain *at least one* probability/utility pair. Finally, we consider the additional assumptions that must be imposed in order to shrink the representation to (the convex hull of) a set of probability/utility pairs, and present a constructive alternative to SSK's indirect reasoning method. We show that although the representing set of probability/utility pairs is nonconvex, it nonetheless has a simple configuration: it is merely the intersection of a convex set of s.d.e.u. functions with the nonconvex surface of state-independent utilities.

The organization of the paper is as follows. Section 2 introduces basic notation and derives a representation of preferences by convex sets of s.d.e.u. functions when neither completeness nor state-independence is assumed. Section 3 incorporates Anscombe and Aumann's state-independence assumption and shows that it requires (only) the existence of at least one agreeing state-independent utility. Section 4 discusses an example of SSK to highlight the implications of different continuity and strictness conditions. Section 5 gives the additional constructive axiom that is needed to obtain a representation purely in terms of probability/utility pairs, illustrated by another example. Section 6 briefly discusses the results.

2 Representation of incomplete preferences

Let S denote a finite set of states and let C denote a finite set of consequences. Let $\mathcal{B} = \{\mathbf{B} : S \times C \mapsto \Re\}$. An element $\mathbf{X} \in \mathcal{B}$ is a *horse lottery* if $\mathbf{X} \geq \mathbf{0}$ and $\forall s,\ \sum_c X(s,c) = 1$, with the interpretation that $X(s,c)$ is the objective probability of receiving consequence c when state s occurs. Henceforth, the symbols $\mathbf{W}$, $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, and $\mathbf{H}$ will be used to denote horse lotteries; the symbol $\mathbf{B}$ will denote an element of $\mathcal{B}$ that is not necessarily a horse lottery (e.g., $\mathbf{B}$ may represent the difference between two horse lotteries). A horse lottery $\mathbf{X}$ is *constant* if the probabilities it assigns to consequences are constant across states—i.e., if $X(s,c) = X(s',c)$ for all s, s', c. The symbol $\succsim$ will denote non-strict preference between horse lotteries: $\mathbf{X} \succsim \mathbf{Y}$ means that $\mathbf{X}$ is preferred or indifferent to $\mathbf{Y}$, which is considered as the behavioral primitive. The domain of $\succsim$ is the set of all horse lotteries. The asymmetric part of $\succsim$ will be denoted by $\succ$.

An *event* is a subset of S. The symbol $\mathbf{E}$ will be used interchangeably as the name for an event and for its indicator function on $S \times C$. That is, $E(s,c) = 1[0]$ for all c if the event $\mathbf{E}$ includes [does not include] state s. $\mathbf{E}_s$ will denote the indicator vector for state s. That is, $\mathbf{E}_s(s',c) = 1$ for all c if $s = s'$ and zero otherwise. If α is a scalar between 0 and 1, then $\alpha\mathbf{X} + (1-\alpha)\mathbf{Y}$ is an *objective mixture* of $\mathbf{X}$ and $\mathbf{Y}$: it yields consequence c in state s with probability $\alpha X(s,c) + (1-\alpha)Y(s,c)$. If $\mathbf{E}$ is an event, then $\mathbf{EX} + (\mathbf{1}-\mathbf{E})\mathbf{Y}$ is a *subjective mixture* of $\mathbf{X}$ and $\mathbf{Y}$: it yields consequence c in state s with probability $X(s,c)$ if $E(s,c) = 1$, and with probability $Y(s,c)$ otherwise.

Assume that C contains a "worst" and a "best" consequence, labeled 0 and 1 respectively.[1] Other consequences are labeled $2, 3, \ldots, K$. The symbols $\mathbf{H}_c$, for $c \in \{0, 1, 2, \ldots, K\}$, and $\mathbf{H}_u$, for $u \in (0,1)$, will be used to denote special "reference" horse lotteries. First, for all $c \in \{0, 1, 2, \ldots, K\}$, let $\mathbf{H}_c$ denote the horse lottery that yields consequence c with probability 1 in every state. That is, $\mathbf{H}_c(s,c') = 1$ if $c = c'$ and $\mathbf{H}_c(s,c') = 0$ otherwise. For example, $\mathbf{H}_2$ is the horse lottery that yields consequence 2 with probability 1 in every state. Next, for all $u \in (0,1)$, let $\mathbf{H}_u$ denote the horse lottery that yields the best and worst consequences with probabilities u and $1-u$ in every state, which is the objective mixture:

$$\mathbf{H}_u \equiv u\mathbf{H}_1 + (1-u)\mathbf{H}_0.$$

For example, $\mathbf{H}_{0.5}$ is the horse lottery that yields consequences 0 and 1 with equal probability. Later on, consequences 0 and 1 will be assigned utilities of 0 and 1, respectively, so that $\mathbf{H}_u$ will have an expected utility of u by definition.

The reference-lottery notation can be stretched further to define $\mathbf{H}_\mathbf{E}$ as the horse lottery that yields the best consequence if event $\mathbf{E}$ occurs and the worst consequence otherwise, i.e., the subjective mixture:

$$\mathbf{H}_\mathbf{E} \equiv \mathbf{E}\mathbf{H}_1 + (\mathbf{1}-\mathbf{E})\mathbf{H}_0.$$

Bounds on subjective probabilities are expressible as preferences between subjective and objective mixtures of $\mathbf{H}_0$ and $\mathbf{H}_1$. For example, a preference of the form $\mathbf{H}_\mathbf{E} \succsim \mathbf{H}_p$ for some event $\mathbf{E}$ and $p \in (0,1)$ means that "the probability of $\mathbf{E}$ is at least p," i.e., that p is a *lower probability* for $\mathbf{E}$. Upper probabilities are defined analogously. If $\mathbf{X}$ is a horse lottery and u is a scalar between 0 and 1, a preference

[1] Our assumption of *a priori* best and worst consequences follows Luce and Raiffa (1957) and Anscombe and Aumman (1963), and it is technically without loss of generality in the sense that the preference order can always be extended to a larger domain that includes two additional consequences which by construction are better and worse, respectively, than all the original consequences. (Such an extension is demonstrated by SSK, Theorem 2.) The best and worst consequences ultimately serve to calibrate the definition and measurement of subjective probabilities, but even so the probabilities remain somewhat arbitrary, as will be shown.

of the form $\mathbf{X} \succsim \mathbf{H}_u$ means that "the expected utility of $\mathbf{X}$ is at least u." Equivalently, we will say that u is a *lower expected utility* for $\mathbf{X}$. Upper expected utilities are defined analogously. Using the terms defined above, we now introduce the first group of axioms that are assumed to govern rational preference:

A1 (Quasi order): $\succsim$ is transitive and reflexive.

A2 (Mixture-independence): $\mathbf{X} \succsim \mathbf{Y} \Leftrightarrow \alpha\mathbf{X} + (1-\alpha)\mathbf{Z} \succsim \alpha\mathbf{Y} + (1-\alpha)\mathbf{Z} \quad \forall\alpha \in (0,1)$.

A3 (Continuity in probability): If $\{\mathbf{X}_n\}$ and $\{\mathbf{Y}_n\}$ are convergent sequences such that $\mathbf{X}_n \succsim \mathbf{Y}_n$, then $\lim \mathbf{X}_n \succsim \lim \mathbf{Y}_n$.

A4 (Existence of best and worst): For all $c > 1$, $\mathbf{H}_1 \succsim \mathbf{H}_c \succsim \mathbf{H}_0$.

A5 (Coherence, or non-triviality): $\mathbf{H}_1 \succ \mathbf{H}_0$ (i.e., *not* $\mathbf{H}_0 \succsim \mathbf{H}_1$).

A1 and A2 are von Neumann and Morgenstern's first two axioms of expected utility, minus completeness[2], as applied to horse lotteries by Anscombe and Aumann (1963); see also Fishburn (1982). A3 is a strong continuity condition used by Garcia del Amo and Rios Insua (2002) that also works in infinite-dimensional spaces. A4 and A5 ensure non-triviality and provide reference points for probability measurement, as noted earlier.

DEFINITION: A collection of preferences $\{\mathbf{X}_n \succsim \mathbf{Y}_n\}$ is a *basis*[3] for $\succsim$ under an axiom system if every preference $\mathbf{X} \succsim \mathbf{Y}$ can be deduced from $\{\mathbf{X}_n \succsim \mathbf{Y}_n\}$ by direct application of those axioms.

The primal geometric representation of $\succsim$ is now given by:

Theorem 1 *$\succsim$ satisfies A1–A5 if and only if there exists a closed convex cone $\mathcal{B}^* \subset \mathcal{B}$, receding from the origin, such that for any horse lotteries $\mathbf{X}$ and $\mathbf{Y}$:*

$$\mathbf{X} \succsim \mathbf{Y} \Leftrightarrow \mathbf{X} - \mathbf{Y} \in \mathcal{B}^*.$$

In particular, if $\{\mathbf{X}_n \succsim \mathbf{Y}_n\}$ is a basis for $\succsim$ under A1–A5, then the cone $\mathcal{B}^*$ is the closed convex hull of the rays whose directions are $\{\mathbf{X}_n - \mathbf{Y}_n\}$ for all n together with $\{\mathbf{H}_1 - \mathbf{H}_c\}$ and $\{\mathbf{H}_c - \mathbf{H}_0\}$ for all c.[4]

Because the direction of preference between two horse lotteries $\mathbf{X}$ and $\mathbf{Y}$ depends only on the direction of the vector $\mathbf{X} - \mathbf{Y}$, it follows that if $\mathbf{EX} + (\mathbf{1} - \mathbf{E})\mathbf{Z} \succsim \mathbf{EY} + (\mathbf{1} - \mathbf{E})\mathbf{Z}$ where $\mathbf{E}$ is an event, then $\mathbf{EX} + (\mathbf{1} - \mathbf{E})\mathbf{Z}' \succsim \mathbf{EY} + (\mathbf{1} - \mathbf{E})\mathbf{Z}'$ for any $\mathbf{Z}'$. Consequently, we will simply write $\mathbf{EX} \succsim \mathbf{EY}$ to indicate that $\mathbf{EX} + (\mathbf{1} - \mathbf{E})\mathbf{Z} \succsim \mathbf{EY} + (\mathbf{1} - \mathbf{E})\mathbf{Z}$ for all $\mathbf{Z}$, or in other words, "$\mathbf{X}$ is preferred to $\mathbf{Y}$ conditional on the event $\mathbf{E}$." This result enables us to give a simple definition of conditional probability or expected utility: if $\mathbf{E}$ is an event and $\mathbf{X}$ is a horse lottery, then the preference $\mathbf{EX} \succsim \mathbf{EH}_u$ means that "the conditional expected utility of $\mathbf{X}$ given $\mathbf{E}$ is at least u."

[2] The completeness assumption asserts that for any $\mathbf{X}$ and $\mathbf{Y}$, either $\mathbf{X} \succsim \mathbf{Y}$ or $\mathbf{Y} \succsim \mathbf{X}$, or both. Here, it is permitted that neither of these conditions holds—i.e., $\mathbf{X}$ and $\mathbf{Y}$ may be incomparable.

[3] Use of the term "basis" in this context is due to SSK.

[4] Proofs have been suppressed in the conference version of the paper but are available in the complete version on the author's web site at http://www.duke.edu/∼rnau.

Now let a state dependent expected utility (s.d.e.u.) function be defined as a function $v : S \times C \mapsto \Re$, with the interpretation that $v(s,c)$ is the expected utility of receiving consequence c with probability 1 if state s obtains and receiving consequence 0 with probability 1 otherwise. Let $U_v(\mathbf{X})$ denote the expected utility assigned to a horse lottery $\mathbf{X}$ by the s.d.e.u. function v:

$$U_v(\mathbf{X}) \equiv \sum_{s \in S, c \in C} X(s,c)v(s,c).$$

DEFINITIONS: A s.d.e.u. function v is a *probability/utility pair* if it can be expressed as the product of a probability distribution on S and a state-independent utility function on C—i.e., if $v(s,c) = p(s)u(c)$ for some functions p and u. A s.d.e.u. function v *agrees* (one way) with $\succsim$ if $\mathbf{X} \succsim \mathbf{Y} \Rightarrow U_v(\mathbf{X}) \geq U_v(\mathbf{Y})$. A set $\mathcal{V}$ of s.d.e.u. functions *represents* $\succsim$ if $\mathbf{X} \succsim \mathbf{Y} \Leftrightarrow U_v(\mathbf{X}) - U_v(\mathbf{Y}) \geq 0 \; \forall \, v \in \mathcal{V}$.

We now have, as the dual to Theorem 1:

Theorem 2 $\succsim$ *satisfies A1–A5 if and only if it is represented by a non-empty closed convex set* $\mathcal{V}^*$ *of s.d.e.u. functions satisfying (w.l.o.g.)* $U_v(\mathbf{H}_0) = 0$ *and* $U_v(\mathbf{H}_1) = 1$.

(The proof relies on a separating hyperplane argument. For a similar result on a more general space, see Rios 1992.) If $\{\mathbf{X}_n \succsim \mathbf{Y}_n\}$ is a basis for $\succsim$, then $\mathcal{V}^*$ is merely the intersection of the linear constraints $\{U_v(\mathbf{X}_n) \geq U_v(\mathbf{Y}_n)\}$, $U_v(\mathbf{H}_0) = 0$, $U_v(\mathbf{H}_1) = 1$, and $0 \leq U_v(\mathbf{H}_c) \leq 1$ for all $c \geq 2$. If the basis is finite, then $\mathcal{V}^*$ is a convex polytope, whose elements need not be probability/utility pairs. Subsequent sections of the paper will discuss the additional assumptions needed to ensure that some points of $\mathcal{V}^*$—especially its extreme points—are probability/utility pairs.

3 The state-independence axiom

We now explore the implications, in the context of incompleteness, of the additional axiom introduced by Anscombe and Aumann[5] to provide the usual separation of subjective probability from utility. First, define the concept of a not-potentially-null event:

DEFINITION: An event $\mathbf{E}$ is *not potentially null* if $\mathbf{H}_\mathbf{E} \succsim \mathbf{H}_p$ for some $p > 0$.

Thus, an event that is not potentially null is precluded from having zero as an upper probability in any extension of $\succsim$ satisfying A1–A5. The final axiom is then:

A6 (State-independence): If $\mathbf{X}$ and $\mathbf{Y}$ are constant and $\mathbf{E}$ is not potentially null, then $\mathbf{EX} \succsim \mathbf{EY} \;\Rightarrow\; \mathbf{E}'\mathbf{X} \succsim \mathbf{E}'\mathbf{Y}$ for every other event $\mathbf{E}'$.

An immediate contribution of A6, in light of A4, is to guarantee that consequences 0 and 1 are best and worst *in every state*. Thus, if A6 holds, any s.d.e.u.

[5] Anscombe-Aumann refer to this assumption as "monotonicity in the prizes" or "substitutability."

function agreeing with $\succsim$ may be considered to belong to the set $\mathcal{V}^{+} \subset \mathcal{V}$ defined by:

$$\mathcal{V}^{+} \equiv \{v : 0 = v(s,0) \leq v(s,c) \leq v(s,1) \leq 1 \,\forall\, s \in S,\ c \geq 2;\ \sum_{s \in S} v(s,1) = 1\}.$$

Henceforth it will be assumed (arbitrarily but w.l.o.g.) that consequences 0 and 1 have the same numerical utilities, namely 0 and 1, in every state as well as unconditionally. Then, regardless of whether v is a probability/utility pair, define

$$p_v(s) \equiv v(s,1)$$

as "the" probability assigned to state s by v, since it is the expected utility of a horse lottery that yields a utility of 1 if state s obtains and 0 otherwise.[6] Correspondingly, if $\mathbf{E}$ is an event,

$$p_v(\mathbf{E}) \equiv U_v(\mathbf{H_E}) = \sum_{s \in \mathbf{E}} p_v(s)$$

is the probability assigned to $\mathbf{E}$ by v. Next define:

$$u_v(s,c) \equiv v(s,c)/v(s,1) \text{ if } v(s,1) > 0,$$

as the utility assigned to consequence c in state s by v. This utility is state-independent if v is a probability/utility pair, otherwise it is state-dependent. In these terms, the expected utility assigned to $\mathbf{X}$ by v can be rewritten as:

$$U_v(\mathbf{X}) = \sum_s p_v(s) \sum_c u_v(s,c) X(s,c).$$

We can now give a dual definition of conditional expected utility in terms of v in the obvious way:

$$U_v(\mathbf{X}|\mathbf{E}) = U_v(\mathbf{XE})/p_v(\mathbf{E}).$$

If the conditional expected utility of $\mathbf{X}$ given $\mathbf{E}$ is at least u by our primal definition—i.e., if $\mathbf{EX} \succsim \mathbf{EH}_u$—then dually we have $U_v(\mathbf{X}|\mathbf{E}) \geq u$ for any v agreeing with $\succsim$ and satisfying $p_v(\mathbf{E}) > 0$, because for any agreeing v:

$$\mathbf{EX} \succsim \mathbf{EH}_u \Rightarrow U_v(\mathbf{EX}) \geq U_v(\mathbf{EH}_u) = u p_v(\mathbf{E}) \Leftrightarrow U_v(\mathbf{X}|\mathbf{E}) \geq u \text{ or else } p_v(\mathbf{E}) = 0.$$

Another consequence of A6, in light of Theorem 1, is the property of stochastic dominance. In particular, if $\mathbf{X}$ is obtained from $\mathbf{Y}$ by shifting probability mass

[6] The same method of defining probabilities is used by Karni (1993). Since this definition is based on the arbitrary assignment of equal utilities to the best and worst outcomes in all states, it should not be interpreted as the "true" probability of a hypothetical decision maker whose preferences are represented by v. The classic definitions of subjective probability given by Savage, Anscombe-Aumann, and others, are all afflicted with the same arbitrariness. The intrinsic impossibility of inferring "true" probabilities from material preferences is discussed by Kadane and Winkler (1988), Schervish et al. (1990), Karni and Mongin (2000) and Nau (1995, 2002).

to consequence 1 from any other consequence, and/or from consequence 0 to any other consequence, in any state, then $\mathbf{X} \succsim \mathbf{Y}$. To see this, note that A6 together with A4 implies that $\mathbf{E}_s\mathbf{H}_1 \succsim \mathbf{E}_s\mathbf{H}_c$ and $\mathbf{E}_s\mathbf{H}_c \succsim \mathbf{E}_s\mathbf{H}_0$ for state s and any $c > 1$. Hence $\mathcal{B}^*$ contains all vectors of the form $\mathbf{E}_s(\mathbf{H}_1 - \mathbf{H}_c)$ and $\mathbf{E}_s(\mathbf{H}_c - \mathbf{H}_0)$. If $\mathbf{X} - \mathbf{Y}$ can be expressed as a non-negative linear combination of these vectors, then $\mathbf{X} - \mathbf{Y} \in \mathcal{B}^*$ and hence $\mathbf{X} \succsim \mathbf{Y}$. To make this result more precise, let the $[\,.\,]_{\min}$ ("minimum s.d.e.u.") operation be defined on $\mathcal{B}$ as follows:

$$[\mathbf{B}]_{\min} \equiv \min_{v \in \mathcal{V}^+} U_v(\mathbf{B}) = \min_{s \in S} [B(s,1) + \sum_{c \geq 2} \min\{0, B(s,c)\}].$$

This quantity is the minimum possible state-dependent expected utility that could be assigned to $\mathbf{B}$: it is achieved by assigning, within each state, a utility of 0 to those consequences $c \geq 2$ for which $\mathbf{B}$ is positive and a utility of 1 to those consequences $c \geq 2$ for which $\mathbf{B}$ is negative, then assigning a subjective probability of 1 to the state in which the conditional expected utility of $\mathbf{B}$ is minimized. Stochastic dominance and the negative orthant in $\mathcal{B}$ can now be defined in a natural way:

DEFINITIONS: $\mathbf{X} \geq^* [>^*]\ \mathbf{Y}$ ("$\mathbf{X}$ [strictly] dominates $\mathbf{Y}$") if $[\mathbf{X} - \mathbf{Y}]_{\min} \geq [>]\ 0$. The open negative orthant $\mathcal{B}^-$ consists of those $\mathbf{B}$ that are strictly dominated by the zero vector, i.e., $\mathcal{B}^- = \{\mathbf{B} \in \mathcal{B} : \mathbf{0} >^* \mathbf{B}\}$.

A6 in conjunction with A1–A5 then implies that $\mathbf{X} \geq^* [>^*]\ \mathbf{Y} \Rightarrow \mathbf{X} \succsim [\succ]\ \mathbf{Y}$. If preferences are complete (i.e., if for any horse lotteries $\mathbf{X}$ and $\mathbf{Y}$, either $\mathbf{X} \succsim \mathbf{Y}$ or $\mathbf{Y} \succsim \mathbf{X}$ or both), then the primal representation $\mathcal{B}^*$ is a half-space, the dual representation $\mathcal{V}^*$ consists of a unique s.d.e.u. function v^*, and axiom A6 requires the latter to be a probability/utility pair, which is the same result obtained by Anscombe and Aumann (1963). (A6 implies that $U_{v^*}(\mathbf{H}_c|\mathbf{E}_s) = U_{v^*}(\mathbf{H}_c)$ independent of the state s.) In the absence of completeness, the contribution of A6 to the separation of probability and utility is weaker, as summarized by:

Theorem 3 . *$\succsim$ satisfies A1–A6 if and only if it is represented by a nonempty convex set $\mathcal{V}^{**} \subseteq \mathcal{V}^+$ of s.d.e.u. functions of which at least one element is a probability/utility pair.*

If $\{\mathbf{X}_n \succsim \mathbf{Y}_n\}$ is a basis for $\succsim$ under axioms A1–A6, then any probability/utility pair v that satisfies $U_v(\mathbf{X}_n) \geq U_v(\mathbf{Y}_n)$ for all n, $v \in \mathcal{V}^+$, belongs to the set $\mathcal{V}^{**}$. Apart from this fact, it is not easy to characterize the set $\mathcal{V}^{**}$ in terms of probability/utility pairs, as will be illustrated in the sequel.

4 Strict vs. non-strict preference: an example

The results of the preceding section establish that a preference relation satisfying A1–A6 is represented by a closed set of s.d.e.u. functions of which at least one is a probability/utility pair. The closedness of the representing set is attributable to the use of non-strict preference as the behavioral primitive, together with a strong continuity assumption. In contrast, SSK use strict preference as the behavioral primitive, together with a weaker continuity assumption, to explicitly allow for the representation of incomplete preferences by open sets that may fail to contain probability/utility pairs.

The differences in these approaches are illustrated by an example of SSK (Example 4.1) comprising two states and three consequences, i.e., $S = \{1,2\}$ and $C = \{0,1,2\}$. Consequences 0 and 1 have state-independent utilities of 0 and 1 by assumption, so that a probability/utility pair is completely parameterized by the probability assigned to state 1 and the utility assigned to consequence 2. Consider, then, the two probability/utility pairs (p_i, u_i) in which $p_0(1) = 0.1$ and $p_1(1) = 0.3$, and $u_0(2) = 0.1$ and $u_1(2) = 0.4$. Let v_0 and v_1 denote the corresponding s.d.e.u. functions—i.e., $v_i(s,c) = p_i(s)u_i(c)$ for $i = 0,1$. Then $U_{v_i}(\mathbf{X})$ denotes the expected utility assigned to horse lottery $\mathbf{X}$ by (p_i, u_i). In particular, $U_{v_0}(\mathbf{H}_2) = 0.1$ and $U_{v_1}(\mathbf{H}_2) = 0.4$. Now let $\succ$ be defined as the preference relation that satisfies a weak Pareto condition with respect to these two probability/utility pairs—i.e., $\mathbf{X} \succ \mathbf{Y} \Leftrightarrow \{U_{v_0}(\mathbf{X}) > U_{v_0}(\mathbf{Y}) \text{ and } U_{v_1}(\mathbf{X}) > U_{v_1}(\mathbf{Y})\}$. Any s.d.e.u. function that is a convex combination of v_0 and v_1 also agrees with $\succ$, so the representing set $\mathcal{V}^{**}$ is the closed line segment whose endpoints are v_0 and v_1, but none of its interior points are probability/utility pairs.

Next SSK extend $\succ$ to obtain a new preference relation $\succ''$ by imposing the additional strict preferences $\mathbf{H}_{0.4} \succ'' \mathbf{H}_2 \succ'' \mathbf{H}_{0.1}$. The effect of this extension is to chop off the two endpoints of the representing set of s.d.e.u. functions, so that $\succ''$ is represented by the *open* line segment connecting v_0 with v_1. SSK point out that, although $\succ''$ satisfies all their axioms, there is no agreeing probability/utility pair for it, since the only two candidates have been deliberately excluded. They proceed to axiomatize the concept of "almost state- independent" utilities, which agree with a strict preference relation and are "within ε" of being state- independent. Clearly, $\succ''$ has an almost-state-independent representation, containing points arbitrarily close to v_0 and v_1.

In our framework, where the language of preference is non-strict, there is no way to implement a constraint such as $\mathbf{H}_2 \succ \mathbf{H}_{0.1}$ (i.e., to chop off v_0) except by asserting that $\mathbf{H}_2 \succsim \mathbf{H}_{0.1+\varepsilon}$ for a specific positive ε. And if this assertion is made, an interesting thing happens: axiom A6 begins to nibble on the v_0 end of the line segment and continues nibbling until the representation collapses to the v_1 end. To illustrate this process, let the non-zero elements of each v be written out as $v = (\{v(s,c)\}) = (v(1,1), v(2,1); v(1,2), v(2,2))$. Thus, $v_0 = (0.1, 0.9; 0.01, 0.09)$ and $v_1 = (0.3, 0.7; 0.12, 0.28)$. (Note that because these are probability/utility pairs, the first two numbers in parentheses are the probabilities of states 1 and 2, and the last two numbers are the same probabilities multiplied by the utility of consequence 2.) Next, let the line segment from v_0 to v_1 be parameterized by $v_\alpha \equiv (1-\alpha)v_0 + \alpha v_1$ for $\alpha \in (0,1)$. In these terms we obtain:

$$v_\alpha = (0.1 + 0.2\alpha, 0.9 - 0.2\alpha; 0.01 + 0.11\alpha, 0.09 + 0.19\alpha),$$

whence:

$$U_{v_\alpha}(\mathbf{H}_2) = v_\alpha(1,2) + v_\alpha(2,2) = 0.1 + 0.3\alpha \tag{4.1}$$

$$U_{v_\alpha}(\mathbf{H}_2|\mathbf{E}_1) = \frac{v_\alpha(1,2)}{v_\alpha(1,1)} = \frac{0.01 + 0.11\alpha}{0.1 + 0.2\alpha} \tag{4.2}$$

$$U_{v_\alpha}(\mathbf{H}_2|\mathbf{E}_2) = \frac{v_\alpha(2,2)}{v_\alpha(2,1)} = \frac{0.09 + 0.19\alpha}{0.9 - 0.2\alpha} \tag{4.3}$$

These are all monotone functions of α for α between 0 and 1, and they all are equal to 0.1 at $\alpha = 0$ and 0.4 at $\alpha = 1$. However, for intermediate values of α, (4.1) is greater than (4.3) and less than (4.2), and by invoking axiom A6, we can play the last two off against each other. In particular, it follows from monotonicity of (4.2) that

$$\alpha \geq \alpha^* \Rightarrow U_{v_\alpha}(\mathbf{H}_2|\mathbf{E}_1) \geq \frac{0.01 + 0.11\alpha^*}{0.1 + 0.2\alpha^*}, \tag{4.4}$$

whereas it follows from monotonicity of (4.3) that

$$U_{v_\alpha}(\mathbf{H}_2|\mathbf{E}_2) \geq u^* \Rightarrow \alpha \geq \frac{0.9u^* - 0.09}{0.2u^* + 0.19} \tag{4.5}$$

Let the set $\mathcal{V}^{**}$ representing the original relation $\succ$ henceforth be parameterized as $\mathcal{V}^{**} = \{v_\alpha | \alpha \in [0,1]\}$. Suppose that we now increase the lower utility of $\mathbf{H}_2$ by $\varepsilon = 0.01$ by adding the preference assertion $\mathbf{H}_2 \succsim \mathbf{H}_{0.11}$ to the basis for $\succ$. This additional assertion imposes the constraint $U_{v_\alpha}(\mathbf{H}_2) \geq 0.11$ for all v_α agreeing with the extended relation, thus excluding v_0 as an agreeing s.d.e.u. function. By application of A6, we may conclude that $U_{v_\alpha}(\mathbf{H}_2|\mathbf{E}_2) \geq 0.11$ as well. Substituting $u^* = 0.11$ in (4.5), it follows that the representing set must consist only of those v_α satisfying $\alpha \geq 0.042453$. But now, substituting $\alpha^* = 0.042453$ back into (4.4), we find that it must also satisfy $U_{v_\alpha}(\mathbf{H}_2|\mathbf{E}_1) \geq 0.135217$. Since $\mathbf{E}_1$ is not potentially null, A6 may be applied again to obtain $U_{v_\alpha}(\mathbf{H}_2) \geq 0.135217$. Thus, if we take one bite out of the line segment by imposing the constraint $U_{v_\alpha}(\mathbf{H}_2) \geq 0.11$, we end up concluding that a larger bite $U_{v_\alpha}(\mathbf{H}_2) \geq 0.135217$ may be taken! If we now repeat the process by substituting $u^* = 0.135217$ in (4.5), we obtain $\alpha^* = 0.146034$, which yields $U_{v_\alpha}(\mathbf{H}_2|\mathbf{E}_1) \geq 0.201721$ when substituted in (4.4). Successive iterations yield u^* values of 0.299288, 0.365247, 0.390144, 0.397381, 0.399317, and so on with rapid convergence to 0.4, which is realized (only) at v_1. The continuity axiom then allows us to assert that $\mathbf{H}_2 \succsim \mathbf{H}_{0.4}$, which together with the original constraint $\mathbf{H}_{0.4} \succsim \mathbf{H}_2$, establishes that the utility of consequence 2 is precisely 0.4.

If instead we start at the other endpoint, adding the constraint $\mathbf{H}_{0.4-\varepsilon} \succsim \mathbf{H}_2$ for $\varepsilon > 0$, the collapse occurs to the 0.1 value. If both constraints are added—i.e., if both endpoints are chopped off by finite margins, the entire interval is annihilated, yielding incoherence (a violation of A5). Hence, this example is unstable in the sense that any *finite* extension of the original preference relation leads to a collapse to one or the other of the original probability/utility pairs, or else to incoherence.

5 The need for stronger state-independence

The original preference relation in SSK’s example is represented by a set of s.d.e.u. functions whose extreme points are both probability/utility pairs. In our framework, if either of these points is excluded, then the intervening points must be excluded as well. Thus, in extending that relation, it is impossible to retain any agreeing state-dependent utilities that are not convex combinations of agreeing state-independent utilities. A second example shows that this is not always the case under axioms A1–A6. In other words, a preference relation can satisfy these axioms and yet not be represented by utilities that are state-independent or even “almost” state-independent.

Let there be three states and three consequences, and let **X** denote the horse lottery that satisfies $X(1,0) = X(2,2) = X(3,1) = 1$. That is, **X** yields consequences 0, 2, and 1 with certainty in states 1, 2, and 3 respectively. Suppose that all states are judged to have probability at least 0.1, and **X** is judged to have an unconditional expected utility of at least 0.5. Furthermore, a coin flip between **X** and {consequence 2 if state 1, otherwise **Z**} is preferred to a coin flip between utility 0.5 and {utility 0.9 if state 1, otherwise **Z**}, but also a coin flip between **X** and {utility 0.1 if state 2, otherwise **Z**} is preferred to a coin flip between utility 0.5 and {consequence 2 given state 2, otherwise **Z**}. (The common alternative **Z** is arbitrary by Theorem 1.) Thus, the basis for $\succsim$ is as follows:

$$\mathbf{H}_{\mathbf{E}} \succsim \mathbf{H}_{0.1} \text{ for } \mathbf{E} = \mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3, \tag{5.1}$$

$$\frac{1}{2}\mathbf{X} + \frac{1}{2}\mathbf{Z} \succsim \frac{1}{2}\mathbf{H}_{0.5} + \frac{1}{2}\mathbf{Z}, \tag{5.2}$$

$$\frac{1}{2}\mathbf{X} + \frac{1}{2}(\mathbf{E}_1\mathbf{H}_2 + (1-\mathbf{E}_1)\mathbf{Z}) \succsim \frac{1}{2}\mathbf{H}_{0.5} + \frac{1}{2}(\mathbf{E}_1\mathbf{H}_{0.9} + (1-\mathbf{E}_1)\mathbf{Z}), \tag{5.3}$$

$$\frac{1}{2}\mathbf{X} + \frac{1}{2}(\mathbf{E}_2\mathbf{H}_{0.1} + (1-\mathbf{E}_2)\mathbf{Z}) \succsim \frac{1}{2}\mathbf{H}_{0.5} + \frac{1}{2}(\mathbf{E}_2\mathbf{H}_2 + (1-\mathbf{E}_2)\mathbf{Z}), \tag{5.4}$$

Notice that (5.3) and (5.4) are obtained from (5.2) by replacing **Z** by subjective mixtures of **Z** with different constant lotteries on the LHS and RHS. These last two preferences imply that the lower bound on the expected utility of **X** among all *probability/utility pairs* agreeing with $\succsim$ must be strictly greater than 0.5. To understand this implication, note that under any s.d.e.u. function that agrees with $\succsim$, the differences in expected utility between the LHS’s and RHS’s of (5.2), (5.3), and (5.4), must all be non-negative. Moreover, if the s.d.e.u. function is a probability/utility pair, then in at least one of the two comparisons (5.3) and (5.4), the difference in expected utility between LHS and RHS must be strictly less than it is in (5.2), a situation that occurs when consequence 2 has a utility strictly greater than 0.1 and/or strictly less than 0.9. If the difference in expected utility between LHS and RHS is non-negative in all cases, then the difference can never be zero in (5.2)—i.e., **X** cannot have a lower expected utility as small as

0.5. In fact, the minimum expected utility of **X** among all probability/utility pairs agreeing with (5.1–5.4) is 0.564314.

The question is whether, by direct application of axioms A1–A6, we can infer that the expected utility of **X** is strictly greater than 0.5. The answer is: we cannot. The problem is that axiom A6 is useless here because of the common nonconstant term **X** in (5.2)–(5.4). In order to apply A6, we must first find non- negative linear combinations of the differences between the LHS's and RHS's of (5.1)–(5.4) that are conditionally constant—i.e., of the form **EB**, where **E** is an event and **B** is constant across states. But the search for such conditionally constant terms is constrained here by the presence of a common nonconstant term $\mathbf{X} - \mathbf{H}_{0.5}$ in the differences between LHS's and RHS's of (5.2)–(5.4). Furthermore, in order for A6 to "bite," **B** needs to have a negative lower expected utility when conditioned on some other event $\mathbf{E}'$. The effect of applying A6 will then be to raise this lower expected utility to zero, which shrinks the set of s.d.e.u. functions representing $\succsim$. In the example, the few conditionally-constant lottery differences **EB** that can be constructed from (5.1)–(5.4) all turn out to satisfy $\mathbf{B} \geq^* 0$, which is completely uninformative. The lower expected utility of **X** therefore remains at 0.5 despite the fact that this value is not realized, *or even closely approached*, by any probability/utility pair agreeing with $\succsim$.

This example shows that when preferences are incomplete, axiom A6 is insufficient to guarantee that they are represented by a set of probability/utility pairs (or their convex hull). Evidently, an additional state-independence condition is needed, such as:

A7 (Stochastic substitution): If

$$\alpha\mathbf{X} + (1-\alpha)(\mathbf{E}\mathbf{X}' + (1-\mathbf{E})\mathbf{Z}) \succsim \alpha\mathbf{Y} + (1-\alpha)(\mathbf{E}\mathbf{Y}' + (1-\mathbf{E})\mathbf{Z})$$

for some $\alpha \in (0,1)$ where $\mathbf{X}'$ and $\mathbf{Y}'$ and **Z** are constant lotteries and **E** is not potentially null, then

$$\alpha\mathbf{X} + (1-\alpha)(p\mathbf{X}' + (1-p)\mathbf{Z}) \succsim \alpha\mathbf{Y} + (1-\alpha)(p\mathbf{Y}' + (1-p)\mathbf{Z})$$

for some $p \in (0,1]$.

In other words, the subjective mixtures of the constant lotteries $\mathbf{X}'$ and $\mathbf{Y}'$ with **Z** can be replaced with objective mixtures *against the background* of a comparison between the nonconstant lotteries **X** and **Y**. In terms of the primal representation $\mathcal{B}^*$, this assumption means that if $\mathbf{B} + \mathbf{E}\mathbf{B}' \in \mathcal{B}^*$, where $\mathbf{B}'$ is constant across states and **E** is not potentially null, then $\mathbf{B} + p\mathbf{B}' \in \mathcal{B}^*$ for some $p > 0$.[7] Note that if a collection of preferences $\{\mathbf{X}_n \succsim \mathbf{Y}_n\}$ satisfies A1–A6, then the imposition of A7 cannot produce a contradiction. A1–A6 require the existence of at least one probability/utility pair agreeing with $\{\mathbf{X}_n \succsim \mathbf{Y}_n\}$, and any probability/utility pair that agrees with the original preferences will also agree with any new preferences generated from them by A7.

[7] **A2** and **A6** imply only that this substitution may be performed in the nonstochastic case $\mathbf{B} = \mathbf{0}$.

The new axiom *does* affect the counterexample discussed above. (5.3) and (5.4) can now be replaced by

$$\frac{1}{2}\mathbf{X}+\frac{1}{2}(p\mathbf{H}_2+(1-p)\mathbf{Z})\succsim\frac{1}{2}\mathbf{H}_{0.5}+\frac{1}{2}(p\mathbf{H}_{0.9}+(1-p)\mathbf{Z}),$$

$$\frac{1}{2}\mathbf{X}+\frac{1}{2}(p'\mathbf{H}_{0.1}+(1-p')\mathbf{Z})\succsim\frac{1}{2}\mathbf{H}_{0.5}+\frac{1}{2}(p'\mathbf{H}_2+(1-p')\mathbf{Z}),$$

for some $p,p'>0$. A mixture of these two comparisons in a ratio of p' to p yields:

$$\frac{1}{2}\mathbf{X}+\frac{1}{2}(\alpha\mathbf{H}_{0.1}+\alpha\mathbf{H}_2+(1-2\alpha)\mathbf{Z})\succsim\frac{1}{2}\mathbf{H}_{0.5}+\frac{1}{2}(\alpha\mathbf{H}_{0.9}+\alpha\mathbf{H}_2+(1-2\alpha)\mathbf{Z}),$$

where $\alpha=pp'/(p+p')$. The LHS must have greater-or-equal expected utility than the RHS, which (because of the $\mathbf{H}_{0.1}$ term on the left and the $\mathbf{H}_{0.9}$ term on the right, and cancellation of the common terms $\mathbf{H}_2$ and $\mathbf{Z}$) means that $\mathbf{X}$ must have strictly greater expected utility than 0.5.

The main result, which generalizes this example, can now be stated as:

Theorem 4 $\succsim$ *satisfies A1–A7 if and only if it is represented by a nonempty set $\mathcal{V}^{***}$ of s.d.e.u. functions that is the convex hull of a set of probability/utility pairs.*

If $\{\mathbf{X}_n\succsim\mathbf{Y}_n\}$ is a basis for $\succsim$ under A1–A7, then $\mathcal{V}^{***}$ is merely the convex hull of the set of probability/utility pairs that satisfy $\{U_v(\mathbf{X}_n)\geq U_v(\mathbf{Y}_n)\}$. If the basis is finite, the construction of $\mathcal{V}^{***}$ can be carried out as follows. First, form the convex polyhedron consisting of the intersection of the constraints $\{U_v(\mathbf{X}_n)\geq U_v(\mathbf{Y}_n)\}$, $v\in\mathcal{V}^{+}$. Now take the intersection of this polyhedron with the nonconvex surface consisting of all probability/utility pairs. (If the latter intersection is empty, the preferences do not satisfy A1–A7: they are incoherent.) Finally, take the convex hull of what remains: this is the set $\mathcal{V}^{***}$.

6 Discussion

It has been shown that, in order to obtain a convenient representation of incomplete preferences by sets of probability/utility pairs, it does not suffice merely to delete the completeness axiom from the standard axiomatic framework of Anscombe and Aumann. This finding is not due to technical problems with limits or null events, but rather to a fundamental weakness of the traditional state-independence axiom in the absence of completeness. Our approach is to introduce an additional state-independence postulate (A7) that has "bite" in the absence of completeness. SSK follow a different approach in their axiomatization of incomplete strict preferences. Instead of directly strengthening the state-independence property, they "fill in the missing preferences" by indirect reasoning, namely, they assume the preference relation has the property that $\neg(\mathbf{X}\succsim\mathbf{Y})\Rightarrow\mathbf{Y}\succ\mathbf{X}$, where "$\neg$" stands for "it is precluded that," meaning that there is no extension of $\succsim$ satisfying the other axioms in which $\mathbf{X} \succsim \mathbf{Y}$ (p. 2204 ff.). SSK's assumption requires that wherever a weak preference is precluded, the opposite strict preference must be affirmed. In our framework, this property of $\succsim$ is not implied by axioms A1–A6, hence it amounts to an additional axiom of rationality. The lack of this property is illustrated by the example of the preceding section, in which it is precluded that $\mathbf{H}_u \succsim \mathbf{X}$ for any $u < 0.5643....$, yet it is not implied by A1–A6 that $\mathbf{X} \succ \mathbf{H}_u$ for any $u > 0.5$. If the "axiom" of indirect reasoning is added to A1–A6, in lieu of A7, the representation of Theorem 4 follows immediately from Theorem 3.

References

[1] Anscombe, F. and R. Aumann. A Definition of Subjective Probability. *Ann. Math. Statist.* **34** 199–205, 1963.

[2] Aumann, R. Utility Theory Without the Completeness Axiom. *Econometrica* **30** 445–462, 1962.

[3] Drèze, J. Decision Theory with Moral Hazard and State-Dependent Preferences. In *Essays on Economic Decisions Under Uncertainty*. Cambridge University Press, 1987.

[4] Fishburn, P.C. *The Foundations of Expected Utility* D. Reidel, Dordrecht, 1982.

[5] Del Amo, A. and D Rios Insua A Note on an Open Problem in the Foundations of Statistics. *Rev. R. Acad. Cinc. Serie A. Mat.* **96**:1 55–61, 2002.

[6] Giron, F.J., and S. Rios. Quasi-Bayesian Behavior: A More Realistic Approach to Decision Making? In J.M. Bernardo et. al. (eds.), *Bayesian Statistics*. University Press, Valencia, Spain, 1980.

[7] Kadane, J.B. and R.L. Winkler. Separating Probability Elicitation from Utilities. *J. Amer. Statist. Assoc.* **83**:402 357–363, 1988.

[8] Karni, E. and P. Mongin. On the Determination of Subjective Probability by Choices. *Management Science* **46** 233–248, 2000.

[9] Karni, E., Schmeidler, D. and K. Vind. On State-Dependent Preferences and Subjective Probabilities. *Econometrica* **51**1021–1031, 1983.

[10] Karni, E. *Decision Making Under Uncertainty: The Case of State-Dependent Preferences*. Harvard University Press, Cambridge, 1985.

[11] Karni, E. A Definition of Subjective Probabilities with State-Dependent Preferences. *Econometrica* **61**187–198, 1993.

[12] Luce, R.D. and H. Raiffa. *Games and Decisions: Introduction and Critical Survey.* Wiley, New York, 1957.

[13] Moskowitz, Preckel and Yang. Decision Analysis with Incomplete Probability and Utility Information. *Operns. Res* **41** 864–879, 1993.

[14] Nau, R. Indeterminate Probabilities on Finite Sets. *Ann. Statist.* **20**:4 1737–1767, 1992.

[15] Nau, R. Coherent Decision Analysis with Inseparable Probabilities and Utilities. *J. Risk and Uncertainty* **10** 71–91, 1995.

[16] Nau, R. De Finetti Was Right: Probability Does Not Exist. *Theory and Decision* **51** 89–124, 2002.

[17] Rios Insua, D. *Sensitivity Analysis in Multiobjective Decision Making*. Springer-Verlag, 1990.

[18] Rios Insua, D. On the Foundations of Decision Making under Partial Information. *Theory and Decision* **33** 83–100, 1992.

[19] Schervish, M.J, T. E. Seidenfeld and J.B. Kadane. State- Dependent Utilities. *J. Amer. Statist. Assoc.* **85**:411 840–847, 1990.

[20] Seidenfeld, T., M. J. Schervish and J. B. Kadane. Decisions Without Ordering. In W. Sieg (ed.) *Acting and Reflecting*, Reidel, Dordrecht, 1990.

[21] Seidenfeld, T.E., M.J. Schervish and J.B. Kadane. A Representation of Partially Ordered Preferences. *Ann. Statist.* **23** 2168–2174, 1995.

[22] Smith, C.A.B. Consistency in Statistical Inference and Decision. *J. Roy. Stat. Soc. B* **23** 1–25, 1961.

[23] Staum, J. Pricing and Hedging in Incomplete Markets: Fundamental Theorems and Robust Utility Maximization. *Math. Finance*, forthcoming.

[24] Suppes, P. The Measurement of Belief. *J. Roy. Stat. Soc. B* **36** 160–175, 1974.

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

[26] Williams, P.M. Indeterminate Probabilities. In M. Przełęcki, K. Szaniawski, and R. Wójcicki (eds.), *Formal Methods in the Methodology of Empirical Sciences* Ossolineum and D. Reidel, Dordecht, Holland 229–246, 1976.

Robert Nau is with the Fuqua School of Business, Duke University, Durham, NC 27708-0120 USA. E-mail: robert.nau@duke.edu

Convex Imprecise Previsions: Basic Issues and Applications

R. PELESSONI
University of Trieste, Italy

P. VICIG
University of Trieste, Italy

Abstract

In this paper we study two classes of imprecise previsions, which we termed convex and centered convex previsions, in the framework of Walley's theory of imprecise previsions. We show that convex previsions are related with a concept of convex natural estension, which is useful in correcting a large class of inconsistent imprecise probability assessments. This class is characterised by a condition of avoiding unbounded sure loss. Convexity further provides a conceptual framework for some uncertainty models and devices, like unnormalised supremum preserving functions. Centered convex previsions are intermediate between coherent previsions and previsions avoiding sure loss, and their not requiring positive homogeneity is a relevant feature for potential applications. Finally, we show how these concepts can be applied in (financial) risk measurement.

Keywords

imprecise previsions, convex imprecise previsions, convex natural extension, risk measures

1 Introduction

Imprecise probability theory is developed by P. Walley in [14] in terms of two major classes of (unconditional) imprecise previsions, relying upon reasonable consistency requirements: *avoiding sure loss* and *coherent* previsions. The condition of avoiding sure loss is less restrictive than coherence but is often too weak.

Coherent imprecise previsions have been studied more extensively, while imprecise previsions that avoid sure loss received less attention, and it is an interesting problem to state whether some special class of previsions avoiding sure loss can be identified, which is such that

(a) its properties are not too far from those of coherent previsions;

(b) it gives further insight into the theory of imprecise previsions or generalises some of its basic aspects;

(c) it may express beliefs which do not match with coherence but which are useful in formalising and dependably modelling certain kinds of problems.

The main aim of this paper is to discuss the properties and some applications of two classes of imprecise previsions, which we termed convex and centered convex previsions and which let us provide some answers to points (a), (b), (c). The paper partly summarises and complements [12], where proofs may be found for those results which are stated without proof here.

After recalling some basic notions in Section 2, we study the larger class of convex lower previsions in Section 3. Although our conclusion is that convexity is an unsatisfactory consistency requirement – for instance, convex previsions do not necessarily avoid sure loss – it is however important as far as (b) is concerned. That is seen in Section 3, where a notion of convex natural extension is discussed which formally parallels the basic concept of natural extension in [14]. We characterise lower previsions whose convex natural extension is finite as those complying with the (mild) requirement of avoiding unbounded sure loss. In this case the convex natural extension indicates a canonical (least-committal) way of correcting them into a convex assessment. As discussed in Section 3.0.1, it is then easy to make a further correction to achieve the stronger (and more satisfactory) property of centered convexity.

Centered convex previsions are discussed in Section 3.0.1, together with generalisations of the important envelope theorem. Centered convex lower previsions are a special class of previsions avoiding sure loss, retaining several properties of coherent imprecise previsions, and hence they appear to fulfil requirement (a).

Section 4 gives some answers to point (c). Here convex previsions provide a conceptual framework for certain kinds of uncertainty models, as shown in Examples 1 (overly prudential assessments) and 2 (supremum preserving functions). These models are sometimes employed in practice, although they cannot usually be regarded as satisfactory. Centered convex previsions do not require the positive homogeneity condition $\underline{P}(\lambda X) = \lambda \underline{P}(X)$, $\forall \lambda > 0$, and hence seem appropriate to capture risk aversion. In Section 4 we focus in particular on risk measurement problems, showing that the results in Section 3 may be used to define convex risk measures (centered or not) for an arbitrary set of random variables $\mathcal{D}$. In particular, the definition of convex risk measure coincides, when $\mathcal{D}$ is a linear space, with the concept of convex risk measure recently introduced in the literature to consider liquidity risks [4, 5, 7]. It appears here that results from the risk measurement area can profitably contribute to the development of imprecise probability theory and viceversa. Section 5 concludes the paper.

2 Preliminaries

Unless otherwise specified, in the sequel we shall denote with $\mathcal{D}$ an *arbitrary* set of bounded random variables (or gambles, in the notation of [14]) and with $\mathcal{L}$ ($\supset \mathcal{D}$) the set of all bounded random variables (on a possibility space). A *lower prevision* $\underline{P}$ (an *upper prevision* $\overline{P}$, a *prevision* P) on $\mathcal{D}$ is a real-valued function with domain $\mathcal{D}$. In particular, if $\mathcal{D}$ contains only indicator functions of events, $\underline{P}$ ($\overline{P}$, P) is termed lower probability (upper probability, probability).

Lower (and upper) previsions should satisfy some consistency requirements: the condition of *avoiding sure loss* and the stronger *coherence* condition [14].

Definition 1 *$\underline{P}: \mathcal{D} \rightarrow \mathbb{R}$ is a lower prevision on $\mathcal{D}$ that* avoids sure loss *iff, for all $n \in \mathbf{N}^+$, $\forall\ X_1, \dots, X_n \in \mathcal{D}$, $\forall\ s_1, \dots, s_n$ real and non-negative, defining $\underline{G} = \sum_{i=1}^{n} s_i(X_i - \underline{P}(X_i))$, $\sup \underline{G} \geq 0$.*

Definition 2 *$\underline{P}: \mathcal{D} \rightarrow \mathbb{R}$ is a* coherent lower prevision *on $\mathcal{D}$ if and only if, for all $n \in \mathbf{N}^+$, $\forall\ X_0, X_1, \dots, X_n \in \mathcal{D}$, $\forall\ s_0, s_1, \dots, s_n$ real and non-negative, defining $\underline{G} = \sum_{i=1}^{n} s_i(X_i - \underline{P}(X_i)) - s_0(X_0 - \underline{P}(X_0))$, $\sup \underline{G} \geq 0$.*

The condition of avoiding sure loss is too weak under many respects: for instance, it does not require that $\underline{P}(X) \geq \inf X$, nor does it impose monotonicity. On the other hand, it is simpler to assess and to check than coherence.

Behaviourally, a lower prevision assessment $\underline{P}(X)$ may be viewed as a supremum buying price for X [14], and $s(X - \underline{P}(X))$ represents an *elementary gain* from a bet on X, with stake s. We shall say that the bet is *in favour* of X if $s \geq 0$, whilst $-s(X - \underline{P}(X))$ ($s \geq 0$) is an elementary gain from a bet *against* X. Definitions 1 and 2 both require that no admissible linear combination $\underline{G}$ of elementary gains originates a sure loss bounded away from zero. The difference is that the concept of avoiding sure loss considers only bets in favour of the X_i, while coherence considers also (at most) one bet against a random variable in $\mathcal{D}$.

We recall the following properties of coherent lower previsions, which hold whenever the random variables involved are in $\mathcal{D}$:

(a) $\underline{P}(\lambda X) = \lambda \underline{P}(X)$, $\forall \lambda > 0$ (positive homogeneity)

(b) $\inf X \leq \underline{P}(X) \leq \sup X$ (internality)

(c) $\underline{P}(X + Y) \geq \underline{P}(X) + \underline{P}(Y)$ (superlinearity).

Coherent precise previsions may be defined by modifying Definition 2 to allow $n \geq 0$ bets in favour of and $m \geq 0$ bets against random variables in $\mathcal{D}$ ($m, n \in \mathbf{N}$). A coherent precise prevision P is necessarily *linear* and *homogeneous*: $P(aX + bY) = aP(X) + bP(Y)$, $\forall a, b \in \mathbb{R}$. In particular $P(0) = 0$.

Coherent lower previsions may be characterised using precise previsions [14]:

Theorem 1 (Lower envelope theorem) *A lower prevision $\underline{P}$ on $\mathcal{D}$ is coherent iff $\underline{P}$ is the lower envelope of some set $\mathcal{M}$ of coherent precise previsions on $\mathcal{D}$, i.e. iff*

$$\underline{P}(X) = \inf_{P \in \mathcal{M}} \{P(X)\}, \forall X \in \mathcal{D} \text{ (inf \textit{is attained).}}$$

Upper and *lower* previsions are customarily related by the *conjugacy* relation $\overline{P}(X) = -\underline{P}(-X)$. An upper prevision $\overline{P}(X)$ may be viewed as an infimum selling price for X and an *elementary gain* from a bet concerning X is written as $s(\overline{P}(X) - X)$. The definitions of coherence and of the condition of avoiding sure loss are modified accordingly.

3 Convex Lower Previsions

Convex Previsions.

Definition 3 *$\underline{P} : \mathcal{D} \rightarrow I\!R$ is a* convex lower prevision *on $\mathcal{D}$ iff, for all $n \in \mathbf{N}^+$, $\forall\, X_0, X_1, \ldots, X_n \in \mathcal{D}$, $\forall\, s_1, \ldots, s_n$ real and non-negative such that $\sum_{i=1}^n s_i = 1$* (convexity condition)*, defining $\underline{G} = \sum_{i=1}^n s_i(X_i - \underline{P}(X_i)) - (X_0 - \underline{P}(X_0))$,* $\sup \underline{G} \geq 0$.[1]

Any coherent lower prevision is convex, since Definition 3 is obtained from Definition 2 adding the constraint $\sum_{i=1}^n s_i = s_0 = 1$ (note that we would get a definition equivalent to Definition 3 requiring only $\sum_{i=1}^n s_i = s_0 > 0$). Conversely, a convex lower prevision does not even necessarily avoid sure loss:

Proposition 1 *Let $\underline{P}$ be a convex lower prevision on $\mathcal{D}$ and let $0 \in \mathcal{D}$. Then $\underline{P}$ avoids sure loss iff $\underline{P}(0) \leq 0$.*

Convexity is characterised by a set of axioms if $\mathcal{D}$ has a special structure:

Theorem 2 *Let $\underline{P} : \mathcal{D} \rightarrow I\!R$.*

(a) If $\mathcal{D}$ is a linear space *containing real constants, $\underline{P}$ is a convex lower prevision iff it satisfies the following axioms:*[2]

(T) $\underline{P}(X + c) = \underline{P}(X) + c, \forall X \in \mathcal{D}, \forall c \in I\!R$ *(translation invariance)*

(M) $\forall X, Y \in \mathcal{D}$*, if* $Y \leq X$ *then* $\underline{P}(Y) \leq \underline{P}(X)$ *(monotonicity)*

(C) $\underline{P}(\lambda X + (1 - \lambda)Y) \geq \lambda \underline{P}(X) + (1 - \lambda)\underline{P}(Y), \forall X, Y \in \mathcal{D}, \forall \lambda \in [0, 1]$ *(concavity).*

(b) If $\mathcal{D}$ is a convex cone*, $\underline{P}$ is a convex lower prevision iff it satisfies (C) and*

(M1) $\forall \mu \in I\!R, \forall X, Y \in \mathcal{D}$*, if* $X \geq Y + \mu$ *then* $\underline{P}(X) \geq \underline{P}(Y) + \mu$.

Proposition 2 *Some properties of convex lower previsions.*

[1] The term 'convex' in 'convex prevision' refers to the convexity condition $\sum_{i=1}^n s_i = 1$ $(s_i \geq 0)$, which distinguishes convex lower (upper) previsions from coherent lower (upper) previsions (cf. Definitions 2, 3 and 7) and convex natural extensions from natural extensions (cf. Definition 4 and Section 3.0.1). The term 'convex prevision' is therefore unrelated with convexity or concavity properties of previsions as real functions.

[2] (T) and (M) can be replaced by $\underline{P}(X) - \underline{P}(Y) \leq \sup(X - Y), \forall X, Y \in \mathcal{D}$.

(a) (Convergence theorem) *Let* $\{\underline{P}_j\}_{j=1}^{+\infty}$ *be a sequence of lower previsions, convex on* $\mathcal{D}$ *and such that* $\forall X \in \mathcal{D}$ *there exists* $\lim_{j\rightarrow+\infty}\underline{P}_j(X) = \underline{P}(X)$*. Then* $\underline{P}$ *is convex on* $\mathcal{D}$*.*

(b) (Convexity theorem) *If* $\underline{P}_1$ *and* $\underline{P}_2$ *are convex lower previsions on* $\mathcal{D}$*, so is* $\underline{P}(X) = \lambda\underline{P}_1(X) + (1-\lambda)\underline{P}_2(X), \forall\lambda \in [0,1]$.

Let $\underline{P}$ *be a convex lower prevision on* $\mathcal{D}$*. The following properties hold (whenever all random variables involved are in* $\mathcal{D}$*):*

(c) If $\underline{P}(0) \geq 0$, $\underline{P}(\lambda X) \geq \lambda\underline{P}(X), \forall\lambda \in [0,1]$ *and* $\underline{P}(\lambda X) \leq \lambda\underline{P}(X), \forall\lambda > 1$

(d) $\underline{P}(0) + \inf X \leq \underline{P}(X) \leq \underline{P}(0) + \sup X$

(e) $\forall\mu \in \mathbb{R}$, $\underline{P}^*(X) = \underline{P}(X) + \mu$ *is convex on* $\mathcal{D}$*.*

Properties (a) and (b), which are quite analogous to corresponding properties of coherent previsions and previsions avoiding sure loss [14], point out ways of obtaining new convex lower previsions from given ones. Property (c) shows that convexity is compatible with lack of positive homogeneity, but requires the condition $\underline{P}(0) \geq 0$. Property (d) highlights a sore point of convexity: $\underline{P}(X)$ need not belong to the closed interval $[\inf X, \sup X]$ (*internality* may fail).[3]

Property (d) suggests that internality could be restored imposing $\underline{P}(0) = 0$, if $0 \notin \mathcal{D}$; by (e), if $0 \in \mathcal{D}$ and $\underline{P}(0) \neq 0$, then $\underline{P}^*(X) = \underline{P}(X) - \underline{P}(0)$ is convex and $\underline{P}^*(0) = 0$. Requiring $\underline{P}(0) = 0$ is also the only choice to make $\underline{P}$ avoid sure loss (Proposition 1), while assuring that (c) holds.

Thinking of the meaning of a lower prevision, it appears extremely reasonable to add condition $\underline{P}(0) = 0$ to convexity: it would be at least weird to give an estimate (even imprecise) of the non-random variable 0 which is other than zero.

Convex Natural Extension. Before considering the stronger class of centered convex previsions, we introduce the notion of convex natural extension, which is strictly related to convexity.

Definition 4 *Let* $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$ *be a lower prevision, Z an arbitrary (bounded) random variable. Define* $g_h = s_h(X_h - \underline{P}(X_h))$, $L = \{\alpha : Z - \alpha \geq \sum_{i=1}^n g_i$*, for some* $n \geq 1, X_i \in \mathcal{D}, s_i \geq 0$*, with* $\sum_{i=1}^n s_i = 1\}$. $\underline{E}_c(Z) = \sup L$ *is termed* convex natural extension[4] *of* $\underline{P}$ *on Z.*

It is clear that L is always non-empty (putting $n = 1$, $s_1 = 1$, $X_1 = X \in \mathcal{D}$ in its definition, $\alpha \in L$ for $\alpha \leq \inf Z - \sup X + \underline{P}(X)$), while $\underline{E}_c(Z)$ can in general be infinite. This situation is characterised in the following Proposition 3.

[3] Non-internality cannot anyway be two-sided: if there exists $X \in \mathcal{D}$ such that $\underline{P}(X) > \sup X$ ($\underline{P}(X) < \inf X$), then $\underline{P}(Y) > \inf Y$ ($\underline{P}(Y) < \sup Y$), $\forall Y \in \mathcal{D}$. This is easily seen applying Definition 3, with $n = 2$, $\{X_0, X_1\} = \{X, Y\}$.

[4] The reason why $\underline{E}_c$ is termed ‘extension’ appears from the later Theorem 3, especially (d).

Definition 5 *$\underline{P} : \mathcal{D} \to \mathbb{R}$ is a lower prevision that* avoids unbounded sure loss *on $\mathcal{D}$ iff there exists $k \in \mathbb{R}$ such that, for all $n \in \mathbf{N}^+$, $\forall\, X_1, \ldots, X_n \in \mathcal{D}$, $\forall\, s_1, \ldots, s_n$ real and non-negative with $\sum_{i=1}^n s_i = 1$, defining $\underline{G} = \sum_{i=1}^n s_i (X_i - \underline{P}(X_i))$, $\sup \underline{G} \geq k$.*

Remark 1 *Definition 5 generalises Definition 1: $\underline{P}$ avoids unbounded sure loss if and only if $\underline{P} + k$ avoids sure loss for some $k \in \mathbb{R}$, since the last inequality in Definition 5 may be written as $\sup \sum_{i=1}^n s_i (X_i - (\underline{P}(X_i) + k)) \geq 0$ and the constraint $\sum_{i=1}^n s_i = 1$ is not restrictive for Definition 1. Note also that if $\underline{P} + k$ avoids sure loss, then so does $\underline{P} + h$, $\forall h \leq k$. Therefore, when $\underline{P}$ avoids unbounded sure loss, defining $\overline{k} = \sup\{k \in \mathbb{R} : \underline{P} + k$ avoids sure loss$\}$, $\underline{P}$ avoids sure loss too whenever $\overline{k} \geq 0$. As a further remark, it can be seen that the constraint $\sum_{i=1}^n s_i = 1$ is essential in Definition 5: wiping it out would make Definition 5 equivalent to Definition 1.*

Proposition 3 *$\underline{E}_c(Z)$ is finite, whatever is Z, iff $\underline{P}$ avoids unbounded sure loss.*

Proof. Suppose first that $\underline{P}$ avoids unbounded sure loss and for an arbitrary Z let $\alpha \in L$. Then $Z - \alpha \geq \sum_{i=1}^n s_i (X_i - \underline{P}(X_i))$ for some $X_1, \ldots, X_n \in \mathcal{D}$ and $s_1, \ldots, s_n \geq 0$ with $\sum_{i=1}^n s_i = 1$, and hence $\sup Z - \alpha \geq \sup \sum_{i=1}^n s_i (X_i - \underline{P}(X_i)) \geq k$, using Definition 5 at the last inequality. Therefore $\underline{E}_c(Z) \leq \sup Z - k$.

Conversely, suppose now that $\underline{P}$ does not avoid unbounded sure loss. Therefore, for each $k \in \mathbb{R}$ there are $X_1, \ldots, X_n \in \mathcal{D}$ and $s_1, \ldots, s_n \geq 0$ with $\sum_{i=1}^n s_i = 1$ such that $\sum_{i=1}^n s_i (X_i - \underline{P}(X_i)) < k \leq Z - (-k + \inf Z)$. This implies, for any Z, $-k + \inf Z \in L$ and, by the arbitrariness of k, $\underline{E}_c(Z) = +\infty$. □

The condition of avoiding unbounded sure loss is rather mild. For instance, it clearly holds whenever $\mathcal{D}$ is finite. It is also implied by convexity, as shown by the following proposition, while the converse implication is generally not true.

Proposition 4 *If $\underline{P} : \mathcal{D} \to \mathbb{R}$ is convex, it avoids unbounded sure loss.*

Proof. Choose arbitrarily $X_1, \ldots, X_n \in \mathcal{D}$ and $s_1, \ldots, s_n \geq 0$ such that $\sum_{i=1}^n s_i = 1$ in Definition 5. Given $X_0 \in \mathcal{D}$, use convexity to write $0 \leq \sup\{\sum_{i=1}^n s_i (X_i - \underline{P}(X_i)) - (X_0 - \underline{P}(X_0))\} \leq \sup\{\sum_{i=1}^n s_i (X_i - \underline{P}(X_i))\} - (\inf X_0 - \underline{P}(X_0))$, and hence $\sup\{\sum_{i=1}^n s_i (X_i - \underline{P}(X_i))\} \geq k = \inf X_0 - \underline{P}(X_0)$. □

We state now some properties of the convex natural extension. An indirect characterisation of the convex natural extension will be given in Theorem 5.

Theorem 3 *Let $\underline{P} : \mathcal{D} \to \mathbb{R}$ be a lower prevision which avoids unbounded sure loss and $\underline{E}_c$ its convex natural extension. Then*

(a) $\underline{E}_c$ is a convex prevision on $\mathcal{L}$ and $\underline{E}_c(X) \geq \underline{P}(X), \forall X \in \mathcal{D}$

(b) $\underline{P}$ is convex if and only if $\underline{E}_c = \underline{P}$ on $\mathcal{D}$

(c) If $\underline{P}^*$ is a convex prevision on $\mathcal{L}$ such that $\underline{P}^*(X) \geq \underline{P}(X)$ $\forall X \in \mathcal{D}$, then $\underline{P}^*(Z) \geq \underline{E}_c(Z), \forall Z \in \mathcal{L}$

(d) If $\underline{P}$ is convex, $\underline{E}_c$ is the minimal convex extension of $\underline{P}$ to $\mathcal{L}$

(e) $\underline{P}$ avoids sure loss on $\mathcal{D}$ if and only if $\underline{E}_c$ avoids sure loss on $\mathcal{L}$.

3.0.1 The Role of the Convex Natural Extension

The properties of $\underline{E}_c$ closely resemble those of the *natural extension* $\underline{E}$ [14] of a lower prevision $\underline{P}$, whose definition differs from that of $\underline{E}_c$ only for the lack of the constraint $\sum_{i=1}^{n} s_i = 1$. In particular, as $\underline{E}$ characterises coherence of $\underline{P}$ ($\underline{P}$ is coherent iff $\underline{E}$ coincides with $\underline{P}$ on $\mathcal{D}$), $\underline{E}_c$ characterises convexity of $\underline{P}$.

Property (d) lets us extend $\underline{P}$ to *any* $\mathcal{D}' \supset \mathcal{D}$ (maintaining convexity) by considering the restriction of $\underline{E}_c$ to $\mathcal{D}'$. Moreover, (e) guarantees that $\underline{E}_c$ inherits the condition of avoiding sure loss when $\underline{P}$ satisfies it.

It is well known that the natural extension is finite iff $\underline{P}$ avoids sure loss, and when finite it can correct $\underline{P}$ into a coherent assessment in a canonical way. Analogously, the convex natural extension is finite iff $\underline{P}$ avoids unbounded sure loss, and can be used to correct $\underline{P}$ into a convex assessment, although property (e) warns us that $\underline{E}_c$ will still incur sure loss if $\underline{P}$ does so. This problem can be solved using Proposition 2, (e): $\underline{P}^*(X) = \underline{E}_c(X) - \underline{E}_c(0)$ is a correction of $\underline{P}$ which avoids sure loss by Proposition 1, as $\underline{P}^*(0) = 0$. This also means that $\underline{P}^*$ is a centered convex prevision by Definition 6 in the next section.

Alternatively, the convex natural extension may be employed to correct an assessment $\underline{P}$ which avoids unbounded sure loss (but not sure loss) into $\underline{P}'$, which avoids sure loss but is not necessarily convex. In fact, $\underline{P} + h$ avoids sure loss $\forall h \leq \overline{k} < 0$ (cf. Remark 1). Since it can be shown that $\overline{k} = -\underline{E}_c(0)$, it ensues that $\underline{E}_c(0)$ is the minimum k to be subtracted from $\underline{P}$ to make $\underline{P}' = \underline{P} - k$ avoid sure loss.

Hence, the convex natural extension points out ways of correcting an assessment incurring (bounded) sure loss into one avoiding sure loss, a problem which cannot be answered using the natural extension. These corrections can be applied in several interesting situations, including, as already noted, the case of a finite $\mathcal{D}$.

Centered Convex Previsions and Envelope Theorems. The considerations at the end of Section 3 lead us naturally to the following stronger notion of centered convexity:

Definition 6 *A lower prevision $\underline{P}$ on domain $\mathcal{D}$ $(0 \in \mathcal{D})$ is* centered convex *(*C-convex*, in short) iff it is convex and $\underline{P}(0) = 0$.*[5]

Proposition 5 *Let $\underline{P}$ be a centered convex lower prevision on $\mathcal{D}$. Then*

(a) $\underline{P}$ has a convex natural extension (hence at least one centered convex extension) on any $\mathcal{D}' \supset \mathcal{D}$

(b) $\underline{P}(\lambda X) \geq \lambda \underline{P}(X), \forall \lambda \in [0,1],\ \underline{P}(\lambda X) \leq \lambda \underline{P}(X), \forall \lambda \in]-\infty, 0[\cup]1, +\infty[$

[5] As shown in [12], we obtain an equivalent definition of centered convex lower prevision by requiring $\underline{P}(0) = 0$ and relaxing the convexity condition $\sum_{i=1}^{n} s_i = s_0 > 0$ to $\sum_{i=1}^{n} s_i \leq s_0$.

(c) $\inf X \le \underline{P}(X) \le \sup X, \forall X \in \mathcal{D}$

(d) $\underline{P}$ *avoids sure loss.*

Besides, the convergence and convexity theorems hold for C-convex previsions too (replacing ‘convex’ with ‘centered convex’ in Proposition 2, (a) and (b)).

Properties (a)÷(d) show that centered convexity is significantly closer to coherence than convexity: C-convex lower previsions are a special class of previsions avoiding sure loss, retaining several properties of coherence and the extension property of convexity, but not requiring positive homogeneity.

A convex prevision $\underline{P}$ which is not centered may still be avoiding sure loss, if $\underline{P}(0) < 0$ (Proposition 1), but in general it is only warranted by Proposition 4 that it avoids unbounded sure loss, a very weak consistency requirement.

Remark 2 (Convexity and n-coherence) *The consistency notion of n-coherence is discussed in [14], Appendix B, illustrating how it can be appropriate for certain ‘bounded rationality’ models. If the model does not require positive homogeneity, n-coherence alone is inadequate: 1-coherence is too weak, being equivalent to the internality condition (c) in Proposition 5, 2-coherence is too strong, as on linear spaces it is equivalent to two axioms, one of which is positive homogeneity [14]. As a matter of fact, C-convex previsions are a special class of 1-coherent (but not necessarily 2-coherent) previsions.*

An indirect comparison among convexity, centered convexity and coherence is given by their corresponding envelope theorems. We firstly recall that it was proved in [14] that any lower envelope of coherent lower previsions is coherent. Here is the parallel statement for convex lower previsions, while the generalisation of Theorem 1 (lower envelope theorem) comes next.

Proposition 6 *Let $\mathcal{P}$ be a set of convex lower previsions all defined on $\mathcal{D}$. If $\underline{P}(X) = \inf_{\underline{Q}\in\mathcal{P}} \left\{\underline{Q}(X)\right\}$ is finite $\forall X \in \mathcal{D}$, $\underline{P}$ is convex on $\mathcal{D}$.*

Theorem 4 (Generalised envelope theorem) *$\underline{P}$ is convex on $\mathcal{D}$ iff there exist a set $\mathcal{P}$ of coherent precise previsions on $\mathcal{D}$ and a function $\alpha : \mathcal{P} \to \mathbb{R}$ such that:*

(a) $\underline{P}(X) = \inf_{P\in\mathcal{P}}\{P(X) + \alpha(P)\}, \forall X \in \mathcal{D}$ *(*inf *is attained).*

Moreover, $\underline{P}$ is centered convex iff ($0 \in \mathcal{D}$ and) both (a) and the following (b) hold:

(b) $\inf_{P\in\mathcal{P}}\{\alpha(P)\} = 0$ *(*inf *is attained).*

A result similar to Theorem 4 was proved in risk measurement theory [4], requiring $\mathcal{D}$ to be a linear space. The proof of Theorem 4, given in [12] in the framework of imprecise prevision theory, is simpler and imposes no structure on $\mathcal{D}$.

Remark 3 *In particular, the constructive implication of the theorem (for convex previsions) enables us to obtain convex previsions as lower envelopes of translated precise previsions. Its proof follows easily from Proposition 6 and Proposition 2, (e): every precise prevision P is convex and so is $P+\alpha(P)$, by Proposition 2, (e);* $\inf_{P\in\mathcal{P}}\{P(X)+\alpha(P)\}$ *is a convex prevision by Proposition 6.*

Remark 4 *Let $\underline{P}$ be a lower prevision and $\mathcal{S}$ the set of all coherent precise previsions on $\mathcal{L}$. Define also $\mathcal{M}(\underline{P}) = \{(Q,r) \in \mathcal{S} \times \mathbb{R} : Q(X) + r \geq \underline{P}(X), \forall X \in \mathcal{D}\}$. It ensues from Theorem 4 that convexity of $\underline{P}$ can be equivalently characterised by the condition $\underline{P}(X) = \inf\left\{Q(X)+r : (Q,r) \in \mathcal{M}(\underline{P})\right\}$ $\forall X \in \mathcal{D}$; C-convexity can be characterised by adding the constraint* $\inf\left\{r : \exists Q \in \mathcal{S} : (Q,r) \in \mathcal{M}(\underline{P})\right\} = 0$ *(cf. also the following Theorem 5, where the lower envelope concerns all $X \in \mathcal{L}$).*

The envelope theorem characterisations of convexity, centered convexity and coherence differ about the role of function α, which is unconstrained with convexity, non-negative and such that $\min\alpha = 0$ with centered convexity, identically equal to zero with coherence (in this case Theorem 4 reduces to Theorem 1).

The result in the next theorem characterises the convex natural extension as the lower envelope of a set of translated coherent precise previsions and can be proved in a way similar to the natural extension theorem in [14], Section 3.4.

Theorem 5 *Let $\underline{P}$ be a lower prevision on $\mathcal{D}$ which avoids unbounded sure loss and define $\mathcal{S}$ and $\mathcal{M}(\underline{P})$ as in Remark 4. Then, $\mathcal{M}(\underline{P}) = \mathcal{M}(\underline{E}_c)$ and $\underline{E}_c(X) =$* $\inf\left\{Q(X)+r : (Q,r) \in \mathcal{M}(\underline{P})\right\}, \forall X \in \mathcal{L}$.

4 Some Applications

We have seen so far that convexity may help in correcting several inconsistent assessments. As noted in Section 3.0.1, its usefulness in this problem is essentially instrumental: we may easily go further and arrive at a centered convex correction, which guarantees a more satisfactory degree of consistency.

Turning to other problems, some uncertainty modelisations give rise to convex previsions, as in the examples which follow. We emphasise that we do not maintain that these models are reasonable, but simply that they are sometimes adopted in practice, and that convexity supplies a conceptual framework for them.

Example 1 (Overly prudential assessments) *Persons or institutions which have to evaluate the random variables in a set $\mathcal{D}$ are often unfamiliar with uncertainty theories. In this case, a solution is to gather n experts and ask each of them to formulate a precise prevision (or an expectation) for all $X \in \mathcal{D}$. Choosing* $\underline{P}(X) = \min_{i=1,\dots,n} P_i(X), \forall X$ *(where P_i is expert i's evaluation) as one's own opinion is an already prudential way of pooling the experts'opinions, and originates a coherent lower prevision. Some more caution or lack of confidence toward some experts may lead to replacing every P_i with $P_i^* = P_i - \alpha_i$ before performing the* *minimum, where* $\alpha_i \geq 0$ *measures in some way the final assessor's personal caution or his/her (partially) distrusting expert i. By Theorem 4,* $\underline{P}^* = \min_{i=1,\ldots,n} P_i^*$ *is convex (cf. Remark 3). More generally,* $\underline{P}^*$ *is of course convex also when the sign of the* α_i *is unconstrained (*$\alpha_i < 0$ *if, for instance, expert i's opinion is believed to be biased and below the 'real' prevision). It is interesting to observe that if* $\alpha_i \geq 0$ *for at least one i,* $\underline{P}^*$ *avoids sure loss too (since then* $\underline{E}_c(0) \leq 0$ *by Theorem 5, hence* $\underline{E}_c$ *avoids sure loss by Proposition 1, and so does* $\underline{P}^*$ *by Theorem 3, (e)). In particular, the following situation may be not unusual with an unexperienced assessor:* $\alpha_i > 0$ *for some i, and* $0 \notin \mathcal{D}$*, because the assessor thinks that no expert is needed to evaluate* 0*, he himself can assign, of course,* $\underline{P}^*(0) = 0$*. If such is the case, the extension of* $\underline{P}^*$ *on* $\mathcal{D} \cup \{0\}$ *keeps on avoiding sure loss, as is easily seen, but is generally not convex (to see this with a simple example, suppose* $X \in \mathcal{D}$*,* $\underline{P}^*(X) < \inf X$ *and use the result in footnote 3 to obtain that* $\underline{P}^*(0) < 0$ *is then necessary for convexity).*

In the following example and in Section 4 we shall refer to upper previsions, to which the theory developed so far extends with mirror-image modifications. We report the conjugates of Definition 3 and Theorem 4.

Definition 7 $\overline{P} : \mathcal{D} \to \mathbb{R}$ *is a* convex upper prevision *on* $\mathcal{D}$ *iff, for all* $n \in \mathbf{N}^+$*,* $\forall X_0, X_1, \ldots, X_n \in \mathcal{D}$*,* $\forall s_1, \ldots, s_n$ *real and non-negative such that* $\sum_{i=1}^n s_i = 1$ (convexity condition)*, defining* $\overline{G} = \sum_{i=1}^n s_i(\overline{P}(X_i) - X_i) - (\overline{P}(X_0) - X_0)$*,* $\sup \overline{G} \geq 0$.

Theorem 6 $\overline{P}$ *is convex on its domain* $\mathcal{D}$ *iff there exist a set* $\mathcal{P}$ *of coherent precise previsions (all defined on* $\mathcal{D}$*) and a function* $\alpha : \mathcal{P} \to \mathbb{R}$ *such that:*

(a) $\overline{P}(X) = \sup_{P \in \mathcal{P}} \{P(X) + \alpha(P)\}, \forall X \in \mathcal{D}$ (sup *is attained).*

Moreover, $\overline{P}$ *is centered convex iff (*$0 \in \mathcal{D}$ *and) both (a) and the following (b) hold:*

(b) $\sup_{P \in \mathcal{P}} \{\alpha(P)\} = 0$ (sup *is attained).*

Example 2 (Supremum preserving functions) *Let* $\mathbb{P} = \{\omega_i\}_{i \in I}$ *be a (not necessarily finite) set of exhaustive non-impossible elementary events or* atoms*, i.e.* $\omega_i \neq \varnothing$ $\forall i \in I$*,* $\cup_{i \in I} \omega_i = \Omega$*,* $\omega_i \cap \omega_j = \varnothing$ *if* $i \neq j$*. Given a function* $\pi : \mathbb{P} \to [0,1]$*, define* $\Pi : 2^{\mathbb{P}} - \{\varnothing\} \to [0,1]$ *(*$2^{\mathbb{P}}$ *is the powerset of* $\mathbb{P}$*) by*

$$\Pi(A) = \sup_{\omega_i \in A} \{\pi(\omega_i)\}, \forall A \in 2^{\mathbb{P}} - \{\varnothing\}. \tag{1}$$

As well-known, if π *is normalised (i.e.,* $\sup \pi = 1$*) and extended to* $\varnothing$ *putting* $\pi(\varnothing)(= \Pi(\varnothing)) = 0$*,* Π *is a normalised possibility measure, a special case of coherent upper probability [3]. Without these additional assumptions,* Π *is a convex upper probability. To see this, define for* $i \in I$*,* $P_i(\omega_i) = 1$*,* $P_i(\omega_j) = 0$ $\forall j \neq i$*,* $\alpha_i = \pi(\omega_i) - 1$*, and extend (trivially) each* P_i *to* $2^{\mathbb{P}}$*. It is not difficult to see that* $\Pi(A) = \sup_{i \in I} \{P_i(A) + \alpha_i\}$*,* $\forall A \in 2^{\mathbb{P}}$ *and therefore* Π *is convex by Theorem 6. If*

$\sup \pi < 1$, Π *has the unpleasant property that* $\Pi(\Omega) < 1$, *and also* $\Pi(\varnothing) < 0$ *(this means that* Π *incurs sure loss and is not C-convex). Functions similar to these kinds of unnormalised possibilities were considered in the literature relating possibility and fuzzy set theories, and their unsatisfactory properties were already pointed out (see e.g. [9], Section 2.6 and the references quoted therein).*

Convex Risk Measures. Further applications of convex imprecise previsions are suggested by the fact that they do not necessarily require positive homogeneity, as appears from Proposition 5, (b). Considering the well-known behavioural interpretation of lower (and upper) previsions [14], it is intuitively clear that applications could be generally related to situations of risk aversion, because of which an agent's supremum buying price for the random quantity λX might be less than λ times his/her supremum buying price for X, when $\lambda > 1$.

In this section we shall discuss an application to (financial) risk measurement. The literature on risk measures is quite large, as this topic is very important in many financial, banking or insurance applications. Formally, a risk measure is a mapping ρ from a set $\mathcal{D}$ of random variables into $\mathbb{R}$. Therefore ρ associates a real number $\rho(X)$ to every $X \in \mathcal{D}$, which should determine how 'risky' X is, and whether it is acceptable to buy or hold X. Intuitively, X should be acceptable (not acceptable) if $\rho(X) \leq 0$ (if $\rho(X) > 0$), and $\rho(X)$ should determine the maximum amount of money which could be subtracted from X, keeping it acceptable (the minimum amount of money to be added to X to make it acceptable).

Traditional risk measures, like Value-at-Risk (*VaR*) – probably the most widespread – require assessing (at least) a distribution function for each $X \in \mathcal{D}$; often, a joint normal distribution is assumed [8]. Quite recently, other risk measures were introduced, which do not require assessing exactly one precise probability distribution for each $X \in \mathcal{D}$, and are therefore appropriate also in situations where conflicting or insufficient information is available. Precisely, coherent risk measures were defined in a series of papers (including [1, 2]) using a set of axioms (among these positive homogeneity), and assuming that $\mathcal{D}$ is a linear space. In these papers, coherent risk measures were not related with imprecise previsions theory, while this was done in [11, 13]; see also [10] for a general approach to these and other theories. Convex risk measures were introduced in [4, 5, 7] as a generalisation of coherent risk measures which does not require the positive homogeneity axiom. We report the definition in [5]:

Definition 8 *Let* $\mathcal{V}$ *be a linear space of random variables which contains real constants.* $\rho : \mathcal{V} \to \mathbb{R}$ *is a* convex risk measure *iff it satisfies the following axioms:*

(T1) $\forall X \in \mathcal{V}, \forall \alpha \in \mathbb{R}, \rho(X + \alpha) = \rho(X) - \alpha$ *(translation invariance)*

(M2) $\forall X, Y \in \mathcal{V}$, *if* $X \leq Y$ *then* $\rho(Y) \leq \rho(X)$ *(monotonicity)*

(C1) $\rho(\lambda X + (1-\lambda)Y) \leq \lambda\rho(X) + (1-\lambda)\rho(Y)\ \forall X, Y \in \mathcal{V}, \lambda \in [0,1]$ *(convexity).*

Convex risk measures are also discussed in [6] and their potential capability of capturing risk aversion is pointed out in [5]. In a risk measurement environment, a motivation for not assuming positive homogeneity is that $\rho(\lambda X)$ may be larger than $\lambda\rho(X)$ for $\lambda > 1$ also because of *liquidity risks*: if we were to sell immediately a large amount λX of a financial investment, we might be forced to accept a smaller reward than λ times the current selling price for X.

It was shown in [11] that risk measures can be encompassed into the theory of imprecise previsions, because a risk measure for X can be interpreted as an upper prevision for $-X$:[6]

$$\rho(X) = \overline{P}(-X). \tag{2}$$

This fact was used in [11, 13] to generalise the notion of coherent risk measures to an arbitrary domain $\mathcal{D}$. An analogue generalisation can be done for convex risk measures [12], as we shall now illustrate.

Definition 9 *$\rho : \mathcal{D} \to I\!R$ is a* convex risk measure *on $\mathcal{D}$ if and only if for all $n \in \mathbf{N}^+$, $\forall\, X_0, X_1, \ldots, X_n \in \mathcal{D}$, $\forall\, s_1, \ldots, s_n$ real and non-negative such that $\sum_{i=1}^n s_i = 1$, defining $\overline{G} = \sum_{i=1}^n s_i(X_i + \rho(X_i)) - (X_0 + \rho(X_0))$, $\sup \overline{G} \geq 0$.*

Note that Definition 9 may be obtained from Definition 7 referring to $-X$ rather than X, for all $X \in \mathcal{D}$.

If $\mathcal{D}$ is a linear space containing real constants, the notion in Definition 9 reduces to that in [4, 5], by the next theorem (cf. also Theorem 2, (a)):

Theorem 7 *Let $\mathcal{V}$ be a linear space of bounded random variables containing real constants. A mapping ρ from $\mathcal{V}$ into $I\!R$ is a convex risk measure according to Definition 9 iff it is a convex risk measures according to Definition 8.*

Definition 9 applies to any set $\mathcal{D}$ of random variables, unlike Definition 8, which, if $\mathcal{D}$ is arbitrary, requires embedding it in a larger linear space.

Results specular to those presented in Section 3 apply to convex risk measures. In particular, the convergence and convexity theorems (Proposition 2, (a) and (b)) hold; convex risk measures can be extended on any $\mathcal{D}' \supset \mathcal{D}$, preserving convexity; they avoid sure loss iff $\rho(0) \geq 0$ (we say that ρ avoids sure loss on $\mathcal{D}$ iff $\overline{P}(-X) = \rho(X)$ avoids sure loss on $\mathcal{D}^- = \{-X : X \in \mathcal{D}\}$).

Like the general case in Section 3, it appears quite appropriate to put $\rho(0) = 0$, and hence to use *centered convex* risk measures: 0 is the unquestionably reasonable selling or buying price for $X = 0$.

Definition 10 *A mapping ρ from $\mathcal{D}$ $(0 \in \mathcal{D})$ into $I\!R$ is a* centered convex risk measure *on $\mathcal{D}$ iff ρ is convex and $\rho(0) = 0$.*

[6] We assume that the time gap between the buying and selling time of X is negligible (if not, we should introduce a discounting factor in (2)). This simplifies the sequel, without substantially altering the conclusions.

Centered convex risk measures have further nice additional properties, corresponding to those of centered convex lower previsions: they always avoid sure loss, and are such that $-\sup X \leq \rho(X) \leq -\inf X, \forall X \in \mathcal{D}$.

This condition corresponds to internality ((c) of Proposition 5), and is a rationality requirement for risk measures: for instance, $\rho(X) > -\inf X$ would mean that to make X acceptable we require adding to it a sure number ($\rho(X)$) higher than the maximum loss X may cause.

A centered convex risk measures ρ is not necessarily positively homogeneous:

$$\rho(\lambda X) \geq \lambda \rho(X), \forall \lambda \geq 1. \tag{3}$$

A notion of convex natural extension may also be given for centered convex (or convex) risk measures and its properties correspond to those listed in Theorem 3. When finite, it gives in particular a standard way of 'correcting' other kinds of risk measures into convex risk measures.[7]

The generalised envelope theorem is obtained from the statement of Theorem 6 replacing $\overline{P}(X)$ and $\underline{P}(X)$ with, respectively, $\rho(X)$ and $\underline{P}(-X)$.

Examples of convex risk measures may be found in [4, 5, 12].

5 Conclusions

In this paper we studied convex and centered convex previsions in the framework of Walley's theory of imprecise previsions. Convex previsions do not necessarily satisfy minimal consistency requirements, but are useful in generalising natural extension-like methods of correcting inconsistent assessments and in providing a conceptual framework for some uncertainty models. Centered convex previsions are in a sense intermediate between avoiding sure loss and coherence: their properties are closer to coherence than those of a generic prevision that avoids sure loss, but are also compatible with lack of positive homogeneity. Because of this, they are potentially useful at least in models which incorporate some forms of risk aversion. We outlined a risk measurement application, where they lead to defining convex risk measures, and believe that several applications of convex imprecise previsions are still to be explored. It might also be interesting to investigate if and how convex previsions can be generalised in a conditional environment, or when allowing unbounded random variables.

References

[1] P. Artzner. Application of coherent risk measures to capital requirements in insurance. *North American Actuarial Journal*, 3:11–26, 1999.

[7] Note that this is always possibile if $\mathcal{D}$ is finite (cf. Section 3.0.1).

[2] P. Artzner, F. Delbaen, S. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9:203–228, 1999.

[3] G. de Cooman, D. Aeyels. Supremum preserving upper probabilities. *Information Sciences*, 118:173–212, 1999.

[4] H. Föllmer, A. Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6:429–447, 2002.

[5] H. Föllmer, A. Schied. Robust preferences and convex measures of risk. In *Advances in Finance and Stochastics*, K. Sandmann, K., P. J. Schönbucher (eds.), Springer-Verlag, 39–56, 2002.

[6] M. Frittelli, E. R. Gianin. Putting order in risk measures. *Journal of Banking and Finance*, 26:1473–1486, 2002.

[7] D. Heath, H. Ku. Market equilibrium with coherent measures of risk. Preprint, Dept. Math. Sciences, Carnegie Mellon University, 2001.

[8] J. C. Hull. Options, Futures and other Derivatives. 4th ed., Prentice Hall Inc., Upper Saddle River, NJ, USA, 2000.

[9] G. J. Klir, M. J. Wierman. Uncertainty-Based Information. Physica-Verlag, Heidelberg, 1998.

[10] S. Maaß. Exact functionals and their core. *Statistical Pap.*, 43:75–93, 2002.

[11] R. Pelessoni, P. Vicig. Coherent risk measures and imprecise previsions. In *Proc. ISIPTA'01*, Ithaca, NY, 307–315, 2001.

[12] R. Pelessoni, P. Vicig. Convex imprecise previsions for risk measurement. Quad. Dip. Mat. Appl. 'B. de Finetti', 5/02, Università di Trieste, available at http://www.univ.trieste.it/∼matappl/quaderni.htm, 2002.

[13] R. Pelessoni, P. Vicig. Imprecise previsions for risk measurement. *Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, to appear.

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

Renato Pelessoni and **Paolo Vicig** are with the Department of Applied Mathematics 'B. de Finetti', University of Trieste, Piazzale Europa 1, I-34127 Trieste, Italy. E-mails: renato.pelessoni@econ.units.it, paolo.vicig@econ.units.it

Reliability Analysis in Geotechnics with Finite Elements — Comparison of Probabilistic, Stochastic and Fuzzy Set Methods

G.M. PESCHL
Graz University of Technology, Austria

H.F. SCHWEIGER
Graz University of Technology, Austria

Abstract

The finite element method is widely used for solving various problems in geotechnical engineering practice. The input parameters required for the calculations are generally imprecise. The paper is devoted to a comparison of probabilistic, stochastic and fuzzy set method for reliability analysis with respect to its applicability for practical problems in geotechnical engineering. Emphasis will be given by comparing the effects of modelling uncertainty using different methods, with special reference to the role of spatial correlation. After introducing some basic notions about the approaches, this article shows that the results obtained with the fuzzy set method for a simple bearing capacity problem agree with the outcomes by a probabilistic and a stochastic method. Advantages and shortcomings of either approach with respect to practical applications will be discussed.

Keywords

finite element method, probabilistic, fuzzy set, stochastic modelling, random field, spatial correlation

1 Introduction

It is well known that material parameters of geomaterials may scatter within a considerable range. Thus, a high degree of uncertainty may be introduced in any type of analysis if material parameters are treated as deterministic values. There is no agreement about what method should be used, to account for these uncertainties especially in practical geotechnical problems where usually not sufficient information is available for a rigorous stochastic analysis, because site investigation and laboratory testing are restricted due to financial und time constraints.

It is still possible to use probabilistic methods in these problems by making suitable assumptions on the statistics of the uncertainties, at least to some extent, by combining different sources of information via Bayes' theorem. However, the numerical values obtained by probabilistic analysis (e.g. probability of failure) are quite sensitive to changes in the input distribution parameters ([1, 13]), but play an important rule in comparative and qualitative studies [14]. On the other hand, Fuzzy set methods provide an appropriate mathematical model which can be used for quantitative assessment.

In the developed methodology point estimate methods (PEM) for probabilistic analyses and fuzzy set method for possibilistic analyses together with a finite element model is used. Emphasis will be given to comparison with methods employing a stochastic model, which means that the parameters are described by spatial random fields (e.g. [7]). This stochastic approach employs the Monte-Carlo method and is used in this paper as a reference.

Both variability and spatial correlation lengths of material properties can affect the reliability of geotechnical systems. In this article, elasto-plastic finite element analysis has been combined with theories mentioned above to investigate the influence of material variability and spatial correlation lengths on the computation of the bearing capacity of a smooth rigid strip footing on a weightless soil with shear strength parameters c and φ under plane strain conditions [14]. The soil stratum is compressed by incrementally displacing the top surface vertically downwards. Geometry and boundary conditions of the problem are shown in figure 1.

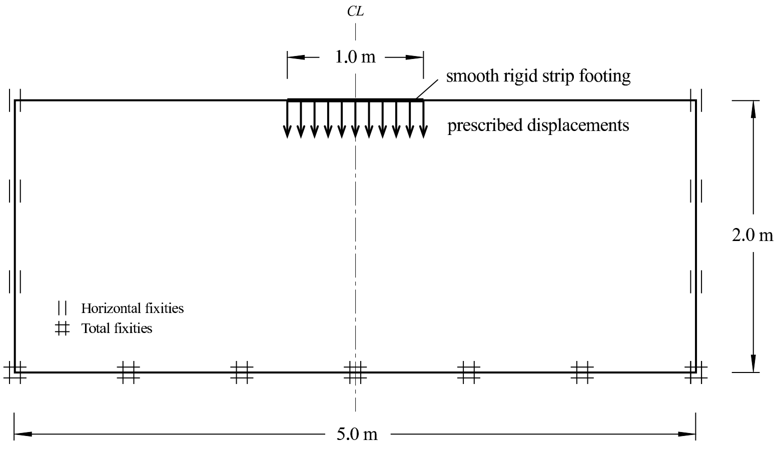


Figure 1: Geometry and boundary conditions

In the simulations, the mean cohesion (μ_c) and mean angle of friction (μ_φ) have been held constant at 100 kN/m^2 and 25° while the coefficient of variation, (COV=σ_c/μ_c), and the spatial correlation length, (Θ), are varied systematically. For this investigation, it is assumed that when the variability in the cohesion is large, the variability in the friction angle will also be large. The material parameters required for the model used are: Young's modulus (E), Poisson's ratio (ν), dilatancy angle (ψ), cohesion (c), and friction angle (φ). In the present study, E, ν and ψ are held constant (at 100.000 kN/m^2, 0.3, and 0, respectively) while c and φ are basic variables. It has to be pointed out that the interaction and cross-correlation between the shear strength parameters is neglected in this study.

The question is how the variability of the shear strength parameters c and φ affects the response given by the dimensionless *bearing capacity factor*, N_c, and consequently the reliability of the structure. The bearing capacity factor is traditionally defined by $N_c = q_f \ / \ c$ where q_f is the computed bearing capacity and c is the cohesion of the soil. The theoretical bearing capacity factor, N_c, for a spatially constant friction angle is given by Sokolovski [19]:

$$N_c = \frac{1}{\tan\varphi}\left[e^{\pi\tan\varphi}\tan^2\left(45+\frac{\varphi}{2}\right)-1\right]$$

2 Spatial variability of soil properties

In principle, the spatial variation of a soil layer can be characterized in detail, but only if a large number of tests can be performed. Thus, for geotechnical purposes a simplification is introduced in which spatial variability is subdivided into two parts, i.e. a linear trend, and a residual variability (stochastic description) about that trend [15]. Figure 2 depicts the value of the soil property, u, at a boring location as a function of depth, z, where $\mu_u(z)$ describes the trend which is represented by a depth-dependent mean value. The stochastic description of the soil property, $u(z)$, consists of the standard deviation, $\sigma_u(z)$, and the scale of fluctuation or autocorrelation length, Θ_u, of $u(z)$.

The spatial correlation length measures the distance within which the property shows relatively strong correlation from point to point. The soil is modelled as a random field ([21, 16]), which is a stochastic process defined by three coordinates in space. This means that the properties of the soil in a specific point are described as a random variable. Rather than a characterization of soil properties at every point, data are used to estimate a smooth trend, and remaining variations are described statistically because of the lack of data.

Spatial averaging.

The mean of large volumes remains the same as the mean of small volumes, but the standard deviation of the average property from one large volume to the next is smaller than the standard deviation of the average property from one small

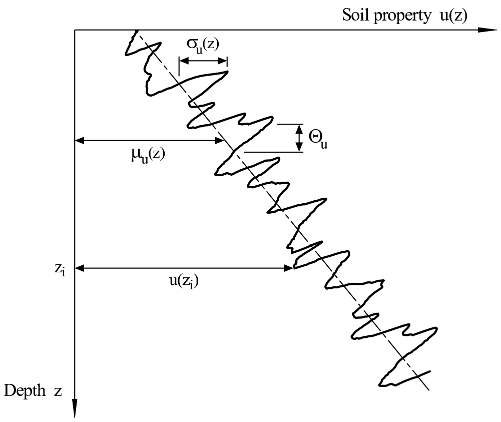


Figure 2: Spatial variability of a soil layer

volume to the next [21]. The extent of averaging of soil properties, $u(z)$, within a large volume depends on the structure of spatial variation. More precisely, the extent of averaging depends on the standard deviation of properties, σ_u, from point to point and on the autocorrelation function. Similarly, the standard deviations of the spatial averages, $u_{\Delta z}$ and u_V, are $\sigma_{u_{\Delta z}}$ and σ_{u_V}, respectively. Therefore, the larger the length Δz or the volume V over which the property is averaged, the more variations of u tends to produce a reduction in the process of spatial averaging. This tends to originate a reduction in standard deviation as the size of the averaged length or volume increases. The so-called reduction factor $\Gamma_u(V)$ is defined as the dimensionless ratio between σ_{u_V} and σ_u ($\Gamma_u(V) = \sigma_{u_V} / \sigma_u$).

The square of the reduction factor, Γ_u^2, will be called the variance function, whereas for the two-dimensional case it will take the form: $\Gamma_u^2(\Delta z) = \Theta_u / \Delta z$ for $\Delta z \geq \Theta_u$. This relationship in fact defines the *scale* Θ_u, and provides a basis for estimating this parameter of $u(z)$ (figure 2). A useful interpretation of this relationship is that Θ_u is the *elementary distance* that can be used to measure Δz. Other assumptions for the determination of this variance reduction factor are presented in e.g. [10, 22].

3 Probabilistic approach

The point estimate method.

An alternative approach for calculating the statistical moments of the limit state function, denoted by G($\boldsymbol{X}$), where $\boldsymbol{X}$ is the collection of random input variables, is the point estimate method (PEM). The method is essentially a weighted average method similar to numerical integration formulas involving *sampling points* and *weighting parameters*. The method seeks to replace a given continuous probability density function, with a discrete function having the same first three central moments (mean value μ, standard deviation σ and skewness ν). The point estimate method is able to account for up to three moments.

The most common point estimate method was developed by Rosenblueth [17]. In addition to Rosenblueth's method, there are many other PEMs developed by various researchers, including the methods of Evans [6], Zhou and Nowak [24], Harr [9] and of Li [11]. In the present study the point estimate methods by Rosenblueth, Harr and Zhou and Nowak are used to obtain the statistical moments of the bearing capacity factor N_c. A brief description of the methods is given below.

PEM by Rosenblueth: Rosenblueth [17] developed a point estimate method which concentrates the probability density of a continuous random variable $\boldsymbol{X}$ into two estimate points. If $\mathrm{G}(\boldsymbol{X})$ is a function of n basic variables whose skewness is zero but which may be correlated, 2^n points are chosen to include all possible combinations so that the value of each variable is one standard deviation above or below its mean value.

PEM by Harr: In particular the point estimate method by Harr [9] extends Rosenblueth's PEM. Harr proposed an alternative method which starts from the correlation matrix of the data. This matrix is a real symmetric matrix of order n, the number of random variables which can be diagonalized by an orthogonal eigenvector matrix. The correlation matrix can be represented by a hypersphere of radius $\sqrt{n}$ centered at the corresponding expected values of x_n in the standardized coordinate system. The eigenvector starts from the origin of expected values in their respective directions and each eigenvector intersects the sphere at two points. These points of intersections provide the $2n$ point estimates for calculating the statistical moments of $\mathrm{G}(\boldsymbol{X})$.

PEM by Zhou and Nowak: In the approach proposed by Zhou and Nowak [24] predetermined points in the standard normal space are used to compute the statistical parameters of a function of multiple random variables $\boldsymbol{X}$. These points must be transformed in the typically correlated and non standard normal distributed space. The integration of $\mathrm{G}(\boldsymbol{X})$ can be achieved using a non-product formula. Zhou and Nowak provide a set of numerical integration formulas. In this work the $2n^2$+1 formula (ZN III) is used which leads to $2n^2$+1 realizations of $\mathrm{G}(\boldsymbol{X})$.

Stochastic modelling of soil properties.

The finite element code [2] used in the proposed approach to calculate the bearing capacity q_f, require the soil profile to be modelled using homogeneous layers with constant soil properties. For this reason soil properties have to be defined not only for a certain point in space, but also for the entire domain which is used in the calculation process. Due to the fact of spatial averaging of soil properties the coefficient of variation is reduced significantly as described above. In this study, the variance reduction factor Γ by Vanmarcke [22] is used and can be obtained by

$$\Gamma^2 = \left[\frac{\Theta}{L_u}\left(1 - \frac{\Theta}{4L_u}\right)\right]$$

for $\Theta/L_u \leq 2$, where Θ is the autocorrelation length and L_u is the length of the potential failure surface. For μ_φ = 25° the length of the failure surface L_u yields a value of approximately 10.5 m.

4 Stochastic approach

The model of Fenton and Griffiths [7] combines random field theory with an elasto-plastic finite element algorithm in a Monte-Carlo framework (RFEM). The spatially varying and cross-correlated random fields are generated using the so-called Local Average Subdivision (LAS) method which produces local arithmetic averages of the lognormally distributed random field over each element. Thus, each element is assigned a random value of ln c (c is the soil cohesion) as a local average, over the element size, of the continuously varying random field having point statistics. The element values thus correctly reflect the variance reduction due to arithmetic averaging over the element as well as the cross-correlation structure dictated by spatial correlation length, $\Theta_{ln\,c}$. For the correlation structure of the underlying generated fields an exponentially decaying isotropic correlation function is assumed, $\rho(\tau) = \exp(-2\tau / \Theta_{ln\,c})$ where τ is the absolute distance between any two points in the field. A typical deformed finite element mesh at failure is shown in figure 3. Lighter regions in the illustration indicate stronger material and darker regions indicate weaker material, which have triggered quite irregular failure mechanisms.

The soil cohesion, c, is assumed to be lognormally distributed with mean μ_c, standard deviation σ_c, and spatial correlation length $\Theta_{ln\,c}$. For the friction angle, φ, a bounded distribution is selected. For each set of statistical properties given in Table 1 according to [7], Monte-Carlo simulations have been performed, which involves 1000 repetitions of the soil property random fields and the subsequent finite element analysis. A different value for the bearing capacity, and after normalization by the mean cohesion μ_c, a different value for the bearing capacity factor, N_{ci}, is obtained for each of the n Monte-Carlo simulations by $N_{ci} = q_{fi} / \mu_c$, $i = 1,2,...,n$. These values are then analysed statistically leading to an expected value E[N_c], and standard deviation, s[N_c].

5 Fuzzy set approach

Zadeh [23] used the theory of fuzzy sets as a basis for possibility to model uncertainties. Although possibility distributions seem to be similar to probability

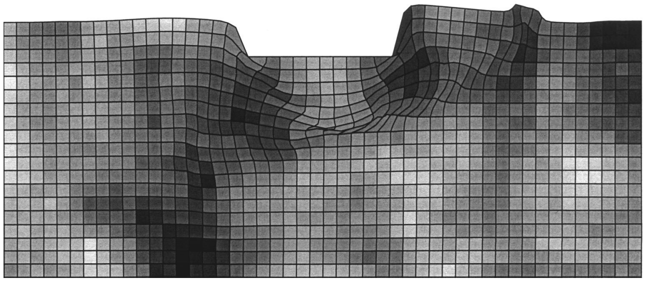


Figure 3: Typical deformed finite element mesh at failure from [7]

distributions, possibility calculus, which is used to derive the membership function of the performance of a system from the membership functions of the uncertain variables, is fundamentally different than probability calculus. The main difference between the axioms of possibility and probability measures is that the possibility of a union of events (disjoint or not) is equal to the maximum of the possibilities of the individual events, whereas the probability of a union of disjoint events is equal to the sum of the probabilities of these events (see e.g. discussion in [4]). Therefore, fuzzy set approach is an alternative to probability.

Fuzzy numbers.

$\mathbb{F}(X)$ denotes the collection of fuzzy subsets of a set X. A fuzzy set $A \in \mathbb{F}(X)$ is characterized by (and can be identified with) its membership function $m_A(x)$, $0 \leq m_A(x) \leq 1$, describing the degree of possibility that the variable A takes the value x of X. The fuzzy sets $[A]^\alpha = x \in X : m_A(x) \geq \alpha$ are the so-called α-level sets of A, i.e. the variable A fluctuates in the range $[A]^\alpha$ with possibility degree α. Given a function $f\colon X \to Y$, the extension principles by Zadeh [23] allows to extend it to a function $f\colon\mathbb{F}(X) \to \mathbb{F}(Y)$ by $m_{f(A)}(y) = \sup\{m_A(x), x \in f^{-1}(y)\}$.

$A \in \mathbb{F}\left(\mathbb{R}^{\mathrm{d}}\right)$ is called a fuzzy vector, if each of its α-level sets is convex and compact ($0 < \alpha < 1$), and $[A]^1$ contains exactly one point. In the case of $d = 1$, A is referred to as a fuzzy number. If $f\colon \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$ is continuous and A a fuzzy vector, the function value $f(A)$ is a fuzzy number, whose level sets are computed by set theoretic evaluation: $[f(A)]^\alpha = f([A]^\alpha)$.

Fuzzification Method .

Dubois and Prade [5] have proposed methods, which are based on judgement and/or on statistical data but there is no commonly accepted procedure for estimating the possibility distribution of a variable. To compare probabilistic and fuzzy set-based methods, we first construct a fuzzy set of an uncertain variable on the basis of a given probability distribution by means of the *least conservative principle* [12]. In this way, we ensure that both models are constructed using the same data. In this paper, the principle is applied to construct a fuzzy set on the basis of a given lognormal probability distribution in such a way that the range between the 5% and the 95%-fractile represents the support (the upper and lower bound value corresponds to α = 0) of the triangular fuzzy number where the ultimate value, the core, respectively is at the modal value, which is the most frequent value (figure 4 and 5). Since the data is based on the lognormal distributions according to section 3.3, it has to be pointed out that autocorrelation is considered already.

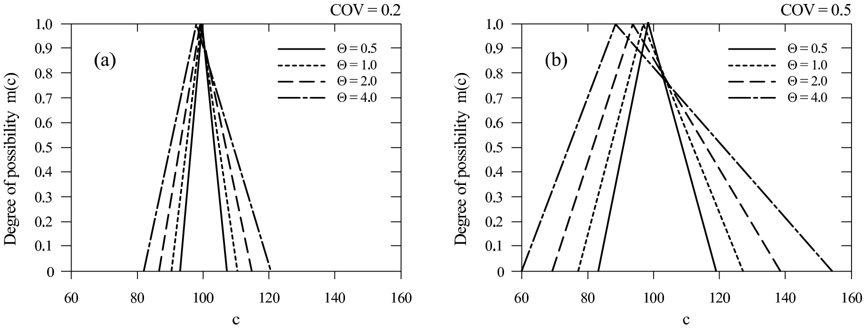


Figure 4: Fuzzy input parameter c, for a) COV of 0.2 and b) COV of 0.5

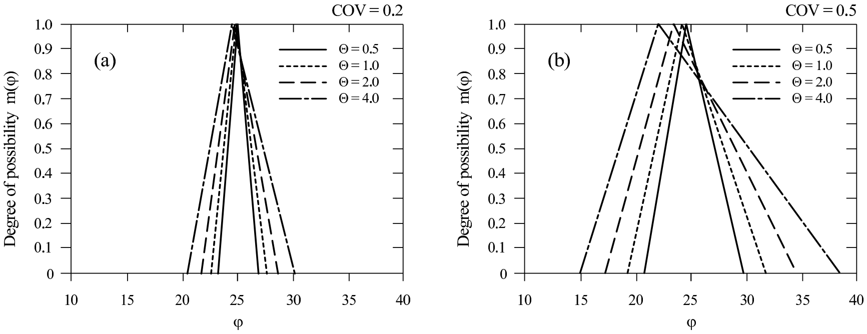


Figure 5: Fuzzy input parameter φ, for a) COV of 0.2 and b) COV of 0.5

Fuzzy finite element analysis.

When the input variables are defined as fuzzy numbers, the computation of the fuzzy response quantity has to be performed. This is achieved by constructing a possibility distribution for the response quantity which is based on the extension principle mentioned above. The principle relates the possibility distribution of fuzzy input variables to the possibility distribution of the fuzzy response function, whereas the α-level concept is used to numerically implement the extension principle. In this approach, the fuzzy function is a finite element model that transforms input fuzzy data to a desired fuzzy output quantity. By replacing the fuzzy numbers in the solution model with intervals, the fuzzy computation reduces to a series of interval analyses, where the minimum and the maximum of the 2^n values define the resulting interval (n is the number of fuzzy input variables). Repeating this process for all selected α-levels, a set of resulting intervals corresponding to the selected α-levels is obtained and define the final output, the response membership function of the dimensionless bearing capacity factor, N_c (figure 6). The higher the number of α-levels under consideration, the greater the accuracy of the possibility distribution of the response. The total number of finite element runs that is involved is $N{\cdot}2^n$, where N is the number of α-levels.

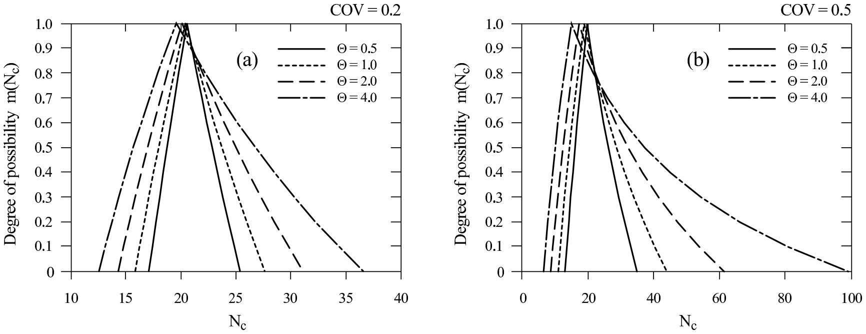


Figure 6: Possibility distribution of the bearing capacity factor, N_c, for a) COV of 0.2 and b) COV of 0.5

Defuzzification Method.

For defuzzification a method based on weighted possibilistic mean and variance of fuzzy numbers is used in this paper. Carlsson and Fuller [3] suggested the notations of weighted possibilistic mean value and variance of fuzzy numbers, which are consistent with the extension principle. Furthermore, they showed that the weighted variance of linear combinations of fuzzy numbers can be computed in a similar manner as in probability theory:

$$\mathrm{E}\left[X^r\right] = \sum_{i=1}^{N} \frac{\alpha_i . x^r_{\alpha_i}}{N}$$

with $x^r_{\alpha_i}$ = 1/2 ($x^r_{\alpha_{i,L}} + x^r_{\alpha_{i,U}}$), where E[X^r] represents the level-weighted r^{th} moment of all α-level sets. α_i denotes the α-level, N the number of α-levels considered and $x^r_{\alpha_i}$ the arithmetic means of all α-level sets, that is, the weight of the arithmetic mean of $x^r_{\alpha_{i,L}}$ and $x_{\alpha_{i,U}}$ is just α.

6 Results and discussion

Figure 7 depicts the influence of Θ and COV$_c$ on the sample coefficient of variation of the estimated bearing capacity factor, COV$_{N_c}$ = s$_{N_c}$/E[N_c] computed by the random field model (RFEM) and by using the probabilistic and the fuzzy set approach.

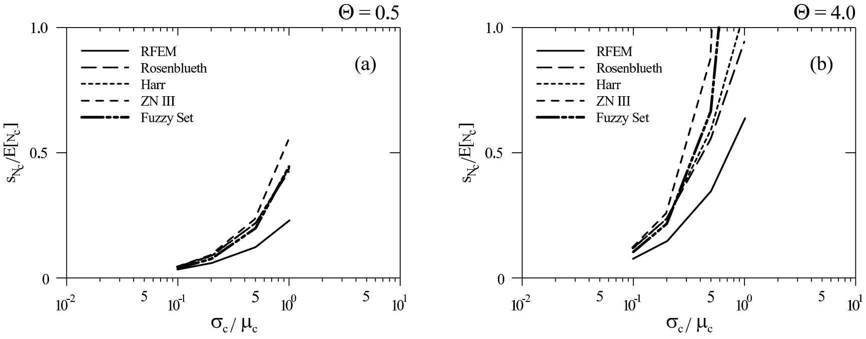


Figure 7: Coefficient of variation of N_c, a) Θ=0.5 and b) Θ=4.0

To have an assessment on the performance of all the approaches, the results from the fuzzy solution are also included in those plots. The figure shows how the bearing capacity factor varies with soil variability, and the spatial correlation length. The plots indicate that COV$_{N_c}$ is positively correlated with both COV$_c$ and Θ, i.e. the variability in E[N_c] increases with the variability in the soil (the higher the spatial correlation length the higher the increase). The results compare well with the COV$_{N_c}$ by the random field method, which represents a more sophisticated method. The PEM methods as well as the fuzzy set method capture the overall behaviour of the analysed ratio and is fairly accurate for moderate magnitudes of the variability in the soil, i.e. COV$<$0.5.

Dubois and Prade [5] have shown that a possibility distribution (fuzzy set A) constructed starting from few statistical data may be used to represent a wide class of probability distributions (compatible with the available information) and to consistently define upper and lower probability distributions, $F_L(x)$ and $F_U(x)$. These bounds may be rewritten in terms of the membership function of the fuzzy set A as $F_L(x)$ = sup$\{m_A(x), x \leq \omega\}$ and $F_U(x)$ = inf$\{$1-$m_A(x), x > \omega\}$, where ω describes the value x with the degree of possibility, $m_A(x)$=1 [8].

Figure 8 shows the possibility and probability of the bearing capacity factor N_c. It can be seen that the possibility is always greater than the probability. Also note that, for this case, the possibility is 1.0 when the probability is 0.5. These results are in line with other studies, e.g. Smith et al. [18] showed that if the fuzzy membership function for a random variable is based on the mean and standard

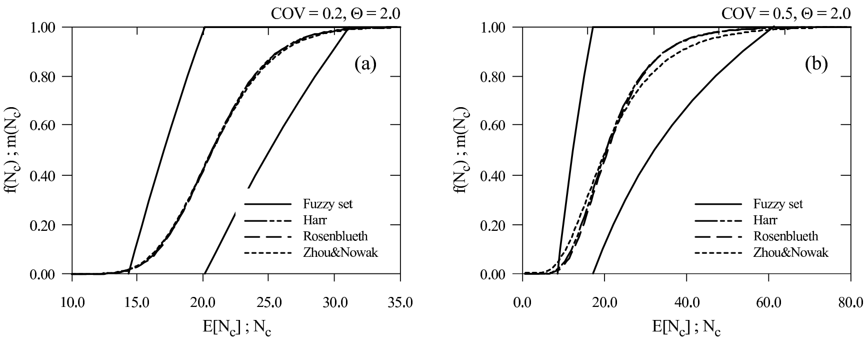


Figure 8: Cumulative distribution functions of E[N_c] assumed as lognormally distributed and membership function of N_c, for a) COV of 0.2 and b) COV of 0.5

deviation of a probabilistic random variable, the possibility of failure is one when the probability of failure is fifty-percent. Therefore, fuzzy set theory may be used to obtain conservative bounds for probability [13].

From a practical point of view, it would be of interest to estimate the probability of *design* failure [7], defined here as occurring when the computed bearing capacity factor, N_c, is less than the deterministic value based on the mean angle of friction divided by a factor of safety F, i.e. 20.7/F (the mean angle of friction φ = μ_φ = 25 degrees, then the deterministic value of N_c yields approximately 20.7).

With the obtained mean value and standard deviation of the performance function based on the PEM assuming a lognormal distribution the probability of design failure (P[N_c<20.7/F]) can be evaluated. For the case where Θ=4.0 figure 9 compares the probability of design failure for two different factors of safety F obtained by probabilistic methods and random field method [14]. The results indicate that the higher the variability (COV) the higher the probability of design failure and show that the proposed method predicts the basic behaviour of relatively simple functions of random variables, but the accuracy is significantly reduced for large coefficients of variation of the input variables.

In order to determine a possibility of design failure the membership functions for the response bearing capacity factor, N_c, are compared with the allowable responses, i.e. 20.7/F as already mentioned. Figure 10 illustrates how the possibility of design failure varies as a function of COV_{N_c} and the ratio of the target value 20.7/F. The fuzzy set method also captures the basic behaviour in terms of the possibility of design failure for the given problem. The outcomes show that the higher the variability (COV) the higher the possibility of design failure. Similar observations can be made about the relations between possibility and probability as described by figure 8, i.e. that the possibility of failure is one when the probability of failure is fifty-percent. However, Stroud et al. [20] reported that even

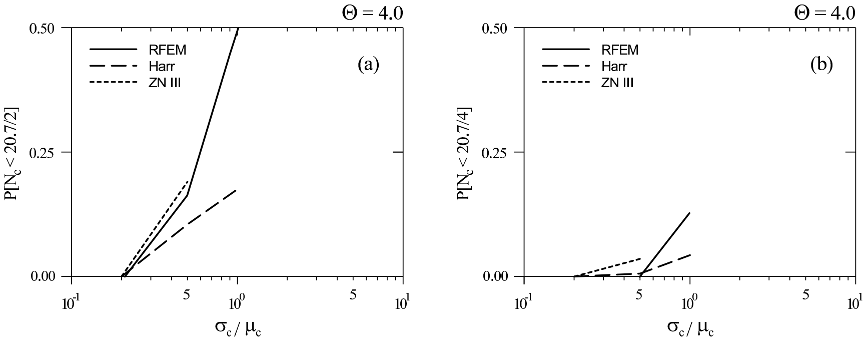


Figure 9: Probabilistic and stochastic approach with Θ=4.0: Influence of factor of safety for a) *F*=2 and b) *F*=4

though the possibility of failure was always greater than the probability of failure for a particular problem with two failure modes, the assumption that possibilistic design is conservative is not a valid assumption when there are many failure modes.

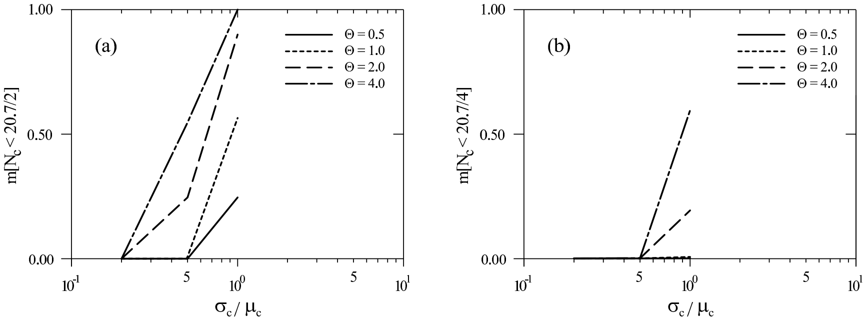


Figure 10: Fuzzy set approach: Influence of factor of safety for a) *F*=2 and b) *F*=4

7 Concluding remarks

The general objective of this paper is to study the differences between probabilistic, stochastic and fuzzy set methods for modelling uncertainties with respect to a simple practical problem for geotechnical engineering. It is argued that the uncertainties associated with material and model parameters are covered in a rational way in the probabilistic and fuzzy set approach. The true probability distributions of the uncertain soil parameters, *c* (cohesion), and, φ (friction angle), are used as the scale to compare the probabilistic and fuzzy set based methods. Generally speaking the outcome of point estimate methods and the fuzzy set method agreed reasonably well with the results obtained by the random field method. An advantage of the fuzzy set approach, from a practical point of view, is the determination of an upper and lower bound to the probability in an efficient way. The results are in line with other studies, even for membership functions as simple as the triangular functions employed here. For the given system and the given data about uncertainties, probabilistic and stochastic analysis yields the probability of failure and fuzzy set analysis yields the possibility of failure, which also varies between zero and one. However, the two measures are not directly comparable, but the results considered were intended to be of easy comprehension and to allow the establishment of a comparison and a correspondence between the methods.

It is acknowledged that the comparisons presented are not rigorous in a mathematical sense and the authors are aware of the discussion on whether the assumptions made in these methods allow a comparison at all. However, from a practical point of view this type of *uncertainty* can be accepted, because it is a significant step forward to be able to account for uncertainties in material parameters using high level numerical methods and keeping the computational effort acceptable. In practice there will always be a trade off between mathematical rigour and practical benefits achievable, which is true in particular in geotechnical engineering. The work presented here should be seen as a step towards a more realistic modelling in geotechnical engineering by demonstrating the applicability of various approaches and should not be seen as a recommendation for one or the other method, at least not at the present stage of developments.

References

[1] Y. Ben-Haim, I. Elishakoff. Convex Models of Uncertainty in Applied Mechanics. Elsevier, Amsterdam, 1990.

[2] R.B.J. Brinkgreve. PLAXIS, Finite element code for soil and rock analyses. *Users manual*, Rotterdam, Balkema, 2002.

[3] C. Carlsson, R. Fuller. On possibilistic mean value and variance of fuzzy numbers. *Fuzzy Sets and Systems*, 122:315-326, 2001.

[4] S. Chen, E. Nikolaidis, H.H. Cudney. Comparison of Probabilistic and Fuzzy Set Methods for Designing under Uncertainty. *American Institute of Aeronautics and Astronautics*, AIAA-99-1579, 1999.

[5] D. Dubois, H. Prade. Possibility Theory. An Approach to Computerized Processing of Uncertainty. Plenum Press, New York, 1988.

[6] D.H. Evans. An application of numerical integration techniques to statistical tolerancing. *Technometrics*, 9:441-456, 1967.

[7] G.A. Fenton, D.V. Griffiths. Bearing capacity of spatially random c - ϕ soils. In *Proc. Int. Conf. on Computer Methods and Advances in Geomechanics*, 1411-1415, Balkema, 2001.

[8] P. Ferrari, M. Savoia. Fuzzy number theory to obtain conservative results with respect to probability. *Comput. Methods Appl. Mech. Engrg.*, 160:205-222, 1998.

[9] M.E. Harr. Probabilistic estimates for multivariate analyses. *Appl. Math. Modelling*, 13(5):313-318, 1989.

[10] S. Lacasse, F. Nadim. Uncertainties in characterising soil properties. NGI-Publication 201, 1997.

[11] K.S. Li. Point-estimate method for calculating statistical moments. *J. Engrg. Mech.*, 118(7):1506-1511, 1992.

[12] E. Nikolaidis, R. Haftka, R. Rosca. Comparison of Probabilistic and Possibility Theory-Based Methods for Designing against Uncertainty. Aerospace and Ocean Engineering Department, Virginia Tech, Blacksburg, VA 24061-0203, 1997.

[13] M. Oberguggenberger, W. Fellin. From probability to fuzzy sets: The struggle for meaning in geotechnical risk assessment. In *Proc. of Int. Conf. on Probabilistics in Geotechnics - Technical and Economic Risk Estimation*, 29-38, 2002.

[14] G.M. Peschl, H.F. Schweiger. Reliability analysis in geotechnics with deterministic finite elements - a comparison of two methods. In *Proc. of 5th European Conference on Numerical Methods in Geotechnical Engineering*, 299-304, 2002.

[15] K.-K. Phoon, F.H. Kulhawy. Characterization of geotechnical variability. *Canadian Geotechnical Journal*, 36:612-624, 1994.

[16] R. Rackwitz. Reviewing probabilistic soils modelling. *Computers and Geotechnics*, 26:199-223, 2000.

[17] E. Rosenblueth. Two-point estimates in probabilities. *Appl. Math. Modelling*, 5(2):329-335, 1981.

[18] S.A. Smith, T. Krishnamurthy, B.H. Mason. Optimized Vertex Method and Hybrid Reliability. *American Institute of Aeronautics and Astronautics*, AIAA-2002-1465, 2002.

[19] V.V. Sokolovski. Statics of Granular Media. Pergamon Press, London, England, 1965.

[20] W.J. Stroud, T. Krishnamurthy, S.A. Smith. Probabilistic and Possibilistic Analyses of the Strength of a Bounded Joint. *American Institute of Aeronautics and Astronautics*, AIAA-2001-1238, 2001.

[21] E.H. Vanmarcke. Probabilistic Modelling of Soil Profiles. *Journal of the Geotechnical Engineering Division, ASCE*, 103:1227-1246, 1977.

[22] E.H. Vanmarcke. Random Fields - Analysis and Syntesis. MIT-Press, Cambridge, Massachusetts, 1983.

[23] L.A. Zadeh. Fuzzy Sets as a Basis for a Theory of Possibility. *Fuzzy Sets and Systems*, 1:3-28, 1978.

[24] J. Zhou, A.S. Nowak. Integration formulas to evaluate functions of random variables. *Structural safety*, 5:267-284, 1988.

Gerd M. Peschl is with the Institute for Soil Mechanics and Foundation Engineering, Graz University of Technology, Graz, Austria. E-mail: peschl@tugraz.at

Helmut F. Schweiger is with the Institute for Soil Mechanics and Foundation Engineering, Graz University of Technology, Graz, Austria. E-mail: helmut.schweiger@tugraz.at

Game-Theoretic Learning Using the Imprecise Dirichlet Model

E. QUAEGHEBEUR
Ghent University, Belgium

G. DE COOMAN
Ghent University, Belgium

Abstract

We discuss two approaches for choosing a strategy in a two-player game. We suppose that the game is played a large number of rounds, which allows the players to use observations of past play to guide them in choosing a strategy.

Central in these approaches is the way the opponent's next strategy is assessed; both a precise and an imprecise Dirichlet model are used. The observations of the opponent's past strategies can then be used to update the model and obtain new assessments. To some extent, the imprecise probability approach allows us to avoid making arbitrary initial assessments.

To be able to choose a strategy, the assessment of the opponent's strategy is combined with rules for selecting an optimal response to it: a so-called best response or a maximin strategy. Together with the updating procedure, this allows us to choose strategies for all the rounds of the game.

The resulting playing sequence can then be analysed to investigate if the strategy choices can converge to equilibria.

Keywords

game theory, fictitious play, equilibria, imprecise Dirichlet model, learning

1 Introduction

In [4] and [5], Fudenberg et al. have proved a number of convergence results concerning methods for learning optimal strategies in a game-theoretic context. They show that these results hold in particular for *fictitious play* in strictly competitive two-player games in strategic form. In this context, a player bases his learning method on the assumption that his opponent uses a fixed, but unknown, mixed strategy. The pure strategies that his opponent actually plays are consequently assumed to be iid observations of the random *multinomial process* that has this mixed strategy as its probability mass function. The player then uses a Bayesian statistical updating scheme, where the prior is chosen from among a class of models that is conjugate with the multinomial likelihood function, namely the Dirichlet priors, mainly because such a choice allows for simple updating rules.

In the present work, we investigate how this learning method is influenced by replacing the Dirichlet priors by so-called imprecise Dirichlet priors, first introduced by Walley [9], and we provide generalisations for Fudenberg's convergence results that can be applied to the new learning method.

The Game. We consider *strictly competitive two-player games in the strategic form*; [3, Chapter 2], [5, Chapter 1]. One player is denoted by i and his opponent by $-i$, where $i \in \{-1,1\}$.

Player i has a finite set $S^i = \{1,\ldots,N^i\}$ of *pure strategies* s^i. After each round of the game, he receives a (possibly negative) *pay-off* $u^i(s^i,s^{-i})$, with $s^i \in S^i$ and $s^{-i} \in S^{-i}$. This pay-off is assumed to be expressed in units of some predetermined linear utility, e.g. probability currency; [7, Sections 13 and 14], [8, Section 2.2.2].

Instead of choosing a pure strategy, player i can also choose a so-called *mixed strategy* σ^i, which is a probability mass function on the set S^i. This amounts to using a randomisation device that chooses a pure strategy from S^i, with the probabilities for each pure strategy defined by the mixed strategy σ^i. These can be written as a vector of length N^i with $\sum_{s^i} \sigma^i(s^i) = 1$. We denote the set of these mixed strategies by Σ^i. In what follows, unless otherwise indicated, s^i will always be an element of S^i and σ^i will always be an element of Σ^i.

When using mixed strategies, only the *expected pay-off* can be calculated,

$$u^i(\sigma^i,\sigma^{-i}) = \sum_{s^i \in S^i} \sum_{s^{-i} \in S^{-i}} u^i(s^i,s^{-i})\sigma^i(s^i)\sigma^{-i}(s^{-i}). \tag{1}$$

It should be clear that pure strategies can be considered as border-case, or degenerate, mixed strategies. The set of all mixed strategies Σ^{-i} can be represented as the unit simplex in $\mathbb{R}^{N^{-i}}$. Pure strategies correspond to the vertices of the simplex. The distance between two strategies is measured using the sup-norm,[1]

$$d(\sigma^{-i},\tau^{-i}) = \sup_{s^{-i} \in S^{-i}} |\sigma^{-i}(s^{-i}) - \tau^{-i}(s^{-i})|.$$

Observe that the convex unit simplex is compact under this norm.

Our Objective. We wish to formulate a procedure that guides the players in their strategy choices in such a way, that, using the information they have at their disposal, their expected pay-off is in some sense optimal.

2 Assessing the Opponent's Strategy

It is essential that the information player i has about the strategy σ^{-i} that his opponent will play, is modelled in a manner that is useful, in light of the objective

[1] This allows for a nice interpretation, but any norm generating the usual topology could be used.

above, for choosing a strategy σ^i in response to σ^{-i}. In this section we describe two uncertainty models for representing such information. The first is a precise probability model, the second is imprecise.

Gambles. The available information about his opponent’s strategy σ^{-i} will lead player i to accept or reject gambles whose outcome depends on σ^{-i}. Both the uncertainty models described later intend to model player i’s behavioural dispositions toward such gambles. A *gamble* X on Σ^{-i} is a bounded real-valued map on Σ^{-i}. It represents an uncertain reward: it yields the amount $X(\sigma^{-i})$ if player $-i$ decides to play the mixed strategy σ^{-i}. The set of all gambles on Σ^{-i} is denoted by $\mathcal{L}(\Sigma^{-i})$; [8, Section 1.5.6]. Two types of gambles are of special interest.

If player i decides to play strategy σ^i, then the game will result in an expected pay-off that still depends on the strategy σ^{-i} that his opponent will play. Thus, we can associate the *strategy gamble* X_{σ^i} on Σ^{-i} with this strategy σ^i by defining $X_{\sigma^i}(\sigma^{-i}) = u^i(\sigma^i, \sigma^{-i})$ for all σ^{-i} in Σ^{-i}. It represents the uncertain expected pay-off for player i if he chooses strategy σ^i. Every gamble in the subset $\mathcal{K}^i = \{X_{\sigma^i} : \sigma^i \in \Sigma^i\}$ of $\mathcal{L}(\Sigma^{-i})$ is thus an *uncertain expected pay-off*. The distance between two strategy gambles is measured using the sup-norm,

$$d(X_{\sigma^i}, X_{\tau^i}) = \sup_{\sigma^{-i} \in \Sigma^{-i}} |X_{\sigma^i}(\sigma^{-i}) - X_{\tau^i}(\sigma^{-i})|.$$

Proposition 1 *The set of strategy gambles $\mathcal{K}^i$ is convex and compact under the* sup-*norm topology on Σ^{-i}.*

Another type of gamble on Σ^{-i}, specifically associated with a pure strategy s^{-i}, is the *evaluation gamble* $Y_{s^{-i}} : \Sigma^{-i} \to [0,1]$ defined by $Y_{s^{-i}}(\sigma^{-i}) = \sigma^{-i}(s^{-i})$. This definition implies that $\sum_{s^{-i}} Y_{s^{-i}} = 1$. Each of these gambles yields the unknown probability mass of the pure strategy s^{-i} defined by (the unknown) probability mass function σ^{-i}. Using this notation, the vector $Y^{-i} = (Y_1, \ldots, Y_{N^{-i}})$ of evaluation gambles returns to the unknown mixed strategy $\sigma^{-i} = Y^{-i}(\sigma^{-i})$ itself.

Using Eq. (1), it is possible to write each strategy gamble as a linear combination of evaluation gambles,

$$X_{\sigma^i} = \sum_{s^{-i} \in S^{-i}} \left(\sum_{s^i \in S^i} u^i(s^i, s^{-i}) \sigma^i(s^i) \right) Y_{s^{-i}}. \tag{2}$$

The Precise Dirichlet Model. First we consider a model that specifies the information available to player i as a *linear prevision* P on some subset of $\mathcal{L}(\Sigma^{-i})$; [2, Chapter 3], [8, Section 2.8]. $P(X)$ is player i’s *fair price*, or *prevision*, for the gamble X, i.e., the unique real number such that he is disposed to buy the gamble X for all prices $p < P(X)$ and to sell X for all prices $p > P(X)$.

If we define $\pi_P = P(Y^{-i}) = (P(Y_1), \ldots, P(Y_{N^{-i}}))$, then the properties of linear previsions allow us to conclude that $\sum_{s^{-i}} \pi_P(s^{-i}) = 1$ and $0 \leq \pi_P(s^{-i}) \leq 1$. We see that π_P is a possible mixed strategy for the opponent. It is player i’s prevision of the strategy that his opponent will play. Using Eq. (2) and the linearity of the operator P, we can write for the prevision of the strategy gamble X_{σ^i}:

$$P(X_{\sigma^i}) = \sum_{s^{-i}\in S^{-i}} \left(\sum_{s^i \in S^i} u^i(s^i, s^{-i})\sigma^i(s^i) \right) \pi_P(s^{-i}) = X_{\sigma^i}(\pi_P), \tag{3}$$

i.e., the expected pay-off if the opponent were actually to play strategy π_P.

The linear prevision P we shall use here is a *precise Dirichlet model* (PDM) $P(\cdot \mid \beta_t, \rho_t)$, where $\beta_t > 0$ and ρ_t is a mixed strategy in the interior $\mathrm{int}(\Sigma^{-i})$ of Σ^{-i}, i.e., $\rho_t(s^{-i}) > 0$ for all $s^{-i} \in S^{-i}$. This PDM is defined for all measurable gambles X on Σ^{-i} by

$$P(X \mid \beta_t, \rho_t) = \frac{1}{B(\beta_t, \rho_t)} \int_{\Sigma^{-i}} X(\sigma^{-i}) f(\sigma^{-i} \mid \beta_t, \rho_t) \mathrm{d}\sigma^{-i}, \tag{4}$$

where f and the normalisation constant B define the parametrised[2] Dirichlet probability density function,

$$f(\sigma^{-i} \mid \beta_t, \rho_t) = \prod_{s^{-i}\in S^{-i}} \sigma^{-i}(s^{-i})^{\beta_t \rho_t(s^{-i}) - 1} \quad \text{and} \quad B(\beta_t, \rho_t) = \frac{\prod_{s^{-i}} \Gamma(\beta_t \rho_t(s^{-i}))}{\Gamma(\beta_t)}.$$

When using such a PDM, the prevision π_P of the strategy his opponent will play coincides with ρ_t:

$$\pi_P = \pi_{P(\cdot|\beta_t,\rho_t)} = P(Y^{-i} \mid \beta_t, \rho_t) = \rho_t.$$

This means that for the calculation of $P(X_{\sigma^i} \mid \beta_t, \rho_t)$ we don't need to use Eq. (4), but that we can use Eq. (3), replacing π_P by ρ_t:

$$P(X_{\sigma^i} \mid \beta_t, \rho_t) = \sum_{s^{-i}\in S^{-i}} \left(\sum_{s^i \in S^i} u^i(s^i, s^{-i})\sigma^i(s^i) \right) \rho_t(s^{-i}) = X_{\sigma^i}(\rho_t).$$

The Imprecise Dirichlet Model. Next, we consider an imprecise probability model for the information player i has about his opponent's strategy. This can always be made to take the form of a coherent *lower prevision* $\underline{P}$ on some subset of $\mathcal{L}(\Sigma^{-i})$; [8, Section 2.3]. $\underline{P}(X)$ specifies player i's supremum acceptable price for buying the gamble X, i.e., it is the greatest real number p such that he is disposed to buying the gamble X for all prices strictly smaller than p.

The lower prevision $\underline{P}$ we shall use here is an *imprecise Dirichlet model* (IDM) $\underline{P}(\cdot \mid \beta_t, M_t)$, where $\beta_t > 0$ and $M_t \subseteq \mathrm{int}(\Sigma^{-i})$; [9]. This IDM is defined for all measurable gambles X on Σ^{-i} as the lower envelope of a set of PDM's (with a common β_t, but each with their own ρ_t),

$$\underline{P}(X \mid \beta_t, M_t) = \inf\{P(X \mid \beta_t, \rho_t) \colon \rho_t \in M_t \subset \Sigma^{-i}\}. \tag{5}$$

[2] We use a non-standard parametrisation, because it is more convenient in this context; [9].

3 Choosing an Optimal Strategy

When choosing an optimal strategy, it is important to be clear on what defines optimality. In this game-theoretic context, it is desirable to attain a pay-off that is as high as possible, but on the other hand it may also be important to limit possible losses. These are the guiding criteria in our search for optimal strategies [6, Section 3.8].

Admissible Strategies, Maximin Strategies, Best Replies. If for two strategies τ^i and σ^i, the pay-off for τ^i is always at least as high as that for σ^i, i.e., $X_{\tau^i} \geq X_{\sigma^i}$ or in other words $(\forall \sigma^{-i} \in \Sigma^{-i})(X_{\tau^i}(\sigma^{-i}) \geq X_{\sigma^i}(\sigma^{-i}))$, we say that τ^i *dominates* σ^i—or that X_{τ^i} dominates X_{σ^i}; [5, Section 1.7.2].

A strategy $\sigma^i \in \Sigma^i$, or its corresponding strategy gamble $X_{\sigma^i} \in \mathcal{K}^i$, is called *inadmissible* if there is another strategy τ^i that strictly dominates it: $X_{\tau^i} \geq X_{\sigma^i}$ and $X_{\sigma^i} \neq X_{\tau^i}$. Otherwise, it is called *admissible*. We consider an admissible strategy to be more optimal than an inadmissible strategy. However, the discussion of, and the results deduced for, the learning models below is not essentially affected when this distinction is not made.

Now suppose that player i knows that his opponent will play some strategy in $M \subseteq \Sigma^{-i}$, but nothing more. When playing σ^i, his expected pay-off will at least be $\inf_{\sigma^{-i} \in M} X_{\sigma^i}(\sigma^{-i})$. An *M-maximin strategy* τ^i maximises this minimal pay-off:

$$\tau^i \in \operatorname*{argmax}_{\sigma^i \in \Sigma^i} \inf_{\sigma^{-i} \in M} X_{\sigma^i}(\sigma^{-i}).$$

Proposition 2 *There are admissible M-maximin strategies for any compact subset M of* Σ^{-i}.

When $M = \Sigma^{-i}$, player i doesn’t have a clue about his opponent’s strategy choice, and the corresponding Σ^{-i}-maximin strategy is simply called a *maximin strategy*.

Corollary 1 *There are always admissible maximin strategies.*

At the other extreme, player i knows his opponent will play a strategy σ^{-i}. Any corresponding $\{\sigma^{-i}\}$-maximin strategy is called a *best reply* to σ^{-i}. The set of all best replies to σ^{-i} is denoted by $BR^i(\sigma^{-i})$.

Corollary 2 *There are always admissible best replies to any strategy* σ^{-i} *in* Σ^{-i}.

This set of best replies has some interesting properties.

Proposition 3 *For all* σ^{-i} *in* Σ^{-i}, $BR^i(\sigma^{-i})$ *is a compact and convex subset of* Σ^i. *Moreover, if* $\sigma^i \in BR^i(\sigma^{-i})$ *and* $\sigma^i(s^i) > 0$ *for some* $s^i \in S^i$, *then* $s^i \in BR^i(\sigma^{-i})$.

For $M \subseteq \Sigma^{-i}$, the collection of best replies to strategies in M is denoted by $BR^i(M)$ and given by

$$BR^i(M) = \bigcup_{\sigma^{-i} \in M} BR^i(\sigma^{-i}).$$

Proposition 4 *For any subset M of Σ^{-i} that is convex and closed, the M-maximin strategies make up a subset of $BR^i(M)$.*

Corollary 3 *There are always admissible best replies to any convex and closed subset M of Σ^{-i}.*

Optimal Strategies and the PDM. When using a linear prevision P, any admissible strategy σ^i that maximises $P(X_{\sigma^i})$ is called a *Bayes strategy*. This name refers to the fact that it is an optimal strategy in the usual Bayesian sense of maximising expected utility; [8, Section 3.9].

Eq. (3) tells us that $P(X_{\sigma^i}) = X_{\sigma^i}(\pi_P)$. This means that τ^i is a Bayes strategy whenever $\tau^i \in \operatorname{argmax}_{\sigma^i} X_{\sigma^i}(\pi_P)$. This gives the following result.

Proposition 5 *The set of the Bayes strategies corresponding to a linear prevision P is given by the admissible strategies of $BR^i(\pi_P)$.*

If player i's model for his opponent's strategy is a PDM $P(\cdot \mid \beta_t, \rho_t)$, we find that his optimal (Bayes) strategies are simply the admissible strategies of $BR^i(\rho_t)$.

Optimal Strategies and the IDM. When using a coherent lower prevision $\underline{P}$, a *maximal strategy* is any admissible strategy σ^i for which $\min_{\tau^i \in \Sigma^i} \overline{P}(X_{\sigma^i} - X_{\tau^i}) \geq 0$; see [8, Section 3.9] for motivation.[3]

We shall use the notation $\mathcal{M}(\underline{P})$ for the set of linear previsions P that dominate $\underline{P}$ on its domain.

Proposition 6 *A strategy σ^i is maximal under $\underline{P}$*

- $\Leftrightarrow$ *σ^i is a Bayes strategy under some P in $\mathcal{M}(\underline{P})$;*
- $\Leftrightarrow$ *σ^i is an admissible best reply to π_P for some $P \in \mathcal{M}(\underline{P})$, i.e., the admissible $\sigma^i \in BR^i(M_{\underline{P}})$, where $M_{\underline{P}} = \{\pi_P \colon P \in \mathcal{M}(\underline{P})\} \subseteq \Sigma^i$.*

Corollary 4 *There are maximal strategies under $\underline{P}$.*

There is another optimality criterion associated with a lower prevision $\underline{P}$: an admissible mixed strategy σ^i is called *$\underline{P}$-maximin* if it maximises the lower prevision $\underline{P}(X_{\tau^i})$ of all strategy gambles X_{τ^i}, i.e., if $\sigma^i \in \operatorname{argmax}_{\tau^i \in \Sigma^i} \underline{P}(X_{\tau^i})$; [8, Section 3.9]. Since a coherent lower prevision $\underline{P}$ is the lower envelope of its set of dominating linear previsions (see [8, Theorem 3.3.3]), we see that

$$\underline{P}(X_{\tau^i}) = \min_{P \in \mathcal{M}(\underline{P})} P(X_{\tau^i}) = \min_{\sigma^{-i} \in M_{\underline{P}}} X_{\tau^i}(\sigma^{-i}),$$

[3] To see that this definition generalises that of a Bayes strategy, consider that

$$\sigma^i \in \operatorname*{argmax}_{\tau^i \in \Sigma^i} P(X_{\tau^i}) \Leftrightarrow P(X_{\sigma^i}) \geq \max_{\tau^i \in \Sigma^i} P(X_{\tau^i}) \Leftrightarrow \min_{\tau^i \in \Sigma^i} P(X_{\sigma^i} - X_{\tau^i}) \geq 0.$$

and consequently, the admissible mixed strategy σ^i is $\underline{P}$-maximin if and only if $\sigma^i \in \operatorname{argmax}_{\tau^i \in \Sigma^i} \min_{\sigma^{-i} \in M_{\underline{P}}} X_{\tau^i}(\sigma^{-i})$, i.e., if it is $M_{\underline{P}}$-maximin. We know from Section 3 that all the $M_{\underline{P}}$-maximin strategies also belong to $BR^i(M_{\underline{P}})$.

Corollary 5 *For any coherent lower prevision $\underline{P}$, there are $\underline{P}$-maximin strategies. They coincide with the admissible $M_{\underline{P}}$-maximin strategies, and are in particular also maximal strategies under $\underline{P}$.*

If player i models his uncertainty about his opponent’s strategy by an IDM $\underline{P}(\cdot \mid \beta_t, M_t)$, we have proved the following results, using the continuity of Y^{-i} and the properties of $\mathcal{M}(\underline{P}(\cdot \mid \beta_t, M_t))$.

Theorem 1 *If M_t is a subset of* $\operatorname{int}(\Sigma^{-i})$*, then the set $M_{\underline{P}(\cdot|\beta_t,M_t)}$ is the closed convex hull $\overline{\operatorname{co}}(M_t)$ of M_t.*

We thus find that the optimal strategies in this imprecise model are the admissible elements of $BR^i(\overline{\operatorname{co}}(M_t))$. Moreover, if player i wants to play it safe (maximise his minimal expected gains), he can use admissible $\overline{\operatorname{co}}(M_t)$-maximin strategies.

4 Playing the Game Over and Over Again

We now turn our attention to how the proposed models, the PDM and the IDM, can be used when a number of rounds of the game are played. We specifically look at the way observations of past play can change the assessments of a player and we formulate an algorithm to guide the players in their strategy choices.

Learning from Past Play. After playing t rounds of the game, player i has observed a so-called *history* $\zeta_t^{-i} \in \mathcal{Z}_t^{-i} = (S^{-i})^t$ of the pure strategies $\zeta_t^{-i}(k)$, $k = 1, \ldots, t$, that his opponent has played.

If player i supposes that his opponent plays a fixed mixed strategy σ^{-i},[4] which is of course not necessarily the case, the order of the strategies in the history does not matter and the observed strategies can be considered as outcomes of a multinomial iid process. As a sufficient statistic for σ^{-i} he can then use the N^{-i}-tuple n^{-i} of *observed occurrences* for which each component $n^{-i}(s^{-i})$ is the number of times his opponent has played $s^{-i} \in S^{-i}$, and which is consequently a random variable with the multinomial distribution. The total number t of rounds played is also equal to $\sum_{s^{-i}} n^{-i}(s^{-i})$. The N^{-i}-tuple of *observed frequencies* $\frac{n^{-i}}{t}$ is denoted by κ_t^{-i} and can be considered to be an element of Σ^{-i}.

The likelihood function for n^{-i} is

$$L_{n^{-i}}(\sigma^{-i}) = \frac{t!}{\prod_{s^{-i}} n^{-i}(s^{-i})!} \prod_{s^{-i} \in S^{-i}} \sigma^{-i}(s^{-i})^{n^{-i}(s^{-i})}.$$

[4] This corresponds to the underlying assumption used in so-called *fictitious play*; [5, Chapter 2].

Using Bayes' rule, we can now update (see e.g. [5, Chapter 2]) a prior Dirichlet density function $f(\sigma^{-i} \mid \beta_0, \rho_0)$ with the observations n^{-i},

$$\begin{aligned} f(\sigma^{-i} \mid \beta_0, \rho_0, n^{-i}) &= \frac{1}{P(L_{n^{-i}} \mid \beta_0, \rho_0)} f(\sigma^{-i} \mid \beta_0, \rho_0) L_{n^{-i}}(\sigma^{-i}) \\ &= f(\sigma^{-i} \mid \beta_0 + t, \frac{\beta_0\rho_0 + n^{-i}}{\beta_0 + t}) \\ &= f(\sigma^{-i} \mid \beta_t, \rho_t). \end{aligned}$$

We see that the posterior density function $f(\sigma^{-i} \mid \beta_t, \rho_t)$ is still a Dirichlet density function. This means that that the Dirichlet density functions constitute a *conjugate* family of density functions for the multinomial sampling likelihood function $L_{n^{-i}}$. Observe that $P(L_{n^{-i}} \mid \beta_0, \rho_0)$ has to be non-zero, which is guaranteed by $\beta_0 > 0$ and $\rho_0 \in \operatorname{int}(\Sigma^{-i})$.

Updating a Dirichlet model. When updating a prior PDM $P(\cdot \mid \beta_0, \rho_0)$ after t rounds, we find that we simply have to update the parameters,

$$\beta_0 \to \beta_t = \beta_0 + t \quad \text{and} \quad \rho_0 \to \rho_t = \frac{\beta_0\rho_0 + n^{-i}}{\beta_0 + t}, \tag{6}$$

to obtain the posterior PDM $P(\cdot \mid \beta_t, \rho_t)$. It is clear that first updating with n^{-i} and then updating the new model with m^{-i} is equivalent to updating the original model with $n^{-i} + m^{-i}$.

When updating a prior IDM $\underline{P}(\cdot \mid \beta_0, M_0)$ after t rounds, the answer is a bit more complicated. It is possible that there are n^{-i} for which $\underline{P}(L_{n^{-i}} \mid \beta_0, M_0) = 0$ even with $M_0 \subseteq \operatorname{int}(\Sigma^{-i})$, i.e., for $\underline{P}(L_{n^{-i}} \mid \beta_0, M_0) > 0$ we need $P(L_{n^{-i}} \mid \beta_0, \rho_0) > 0$ for all $\rho_0 \in \overline{\text{co}}(M_0)$. However, using the notion of *regular extension*, we can find a unique posterior IDM that is coherent with $\underline{P}(\cdot \mid \beta_0, M_0)$ and that satisfies the additional rationality axiom of *regularity*; [8, Appendix J]. This posterior lower prevision turns out to be the lower envelope of the updated PDM's,

$$\inf_{\rho_0 \in M_0} P(X \mid \beta_0 + t, \frac{\beta_0\rho_0 + n^{-i}}{\beta_0 + t}) = \inf_{\rho_t \in M_t} P(X \mid \beta_t, \rho_t) = \underline{P}(X \mid \beta_t, M_t),$$

where β_t and M_t are the parameters of the updated *IDM*,

$$\beta_0 \to \beta_t = \beta_0 + t \quad \text{and} \quad M_0 \to M_t = \{\frac{\beta_0\rho_0 + n^{-i}}{\beta_0 + t} : \rho_0 \in M_0\}. \tag{7}$$

Iterative Playing Algorithm: Assess, Decide and Update. Our generic guiding algorithm for player *i* playing multiple rounds of a strictly competitive two-player game consists of three steps; [4, Section 3]. Assume that t rounds have already been played, and that the history ζ_t^{-i} of the pure strategies played by the opponent during these rounds is available to player *i*. He is about to play a new round and uses some model to describe the information he has.

1. Player i has to make an assessment $\mu^i(\zeta_t^{-i})$ about the data that are relevant for his strategy choice: to this end, he uses an *assessment rule* μ^i.

2. Player i has to use a *decision rule* ϕ^i to choose a strategy $\phi^i(\zeta_t^{-i})$ to play, using his assessments $\mu^i(\zeta_t^{-i})$.

3. After the round is played, player i should use the observation of his opponent’s strategy to *update* his information.

Let us now see what this algorithm becomes for the two types of uncertainty models described above.

When using a PDM $P(\cdot \mid \beta_t, \rho_t)$, we can formulate the following implementation of the algorithm.

1. Let $\mu^i(\zeta_t^{-i}) = \rho_t = \pi_{P(\cdot|\beta_t,\rho_t)}$, the prevision of the opponent’s strategy.

2. Let $\phi^i(\zeta_t^{-i})$ be some (admissible) element of $BR^i(\rho_t)$.

3. Update the PDM to $P(\cdot \mid \beta_{t+1}, \rho_{t+1})$ using Eq. (6).

Initially, player i has to choose a ρ_0 and β_0. The parameter β_0 can be interpreted as the number of *pseudocounts*[5] associated with the initial prevision of his opponent’s strategy ρ_0, for which any choice is arbitrary (if it is not based on some information).

When using an IDM $\underline{P}(\cdot \mid \beta_t, M_t)$, we can formulate two different implementations of the algorithm, different only in their choice of behaviour rule.

1. Let $\mu^i(\zeta_t^{-i}) = \overline{\text{co}}(M_t) = M_{\underline{P}(\cdot|\beta_t,M_t)}$.

2. (a) If we consider maximality as the optimality criterion, then let $\phi^i(\zeta_t^{-i})$ be some (admissible) element of $BR^i(\overline{\text{co}}(M_t))$.

   (b) If we consider maximinity as the optimality criterion, then let $\phi^i(\zeta_t^{-i})$ be some (admissible) $\overline{\text{co}}(M_t)$-maximin strategy.

3. Update the IDM to $\underline{P}(\cdot \mid \beta_{t+1}, M_{t+1})$ using Eq. (7).

Initially, player i has to choose an M_0 and a number of pseudocounts β_0. When he has no information available, an obvious choice for M_0 is $\text{int}(\Sigma^{-i})$, which corresponds to so-called near-ignorance [8, Section 4.6.9]. The choice for the best reply behaviour rule or the maximin behaviour rule will not influence the results of Section 5 in any way.

[5] In the literature, the values 1 and 2 are found for prior models that are not based on any information; [9].

5 Equilibria and Convergence

Now that we have two learning models, the PDM and the IDM, at our disposal, we can investigate the game-play that results from using them. We start by giving some definitions that are essential for the ensuing analysis.

Strategy Profiles and Equilibria. To be able to analyse the game-play that results from the assessment and behaviour rules discussed in Section 4, we introduce some new notation and recall the concept of an equilibrium.

A couple of strategies of the players is called a *strategy profile*, which can be pure $s = (s^i, s^{-i}) \in S = S^i \times S^{-i}$, or mixed $\sigma = (\sigma^i, \sigma^{-i}) \in \Sigma = \Sigma^i \times \Sigma^{-i}$. A corresponding *profile history* after t rounds of play is denoted by $\zeta_t \in \mathcal{Z}_t = S^t$.

The notation $\sigma(s)$ corresponds to $(\sigma^i(s^i), \sigma^{-i}(s^{-i}))$. Likewise, we write

$$\begin{aligned} BR(\sigma) &= BR^i(\sigma^{-i}) \times BR^{-i}(\sigma^i) \subseteq \Sigma, \\ \mu(\zeta_t) &= \mu^i(\zeta_t^{-i}) \times \mu^{-i}(\zeta_t^i) \subseteq \Sigma, \\ \phi(\zeta_t) &= (\phi^i(\zeta_t^{-i}), \phi^{-i}(\zeta_t^i)) \in \Sigma. \end{aligned}$$

An *equilibrium* is a strategy profile for which the pay-off for both players cannot be increased if one of them changes his strategy, while his opponent's strategy remains unchanged; [3]. This means that

$$\sigma_* \text{ is an equilibrium } \Leftrightarrow (\forall i)\left(u^i(\sigma_*) = \max_{\tau^i \in \Sigma^i} u^i(\tau^i, \sigma_*^{-i})\right) \Leftrightarrow \sigma_* \in BR(\sigma_*).$$

If $s_* = BR(s_*)$, then s_* is a *strict equilibrium*.[6] A game can have multiple (strict) equilibria.[7]

Assessment Rules. The definitions in this section and in the next are generalisations of the definitions given by Fudenberg and Kreps in [4] to learning models with assessments $\mu^i(\zeta_t^{-i})$ that are set-valued rather than point-valued.

An important characterisation of possible assessment rules can be made by looking at what the influence is of different parts of a history.

We say that a assessment rule μ^i is *adaptive* if it attaches diminishing importance to earlier parts of the history, as the number of rounds t increases. This means that for all t and all $\varepsilon > 0$,

$$(\exists T > t)\big(\forall t' > T - t\big)\big(\forall \zeta_{t+t'}^{-i} \in \mathcal{Z}_{t+t'}^{-i}\big)\big(\forall \sigma^i \in \mu^i(\zeta_{t+t'}^{-i})\big)\big(\sigma^i(s^i) < \varepsilon\big),$$

for every pure strategy s^i that was not played in the last t' rounds (did not appear in the t' last components of $\zeta_{t+t'}^{-i}$).

A specific subcategory of the adaptive assessment rules can be defined using the observed frequencies κ_t^{-i} of strategies played by the opponent. An assessment

[6] By Proposition 3, only pure strategy profiles can be strict equilibria.

[7] When only admissible strategies are considered optimal, some equilibria might not be playable. There is always at least one admissible equilibrium.

rule μ^i is called *asymptotically empirical* if for every *infinite history* $\zeta_\infty^{-i} \in \mathcal{Z}_\infty^{-i}$ it holds that $\lim_{t\to\infty}\sup_{\sigma^{-i}\in\mu^i(\zeta_t^{-i})} d(\sigma^{-i},\kappa_t^{-i}) = 0$, where the ζ_t^{-i} are partial histories of the selected infinite history ζ_∞^{-i}.

Using the updating formulae (6) and (7), we obtain the following result.

Theorem 2 *The assessment rules of the PDM and the IDM are asymptotically empirical, and thus adaptive.*

Behaviour Rules. It is clear that the behaviour rules ϕ determine which histories are possible. A history is called *compatible* with the behaviour rules ϕ used by the players if it can be generated (with non-zero probability) by these behaviour rules. Explicitly, this means that for every pure profile $\zeta_t(k)$, $k = 1,\dots,t$, that is a component of a compatible profile history, both components of $\phi(\zeta_{k-1})(\zeta_t(k))$ are strictly positive, so the randomisation devices used by the players can select the pure strategies $\zeta_t^i(k)$ and $\zeta_t^{-i}(k)$ with non-zero probability.

It is useful to know to what degree the behaviour rules ϕ used by the players succeed in attaining the objective of optimality (see Section 1). The characterisations in this section do just this, and give a clear interpretation of this objective, keeping in mind that the players suppose that their opponent plays an unknown, but fixed, mixed strategy.

We call a behaviour rule ϕ^i *set-myopic* relative to the assessment rule μ^i if, for all t and histories ζ_t^{-i}, it holds that $\phi^i(\zeta_t^{-i}) \in BR^i(\mu^i(\zeta_t^{-i}))$. When the assessments $\mu^i(\zeta_t^{-i})$ are point-valued, the prefix ‘set’ in set-myopic is dropped.

We now define a weakening of the notion of a set-myopic behaviour rule. We call a behaviour rule ϕ^i *strongly asymptotically set-myopic* relative to the assessment rule μ^i if, for some sequence $\varepsilon_t > 0$ with $\lim_{t\to\infty}\varepsilon_t = 0$ and for all t and histories ζ_t^{-i}, it holds that

$$\left(\forall\sigma^{-i}\in\mu^i(\zeta_t^{-i})\right)\left(\forall\hat{s}^i\in S^i \text{ such that } \phi^i(\zeta_t^{-i})(\hat{s}^i) > 0\right)$$
$$\left(u^i(\hat{s}^i,\sigma^{-i}) + \varepsilon_t \geq \max_{s^i\in S^i} u^i(s^i,\sigma^{-i})\right).$$

Using the definitions from Section 4 the next result is immediate.

Theorem 3 *The behaviour rules for the IDM are set-myopic and the behaviour rule for the PDM is myopic.*

Convergence to equilibria. An interesting theorem about strict equilibria follows directly from the definitions of a strict equilibrium and of a myopic behaviour rule; [4].

Theorem 4 (absorption to a strict equilibrium) *If there is a strict equilibrium s_* that is played in some round t of a profile history ζ_t compatible with a myopic behaviour rule ϕ, then s_* will be played during all subsequent rounds $t' > t$.*

This theorem holds for the PDM (with myopic behaviour rules), due to Theorem 3, but not for the IDM (with set-myopic behaviour rules), because we have been able to show that the selected mixed strategy $\phi^i(\zeta_t^{-i})$ under both optimisation criteria can still be different from s_*^i due to the fact that $\mu^i(\zeta_t^{-i}) = \overline{\text{co}}(M_t)$ is a set. It is possible to tighten the conditions, to obtain a result that also works for the IDM, i.e., to make sure that best reply only contains s_*.

Theorem 5 (conditional absorption to a strict equilibrium) *If, for some profile history ζ_t compatible with set-myopic behaviour rules ϕ, the strategy profile $\phi(\zeta_t)$ cannot be different from the strict equilibrium s_*, then s_* will be played during all subsequent rounds $t' > t$.*

For equilibria s_* of pure strategies that aren't necessarily strict, the following result is found.

Theorem 6 (repeated play of a pure strategy profile) *Consider an infinite history ζ_∞ in $\mathcal{Z}_\infty$ such that for some t, a pure strategy profile s_* is played in all subsequent rounds. If ζ_∞ is compatible with behaviour rules ϕ that are strongly asymptotically set-myopic relative to the adaptive assessment rules μ, then s_* is an equilibrium.*

This theorem can be used for both the PDM and the IDM, due to Theorems 2 and 3, even if both players don't use the same model. For example, one player can use the IDM and his opponent the PDM, or two players can use the IDM, each using a different optimality criterion.

For mixed equilibria σ_*, the following result about the convergence of the observed game-play to a mixed equilibrium, is found.

Theorem 7 (repeated play of a mixed strategy profile) *Let the infinite history ζ_∞ in $\mathcal{Z}_\infty$ be such that for some mixed strategy profile σ_*, it holds that for both players $i \in \{-1, 1\}$*

$$\lim_{t\to\infty} \kappa_t^{-i} = \sigma_*^{-i}.$$

If the infinite history ζ_∞ is compatible with behaviour rules ϕ that are strongly asymptotically set-myopic relative to the assessment rules μ that are asymptotically empirical, then σ_* is an equilibrium.

As before, due to Theorems 2 and 3, this theorem can be used for both the PDM and the IDM.

Theorems 6 and 7 can only say that convergence has occurred, but do not indicate when convergence will occur. They could be useful for finding equilibria in large games. As these theorems are generalisations to set-valued assessment rules of theorems found in [4], their proofs are (not always trivial) modifications of the ones found there.

6 Conclusions

General Remarks. Both the learning models discussed above accomplish our objective of optimality of the expected pay-off quite well. Their convergence properties also favour their use in game theory, notably in the search for equilibria.

The PDM has already been studied in the literature and the learning model based on it is often called *fictitious play* in a game-theoretic setting. The IDM has also been used in different contexts; see [9] for the presentation of the IDM itself and [1], [10], [11] and [12] for examples of possible applications in other areas.

In Section 5 we have in fact generalised the results of Fudenberg and Kreps in [4, Sections 3 and 4], where point-valued assessments $\mu^i(\zeta_t^{-i})$ are used, to set-valued assessments. This is why we formulated Theorems 6 and 7 for a broader class of learning models than our Dirichlet models, which allows Section 5 to be seen as a generalisation of [4, Sections 3 and 4], and not only as a group of results for the PDM and IDM.

We haven’t discussed the choice of a specific strategy $\phi^i(\zeta_t^{-i})$ from among the optimal ones. But, if for a specific application other, additional, criteria are available, then using them at this stage will not influence the convergence results in any way.

PDM vs. IDM. If we compare the PDM to the IDM, the first thing to be said is that the PDM is a special case of the IDM, where $M_0 = \{\rho_0\}$. This immediately indicates the most important advantage of the IDM over the PDM, the possibility of not having to make an arbitrary initial choice, as there is no need to choose one specific prior.

The second advantage of the learning model using the IDM is that it reflects, in its assessment $\mu^i(\zeta_t^{-i})$, the amount of information on which it is based. This corresponds to the fact that the distances between elements of M_t shrink with increasing t. So the model becomes more precise as more observations come in, in the sense that all elements of M_t will lie closer and closer to the ρ_t of any PDM that could have been used.[8]

One disadvantage of the IDM is that it is a more complex model (the player has to work with sets of strategies instead of one strategy). This difference could be reflected by the calculation load for both models.

Acknowledgements

This paper presents research results of the Belgian Programme on Interuniversity Poles of Attraction initiated by the Belgian state, Prime Minister’s Office for Science, Technology and Culture. It was also financed by a Ph.D. grant of the Insti-

[8] It is interesting to note that the successive M_t are similar to one another, because the updating procedure of Eq. (7) corresponds to a contraction and a translation of M_t on the simplex Σ^{-i}.

tute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). The scientific responsibility rests with the authors.

References

[1] BERNARD, J.-M. Non-parametric inference about an unknown mean using the imprecise dirichlet model. In *ISIPTA '01 – Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, G. de Cooman, T. L. Fine, and T. Seidenfeld, Eds. Shaker Publishing, Maastricht, 2000, pp. 40–50.

[2] DE FINETTI, B. *Theory of Probability*, vol. 1. John Wiley & Sons, Chichester, 1974. English Translation of *Teoria delle Probabilità*.

[3] FRIEDMAN, J. W. *Game Theory with Applications to Economics*. Oxford University Press, New York, 1989.

[4] FUDENBERG, D., AND KREPS, D. M. Learning mixed equilibria. *Games and Economic Behaviour 5* (1993), 320–367.

[5] FUDENBERG, D., AND LEVINE, D. K. *The Theory of Learning in Games*, vol. 2 of *The MIT Press Series on Economic Learning and Social Evolution*. The MIT Press, Cambridge, Massachusets and London, England, 1998.

[6] MYERSON, R. B. *Game Theory: Analysis of Conflict*. Harvard University Press, Cambridge, Massachusetts, 1991.

[7] SMITH, C. A. B. Consistency in statistical inference and decision. *Journal of the Royal Statistical Society, Series A 23* (1961), 1–37.

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

[9] WALLEY, P. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B 58* (1996), 3–57. With discussion.

[10] WALLEY, P., AND BERNARD, J.-M. Imprecise probabilistic prediction for categorical data. Tech. Rep. CAF-9901, Laboratoire Cognition et Activités Finalisées, Université de Paris 8, Paris, January 1999.

[11] ZAFFALON, M. Statistical inference of the naive credal classifier. In *ISIPTA '01 – Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, G. de Cooman, T. L. Fine, and T. Seidenfeld, Eds. Shaker Publishing, Maastricht, 2000, pp. 384–393.

[12] ZAFFALON, M. The naive credal classifier. *Journal of Statistical Planning and Inference 105* (2002), 5–21.

Erik Quaeghebeur is a member of the SYSTeMS research group at Ghent University.
address: Technologiepark – Zwijnaarde 914, 9052 Zwijnaarde, Belgium.
e-mail: erik.quaeghebeur@ugent.be

Gert de Cooman is a member of the SYSTeMS research group at Ghent University.
address: Technologiepark – Zwijnaarde 914, 9052 Zwijnaarde, Belgium.
e-mail: gert.decooman@ugent.be

A Sensitivity Analysis for the Pricing of European Call Options in a Binary Tree Model[*]

H. REYNAERTS
Ghent University, Belgium

M. VANMAELE
Ghent University, Belgium

Abstract

The European call option prices have well-known formulae in the Cox-Ross-Rubinstein model [2], depending on the volatility of the underlying asset. Nevertheless it is hard to give a precise estimate of this volatility. S. Muzzioli and C. Toricelli [6] handle this problem by using possibility distributions. In the first part of our paper we make some critical comments on their work. In the second part we present an alternative solution to the problem by performing a sensitivity analysis for the pricing of the option. This method is very general in the sense that it can be applied if one describes the uncertainty in the volatility by confidence intervals as well as if one describes it by fuzzy numbers. The conclusion is that the price of the option is not necessarily a strictly increasing function of the volatility.

Keywords

fuzzy sets, option pricing, sensitivity analysis

1 Introduction

In the first section of this paper we introduce the Cox-Ross-Rubinstein model [2] for the pricing of a European call option and the assumptions which are made. The well-known formula for the option price depends on the volatility of the underlying asset. However in practice it is hard to give a precise estimate of this volatility. S. Muzzioli and C. Toricelli [6] handle this problem by using possibility distributions. In the first part of our paper we make some critical comments on their work. In the second part we present an alternative solution to the problem

[*] This work has been supported by the BOF-project 001104599 of Ghent University

by performing a sensitivity analysis for the pricing of the option. This method is very general in the sense that it can be applied if one describes the uncertainty in the volatility by confidence intervals as well as if one describes it by fuzzy numbers. Indeed, in both approaches the imprecise volatility results in imprecise up and down factors. Those factors are modelled by a fuzzy quantity or are said to belong to a confidence interval.
We consider the case where the down factor is the inverse of the up factor. The lifetime of the option is divided into N steps of length T/N. Then we need to study the behaviour of the option price as a function of the up factor in an interval, which is a subset of $](1+r)^{T/N}, +\infty[$, where r stands for the risk-free interest rate. Therefore we study the functional behaviour of the option price for all possible values of the up factor.
Finally, we illustrate the method by an example with a fuzzy up factor.

2 The binary tree model

The binary tree model of Cox-Ross-Rubinstein [2] can be considered as a discrete-time version of the Black & Scholes model [1]. The following assumptions are made:

- The markets have no transaction costs, no taxes, no restrictions on short sales, and assets are infinitely divisible.
- The lifetime T of the option is divided into N steps of length T/N.
- The market is complete.
- No arbitrage opportunities are allowed which implies for the risk-free rate of interest r, that $d < (1+r)^{T/N} < u$, where u is the up factor and d the down factor.

The European call option price at time zero, has a well-known formula in this model:

$$EC(K,T) = \frac{1}{(1+r)^T} \sum_{j=0}^{N} \binom{N}{j} p_u^j (1-p_u)^{N-j} \left(S_0 u^j d^{N-j} - K\right)_+ \tag{1}$$

where K is the exercise (or strike) price, S_0 is the price of the underlying asset at time the contract begins, p_u the risk-neutral probability that the price goes up with the factor $u = \exp(\sigma\sqrt{T/N})$, with σ the volatility of the underlying asset. Let p_d be the risk-neutral probability that the price goes down with the factor d. We assume that $d = 1/u$. It is known that p_u and p_d are solutions to the system:

$$\begin{cases} p_u + p_d = 1 \\ dp_d + up_u = (1+r)^{T/N}. \end{cases} \tag{2}$$

The solutions are:

$$p_u = \frac{(1+r)^{T/N} - d}{u-d} = \frac{(1+r)^{T/N}u - 1}{u^2 - 1} \quad (3)$$

$$p_d = \frac{u - (1+r)^{T/N}}{u-d} = \frac{u^2 - (1+r)^{T/N}u}{u^2 - 1}. \quad (4)$$

3 Critical Analysis of the paper 'A Multiperiod Binomial Model for Pricing Options in an Uncertain World' by S. Muzzioli and C. Torricelli

S. Muzzioli and C. Torricelli [6] state: 'There are different methods for estimating volatility either from historical data, or from option prices. Sometimes it is hard to give a precise estimate of the volatility of the underlying asset and it may be convenient to let it take interval values. Moreover, it may be the case that not all members of the interval have the same reliability, as central members are more possible then the ones near the borders. This is exactly the idea behind our model, but instead of modelling volatility as a fuzzy quantity, we directly model the up and down jumps of the stock price.'

Instead of modelling the volatility as a fuzzy quantity, S. Muzzioli and C. Torricelli model directly the up and down factors u and d as the fuzzy numbers (u_1, u_2, u_3) and (d_1, d_2, d_3), where u_1 (resp d_1) is the minimum possible value, u_3 (resp d_3) is the maximum possible value and u_2 (resp d_2) is the most possible value. A triangular fuzzy number (a_1, a_2, a_3) can alternatively be defined by its α-cuts $[a_1(\alpha_2), a_3(\alpha_2)]$, $\alpha \in [0,1]$:

$$[a_1(\alpha_2), a_3(\alpha_2)] = [a_1 + \alpha(a_2 - a_1), a_3 - \alpha(a_3 - a_2)].$$

In fact a fuzzy quantity is completely defined by its α-cuts. Consider intervals $[a_1(\alpha), a_3(\alpha)]$, $\alpha \in [0,1]$, where

$$\alpha_1 \leq \alpha_2 : [a_1(\alpha_1), a_3(\alpha_1)] \subseteq [a_1(\alpha_2), a_3(\alpha_2)],$$

then the intervals $[a_1(\alpha), a_3(\alpha)]$ are the α-cuts of the fuzzy quantity a,

$$a(x) = \sup_{\alpha \in [0,1]} \min\{\alpha, 1_{[a_1(\alpha), a_3(\alpha)]}(x)\}, \qquad x \in \mathbb{R}.$$

Since the α-cuts of a triangular fuzzy number are compact intervals of the set of real numbers, the interval calculus of Moore [5] can be applied to them. Thus every binary operation in $\mathbb{R}$ can be extended to a binary operation on the set of fuzzy numbers.

S. Muzzioli and C. Torricelli consider a binary tree with 1 period, i.e. $T = N = 1$.

A fuzzy version of the two equations of the system (2) should be introduced. This can be done (for each equation) in two different ways:

$$\begin{aligned} p_u + p_d &= (1,1,1) \\ p_u &= (1,1,1) - p_d \end{aligned}$$

respectively

$$\begin{aligned} dp_d + up_u &= (1+r, 1+r, 1+r) \\ up_u &= (1+r, 1+r, 1+r) - dp_d \end{aligned}$$

where p_u and p_d are the fuzzy up and down probabilities $((p_u)_1, ((p_u)_2, ((p_u)_3)$ and $(((p_d)_1, ((p_d)_2, ((p_d)_3)$.
S. Muzzioli and C. Torricelli choose for both equations the first form. However, one has to take into account that p_u and p_d are fuzzy probabilities and therefore one should use the second form for the first equation. For the second equation in (2) the first form should be taken since the left-hand side is an expectation.
Thus the correct solution is obtained by extending the system (2) to

$$\begin{cases} p_u = (1,1,1) - p_d \\ dp_d + up_u = (1+r, 1+r, 1+r). \end{cases}$$

Expressed in α-cuts and keeping in mind that the operations are binary operations on fuzzy numbers, see e.g. E. Kerre [3], this leads to the system:

$$\begin{cases} [(p_u)_1(\alpha), (p_u)_3(\alpha)] = [1,1] - [(p_d)_1(\alpha), (p_d)_3(\alpha)] \\ [d_1(\alpha), d_3(\alpha)][(p_d)_1(\alpha), (p_d)_3(\alpha)] + [u_1(\alpha), u_3(\alpha)][(p_u)_1(\alpha), (p_u)_3(\alpha)] \\ \qquad = [1+r, 1+r] \end{cases}$$

or

$$\begin{cases} (p_u)_1(\alpha) = 1 - (p_d)_3(\alpha) \\ (p_u)_3(\alpha) = 1 - (p_d)_1(\alpha) \\ d_1(\alpha)(p_d)_1(\alpha) + u_1(\alpha)(p_u)_1(\alpha) = 1 + r \\ d_3(\alpha)(p_d)_3(\alpha) + u_3(\alpha)(p_u)_3(\alpha) = 1 + r. \end{cases}$$

The correct solution to this system is:

$$\begin{cases} (p_u)_1(\alpha) = \frac{d_1(\alpha)(d_3(\alpha)+u_3(\alpha))-(1+r)(d_1(\alpha)+u_3(\alpha))}{d_1(\alpha)d_3(\alpha)-u_1(\alpha)u_3(\alpha)} \\ (p_u)_3(\alpha) = \frac{d_3(\alpha)(d_1(\alpha)+u_1(\alpha))-(1+r)(d_3(\alpha)+u_1(\alpha))}{d_1(\alpha)d_3(\alpha)-u_1(\alpha)u_3(\alpha)} \\ (p_d)_1(\alpha) = \frac{(1+r)(d_3(\alpha)+u_1(\alpha))-u_1(\alpha)(d_3(\alpha)+u_3(\alpha))}{d_1(\alpha)d_3(\alpha)-u_1(\alpha)u_3(\alpha)} \\ (p_d)_3(\alpha) = \frac{(1+r)(d_1(\alpha)+u_3(\alpha))-u_3(\alpha)(d_1(\alpha)+u_1(\alpha))}{d_1(\alpha)d_3(\alpha)-u_1(\alpha)u_3(\alpha)}. \end{cases}$$

One can easily prove that for $\alpha = 1$:

$$\begin{cases} (p_u)_2 = \frac{(1+r)-d_2}{u_2-d_2} \\ (p_d)_2 = \frac{u_2-(1+r)}{u_2-d_2} \end{cases}$$

and for $\alpha = 0$:

$$\begin{cases} (p_u)_1 = \frac{d_1(d_3+u_3)-(1+r)(d_1+u_3)}{d_1d_3-u_1u_3} \\ (p_u)_3 = \frac{d_3(d_1+u_1)-(1+r)(d_3+u_1)}{d_1d_3-u_1u_3} \\ (p_d)_1 = \frac{(1+r)(d_3+u_1)-u_1(d_3+u_3)}{d_1d_3-u_1u_3} \\ (p_d)_3 = \frac{(1+r)(d_1+u_3)-u_3(d_1+u_1)}{d_1d_3-u_1u_3}. \end{cases}$$

Next S. Muzzioli and C. Torricelli calculate the price of the option in the one period model. They assume that the exercise price is between the highest value of the underlying asset in state down and the lowest value of the underlying asset in state up,

$$S_0d_3 \leq K \leq S_0u_1 \tag{5}$$

in which case the calculations are very simple. The aim of their next section is to extend the pricing methodology to a two period and then to a multi period binary model. The condition (5) is extended as follows:

$$S_0d_3^{j+1}u_3^{N-j-1} \leq K \leq S_0d_1^ju_1^{N-j} \;\; j = 0,\ldots,N-1$$

which is impossible since K can not be an element of those N intervals. Even if one changes the condition to

$$\exists j \in \{0,\ldots,N-1\} : S_0d_3^{j+1}u_3^{N-j-1} \leq K \leq S_0d_1^ju_1^{N-j}$$

the condition is not always fulfilled since one can easily prove (for example in the crisp case with $d = 1/u$) that $S_0d_3^{j+1}u_3^{N-j-1}$ is not always less then $S_0d_1^ju_1^{N-j}$. Even if this is the case, there are no economic reasons why the exercise price would not be out of the mentioned intervals.

Finally, they calculate the price of the option in one special situation of the two period model and remark that the extension to N periods is straightforward, which is not the case as we will see in what follows.

A last remark concerns the number of periods. S. Muzzioli and C. Torricelli extend the number of periods without explicitly mentioning that at the same time one should fix the lifetime of the option. Otherwise when the lifetime equals the number N of periods and N is increased, another option is considered at each step. Hence, if one models one and the same option, one has to fix the lifetime T and divide T in N subperiods of length T/N. Then increasing the number N of steps implies at the same time a decrease of the steplength.

This is also the way to proceed in order to be able to consider the important limit problem.

4 Imprecise volatility and the pricing of a European Call Option

The change of the price S_t of the underlying asset at time t can be modelled as in [4] by

$$S_{t+1} = \xi_{t+1} S_t$$

where ξ_{t+1} is a sequence taking values in a compact set M. We are interested in the special case where M consists only of two elements, its upper and lower bounds u and d. Those up and down factors depend on the volatility σ. As we already mentioned, it is often hard to give a precise estimate of the volatility. This problem can be avoided either by giving a confidence interval of the volatility or by modelling the volatility by a fuzzy quantity.

Imprecise volatility implies imprecision in the up (and down) factors. Under the assumptions of section 2 the (confidence or α-cut) intervals, to which the up factor, belongs, are subsets of $](1+r)^{T/N}, +\infty[$. We study the behaviour of the price of a European call option for all possible values of the up factor. In sections 5, 6 and 7 we also need to include the border case where the up factor equals $(1+r)^{T/N}$. Therefore we define the up factor as u_λ:

$$u_\lambda = (1+r)^{T/N} + \lambda, \qquad \lambda \in \mathbb{R}^+.$$

If we invoke (3), the risk-neutral probability, p_λ, that the price goes up, is

$$p_\lambda = \frac{(1+r)^{T/N} u_\lambda - 1}{u_\lambda^2 - 1}.$$

The price $C_\lambda(K)$ of the option is:

$$\begin{aligned} C_\lambda(K) &= \frac{1}{(1+r)^T} E[(S_\lambda^T - K)_+] \\ &= \frac{1}{(1+r)^T} \sum_{j=0}^{N} (S_0 u_\lambda^{2j-N} - K)_+ \cdot P[X_\lambda^N = j] \\ &= \frac{1}{(1+r)^T} \sum_{j=j_\lambda^*}^{N} (S_0 u_\lambda^{2j-N} - K) \binom{N}{j} p_\lambda^j (1-p_\lambda)^{N-j} \qquad (6) \end{aligned}$$

where X_λ^N is the number of ups in the lifetime T and $S_0 u_\lambda^{2j-N} - K$ is positive for $j \geq j_\lambda^*$.

Consider a confidence interval, $[u_0, u_1] \subset](1+r)^{T/N}, +\infty[$, of the up factor with

$$\begin{aligned} u_0 &= (1+r)^{T/N} + \lambda_0 \\ u_1 &= (1+r)^{T/N} + \lambda_1. \end{aligned}$$

If $u_\mu \in [u_0, u_1], \mu \in [0,1]$ then

$$\begin{aligned}
u_\mu &= \mu u_0 + (1-\mu)u_1 \\
&= \mu((1+r)^{T/N} + \lambda_0) + (1-\mu)((1+r)^{T/N} + \lambda_1) \\
&= (1+r)^{T/N} + [\mu\lambda_0 + (1-\mu)\lambda_1] \\
&= (1+r)^{T/N} + \lambda^*(\mu).
\end{aligned}$$

The price of the option belongs to the interval

$$[\min_{\mu\in[0,1]} C_{\lambda^*(\mu)}(K), \max_{\mu\in[0,1]} C_{\lambda^*(\mu)}(K)].$$

Suppose the imprecise volatility is described by using a fuzzy quantity, (u_1, u_2, u_3), $u_1, u_2, u_3 \in](1+r)^{T/N}, +\infty[$, with

$$\begin{aligned}
u_1 &= (1+r)^{T/N} + \lambda_1 \\
u_2 &= (1+r)^{T/N} + \lambda_2 \\
u_3 &= (1+r)^{T/N} + \lambda_3,
\end{aligned}$$

for the up factor. An α-cut, $\alpha \in [0,1]$, is the interval:

$$\begin{aligned}
&[u_1 + (u_2 - u_1)\alpha, u_3 + (u_2 - u_3)\alpha] = \\
&[(1+r)^{T/N} + \lambda_1 + \alpha(\lambda_2 - \lambda_1), (1+r)^{T/N} + \lambda_3 + \alpha(\lambda_2 - \lambda_3)].
\end{aligned}$$

An element of this interval can be described by

$$\begin{aligned}
&\mu[(1+r)^{T/N} + (\lambda_1 + \alpha(\lambda_2 - \lambda_1))] + (1-\mu)[(1+r)^{T/N} + \lambda_3 + \alpha(\lambda_2 - \lambda_3)] \\
&= (1+r)^{T/N} + \mu(\lambda_1 + \alpha(\lambda_2 - \lambda_1)) + (1-\mu)(\lambda_3 + \alpha(\lambda_2 - \lambda_3)) \\
&= (1+r)^{T/N} + \lambda^*_\alpha(\mu), \qquad \mu \in [0,1].
\end{aligned}$$

The α-cut, $\alpha \in [0,1]$, of the option price is:

$$[\min_{\mu\in[0,1]} C_{\lambda^*_\alpha(\mu)}(K), \max_{\mu\in[0,1]} C_{\lambda^*_\alpha(\mu)}(K)]. \tag{7}$$

It is clear that, for the method with confidence intervals as well as for the method using fuzzy quantities, the behaviour of $C_\lambda(K)$ as function of u_λ should be studied. This is the subject of the following sections.

5 Definitions, notations and lemmas

The function is broken up in its basic elements: first the (up and down) probabilities are considered, then their products and finally their products with the up and down factors. The risk-neutral probability, p_λ, is a decreasing function of u_λ. For $u_\lambda = (1+r)^{T/N}$ this probability is one and

$$\lim_{u_\lambda \to +\infty} p_\lambda = 0.$$

And one obtains $p^* = 0.5$ for

$$u_\lambda = u^* = (1+r)^{T/N} + \sqrt{(1+r)^{2T/N} - 1}.$$

The probability $1 - p_\lambda$ is an increasing function of u_λ.
The function $p_\lambda(1-p_\lambda)$ has a maximum for $u_\lambda = u^*$. It is zero for $u_\lambda = (1+r)^{T/N}$ and in the limit for $u_\lambda \to +\infty$.
The function $u_\lambda p_\lambda$ attains a minimum for $u_\lambda = u^*$. It is equal to $(1+r)^{T/N}$ for $u_\lambda = (1+r)^{T/N}$ and in the limit for $u_\lambda \to +\infty$.
The function $u_\lambda^{-1}(1-p_\lambda)$ attains a maximum for $u_\lambda = u^*$. It is zero for $u_\lambda = (1+r)^{T/N}$ and in the limit for $u_\lambda \to +\infty$.
One can prove that

$$(u_\lambda p_\lambda)' = \frac{1 - 2p_\lambda}{u_\lambda^2 - 1} = -(u_\lambda^{-1}(1-p_\lambda))'.$$

6 Functional behaviour of the functions $C_1(\lambda, j)$ and $C_2(\lambda, j, K)$

In the next section we will examine the functional behaviour of each term in the sum (6). Those terms consist of two parts, namely $C_1(\lambda, j) = S_0 u_\lambda^{2j-N} \binom{N}{j} p_\lambda^j (1-p_\lambda)^{N-j}$ and $C_2(\lambda, j, K) = -K\binom{N}{j} p_\lambda^j (1-p_\lambda)^{N-j}$. Those functions are first examined separately, regardless the sign of their sum.
The derivative of the function $C_1(\lambda, j)$ with respect to u_λ, $u_\lambda \in [(1+r)^{T/N}, +\infty]$, is:

$$\begin{aligned}
&(C_1(\lambda, j))' \qquad (8)\\
&= \frac{S_0\binom{N}{j}}{u_\lambda}(u_\lambda p\lambda)^{j-1}(u_\lambda^{-1}(1-p_\lambda))^{N-j-1}(j(1-p_\lambda+u_\lambda^2 p_\lambda) - N u_\lambda^2 p_\lambda)\frac{1-2p_\lambda}{u_\lambda^2 - 1}
\end{aligned}$$

which implies that:

- If $j \leq N/2$ then $j(1-p_\lambda+u_\lambda^2 p_\lambda) - N u_\lambda^2 p_\lambda < 0$ and the function $C_1(\lambda, j)$ attains a maximum for $u_\lambda = u^*$. It is zero for $u_\lambda = (1+r)^{T/N}$ and in the limit for $u_\lambda \to +\infty$.
- If $N/2 < j < N$ then the expression $j(1-p_\lambda+u_\lambda^2 p_\lambda) - N u_\lambda^2 p_\lambda$

- is negative for all u_λ if moreover $(1+r)^{T/N} \leq \frac{N}{2\sqrt{(N-j)j}}$ and the function $C_1(\lambda, j)$ attains a maximum for $u_\lambda = u^*$.
- if $(1+r)^{T/N} > \frac{N}{2\sqrt{(N-j)j}}, \qquad j = \text{floor}(N/2+1)$

  (a) the expression is negative for $j > \frac{Nu^*}{2(1+r)^{T/N}}$ and the function $C_1(\lambda, j)$ attains a maximum for $u_\lambda = u^*$

  (b) the expression has two roots for $j \leq \frac{Nu^*}{2(1+r)^{T/N}}$

  Those roots are:

$$\begin{aligned} u_1(j) &= \frac{N + \sqrt{N^2 - 4(N-j)j(1+r)^{2T/N}}}{2(N-j)(1+r)^{T/N}} \\ u_2(j) &= \frac{N - \sqrt{N^2 - 4(N-j)j(1+r)^{2T/N}}}{2(N-j)(1+r)^{T/N}} \end{aligned}$$

  with $u_1(j) \geq u^* \geq u_2(j) \geq (1+r)^{T/N}$
  and if $u_1(j) = u_2(j)$ then $u_1(j) = u_2(j) = u^*$.
  The function $C_1(\lambda, j)$ attains a maximum for $u_\lambda = u_2(j)$ and for $u_\lambda = u_1(j)$. It attains a minimum for $u_\lambda = u^*$.
- The function $C_1(\lambda, j)$ is zero for $u_\lambda = (1+r)^{T/N}$ and in the limit for $u_\lambda \to +\infty$.

- if $j = N$ then the function equals $S_0(u_\lambda p_\lambda)^N$ and it attains a minimum for $u_\lambda = u^*$. The function $C_1(\lambda, j)$ is equal to $S_0(1+r)^T$ for $u_\lambda = (1+r)^{T/N}$ and in the limit for $u_\lambda \to +\infty$.

The derivative of the function $C_2(\lambda, j, K), 0 < j < N$, with respect to u_λ is:

$$(C_2(\lambda, j, K))' = -K\binom{N}{j}(j - Np_\lambda)p_\lambda^{j-1}(1-p_\lambda))^{N-j-1}(p_\lambda)' \qquad (9)$$

The factor $(j - Np_\lambda)$ has two roots for all j:

$$\begin{aligned} u_1^*(j) &= \frac{N(1+r)^{T/N} + \sqrt{N^2(1+r)^{2T/N} - 4j(N-j)}}{2j} \\ u_2^*(j) &= \frac{N(1+r)^{T/N} - \sqrt{N^2(1+r)^{2T/N} - 4j(N-j)}}{2j} \end{aligned}$$

but $u_2^*(j) < (1+r)^{T/N}$.
The function attains a minimum for $u_\lambda = u_1^*(j)$. If $j \leq N/2$ then $u_1^*(j) > u^*$ and if $j \geq N/2$ then $u_1^*(j) < u^*$. The function is zero for $u_\lambda = (1+r)^{T/N}$ and in the limit for $u_\lambda \to +\infty$.
The function is decreasing for $j = 0$. It is zero for $u_\lambda = (1+r)^{T/N}$ and equal to

$-K$ in the limit for $u_\lambda \to +\infty$.
The function $C_2(\lambda, j, K)$ increases for $j = N$. It is equal to $-K$ for $u_\lambda = (1+r)^{T/N}$ and zero in the limit for $u_\lambda \to +\infty$.

7 The branches of the binary tree considered separately

Each term in the sum corresponds to a branch in the binary tree. Such a term depends on j, $j = 0, \dots, N$, and K, and is function of λ:

$$C_\lambda(j,K) = (S_0 u^{2j-N} - K)\binom{N}{j} p_\lambda^j (1-p_\lambda)^{N-j}.$$

The functional behaviour of $C_\lambda(j,K)$ is examined regardless of its sign.
Noting that

$$C_\lambda(j,K) = C_1(\lambda, j) + C_2(\lambda, j, K)$$

the derivative of $C_\lambda(j,K)$ with respect to u_λ can be calculated by invoking (8) and (9):

$$S_0\binom{N}{j}(u_\lambda p\lambda)^{j-1}(u_\lambda^{-1}(1-p_\lambda))^{N-j-1}u_\lambda^{-1}(j(1-p_\lambda+u_\lambda^2 p_\lambda) - Nu_\lambda^2 p_\lambda)\frac{1-2p_\lambda}{u_\lambda^2-1}$$
$$-K\binom{N}{j}(j-Np_\lambda)p_\lambda^{j-1}(1-p_\lambda))^{N-j-1}(p_\lambda)'.$$

In those intervals where both derivatives (8) and (9) have the same sign or for those values of u_λ where one of the derivatives is zero, one can immediately conclude from section 6 if the term is decreasing or increasing.
On the other hand we can draw conclusions about the functional behaviour of the term by remarking that we studied the functional behaviour of $S_0 u_\lambda^{2j-N}$ multiplied by $\binom{N}{j} p_\lambda^j (1-p_\lambda^{j-N})$ and that the term can be calculated in two steps: first subtract K from $S_0 u_\lambda^{2j-N}$ and then multiply the result by $\binom{N}{j} p_\lambda^j (1-p_\lambda^{j-N})$.
This leads to the following conclusions:

$\boxed{j = 0}$

- $C_0(0,K) = 0$ and $C_\lambda(0,0) > 0$,
  $\lim\limits_{u_\lambda \to +\infty} C_\lambda(0,K) = -K$.
- If $K \geq S_0(1+r)^{-T}$ then $C_\lambda(0,K)$ is negative for all u_λ.

- If $0 < K < S_0(1+r)^{-T}$ then $C_\lambda(0,K)$ has a root, $u^*(0) = (\frac{S_0}{k})^{\frac{1}{N}}$. The function is negative for all $u_\lambda > u^*(0)$, it attains a maximum in the interval $](1+r)^{T/N}, u^*[$. The root and the maximum decrease as K increases.

$\boxed{0 < j < \frac{N}{2}}$

- $C_0(j,K) = 0$ and $C_\lambda(j,0) > 0$,
  $\lim_{u_\lambda \to +\infty} C_\lambda(j,K) = 0$
- If $K \geq S_0(1+r)^{\frac{(2j-N)T}{N}}$ then $C_\lambda(j,K)$ is negative for all u_λ.
- If $K < S_0(1+r)^{\frac{(2j-N)T}{N}}$ then $C_\lambda(j,K)$ has a root, $u^*(j) = (\frac{K}{S_0})^{\frac{1}{2j-N}}$. The function is negative for all $u_\lambda > u^*(j)$, it attains a maximum in the interval $](1+r)^{T/N}, u^*[$. The root and the maximum decrease as K increases.
- Since the function converges to zero it attains a minimum in the interval $]u^*(j), +\infty[$.

$\boxed{j = N/2}$, j is odd
Since $C_\lambda(j,K) = (S_0 - K)\binom{N}{j}(p_\lambda(1-p_\lambda))^{N/2}$,

- $C_0(j,K) = 0$ and $\lim_{u_\lambda \to +\infty} C_\lambda(j,K) = 0$.
- The function is positive for all u_λ if $S_0 > K$ and negative for all u_λ if $S_0 < K$.
- The function attains a maximum for $u_\lambda = u^*$ if $S_0 > K$ and a minimum for $u_\lambda = u^*$ if $S_0 < K$.

$\boxed{\frac{N}{2} < j < N}$

- $C_0(j,K) = 0$ and $\lim_{u_\lambda \to +\infty} C_\lambda(j,K) = 0$
- If $(1+r)^{T/N} \leq \frac{N}{2\sqrt{(N-j)j}}$
  or
  $(1+r)^{T/N} > \frac{N}{2\sqrt{(N-j)j}}$ and $\frac{N}{2} < j \leq \frac{Nu^*}{2(1+r)^{T/N}}$
  then
  - If $K \leq S_0(1+r)^{\frac{(2j-N)T}{N}}$ the function is positive for all u_λ and attains a maximum, larger then u^*. The maximum increases as K increases.

  - If $K > S_0(1+r)^{\frac{(2j-N)T}{N}}$ the function has a root, $u^*(j) = (\frac{K}{S_0})^{\frac{1}{2j-N}}$. The function is positive for all $u_\lambda > u^*(j)$. It attains a maximum, larger then u^*. The maximum and the root increase as K increases. It attains a minimum, smaller then $u_1^*(j)$, between $(1+r)^{T/N}$ and the root.

- $(1+r)^{T/N} > \frac{N}{2\sqrt{(N-j)j}}$ and $\frac{Nu^*}{2(1+r)^{T/N}} < j < N$

  - If $K \leq S_0(1+r)^{\frac{T(2j-N)}{N}}$ then the function is positive for all u_λ. It attains a maximum and a minimum in $]u_2(j), u^*[$. If K increases the difference between the maximum and the minimum becomes insignificant. It also attains a maximum which is larger then $u_1(j)$.
  - If $S_0(1+r)^{\frac{T(2j-N)}{N}} < K$ then the function has a root, $u^*(j) = (\frac{K}{S_0})^{\frac{1}{2j-N}}$, and is negative for all $u_\lambda < u^*(j)$. It attains a minimum, smaller then $u_1^*(j)$, between $(1+r)^{T/N}$ and the root.

$\boxed{j = N}$

- $C_0(N,K) = S_0(1+r)^T - K$ and $\lim\limits_{u_\lambda \to \infty} C_\lambda(N,K) = S_0(1+r)^T$.

- If $K \geq S_0(1+r)^T$ then $C_\lambda(N,K)$ has a root, $u^*(N) = (\frac{K}{S_0})^{\frac{1}{N}}$. The function is positive for all $u_\lambda > u^*(j)$. The root decreases when K decreases. The function increases.

- If $S_0(1+r)^{T-\frac{2T}{N}} \leq K < S_0(1+r)^T$ then the function is positive and increasing for all u_λ.

- If $0 \leq K < S_0(1+r)^{T-\frac{2T}{N}}$ then the function is positive for all u_λ and attains a minimum in $](1+r)^{T/N}, u^*[$. The minimum decreases as K increases.

8 Procedure for the Pricing of the European Call Option

Suppose that K is such that $C_\lambda(j,K)$ is positive for all j, then the price $EC(K,T)$ (1) or (6) reads

$$\begin{aligned} C_\lambda(K) &= \frac{1}{(1+r)^T}\sum_{j=0}^{N}(S_0 u_\lambda^{2j-N}-K)\binom{N}{j}p_\lambda^j(1-p_\lambda)^{N-j} \\ &= \frac{S_0}{(1+r)^T}\sum_{j=0}^{N}(u_\lambda p_\lambda)^j(u_\lambda^{-1}(1-p_\lambda))^{N-j} \\ &\quad -\frac{K}{(1+r)^T}\sum_{j=0}^{N}p_\lambda^j(1-p_\lambda)^{N-j} \\ &= \frac{S_0}{(1+r)^T}(u_\lambda p_\lambda+u_\lambda^{-1}(1-p_\lambda))^N-\frac{K}{(1+r)^T} \\ &= S_0-\frac{K}{(1+r)^T}, \end{aligned}$$

where in the last equality we applied (2).

This case is only possible if $K < S_0$, since otherwise the terms for $j < N/2$ are not in the sum. If this condition is fulfilled for K, then all terms for $j \geq N/2$ are in the sum. Therefore we concentrate on the terms with $j < N/2$. The expression $C_\lambda(j,K)$ is positive for all $u_\lambda < (\frac{S_0}{K})^{\frac{1}{N-2j}}$.

The smallest root is $(\frac{S_0}{K})^{\frac{1}{N}}$. This root is larger then $(1+r)^{T/N}$ if $0 < K \leq S_0(1+r)^{T/N}$. If, in this case,

$$(1+r)^{T/N} < u_\lambda < (\frac{S_0}{K})^{\frac{1}{N}}$$

then all terms are in the sum and $C_\lambda(K)$ is constant for those values of u_λ, namely $C_\lambda(K) = S_0(1+r)^{T/N} - K$.

If u_λ increases:

$$(\frac{S_0}{K})^{\frac{1}{N}} \leq u_\lambda < (\frac{S_0}{K})^{\frac{1}{N-2}}$$

then $C_\lambda(0,K) < 0$ and the corresponding term is not in the sum. Thus $C_\lambda(K) = S_0 - K(1+r)^{-T} - C_\lambda(0,K)(1+r)^{-T}$. Since $C_\lambda(0,K)$ is negative and decreasing for $(\frac{S_0}{K})^{\frac{1}{N}} \leq u_\lambda < (\frac{S_0}{K})^{\frac{1}{N-2}}$, $C_\lambda(K)$ increases for those values of u_λ.

This procedure can be extended for all values of u_λ and K.

Finally, we illustrate the procedure by an example in the case the imprecise volatility is described by a fuzzy quantity.

Let $0 < K \leq S_0(1+r)^{T/N}$ and

$$\begin{array}{rcl} u_1 & \in &](1+r)^{T/N}, (\frac{S_0}{K})^{\frac{1}{N}}[\\ u_2 & = & (\frac{S_0}{K})^{\frac{1}{N}} \\ u_3 & \in &](\frac{S_0}{K})^{\frac{1}{N}}, \frac{S_0}{K})^{\frac{1}{N-2}}[. \end{array}$$

then, by applying (7), the α-cuts of the option price are

$$[S_0 - \frac{K}{(1+r)^T}, S_0 - \frac{K}{(1+r)^T} - \frac{C_{\lambda^*_\alpha(1)}(0,K)}{(1+r)^T}].$$

9 Conclusions

In the continuous Black & Scholes model, of which the binary tree model is a discrete time version, the price of a European call option is a strictly increasing function of the volatility, since the hedging parameter vega, i.e. the derivative of the price with respect to the volatility, is strictly positive.

In the discrete case we studied the functional behaviour of the price in order to model the uncertainty in the volatility. We can conclude that in the binary tree model the price is not necessarily a strictly increasing function of the volatility. As further research we will investigate the functional behaviour when this discrete time model converges to the Black & Scholes model.

References

[1] F. Black and M. Scholes. The Pricing of Options and Corporate Liabilities. *Journal of Political Economics*, 7:637–659, 1973.

[2] C. Cox and M. Rubinstein. Option pricing. A simplified approach. *Journal of Financial Economics*, 81:229–263, 1979.

[3] E. Kerre. Fuzzy Sets and Approximate Reasoning. *Xian Jiaotong University Press*, pp 254, 1998.

[4] V. Kolokotsov. Nonexpansive maps and option pricing theory. *Kibernetica*, 34:713–724, 1998.

[5] R. Moore. Interval Analysis. *Prentice-Hall*, 1966.

[6] S. Muzzioli and C. Torricelli. A Multiperiod Binomial Model for Pricing Options in an Uncertain World. *Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, 255–264, 2001.

Huguette Reynaerts is with the Ghent University, Ghent, Belgium. E-mail: huguette.reynaerts@rug.ac.be

Michèle Vanmaele is with the Ghent University, Ghent, Belgium. E-mail: michele.vanmaele@rug.ac.be

Inference in Credal Networks with Branch-and-Bound Algorithms[*]

JOSÉ CARLOS FERREIRA DA ROCHA
Escola Politécnica, Universidade de São Paulo
Universidade Estadual de Ponta Grossa, Brazil

FABIO GAGLIARDI COZMAN
Escola Politécnica, Universidade de São Paulo, Brazil

Abstract

A credal network associates sets of probability distributions with directed acyclic graphs. Under strong independence assumptions, inference with credal networks is equivalent to a signomial program under linear constraints, a problem that is NP-hard even for categorical variables and polytree models. We describe an approach for inference with polytrees that is based on branch-and-bound optimization/search algorithms. We use bounds generated by Tessem's A/R algorithm, and consider various branch-and-bound schemes.

Keywords

credal networks, strong independence, probability intervals, inference, branch-and-bound algorithms

1 Introduction

A credal network provides a representation for imprecise probabilistic knowledge through direct acyclic graphs (DAGs) [1]. In this formalism, each node in a DAG represents a random variable, and each variable is associated with convex sets of probability distributions. The structure of the graph indicates relations of probabilistic independence between variables. In this paper we interpret independence relations as statements of strong independence [1, 2].

A credal network can be viewed as a Bayesian network [3] with relaxed numerical statements. Credal networks can be used to study the robustness of

[*]The first author is supported in part by CAPES. The work has received substantial support from HP Labs through "Convênio Redes Bayesianas para Aprendizado."

Bayesian networks [4], or to represent vague or incomplete probability statements.

An *inference* with a credal network is the computation of upper and lower probability values for each category of a *query* variable. This computation is NP-hard even for polytrees [5], and it can be viewed as a signomial program under linear constraints [6]. Exact and approximate inference algorithms have been proposed in the literature, but no algorithm can handle large credal networks exactly.

In this article we propose new algorithms for inferences in polytrees. The idea is to use branch-and-bound search/optimization techniques to produce inferences. We explore Tessem's A/R algorithm [7] as a bound generation mechanism. We show how this approach can generate exact and approximate inferences, illustrating the main ideas with of examples.

The organization of the text is as follows. Sections 2 and 3 present a summary of credal networks and branch-and-bound techniques. Section 4 describes how exact and approximate inference can be performed with branch-and-bound techniques and the A/R algorithm. Section 5 shows how exact and approximate techniques can be combined through decomposition of networks. Section 6 discusses the proposed algorithms and results.

2 Credal sets, credal networks and inference

A convex set of probability distributions is called a credal set [8].[1] Denote the probability density of a categorical random variable X by $p(X)$. A credal set for X is denoted by $K(X)$; we assume that every credal set has a finite number of vertices. We can represent such a set just enumerating its vertices. A conditional credal set is a set of conditional distributions. We obtain a conditional credal set applying Bayes rule to each distribution in a joint credal set.

Given a number of marginal and conditional credal sets, an *extension* of these sets is a joint credal set with the given marginal and conditional credal sets. A collection of marginal and conditional credal sets can have more than one extension. In this paper we are always interested in computing the largest possible extension for a given collection of marginal and conditional credal sets.

Credal networks associate credal sets with a direct acyclic graph. In analogy to Bayesian networks, in a credal network every node of a directed acyclic graph is associated with a variable,[2] and every variable is associated with a collection of *local* credal sets $K(X|\text{pa}(X))$, where $\text{pa}(X)$ denotes the parents of variable X in the graph. That is, a node stores the credal sets

$$\{K(X|\text{pa}(X) = \pi_1), \ldots, K(X|\text{pa}(X) = \pi_m)\},$$

[1] We deal only with convex sets.

[2] To simplify the text, we represent a node and its variable with the same symbol.

where $\{\pi_1, \ldots, \pi_m\}$ are the instances of $\mathrm{pa}(X)$. A root node has only one credal set associated with it.

The sets $K(X|\mathrm{pa}(X))$ are called *separately specified* when there is no relationship between them for different values of $\mathrm{pa}(X)$. In this paper we assume that local credal sets are always separately specified.

The basic assumption in a credal network is that every variable is independent of its nondescendants nonparents given its parents. Obviously the import of such a condition depends on which concept of independence for credal sets is adopted [1, 2, 9]. In this paper we adopt the concept of *strong independence*: two variables X and Y are strongly independent when every extreme point of $K(X,Y)$ satisfies stochastic independence of X and Y (that is, each vertex $p(X,Y) \in K(X,Y)$ satisfies $p(X|Y) = p(X)$ and $p(Y|X) = p(Y)$ for all possible conditioning values) [10].

The *strong extension* of a credal network is the largest joint credal set such that every variable is strongly independent of its nondescendants nonparents given its parents. The strong extension of a credal network is the joint credal set that contains every possible combination of vertices for all credal sets in the network, such that the vertices are combined as follows [1]:

$$p(X_1, \ldots, X_n) = \prod_i p(X_i|\mathrm{pa}(X_i)) \,. \tag{1}$$

Figure 1 shows the structure of a credal network that is latter used in examples.

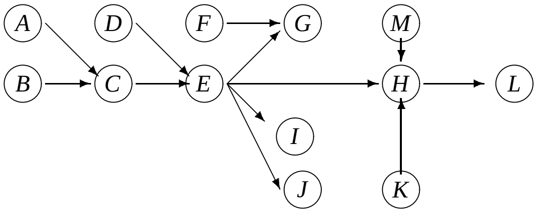


Figure 1: A polytree credal network.

An *inference* in a credal network is the computation of tight bounds for probability values in an extension of the network. These bounds are called *upper* and *lower* probabilities. If X_q is a *query* variable and $\mathbf{X}_E$ represents a set of *observed* variables, then an inference is the computation of tight bounds for $p(X_q|\mathbf{X}_E)$ for one or more values of X_q.

Algorithms for exact inference in strong extensions can be found in [1, 5, 11, 12]. The only known polynomial algorithm for strong extensions is the 2U algorithm, which processes polytrees with binary variables [13]. In general, exact inference in credal networks is a NP-hard problem (even for polytrees), so approximate algorithms are a natural solution. We distinguish *outer* approximations from *inner* ones; the former are produced when the correct interval between lower and upper probabilities is enclosed in the approximate interval; the latter approximations are produced when the correct interval encloses the approximate interval. Outer approximations can be found in [7, 14, 15], and inner approximate algorithms can be found in [4, 16, 17]. Generally speaking, inner algorithms are obtained by local optimization methods.

In this paper we are interested in inferences with strong extensions. The difficulty faced by inference algorithms is the potentially enormous number of vertices that a strong extension can have — even a relatively small network can dwarf the best exact algorithms. Consider the following example, taken from [5]:

Example 1 *Consider a network with four variables X, Y, Z and W; W is the sole child of X, Y and Z, and there are no other arrows in the network. Suppose that all variables have three values and that every local credal set has only three vertices. The vertices of the strong extension $K(X,Y,Z,W)$ factorize as $p(W,X,Y,Z) = p(W|X,Y,Z)\,p(X)\,p(Y)\,p(Z)$. Now, W is associated with 27 credal sets; therefore there are 3^{27} ways to combine the vertices of these credal sets. These 3^{27} vertices must be combined with every combination of vertices of $K(X)$, $K(Y)$ and $K(Z)$. So, the potential number of vertices in $K(X,Y,Z,W)$ is 3^{30}.*

We note that inference in credal networks is an optimization problem. Consider the computation of an upper probability:

- The goal is to find a distribution $p(X_i|\text{pa}(X_i))$ in $K(X_i|\text{pa}(X_i))$, for each variable X_i, so as to maximize the probability value $p(X_q|\mathbf{X}_E)$.
- The objective function $p(X_q|\mathbf{X}_E)$ is a fraction of multilinear expressions:

$$p(X_q|\mathbf{X}_E) = \frac{\sum_{X_1,\ldots,X_n\setminus\{X_q,\mathbf{X}_E\}} \prod_i p(X_i|\text{pa}(X_i))}{\sum_{X_1,\ldots,X_n\setminus\mathbf{X}_E} \prod_i p(X_i|\text{pa}(X_i))}.$$

- The maximization is subject to linear constraints, given our assumption of credal sets with finitely many vertices.

This maximization problem belongs to the field of *signomial* programming [6], as observed independently by [4, 12, 18]. Signomial programs are generally solved dividing the feasible set ("branching" on various subsets) and obtaining outer approximations ("bounding" the objective function in each subset) [6, 19]. That is, signomial programming is solved by branch-and-bound procedures. The great advantage of signomial programming over more general optimization problems is that it is possible to obtain bounds for signomial programs using geometric programming — a well establish field that can be tackled efficiently through convex programming [20]. However, direct application of geometric programming bounds to strong extensions seems to face difficulties. First, the inference problem is an “implicit” signomial programming, as the objective function is encoded in the graph through Expression (1); each combination of variables in the credal network would be a maximizer in the geometric program. Second, and perhaps more importantly, the “degree of difficulty” of a geometric program depends on the number of polynomial terms in the program — note that Expression (1) summarizes a large number of terms.

In this paper we adopt the basic idea of branching and bounding to compute lower and upper probabilities, but instead of relying on properties of geometric programming, we use bounds that have been specifically developed for strong extensions.

3 Branch-and-bound search and optimization

Branch-and-bound techniques appear in artificial intelligence, optimization and constraint satisfaction [21, 22]. The basic purpose of a branch-and-bound algorithm is to optimize a function. For example, take a problem P stated as:

$$\begin{array}{ll} (P) & \max f(w) \\ & \quad \text{s.t.} \qquad g(w) \leq 0, w \in \mathbf{W}, \end{array}$$

where $\mathbf{W} \subseteq \Re^n$, f is a real valued function, and the image of g is contained in $\Re^m$. A branch-and-bound technique is suitable for P whenever it is possible to divide P in sub-instances that are easier to solve or approximate than P itself, and such that the solution for P is present in one of these sub-instances [23, 24]. Additionally, a branch-and-bound technique requires a bound r (overestimation) for the solution of P. This upper bound is usually obtained from a relaxed version of P, indicated by R. Obviously, R must be easier and faster to solve than P, and must give a good approximation for P. The relaxed bound for P is denoted by $r(P)$.

In our implementation we use the following version of branch-and-bound [25]; several variants exist for it [23].

Algorithm 1 - Depth-first branch-and-bound

- *Input:* a problem P.
- *Output:* the value of $\max f(w)$, denoted by $\overline{p}$.

1. Initialize $\hat{p}$ with a small value (necessarily smaller than $\overline{p}$).
2. If $\mathbf{W}$ contains a single value w then: update $\hat{p}$ with $f(w)$ when $f(w) > \hat{p}$;
3. else:

   (a) using decomposition, obtain a list L of sub-instances of P; each sub-instance is denoted by P_h and has feasible region $\mathbf{W}_h$.

   (b) for each P_h do

      i. if $\mathbf{W}_h$ is feasible in the original problem and $r(P_h) > \hat{p}$, call recursively depth-first branch-and-bound over P_h.

4. Take the last $\hat{p}$ as $\overline{p}$.

This algorithm can be viewed as a search in a tree where the root node contains P and descendant nodes contain sub-instances of P. The leaf nodes contain problems that can be exactly solved. When a leaf node l is reached, the value for f at l is computed; if this value is the largest one up to that moment, it is retained. Non-leaf nodes are processed by relaxing the original problem and producing bounds. Every non-leaf node is expanded by decomposition into sub-instances, as long as its bound is larger than the current best value.

4 Branch-and-bound inference in strong extensions

This section contains the central ideas in this paper. We use a branch-and-bound procedure where

- branching occurs at every vertex of credal sets, and
- bounding is achieved by Tessem's A/R algorithm [7].

Given a query variable X and a credal network $\mathcal{N}$, a single run of the branch-and-bound procedure computes the lower or upper probability for a single state of X, denoted by x.

The first step is to discard variables that are not used to compute the inference; this can be done using d-separation [26]. The resulting network is denoted by $\mathcal{N}_0$.

Branching.

The root node in the branch-and-bound search tree is $\mathcal{N}_0$. The root node $\mathcal{N}_0$ is then divided into several simpler credal networks $\{\mathcal{N}_{01}, \ldots, \mathcal{N}_{0q}\}$. Each one of these networks is obtained as follows. We select one credal set in $\mathcal{N}_0$, and produce as many networks as there are vertices in this credal set — each network is associated with a single vertex of the selected credal set. This decomposition procedure is then applied recursively, following the branch-and-bound algorithm. At each step, a credal set is "expanded". Using this decomposition strategy, a leaf node contains a Bayesian network, obtained by a particular selection of vertices in all credal sets in the credal network. When a leaf node is reached, a variable elimination algorithm is used to perform inference in the Bayesian network defined by the leaf [27]. Such an algorithm produces a probability value $p(x|\mathbf{X}_E)$; if $p(x|\mathbf{X}_E)$ is greater than the current maximum probability, the latter value is updated.

We always select the non-expanded credal set nearest to the queried variable, but we always keep the query variable to be processed at last (a similar criterion is used in [28] to deal with partial evaluation of belief nets). We have tried several criteria for the selection of the credal sets that are expanded, and we found that the procedure just described is quite appropriate.

Bounding.

For non-leaf nodes in the search tree, we run the A/R algorithm as a relaxation of exact inference [7], because this algorithm produces outer bounds rather quickly. The A/R algorithm focuses on polytrees, even though it can be modified to handle more general networks [14].

The A/R algorithm assumes that every credal set is approximated by a collection of probability intervals. So we must convert the credal network to an interval-based Bayesian network (conditional probability tables contain intervals). Obviously the replacement of credal sets by probability intervals introduces potential inaccuracies into the process.

The A/R algorithm mimics the dynamics of Pearl’s belief propagation algorithm [3]. The functions λ, π and the messages used in BP are still defined with identical purposes, but they are now interval-valued functions. The idea is to manipulate these intervals using interval arithmetic and two additional techniques called by Tessem *annihilation* and *reinforcement*.

We can understand the basic ideas in the A/R algorithm by looking at the computation of the interval-valued message $\pi(X)$ — this message is computed at a node X with parents $Y_0,\ldots,Y_k$. Consider then the computation of $\pi_*(x_j)$, the lower bound of $\pi(x_j)$ for a particular value x_j:

1. Construct a interval-valued function $\beta(Y_0,\ldots,Y_k)$ by interval-multiplication of the messages $\pi_X(Y_i)$ received by X (these messages are also interval-valued).

2. Construct a distribution $p(Y_0,\ldots,Y_k)$ that is consistent with the intervals in $\beta(Y_0,\ldots,Y_k)$, such that $p(Y_0,\ldots,Y_k)$ minimizes the sum

$$\sum_{Y_0,\ldots,Y_k} \underline{p}(x_j|Y_0,\ldots,Y_k)\, p(Y_0,\ldots,Y_k)\,,$$

   where $\underline{p}(x_j|Y_0,\ldots,Y_k)$ is the lower value for $p(x_j|Y_0,\ldots,Y_k)$; the minimum of the sum is $\pi_*(x_j)$.

These operations are efficient because it is not hard to find $p(Y_0,\ldots,Y_k)$ in step 2: sort $\underline{p}(x_j|Y_0,\ldots,Y_k)$ in increasing order, and distribute probability mass (consistently with $\beta(Y_0,\ldots,Y_k)$) from the smallest to the largest value of $\underline{p}(x_j|Y_0,\ldots,Y_k)$. The same operations can be adapted to compute the upper bound $\overline{\pi}^*(x_j)$.

The A/R algorithm prescribes similar operations for computation of $\lambda_X(Y_i)$ and $\pi_{Z_i}(X)$ (where Z_i is a child of X). The function $\lambda(X)$ is obtained by direct interval multiplication. Finally, the algorithm uses annihilation or reinforcement

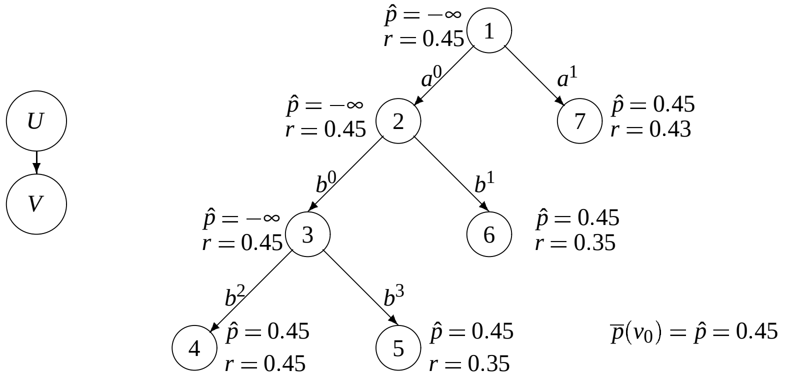


Figure 2: An example of branch-and-bound based inference. Left: a simple credal network, where $K(U)$ is the convex hull of $\{a^0, a^1\}$, with $a^0 = (0.5, 0.5)$ and $a^1 = (0.3, 0.7)$; $K(V|u_0)$ is the convex hull of $\{b^0, b^1\}$, with $b^0 = (0.5, 0.5)$ and $b^1 = (0.3; 0.7)$; $K(V|u_1)$ is the convex hull of $\{b^2, b^3\}$, with $b^2 = (0.4, 0.6)$ and $b^3 = (0.2, 0.8)$. Right: Search tree for computation of $\overline{p}(v_0)$.

operations to "normalize" the functions $\lambda_X(Y_i)$, $\pi_{Z_i}(X)$, and the product $\pi(X)\lambda(X)$ — "normalization" means simply computing bounds that take into account the fact that probability distributions add up to one.

In our branch-and-bound procedure, the deeper a node is in the search tree, the more point probabilities are manipulated by the A/R algorithm.

Example 2 *Figure 2 shows a very simple network and the the basic steps of our branch-and-bound algorithm when computing $\overline{p}(v_0)$. Nodes in the search tree represent credal networks; the numbering inside nodes indicates the order in which nodes are visited. The value r is obtained by the A/R algorithm. Close to each arc in the search tree we indicate which vertex (and for which credal set) was expanded.*

Experiments.

We have implemented the branch-and-bound scheme in a Java program, using Pentium IV machines to run tests. We ran experiments with networks containing variables with three and four states. Each configuration was tested against several randomly generated credal nets [29]. Experiments discussed in this section have no evidence ($\mathbf{X}_E = \emptyset$); this restriction simplifies the presentation with no loss in generality.

Table 1: Cost for exact inference for E in the network of Figure 1.

# states per variable	# vertices per credal sets	Potential size of the strong extension	Visited nodes (mean)	Samples (networks)
03	02	2^{21}	5499	35
03	03	3^{21}	284912	10
04	02	2^{35}	559255	10

We have observed that the size of the search tree explored by branch-and-bound is usually a small fraction of the potential vertices of the strong extension. As an example, consider the network in Figure 1. Table 1 shows relevant results for query variable E, indicating the number of states for variables, the number of vertices for each credal set in the network, and the potential number of vertices of the strong extension. The table indicates how many networks of each type were tested, and the mean number of visited nodes during branch-and-bound. Note the enormous difference between the potential number of vertices and the number of effectively expanded nodes.

As another instructive example, we applied the branch-and-bound scheme to the network described in Example 1. We tested thirty randomly generated credal networks with the same structure and different credal sets; in each one of them we computed the lower and upper probabilities for w_0. These sample networks had ternary variables and three distributions in each credal set. The branch-and-bound search was always able to quickly compute the exact inference, on average exploring 243 nodes per inference.

Consider another example. We took the polytree structure of the well-known Bayesian network called “Car Starts”[3] and set all of its variables as ternary. We assumed that in practice it would be unusual to have credal sets associated to all variables in a credal network — some distributions could be obtained with greater precision, and in any case the specification of dozens of credal sets is not an easy matter. We therefore introduced credal sets in all root nodes and in the node called *BatteryState*, using ε-contaminated models with $\varepsilon = 0.2$ [30]. The resulting strong extension has 3^{18} potential vertices (about 387 million potential vertices). We ran branch-and-bound inference for all states of the variable *Starts* and obtained exact values after evaluating 1,139,717 nodes (less than 0.3% of the number of potential vertices were explored). An interesting test was made with the branching strategy. We ran the same inference using a “reverse ordering” for branching; that is, we first expanded the credal sets that were farthest away from the query node. Using this strategy, the branch-and-bound algorithm found

[3] Microsof Research: http://www.research.microsoft.com/research/dtg/bnformat/autoxml.html.

Table 2: The probability error in underestimated approximate reasoning for E in the network of Figure 1.

# states per variable	# vertices per credal set	Fixed number of visited nodes	Mean relative error
03	03	150000	0.0013
03	03	30000	0.0067
04	02	200000	0.0061
04	02	50000	0.0097

the exact values after expanding 4,546,943 nodes. This simple test reinforces the intuition that the most relevant probability values in an inference are the values that are "close" to the query variable.

It is also possible to look at the branch-and-bound scheme not only as an exact algorithm, but also as an algorithm that can be stopped at any time to generate approximate results. We tested this idea by running the branch-and-bound algorithm with a fixed number of nodes. Table 2 shows the mean relative error in inferences (each row is the mean of ten random networks). The relative error is computed using the approximate and the exact values for $\overline{P}(E = e_0)$.

5 Inference with network fragments

If the credal network $\mathcal{N}$ is large, it may not be possible to run the branch-and-bound algorithm to optimality. In this section we propose strategies to handle such problems. The basic idea is to divide the credal network in parts and to run branch-and-bound in these sub-networks, in some suitable order. We illustrate this idea through an example.

Consider the network in Figure 1, with ternary variables and two vertices in each credal set. Suppose that we want to compute exact lower and upper probabilities for variable G and that our space and time constraints allow us to perform an exact inference just for E, but not for G. We then run branch-and-bound and obtain lower and upper probabilities for E. In a particular instance of the network shown in Figure 1, we obtained $p(e_0) \in [0.199; 0.587]$, $p(e_1) \in [0.084; 0.375]$ and $p(e_2) \in [0.212; 0.604]$. We can easily generate the largest credal set that is consistent with these intervals. We obtain $K(E)$ defined by the vertices

$$\{(0.413; 0.375; 0.212), (0.312; 0.084; 0.604), (0.587; 0.084; 0.329),$$

$$(0.199; 0.197; 0.604), (0.587; 0.201; 0.212), (0.199; 0.375; 0.426)\}.$$

Now we can remove E and its antecedents from the network, and replace E by a new node E' that has the marginal credal set of E as its marginal credal set.

The transformed network is displayed in Figure 3. We then run exact branch-and-bound based inference for G, obtaining the intervals $p(g_0) \in [0.091; 0.447]$, $p(g_1) \in [0.157; 0.564]$ and $p(g_2) \in [0.208; 0.591]$. Incidentally, we computed the same inferences with an exhautive algorithm in the JavaBayes system[4] and got the same values.

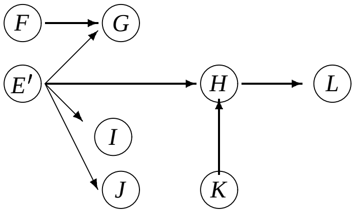


Figure 3: Transformed polytree credal network.

If inferences in the transformed credal network are still unfeasible, we can run an approximate inference algorithm in the transformed credal network. Consider running Tessem's algorithm in the network in Figure 3. We obtain the intervals $p(g_0) \in [0.053, 0.502]$, $p(g_1) \in [0.116, 0.663]$ and $p(g_2) \in [0.128, 0.644]$.

In closing, we note that Tessem's algorithm alone in the complete example network produced the intervals $p(g_0) \in [0.040, 0.524]$, $p(g_1) \in [0.106, 0.698]$ and $p(g_2) \in [0.097, 0.667]$.

6 Discussion

Any branch-and-bound algorithm is highly dependent on the quality of the bounds it employs. We have found that Tessem's bounds, while fast to compute and reasonably accurate, are quite wide — usually the search tree is expanded to a large depth before some of its branches are discarded. To give an example, in the computation of inferences for variable E in our samples with ternary variables and three vertices, the branch-and-bound algorithm explored the search tree almost completely down to levels 12 or 13 (the complete search tree has 21 levels).

As an aside, we have also implemented a breadth-first version of branch-and-bound [22], but we have found that the need to store the expanded frontier in such algorithms makes them unfeasible. Breadth-first branch-and-bound will only become a reality if better bounds than Tessem's are found.

Generally speaking, we can say that the branch-and-bound algorithm needs to explore a tiny fraction of potential vertices of the strong extension, and is faster than the best existing exact algorithms [5]. For really small credal networks (with

[4] Free software, site http://www.cs.cmu.edu/ javabayes.

a few thousand potential vertices in the strong extension), the overhead of branching and bounding can be significant, and in those cases enumeration algorithms may be faster.

Clearly, the branch-and-bound algorithm with Tessem bounds cannot cope with arbitrarily large problems, and it can face difficulties even in seemingly simple situations. In the network in Figure 1, inferences for variable L could not be found exactly, even after extensive tests.

7 Conclusion

This paper can be best understood as proposing a *family* of solutions for inference in strong extensions, using branch-and-bound algorithms as a unifying idea in such solutions. We have restricted ourselves to polytrees, but branch-and-bound techniques can be used for general inference; we have stressed the use of Tessem bounds, but any bounding scheme can be used.

We believe that our ideas are the first explicit formulation and implementation of inference in credal networks as a search procedure that runs to optimality. Branch-and-bound techniques are rather suitable for this purpose; the experiments show that inference with branch-and-bound and Tessem bounds is a definite improvement over existing algorithms.

We also would like to emphasize the possibility that a network is processed in pieces, using different levels of accuracy in each one of the partial inferences. Such a strategy seems to be appropriate for large networks. Our future research will be focused on developing and implementing general algorithms for decomposing networks and processing fragments with different strategies.

References

[1] Cozman, F.G.: Credal Networks. Artificial Intelligence **120** (2000) 199–233

[2] Couso, I. Moral, S., Walley, P.: Examples of Independence for Imprecise Probabilities. Proc. of the 1st ISIPTA, Ghent Belgium (1999) 121–130

[3] Pearl, J.: Probabilistic Reasoning in Intelligent Systems. Networks of Plausible Inference. Morgan Kaufmann Publishers, San Mateo CA (1988)

[4] Cozman, F.G.: Robustness Analysis of Bayesian Networks with Local Convex Sets of Distributions. Proc. of the 13th Annual Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA, (1997) 393-405

[5] Rocha, J.C.F.; Cozman, F.G.: Inference with Separately Specified Sets of Probabilities in Credal Networks. Proc. of the 18th Annual Confence on Un-

certainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA , (2002) 430-437

[6] Avriel, M.: Advances in Geometric Programming, Plenum Press, New York, 1980.

[7] Tessem, B.: Interval Probability Propagation. Int. Journal of Approximate Reasoning **7**, (1992) 95–120

[8] Levi, I. The Enterprise of Knowledge. MIT Press, Cambridge, Mass, (1980)

[9] Campos, L. M. de, Moral, S.: Independence Concepts for Convex Sets of Probabilities. Proc. of the XI Conference on Artificial Intelligence, Morgan Kaufmann, San Francisco, CA, (1995) 108–115

[10] Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London, (1991)

[11] Cano, J.E., Delgado, M., Moral, S.: An Axiomatic Framework for the Propagation of Uncertainty in Directed Acyclic Networks. Int. Journal of Approximate Reasoning **8**, (1993) 253–280.

[12] Zaffalon, M.: Inferenze e Decisioni in Condizioni di Incertezza con Modelli Grafici Orientati. Ph.D. Thesis, Università di Milano, Milan, Italy, (1997) (in Italian)

[13] Fagiuoli, E., Zaffalon, M. (1998).: 2U - An Exact Interval Propagation Algorithm for Polytrees with Binary Variables. Artificial Intelligence **106**(1), 77-107

[14] Ha, V.A. *et al*: Geometric Foundations for Interval-Based Probabilities. Annals of Mathematics and Artificial Intelligence, Vol.24, **1-4** (1998) 1–21

[15] Cano, A., Moral, S.: Using Probabilistic Trees to Compute Marginals with Imprecise Probabilities. TR-DECSAI-00-02-14, University of Granada, (2000)

[16] Cano, A., Moral, S.: A Genetic Algorithm to Approximate Convex Sets of Probabilities. 7th Int. Conf. IPMU-96, (Paris, France, July, 1994), (1994) pp 859–864

[17] Cano, A., Moral, S.: Convex Sets of Probabilities Propagation by Simulated Annealing. 5th Int. Conf. IPMU-94, (Paris, France, July, 1994), (1994) pp 4–8

[18] Andersen, K.A., Hooker, J.N.: Bayesian Logic. Decision Support Systems **11**, (1994) 191-210.

[19] Duffin R.J., Peterson, E.L.: Geometric Programming with Signomials. Journal of Optimization Theory and Applications, **11**(1), (1973) 3–35"

[20] Duffin, R.J., Peterson, E.L., Zener, C.: Geometric Programming, Theory and Application. John Willey and Sons, New York, (1967)

[21] Norvig, P., Russell, S.: Artificial Intelligence: a Modern Approach. Prentice Hall, Englewoods, (1995)

[22] Bertsekas, D.P.: Dynamic Programming, Deterministic and Stochastic Models. Prentice Hall, Englewoods, (1987)

[23] Papadimitriou, C., Steiglitz, I.: Combinatorial Optimization, Algorithms and Complexity. Prentice-Hall, Englewood Cliffs, New Jersey, (1982)

[24] Sahinidis, N.V.: BARON, Branch and Reduce Optimization Navigator, User's manual 4.0. University of Illinois at Urbana-Champaign, (2000)

[25] Preiss, B.R.: Data Structures and Algorithms with Object-Oriented Design Patterns in Java. Wiley, New York, (2000)

[26] Cozman, F.G.: Irrelevance and Independence in Quasi-Bayesian Networks. Proc. XIV Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA, (1998) 86–96

[27] Cozman, F.G.: Generalizing Variable Elimination in Bayesian Networks. Workshop on Probabilistic Reasoning in Artificial Intelligence, Editora Tec Art, São Paulo, (2000) 27–32

[28] Draper, D.L., Hanks, S.: Localized Partial Evaluation of Belief Networks. XV Conference on UAI, (1995) 170-177

[29] Ide, J.S., Cozman, F.G.: Random Generation of Bayesian Networks. Proc. of the XVI Braziliam Symposium on Artificial Intelligence, Springer-Verlag, (2002)

[30] Berger, J.O.: Statistical Decision Theory and Bayesian Analysis. Springer, Berlin, (1985)

José Carlos Ferreira da Rocha is a PhD student at the Engineering School (Escola Politécnica), University of São Paulo, and a teaching assistant at UEPG, Deinfo, Ponta Grossa, PR, Brasil, CEP 84030-900 E-mail: jrocha@uepg.br

Fabio Gagliardi Cozman is with the Engineering School (Escola Politécnica), University of São Paulo, São Paulo, SP, Brazil, CEP 05508-900. E-mail: fgcozman@usp.br

Extensions of Expected Utility Theory and Some Limitations of Pairwise Comparisons[*]

M. J. SCHERVISH
Carnegie Mellon University, USA

T. SEIDENFELD
Carnegie Mellon University, USA

J. B. KADANE
Carnegie Mellon University, USA

I. LEVI
Columbia University, USA

Abstract

We contrast three decision rules that extend Expected Utility to contexts where a convex set of probabilities is used to depict uncertainty: Γ-Maximin, Maximality, and *E*-admissibility. The rules extend Expected Utility theory as they require that an option is inadmissible if there is another that carries greater expected utility for each probability in a (closed) convex set. If the convex set is a singleton, then each rule agrees with maximizing expected utility. We show that, even when the option set is convex, this pairwise comparison between acts may fail to identify those acts which are Bayes for some probability in a convex set that is not closed. This limitation affects two of the decision rules but not *E*-admissibility, which is not a pairwise decision rule. *E*-admissibility can be used to distinguish between two convex sets of probabilities that intersect all the same supporting hyperplanes.

1 Introduction

This paper offers a comparison among three decision rules for use when uncertainty is depicted by a non-trivial, convex set of probability functions $\mathcal{P}$. This setting for uncertainty is different from the canonical Bayesian decision theory of expected utility, which uses a singleton set, just one probability function, to represent a decision maker's uncertainty. Justifications for using a non-trivial set of probabilities to depict uncertainty date back at least a half century [4] and a

[*]The research of the first three authors is supported by NSF grant DMS-0139911

foreshadowing of that idea can be found even in [7], where he allows that not all hypotheses may be comparable by qualitative probability – in accord with, e.g., the situation where the respective intervals of probabilities for two events merely overlap with no further (joint) constraints, so that neither of the two events is more, or less, or equally probable compared with the other.

We study decision rules that are extensions of canonical Subjective Expected Utility [SEU] theory, using sets of probabilities, in the following sense. The decision rules we consider all satisfy the following pairwise comparison between two options.

Criterion 1 *For a pair of options f and g, if for each probability $P \in \mathcal{P}$, f has greater expected utility than g, then g is inadmissible whenever f is available.*

This pairwise comparison itself creates a strict partial order. It (or a similar relation) has been the subject of representation theorems by, e.g., [3, 14, 15]. Note that when $\mathcal{P}$ is a singleton set, then the partial order is a weak order that satisfies SEU theory. In this sense, a decision rule that embeds this partial order extends SEU theory.

Here, we avail ourselves of four simplifying assumptions:

1. The decision maker's values for outcomes are determinate and are depicted by a (cardinal) utility function.

   Reason: We use circumstances under which convexity of $\mathcal{P}$ is not controversial. [1]

2. The algebra of uncertainty is finite, with finite state space $\Omega = \{\omega_1, \omega_2, ..., \omega_k\}$.

   Reason: We avoid the controversies surrounding countable versus finite additivity, which arise with infinite algebras.

3. Acts (or options) are *gambles*, i.e. functions from states to utilities, $f : \Omega \longrightarrow \mathbb{R}$.

   Reason: This assumption is commonplace and affords us an opportunity to contrast a variety of decision rules.

4. Each decision problem presents the decision maker a *uniformly bounded* choice set $\mathcal{A}$ of *gambles.*

   Reason: We avoid complications with unbounded utilities. Moreover, by considering the convex hull of a family of such gambles, we are assured of achieving the infimum and supremum operations with respect to expected utilities calculated with respect to the set $\mathcal{P}$.

[1] The issue of convexity of $\mathcal{P}$ is controversial. See [14] for a representation of partially ordered strict preferences that does not require convexity unless the decision maker has a determinate (cardinal) utility for outcomes. Rebuttal is presented in Section 7 of [11].

Of the three decision rules we discuss, perhaps the most familiar one is Γ-Maximin[2]. This rule requires that the decision maker ranks a gamble by its lower expected value, taken with respect to a closed, convex set of probabilities, $\mathcal{P}$, and then to choose an option from $\mathcal{A}$ whose lower expected value is maximum. This decision rule (as simplified by the assumptions, above) was given a representation in terms of a binary preference relation over Anscombe-Aumann horse lotteries [2], has been discussed by, e.g., Section 4.7.6 of [1] and recently by [5], who defend it as a form of Robust Bayesian decision theory. The Γ-Maximin decision rule creates a preference ranking of options independent of the alternatives available in $\mathcal{A}$: it is context independent in that sense. But Γ-Maximin corresponds to a preference ranking that fails the so-called (von Neumann-Morgenstern's) "Independence" or (Savage's) "Sure-thing" postulate of SEU theory. In Section 2 of [15], we note that such theories suffer from *sequential incoherence* in particular sequential decision problems.

The second decision rule that we consider, called *E*-admissibility ('*E*' for "expectation"), was formulated in [8, 9]. *E*-admissibility constrains the decision maker's admissible choices to those gambles in $\mathcal{A}$ that are Bayes for at least one probability $P \in \mathcal{P}$. That is, given a choice set $\mathcal{A}$, the gamble f is *E*-admissible on the condition that, for at least one $P \in \mathcal{P}$, f maximizes subjective expected utility with respect to the options in $\mathcal{A}$.[3] Section 7.2 of [12][4] defends a precursor to this decision rule in connection with cooperative group decision making. *E*-admissibility does not support an ordering of options, real-valued or otherwise, so that it is inappropriate to characterize *E*-admissibility by a ranking of gambles independent of the set $\mathcal{A}$ of feasible options. However, the distinction between options that are and are not *E*-admissible does support the "Independence" postulate. For example, if neither option f nor g is *E*-admissible in a given decision problem $\mathcal{A}$, then the convex combination, the mixed option $h = \alpha f \oplus (1\text{-}\alpha)g$ ($0 \leq \alpha \leq 1$) likewise is *E*-inadmissible when added to $\mathcal{A}$. This is evident from the basic SEU property: the expected utility of a convex combination of two gambles is the corresponding weighted average of their separate expected utilities; hence, for a given $P \in \mathcal{P}$ the expected utility of the mixture of two gambles is bounded above by the maximum of the two expected utilities. The assumption that neither of two gambles is *E*-admissible entails that their mixture has *P*-expected utility less than some *E*-admissible option in $\mathcal{A}$.

The third decision rule we consider is called *Maximality* by Walley in [17][5],

[2] When outcomes are cast in terms of a (statistical) loss function, the rule is then Γ-Minimax: rank options by their maximum expected risk and choose an option whose maximum expected risk is minimum.

[3] Levi's decision theory is lexicographic, in which the first consideration is *E*-admissibility, followed by other considerations, e.g. what he calls a Security index. Here, we attend solely to *E*-admissibility.

[4] Savage's analysis of the decision problem depicted by his Figure 1, p. 123, and his rejection of option *b*, p. 124 is the key point.

[5] There is, for our discussion here, a minor difference with Walley's formulation of Maximality

who appears to endorse it (p. 166). *Maximality* uses the strict partial order (above) to fix the admissible gambles from $\mathcal{A}$ to be those that are not strictly preferred by any other member of $\mathcal{A}$. That is, f is a *Maximal* choice from $\mathcal{A}$ provided that there is no other element $g \in \mathcal{A}$ that, for each $P \in \mathcal{P}$, carries greater expected utility than f does. *Maximality* (under different names) has been studied, for example, in [6, 8, 10, 13, 16]. Evidently, the E-admissible gambles in a decision problem are a subset of the Maximally admissible ones.

The three rules have different sets of admissible options. Here is a heuristic illustration of that difference.

Example 1 *Consider a binary-state decision problem,* $\Omega = \{\omega_1, \omega_2\}$*, with three feasible options. Option* f *yields an outcome worth 1 utile if state* ω_1 *obtains and an outcome worth 0 utiles if* ω_2 *obtains. Option g is the mirror image of* f *and yields an outcome worth 1 utile if* ω_2 *obtains and an outcomes worth 0 utiles if* ω_1 *obtains. Option h is constant in value, yielding an outcome worth 0.4 utiles regardless whether* ω_1 *or* ω_2 *obtains. Figure 1 graphs the expected utilities for these three acts. Let* $\mathcal{P} = \{P\colon 0.25 \leqslant P(\omega_1) \leqslant 0.75\}$*. The surface of Bayes solutions is highlighted. The expected utility for options* f *and* g *each has the interval of values [0.25, 0.75], whereas h of course has constant expected utility of 0.4. From the choice set of these three options* $\mathcal{A} = \{f, g, h\}$ *the* Γ*-Maximin decision rule determines that h is (uniquely) best, assigning it a value of 0.4, whereas* f *and g each has a* Γ*-Maximin value of 0.25. By contrast, under E-admissibility, only the option h is E-*inadmissible *from the trio. Either of* f *or g is E-admissible. And, as no option is strictly preferred to any other by expectations with respect to* $\mathcal{P}$*, all three gambles are admissible under* Maximality.

What normative considerations can be offered to distinguish among these three rules? For example, all three rules are immune to a Dutch Book, in the following sense:

Definition 1 *Call an option* favorable *if it is uniquely admissible in a pairwise choice against the status-quo of "no bet," which we represent as the constant 0.*

Proposition 1 *For each of the three decision rules above, no finite combination of favorable options can result in a Dutch Book, i.e., a sure loss.*

Proof. Reason indirectly. Suppose that the sum of a finite set of favorable gambles is negative in each state ω. Choose an element P from $\mathcal{P}$. The probability P is a convex combination of the extreme (0-1) probabilities, corresponding to a convex combination of the finite partition by states. The expectation of a convex

involving null-events. Walley's notion of Maximality requires, also, that an admissible gamble be classically admissible, i.e., not weakly dominated with respect to state-payoffs. This means that, e.g., our Theorem 1(i) is slightly different in content from Walley's corresponding result.

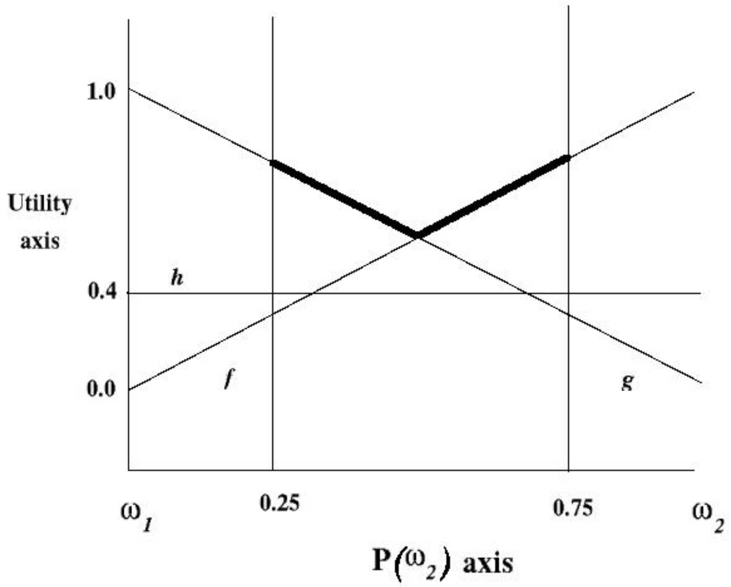


Figure 1: Expected utilities for three acts in Example 1. The thicker line indicates the surface of Bayes solutions.

combination of probabilities is the convex combination of the individual expectations. This makes the *P*-expectation of the sum of the finite set of favorable options negative. But the *P*-expectation of the sum cannot be negative unless at least one of the finitely many gambles has a negative *P*-expectation. Then that gamble cannot be favorable under any of the three decision rules. Thus, none of these three decision rules is subject to sure loss. □

In this paper, we develop an additional criterion for contrasting these decision rules. In Section 2 we address the question of what operational content the rules give to *distinguishing among different (convex) sets of probabilities*. That is, we are concerned to understand which convex sets of probabilities are treated as equivalent under a given decision rule. When do two convex sets of probabilities lead to all the same admissible options for a given decision rule? Γ-Maximin and Maximality are based solely on pairwise comparisons. Not so for *E*-admissibility. Even when the choice set $\mathcal{A}$ of feasible options is convex (e.g., closed under mixed strategies), these rules have distinct classes of admissible options.

2 Gambles and pairwise choice rules

It is evident that for Γ-Maximin generally to satisfy Criterion 1, the convex set of probabilities $\mathcal{P}$ must be closed. For an illustration why, if Example 1 is modified so that $\mathcal{P}' = \{P : 0.4 < P(\omega_1) \leqslant 0.75\}$ then, even though f and h both have the same infimum, 0.4, of expectations with respect to $\mathcal{P}'$, for each $P \in \mathcal{P}'$ f has greater expected utility than does h. Thus, from the perspective of operational content, the Γ-Maximin rule fails to distinguish between different convex sets of probabilities that differ with respect to Criterion 1, although each of Maximality and E-admissibility does distinguish the two sets $\mathcal{P}$ and $\mathcal{P}'$.

In order to contrast Maximality and E-admissibility, first we ask when do they lead to the same choices? Walley's Theorem 3.9.5 of [17] shows that, when the option space $\mathcal{A}$ is convex and the convex set of probabilities $\mathcal{P}$ is closed, the two rules are equivalent, i.e. both E-admissibility and Maximality reduce to a pairwise comparison of options according to Criterion 1. In this circumstance, an option is admissible, under either rule, just in case there is no other option that makes it inadmissible under Criterion 1. Then, with decision problems using convex sets of options, the two rules are capable of distinguishing between any two closed convex sets of probabilities, since distinct closed convex sets have distinct sets of supporting hyperplanes.

In Corollary 1 we re-establish Walley's result, and we extend the equivalence to decision problems in which $\mathcal{P}$ is open and $\mathcal{A}$ is finitely generated. The example following Theorem 1 establishes that for part (ii), the restriction to a finite (or finitely generated) option set, $\mathcal{A}$, is necessary. More important, however, we think is the second example following Theorem 1. That example is of a finite decision problem with a convex set of probabilities $\mathcal{P}$ (neither closed nor open) where, even though the option set is made convex, some Maximal options are not Bayes with respect to $\mathcal{P}$. Hence, even when the option space is convex, E-admissibility does not in general reduce to pairwise comparisons.

We preface Theorem 1 with a restatement of the structural assumptions for decision problems that we use in this paper. Let Ω be a finite state space with k states. Let $\mathcal{A}$ be a uniformly bounded collection of acts or gambles (real-valued functions from Ω). Let $\mathcal{C}$ be the convex hull of $\mathcal{A}$. For each probability vector $P = (p_1, \dots, p_k) \in \mathcal{P}$ and each $f \in \mathcal{C}$ there is a point $(p_1, \dots, p_{k-1}, E_p(f)) \in \mathbb{R}^k$, where $E_p(f) = \sum_{j=1}^{k} p_j f(\omega_j)$. For each $f \in \mathcal{C}$ there is a hyperplane that contains all of the points of the form $(p_1, \dots, p_{k-1}, E_p(f))$. For each $f \in \mathcal{C}$, the halfspace at or above its corresponding hyperplane is

$$\{x \in \mathbb{R}^k : \alpha_f^\top x \geq c_f\},$$

where

$$\alpha_f = (f(\omega_k) - f(\omega_1), \dots, f(\omega_k) - f(\omega_{k-1}), 1),$$

and $c_f = f(\omega_k)$.

Definition 2 *Let $\mathcal{P}$ be a convex set of probability vectors. We say that $f \in \mathcal{A}$ is* Bayes with respect to $\mathcal{P}$ *if there exists $p \in \mathcal{P}$ such that $E_p(f) \geq E_p(g)$ for all $g \in \mathcal{A}$.*

Theorem 1 *Let $\mathcal{B}$ be the set of all $f \in \mathcal{A}$ such that f is Bayes with respect to $\mathcal{P}$. Suppose that $g \in \mathcal{A} \setminus \mathcal{B}$. Assume either*

(i) that $\mathcal{P}$ is closed, or

(ii) that $\mathcal{A}$ is finite and that $\mathcal{P}$ is open. That is,

$$\{(p_1,\ldots,p_{k-1}) : (p_1,\ldots,p_k) \in \mathcal{P}\}$$

is an open subset of $\mathbb{R}^{k-1}$.

Then there exists h in the convex hull of $\mathcal{B}$ such that $E_p(h) > E_p(g)$ for all $p \in \mathcal{P}$.

Corollary 1 *Assume that $\mathcal{A}$ is closed and convex. Let $\mathcal{B}$ be the set of all $f \in \mathcal{A}$ such that f is Bayes with respect to $\mathcal{P}$. Suppose that $g \in \mathcal{A} \setminus \mathcal{B}$ is not Bayes with respect to $\mathcal{P}$. Assume either*

(i) that $\mathcal{P}$ is closed, or

(ii) that $\mathcal{A}$ is the convex hull of finitely many acts and that $\mathcal{P}$ is open.

Then there exists $h \in \mathcal{B}$ such that $E_p(h) > E_p(g)$ for all $p \in \mathcal{P}$.

The proofs of Theorem 1 and Corollary 1 rely on a series of results about convex sets and are given in Appendix A.

Example 2 *The following example illustrates that Theorem 1(ii) does not hold if $\mathcal{A}$ is allowed to be infinite. Let Ω have only $k = 2$ states. Let $\mathcal{A}$ consist of the gambles $\{f_\theta : 0 \leq \theta \leq \pi/4\}$ where*

$$f_\theta = (0.4 + 0.8\tan(\theta) - 0.2\sec(\theta), 0.4 - 0.2\tan(\theta) - 0.2\sec(\theta)).$$

Notice that $f_0 = (0.2, 0.2)$. Let

$$\mathcal{P} = \{(p_1, p_2) : p_1 > 0.2\}.$$

For each $p_1 \in (0.2, 0.3)$, the act f_θ is Bayes with respect to $\mathcal{P}$ when $\theta = 0.5\sin^{-1}(10[p_1 - 0.2])$. For $p_1 \geq 0.3$, $f_{\pi/4}$ is Bayes with respect to $\mathcal{P}$. Let $g = f_0$, which is not Bayes with respect to $\mathcal{P}$. Notice that, for every θ,

$$E_p(f_\theta) = (p_1 - 0.2)\tan(\theta) + 0.4 - 0.2\sec(\theta).$$

So, $E_p(f_\theta) < 0.2$ when $p_1 = 0.2$. Since $E_p(f_\theta)$ is a continuous function of p, $E_p(f_\theta) < 0.2$ for p in an open set around $(0.2, 0.8)$, which includes part of $\mathcal{P}$. It follows that every convex combination h of f_θ’s has $E_p(h) < 0.2$ somewhere inside of $\mathcal{P}$.

Example 3 *This example illustrates why we assume that $\mathcal{P}$ is closed in Theorem 1(i). Let Ω consist of three states. Let*

$$\begin{aligned}\mathcal{P} &= \{(p_1,p_2,p_3) : p_2 < 2p_1 \textit{ for } p_1 \leq 0.2\} \\ &\quad \cup\{(p_1,p_2,p_3) : p_2 \leq 2p_1 \textit{ for } 0.2 < p_1 \leq 1/3\}.\end{aligned}$$

The set of acts $\mathcal{A}$ contains only the following three acts (each expressed as a vector of its payoffs in the three states):

$$\begin{aligned}f_1 &= (0.2, 0.2, 0.2), \\ f_2 &= (1, 0, 0), \\ g &= (-1.8, 1.2, .2).\end{aligned}$$

Notice that $E_p(f_2)$ is the highest of the three whenever $p_1 \geq 0.2$, $E_p(f_1)$ is the highest whenever $p_1 \leq 0.2$, and $E_p(g)$ is never the highest. So, $\mathcal{B} = \{f_1, f_2\}$ and g is not Bayes with respect to $\mathcal{A}$. For each $0 \leq \alpha \leq 1$, we compute

$$\begin{aligned}E_p(\alpha f_1 + (1-\alpha) f_2) &= p_1(1-\alpha) + 0.2\alpha, \\ E_p(g) &= -2p_1 + p_2 + 0.2.\end{aligned}$$

Notice that $E_p(\alpha f_1 + (1-\alpha) f_2)$ is strictly greater than $E_p(g)$ if and only if $p_2 < (3-\alpha)p_1 - 0.2(1-\alpha)$. There is no α such that this inequality holds for all $p \in \mathcal{P}$.

Remark 1 *Note that is it irrelevant to this example that $p_2 = 0$ for some $p \in \mathcal{P}$.*

Definition 3 *Say that two convex sets* intersect all the same supporting hyperplanes *if they have the same closure and a supporting hyperplane intersects one convex set if and only if it intersects the other.*

In addition to showing that E-admissibility does not reduce to pairwise comparisons even when the option set is convex, this example also brings out the important point the E-admissibility (but not Maximality) can distinguish between some convex sets that intersect all the same supporting hyperplanes. As we noted some years ago (Section III of [15]), the strict preference relation induced by Criterion 1 cannot distinguish between pairs of convex sets that intersect all the same supporting hyperplanes. Of course, Γ-Maximin does even worse than Maximality, as it cannot distinguish open convex sets from their closure.

Figure 2 illustrates Example 3 and that the presence or absence of probability point $D = (0.2, 0.2, 0.4)$ determines whether or not act g is Bayes from the choice set $\mathcal{A} = \{f_1, f_2, g\}$. The closure of the convex set $\mathcal{P}$ is the triangle with extreme points $A = (1/3, 0, 2/3)$, $B = (1/3, 2/3, 0)$, and $C = (0, 0, 1)$. In Example 3, set $\mathcal{P}$ is the result of removing the closed line segment $[C, D]$ from the left face $[B, C]$ of the triangle ABC, leaving the half-open line segment $[B, D)$ along that face. The convex set $\mathcal{P}^*$ is the set of probabilities that results by adding point D to

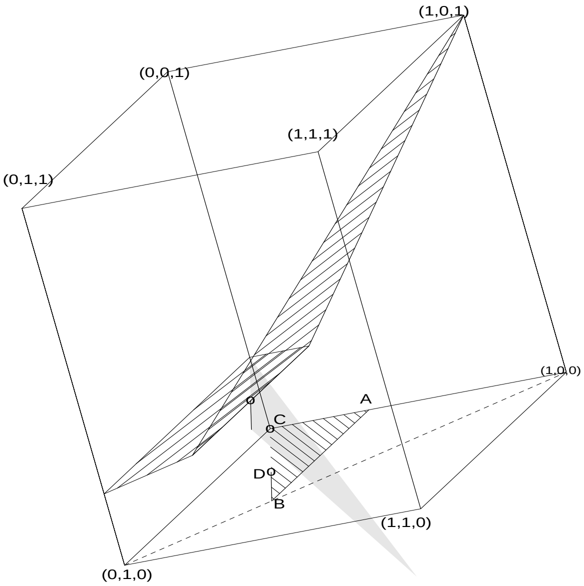


Figure 2: Illustration for Example 3. The set of (p_1, p_2) such that $(p_1, p_2, 1-p_1-p_2) \in \mathcal{P}$ is the diagonally shaded set inside the probability triangle at the bottom of the figure with the points *A*, *B*, *C*, and *D* that are discussed in the text labeled. The diagonally shaded surface is the surface of Bayes solutions for all probabilities (not just those in $\mathcal{P}$). The solid shaded set is $\{(p_1, p_2, E_p(g)) : p \in \mathcal{P}\}$. The points $(0.2, 0.4)$, $(0, 0)$, and $(0.2, 0.4, E_p(g))$ are indicated by open circles.

set $\mathcal{P}$. Point *D* then is an extreme but not exposed point in $\mathcal{P}^*$. Evidently, $\mathcal{P}$ and $\mathcal{P}^*$ intersect all the same supporting hyperplanes. Next, we indicate how to use *E*-admissibility to distinguish between these two convex sets of probabilities.

For this exercise, we bypass the details of what can easily be done with pairwise comparisons to fix the common boundaries of $\mathcal{P}$ and $\mathcal{P}^*$. Specifically, binary comparisons suffice to fix the closed interval [*A*,*B*] belongs to both sets, as the upper probability $\overline{P}(\omega_1)$ = 1/3; they suffice to fix that point *C* does *not* belong to either set, as the lower probability $\underline{P}(\omega_1) > 0$; they suffice to fix the half-open interval $[A,C)$ belongs to both sets, as the lower probability $\underline{P}(\omega_2) = 0$, and they suffice to fix the half open interval $[B,C)$ as a boundary for both sets, as the upper called-off (conditional) odds ratio $\overline{P}(\omega_1|\{\omega_1, \omega_2\}) > 1/3$. But pairwise comparisons according to Criterion 1, alone, cannot determine *how much* of the half-open interval $[B,C)$ belongs to either set $\mathcal{P}$ or $\mathcal{P}^*$. For that, we use non-binary choice problems and *E*-admissibility.

In order to establish that the half open line segment [*C*, *D*) does not belong to either set $\mathcal{P}$ or $\mathcal{P}^*$, consider the family of decision problems defined by the three-way choices: $\mathcal{A}_{-\varepsilon} = \{f_1, f_2, g_{-\varepsilon}\}$, where $g_{-\varepsilon}$ is the act with payoffs $(1.8, 1.2 - \varepsilon, 0.2)$. For each $\varepsilon > 0$, only the pair $\{f_1, f_2\}$ is *E*-admissible from such a three-way choice, with respect to each of the two convex sets of probabilities.

Likewise, in order to establish that the half-open line segment $(D, B]$ belongs to both sets, $\mathcal{P}$ and $\mathcal{P}^*$, consider the family of decision problems defined by the three-way choices: $\mathcal{A}_{+\varepsilon} = \{f_1, f_2, g_{+\varepsilon}\}$, where $g_{+\varepsilon}$ is the act with payoffs $(1.8, 1.2 + \varepsilon, .2)$. For each $\varepsilon > 0$, all three options are E-admissible with respect to each of the two convex sets of probabilities.

However, in the decision problem with options $\mathcal{A} = \{f_1, f_2, g\}$, as shown above, only the pair $\{f_1, f_2\}$ is *E*-admissible with respect to the convex set $\mathcal{P}$, whereas all three options are *E*-admissible with respect to the convex set $\mathcal{P}^*$.

By contrast, given a choice set, Maximality makes the same ruling about which options are admissible from that choice set, regardless whether convex set $\mathcal{P}$ or convex set $\mathcal{P}^*$ is used. That is, Maximality cannot distinguish between these two convex sets of probabilities in terms of admissibility of choices, as the two convex sets of probabilities intersect all the same supporting hyperplanes.

3 Summary

The discussion here contrasts three decision rules that extend Expected Utility and which apply when uncertainty is represented by a convex set of probabilities, $\mathcal{P}$, rather than when uncertainty is represented only by a single probability distribution. The decision rules are: Γ-Maximin, Maximality, and *E*-admissibility. We show that these decision rules have different operational content in terms of their ability to distinguish different convex sets of probabilities. When do the admissible choices differ for different convex sets of probabilities? Γ-Maximin is least sensitive among the three in this regard. We show that, even when the option set is convex, one decision rule (*E*-admissibility) distinguishes among more convex sets than the other two. This is because it alone among these three is not based on pairwise comparisons among options. The upshot it that it, but neither of the other two rules, can distinguish between two convex sets of probabilities that intersect all the same supporting hyperplanes.

A Proofs of Theorem 1 and Corollary 1

The proofs rely on some lemmas about convex sets.

Lemma 1 *Let k be a positive integer. Let C be a closed convex subset of $\mathbb{R}^k$ that contains the origin. There exists a unique closed convex subset D of $\mathbb{R}^{k+1}$ with the following properties:*

- $C = \{x \in \mathbb{R}^k : \alpha^\top x \geq c,$ *for all* $(\alpha, c) \in D\}$.
- $(\alpha, c) \in D$ *implies* $(a\alpha, ac) \in D$ *for all* $a \geq 0$,
- $(\alpha, c) \in D$ *implies* $(\alpha, c - a) \in D$ *for all* $a > 0$,

Also, for each $(\alpha, c) \in D$, $c \leq 0$.

Proof. To see that $(\alpha, c) \in D$ implies $c \leq 0$, let $\mathbf{0}$ be the origin. Then $\alpha^\top \mathbf{0} = 0 \geq c$. Define the following set

$$D_0 = \{(\alpha, c) : \alpha^\top x \geq c, \text{ for all } x \in C\}. \tag{1}$$

To see that D_0 is convex, let (γ_1, d_1) and (γ_2, d_2) be in D_0 and $0 \leq \beta \leq 1$. Then, for all $x \in C$,

$$(\beta\gamma_1 + [1 - \beta]\gamma_1)^\top x \geq \beta d_1 + (1 - \beta) d_2.$$

This means that $\beta(\gamma_1, d_1) + [1 - \beta](\gamma_2, d_2) \in D_0$, and D_0 is convex. To see that D_0 is closed, notice that $D_0 = \bigcap_{x \in C} D_x$, where $D_x = \{(\alpha, c) : \alpha^\top c \geq c\}$ and each D_x is closed. It is clear that D_0 has the last two properties in the itemized list. For the first condition, let E be the set defined in the first condition. It is clear that $C \subseteq E$. Suppose that there is $x_0 \in E$ such that $x_0 \not\in C$. Then there is a hyperplane that separates $\{x_0\}$ from C. That is, there is $\gamma \in \mathbb{R}^k$ and d such that $\gamma^\top x \geq d$ for all $x \in C$ and $\gamma^\top x_0 < d$. It follows that $(\gamma, d) \in D_0$, but then $x_0 \not\in E$, a contradiction.

To see that the set that satisfies the conditions is unique, suppose that D and F are both sets satisfying the listed conditions. If $F \neq D$, then there is (α, c) either in $D \setminus F$ or in $F \setminus D$. We will show, by way of contradiction, that neither of these cases can occur. The two cases are handled the same way. We will do only the first. In the first case, there is a hyperplane separating $\{(\alpha, c)\}$ from F. That is, there is (γ, d, f) with $\gamma \in \mathbb{R}^k$ and $d, f \in \mathbb{R}$ such that

$$\gamma^\top \delta + dg \geq f, \text{ for all } (\delta, g) \in F, \tag{2}$$

and $\gamma^\top \alpha + dc < f$. It follows that $a\gamma^\top \delta + da(g - b) \geq f$ for all $(\delta, g) \in F$ and all $a, b > 0$. As $a \to 0$, we see that $f \leq 0$ is required. As $b \to \infty$, we see that $d \leq 0$ is required. As $a \to \infty$ we see that $\gamma^\top \delta + dg \geq 0$ for all $(\delta, g) \in F$, hence we can assume that $f = 0$. Because $d, c \leq 0$ and $\gamma^\top \alpha + dc < 0$ it follows that $\gamma^\top \alpha < 0$ and there exists $d_0 < 0$ such that $\gamma^\top \alpha + d_0 c < 0$. Because $g \leq 0$ for all $(\delta, g) \in F$, we see that, even if $d = 0$, $\gamma^\top \delta + d_0 g \geq 0$ for all $(\delta, g) \in F$. Hence, we can assume that the separating hyperplane has the form $(\gamma, d_0, 0)$ with $d_0 < 0$. Define $\gamma_0 = \gamma/(-d_0)$. It follows from (2) that $\delta^\top \gamma_0 \geq g$ for all $(\delta, g) \in F$ and so $\gamma_0 \in C$. Hence $\alpha^\top \gamma_0 \geq c$ which contradicts $\gamma^\top \alpha + d_0 c < 0$. □

Lemma 2 *Let V be a closed convex subset of $\mathbb{R}^{k+1}$, and express elements of V as (α, d) where $\alpha \in \mathbb{R}^k$ and d is real. Define*

$$A = \{x \in \mathbb{R}^k : \alpha^\top x \geq d, \textit{ for all } (\alpha, d) \in V\}.$$

Assume that A is nonempty. Define D to be the set of all vectors in $\mathbb{R}^{k+1}$ of the form $(a\alpha, ad-b)$ with $a,b \geq 0$ and $(\alpha,d) \in V$. Then $D = \{(\alpha,d) \in \mathbb{R}^{k+1} : \alpha^\top x \geq d,$ for all $x \in A\}$.

Proof. Let $x_0 \in A$, and define

$$\begin{aligned} C &= \{x - x_0 : x \in A\}, \\ V' &= \{(\alpha, d - \alpha^\top x_0) : (\alpha,d) \in V\}. \end{aligned}$$

It follows that

$$C = \{x \in \mathbb{R}^k : \alpha^\top x \geq c, \text{ for all } (\alpha,c) \in V'\}, \tag{3}$$

and C contains the origin and is a closed convex set. Define $D_1 = \{(\alpha, d - \alpha^\top x_0) : (\alpha,d) \in D\}$. In other words, D_1 is the convex closed convex set of all vectors in $\mathbb{R}^{k+1}$ of the form $(a\alpha, ac-b)$ with $a,b \geq 0$ and $(\alpha,c) \in V'$. The definitions of D and D_1 were rigged so that D_1 satisfies all the conditions required of the set called D in Lemma 1 except possibly the first condition in the itemized list. To verify this condition, define

$$C' = \{x \in \mathbb{R}^k : \alpha^\top x \geq c, \text{ for all } (\alpha,c) \in D_1\}.$$

To see that $C \subseteq C'$, let $x \in C$. Then $a\alpha^\top x \geq ac - b$ for all $(\alpha,c) \in V'$ and all $a,b \geq 0$. Hence, $\alpha^\top x \geq c$ for all $(\alpha,c) \in D_1$. To see that $C' \subseteq C$, let $x \in C'$. Since $(\alpha,c) \in V'$ implies $(\alpha,c) \in D_1$, we have $\alpha^\top x \geq c$ for all $(\alpha,c) \in D_1$ and $x \in C$. It follows from Lemma 1 that D_1 is the set D_0 defined in (1) and D is the claimed set as well. □

Proof of Theorem 1. (i) Let

$$U = \left\{ x \in \mathbb{R}^k : \left(x_1, \ldots, x_{k-1}, 1 - \sum_{j=1}^{k-1} x_j \right) \in \mathcal{P} \right\}.$$

Let $\mathcal{C}'$ be the convex hull of $\mathcal{B}$. Let V consist of all points of the form (α_f, c_f) where $f \in \mathcal{C}'$. Let A be as defined in the statement of Lemma 2. Since g is not Bayes with respect to $\mathcal{P}$, the set $H_g = \{x \in U : \alpha_g^\top x = c_g\}$ does not intersect A. Now, notice that H_g and A are disjoint closed convex sets, hence there is a separating hyperplane. That is, there exists a nonzero $\gamma \in \mathbb{R}^k$ and c such that $\gamma^\top x \geq c$ for all $x \in A$ and $\gamma^\top y < c$ for all $y \in H_g$. Because $\gamma^\top x \geq c$ for all $x \in A$, it follows from Lemma 2 that (γ, c) is in the set D defined in the statement of Lemma 2. Hence, $\gamma = a\alpha$ and $c = ad - b$ for some $(\alpha, c) \in V$ and some $a,b \geq 0$. Because γ is nonzero, we have $a > 0$ and we can assume without loss of generality that $a = 1$ and $(\gamma, d-b) \in V$. So, $\gamma = \alpha_h$ for some $h \in \mathcal{C}'$ and $c = c_h - b$, and we can assume without loss of generality that $b = 0$ and $c = c_h$. Now, for all real t,

$$\alpha_h^\top (p_1, \ldots, p_{k-1}, t) = h(\omega_k) - E_p(h) + t = c_h - E_p(h) + t.$$

So, for all $x \in H_g$,

$$c_h > \alpha_h^\top x = c_h - E_p(h) + E_p(g).$$

It follows that, for all $p \in \mathcal{P}$, $E_p(h) > E_p(g)$.

(ii) Define U, $\mathcal{C}'$, V, A, and H_g exactly as in the proof of part (i). It is still true that H_g and A are convex, that A is closed, and that H_g does not intersect A. But H_g is now relatively open. That is, it is the intersection of the hyperplane $H'_g = \{x : \alpha_g^\top x = c_g\}$ with an open set. For this reason, H'_g is the unique hyperplane that contains H_g. Of course, if the closure of H_g fails to intersect A, the rest of the proof of part (i) continues to work. So, suppose that the closure $\overline{H_g}$ of H_g intersects A. Even so, there is a weakly separating hyperplane (γ, c), i.e., there is a $\gamma \in \mathbb{R}^k$ and c such that $\gamma^\top x \geq c$ for all $x \in A$ and $\gamma^\top y \leq c$ for all $y \in H_g$. We need to show that among all such separating hyperplanes, there is at least one such that the second inequality is strict, i.e., at least one of the separating hyperplanes fails to intersect H_g. Then the rest of the proof of part (i) will finish the proof.

Because $\mathcal{A}$ is finite, A is the intersection of finitely many closed halfspaces, and each of these halfspaces is of the form $\{x : \alpha_f^\top x \geq c_f\}$ for some $f \in \mathcal{B}$. Now, $\overline{H_g}$ intersects A in some convex subset of the union of the hyperplanes that determine these halfspaces. No subset of the union of finitely many distinct hyperplanes can be convex unless it is contained in the intersection of one or more of the hyperplanes. (Just check that $\alpha x + (1-\alpha)y$ is in the same hyperplane with x if and only if y is as well.) Hence, $A \cap \overline{H_g}$ is a subset of the intersection of one or more of the hyperplanes of the form $H'_f = \{x : \alpha_f^\top x = c_f\}$ for some $f \in \mathcal{B}$. Define

$$W = \{f \in \mathcal{B} : A \cap \overline{H_g} \subset H'_f\}.$$

If $W = \mathcal{B}$, then $H_g \subset A$, a contradiction. Let $f_0 \in \mathcal{B} \setminus W$ be such that H'_{f_0} is closest to $A \cap \overline{H_g}$. Such f_0 exists because $\mathcal{B}$ is finite. Let ε be one-half of the distance from H'_{f_0} to $A \cap \overline{H_g}$, and define

$$O = \{x : \|x - A \cap \overline{H_g}\| < \varepsilon\}.$$

Then

$$T = O \cap \left(\bigcap_{f \in W} \{x : \alpha_f^\top x \geq c_f\} \right) \subset A.$$

For each $f \in W$, define $M_f = \{x \in H_g : \alpha_f^\top x \geq c_f\}$. If at least one $M_f = \emptyset$, then H'_f fails to intersect H_g, and the proof is complete. So assume, to the contrary, that every $M_f \neq \emptyset$. Then for each f, the closure $\overline{M_f}$ of M_f contains $A \cap \overline{H_g}$. It follows that each M_f contains points in every neighborhood of $A \cap \overline{H_g}$, including O. Hence, for each f, there exists $x \in T \cap M_f$. Each such $x \in H_g \cap A$, a contradiction. □

Proof of Corollary 1. Let $\mathcal{C}$ be the convex hull of $\mathcal{B}$. Either assumption (i) or (ii) is strong enough to imply that Theorem 1 applies, hence there is $h' \in \mathcal{C}$ such that $E_p(h') > E_p(g)$ for all $p \in \mathcal{P}$. If $h' \notin \mathcal{B}$ let $h_1 = h'$, and apply Theorem 1 repeatedly in a transfinite induction as follows. At each successor ordinal $\gamma+1$, find $h_{\gamma+1} \in \mathcal{C}$ such that $E_p(h_{\gamma+1}) > E_p(h_\gamma)$ for all $p \in \mathcal{P}$. At a countable limit ordinal γ choose any countable sequence $\{\gamma_n\}_{n=1}^\infty$ of ordinals that is cofinal with γ. By the induction hypothesis, $E_p(h_{\gamma_i}) < E_p(h_{\gamma_j})$ for all $p \in \mathcal{P}$ if $i < j$. The sequence $\{h_{\gamma_n}\}_{n=1}^\infty$ belongs to the closed bounded set $\mathcal{A}$, hence it has a limit h_γ and

$$E_p(h_\gamma) = \lim_{n\to\infty} E_p(h_{\gamma_n}) = \sup_n E_p(h_{\gamma_n}),$$

for all p, and hence does not depend on which limit point we take. Also, $\sup_n E_p(h_\gamma) > E_p(h_\alpha)$ for all $\alpha < \gamma$, so we continue to satisfy the induction hypothesis. Since $\mathcal{A}$ is bounded, there cannot exist an uncountable increasing sequence of $E_p(h_\gamma)$ values, hence the transfinite induction terminates at some countable ordinal γ_0 with $h_{\gamma_0} \in \mathcal{B}$.

References

[1] Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*, 2 ed. Springer-Verlag, New York, 1985.

[2] Gilboa, I., and Schmeidler, D. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics 18* (1989), 141–153.

[3] Giron, F. J., and Rios, S. Quasi-Bayesian behavior: A more realistic approach to decision making? In *Bayesian Statistics*, e. a. J. M. Bernardo, Ed. University of Valencia Press, Valencia, 1980, pp. 17–38.

[4] Good, I. J. Rational decisions. *Journal of the Royal Statistical Society 14* (1952), 117–114.

[5] Grunwald, P. D., and Dawid, A. P. Game theory, maximum entropy, minimum discrepancy, and robust bayesian decision theory. Tech. Rep. 223, University College London, 2002.

[6] Herzberger, H. Ordinal preferences and rational choice. *Econometrica 41* (1973).

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

[8] Levi, I. On indeterminate probabilities. *Journal of Philosophy 71* (1974), 391–418.

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

[10] Levi, I. Imprecise and indeterminate probabilities. In *Proceedings of the 1st International Symposium on Imprecise Probabilities and Their Applications*. University of Ghent, Ghent, Belgium, 1999.

[11] Levi, I. Value commitments, value conflict, and the separability of belief and value. *Philosophy of Science 66* (1999), 509–533.

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

[13] Seidenfeld, T. Comment on “Probability, evidence, and judgment” by A. P. Dempster. In *Bayesian Statistics 2: Proceedings of the Second Valencia International Meeting, 1983*, e. a. J. M. Bernardo, Ed. North-Holland, Amsterdam, 1985, pp. 127–129.

[14] Seidenfeld, T., and Schervish, M. J.and Kadane, J. A representation of partially ordered preferences. *Annals of Statistics 23* (1995), 2168–2217.

[15] Seidenfeld, T., Schervish, M. J., and Kadane, J. Decisions without ordering. In *Acting and Reflecting*, W. Sieg, Ed. Kluwer Academic Publishers, Dordrecht, 1990, pp. 143–170.

[16] Sen, A. K. Social choice theory: a re-examination. *Econometrica 45* (1977), 53–89.

[17] Walley, P. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1990.

M. J. Schervish is with Carnegie Mellon University, USA

T. Seidenfeld is with Carnegie Mellon University, USA

J. B. Kadane is with Carnegie Mellon University, USA

I. Levi is with Columbia University, USA

Subjective Probability and Lower and Upper Prevision: A New Understanding[*]

G. SHAFER
Rutgers Business School, USA

P. R. GILLETT
Rutgers Business School, USA

R. B. SCHERL
Monmouth University, USA

Abstract

This article introduces a new way of understanding subjective probability and its generalization to lower and upper prevision. Instead of asking whether a person is willing to pay given prices for given risky payoffs, we ask whether the person believes he can make a lot of money at those prices. If not—if the person is convinced that no strategy for exploiting the prices can make him very rich in the long run—then the prices measure his subjective uncertainty about the events involved.

This new understanding justifies Peter Walley's updating principle, which applies when new information is anticipated exactly. It also justifies a weaker principle that is more useful for planning because it applies even when new information is not anticipated exactly. This weaker principle can serve as a basis for flexible probabilistic planning in event trees.

Keywords

subjective probability, upper and lower prevision, updating, event trees

1 Introduction

In the established understanding of subjective probability, set out by Bruno de Finetti [2] and his followers, a person's beliefs are revealed by the bets he is willing to make. The odds at which he is willing to bet define his probabilities.

We develop a somewhat different understanding of subjective probability, using Shafer and Vovk's game-theoretic framework [8]. In this framework, probability is understood in terms of two players: one who offers bets, and one to whom

[*] Research for this article was supported by NSF grant SES-9819116.

the bets are offered. We call these two players *House* and *Gambler*, respectively. The established understanding seems to be concerned with House’s uncertainty, since he is the one stating odds and offering to bet. But following Shafer and Vovk, we take Gambler’s point of view. Gambler is trying to beat the odds, and Shafer and Vovk’s work suggests that what makes odds expressions of a person’s uncertainty is his conviction that he cannot beat them.

We briefly introduce our new understanding of subjective probability in Section §2 and immediately generalize it to lower and upper prevision in Section §3.

2 Subjective Probability

Suppose House states odds $p : (1-p)$ and offers Gambler the opportunity to bet any amount he chooses for or against E at these odds. This means that House offers Gambler the payoff

$$\begin{cases} \alpha(1-p) & \text{if } E \text{ happens} \\ -\alpha p & \text{if } E \text{ fails} \end{cases} \tag{1}$$

for any real number α, which Gambler must choose immediately, before any other information becomes available. The absolute value of α is the total stakes for the bet, and the sign of α indicates which side Gambler is taking:

- If α is positive, then Gambler is betting on E happening. Gambler puts up αp, which he loses to House if E fails, while House puts up $\alpha(1-p)$, which he loses to Gambler if E happens. The total stakes are $\alpha p + \alpha(1-p)$, or α.
- If α is negative, then Gambler is betting against E happening. Gambler puts up $-\alpha(1-p)$, which he loses to House if E happens, while House puts up $-\alpha p$, which he loses to Gambler if E fails. The total stakes are $-\alpha(1-p) - \alpha p$, or $-\alpha$.

No principle of logic requires House to state odds at which Gambler can take either side. But mathematical probability has earned our attention by its practical successes over several centuries, and if we follow de Finetti in rejecting as defective all past attempts to provide objective interpretations of probability, then we seem to be left with (1) as the only viable way of interpreting this successful mathematical theory.

De Finetti developed this interpretation from the viewpoint of the player we are calling House. The principle that House should avoid sure loss to Gambler was fundamental to this development. If we agree that House should offer Gambler (1) for some p, then the principle that House should avoid sure loss leads immediately to the conclusion that p should be unique. If House offers (1) for both p_1 and p_2, where $p_1 < p_2$, then Gambler can accept the p_1-offer with $\alpha = 1$ and the p_2-offer with $\alpha = -1$, and this produces a sure gain of $p_2 - p_1$ for Gambler, no matter whether E happens or fails.

Protocols.

From a thoroughly game-theoretic point of view, the game between House and Gambler also involves a third player, who decides the outcomes on which they are betting. Calling this third player *Reality*, we can lay out an explicit protocol for the game in the style of Shafer and Vovk [8].

PROBABILITY FORECASTING

House announces $p \in [0,1]$.
Gambler announces $\alpha \in \mathbb{R}$.
Reality announces $x \in \{0,1\}$.
$\mathcal{K}_1 := \mathcal{K}_0 + \alpha(x-p)$.

This is a perfect-information protocol; the players move in the order indicated (not simultaneously), and each player sees the other players' moves as they are made. We have written $\mathcal{K}_0$ for Gambler's initial capital and $\mathcal{K}_1$ for his final capital. Reality's announcement indicates the happening or failure of E: $x = 1$ means E happens, and $x = 0$ means E fails. Thus $\alpha(x-p)$ is the same as (1). This is Gambler's net gain, which we can think of as the result of his paying αp for αx; Gambler buys α units of x for p per unit.

We may, for example, present de Finetti's argument for Additivity in this format. Consider the following protocol, where House gives probabilities for the three events E, F, and $E \cup F$, where E and F are disjoint:

MULTIPLE PROBABILITY FORECASTING

House announces $p_E, p_F, p_{E\cup F} \in [0,1]$.
Gambler announces $\alpha_E, \alpha_F, \alpha_{E\cup F} \in \mathbb{R}$.
Reality announces $x_E, x_F, x_{E\cup F} \in \{0,1\}$.
$\mathcal{K}_1 := \mathcal{K}_0 + \alpha_E(x_E - p_E) + \alpha_F(x_F - p_F) + \alpha_{E\cup F}(x_{E\cup F} - p_{E\cup F})$.

Constraint on Reality: Reality must make $x_{E\cup F} = x_E + x_F$ (this expresses the assumptions that E and F are disjoint and that $E \cup F$ is their disjunction).

The constraint on Reality is part of the rules of the game. Like the other rules, it is known to the players at the outset.

To see that House must make $p_{E\cup F} = p_E + p_F$ in order to avoid sure loss in this protocol, set

$$\delta := \begin{cases} 1 & \text{if } p_{E\cup F} > p_E + p_F \\ 0 & \text{if } p_{E\cup F} = p_E + p_F \\ -1 & \text{if } p_{E\cup F} < p_E + p_F \end{cases}$$

and consider the strategy for Gambler in which α_E and α_F are equal to δ and $\alpha_{E\cup F}$ is equal to $-\delta$. Gambler's net gain with this strategy is

$$\delta(x_E - p_E) + \delta(x_F - p_F) - \delta(x_{E\cup F} - p_{E\cup F}) = \delta(p_{E\cup F} - (p_E + p_F)),$$

which is positive unless $p_E + p_F = p_{E\cup F}$.

This argument readily generalizes to an argument for the rule that relates the expected value (or prevision) of a payoff to the probability of the events that determine the payoff.

Cournot's Principle.

The rules of probability can be derived from House's motivation to avoid sure loss. But a clear understanding of how subjective probabilities should be updated over time requires that we shift to Gambler's viewpoint and invoke Cournot's principle. When we assert that certain numbers are valid as objective probabilities, we are asserting that they do not offer anyone any opportunity to get very rich. When we advance them as our subjective probabilities, we are saying something only a little different: we are asserting that they do not offer us, with the knowledge we have, any opportunity to get very rich. When we say this, we put ourselves in the role of Gambler, not in the role of House. The point is not how we got the numbers: the point is what we think we can do with them.

A probability for a single event, if it is not equal to 0 or 1, can hardly be refuted. Even if Gambler chooses the winning side, with stakes high enough to make a lot of money, we will hesitate to conclude that the probability was wrong. Gambler may simply have been lucky. On the other hand, if House announces probabilities for a sequence of events, and Gambler consistently manages to make money, then the validity of the probabilities will be cast in doubt.

Shafer and Vovk [8] have shown that we can make this notion of testing precise within the following protocol, where House announces probabilities $p_1, p_2, \ldots$ for a series of events $E_1, E_2, \ldots$ with indicator variables $x_1, x_2, \ldots$:

```
SEQUENTIAL PROBABILITY FORECASTING
  $\mathcal{K}_0 := 1$.
  For $n = 1, 2, \ldots$:
    House announces $p_n \in [0,1]$.
    Gambler announces $\alpha_n \in \mathbb{R}$.
    Reality announces $x_n \in \{0,1\}$.
    $\mathcal{K}_n := \mathcal{K}_{n-1} + \alpha_n(x_n - p_n)$.
```

In this protocol, Gambler can test House's probabilities by trying to get infinitely rich ($\lim_{n\to\infty} \mathcal{K}_n = \infty$) without ever risking bankruptcy (without giving Reality an opportunity to make $\mathcal{K}_n$ negative for any n). If Gambler succeeds in doing this, he has refuted an infinite subset of the set of given probabilities.

Shafer and Vovk use the name *Cournot's principle* for the hypothesis that Reality will not allow Gambler to become infinitely rich without risking bankruptcy. This principle says that no matter what bankruptcy-free strategy for Gambler we specify (in addition to House's and Reality's previous moves, such a strategy may also use other information available to Gambler), we can be confident that Reality will move in such a way that the strategy will not make Gambler infinitely rich. This is an empirical hypothesis—a hypothesis about how Reality will behave, not a rule of the game.

If given probabilities satisfy Cournot's principle for any potential gambler, no matter how much information that gambler has, then we might call them *objective* or *causal* probabilities [5, 6]. On the other hand, if they satisfy Cournot's principle only for gamblers with a certain level of information, then we might call them *subjective* probabilities for that level of information. An individual who believes that the probabilities provided to him by some source or method do not permit any bankruptcy-free strategy to make him very rich might reasonably call them his personal subjective probabilities.

Under this interpretation, a person with subjective probabilities is not merely saying that he does not know how to get very rich betting at these probabilities. He is saying that he is convinced that there is no bankruptcy-free strategy that will make him very rich.

3 Subjective Lower and Upper Prevision

In recent decades, there has been great interest in supplementing subjective probability with more flexible representations of uncertainty. Some of the representations studied emphasize evidence rather than gambling [4, 9, 12]; others use a concept of partial possibility [3]. But many scholars prefer to generalize the story about betting that underlies subjective probability. The first step of such a generalization is obvious. Instead of requiring a person to set odds at which he will take either side of a bet, allow him to set separate odds for the two sides. This leads to lower and upper probabilities and lower and upper previsions rather than additive probabilities and expected values. See the early work of C. A. B. Smith [10, 11] and Peter Williams [17, 18, 19], the influential work of Peter Walley [13, 14, 15, 16], and the recent work of the imprecise probabilities project [1].

In this section, we look at lower and upper previsions from the point of view developed in the preceding section. This leads to a better understanding of how these measures of subjective uncertainty should change with new information, both when the new information is *exact* (i.e., when it is the *only* additional information) and when it is not.

Pricing Events and Payoffs.

Whereas probabilities for events determine expected values for payoffs that depend on those events (see §2), lower and upper probabilities are not so informative. The rates at which a person is willing to bet for or against given events do not necessarily determine the prices at which he is willing to buy or sell payoffs depending on those events. We need more than a theory of lower and upper probabilities for events: we need a theory of lower and upper previsions for payoffs.

3.0.1 Lower and Upper Probabilities

Suppose House expresses his uncertainty about E by specifying two numbers, p_1 and p_2. He offers to pay Gambler

$$-\alpha_1(x-p_1)=\begin{cases} -\alpha_1(1-p_1) & \text{if } E \text{ happens} \\ \alpha_1 p_1 & \text{if } E \text{ fails} \end{cases} \tag{2}$$

for any $\alpha_1 \geq 0$, and he also offers to pay Gambler

$$\alpha_2(x-p_2)=\begin{cases} \alpha_2(1-p_2) & \text{if } E \text{ happens} \\ -\alpha_2 p_2 & \text{if } E \text{ fails} \end{cases} \tag{3}$$

for any $\alpha_2 \geq 0$. In (2), Gambler sells α_1 units of x for p_1 per unit, while in (3), he buys α_2 units of x for p_2 per unit. Here is the protocol for this:

```
FORECASTING WITH LOWER AND UPPER PROBABILITIES
    House announces $p_1, p_2 \in [0,1]$.
    Gambler announces $\alpha_1, \alpha_2 \in [0,\infty)$.
    Reality announces $x \in \{0,1\}$.
    $\mathcal{K}_1 := \mathcal{K}_0 - \alpha_1(x-p_1) + \alpha_2(x-p_2)$.
```

To avoid sure loss, House must make $p_1 \leq p_2$. If $p_1 > p_2$, then Gambler can make money for sure by making α_1 and α_2 strictly positive and equal.

House would presumably be willing to increase his own payoffs by decreasing p_1 in (2) and increasing p_2 in (3). The natural remaining question is how high House will make p_1 and how low he will make p_2. We may call p_1 and p_2 *House’s lower and upper probabilities*, respectively, if House will not offer (2) for any value higher than p_1 and will not offer (3) for any value lower than p_2.

When we model our beliefs by putting ourselves in the role of House, we have some flexibility in the meaning we give our refusal to offer higher values of p_1 or lower values of p_2. Perhaps we are certain that we do not want to make additional offers, perhaps we are hesitating, or perhaps we are providing merely an incomplete model of our beliefs (Walley [14], pp. 61–63).

When we instead model our beliefs by putting ourselves in the role of Gambler, the question is what values of p_1 and p_2 we believe will satisfy Cournot’s principle. In the context of a sequence of forecasts, we might call p_1 and p_2 *Gambler’s lower and upper probabilities* when (1) Gambler believes that no strategy for buying and selling will make him very rich in the long run when he can sell x for p_1 or buy it for p_2 but (2) Gambler is not confident about this in the case where he is allowed to sell x for more than p_1 or buy it for less than p_2.

Clause (2) can be made precise in more than one way. Gambler might be unsure about whether he can get very rich with better values of p_1 or p_2, or he might believe that a strategy available to him would succeed with such values.

3.0.2 Lower and Upper Previsions

The following protocol allows us to price a payoff x that depends on the outcome of more than one event:

FORECASTING WITH LOWER AND UPPER PREVISIONS
House announces $p_1, p_2 \in \mathbb{R}$.
Gambler announces $\alpha_1, \alpha_2 \in [0, \infty)$.
Reality announces $x \in \mathbb{R}$.
$\mathcal{K}_1 := \mathcal{K}_0 - \alpha_1(x - p_1) + \alpha_2(x - p_2)$.

Again, Gambler is allowed to sell x for p_1 and buy it for p_2. If p_1 is the highest price at which Gambler can sell x (either the highest price House will offer or the highest price at which Gambler believes Cournot's principle, depending on our viewpoint), we may call it the *lower prevision* of x. Similarly, if p_2 is the lowest price at which Gambler can buy x, we may call it the *upper prevision* of x.

House may have more to say about x than the lower and upper previsions p_1 and p_2, and even the statement that these are lower and upper previsions is not exactly a statement about the protocol itself. We now turn to a more abstract approach, better suited to general discussion.

Forecasting in General.

Consider a set $\mathbf{R}$, and consider a set $\mathbf{H}$ of real-valued functions on $\mathbf{R}$. We call $\mathbf{H}$ a *belief cone* on $\mathbf{R}$ if it satisfies these two conditions:

1. If $\mathbf{g}$ is a real-valued function on $\mathbf{R}$ and $\mathbf{g}(\mathbf{r}) \leq 0$ for all $\mathbf{r} \in \mathbf{R}$, then $\mathbf{g}$ is in $\mathbf{H}$.
2. If $\mathbf{g}_1$ and $\mathbf{g}_2$ are in $\mathbf{H}$ and a_1 and a_2 are nonnegative numbers, then $a_1\mathbf{g}_1 + a_2\mathbf{g}_2$ is in $\mathbf{H}$.

We write $\mathcal{C}_{\mathbf{R}}$ for the set of all belief cones on $\mathbf{R}$.

Intuitively, a belief cone is a set of payoffs that House might offer Gambler. Thus if $(\alpha - \mathbf{g}) \in \mathbf{H}$, House is willing to buy $\mathbf{g}$ for α; and if $(\mathbf{g} - \alpha) \in \mathbf{H}$, House is willing to sell $\mathbf{g}$ for α. Condition 1 says that House will at least offer any contract that does not require him to risk a loss. Condition 2 says House will allow Gambler to combine any two of his offers, in any amounts. These conditions are, of course, closely related to Walley's concept of *desirability*.

The following abstract protocol is adapted from p. 90 of [8].

FORECASTING
Parameters: $\mathbf{R}$ and $\mathcal{C} \subseteq \mathcal{C}_{\mathbf{R}}$
Protocol:
House announces $\mathbf{H} \in \mathcal{C}$.
Gambler announces $\mathbf{g} \in \mathbf{H}$.
Reality announces $\mathbf{r} \in \mathbf{R}$.
$\mathcal{K}_1 := \mathcal{K}_0 + \mathbf{g}(\mathbf{r})$.

We call any protocol obtained by a specific choice of **R** and $\mathcal{C}$ a *forecasting protocol*. We call **R** the *sample space*.

We call a real-valued function on the sample space **R** a *variable*. House’s move **H**, itself a set of variables, determines lower and upper previsions for all variables. The *lower prevision* for a bounded variable x is

$$\underline{\mathbb{E}}_{\mathbf{H}}\, x := \sup\{\alpha \mid (\alpha - x) \in \mathbf{H}\}, \tag{4}$$

and the *upper prevision* is

$$\overline{\mathbb{E}}_{\mathbf{H}}\, x := \inf\{\alpha \mid (x - \alpha) \in \mathbf{H}\}. \tag{5}$$

These definitions are similar to those given by Walley ([14], pp. 64–65), with a difference in sign because Walley considers a collection $\mathcal{D}$ of payoffs that House is willing to accept for himself rather than a collection **H** that House offers.

The condition $(\alpha - x) \in \mathbf{H}$ in (4) means that Gambler can sell x for α. So roughly speaking, the lower prevision $\underline{\mathbb{E}}_{\mathbf{H}}\, x$ is the highest price at which Gambler can sell x. We say “roughly speaking” because (4) tells us only that Gambler can obtain $\alpha - x$ for α arbitrarily close to $\underline{\mathbb{E}}_{\mathbf{H}}\, x$, not that he can obtain $(\underline{\mathbb{E}}_{\mathbf{H}}\, x) - x$. Similarly, the upper prevision $\overline{\mathbb{E}}_{\mathbf{H}}\, x$ is roughly the lowest price at which Gambler can buy x.

Once we know lower previsions for all variables, we also know upper previsions for all variables, and vice versa, because

$$\overline{\mathbb{E}}_{\mathbf{H}}\, x = -\,\underline{\mathbb{E}}_{\mathbf{H}}(-x)$$

for every variable x. For additional general properties of lower and upper previsions, see Chapter 2 of Walley [14] and Chapters 1 and 8 of [8].

3.0.3 Regular Protocols

Given $\mathbf{H} \in \mathcal{C}_{\mathbf{R}}$, set

$$\mathbf{H}^* := \{x : \mathbf{R} \mapsto \mathbb{R} \mid \overline{\mathbb{E}}_{\mathbf{H}}\, x \leq 0\}.$$

The following facts can be verified straightforwardly:

- $\mathbf{H}^*$ is also a belief cone ($\mathbf{H}^* \in \mathcal{C}_{\mathbf{R}}$),
- $\mathbf{H} \subseteq \mathbf{H}^*$,
- $\overline{\mathbb{E}}_{\mathbf{H}}\, x = \overline{\mathbb{E}}_{\mathbf{H}^*}\, x$ and $\underline{\mathbb{E}}_{\mathbf{H}}\, x = \underline{\mathbb{E}}_{\mathbf{H}^*}\, x$ for every variable x, and
- $(\mathbf{H}^*)^* = \mathbf{H}^*$.

Intuitively, if House offers Gambler all the payoffs in **H**, then he might as well also offer the other payoffs in $\mathbf{H}^*$, because for every payoff in $\mathbf{H}^*$, there is one in **H** that is arbitrarily close.

We call a forecasting protocol *regular* if $\mathbf{H} = \mathbf{H}^*$ for every $\mathbf{H}$ in $\mathcal{C}$. Because any forecasting protocol can be replaced with a regular one with the same lower and upper previsions (enlarge each $\mathbf{H}$ in $\mathcal{C}$ to $\mathbf{H}^*$), little generality is lost when we assume regularity. This assumption allows us to remove the "roughly speaking" from the statements that the lower prevision of x is the highest price at which Gambler can sell x and the upper prevision the lowest price at which he can buy it. It also allows us to say that $\mathbf{H}$ is completely determined by its upper previsions (and hence also by its lower previsions):

$$x \in \mathbf{H} \text{ if and only if } \overline{\mathbb{E}}_{\mathbf{H}}\, x \leq 0.$$

The condition $x \in \mathbf{H}$ says that House will give x to Gambler. The condition $\overline{\mathbb{E}}_{\mathbf{H}}\, x \leq 0$ says that House will sell x to Gambler for 0 or less.

3.0.4 Interpretation

Both interpretations of lower and upper previsions we discussed in §3 generalize to forecasting protocols in general. We can put ourselves in the role of House and say that our beliefs are expressed by the prices we are willing to pay—our lower and upper previsions. Or, as we prefer, we can put ourselves in the role of Gambler and subscribe to these prices in the sense of believing that they will not allow us to become very rich in the long run, no matter what strategy we follow.

The reference to the long run in the second interpretation must be understood in terms of a sequential version of our abstract protocol. If we suppose, for simplicity, that Reality and House have the same choices of belief cones and payoffs on every move, this sequential protocol can be written as follows:

```
SEQUENTIAL FORECASTING
Parameters: $\mathbf{R}$ and $\mathcal{C} \subseteq \mathcal{C}_{\mathbf{R}}$
Protocol:
    $\mathcal{K}_0 := 1$.
    For $n = 1, 2, \ldots$:
        House announces $\mathbf{H}_n \in \mathcal{C}$.
        Gambler announces $\mathbf{g}_n \in \mathbf{H}_n$.
        Reality announces $\mathbf{r}_n \in \mathbf{R}$.
        $\mathcal{K}_n := \mathcal{K}_{n-1} + \mathbf{g}_n(\mathbf{r}_n)$.
```

The ambiguities we discussed in §3 also arise here. If we take House's point of view, we may or may not be categorical about our unwillingness to offer riskier payoffs than those in $\mathbf{H}_n$. If we take Gambler's point of view, we may be more or less certain about whether larger $\mathbf{H}_n$ would also satisfy Cournot's principle.

Walley's Updating Principle.

We turn now to Peter Walley's updating principle. This principle can be shown to entail the rule of conditional probability when it is applied to subjective probability. Here we apply it to our abstract framework for lower and upper previsions.

TWO-STAGE FORECASTING
Parameters: $\mathbf{R}$, a disjoint partition $\mathbf{B}_1,\ldots,\mathbf{B}_k$ of $\mathbf{R}$, $\mathcal{C} \subseteq \mathcal{C}_{\mathbf{R}}$
Protocol:

At time 0:
House announces $\mathbf{H}_0 \in \mathcal{C}$.
Gambler announces $\mathbf{g}_0 \in \mathbf{H}_0$.
Reality announces $i \in \{1,2,\ldots,k\}$.
At time t:
House announces $\mathbf{H}_t \in \mathcal{C}_{\mathbf{B}_i}$.
Gambler announces $\mathbf{g}_t \in \mathbf{H}_t$.
Reality announces $\mathbf{r}_t \in \mathbf{B}_i$.
$\mathcal{K}_t := \mathcal{K}_0 + \mathbf{g}_0(\mathbf{r}_t) + \mathbf{g}_t(\mathbf{r}_t)$.

Because we are considering here how House should make his second move, we leave this move unconstrained by the protocol. In Sections 3.0.4 and 3.0.4 below we consider two specific alternatives for this choice. Here, House can choose any belief cone on the reduced sample space $\mathbf{B}_i$.

Walley's updating principle says that if House knows at time 0 that Reality's announcement of i will be House's only new information when he moves at time t, then at time 0, as he makes his move $\mathbf{H}_0$, House should intend for his move $\mathbf{H}_t$ to be the belief cone $\mathbf{w}_t^i$ on $\mathbf{B}_i$ given by

$$\mathbf{w}_t^i := \{\mathbf{g} : \mathbf{B}_i \mapsto \mathbb{R} \mid \mathbf{g}^{\uparrow} \in \mathbf{H}_0\}, \tag{6}$$

where $\mathbf{g}^{\uparrow}$ is defined by

$$\mathbf{g}^{\uparrow}(\mathbf{r}) := \begin{cases} \mathbf{g}(\mathbf{r}) & \text{if } \mathbf{r} \in \mathbf{B}_i \\ 0 & \text{if } \mathbf{r} \notin \mathbf{B}_i. \end{cases} \tag{7}$$

In words: House should intend to offer a given payoff at the second stage after Reality announces i if and only if he is already offering that payoff at the first stage contingent on that value of i. This produces simple formulae relating the new lower and upper previsions to the old ones:

$$\underline{\mathbb{E}}_{\mathbf{w}_t^i}\, x = \sup\{\alpha \mid \underline{\mathbb{E}}_{\mathbf{H}_0}(x-\alpha)^{\uparrow} \geq 0\} \tag{8}$$

and

$$\overline{\mathbb{E}}_{\mathbf{w}_t^i}\, x = \inf\{\alpha \mid \overline{\mathbb{E}}_{\mathbf{H}_0}(x-\alpha)^{\uparrow} \leq 0\} \tag{9}$$

for every variable x on the reduced sample space $\mathbf{B}_i$.

Announcing Future Beliefs in Advance.

We now consider House's second move being constrained by announcing his future beliefs in advance. The rule of conditional probability can be shown to be mandated by the principle of House's avoiding sure loss when he announces future subjective probabilities in advance. What if House announces in advance future beliefs that determine only lower and upper previsions?

ADVANCE FORECASTING
Parameters: $\mathbf{R}$, a disjoint partition $\mathbf{B}_1, \ldots, \mathbf{B}_k$ of $\mathbf{R}$, $\mathcal{C} \subseteq \mathcal{C}_{\mathbf{R}}$.
Protocol:
At time 0:
House announces $\mathbf{H}_0 \in \mathcal{C}$ and $\mathbf{H}_t^j \in \mathcal{C}_{\mathbf{B}_j}$ for $j = 1, \ldots, k$.
Gambler announces $\mathbf{g}_0 \in \mathbf{H}_0$.
Reality announces $i \in \{1, 2, \ldots, k\}$.
At time t:
Gambler announces $\mathbf{g}_t \in \mathbf{H}_t^i$.
Reality announces $\mathbf{r}_t \in \mathbf{B}_i$.
$\mathcal{K}_t := \mathcal{K}_0 + \mathbf{g}_0(\mathbf{r}_t) + \mathbf{g}_t(\mathbf{r}_t)$.

Consider House's $\mathbf{H}_0$ and his $\mathbf{H}_t^j$ for some particular j. Suppose the variable $\mathbf{g}$ is in $\mathbf{H}_t^j$, but $\mathbf{g}^{\uparrow}$ is not in $\mathbf{H}_0$. Then it would make no difference in what Gambler can do if House were to enlarge $\mathbf{H}_0$ by adding $\mathbf{g}^{\uparrow}$ to it. He can already get the effect of $\mathbf{g}^{\uparrow}$ at time 0 by planning in advance to announce $\mathbf{g}$ at time t.

So we can assume, without changing what Gambler can accomplish, that if $\mathbf{g} \in \mathbf{H}_t^j$, then $\mathbf{g}^{\uparrow} \in \mathbf{H}_0$. This assumption implies $\mathbf{H}_t^j \subseteq \mathbf{w}_t^j$ by (6) and then

$$\underline{\mathbb{E}}_{\mathbf{H}_t^j} \leq \underline{\mathbb{E}}_{\mathbf{w}_t^j} \tag{10}$$

by (4). *The lower prevision at time t that is foreseen and announced at time* 0 *should not exceed the lower prevision given by Walley's updating principle.* Writing simply $\underline{\mathbb{E}}_0\, x$ for $\underline{\mathbb{E}}_{\mathbf{H}_0}\, x$ and $\underline{\mathbb{E}}_t\, x$ for $\underline{\mathbb{E}}_{\mathbf{H}_t^i}\, x$ (the lower previsions that House's time-0 announcements imply for time 0 and t, respectively) and recalling (8), we can write (10) in the form

$$\underline{\mathbb{E}}_t\, x \leq \sup\{\alpha \mid \underline{\mathbb{E}}_0 (x - \alpha)^{\uparrow} \geq 0\}, \tag{11}$$

where x is a variable on the reduced sample space $\mathbf{B}_i$.

The argument for (11) relies on the new viewpoint developed in this article, according to which a person's uncertainty is measured by prices he believes he cannot beat, not by prices he is disposed to offer. We expect (11) to hold because if it did not, the time 0 lower previsions would need to be increased to reflect stronger betting offers that Gambler cannot beat. Strictly speaking, of course, talk about Gambler not being able to beat given prices is talk about the long run, and so a complete exposition of the argument would involve a sequential protocol. We leave this further elaboration of the argument to the reader.

The argument does *not* rely on any assumption about exact information. Possibly House and Gambler will learn more than $\mathbf{B}_i$ by time t. $\underline{\mathbb{E}}_t\, x$, in (11), is not necessarily the lower prevision at time t. It is merely the lower prevision at time t to which House commits himself at time 0. This commitment does not exclude the possibility that House and Gambler will acquire additional unanticipated information and that House will thus offer Gambler more variables at time t than those to which he committed himself at time 0. In this case, the actual lower prevision for x at time t may come out higher than $\underline{\mathbb{E}}_{\mathbf{H}_t^i}\, x$ and even higher than $\underline{\mathbb{E}}_{\mathbf{w}_t^i}\, x$.

For planning at time 0, we are interested in what we can count on already at time 0. This is why the upper bound in (11) is interesting. When time t comes around, positive unanticipated information may lead us to give x a lower prevision exceeding this upper bound, but there is also the possibility of negative unanticipated information, and the upper bound can be thought of as telling us how conservative we need to be in our advance commitments in order to hedge against the possible negative information.

Updating with Exact Information.

Although the case in Section 3.0.4 above where commitments are made in advance in the face of possible *unanticipated* new information seems to us to have greater practical importance, it is also of interest to consider the case where new information is *anticipated exactly*. This is where Walley’s principle applies.

Extending the protocol of §3.0.4, we obtain the following sequential protocol:

```
SEQUENTIAL TWO-STAGE FORECASTING
  $\mathcal{K}_0 := 1$.
  For $n = 1, 2, \dots$
    At time $n$:
      House announces $\mathbf{H}_{n0} \in \mathcal{C}$.
      Gambler announces $\mathbf{g}_{n0} \in \mathbf{H}_{n0}$.
      Reality announces $i_n \in \{1, 2, \dots, k\}$.
    At time $n + 1/2$:
      House announces $\mathbf{H}_{n1} \in \mathcal{C}_{\mathbf{B}_{i_n}}$.
      Gambler announces $\mathbf{g}_{n1} \in \mathbf{H}_{n1}$.
      Reality announces $\mathbf{r_n} \in \mathbf{B}_{i_n}$.
    $\mathcal{K}_n := \mathcal{K}_{n-1} + \mathbf{g}_{n0}(\mathbf{r_n}) + \mathbf{g}_{n1}(\mathbf{r_n})$.
```

First, we make the following assumptions:

1. House’s $\mathbf{H}_{n0}$ satisfy Cournot’s principle.
2. House agrees in advance to follow Walley’s updating principle: $\mathbf{H}_{n1} = \mathbf{w}_n^{i_n}$, where $\mathbf{w}_n^j := \{\mathbf{g} : \mathbf{B}_j \mapsto \mathbb{R} \mid \mathbf{g}^{\uparrow} \in \mathbf{H}_{n0}\}$.
3. The only new information Gambler acquires between his move at time n and his move at time $n + 1/2$ is Reality’s choice of of i_n. (By the preceding assumption, he already knows House’s move $\mathbf{H}_{n1}$.)

4. Reality disregards Gambler's moves when she chooses her own moves.

Will all of House's announcements (the $\mathbf{H}_{n0}$ and $\mathbf{H}_{n1}$) satisfy Cournot's principle as a group? It is reasonable to conclude that they will. If they did not, then Gambler would have a bankruptcy-free strategy $\mathcal{S}$ that would make him infinitely rich. This strategy would specify $\mathbf{g}_{n0} \in \mathcal{C}$ for $n = 1,2,\ldots$ and $\mathbf{g}_{n1}^{j} \in \mathbf{w}_{n}^{j}$ for $n = 1,2,\ldots$ and $j = 1,\ldots,k$. Because Reality's moves do not depend on what Gambler does (Assumption 4) and House will follow Walley's recommendation for $\mathbf{H}_{n1}$ (Assumption 2), Gambler has a strategy $\mathcal{S}'$ for choosing the $\mathbf{g}_{n0}$ alone that makes his capital grow exactly as $\mathcal{S}$ does: to duplicate the effect of $\mathcal{S}$'s move $\mathbf{g}_{n1}$, he adds $(\mathbf{g}_{n1}^{j})^{\uparrow}$ to $\mathcal{S}$'s $\mathbf{g}_{n0}$ for $j = 1,\ldots,k$. This strategy does not require knowledge of i_n, and so Gambler would have the information needed to implement it (Assumption 3). So $\mathcal{S}'$ would also make Gambler infinitely rich, contradicting Assumption 1.

This result is a long-run justification for Walley's updating principle in its full generality.

4 Summary and Prospects

In this article, we set forth a new way of understanding probabilities and previsions in which we considered Gambler's viewpoint, and adopted Cournot's principle, in a series of game-theoretic protocols.

The proper handling of updating depends on whether we can exactly anticipate new information.

- We learned in §3.0.4 that if we can exactly anticipate new information—i.e., if we have an exhaustive advance list $\mathbf{B}_1,\ldots,\mathbf{B}_k$ of possibilities for exactly what all our new information will be, then we can follow Walley's updating principle, deriving new lower previsions from old ones using the formula

$$\underline{\mathbb{E}}_t\, x = \sup\{\alpha \mid \underline{\mathbb{E}}_0 (x-\alpha)^{\uparrow} \geq 0\}. \tag{12}$$

- We learned in §3.0.4 that if we cannot exactly anticipate new information, but we do know that we will learn which of the mutually exclusive events $\mathbf{B}_1,\ldots,\mathbf{B}_k$ has happened, and we commit ourselves in advance to lower previsions that depend on which $\mathbf{B}_i$ happens, then these preannounced lower previsions should satisfy the upper bound

$$\underline{\mathbb{E}}_t\, x \leq \sup\{\alpha \mid \underline{\mathbb{E}}_0 (x-\alpha)^{\uparrow} \geq 0\}. \tag{13}$$

The requirement of exact new information is very strong. The inequality (13) depends only on the weaker condition that we learn which of the $\mathbf{B}_1,\ldots,\mathbf{B}_k$ happens. There is no requirement that this be all we learn. On the other hand, the inequality only bounds the new lower prevision that can be guaranteed at the outset, at the planning stage. Unanticipated information may produce a higher lower prevision.

In this article, we have invoked Cournot’s principle using a relatively simple protocol, in which Reality has a binary choice at each step. This principle can also be adopted, however, when Reality sometimes has more than two choices, and when the choices available to her may depend on what she has done previously. This brings us to the generality of an event tree [5], offering additional flexibility that is needed in planning. Here it may be convenient to suppress the role of House in favor of a formal rule for determining the probabilities offered to Gambler, and to allow for unanticipated information and the refinement of beliefs. We explore these questions in [7].

References

[1] DE COOMAN, G., AND WALLEY, P. The imprecise probabilities project. `http://ippserv.rug.ac.be/home/ipp.html`.

[2] DE FINETTI, B. *Teoria Delle Probabilità*. Einaudi, Turin, 1970. An English translation, by Antonio Machi and Adrian Smith, was published as *Theory of Probability* by Wiley (London, England) in two volumes in 1974 and 1975.

[3] DUBOIS, D., AND PRADE, H. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York, 1988.

[4] SHAFER, G. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, New Jersey, 1976.

[5] SHAFER, G. *The Art of Causal Conjecture*. The MIT Press, Cambridge, Massachusetts, 1996.

[6] SHAFER, G. Causality and responsibility. *Cardozo Law Review 22* (2001), 101–123.

[7] SHAFER, G., GILLETT, P. R., AND SCHERL, R. B. A new understanding of subjective probability and its generalization to lower and upper prevision. *International Journal of Approximate Reasoning 33* (2003), 1–49.

[8] SHAFER, G., AND VOVK, V. *Probability and Finance: It’s Only a Game!* Wiley, New York, 2001. For more information on this book, see `http://www.cs.rhul.ac.uk/˜vovk/book/`.

[9] SMETS, P. Varieties of ignorance. *Information Sciences 57–58* (1991), 135–144.

[10] SMITH, C. A. B. Consistency in statistical inference and decision. *Journal of the Royal Statistical Society, Series B 23* (1961), 1–25.

[11] SMITH, C. A. B. Personal probabilities and statistical analysis. *Journal of the Royal Statistical Society, Series A 128* (1965), 469–499.

[12] SRIVASTAVA, R. P., AND MOCK, T. J., Eds. *Belief Functions in Business Decisions*. Springer Verlag, New York, 2002.

[13] WALLEY, P. Coherent lower (and upper) probabilities. Tech. Rep. 22, Department of Statistics, University of Warwick, Coventry, UK, 1981.

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

[15] WALLEY, P. Inferences from multinomial data: learning about a bag of marbles (with discussion). *Journal of the Royal Statistical Society, Series B 58* (1996), 3–57.

[16] WALLEY, P. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning 24* (2000), 125–148.

[17] WILLIAMS, P. M. Coherence, strict coherence and zero probabilities. In *Fifth International Congress of Logic, Methodology and Philosophy of Science* (1975), pp. VI, 29–33.

[18] WILLIAMS, P. M. Notes on conditional previsions. Research Report, School of Mathematical and Physical Sciences, University of Sussex, February 1975.

[19] WILLIAMS, P. M. Indeterminate probabilities. In *Formal Methods in the Methodology of Empirical Sciences*, M. Przełęcki, K. Szaniawski, and R. Wojcíckí, Eds. Ossolineum & Reidel, 1976, pp. 229–246.

Glenn Shafer is on the faculty of the Rutgers Business School—Newark and New Brunswick, NJ, USA. E-mail: gshafer@andromeda.rutgers.edu.

Peter R. Gillett is on the faculty of the Rutgers Business School—Newark and New Brunswick, NJ, USA. E-mail: gillett@business.rutgers.edu.

Richard B. Scherl is on the faculty of Monmouth University's Department of Computer Science, NJ, USA. E-mail: rscherl@monmouth.edu.

Products of Capacities Derived from Additive Measures

Extended abstract

DAMJAN ŠKULJ
University of Ljubljana, Slovenia

Abstract

A new approach to define a product of capacities is presented. It works for capacities that are in a certain relation with additive measures, most often this means that they are somehow derived from additive measures. The product obtained is not unique, but rather, lower and upper bound are given.

Keywords

products of capacities, capacities, non-additive measures, increasing capacities, distorted measures

1 Introduction

It is a well known fact that there is no straightforward unique way to generalize the product of additive measures to the non-additive case. Several approaches to define a product for a specific family of non-additive measures, also called capacities, have already been proposed (see [3, 4, 6]). In this paper a new approach is presented to define a product for a family of capacities related to additive measures. The product of capacities defined here is in a close relation with the product of the corresponding additive measures.

Let us first explain the terminology used in this paper. Let S be a nonempty set and $\mathcal{A}$ a σ-algebra of its subsets. A *capacity* is a monotone function $v\colon \mathcal{A} \to \mathbb{R}$, such that $v(\emptyset) = 0$ and $v(S) < \infty$. Additive measures used here are assumed to be finite and defined on the same algebras as the capacities. We will also use the standard terminology for the products in additive case. So $\mu \times \lambda$ will be the usual additive product of two additive measures μ and λ, and $\mathcal{A} \times \mathcal{B}$ will be the usual product algebra.

A *product* of capacities u and v on σ-algebras $\mathcal{A}$ and $\mathcal{B}$ respectively, is any capacity $w\colon \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ such that

$$w(A \times B) = u(A)v(B). \tag{1}$$

In additive measure theory the above condition uniquely determines the product of measures. Uniqueness crucially depends on additivity, moreover, without additivity requirement uniqueness of product can in general not be achieved. However, there always exist product capacities that satisfy (1), as shown in [4] where the lower and the upper bound are also given. But the set of all products if monotonicity of the product alone is required is far too big and their values differ too much on non-rectangular sets.

In order to reduce the set of all possible product capacities, the products are sought within some class of capacities with some additional properties that are preserved by multiplication. In [4] Hendon et al. define a product of *belief functions* using the idea of *Möbius representation* of capacities. Another definition of a product was proposed by Koshevoy in [6] using triangulation of geometrical realizations of distributive lattices. Denneberg in [3] joins both ideas to obtain a definition of a product for general monotone capacities which coincides with the Möbius product for the class of belief functions.

Instead of restricting to a special class of capacities Ghirardato in [5] restricts to a special class of functions for which the Fubini theorem for capacities holds. This class contains characteristic functions for a family of sets that he calls *comonotonic sets*. For these sets the double integral of their characteristic functions is a natural definition of a product of the capacities.

Although the existing definitions of products cover a very general class of capacities, most of them are still limited to discrete capacities. In this paper I present a definition of a product of capacities that seems to work better for continuous capacities, however, the results are valid for discrete case as well. The class of capacities it covers is rather restricted, but I think there are ways open to generalize this idea.

2 Increasing Capacities

The product of capacities defined here works for a family of capacities that are in a certain way related to additive measures. Before defining this relation, we will observe it in the case of a supermodular distorted measure. A capacity v is a *distorted measure* if it can be expressed as a composite $f \circ \mu$, where μ is an additive measure and the distortion f is an increasing real function with $f(0) = 0$. It is well known that a distorted measure is submodular or supermodular if the distortion is concave or convex respectively (see [2]). Suppose now that v is a supermodular distorted measure with distortion f applied to measure μ. Since f is a convex function, graph of a linear function intersects its graph in at most two different points. Using this fact, one can easily observe that for each pair of subsets $A \subseteq B$, $v(A)/\mu(A) \leq v(B)/\mu(B)$ holds. This leads to the next definition.

Definition 1 *Let μ be an additive measure on a σ-algebra $\mathcal{A}$ and v a capacity on the same algebra. The capacity v is* increasing *with respect to μ if the following* *is true: If $A \subseteq B$ and $\mu(A) > 0$ then $v(A)/\mu(A) \leq v(B)/\mu(B)$ and if $\mu(A) = 0$ then also $v(A) = 0$.*

If $\mu\colon \mathcal{A} \to \mathbb{R}$ is an additive measure, $I(\mu)$ will denote the set of all increasing capacities with respect to μ.

Further, we define quotient $m_v(A) := v(A)/\mu(A)$, where $m_v(A) = 0$ for all A with $v(A) = \mu(A) = 0$, and for each $t \in \mathbb{R}$, $\mathcal{A}_{v,t} := \left\{A \middle| \, t \leq \frac{v(A)}{\mu(A)}\right\}$. According to Definition 1, $m_v\colon \mathcal{A} \to \mathbb{R}$ is an increasing set function and it will be used to define the product of capacities. Thus, the product of two increasing capacities u and v will be defined by defining the corresponding $m_{u\times v}$.

We will also generalize the concept of comonotonicity for the case of increasing capacities. (For definition of comonotonicity for real functions see e.g. [2]). If v_1 and v_2 are capacities on a σ-algebra $\mathcal{A}$, increasing with respect to an additive measure μ, then we say v_1 and v_2 are *comonotonic* if the union $\{\mathcal{A}_{v_1,t}|t \in \mathbb{R}\}$ $\cup\{\mathcal{A}_{v_2,s}|s \in \mathbb{R}\}$ forms a chain of subsets of $\mathcal{A}$. Equivalently, capacities v_1 and v_2 are comonotonic exactly when m_{v_1} and m_{v_2} are comonotonic as real functions on $\mathcal{A}$ in the usual sense.

3 Products of Increasing Capacities

Given a set $C \in \mathcal{A} \times \mathcal{B}$, we will first define two Borel measurable sets in $\mathbb{R}^2$ whose Lebesgue measures are the minimum and the maximum value for the function $m_{u\times v}$. These sets can be considered as some kind of products of m_u and m_v.

Definition 2 *Let u and v be increasing capacities with respect to measures μ and λ respectively and defined on σ-algebras $\mathcal{A}$ and $\mathcal{B}$. Let $\mathcal{A} \times \mathcal{B}$ be the algebra of all measurable sets with respect to the product measure $\mu \times \lambda$. Define functions $\underline{\varphi}_{u,v}$ and $\overline{\varphi}_{u,v}\colon \mathcal{A} \times \mathcal{B} \to 2^{\mathbb{R}^2}$ with*

$$\begin{aligned}(x,y) \in \underline{\varphi}_{u,v}(C) \quad &\Longleftrightarrow \quad \textit{If there exist } A \in \mathcal{A}_{u,x} \textit{ and } B \in \mathcal{B}_{v,y}\\ &\qquad \textit{such that } A \times B \subseteq C, x > 0, y > 0\\ (x,y) \in \overline{\varphi}_{u,v}(C) \quad &\Longleftrightarrow \quad \textit{If for all } A \in \mathcal{A} \textit{ and } B \in \mathcal{B} \textit{ such that } A \times B \supseteq C\\ &\qquad A \in \mathcal{A}_{u,x} \textit{ and } B \in \mathcal{B}_{v,y} \textit{ holds, } x > 0, y > 0\end{aligned}$$

It is easy to see that $\underline{\varphi}_{u,v}(C)$ and $\overline{\varphi}_{u,v}(C)$ are Borel measurable sets in $\mathbb{R}^2$ for all $C \in \mathcal{A} \times \mathcal{B}$. However, there is a substantial asymmetry between both sets. While $\overline{\varphi}_{u,v}(C)$ is only a rectangle that represents the smallest rectangular set (with respect to m_u and m_v) that contains C, $\underline{\varphi}_{u,v}(C)$ is a union of rectangles representing the family of the largest rectangular sets that are contained in C. Clearly, the latter set therefore characterizes C much more precisely, in general, than the former one.

The definition of the lower and the upper bound for a product follows.

Definition 3 *Let u and v be increasing capacities with respect to measures μ and λ. We define the* lower product *of u and v as*

$$(\underline{u\times v})(C) = \mu_{\mathbb{R}^2}\left(\underline{\varphi}_{u,v}(C)\right)(\mu\times\lambda)(C)$$

and their upper product *as*

$$(\overline{u\times v})(C) = \mu_{\mathbb{R}^2}\left(\overline{\varphi}_{u,v}(C)\right)(\mu\times\lambda)(C).$$

The products $\underline{u\times v}$ and $\overline{u\times v}$ turn out to be the lower and the upper bound for a product of capacities under some additional natural assumptions. But first we state some properties of the products just defined.

Proposition 1 *The following statements hold for $u, u', u_i \in I(\mu)$ and $v \in I(\lambda)$.*

(i) If $u \leq u'$ then $\underline{u\times v} \leq \underline{u'\times v}$.

(ii) $\underline{(u+u')\times v} \leq \underline{u\times v} + \underline{u'\times v}$, equality holds if u and u' are comonotonic.

(iii) If $u_i \nearrow u$ then $\underline{u_i\times v} \nearrow \underline{u\times v}$.

and

(i)' If $u \leq u'$ then $\overline{u\times v} \leq \overline{u'\times v}$.

(ii)' $\overline{(u+u')\times v} = \overline{u\times v} + \overline{u'\times v}$

(iii)' If $u_i \nearrow u$ then $\overline{u_i\times v} \nearrow \overline{u\times v}$.

Because of symmetry of the product all of the above properties also hold for the second term.

The above properties also show that the upper and the lower product are not symmetric, as one might expect. While the upper product is additive, the lower is only comonotonically additive.

In order to prove that the lower and the upper product are indeed lower and upper bound in a family of product operators, we define operators $\underline{\Phi}$ and $\overline{\Phi}$: $I(\mu)\times I(\lambda) \to I(\mu\times\lambda)$ with $\underline{\Phi}(u,v) = \underline{u\times v}$ and $\overline{\Phi}(u,v) = \overline{u\times v}$.

Proposition 1 implies that the operators $\underline{\Phi}$ and $\overline{\Phi}$ are monotonic and continuous from below (in the sense of [2]) in both terms. The upper product operator $\overline{\Phi}$ is also biadditive, while the lower product operator $\underline{\Phi}$ is subadditive in both terms, however, when applied to sum of comonotonic capacities it is additive as well. Usually such operators are said to be *comonotonically additive*.

The following two theorems are the main results of this paper.

Theorem 1 *Let μ and λ be positive measures on σ-algebras $\mathcal{A}$ and $\mathcal{B}$ respectively. Let $\Phi\colon\, I(\mu)\times I(\lambda)\to I(\mu\times\lambda)$ be an operator that is comonotonically additive, positively homogeneous and continuous in both terms and such that*

$$\Phi(u,v)(A\times B)=u(A)v(B),$$

for all $A\in\mathcal{A}$ and $B\in\mathcal{B}$. Then $\underline{\Phi}\leq\Phi\leq\overline{\Phi}$ holds.

Proof Sketch. To prove this and also the next theorem, we define a family of simple increasing capacities that we call *cut measures*. Let $\mathcal{A}'\subseteq\mathcal{A}$ be a family of sets such that for each pair of sets $A\subseteq B$, $A\in\mathcal{A}'$ implies $B\in\mathcal{A}'$. Then we define cut measure $\mu|_{\mathcal{A}'}$ by

$$\mu|_{\mathcal{A}'}(A):=\begin{cases}\mu(A) & \text{if } A\in\mathcal{A}'\\ 0 & \text{otherwise}\end{cases}$$

The first step of proof is to verify that $\underline{u\times v}$ and $\overline{u\times v}$ are the smallest and the greatest product measures in case where u and v are cut measures, say $u=\mu|_{\mathcal{A}'}$ and $v=\lambda|_{\mathcal{B}'}$. It turns out that cut measures $(\mu\times\lambda)|_{\mathcal{C}'}$ and $(\mu\times\lambda)|_{\mathcal{C}''}$ are their smallest and largest product capacities increasing with respect to $\mu\times\lambda$, where $\mathcal{C}'=\{C|\text{ there exist } A\in\mathcal{A} \text{ and } B\in\mathcal{B} \text{ such that } A\times B\subseteq C\}$ and $\mathcal{C}''=\{C|\text{ if for all } A\times B\supseteq C,\ A\in\mathcal{A} \text{ and } B\in\mathcal{B} \text{ holds}\}$. These two cut measures turn out to be equal to $\underline{u\times v}$ and $\overline{u\times v}$ respectively.

The second step is to show that an increasing capacity can be uniformly approximated by sums of comonotonic cut measures. Using first step, comonotonic additivity and continuity of Φ we get the desired inequality. □

Next important property that a product should have is associativity.

Theorem 2 *Let u,v and w be increasing capacities, with respect to μ,λ and η. Then the following equalities hold:*

$$\underline{\underline{u\times v}\times w}=\underline{u\times\underline{v\times w}}=:\underline{u\times v\times w}$$

and

$$\overline{\overline{u\times v}\times w}=\overline{u\times\overline{v\times w}}=:\overline{u\times v\times w}.$$

The proof of this theorem also consists of two steps, the first being proof that it holds for the case of cut measures and the second one is extension to general case, using comonotonic additivity and continuity of Φ.

4 Conclusion

The results presented here, should be extended to more general families of capacities. One idea is to extend the product to differences of increasing measures.

That is, if capacities u and v can be written as $u = u_1 - u_2$ and $v = v_1 - v_2$, where u_1, u_2, v_1, v_2 are increasing with respect to some additive measure μ and λ respectively, an obvious way to extend present definition of the product would be, to define the lower product $\underline{u \times v} = \underline{u_1 \times v_1} + \underline{u_2 \times v_2} - \underline{u_1 \times v_2} - \underline{u_2 \times v_1}$. Such a definition unfortunately does not provide uniqueness of the product. A topic of further study is therefore searching for alternative generalizations.

The main disadvantage of the product defined here is, that it depends on the underlying additive measure. If we, on the other hand, modified the definition to allow all additive measures and apply minimum or maximum on it, we would probably obtain a trivial result. A compromise would be, to consider a proper family of additive measures. Such a family could depend on the type of considered capacities.

References

[1] A. Chateauneuf, Decomposable Capacities, Distorted Probabilities and Concave Capacities, Mathematical Social Sciences **31**, (1996), 19–37.

[2] D. Denneberg, Non-Additive Measure and Integral, Kluwer Academic Publishers, Dordrecht, 1997.

[3] D. Denneberg, Totally Monotone Core and Products of Monotone Measures, International Journal of Approximate Reasoning **24**, (2000), 273–281.

[4] E. Hendon, H. J. Jacobsen, B. Sloth and T. Trana.e.s, The Product of Capacities and Belief Functions, Mathematical Social Sciences **32**, (1996), 95–108.

[5] P. Ghirardato, On independence for non-additive measures, with a Fubini theorem, Journal of Economic Theory **73**, (1997), 261–291.

[6] G. A. Koshevoy, Distributive Lattices and Products of Capacities, Journal of Mathematical Analysis and Applications **219**, (1998), 427–441.

Damjan Škulj Faculty of Social Sciences, University of Ljubljana, Slovenia, E-Mail: Damjan.Skulj@Uni-lj.si

A Second-Order Uncertainty Model of Independent Random Variables: An Example of the Stress-Strength Reliability

L.V. UTKIN
University of Munich, Germany
St.Petersburg State Forest Technical Academy, Russia

Abstract

A second-order hierarchical uncertainty model of a system of independent random variables is studied in the paper. It is shown that the complex non-linear optimization problem for reducing the second-order model to the first-order one can be represented as a finite set of simple linear programming problems with a finite number of constraints. The stress-strength reliability analysis by unreliable information about statistical parameters of the stress and strength exemplifies the model. Numerical examples illustrate the proposed algorithm for computing the stress-strength reliability.

Keywords

stress-strength reliability, imprecise probabilities, second-order uncertainty, natural extension, previsions, linear programming

1 Introduction

By processing unreliable information, much attention have been focused on the *second-order uncertainty models* (*hierarchical uncertainty models*) due to their quite commonality. These models describe the uncertainty of a random quantity by means of two levels. Various second-order models and their applications can be found in the literature [4, 6, 7, 12, 13, 22], and a comprehensive review of hierarchical models is given in [5], where it is argued that the most common hierarchical model is the Bayesian one [2, 8, 14]. At the same time, the Bayesian hierarchical model is unrealistic in problems where there is available only partial information about the system behavior.

The main shortcoming of most proposed second-order hierarchical models (from the informational point of view) is the necessity to assume the certain type of the second-order probability or possibility distributions defined on the first-order level. This information is usually absent in many applications and additional assumptions may lead to some inaccuracy in results. The study of some tasks related to homogeneous second-order models without any assumptions about probability distributions has been illustrated by Kozine and Utkin [10]. However, these models are of limited use due to the homogeneity of gambles considered on the first-order level, i.e., the initial information is restricted by previsions of identical gambles. A new hierarchical uncertainty model for combining different types of evidence was proposed by Utkin [17, 16], where the second-order probabilities can be regarded as confidence weights and the first-order uncertainty is modelled by lower and upper previsions of different gambles [21]. However, the proposed model [17, 16] supposes that initial information is given only for one random variable. At the same time, many applications use a set of random variables described by a second-order uncertainty model, and it is necessary to find a model for some function of these variables. For example, reliability analysis demands to compute the reliability of a system under uncertain information about its components. An imprecise hierarchical model of a number of random variables has been studied by Utkin [18], but this model supposes that there is no information about independence of random variables. It should be noted that the condition of independence takes place in many applications. This condition makes the natural extension to be non-linear and, as a result, the corresponding hierarchical model becomes very complex.

An efficient approach to solve this problem is proposed in the paper. In order to show the practical relevance of the proposed approach, it is applied to the stress-strength reliability analysis by the independent stress and strength.

2 Imprecise Stress-Strength Reliability

A probabilistic model of structural reliability can be formulated as follows. Let Y represent a random variable describing the strength of a system and let X represent a random variable describing the stress or load placed on the system. System failure occurs when the stress on the system exceeds the strength of the system. Then the reliability of the system is determined as $R = \Pr\{X \leq Y\}$. A general approach to the structural reliability analysis based on the imprecise probability theory [11, 21, 23] was proposed in [19, 20]. Let us briefly consider this approach. Suppose that available information about the random stress X and the random strength Y is given as a set of n lower $\underline{\mathbb{E}}h_i$ and upper $\overline{\mathbb{E}}h_i$ previsions of gambles $h_i(X,Y)$ (unbounded gambles are considered in [15]) such that

$$\underline{\mathbb{E}}h_i \leq \mathbb{E}_{p(x,y)}\, h_i(X,Y) \leq \overline{\mathbb{E}}h_i,\ i = 1,...,n.$$

Here $p(x,y)$ is a joint density of the stress and strength. It is assumed that there exist a set of density functions such that linear previsions $\mathbb{E}_{p(x,y)}\, h_i$ can be regarded as expectations of h_i. Taking into account that

$$R = \Pr\{X \leq Y\} = \mathbb{E}_{p(x,y)}\, I_{[0,\infty)}(Y - X), \tag{1}$$

we can write the following optimization problems (natural extension) for computing the lower $\underline{R}$ and upper $\overline{R}$ stress-strength reliability as follows:

$$\underline{R}\,\langle\overline{R}\rangle = \inf_{p}\left\langle \sup_{p}\right\rangle \int_{\mathbb{R}_+^2} I_{[0,\infty)}(y-x)\,p(x,y)\mathrm{d}x\mathrm{d}y, \tag{2}$$

subject to

$$\underline{\mathbb{E}}h_i \le \int_{\mathbb{R}_+^2} h_i(x,y)\,p(x,y)\mathrm{d}x\mathrm{d}y \le \overline{\mathbb{E}}h_i,\ i=1,...,n. \tag{3}$$

Here the infimum and supremum are taken over the set of all possible densities $\{p(x,y)\}$ satisfying conditions (3), $I_{[0,\infty)}(Y-X)$ is the indicator function taking the value 1 if $Y \ge X$ and 0 otherwise. If random variables X and Y are independent, then the constraint $p(x,y) = p_X(x)p_Y(y)$ is added to constraints (3), where p_X and p_Y are densities of X and Y, respectively.

The natural extension is a powerful tool for analyzing the reliability on the basis of available partial information. However, it has a shortcoming. Let us imagine that two experts provide the following judgements about the stress: (i) mean value of the stress is not greater than 10; (ii) mean value of the stress is not less than 10 hours. The natural extension produces the resulting mean value $[0,10]\cap[10,\infty) = 10$. In other words, the absolutely precise measure is obtained from too imprecise initial data. This is unrealistic in practice of reliability analysis. The reason of such results is that probabilities of judgements are assumed to be 1. If we assign some different probabilities to judgements, then we obtain more realistic assessments. For example, if the belief to each judgement is 0.5, then, according to [9], the resulting mean value is greater than 5 hours. Therefore, in order to obtain the accurate and realistic reliability assessments, it is necessary to take into account some vagueness of information about characteristics of the stress and strength.

3 Second-Order Model. Problem Statement

Suppose that there exist n judgements about the stress X:

$$\mathbb{E}f_j(X) \in T_j = [\underline{t}_j, \overline{t}_j],\ j=1,...,n,$$

and l judgements about the strength Y:

$$\mathbb{E}h_j(Y) \in S_j = [\underline{s}_j, \overline{s}_j],\ j=1,...,l.$$

Here f_j and h_j are gambles corresponding to the available judgements about X and Y. Moreover, it is known that

$$\underline{\alpha}_j \le \Pr\{\mathbb{E}f_j \in T_j\} \le \overline{\alpha}_j,\ j=1,...,n,$$
$$\underline{\beta}_j \le \Pr\{\mathbb{E}h_j \in S_j\} \le \overline{\beta}_j,\ j=1,...,l.$$

The second-order probabilities $[\underline{\alpha}_j, \overline{\alpha}_j]$ and $[\underline{\beta}_j, \overline{\beta}_j]$ are interpreted as a model for uncertainty about "correct" values of partially known measures of X and Y. Let us briefly discuss the sense of beliefs to the expert judgements. If we know that an expert provides $100 \cdot \alpha\%$ of "correct" judgements, this means that, by giving finitely many intervals, say n, for an unknown parameter, approximately $n\alpha$ intervals cover some "correct" value of the parameter. But if we have only the $(n+1)$-st interval and do not know anything about previous n intervals, then we can only say that the "correct" value of the parameter lies in this interval with probability α and outside this interval with probability $1-\alpha$. If we would have all aforementioned n intervals, some probability distribution of the parameter could be constructed and well-known Bayesian methods could be used. In this case, there is no need to apply imprecise probabilities.

The term "expert information" may be used in a more general sense. In particular, confidence intervals of parameters elicited as a result of statistical inference with corresponding confidence probabilities may be regarded as "beliefs to experts". For example, if we have one confidence interval for the expectation of a probability distribution, then we can only assert, that the "correct" value of the expectation is in the interval with the confidence interval probability $[\alpha, 1]$ and outside the confidence interval with the probability $[0, 1-\alpha]$.

How to find average values of $\underline{R}$ and $\overline{R}$, i.e., to reduce the second-order model to the first-order one? Roughly speaking, if we have second-order probabilities defined for different intervals of $\mathbb{E}f_j$ and $\mathbb{E}h_j$, then there exist a set of second-order distributions of $\mathbb{E}f_j$, $\mathbb{E}h_j$, and $\mathbb{E}I_{[0,\infty)}(Y-X)$ produced an interval of lower and upper expectations of $\mathbb{E}I_{[0,\infty)}(Y-X)$, i.e., $\underline{R}$ and $\overline{R}$. We will call this interval "average" to distinguish expectations (previsions) on the first and second levels of the considered second-order uncertainty model. In fact, the "average" interval allows us to get rid of the more complex second-order model and to deal with the first-order model. This problem is especially difficult if the stress and strength are independent. At that, a special type of independence called by the free product [11] is studied in the paper. This type of independence is like to the epistemic irrelevance [3] and, generally, is asymmetric.

In order to give the reader the essence of the subject analyzed and make all the formulas more readable, we will mainly consider only the lower bound $\underline{R}$.

Let $v_i = \mathbb{E}f_i$ and $w_i = \mathbb{E}h_i$ be values of random variables V_i and W_i defined on sample spaces Ω_i and Λ_i, respectively. Let $V = (V_1, ..., V_n)$, $W = (W_1, ..., W_n)$ and $\mathbf{V} = (v_1, ..., v_n)$, $\mathbf{W} = (w_1, ..., w_l)$ be the vectors of random variables V_i, W_i and their values, respectively. Denote $N = \{1, ..., n\}$ and $L = \{1, ..., l\}$. Then the natural extension for computing $\underline{R}$ can be written as a sequence of lower expectations:

$$\underline{R} = \underline{\mathbb{E}}^{\mathbf{W}} \left\{ \underline{\mathbb{E}}^{\mathbf{V}|\mathbf{W}} \left(\mathbb{E}I_{[0,\infty)}(Y-X) \right) \right\}$$

by given lower and upper previsions

$$\underline{\mathbb{E}}I_{T_i}(v_i) = \underline{\alpha}_i,\ \overline{\mathbb{E}}I_{T_i}(v_i) = \overline{\alpha}_i,\ \underline{\mathbb{E}}I_{S_i}(w_i) = \underline{\beta}_i,\ \overline{\mathbb{E}}I_{S_i}(w_i) = \overline{\beta}_i. \tag{4}$$

By introducing a random variable Z having values $z(\mathbf{V},\mathbf{W}) = \mathbb{E}I_{[0,\infty)}(Y-X)$ and assuming that there exists a set of densities $\varphi(\mathbf{V})$ and $\psi(\mathbf{W})$ of vectors V and W, respectively, we can write

$$\underline{R} = \inf_{\psi} \int_{\Lambda} \left(\inf_{\varphi} \int_{\Omega} z(\mathbf{V},\mathbf{W})\varphi(\mathbf{V})\mathrm{d}\mathbf{V} \right) \psi(\mathbf{W})\mathrm{d}\mathbf{W}, \tag{5}$$

subject to

$$\underline{\alpha}_i \leq \int_{\Omega} I_{T_i}(v_i)\varphi(\mathbf{V})\mathrm{d}\mathbf{V} \leq \overline{\alpha}_i,\ i \in N,\ \underline{\beta}_i \leq \int_{\Lambda} I_{S_i}(w_i)\psi(\mathbf{W})\mathrm{d}\mathbf{W} \leq \overline{\beta}_i,\ i \in L. \tag{6}$$

Here $\Omega = \Omega_1 \times \cdots \times \Omega_n$, $\Lambda = \Lambda_1 \times \cdots \times \Lambda_l$. The sample spaces Ω_i and Λ_j are determined by sets of values $\mathbb{E}f_i$ and $\mathbb{E}h_j$, i.e.,

$$\Omega_i = [\inf \mathbb{E}f_i, \sup \mathbb{E}f_i],\ \Lambda_j = [\inf \mathbb{E}h_j, \sup \mathbb{E}h_j].$$

A dual optimization problem can not be written as it has been made in [18] because the initial problem is non-linear. Our aim is to find $\underline{R}$, i.e., to solve (5)-(6).

4 Solution of Problem (5)-(6)

A Set of Linear Programming Problems.

Let $\mathbf{W}^* = (w_1^*,...,w_n^*) \in \Lambda$ be a realization of the vector $\mathbf{W}$. Denote $R(\mathbf{W}^*) = \mathbb{E}_{\varphi}z(\mathbf{V},\mathbf{W}^*)$. Problem (5)-(6) can be represented as follows:

$$\begin{aligned} \underline{R} &= \inf_{\psi} \int_{\Lambda} \left(\inf_{\varphi} \int_{\Omega} z(\mathbf{V},\mathbf{W})\varphi(\mathbf{V})\mathrm{d}\mathbf{V} \right) \psi(\mathbf{W})\mathrm{d}\mathbf{W} \\ &= \inf_{\psi} \int_{\Omega} \inf_{\varphi} R(\mathbf{W}^*)\psi(\mathbf{W})\mathrm{d}\mathbf{W} = \inf_{\psi} \int_{\Omega} \underline{R}(\mathbf{W}^*)\psi(\mathbf{W})\mathrm{d}\mathbf{W}, \end{aligned} \tag{7}$$

subject to

$$\underline{\beta}_i \leq \mathbb{E}_{\psi} I_{S_i}(w_i) \leq \overline{\beta}_i,\ i \in L. \tag{8}$$

Here

$$\underline{R}(\mathbf{W}^*) = \inf_{\varphi} \mathbb{E}_{\varphi} z(\mathbf{V},\mathbf{W}^*), \tag{9}$$

subject to

$$\underline{\alpha}_i \leq \mathbb{E}_{\varphi} I_{T_i}(v_i) \leq \overline{\alpha}_i,\ i = 1,...,n. \tag{10}$$

Problems (7)-(8) and (9)-(10) are linear and dual optimization problems can be written, i.e., we have a set of the following problems for each $\mathbf{W}^* \in \Lambda$:

$$\underline{R}(\mathbf{W}^*) = \sup \left(c_0 + \sum_{i \in N} (c_i \underline{\alpha}_i - d_i \overline{\alpha}_i) \right), \tag{11}$$

subject to $c_i, d_i \in \mathbb{R}_+$, $c_0 \in \mathbb{R}$, $i \in N$, and $\forall \mathbf{V} \in \Omega$,

$$c_0 + \sum_{i \in N} (c_i - d_i)\, I_{T_i}(v_i) \leq z(\mathbf{V}, \mathbf{W}^*), \tag{12}$$

and one linear programming problem

$$\underline{R} = \sup \left(c_0 + \sum_{i \in L} \left(c_i \underline{\beta}_i - d_i \overline{\beta}_i \right) \right), \tag{13}$$

subject to $c_i, d_i \in \mathbb{R}_+$, $c_0 \in \mathbb{R}$, $i \in L$, and $\forall \mathbf{W}^* \in \Lambda$,

$$c_0 + \sum_{i \in L} (c_i - d_i)\, I_{S_i}(w_i) \leq \underline{R}(\mathbf{W}^*). \tag{14}$$

The dual problems have been introduced in order to get rid of densities $\varphi(\mathbf{V})$ and $\psi(\mathbf{W})$.

Solution of Problem (11)-(12).

An algorithm and an approach to solving a problem similar to (11)-(12) are given in [16, 17]. But problem (11)-(12) has some difference. To solve this problem, it is necessary to define what $z(\mathbf{V}, \mathbf{W}^*)$ is.

Let J be a set of indices and $J \subseteq N$. Introduce the following sets of constraints:

$$\mathcal{T}_J = \{T_i,\ i \in J\},\ \mathcal{T}_J^c = \{T_i^c,\ i \in J\},\ T_i^c = \Omega_i \backslash T_i.$$

Then constraints (12) can be rewritten as

$$c_0 + \sum_{i=1}^{n} (c_i - d_i)\, I_{T_i}(\mathbb{E}_{p_X} f_i) \leq z(\mathbf{V}, \mathbf{W}^*),\ p_X \in \mathcal{P}. \tag{15}$$

Here $\mathcal{P}$ is the set of all densities $\{p_X\}$. Let us consider these constraints in detail and define $z(\mathbf{V}, \mathbf{W}^*)$. Note that

$$z(\mathbf{V}, \mathbf{W}^*) = \mathbb{E}_{p_X p_Y} I_{[0,\infty)}(Y - X). \tag{16}$$

However, we fixed the vector $\mathbf{W}^* = (\mathbb{E}^* h_1, ..., \mathbb{E}^* h_l)$. This means that the set of probability densities $p_Y(y)$ is restricted as follows:

$$\mathbb{E}^*_{p_Y} h_1 = w_1^*, ..., \mathbb{E}^*_{p_Y} h_l = w_l^*. \tag{17}$$

So, $z(\mathbf{V}, \mathbf{W}^*)$ can be found by solving the optimization problem with objective function (16), constraints (17), and constraints for p_X, which will be considered below.

In order to compute the indicator functions in (15), it is necessary to substitute different functions p_X from $\mathcal{P}$ and to calculate the corresponding values of $\mathbb{E}_{p_X} f_i$ and $I_{T_i}(\mathbb{E}_{p_X} f_i)$. Moreover, it is necessary to solve problem (16)-(17) for each $p_X \in$

$\mathcal{P}$. Obviously, this task can not be practically solved. Therefore, another way for solving the optimization problem is proposed.

We call the set $\mathcal{T}^c_{N\backslash J} \cup \mathcal{T}_J$ *consistent* if there is at least one density p_X such that $\mathbb{E}_{p_X} f_i \in T_i$, $i \in J$, $\mathbb{E}_{p_X} f_j \in T^c_j$, $j \in N\backslash J$. Now we can see that if the set $\mathcal{T}_J \cup \mathcal{T}^c_{N\backslash J}$ is consistent, then $I_{T_i}(\mathbb{E}_{p_X} f_i) = 1$ if $i \in J$, and $I_{T_i}(\mathbb{E}_{p_X} f_i) = 0$ if $i \in N\backslash J$. In other words, if the set $\mathcal{T}^c_{N\backslash J} \cup \mathcal{T}_J$ is consistent, then there exists at least one density p_X such that all linear previsions $\mathbb{E}_{p_X} f_i$, $i \in J$, are in intervals T_i and their indicator functions are equal to 1, all linear previsions $\mathbb{E}_{p_X} f_j$, $j \in N\backslash J$, do not belong to intervals T_i and their indicator functions are equal to 0. In this case, we will say that p_X belongs to a set $\mathcal{P}_J$. So, to simplify constraints (15), it is necessary to look over all consistent sets $\mathcal{T}^c_{N\backslash J} \cup \mathcal{T}_J$. Then constraints (15) can be rewritten for all $J \subseteq N$, such that $\mathcal{T}^c_{N\backslash J} \cup \mathcal{T}_J$ are consistent, as follows:

$$c_0 + \sum_{i\in J}(c_i - d_i) \leq z(\mathbf{V}, \mathbf{W}^*). \tag{18}$$

If $\mathcal{T}^c_{N\backslash J} \cup \mathcal{T}_J$ is inconsistent, then corresponding inequality (18) is excluded from the list of all constraints.

But how to determine the consistency of sets $\mathcal{T}^c_{N\backslash J} \cup \mathcal{T}_J$? The set $\mathcal{T}^c_{N\backslash J} \cup \mathcal{T}_J$ is consistent if an optimization problem with constraints produced by $\mathcal{T}^c_{N\backslash J} \cup \mathcal{T}_J$ has any solution. At that, the objective function may be arbitrary. In other words, for determining the consistency of $\mathcal{T}^c_{N\backslash J} \cup \mathcal{T}_J$, it is necessary to solve the following optimization problem:

$$\inf_{p_X}\left(\sup_{p_X}\right)\mathbb{E}_{p_X} u(x),$$

subject to $\mathbb{E}_{p_X} f_i \in T_i$, $i \in J$, $\mathbb{E}_{p_X} f_j \in T^c_j$, $j \in N\backslash J$. Here u is an arbitrary function.

Let $p^{(1)}_X \in \mathcal{P}_J$ and $p^{(2)}_X \in \mathcal{P}_J$. Then

$$I_{T_i}(\mathbb{E}_{p^{(1)}_X} f_i) = I_{T_i}(\mathbb{E}_{p^{(2)}_X} f_i).$$

Let

$$z^{(2)}(\mathbf{V}, \mathbf{W}^*) = \mathbb{E}_{p^{(2)}_X p_Y} I_{[0,\infty)}(Y - X) \leq \mathbb{E}_{p^{(1)}_X p_Y} I_{[0,\infty)}(Y - X) = z^{(1)}(\mathbf{V}, \mathbf{W}^*).$$

Then the constraint

$$c_0 + \sum_{i\in N}(c_i - d_i) I_{T_i}(\mathbb{E}_{p^{(1)}_X} f_i) \leq z^{(1)}(\mathbf{V}, \mathbf{W}^*)$$

follows from the constraint

$$c_0 + \sum_{i\in N}(c_i - d_i) I_{T_i}(\mathbb{E}_{p^{(2)}_X} f_i) \leq z^{(2)}(\mathbf{V}, \mathbf{W}^*)$$

and can be removed. This implies that (15) is equivalent to

$$c_0 + \sum_{i \in J} (c_i - d_i) \leq \inf_{\mathcal{P}_J} z(\mathbf{V}, \mathbf{W}^*), \tag{19}$$

where

$$\inf_{\mathcal{P}_J} z(\mathbf{V}, \mathbf{W}^*) = \inf_{p_X, p_Y} \mathbb{E}_{p_X p_Y} I_{[0,\infty)}(Y - X), \tag{20}$$

subject to

$$\mathbb{E}_{p_X} f_i \in \left\{ \begin{array}{ll} T_i, & i \in J \\ T_i^c, & i \in N \backslash J \end{array} \right., \ i \in N, \tag{21}$$

$$\mathbb{E}_{p_Y} h_i = w_i^*, \ i \in L. \tag{22}$$

So, an infinite number of constraints has been reduced to at most 2^n constraints (19). Since the function u is arbitrary, then $\inf_{\mathcal{P}_J} z(\mathbf{V}, \mathbf{W}^*)$ may be used in place of u. There exist exact analytical solutions to problem (20)-(22) for various types of initial information [19].

Solution of Problem (13)-(14).

Now we have the values of $\underline{R}(\mathbf{W}^*)$ for each $\mathbf{W}^* \in \Lambda$. Let us introduce the sets

$$\mathcal{S}_K = \{S_i, \ i \in K\}, \ \mathcal{S}_K^c = \{S_i^c, \ i \in K\}, \ K \subseteq L = \{1, 2, ..., l\}.$$

For solving problem (13)-(14), we apply an algorithm which is similar to the considered one in the previous subsection, i.e.,

$$\underline{R} = \sup \left(c_0 + \sum_{i \in L} \left(c_i \underline{\beta}_i - d_i \overline{\beta}_i \right) \right), \tag{23}$$

subject to $c_i, d_i \in \mathbb{R}_+$, $c_0 \in \mathbb{R}$, $i \in L$, and $\forall K \subseteq L$, $\forall \mathbf{W}^* \in \Lambda$,

$$c_0 + \sum_{i \in K} (c_i - d_i) \leq \inf_{\mathbf{W}^* \in \mathcal{S}_{L \backslash K}^c \cup \mathcal{S}_K} \underline{R}(\mathbf{W}^*). \tag{24}$$

This is a simple linear programming problem with at most 2^l constraints.

5 Exact Bounds for the Reliability

It can be seen from results of the previous section that complex non-linear optimization problem (5)-(6) is reduced to a set of linear programming problems with finitely many constraints and non-linear problems (20)-(22) which can be numerically solved or have exact solutions [19] for the most important types of initial information (points of probability distribution functions of X and Y, moments of X and Y, probabilities defined on nested intervals). However, these optimization problems have to be solved for all values of $\mathbf{W}^* \in \Lambda$ whose number may be infinite. This leads only to the approximate solution and makes the task to be rather difficult from the computational point of view even by a small number of initial judgements. It turns out that optimization problem (5)-(6) can be exactly solved. Therefore, an interesting and efficient solution of the problem is proposed in this section.

Let us consider constraints (24). Suppose that $\underline{R}(\mathbf{W}^*)$ achieves its minimum at $\mathbf{W}^* = \mathbf{W}_o^*(K) \in \mathcal{S}_{L\backslash K}^c \cup \mathcal{S}_K$. Then all vectors $\mathbf{W}^* \in \mathcal{S}_{L\backslash K}^c \cup \mathcal{S}_K$ except $\mathbf{W}_o^*(K)$ are not used in constraints to problem (23)-(24). This implies that we do not need to look over all possible vectors $\mathbf{W}^*$. By returning to problem (11)-(12), it is necessary to solve it only for $\mathbf{W}_o^*(K)$, $K \subseteq L$. This implies that the number of solved optimization problems is finite and depends on numbers n and l of initial judgements about X and Y. Moreover, we can obtain exact bounds for the stress-strength reliability in this case. However, we do not know points $\mathbf{W}_o^*(K)$ before solving problem (11)-(12). Let us show how to overcome this difficulty.

It follows from (11)-(12) that $\underline{R}(\mathbf{W}^*)$ decreases as $z(\mathbf{V}, \mathbf{W}^*)$ decreases. Moreover, the left sides of constraints (19) and (24) do not depend on special values of $\mathbf{W}^*$ and are determined only by the set $\mathcal{S}_{L\backslash K}^c \cup \mathcal{S}_K$. This implies that we do not need to know an optimal value of the vector $\mathbf{W}^* = \mathbf{W}_o^*(K)$. It is enough to know that this value belongs to the set $\mathcal{S}_{L\backslash K}^c \cup \mathcal{S}_K$ (this allows us to construct the K-th constraint in (24)) and makes $z(\mathbf{V}, \mathbf{W}^*)$ and $\underline{R}(\mathbf{W}^*)$ to be minimal for at least one $\mathbf{W}^* \in \mathcal{S}_{L\backslash K}^c \cup \mathcal{S}_K$. Therefore, constraints (22) have to be replaced by constraints

$$\mathbb{E}_{p_Y} h_i(Y) \in \left\{ \begin{array}{ll} S_i, & i \in K \\ S_i^c, & i \in L\backslash K \end{array} \right., \ i \in L, \tag{25}$$

where intervals S_i, S_i^c are defined by the set $\mathcal{S}_{L\backslash K}^c \cup \mathcal{S}_K$.

Indeed, $\inf_{\mathbf{W}^* \in \mathcal{S}_{L\backslash K}^c \cup \mathcal{S}_K} \underline{R}(\mathbf{W}^*)$ corresponds to $\inf_{\mathbf{W}^* \in \mathcal{S}_{L\backslash K}^c \cup \mathcal{S}_K} \inf_{\mathcal{P}_J} z(\mathbf{V}, \mathbf{W}^*)$. At the same time, this is equivalent to the problem $\inf_{\mathcal{P}_J} z(\mathbf{V}, K)$ subject to

$$\mathbb{E}_{p_X} f_i \in \left\{ \begin{array}{ll} T_i, & i \in J \\ T_i^c, & i \in N\backslash J \end{array} \right., \ i \in N, \ \mathbb{E}_{p_Y} h_i = \left\{ \begin{array}{ll} S_i, & i \in K \\ S_i^c, & i \in L\backslash K \end{array} \right., \ i \in L,$$

because constraints (25) contain all points $\mathbf{W}^* \in \mathcal{S}_{L\backslash K}^c \cup \mathcal{S}_K$ and $\inf_{\mathcal{P}_J} z(\mathbf{V}, \mathbf{W}^*)$ is achieved at p_Y satisfying one of the values $\mathbf{W}^*$.

So, $z(\mathbf{V}, \mathbf{W}^*)$ and $\underline{R}(\mathbf{W}^*)$ can be replaced by $z(J, K)$ and $\underline{R}(K)$. This means that values of $\mathbf{V}$ and $\mathbf{W}$ are taken from the sets $\mathcal{T}_{N\backslash J}^c \cup \mathcal{T}_J$ and $\mathcal{S}_{L\backslash K}^c \cup \mathcal{S}_K$, respectively.

It is worth noticing that this subtle technique allows us to solve a problem of consistency of judgements (22). It is obvious that constraints (22) may be inconsistent by some values of w_i^*, and it is not clear what to do in this case. After introducing constraints (25), the inconsistency means that the corresponding constraint in (24) is removed from the list of constraints to problem (23)-(24).

6 Algorithm for Computing $\underline{R}$

Let us write a final algorithm for computing $\underline{R}$.

Step 1. Choosing a set $\mathcal{S}^c_{L\backslash K_i} \cup \mathcal{S}_{K_i}$ from the possible sets $\mathcal{S}^c_{L\backslash K} \cup \mathcal{S}_K$, $K \subseteq L$.

Step 2. Choosing a set $\mathcal{T}^c_{N\backslash J_j} \cup \mathcal{T}_{J_j}$ from the possible sets $\mathcal{T}^c_{N\backslash J} \cup \mathcal{T}_J$, $J \subseteq N$.

Step 3. Solving the optimization problem with objective function (20) and constraints (21) and (25) by T_i and S_i taken from sets $\mathcal{T}^c_{N\backslash J_i} \cup \mathcal{T}_{J_i}$ and $\mathcal{S}^c_{L\backslash K_i} \cup \mathcal{S}_{K_i}$ defined on Steps 1 and 2, respectively. The result of this step is the value $z(J_j, K_i)$. If $z(J, K_i)$ are obtained for all possible $J \subseteq N$, then go to Step 4, else go to Step 2.

Step 4. Solving linear programming problem (11)-(12) by using the consistent values of $z(J, K_i)$ computed on Step 3. The result of this step is the value $\underline{R}(K)$. If $\underline{R}(K)$ are obtained for all possible $K \subseteq L$, then go to Step 5, else go to Step 1.

Step 5. Solving linear programming problem (23)-(24) by using the consistent values of $\underline{R}(K)$ computed on Step 4. The result of this step is $\underline{R}$.

According to the algorithm, it is necessary to solve $2^l + 1$ linear programming problems (Steps 4 and 5) and 2^{nl} non-linear optimization problems (Step 3). Step 3 can be realized by means of results given in [19]. For solving this non-linear problem in a case of arbitrary judgements, a software program has been developed.

7 Numerical Example 1

Suppose that two experts provide probabilities of events concerning the stress and strength. The first expert: 0.9 and 1 are bounds for the probability that the stress is less than $x_1 = 18$. The second expert: 0 and 0.2 are bounds for the probability that the strength is less than $y_1 = 14$; 0.75 and 1 are bounds for the probability that the strength is less than $y_2 = 20$. The beliefs to experts are 0.9 and $[0.6, 0.8]$, respectively. The beliefs $[a, b]$ mean that the expert provides between a% and b% of true judgements. This information can be formally represented as

$$\Pr\{0.9 \leq \mathbb{E}I_{[0,18]}(X) \leq 1\} = 0.9,$$
$$\Pr\{0 \leq \mathbb{E}I_{[0,14]}(Y) \leq 0.2\} \in [0.6, 0.8],$$
$$\Pr\{0.75 \leq \mathbb{E}I_{[0,20]}(Y) \leq 1\} \in [0.6, 0.8].$$

Here $N = \{1\}$, $L = \{1, 2\}$. Let us find $\underline{R} = \underline{\mathbb{E}}\mathbb{E}I_{[0,\infty)}(Y - X)$. Define sets

$$K = \{1, 2\},\ \mathcal{S}^c_{L\backslash K} \cup \mathcal{S}_K = \{S_1, S_2\} = \{[0, 0.2], [0.75, 1]\},$$
$$K = \{1\},\ \mathcal{S}^c_{L\backslash K} \cup \mathcal{S}_K = \{S_1, S^c_2\} = \{[0, 0.2], [0, 0.75]\},$$
$$K = \{2\},\ \mathcal{S}^c_{L\backslash K} \cup \mathcal{S}_K = \{S^c_1, S_2\} = \{[0.2, 1], [0.75, 1]\},$$
$$K = \{\varnothing\},\ \mathcal{S}^c_{L\backslash K} \cup \mathcal{S}_K = \{S^c_1, S^c_2\} = \{[0.2, 1], [0, 0.75]\}.$$

and

$$J=\{1\},\ \mathcal{T}^c_{N\backslash J}\cup\mathcal{T}_J=\{T_1\}=\{[0.9,1]\},$$
$$J=\{\varnothing\},\ \mathcal{T}^c_{N\backslash J}\cup\mathcal{T}_J=\{T^c_1\}=\{[0,0.9]\}.$$

Let us compute $z(J,K)$ for each K and J. According to [19], there holds

$$z=\underline{t}_1(1-\overline{s}_{j(1)}),\ j(i)=\min\{j:x_i\leq y_j\}.$$

Hence $j(1)=2$, and the following hold for $J=\{1\}\subseteq\{1\}$

$$K=\{1,2\},\ z(J,K)=0,$$
$$K=\{1\},\ z(J,K)=0.225,$$
$$K=\{2\},\ z(J,K)=0,$$
$$K=\{\varnothing\},\ z(J,K)=0.225.$$

If $J=\{\varnothing\}$, then $z(J,K)=0$ for all $K\subseteq\{1,2\}$ because $\inf T^c_1=0$. Let us solve problem (11)-(12) for each $K\subseteq L$. For example, if $K=\{1\}$, then

$$\underline{R}(\{1\})=\sup\left(c_0+0.9c_1-0.9d_1\right),$$

subject to $c_1,d_1\in\mathbb{R}_+$, $c_0\in\mathbb{R}$, $c_0+(c_1-d_1)\leq 0.225$, $c_0\leq 0$.

Hence $\underline{R}(\{1\})=0.2025$. Similarly, we can get $\underline{R}(\{1,2\})=0$, $\underline{R}(\{2\})=0$, $\underline{R}(\{\varnothing\})=0.2025$. Let us solve problem (23)-(24)

$$\underline{R}=\sup\left(c_0+0.6c_1-0.8d_1+0.6c_2-0.8d_2\right),$$

subject to $c_1,d_1,c_2,d_2\in\mathbb{R}_+$, $c_0\in\mathbb{R}$,

$$\begin{aligned}c_0+(c_1-d_1)+(c_2-d_2)&\leq 0,\\ c_0+(c_1-d_1)&\leq 0.2025,\\ c_0+(c_2-d_2)&\leq 0,\ c_0\leq 0.2025.\end{aligned}$$

Hence $\underline{R}=0.0405$. The upper stress-strength reliability $\overline{R}=0.9996$ can be computed in the same way by taking into account that there holds $z=1-\underline{s}_1(1-\overline{t}_1)$.

How to use the obtained interval? This depends on a decision maker and the system purposes (consequences of failures). The values 0.0405 and 0.9996 can be interpreted as pessimistic and optimistic assessments of the stress-strength reliability, respectively. If consequences of the system failure are catastrophic (transport systems, nuclear power plants), then the lower bound (pessimistic decision) for the system reliability has to be determinative and is compared with a required level of the system reliability. If the system failure does not imply major consequences, then the upper bound (optimistic decision) can be used. Generally, the decision maker may use a caution parameter η [1] on the basis of his (her) own experience, various conditions of the system functioning, etc. In this case, the precise value of the system reliability is determined as the linear combination $\eta\underline{R}+(1-\eta)\overline{R}$. At the same time, it can be seen from the example that the obtained interval $[\underline{R},\overline{R}]$ is very wide and the results are too imprecise to make a useful decision concerning the reliability.

8 Numerical Example 2

Suppose that information about the stress and strength is represented as the following set of confidence intervals for two moments: the first and second moments of the stress are in intervals $[7,8]$ and $[40,50]$, respectively, with the confidence probability 0.95; the first and second moments of the strength are in intervals $[12,13]$ and $[150,160]$, respectively, with the confidence probability 0.9. By assuming that all values of the stress and strength are in the interval $[0,50]$ (the sample space), this information can be formally represented as

$$\Pr\{7 \leq \mathbb{E}X \leq 8\} \in [0.95,1], \ \ \Pr\{40 \leq \mathbb{E}X^2 \leq 50\} \in [0.95,1],$$
$$\Pr\{12 \leq \mathbb{E}Y \leq 13\} \in [0.9,1], \ \ \Pr\{150 \leq \mathbb{E}Y^2 \leq 160\} \in [0.9,1].$$

Here $N=\{1,2\}$, $L=\{1,2\}$. Results of computing $z(J,K)$ for each K and J are shown in Table 1.

Table 1: Values of $z(J,K)$

	$K=\{1,2\}$	$K=\{1\}$	$K=\{2\}$	$K=\{\varnothing\}$
$J=\{1,2\}$	0.62	0.122	0.04	0
$J=\{1\}$	0.265	0.085	0.03	0
$J=\{2\}$	0.5	0.12	0.042	0
$J=\{\varnothing\}$	0	0	0	0

Let us solve (11)-(12) for each $K \subseteq L$. For example, if $K=\{1,2\}$, then

$$\underline{R}(\{1,2\}) = \sup\left(c_0 + 0.95c_1 - 1d_1 + 0.95c_1 - 1d_1\right),$$

subject to $c_1,d_1,c_2,d_2 \in \mathbb{R}_+$, $c_0 \in \mathbb{R}$,

$$\begin{aligned} c_0 + (c_1-d_1) + (c_2-d_2) &\leq 0.62, \\ c_0 + (c_1-d_1) &\leq 0.265, \\ c_0 + (c_2-d_2) &\leq 0.5, \ \ c_0 \leq 0. \end{aligned}$$

Hence $\underline{R}(\{1,2\}) = 0.589$. Similarly, we can get $\underline{R}(\{1\}) = 0.116$, $\underline{R}(\{2\}) = 0.038$, $\underline{R}(\{\varnothing\}) = 0$. Let us solve problem (23)-(24)

$$\underline{R} = \sup\left(c_0 + 0.9c_1 - 1d_1 + 0.9c_2 - 1d_2\right),$$

subject to $c_1, d_1, c_2, d_2 \in \mathbb{R}_+$, $c_0 \in \mathbb{R}$,

$$\begin{aligned} c_0 + (c_1 - d_1) + (c_2 - d_2) &\leq 0.589, \\ c_0 + (c_1 - d_1) &\leq 0.116, \\ c_0 + (c_2 - d_2) &\leq 0.038, \quad c_0 \leq 0. \end{aligned}$$

Hence $\underline{R} = 0.487$. The upper bound is $\overline{R} = 1$. If we assume that the intervals for moments of the stress and strength have probabilities 1 (the first-order model), then lower and upper bounds for the stress-strength reliability are 0.62 and 1, respectively.

9 Conclusion

The efficient algorithm for computing the stress-strength reliability by the second-order initial information about the stress and strength has been proposed in the paper. This algorithm uses the imprecise stress-strength reliability models obtained in [19]. Its main virtue is that complex non-linear optimization problem (5)-(6) is reduced to a finite set of simple problems whose solution presents no difficulty. Therefore, this approach might be a basis for developing similar algorithms for reliability analysis of various systems where random variables describing the system reliability behavior are independent. The upper bound for the stress-strength reliability can be similarly computed. In this case, the “inf” is replaced by “sup” in optimization problems and vice versa.

It should be noted also a shortcoming of the algorithm. The joint judgements about the stress and strength can not be used because optimization problem (5)-(6) in this case can not be decomposed into a set of linear programming problems. Therefore, further study is needed to develop methods and efficient algorithms for processing the second-order imprecise probabilities by this type of initial information.

Acknowledgement

This work was supported by the Alexander von Humboldt Foundation. I am very grateful to Prof. Dr. Kurt Weichselberger (Munich University, Germany) for his scientific support and help. I have would like to thank Dr. Anton Wallner (Munich University, Germany) for a very useful discussion and comments related to the independence conditions. I also thank three referees for useful and detailed suggestions that improved the paper.

References

[1] T. Augustin. On decision making under ambiguous prior and sampling information. In G. de Cooman, T.L. Fine, and T. Seidenfeld, editors, *Imprecise Probabilities and Their Applications. Proc. of the 2nd Int. Symposium ISIPTA'01*, pages 9–16, Ithaca, USA, June 2001. Shaker Publishing.

[2] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 1985.

[3] I. Couso, S. Moral, and P. Walley. Examples of independence for imprecise probabilities. In G. de Cooman, F.G. Cozman, S. Moral, and P. Walley, editors, *ISIPTA '99 - Proceedings of the First International Symposium on Imprecise Probabilities and Their Applications*, pages 121–130, Zwijnaarde, Belgium, 1999.

[4] G. de Cooman. Precision–imprecision equivalence in a broad class of imprecise hierarchical uncertainty models. *Journal of Statistical Planning and Inference*, 105(1):175–198, 2002.

[5] G. de Cooman and P. Walley. A possibilistic hierarchical model for behaviour under uncertainty. *Theory and Decision*, 52(4):327–374, 2002.

[6] L. Ekenberg and J. Thorbiörnson. Second-order decision analysis. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9:13–38, 2 2001.

[7] P. Gärdenfors and N.-E. Sahlin. Unreliable probabilities, risk taking, and decision making. *Synthese*, 53:361–386, 1982.

[8] I.J. Good. Some history of the hierarchical Bayesian methodology. In J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, editors, *Bayesian Statistics*, pages 489–519. Valencia University Press, Valencia, 1980.

[9] I.O. Kozine and L.V. Utkin. Constructing coherent interval statistical models from unreliable judgements. In E. Zio, M. Demichela, and N. Piccini, editors, *Proceedings of the European Conference on Safety and Reliability ESREL2001*, volume 1, pages 173–180, Torino, Italy, September 2001.

[10] I.O. Kozine and L.V. Utkin. Processing unreliable judgements with an imprecise hierarchical model. *Risk Decision and Policy*, 7(3):325–339, 2002.

[11] V. P. Kuznetsov. *Interval Statistical Models*. Radio and Communication, Moscow, 1991. in Russian.

[12] R. F. Nau. Indeterminate probabilities on finite sets. *The Annals of Statistics*, 20:1737–1767, 1992.

[13] H.T. Nguyen, V. Kreinovich, and L. Longpre. Second-order uncertainty as a bridge between probabilistic and fuzzy approaches. In *Proceedings of the 2nd Conference of the European Society for Fuzzy Logic and Technology EUSFLAT'01*, pages 410–413, England, September 2001.

[14] C.P. Robert. *The Bayesian Choice*. Springer, New York, 1994.

[15] M.C.M. Troffaes and G. de Cooman. Lower previsions for unbounded random variables. In P. Grzegorzewski, O. Hryniewicz, and M.A. Gil, editors, *Soft Methods in Probability, Statistics and Data Analysis*, pages 146–155. Phisica-Verlag, Heidelberg, New York, 2002.

[16] L.V. Utkin. A hierarchical uncertainty model under essentially incomplete information. In P. Grzegorzewski, O. Hryniewicz, and M.A. Gil, editors, *Soft Methods in Probability, Statistics and Data Analysis*, pages 156–163. Phisica-Verlag, Heidelberg, New York, 2002.

[17] L.V. Utkin. Imprecise second-order hierarchical uncertainty model. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 11(3), 2003. To Appear.

[18] L.V. Utkin. A second-order uncertainty model for the calculation of the interval system reliability. *Reliability Engineering and System Safety*, 79(3):341–351, 2003.

[19] L.V. Utkin and I.O. Kozine. Stress-strength reliability models under incomplete information. *International Journal of General Systems*, 31(6):549 – 568, 2002.

[20] L.V. Utkin and I.O. Kozine. Structural reliability modelling under partial source information. In H. Langseth and B. Lindqvist, editors, *Proceedings of the Third Int. Conf. on Mathematical Methods in Reliability (Methodology and Practice)*, pages 647–650, Trondheim, Norway, June 2002. NTNU.

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

[22] P. Walley. Statistical inferences based on a second-order possibility distribution. *International Journal of General Systems*, 9:337–383, 1997.

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

Lev Utkin is with the Department of Computer Science, St.Petersburg Forest Technical Academy, Institutski per. 5, 194021 St.Petersburg, Russia.
Phone +7/812/2476937
E-mail: lvu@utkin.usr.etu.spb.ru

Decision Making with Imprecise Second-Order Probabilities

L.V. UTKIN
St.Petersburg State Forest Technical Academy, Russia

TH. AUGUSTIN
University of Munich, Germany

Abstract

In this paper we consider decision making under hierarchical imprecise uncertainty models and derive general algorithms to determine optimal actions. Numerical examples illustrate the proposed methods.

Keywords

decision making, generalized expected utility, imprecise probabilities, second-order uncertainty, natural extension, linear programming

1 Introduction

Consider the basic model of decision theory: One has to choose an *action* from a non-empty, finite set $\mathbb{A} = \{a_1, ..., a_n\}$ of possible actions. The consequences of every action depend on the true, but unknown *state* of nature $\vartheta \in \Theta = \{\vartheta_1, ..., \vartheta_m\}$. The corresponding outcome is evaluated by the *utility function*

$$\begin{aligned} u: \quad & (\mathbb{A} \times \Theta) \to \mathbb{R} \\ & (a, \vartheta) \longmapsto u(a, \vartheta) \end{aligned}$$

and by the associated random variable $\mathbf{u}(\mathbf{a})$ on $(\Theta, \mathcal{P}o(\Theta))$ taking the values $u(a, \vartheta)$. Often it makes sense to study randomized actions, which can be understood as a probability measure $\lambda = (\lambda_1, ..., \lambda_n)$ on $(\mathbb{A}, \mathcal{P}o(\mathbb{A}))$. Then $u(\cdot)$ and $\mathbf{u}(\cdot)$ are extended to randomized actions by defining $u(\lambda, \vartheta) := \sum_{s=1}^{n} u(a_s, \vartheta)\lambda_s$.

This model contains the essentials of every (formalized) decision situation under uncertainty and is applied in a huge variety of disciplines. If the states of nature are produced by a perfect random mechanism (e.g. an ideal lottery), and the corresponding probability measure $\pi(\cdot)$ on $((\Theta, \mathcal{P}o(\Theta)))$ is completely known, the Bernoulli principle is nearly unanimously favored. One chooses that action λ^* which maximizes the expected utility $\mathbb{E}_\pi \mathbf{u}(\lambda) := \sum_{j=1}^{m} \left(u(\lambda, \vartheta_j) \cdot \pi(\vartheta_j) \right)$ among all λ.

In most practical applications, however, the true state of nature can not be understood as arising from an ideal random mechanism. And even if so, the corresponding probability distribution will be not known exactly. An efficient approach for solving this problem in the framework of imprecise probability theory (Kuznetsov [13], Walley [18], Weichselberger [20]) has been proposed by Augustin in [1, 2].

A related, quite commonly used, way to deal with complex uncertainty is to apply *second-order uncertainty models* (*hierarchical uncertainty models*). These models describe the uncertainty of a random quantity by means of two levels. Many papers are devoted to the theoretical [4, 5, 11, 14, 19] and practical [7, 9, 12] aspects of second-order uncertainty models. A comprehensive review of hierarchical models is given in [6] where it is argued that the most common hierarchical model is the Bayesian one [3, 10, 21]. At the same time, the Bayesian hierarchical model is unrealistic in applications where there is available only partial information about the system behavior.

Most proposed second-order uncertainty models assume that there is a precise second-order probability distribution (or possibility distribution). Unfortunately, such information is often absent and making additional assumptions may lead to wrong results. A new hierarchical uncertainty model for combining different types of evidence was proposed by Utkin [15, 16], where the second-order probabilities can be regarded as confidence weights and the first-order uncertainty is modelled by lower and upper previsions of different gambles. We will call these hierarchical models second-order probabilities of type 1.

It is worth noticing that there are cases when the type of the probability distribution of the states of nature is known, for example, from their physical nature, but parameters or a part of the parameters of the distribution are defined by experts. In reality, there is some degree of our belief to each expert’s judgement whose value is determined by experience and competence of the expert. Therefore, it is necessary to take into account the available information about experts to obtain more credible decisions. This model can be also considered in the framework of hierarchical models and will be called second-order probabilities of type 2.

Decision making for both models of type 1 and type 2 are studied in the paper. In particular, we give general and efficient algorithms for calculating optimal actions and illustrate them in detailed examples.

One should note explicitly that throughout the paper we assume the utility and the description of the uncertainty on the state of nature are given. Alternatively, there are quite sophisticated approaches directly extending the Neumann-Morgenstern point of view. They *construct* separated utility and imprecise probability from axioms on behaviour and preferences (see, e.g., the work of [8] and the references therein).

2 Second-Order Probabilities of Type 1

Suppose that there is a set of weighted expert judgements related to some measures of the states of nature $\mathbb{E}f_i(\vartheta_j)$, $i = 1,...,r$, i.e., there are values $\underline{b}_i$, $\overline{b}_i$ of lower and upper previsions. Suppose that the credibility of each of r experts is characterized by a subjective probability γ_i or interval of probabilities $[\underline{\gamma}_i, \overline{\gamma}_i]$, $i = 1,...,r$. It should be noted that the second-order probabilities $\underline{\gamma}_i$ and $\overline{\gamma}_i$ form an imprecise probability, described by a set $\mathcal{N}$ of distributions on the set $\mathcal{M}$ of all distributions π on $(\Theta, \mathcal{P}o(\Theta))$. We assume that the second-order imprecise probability is avoiding sure loss, i.e., $\mathcal{N}$ is not empty. Denote for any gamble f the lower (upper) second-order expectations by ${}^L\mathbb{E}_{\mathcal{N}} f$ (${}^U\mathbb{E}_{\mathcal{N}} f$), respectively. Generally, the judgements can be written as follows:

$$\Pr\left\{\underline{b}_i \leq \mathbb{E}_\pi f_i \leq \overline{b}_i\right\} \in [\underline{\gamma}_i, \overline{\gamma}_i],\ i = 1,...,r, \tag{1}$$

or

$${}^L\mathbb{E}_{\mathcal{N}} I_{B_i}(\mathbb{E}_\pi f_i) = \underline{\gamma}_i,\ \ {}^U\mathbb{E}_{\mathcal{N}} I_{B_i}(\mathbb{E}_\pi f_i) = \overline{\gamma}_i,\ i = 1,...,r.$$

Here the set $\{\underline{b}_i, \overline{b}_i\}$ contains the first-order previsions, $B_i = [\underline{b}_i, \overline{b}_i]$, the set $\{\underline{\gamma}_i, \overline{\gamma}_i\}$ contains the second-order probabilities and $\mathbb{E}_\pi f_i = \sum_{j=1}^m f_i(\vartheta_j)\pi(\vartheta_j)$.

The problem here is that the resulting set of distributions may be rather complex because the functions f_i are different, especially, if the value of m is large.

Decision Making.

Since there exists the set $\mathcal{N}$ of distributions on the set $\mathcal{M}$ of all distributions π, the expected utility $\mathbb{E}_\pi \mathbf{u}(\lambda)$ can be considered as a random variable described by distributions from $\mathcal{N}$, and there exist lower ${}^L\mathbb{E}_{\mathcal{N}}(\mathbb{E}_\pi \mathbf{u}(\lambda))$ and upper ${}^U\mathbb{E}_{\mathcal{N}}(\mathbb{E}_\pi \mathbf{u}(\lambda))$ expectations of this random variable, which depend on the action λ. These expectations can be roughly called also by lower and upper "average" expected utilities. With this respect, we can assert that every action is evaluated by its minimal "average" expected utility. By representing the interval $[{}^L\mathbb{E}_{\mathcal{N}}(\mathbb{E}_\pi \mathbf{u}(\lambda)), {}^U\mathbb{E}_{\mathcal{N}}(\mathbb{E}_\pi \mathbf{u}(\lambda))]$ by the lower interval limit alone, we can write the criterion of decision making.

Throughout the paper we evaluate interval-valued expectations by their lower interval-limits only — more complex interval orderings are a topic of further research, see also Section 4. Therefore, an action λ^* is optimal iff for all λ

$${}^L\mathbb{E}_{\mathcal{N}}(\mathbb{E}_\pi \mathbf{u}(\lambda^*)) \geq {}^L\mathbb{E}_{\mathcal{N}}(\mathbb{E}_\pi \mathbf{u}(\lambda)). \tag{2}$$

Then the optimal action λ^* can be obtained by maximizing ${}^L\mathbb{E}_{\mathcal{N}}(\mathbb{E}_\pi \mathbf{u}(\lambda))$ subject to $\sum_{s=1}^n \lambda_s = 1$, $\lambda_s \geq 0$, $s = 1,...,n$. In other words, the following optimization problem has to be solved:

$${}^L\mathbb{E}_{\mathcal{N}}(\mathbb{E}_\pi \mathbf{u}(\lambda)) \to \max_{\lambda_s}$$

under the constraints

$$\sum_{s=1}^{n} \lambda_s = 1,\ \lambda_s \geq 0,\ s = 1,...,n.$$

Due to arguments similar to those used in [17], this problem can be rewritten as

$$^{L}\mathbb{E}_{\mathcal{N}}\left(\mathbb{E}_{\pi}\mathbf{u}(\lambda^{*})\right) = \max_{c\in\mathbb{R},c_k\in\mathbb{R}_+,d_k\in\mathbb{R}_+,\lambda_s\in\mathbb{R}_+} \left\{ c + \sum_{k=1}^{r}\left(c_k \underline{\gamma}_k - d_k \overline{\gamma}_k \right) \right\} \tag{3}$$

subject to

$$c + \sum_{k=1}^{r} (c_k - d_k)\, I_{B_k}\left(\mathbb{E}_{\pi} f_k\right) \leq \mathbb{E}_{\pi}\mathbf{u}(\lambda), \tag{4}$$

$$\sum_{s=1}^{n} \lambda_s = 1. \tag{5}$$

By substituting the expressions for $\mathbb{E}_{\pi} f_i$ and $\mathbb{E}_{\pi}\mathbf{u}(\lambda^{*})$ into the constraints, we get

$$c + \sum_{k=1}^{r} (c_k - d_k)\, I_{B_k}\left(\sum_{j=1}^{m} f_k(\vartheta_j)\pi(\vartheta_j) \right) \leq \sum_{j=1}^{m} \left(u(\lambda,\vartheta_j)\cdot\pi(\vartheta_j) \right),\ \forall \pi \in \mathcal{M}. \tag{6}$$

It is worth noticing that the maximal number of different expressions for the left sides of the constraints (6) is 2^r because they involve indicator functions. Let us write a vector $\mathbf{i} = (i_1,...,i_r)$, $i_j \in \{0,1\}$, whose values correspond to those situations. In accordance with possible values of the binary vector $\mathbf{i}$, the set $\mathcal{M}$ can be divided into 2^r subsets $\mathcal{M}_1,...,\mathcal{M}_{2^r}$ such that the i-th subset is formed by the set of constraints

$$\mathbb{E}_{\pi} f_k \in \begin{cases} B_k, & i_k = 1 \\ B_k^c, & i_k = 0 \end{cases},\ k = 1,...,r. \tag{7}$$

Here $B_k^c = [\inf \mathbb{E}_{\pi} f_k, \sup \mathbb{E}_{\pi} f_k] \backslash B_k$ is the (relative) complement of the interval B_k.

Introduce the set $K_j \subseteq \{1,...,r\}$ corresponding to the set $\mathcal{M}_j$ such that for any $\pi \in \mathcal{M}_j$ and $k \in K_j$ there holds $I_{B_k}(\mathbb{E}_{\pi} f_k) = 1$, and for $l \notin K_j$ there holds $I_{B_l}(\mathbb{E}_{\pi} f_l) = 0$.

Let $\pi = (\pi(\vartheta_1),...,\pi(\vartheta_m))$ be a probability distribution belonging to $\mathcal{M}_j$. It should be noted that some elements from the set $\{\mathcal{M}_j, j = 1,...,2^r\}$ may be empty, i.e., there are no such distributions π that satisfy all constraints (7). This means that the corresponding vector of indices $\mathbf{i}$ provides inconsistent judgements (7) and corresponding constraints (4) must be removed from the list of 2^r constraints. Therefore, as the first step, it is necessary to determine the consistency of judgements. The consistency of the set of constraints, corresponding to a realization of the vector $\mathbf{i}$, can be determined by solving a linear programming problem with an arbitrary objective function and constraints (7). If any solution exists, then the feasible region is non-empty and there exists at least one probability distribution π satisfying all constraints (7), i.e., $\mathcal{M}_j \neq \emptyset$. Otherwise, $\mathcal{M}_j = \emptyset$ and the corresponding constraint (4) must be removed.

Let $L \subseteq \{1,...,2^r\}$ be a set of indices for all consistent constraints or all non-empty sets. Suppose that $\pi_1 \in \mathcal{M}_j$ and $\pi_2 \in \mathcal{M}_j$ are two distributions from $\mathcal{M}_j$, $j \in L$, such that $\mathbb{E}_{\pi_1}\mathbf{u}(\lambda) \geq \mathbb{E}_{\pi_2}\mathbf{u}(\lambda)$. Since $\pi_1 \in \mathcal{M}_j$ and $\pi_2 \in \mathcal{M}_j$, then the constraint

$$c + \sum_{k \in K_j} (c_k - d_k) \leq \mathbb{E}_{\pi_1}\mathbf{u}(\lambda),$$

follows from the constraint

$$c + \sum_{k \in K_j} (c_k - d_k) \leq \mathbb{E}_{\pi_2}\mathbf{u}(\lambda),$$

because the left sides of constraints are the same. This implies that from all constraints, corresponding to the set $\mathcal{M}_j$, we have to keep only one constraint

$$c + \sum_{k \in K_j} (c_k - d_k) \leq \min_{\pi \in \mathcal{M}_j} \mathbb{E}_{\pi}\mathbf{u}(\lambda).$$

So, problem (3)-(5) becomes

$$^{L}\mathbb{E}_{\mathcal{N}}\left(\mathbb{E}_{\pi}\mathbf{u}(\lambda^*)\right) = \max_{c \in \mathbb{R}, c_k \in \mathbb{R}_+, d_k \in \mathbb{R}_+, \lambda_s \in \mathbb{R}_+} \left\{ c + \sum_{k=1}^{r} \left(c_k \underline{\gamma}_k - d_k \overline{\gamma}_k \right) \right\} \tag{8}$$

subject to

$$c + \sum_{k \in K_j} (c_k - d_k) \leq \min_{\pi \in \mathcal{M}_j} \mathbb{E}_{\pi}\mathbf{u}(\lambda),\ \forall j \in L, \tag{9}$$

$$\sum_{s=1}^{n} \lambda_s = 1. \tag{10}$$

Write $G_j = \min_{\pi \in \mathcal{M}_j} \mathbb{E}_{\pi}\mathbf{u}(\lambda)$, $j \in L$. Then there holds

$$^{L}\mathbb{E}_{\mathcal{N}}\left(\mathbb{E}_{\pi}\mathbf{u}(\lambda^*)\right) = \max_{c \in \mathbb{R}, c_k \in \mathbb{R}_+, d_k \in \mathbb{R}_+, \lambda_s \in \mathbb{R}_+, G_j} \left\{ c + \sum_{k=1}^{r} \left(c_k \underline{\gamma}_k - d_k \overline{\gamma}_k \right) \right\} \tag{11}$$

subject to

$$c + \sum_{k \in K_j} (c_k - d_k) \leq G_j, \tag{12}$$

$$\mathbb{E}_{\pi}\mathbf{u}(\lambda) \geq G_j,\ \pi \in \mathcal{M}_j,\ \forall j \in L,\ \sum_{s=1}^{n} \lambda_s = 1. \tag{13}$$

One can see that the variables G_k are linear for all $k \in L$. This implies that the optimization problem (11)-(13) is linear, but, in the way it is written, it contains infinitely many constraints. In order to overcome this difficulty, note, however, that the set of distributions $\mathcal{M}_j$ for every j can be viewed as a simplex in a finite dimensional space. According to some general results from linear programming theory, an optimal solution to the above problem is achieved at extreme points of the simplex, and the number of its extreme points is finite. This implies, similar to the solution in the first-order decision problem [1, 2], that the infinite set of constraints (13) is reduced to some finite number, and standard routines for linear programming can be used to determine optimal actions. If one wants to concentrate on unrandomized actions (pure actions), where $\lambda_s \in \{0,1\}$, then Boolean optimization can be used.

Numerical Example.

Suppose that 2 experts evaluate 3 states $\{1,2,3\}$ of nature as follows: the probability that either the first state or the second one is true is less than 0.4; the mean value of states is between 1 and 2. The belief to the first expert is 0.5. This means that he (she) provides 50% of true judgements. The belief to the second expert is between 0.3 and 1. This means that he (she) provides more than 30% of true judgements. Values of the utility function $u(a_s,\vartheta_j)$ are given in Table 1.

Table 1: Values of the utility function $u(a_s,\vartheta_j)$

	ϑ_1	ϑ_2	ϑ_3
a_1	6	3	1
a_2	2	7	4

Table 2: Consistency of constraints

i	set	consistent
$(1,1)$	$\mathbb{E}_\pi I_{\{1,2\}}(\vartheta) \in [0,0.4]$, $\mathbb{E}_\pi \vartheta \in [1,2]$	no
$(1,0)$	$\mathbb{E}_\pi I_{\{1,2\}}(\vartheta) \in [0,0.4]$, $\mathbb{E}_\pi \vartheta \in [2,3]$	yes
$(0,1)$	$\mathbb{E}_\pi I_{\{1,2\}}(\vartheta) \in [0.4,1]$, $\mathbb{E}_\pi \vartheta \in [1,2]$	yes
$(0,0)$	$\mathbb{E}_\pi I_{\{1,2\}}(\vartheta) \in [0.4,1]$, $\mathbb{E}_\pi \vartheta \in [2,3]$	yes

The above judgements can be written in the formal form as follows:

$$\Pr\left\{0 \le \mathbb{E}_\pi I_{\{1,2\}}(\vartheta) \le 0.4\right\} = 0.5, \quad \Pr\{1 \le \mathbb{E}_\pi \vartheta \le 2\} \in [0.3,1].$$

Let us find the set $L \subseteq \{1,2,3,4\}$. It can be seen from Table 2 that $L = \{2,3,4\}$. Let us find the optimal strategies λ_1^*, λ_2^*. For doing so, it is necessary to find extreme points for subsets $\mathcal{M}_2$, $\mathcal{M}_3$, $\mathcal{M}_4$.

Subset 2:

$$\begin{gathered}
\{\pi_1 = 0, \pi_2 = 0, \pi_3 = 1\} \\
\{\pi_1 = 0, \pi_2 = 0.4, \pi_3 = 0.6\} \\
\{\pi_1 = 0.4, \pi_2 = 0, \pi_3 = 0.6\}
\end{gathered}$$

Subset 3:

$$\begin{gathered}
\{\pi_1 = 1, \pi_2 = 0, \pi_3 = 0\} \\
\{\pi_1 = 0, \pi_2 = 1, \pi_3 = 0\} \\
\{\pi_1 = 0.5, \pi_2 = 0, \pi_3 = 0.5\}
\end{gathered}$$

Subset 4:

$$\begin{gathered}
\{\pi_1 = 0, \pi_2 = 1, \pi_3 = 0\} \\
\{\pi_1 = 0.5, \pi_2 = 0, \pi_3 = 0.5\} \\
\{\pi_1 = 0, \pi_2 = 0.4, \pi_3 = 0.6\} \\
\{\pi_1 = 0.4, \pi_2 = 0, \pi_3 = 0.6\}
\end{gathered}$$

So, the following optimization problem has to be considered:

$$^{L}\mathbb{E}_{\mathcal{N}}\left(\mathbb{E}_{\pi}\mathbf{u}(\lambda^*)\right) = \max_{c,c_k,d_k,\lambda_s,G_j} \{c + 0.5c_1 - 0.5d_1 + 0.3c_2 - 1d_2\}$$

subject to $c_i \geq 0, d_i \geq 0, \lambda_i \geq 0$, $i = 1, 2$,

$$\begin{aligned}
c + 1 \cdot (c_1 - d_1) + 0 \cdot (c_2 - d_2) &\leq G_2, \\
c + 0 \cdot (c_1 - d_1) + 1 \cdot (c_2 - d_2) &\leq G_3, \\
c + 0 \cdot (c_1 - d_1) + 0 \cdot (c_2 - d_2) &\leq G_4,
\end{aligned}$$

$$\begin{aligned}
(\lambda_1 + 4\lambda_2) \cdot 1 &\geq G_2, \\
(3\lambda_1 + 7\lambda_2) \cdot 0.4 + (\lambda_1 + 4\lambda_2) \cdot 0.6 &\geq G_2, \\
(6\lambda_1 + 2\lambda_2) \cdot 0.4 + (\lambda_1 + 4\lambda_2) \cdot 0.6 &\geq G_2, \\
(6\lambda_1 + 2\lambda_2) \cdot 1 &\geq G_3, \\
(3\lambda_1 + 7\lambda_2) \cdot 1 &\geq G_3, \\
(6\lambda_1 + 2\lambda_2) \cdot 0.5 + (\lambda_1 + 4\lambda_2) \cdot 0.5 &\geq G_3, \\
(3\lambda_1 + 7\lambda_2) \cdot 1 &\geq G_4, \\
(6\lambda_1 + 2\lambda_2) \cdot 0.5 + (\lambda_1 + 4\lambda_2) \cdot 0.5 &\geq G_4, \\
(3\lambda_1 + 7\lambda_2) \cdot 0.4 + (\lambda_1 + 4\lambda_2) \cdot 0.6 &\geq G_4, \\
(6\lambda_1 + 2\lambda_2) \cdot 0.4 + (\lambda_1 + 4\lambda_2) \cdot 0.6 &\geq G_4, \\
\lambda_1 + \lambda_2 &= 1.
\end{aligned}$$

Solution of the problem: $c = 3.143$, $G_2 = G_3 = G_4 = 3.143$, $c_1 = c_2 = d_1 = d_2 = 0$, $\lambda_1 = 0.2857$, $\lambda_2 = 0.7143$.

3 Second-Order Probabilities of Type 2

Suppose that the states of nature are described by a discrete probability distribution of a certain type, for example, binomial, hypergeometric or Poisson distributions. The certain type of the distribution is often known from some physical properties of the considered object. However, the parameters of the corresponding distribution may be uncertain. Denote by $\alpha = (\alpha_1, ..., \alpha_h)$ a vector of parameters for some discrete distribution $\pi(\vartheta, \alpha)$. Consider a case of continuous real parameters, i.e., $\alpha_i \in \mathbb{R}$. If we suppose that the experts provide some evidence about parameters, then the vector α can be considered, just as in classical Bayesian statistics, as a random variable. This is due to the following reasons: First, experts may provide some information about statistical characteristics of parameters, for example, about intervals of mean values or about some probability that the i-th parameter is in an interval. Second, even if experts provide only information about intervals of possible values of parameters, we can not totally believe in the experts because they may be unreliable. This implies that every expert is characterized by a probability or by an interval-valued probability of producing correct judgements. Generally, if we suppose that the vector of parameters is governed by some unknown joint density ρ, then the expert judgements can be formally written as follows:

$$\underline{\gamma}_{ij} \leq \mathbb{E}_\rho f_{ij}(\alpha_i) \leq \overline{\gamma}_{ij}, \ i = 1, ..., h, \ j = 1, ..., r_i. \tag{14}$$

Here r_i is a number of judgements related to i-th parameter; f_{ij} is a function corresponding to information about the i-th parameter provided by the j-th expert. For example, if an expert offers information about the probability that the i-th parameter is in an interval B, then $f_{ij}(\alpha_i)$ is the indicator function of the event B, i.e., $f_{ij}(\alpha_i) = I_B(\alpha_i)$. If the expert provides the mean value of the i-th parameter, then there holds $f_{ij}(\alpha_i) = \alpha_i$. The values $\underline{\gamma}_{ij}$ and $\overline{\gamma}_{ij}$ are the bounds for the provided characteristic $\mathbb{E}_\rho f_{ij}(\alpha_i)$ of the i-th parameter[1].

Decision Making.

We assume that there are some bounds for all parameters $[\underline{\alpha}_i, \overline{\alpha}_i]$, $i = 1, ..., h$. This means that the i-th parameter belongs to the interval $[\underline{\alpha}_i, \overline{\alpha}_i]$ with probability 1. Inside this interval, the parameter is distributed according to an unknown probability density ρ_i.

So, we have some infinite set of discrete probability distributions $\pi(\vartheta_j, \alpha)$ defined by different parameters. Then the expected utility corresponding to one

[1] For simplicity, it is assumed that either experts with weights provide intervals for unknown parameters or experts without weights provide some statistical characteristics of random parameters. Of course, we could consider more complex cases when experts with weights provide statistical characteristics of random parameters, but the study of these, so-to-say third-order level, cases may hide the main results behind complex notation.

realization of the vector α is

$$\mathbb{E}_{\pi}\mathbf{u}(\lambda,\alpha)=\sum_{j=1}^{m}\left(u(\lambda,\vartheta_j)\cdot\pi(\vartheta_j,\alpha)\right).$$

By averaging the expected utilities $\mathbb{E}_{\pi}\mathbf{u}(\lambda,\alpha)$ over all possible vectors α, we get

$$\mathbb{E}_{\rho}\mathbb{E}_{\pi}\mathbf{u}(\lambda,\alpha)=\int_{\Omega^h}\left(\sum_{j=1}^{m}\left(u(\lambda,\vartheta_j)\cdot\pi(\vartheta_j,\alpha)\right)\right)\rho(\alpha)\mathrm{d}\alpha.$$

Here Ω^h is a sample space and $\Omega^h=[\underline{\alpha}_1,\overline{\alpha}_1]\times...\times[\underline{\alpha}_h,\overline{\alpha}_h]$.

Now we define an optimal action. An action λ^* is optimal iff

$$^{L}\mathbb{E}_{\mathcal{P}}\left(\mathbb{E}_{\pi}\mathbf{u}(\lambda^*,\alpha)\right)\geq{}^{L}\mathbb{E}_{\mathcal{P}}\left(\mathbb{E}_{\pi}\mathbf{u}(\lambda,\alpha)\right). \tag{15}$$

Here $\mathcal{P}$ is a set of all possible density functions $\rho(\alpha)$ satisfying the constraints

$$\underline{\gamma}_{ij}\leq\mathbb{E}_{\rho}f_{ij}(\alpha_i)\leq\overline{\gamma}_{ij},\ i=1,...,h,\ j=1,...,r_i,$$

or

$$\underline{\gamma}_{ij}\leq\int_{\underline{\alpha}_i}^{\overline{\alpha}_i}f_{ij}(\alpha_i)\rho_i(\alpha_i)\mathrm{d}\alpha_i\leq\overline{\gamma}_{ij},\ i=1,...,h,\ j=1,...,r_i.$$

Then the optimal action λ^* can be obtained by maximizing $^{L}\mathbb{E}_{\mathcal{P}}\left(\mathbb{E}_{\pi}\mathbf{u}(\lambda,\alpha)\right)$ subject to $\sum_{s=1}^{n}\lambda_s=1$, $\lambda_s\geq 0$, $s=1,...,n$. In other words, the following optimization problem has to be solved:

$$^{L}\mathbb{E}_{\mathcal{P}}\left(\mathbb{E}_{\pi}\mathbf{u}(\lambda^*,\alpha)\right)\to\max_{\lambda_s} \tag{16}$$

under the constraints

$$\sum_{s=1}^{n}\lambda_s=1,\ \lambda_s\geq 0,\ s=1,...,n. \tag{17}$$

If we assume that there is no information about independence of parameters, i.e., the joint density $\rho(\alpha)$ can not be represented as a product of marginal ones, then problem (16)-(17) can be rewritten as

$$^{L}\mathbb{E}_{\mathcal{P}}\left(\mathbb{E}_{\pi}\mathbf{u}(\lambda^*,\alpha)\right)=\max_{c\in\mathbb{R},c_{kj}\in\mathbb{R}_+,d_{kj}\in\mathbb{R}_+,\lambda_s}\left\{c+\sum_{k=1}^{h}\sum_{j=1}^{r_k}\left(c_{kj}\underline{\gamma}_{kj}-d_{kj}\overline{\gamma}_{kj}\right)\right\} \tag{18}$$

subject to

$$c+\sum_{k=1}^{h}\sum_{j=1}^{r_k}\left(c_{kj}-d_{kj}\right)f_{kj}(\alpha_i)\leq\mathbb{E}_{\pi}\mathbf{u}(\lambda,\alpha),\ \forall\alpha\in\Omega^h, \tag{19}$$

$$\sum_{s=1}^{n}\lambda_s=1,\ \lambda_s\geq 0. \tag{20}$$

This is a linear programming problem having an infinite number of constraints. However, for many special cases problem (18)-(20) can be simplified. Let us consider the most important and realistic case when experts provide h intervals $B_1,...,B_r$ for unknown parameters and each expert is characterized by some probability γ_{kj} or interval-valued probability $[\underline{\gamma}_{ij},\overline{\gamma}_{ij}]$. Moreover, in order to give the reader the essence of the subject analyzed and make all the formulas more readable, we will also assume that $h=1$ and $\alpha=(\alpha)$, i.e., there is only one parameter of the distribution $\pi(\vartheta_j,\alpha)$. We also denote r_1 by r. In other words, constraints (14) are represented as

$$\underline{\gamma}_j \leq \int_{\underline{\alpha}}^{\overline{\alpha}} I_{B_j}(\alpha)\rho(\alpha)\mathrm{d}\alpha \leq \overline{\gamma}_j,\ j=1,...,r. \tag{21}$$

Then problem (18)-(20) can be rewritten as

$$^{L}\mathbb{E}_{\mathcal{P}}\left(\mathbb{E}_{\pi}\mathbf{u}(\lambda^{*},\alpha)\right)=\max_{c\in\mathbb{R},c_k\in\mathbb{R}_{+},d_k\in\mathbb{R}_{+},\lambda_s}\left\{c+\sum_{k=1}^{r}\left(c_k\underline{\gamma}_k-d_k\overline{\gamma}_k\right)\right\} \tag{22}$$

subject to

$$c+\sum_{k=1}^{r}(c_k-d_k)\,I_{B_k}(\alpha)\leq\mathbb{E}_{\pi}\mathbf{u}(\lambda,\alpha),\ \forall\alpha\in[\underline{\alpha},\overline{\alpha}], \tag{23}$$

$$\sum_{s=1}^{n}\lambda_s=1,\ \lambda_s\geq 0. \tag{24}$$

Denote $\mathbf{i}=(i_1,...,i_r)$, $i_j\in\{0,1\}$. In accordance with possible values of the binary vector $\mathbf{i}$, the interval $B=[\underline{\alpha},\overline{\alpha}]$ of all values α can be divided into 2^r subintervals $B^{(1)},...,B^{(2^r)}$ such that the i-th subinterval is formed by

$$B^{(i)}=\bigcap_{k=1}^{r}\left\{\begin{array}{ll} B_k, & \mathrm{i}_k=1\\ B_k^c, & \mathrm{i}_k=0\end{array}\right. . \tag{25}$$

Let $L\subseteq\{1,...,2^r\}$ be a set of indices for all non-empty subintervals $B^{(j)}\neq\emptyset$. Then from all constraints corresponding to the subinterval $B^{(j)}$, we have to keep only one constraint

$$c+\sum_{k=1}^{r}(c_k-d_k)\,i_k\leq\min_{\alpha\in B^{(j)}}\mathbb{E}_{\pi}\mathbf{u}(\lambda,\alpha).$$

So, problem (22)-(24) becomes

$$^{L}\mathbb{E}_{\mathcal{P}}\left(\mathbb{E}_{\pi}\mathbf{u}(\lambda^{*},\alpha)\right)=\max_{c\in\mathbb{R},c_k\in\mathbb{R}_{+},d_k\in\mathbb{R}_{+},\lambda_s}\left\{c+\sum_{k=1}^{r}\left(c_k\underline{\gamma}_k-d_k\overline{\gamma}_k\right)\right\} \tag{26}$$

subject to

$$c + \sum_{k=1}^{r} (c_k - d_k)\, i_k \leq \min_{\alpha \in B^{(j)}} \mathbb{E}_{\pi}\mathbf{u}(\lambda,\alpha),\ \forall \mathbf{i}, \tag{27}$$

$$\sum_{s=1}^{n} \lambda_s = 1,\ \lambda_s \geq 0. \tag{28}$$

Let us introduce the variable $G_j = \min_{\alpha \in B^{(j)}} \mathbb{E}_{\pi}\mathbf{u}(\lambda,\alpha)$. Then problem (26)-(28) can be rewritten as

$${}^{L}\mathbb{E}_{\mathcal{P}}\left(\mathbb{E}_{\pi}\mathbf{u}(\lambda^{*},\alpha)\right) = \max_{c\in\mathbb{R}, c_k\in\mathbb{R}_+, d_k\in\mathbb{R}_+, \lambda_s, G_j} \left\{ c + \sum_{k=1}^{r} \left(c_k \underline{\gamma}_k - d_k \overline{\gamma}_k \right) \right\} \tag{29}$$

subject to

$$c + \sum_{k=1}^{r} (c_k - d_k)\, i_k \leq G_j,\ \forall \mathbf{i}, \tag{30}$$

$$\mathbb{E}_{\pi}\mathbf{u}(\lambda,\alpha) \geq G_j,\ \forall \alpha \in B^{(j)},\ \forall \mathbf{i}, \tag{31}$$

$$\sum_{s=1}^{n} \lambda_s = 1,\ \lambda_s \geq 0. \tag{32}$$

In this case, we obtain the linear programming problem with infinite number of constraints. However, if it is known that the function $\mathbb{E}_{\pi}\mathbf{u}(\lambda,\alpha)$ is monotone with α, then it is sufficient to consider only boundary points of intervals $B^{(j)}$. Constraints (31) can be written as

$$\sum_{j=1}^{m} \left(\sum_{s=1}^{n} (u(a_s,\vartheta_j)\lambda_s) \cdot \pi(\vartheta_j) \right) \leq G_j,$$

or

$$\sum_{s=1}^{n} \left(\sum_{j=1}^{m} (u(a_s,\vartheta_j)\pi(\vartheta_j)) \right) \lambda_s \leq G_j.$$

Hence it is obvious that the constraints are linear with λ_s.

Numerical Example.

Suppose that 3 states $\{1,2,3\}$ of nature are governed by the binomial distribution

$$\pi(\vartheta_j,\alpha) = \binom{3-1}{j-1} \alpha^{j-1} (1-\alpha)^{3-j-1},\ j = 1,2,3.$$

Two experts provide their judgements about the parameter $\alpha \in [0,1]$ as follows:

1. the parameter α is in interval $[0.8,1]$;

2. the parameter α is in interval $[0.7,1]$.

The belief in the correctness of the first expert is 0.5. The belief in the second expert is between 0.3 and 1 (see Section 2). The above judgements can be written in the formal form as follows:

$$\int_0^1 I_{[0.8,1]}(\alpha)\rho(\alpha)\mathrm{d}\alpha = 0.5,\ \int_0^1 I_{[0.7,1]}(\alpha)\rho(\alpha)\mathrm{d}\alpha \in [0.3,1].$$

Let us find the set $L \subseteq \{1,2,3,4\}$.

i	intervals	non-empty
$(1,1)$	$[0.8,1]\cap[0.7,1]$	yes
$(1,0)$	$[0.8,1]\cap[0,0.7]$	no
$(0,1)$	$[0,0.8]\cap[0.7,1]$	yes
$(0,0)$	$[0,0.8]\cap[0,0.7]$	yes

Table 3: Intersections of intervals

It can be seen from Table 3 that $L = \{1,3,4\}$.

Let us find λ_1, λ_2. In this case, there holds

$$^{L}\mathbb{E}_{\mathcal{P}}\left(\mathbb{E}_{\pi}\mathbf{u}(\lambda^*,\alpha)\right) = \max_{c\in\mathbb{R},c_k\in\mathbb{R}_+,d_k\in\mathbb{R}_+,\lambda_s,G_j}\{c+0.5c_1-0.5d_1+0.3c_2-1d_2\}$$

subject to

$$\begin{aligned} c+1\cdot(c_1-d_1)+1\cdot(c_2-d_2) &\leq G_1,\\ c+0\cdot(c_1-d_1)+1\cdot(c_2-d_2) &\leq G_3,\\ c+0\cdot(c_1-d_1)+0\cdot(c_2-d_2) &\leq G_4,\\ \left(\alpha^2-6\alpha+6\right)\lambda_1+\left(10\alpha-8\alpha^2+2\right)\lambda_2 &\geq G_1,\ \alpha\in[0.8,1],\\ \left(\alpha^2-6\alpha+6\right)\lambda_1+\left(10\alpha-8\alpha^2+2\right)\lambda_2 &\geq G_3,\ \alpha\in[0.7,0.8],\\ \left(\alpha^2-6\alpha+6\right)\lambda_1+\left(10\alpha-8\alpha^2+2\right)\lambda_2 &\geq G_3,\ \alpha\in[0,0.7],\\ \lambda_1+\lambda_2 &= 1,\ \lambda_1\geq 0,\lambda_2\geq 0. \end{aligned}$$

By solving this problem approximately (for a finite number of values of α), we get $c = 3.636$, $G_1 = 2.773$, $G_3 = G_4 = 3.636$, $c_1 = c_2 = d_2 = 0$, $d_1 = 0.864$, $\lambda_1 = 0.409$, $\lambda_2 = 0.591$.

4 Concluding Remarks

Two models of decision making based on different types of initial hierarchical information about states of nature have been studied in the paper. We have shown that both models can be brought into a form which allows us to give general algorithms to determine optimal solutions.

It should be noted that we have focused in this paper on the basic decision problem. However, the fundamental ideas of this paper should be also applicable to more complex decision problems, like multi-criteria decision making and data-based decision problems. Another topic of furhter research is to extend the results obtained here to other optimality criteria which are more sophisticated than the criteria from (2) and (15), which take into account only the lower interval limits.

Acknowledgement

The first author gratefully acknowledges financial support of the Alexander von Humboldt Foundation. We both thank Kurt Weichselberger, Franz Ferschl, Martin Gümbel, Bernhard Rüger, Heinrich Strecker and Anton Wallner for numerous comments and stimulating discussions during the Munich weekly seminar on interval probability.

References

[1] AUGUSTIN, T. On decision making under ambiguous prior and sampling information. In *Imprecise Probabilities and Their Applications. Proc. of the 2nd Int. Symposium ISIPTA'01* (Ithaca, USA, June 2001), G. de Cooman, T. Fine, and T. Seidenfeld, Eds., Shaker Publishing, pp. 9–16.

[2] AUGUSTIN, T. Expected utility within a generalized concept of probability - a comprehensive framework for decision making under ambiguity. *Statistical Papers 43* (2002), 5–22.

[3] BERGER, J. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 1985.

[4] DE COOMAN, G. Possibilistic previsions. In *Proceedings of IPMU'98* (Paris, 1998), vol. 1, Editions EDK, pp. 2–9.

[5] DE COOMAN, G. Precision–imprecision equivalence in a broad class of imprecise hierarchical uncertainty models. *Journal of Statistical Planning and Inference 105*, 1 (2002), 175–198.

[6] DE COOMAN, G., AND WALLEY, P. A possibilistic hierarchical model for behaviour under uncertainty. *Theory and Decision 52*, 4 (2002), 327–374.

[7] EKENBERG, L., AND THORBIÖRNSON, J. Second-order decision analysis. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 9* (2 2001), 13–38.

[8] GHIRARDATO, P., AND MARINACCI, M. Risk, ambiguity, and the separation of utility and beliefs. *Mathematics of Operations Research 26* (2001), 864–890.

[9] GILBERT, L., DE COOMAN, G., AND KERRE, E. Practical implementation of possibilistic probability mass functions. In *Proceedings of Fifth Workshop on Uncertainty Processing (WUPES 2000)* (Jindvrichouv Hradec, Czech Republic, June 2000), pp. 90–101.

[10] GOOD, I. Some history of the hierarchical Bayesian methodology. In *Bayesian Statistics*, J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, Eds. Valencia University Press, Valencia, 1980, pp. 489–519.

[11] GOODMAN, I. R., AND NGUYEN, H. T. Probability updating using second order probabilities and conditional event algebra. *Information Sciences 121*, 3-4 (1999), 295–347.

[12] KOZINE, I., AND UTKIN, L. Processing unreliable judgements with an imprecise hierarchical model. *Risk Decision and Policy 7*, 3 (2002), 325–339.

[13] KUZNETSOV, V. P. *Interval Statistical Models*. Radio and Communication, Moscow, 1991. in Russian.

[14] NAU, R. F. Indeterminate probabilities on finite sets. *The Annals of Statistics 20* (1992), 1737–1767.

[15] UTKIN, L. A hierarchical uncertainty model under essentially incomplete information. In *Soft Methods in Probability, Statistics and Data Analysis*, P. Grzegorzewski, O. Hryniewicz, and M. Gil, Eds. Phisica-Verlag, Heidelberg, New York, 2002, pp. 156–163.

[16] UTKIN, L. Imprecise second-order hierarchical uncertainty model. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 11*, 3 (2003). To Appear.

[17] UTKIN, L., AND KOZINE, I. Different faces of the natural extension. In *Imprecise Probabilities and Their Applications. Proc. of the 2nd Int. Symposium ISIPTA'01* (Ithaca, USA, June 2001), G. de Cooman, T. Fine, and T. Seidenfeld, Eds., Shaker Publishing, pp. 316–323.

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

[19] WALLEY, P. Statistical inferences based on a second-order possibility distribution. *International Journal of General Systems 9* (1997), 337–383.

[20] WEICHSELBERGER, K. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung*, vol. I Intervallwahrscheinlichkeit als umfassendes Konzept. Physika, Heidelberg, 2001.

[21] ZELLNER, A. *An introduction to Bayesian Inference in Econometrics*. Wiley, New York, 1971.

Lev Utkin is with the Department of Computer Science, St.Petersburg Forest Technical Academy, Institutski per. 5, 194021 St.Petersburg, Russia.
Phone +7/812/247693
E-mail: lvu@utkin.usr.etu.spb.ru

Thomas Augustin is with the Department of Statistics, Ludwig-Maximilians-University of Munich, Ludwigstr. 33, D-80539 Munich, Germany.
Phone +49/89/2180-3520 Fax +49/89/2180-5044
E-mail: augustin@stat.uni-muenchen.de

Graphical Representation of Asymmetric Graphoid Structures

B. VANTAGGI
Università "La Sapienza," Roma, Italy

Abstract

Independence models induced by some uncertainty measures (e.g. conditional probability, possibility) do not obey the usual graphoid properties, since they do not satisfy the symmetry property. They are efficiently representable through directed acyclic l-graphs by using L-separation criterion.

In this paper, we show that in general there is not a l-graph which describes completely all the independence statements of a given model; hence we introduce in this context the notion of minimal I-map and we show how to build it, given an ordering on the variables. In addition, we prove that, for any ordering, there exists an I-map for any asymmetric graphoid structure.

Keywords

conditional independence models, directed acyclic graph, L-separation criterion, I-map

1 Introduction

The use of graphs to describe conditional independence structures (the set of conditional independence statements "X is independent of Y given Z") induced by probability distributions has a long and rich tradition; one can distinguish three main classic approaches based on *undirected graphs* [12], *directed acyclic graphs* [14], or *chain graphs* [15]. These graphical structure obey graphoid properties (symmetry, decomposition, weak union, contraction, intersection). On the other hand, the independence models based on the classic definition of stochastic independence in the usual probabilistic setting, have semi-graphoid structure (they satisfy all graphoid properties except intersection). However, if the probability distribution is strictly positive, the independence model has a graphoid structure. Hence, the lack of intersection property is due to zero probability on some of the possible events. Actually, it is well-known (see, for example, [4, 6]) that the classic definition of stochastic independence presents counter-intuitive situations when zero or one probability events are involved: for example, a possible event with zero or one probability is independent of itself.

We stress that zero probability values are interesting not only from a merely theoretical point of view, but they are met in many real problems, for example in medical diagnosis [7] , statistical mechanics, physics, etc. [11].

The counter-intuitive situations cannot be avoided within the usual framework of conditional probability. In the more general framework (de Finetti [8], Dubins [9]), a definition of stochastic independence (called cs-independence), which avoids these critical situations, has been introduced in [4] and the main properties have been studied. We recall that the aforementioned definition agrees with the classic one when the probabilities of the relevant events are different from 0 and 1.

The main properties connected with graphoid structures were proved in [16]: these independence models generally are not closed with respect to the symmetry property. Hence, the classic separation criterion are not apt to represent asymmetric independence statements, so in [17] a new separation criterion (called L-separation) for directed acyclic l-graphs has been introduced. It has been shown also that L-separation criterion satisfy *asymmetric graphoid* properties (graphoid properties except symmetry).

In this paper we deepen the problem of representing such cs-independence model, together with the logical constraints, using L-separation criterion in directed acyclic l-graphs. In particular, Example 1 shows that cs-independence structures are richer than the graphical ones, i.e. for some independence model there is no graph able to describe all the independence statements. Hence, in Section 5 we define in this context (analogously to [14, 10]) the notion of minimal I-map for a given independence model $\mathcal{M}$: a directed acyclic l-graph such that every statement represented by it is in $\mathcal{M}$, while the graph obtained by removing any arrow from it would represent an independence statement not in $\mathcal{M}$.

Moreover, in Section 5 we show how to build such minimal I-maps underling the differences arising from the lack of symmetry property, and, in addition, we prove that any ordering on the variables gives rise to an I-map for any independence model $\mathcal{M}$ obeying to asymmetric graphoid properties.

On the other hand, the ordering has a crucial role: in fact, if a perfect I-map (able to describe all the independence statements) exists, it can be built using only some specific ordering on the variables.

2 Independence in a coherent probability setting

It is well known that the classic definition of stochastic independence of two events

$$P(A \wedge B) = P(A)P(B) \tag{1}$$

gives rise to counter-intuitive situations when one of the events has probability 0 or 1. For instance an event A with $P(A) = 0$ is stochastically independent of itself, while it is natural (due to the intuitive meaning of independence) to require for any event to be dependent on itself. Other classic formulations are $P(A|B) = P(A)$ and

$P(A|B) = P(A|B^c)$, that are equivalent to (1) for events such that the probability of B is different from 0 and 1, but in that "extreme" cases (without positivity assumption) they may even lack meaning in the Kolmogorovian approach.

Anyway, some critical situations related to logical dependence continue to exist (see [16]) also considering the last stronger formulation in the more general framework of de Finetti [8]:

Definition 1 *Given a Boolean algebra $\mathcal{A}$, a conditional probability on $\mathcal{A} \times \mathcal{A}^0$ (with $\mathcal{A}^0 = \mathcal{A} \setminus \{\emptyset\}$) is a function $P(\cdot|\cdot)$ into $[0,1]$, which satisfies the following conditions:*
(i) $P(\cdot|H)$ is a finitely additive probability on $\mathcal{A}$ for any $H \in \mathcal{A}^0$
(ii) $P(H|H) = 1$ for every $H \in \mathcal{A}^0$
(iii) $P(E \wedge A|H) = P(E|H)P(A|E \wedge H)$, whenever $E, A \in \mathcal{A}$ and $H, E \wedge H \in \mathcal{A}^0$

Note that (iii) reduces, when $H = \Omega$ (where Ω is the *certain* event), to the classic "chain rule" for probability $P(E \wedge A) = P(E)P(A|E)$. In the case $P_0(\cdot) = P(\cdot|\Omega)$ is strictly positive on $\mathcal{A}^0$, any conditional probability can be derived as a ratio (Kolmogorov's definition) by this unique "unconditional" probability P_0.

As proved in [6], in all other cases to get a similar representation we need to resort to a finite family $\mathcal{P} = \{P_0, \ldots, P_k\}$ of unconditional probabilities:
- every P_α is defined on a proper set of events (taking $\mathcal{A}_0 = \mathcal{A}$)

$$\mathcal{A}_\alpha = \{E \in \mathcal{A}_{\alpha-1} : P_{\alpha-1}(E) = 0\}$$

- for each event $B \in \mathcal{A}^0$ there exists an unique α such that $P_\alpha(B) > 0$ and for every conditional event $E|H$ one has $P(E|H) = \frac{P_\alpha(E \wedge H)}{P_\alpha(H)}$ with $P_\alpha(H) > 0$.

The class of probabilities $\mathcal{P} = \{P_0, \ldots, P_k\}$ is said to *agree* with the conditional probability $P(\cdot|\cdot)$.

Such theory of conditional probability allows to handle also *partial* probability assessment on an arbitrary set of conditional events $\mathcal{F} = \{E_1|H_1, \ldots, E_n|H_n\}$ through the concept of coherence: an assessment is *coherent* if it is the restriction of a conditional probability defined on $\mathcal{A} \times \mathcal{A}^0$, where $\mathcal{A}$ is the algebra generated by $\{E_1, H_1, \ldots, E_n, H_n\}$. A characterization of coherence was proven in [3]:

Theorem 1 *Let $\mathcal{F}$ be an arbitrary finite family of conditional events and $\mathcal{C}$ denote the set of atoms C_r generated by the events $E_1, H_1, \ldots, E_n, H_n$. For a real function P on $\mathcal{F}$ the following two statements are equivalent:*
(i) P is a coherent conditional probability on $\mathcal{F}$;
(ii) there exists a class of unconditional probabilities $\{P_0, \ldots P_k\}$, with P_0 defined on $\mathcal{A}_0$ and P_α $(\alpha > 0)$ being defined on $\mathcal{A}_\alpha = \{E \in \mathcal{A}_{\alpha-1} : P_{\alpha-1}(E) = 0\}$, such that for any $E_i|H_i \in \mathcal{F}$ there is a unique P_α, with $P_\alpha(H_i) > 0$, and

$$P(E_i|H_i) = \frac{\sum\limits_{C_r \subseteq E_i \wedge H_i} P_\alpha(C_r)}{\sum\limits_{C_r \subseteq H_i} P_\alpha(C_r)} .$$

The class of probabilities $\mathcal{P} = \{P_0, \ldots, P_k\}$ *agreeing* with the given coherent assessment P is not unique. But, given one class $\mathcal{P} = \{P_0, \ldots, P_k\}$, for each event H there is a unique α such that $P_\alpha(H) > 0$ and α is said *zero-layer* of H according to $\mathcal{P}$, and it is denoted by the symbol $\circ(H)$. In particular, for every probability we have $\circ(\Omega) = 0$, while we define $\circ(\emptyset) = \infty$. The *zero-layer of a conditional event* $E|H$ is defined (see [4]) as

$$\circ(E|H) = \circ(E \wedge H) - \circ(H).$$

In the sequel, to avoid cumbersome notation, the conjunction symbol $\wedge$ among events is omitted.

In this framework the following definition of stochastic independence has been proposed in [4] and extended to conditional independence in [16]:

Definition 2 *Given a coherent conditional probability P, defined on a family $\mathcal{F}$ containing $\mathcal{D} = \{A|BC, A|B^cC, A^c|BC, A^c|B^cC, B|AC, B|A^cC, B^c|AC, B^c|A^cC\}$, A is conditionally independent of B given C with respect to P (in symbol $A \perp\!\!\!\perp_{cs} B|C$) if both the following conditions hold:*

(i) $P(A|BC) = P(A|B^cC)$;

(ii) there exists a class $\{P_\alpha\}$ of probabilities agreeing with the restriction of P to the family $\mathcal{D}$, such that

$$\circ(A|BC) = \circ(A|B^cC) \quad \text{and} \quad \circ\,(A^c|BC) = \circ(A^c|B^cC)\,.$$

Note that if $0 < P(A|BC) = P(A|B^cC) < 1$ (so $0 < P(A^c|BC) = P(A^c|B^cC) < 1$), then both equalities in condition (ii) are trivially satisfied

$$\circ(A|BC) = 0 = \circ(A|B^cC) \quad \text{and} \quad \circ\,(A^c|BC) = 0 = \circ(A^c|B^cC).$$

Hence, in this case condition (i) completely characterizes conditional cs-independence, and, in addition, this definition coincides with the classic formulations when also $P(B|C)$ and $P(C)$ are in $(0,1)$. However, in the other cases (when $P(A|BC)$ is 0 or 1) condition (i) needs to be "reinforced" by the requirement that also their zero-layers must be equal, otherwise we can meet critical situations (see, e.g. [6]).

Observation 1 *Even if different agreeing classes generated by the restriction of P on $\mathcal{D}$ may give rise to different zero-layers, it has been proved in [5, 6] that condition (ii) of Definition 2 either holds for all the agreeing classes of P or for none of them.*

Notice that for every event A this notion of stochastic independence is always irreflexive (also when the probability of A is 0 or 1) because $\circ(A|A) = 0$, while $\circ(A|A^c) = \infty$. Moreover, conditional independence of two possible events A and B imply the *logical independence* of A and B, i.e. all the events of the kind $A^* \wedge B^*$ is possible, with A^* - analogously B^* - is either A or A^c. (see [4]).

In [4, 16] theorems characterizing stochastic and conditional independence of two logically independent events A and B in terms of probabilities $P(B|C), P(B|AC)$ and $P(B|A^cC)$ is given, giving up any direct reference to the zero-layers.

Theorem 2 *Let A, B be two events logically independent with respect to the event C. If P is a coherent conditional probability such that $P(A|BC) = P(A|B^cC)$, then $A \perp\!\!\!\perp_{cs} B \mid C$ if and only if one of the following conditions holds:*

(a) $0 < P(A|BC) < 1$*;*

(b) $P(A|BC) = 0$ *and the extension of P to $B|C$ and $B|AC$ satisfies one of the following conditions*

1. $P(B|C) = 0, P(B|AC) = 0,$

2. $P(B|C) = 1, P(B|AC) = 1,$

3. $0 < P(B|C) < 1, 0 < P(B|AC) < 1$*;*

(c) $P(A|BC) = 1$ *and the extension of P to $B|C$ and $B|A^cC$ satisfies one of the following conditions*

1. $P(B|C) = 0, P(B|A^cC) = 0,$

2. $P(B|C) = 1, P(B|A^cC) = 1,$

3. $0 < P(B|C) < 1, 0 < P(B|A^cC) < 1.$

Indeed, in [16] the definition of cs-independence has been extended to the case of finite sets of events and to finite random variables.

Definition 3 *Let $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ be three different partitions of Ω such that $\mathcal{E}_2$ is not trivial. The partition $\mathcal{E}_1$ is stochastically independent of $\mathcal{E}_2$ given $\mathcal{E}_3$ with respect to a coherent conditional probability P (in symbols $\mathcal{E}_1 \perp\!\!\!\perp_{cs} \mathcal{E}_2 | \mathcal{E}_3\ [P]$) iff $C_{i_1} \perp\!\!\!\perp_{cs} C_{i_2} | C_{i_3}\ [P]$ for every $C_{i_1} \in \mathcal{E}_1, C_{i_2} \in \mathcal{E}_2, C_{i_3} \in \mathcal{E}_3$ such that $C_{i_2} \wedge C_{i_3} \neq \emptyset$.*

Let $X = (X_1, \ldots, X_n)$ be a random vector with values in $R_X \subseteq \mathbb{R}^n$. The partition $\mathcal{E}$ of the sure event Ω generated by X is denoted by $\mathcal{E}_X = \{X = x : x \in R_X\}$.

Definition 4 *Let (X, Y, Z) be a finite discrete random vector with values in $R \subseteq R_X \times R_Y \times R_Z$ and $\mathcal{E}_X$, $\mathcal{E}_Y$, $\mathcal{E}_Z$ be the partitions generated by X, Y and Z, respectively. Let P be a coherent conditional probability on $\mathcal{F}$ containing $\{A|BC : A \in \mathcal{E}_X, B \in \mathcal{E}_Y, C \in \mathcal{E}_Z\}$: then X is stochastically cs-independent of Y given Z with respect to P (in symbol $X \perp\!\!\!\perp_{cs} Y|Z\ [P]$) iff $\mathcal{E}_X \perp\!\!\!\perp_{cs} \mathcal{E}_Y | \mathcal{E}_Z\ [P]$.*

Note that in Definition 4 it is not required that the domain of the random vector (X, Y, Z) must be $R = R_X \times R_Y \times R_Z$, so logical constraints among the variables can be considered.

The set $\mathcal{M}_P$ of cs-independence statements induced by a coherent conditional probability P of the form $X_I \perp\!\!\!\perp_{cs} X_J | X_K$, where I, J and K are three disjoint subsets, is called *cs-independence model*.

Every cs-independence model induced by P is closed with respect to the following properties (for the proof see [16]):

Decomposition property

$X_I \perp\!\!\!\perp_{cs} [X_J, X_K] | X_W\ [P] \Longrightarrow X_I \perp\!\!\!\perp_{cs} X_J | X_W\ [P];$

Reverse decomposition property

$[X_I, X_J] \perp\!\!\!\perp_{cs} X_W | X_K\ [P] \Rightarrow X_I \perp\!\!\!\perp_{cs} X_W | X_K\ [P]$;

Weak union property

$X_I \perp\!\!\!\perp_{cs} [X_J, X_K] | X_W\ [P] \Rightarrow X_I \perp\!\!\!\perp_{cs} X_J | [X_W, X_K]\ [P]$;

Contraction property

$X_I \perp\!\!\!\perp_{cs} X_W | [X_J, X_K]\ [P] \,\&\, X_I \perp\!\!\!\perp_{cs} X_J | X_K\ [P] \Rightarrow\ X_I \perp\!\!\!\perp_{cs} [X_J, X_W] | [X_K]\ [P]$;

Reverse contraction property

$X_I \perp\!\!\!\perp_{cs} X_W | [X_J, X_K]\ [P] \,\&\, X_J \perp\!\!\!\perp_{cs} X_W | X_K\ [P]\ \Rightarrow\ [X_I, X_J] \perp\!\!\!\perp_{cs} X_W | [X_K]\ [P]$;

Intersection property

$X_I \perp\!\!\!\perp_{cs} X_J | [X_W, X_K]\ [P]\ \&\, X_I \perp\!\!\!\perp_{cs} X_W | [X_J, X_K]\ [P]\ \Rightarrow\ X_I \perp\!\!\!\perp_{cs} [X_J, X_W] | [X_K]\ [P]$;

Reverse intersection property

$X_I \perp\!\!\!\perp_{cs} X_W | [X_J, X_K]\ [P] \,\&\, X_J \perp\!\!\!\perp_{cs} X_W | [X_I, X_K]\ [P] \Rightarrow\ [X_I, X_J] \perp\!\!\!\perp_{cs} X_W | [X_K]\ [P]$.

Hence, these models satisfy all graphoid properties (see [14],[15]) except the symmetry property

$$X_I \perp\!\!\!\perp_{cs} X_J | X_K\ [P] \Rightarrow\ X_J \perp\!\!\!\perp_{cs} X_I | X_K\ [P]$$

and reverse weak union property

$$[X_J, X_W] \perp\!\!\!\perp_{cs} X_I | [X_K]\ [P] \Rightarrow X_J \perp\!\!\!\perp_{cs} X_I | [X_W, X_K]\ [P].$$

In [16] the models closed with respect to reverse weak union property, but not necessarily with respect to symmetry, (called *a-graphoid*) were classified. The possible lack of symmetry is not counterintuitive (see [4, 6]). Obviously, when the probability P is strictly positive on possible events, the cs-independence model induced by P is closed with respect to graphoid properties.

3 Basic graphical concepts

A *l-graph* G is a triplet $(V, E, \mathcal{B})$, where V is a finite set of *vertices*, E is a set of *edges* (i.e. a subset of ordered pairs of distinct vertices of $V \times V \setminus \{(v, v)\ : v \in V\}$) and $\mathcal{B}$ is a family (possibly empty) of subsets of vertices. The elements of the family $\mathcal{B} = \{B\ ,\ \ B \subseteq V\}$ are represented graphically by boxes enclosing the vertices in B. If $\mathcal{B}$ is empty, then the l-graph is a graph.

The attention in the sequel will be focused on *directed acyclic* l-graphs, and to introduce this kind of l-graphs we need to recall some basic notion from graph theory. A directed l-graph is a l-graph whose set of vertices E satisfies the following property: $(u, v) \in E \Rightarrow (v, u) \notin E$. A directed edge $(u, v) \in E$ is represented by an arrow pointing from u to v, $u \to v$. We say that u is a *parent* of v and v a *child* of u. The set of parents of v is denoted by $pa(v)$ and the set of children of u by $ch(u)$.

A *path* from u to v is a sequence of distinct vertices $u = u_1, \dots, u_n = v$, $n \geq 1$ such that either $u_i \rightarrow u_{i+1}$ or $u_{i+1} \rightarrow u_i$ for $i = 1, \dots, n-1$. A *directed path* from u to v is a sequence $u = u_1, \dots, u_n = v$ of distinct vertices such that $u_i \rightarrow u_{i+1}$ for all $i = 1, \dots, n-1$. If there is a directed path from u to v, we say that u is an ancestor of v or v a descendant of u and we write $u \mapsto v$. The symbols $an(v)$ and $ds(u)$ denote the set of *ancestors* of v and the set of *descendants* of u (vertices that $u \in an(v)$ and $v \in ds(u)$), respectively. Note that, according to our definition, a sequence consisting of one vertex is a directed path of length 0, and therefore every vertex is its own descendent and ancestor, i.e. $u \in an(u), u \in ds(u)$.

A *reverse directed path* from u to v is a sequence $u = u_1, \dots, u_n = v$ of distinct vertices such that $u_i \leftarrow u_{i+1}$ for all $i = 1, \dots, n-1$.

A *n-cycle* is a sequence of $u_1, \dots, u_n$, with $n > 3$, such that $u_n \rightarrow u_1$ and $u_1, \dots, u_n$ is a directed path. A directed graph is *acyclic* if it contains no cycles.

Given an acyclic directed graph G, the relation $\mapsto$ defines a *partial ordering* $\prec_G$ on the set of vertices, in particular for any $u, v \in V$ we have that if $u \in an(v)$, then $u \prec_G v$, while if $u \in ds(v)$, then $v \prec_G u$.

L-graphs and logical constraints. In Section 2 the relationship between logical independence and stochastic cs-independence has been shown, so we need to visualize which variables are linked by a logical constraint, and for this purpose we refer to the family $\mathcal{B}$ of subsets of vertices. Since, given a random vector $X = (X_1, \dots, X_n)$, a vertex i is associated with each random variable X_i, by means of the boxes $B \in \mathcal{B}$, we visualize the sets of random variables linked by a logical constraint (more precisely, a logical constraint involves the events of the partitions generated by the random variables). Recall that the partitions $\mathcal{E}_1, \dots, \mathcal{E}_n$ are *logically independent* if for every choice $C_i \in \mathcal{E}_i$, with $i = 1, ..., n$, the conjunction $C_1 \wedge \dots \wedge C_n \neq \emptyset$.

Obviously, if n partitions are logically independent, then arbitrary subsets of these partitions are logically independent.

However, n partitions $\mathcal{E}_1, \dots, \mathcal{E}_n$ need not be logically independent, even if every $n-1$ partitions can be logically independent; it follows that there is a *logical constraint* such that an event of the kind $C_1 \wedge \dots \wedge C_n$ is impossible, with $C_i \in \mathcal{E}_i$. For example, suppose $\mathcal{E}_1 = \{A, A^c\}$, $\mathcal{E}_2 = \{B, B^c\}$ and $\mathcal{E}_3 = \{C, C^c\}$ are three distinct partitions of Ω with $A \wedge B \wedge C = \emptyset$. All the couples of that partitions are logically independent, but they are not logically independent. Actually, the partition $\mathcal{E}_1$ is not logically independent of the partition generated by $\{\mathcal{E}_2, \mathcal{E}_3\}$. The same conclusion is reached replacing $\mathcal{E}_1$ by $\mathcal{E}_2$ or $\mathcal{E}_3$.

Given n partitions and some logical constraints among such partitions, it is possible, for each constraint, to find the *minimal subset* $\{\mathcal{E}_1, \dots, \mathcal{E}_k\}$ of partitions generating it. Actually, $\mathcal{E}_1, \dots, \mathcal{E}_k$ are such that $C_1 \wedge \dots \wedge C_k = \emptyset$, with $C_i \in \mathcal{E}_i$, and, in addition, for all $j = 1, \dots, k$, $C_1 \wedge \dots \wedge C_{j-1} \wedge C_{j+1} \wedge \dots \wedge C_k \neq \emptyset$. Such set of partitions $\{\mathcal{E}_1, \dots, \mathcal{E}_k\}$ is said the *minimal set* generating the given logical constraint, and it is singled-out graphically by the box $B = \{1, \dots, k\}$, which includes exactly the vertices associated to the corresponding random variables $X_1,\dots,X_k$. Then, in the sequel we call the boxes $B \in \mathcal{B}$ *logical components*.

4 Separation criterion for directed acyclic graphs

To represent conditional cs-independence models we need to recall L-separation criterion . In fact, the classic separation criterion for directed acyclic graphs (see [14]), known as d-separation (where d stands for directional), is not suitable for our purposes, because it induces a graphoid structure, and so it is not useful to describe a model where symmetry property may not hold (see Example 1).

Definition 5 *Let G be an acyclic directed graph. A path $u_1,\dots,u_n$, $n \geq 1$ in G is blocked by a set of vertices $S \subset V$, whenever there exists $1 < i < n$ such that one of the following three condition holds:*

1. $u_{i+1} \rightarrow u_i \rightarrow u_{i-1}$ *(i.e. u_{i-1}, u_i, u_{i+1} is the reverse directed path) and $u_i \in S$*
2. $u_{i-1} \leftarrow u_i \rightarrow u_{i+1}$ *and $u_i \in S$*
3. $u_{i-1} \rightarrow u_i \leftarrow u_{i+1}$ *and $ds(u_i) \notin S$*

The three conditions of Definition 5 are illustrated by Figure 1 (the grey vertices belong to S).

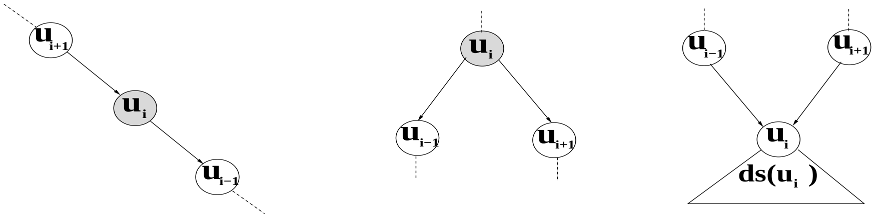


Figure 1: Blocked paths

Note that the definition of blocked path strictly depends on the direction of the path, in fact the main difference between our notion and that used in d-separation criterion [14] consists essentially in condition 1. of Definition 5. The path u_{i-1}, u_i, u_{i+1} drawn in the left-side of Figure 1 is blocked by u_i, while its reverse is not blocked by u_i because of the direction. Hence, the reverse path of a blocked one is not necessarily blocked according to our definition, so the blocking path notion does not satisfy the symmetry property.

The second and third cases of Definition 5 are like in d-separation criterion.

Definition 6 *Let G be a directed acyclic l-graph and let U, W and S be three pairwise disjoint sets of vertices of V. We say that U is L-separated from W by S in G and write symbol $(U,W|S)^l_G$, whenever every path in G from U to W is blocked by S and moreover, the following "logical separation" condition holds*

$$\forall B \in \mathcal{B} \text{ s.t. } B \subseteq U \cup W \cup S \text{ one has either } B \cap U = \emptyset \text{ or } B \cap W = \emptyset. \qquad (2)$$

Figure 2 clarifies when condition (2) holds (the set of vertices V_i and S are represented as ovals).

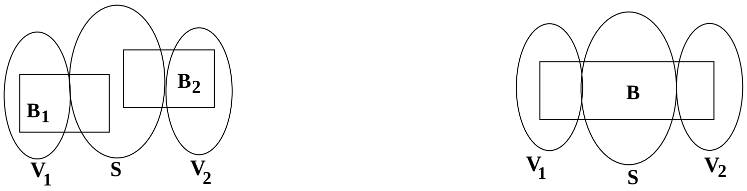


Figure 2: Representation of logical components: in the left-side V_1 and V_2 are not connected, in the right-side they are connected by B

Since the notion of blocked path is not necessarily symmetric, it follows that $(U,W|S)^l_G \not\Rightarrow (W,U|S)^l_G$. Actually, the lack of symmetry property depends on the notion of blocked path and not on the condition of logical separation (2).

Theorem 3 *[17] Let $G = (V,E,\mathcal{B})$ be a graph. The following properties hold*

1. *(Decomposition property)*
$$(U,W \cup Z|S)^l_G \Longrightarrow (U,W|S)^l_G$$
2. *(Reverse decomposition property)*
$$(U \cup Z,W|S)^l_G \Longrightarrow (U,W|S)^l_G$$
3. *(Weak union property)*
$$(U,W \cup Z|S)^l_G \Longrightarrow (U,W|Z \cup S)^l_G$$
4. *(Reverse weak union property)*
$$(U \cup Z,W|S)^l_G \Longrightarrow (U,W|Z \cup S)^l_G.$$
5. *(Contraction property)*
$$(U,W|S)^l_G \ \& \ (U,Z|W \cup S)^l_G \Longrightarrow (U,W \cup Z|S)^l_G$$
6. *(Reverse contraction property)*
$$(U,W|S)^l_G \ \& \ (Z,W|U \cup S)^l_G \Longrightarrow (U \cup Z,W|S)^l_G$$
7. *(Intersection property)*
$$(U,W|Z \cup S)^l_G \ \& \ (U,Z|W \cup S)^l_G \Longrightarrow (U,W \cup Z|S)^l_G$$
8. *(Reverse intersection property)*
$$(U,W|Z \cup S)^l_G \ \& \ (Z,W|U \cup S)^l_G \Longrightarrow (U \cup Z,W|S)^l_G$$

5 Minimal I-map

Given an independence model $\mathcal{M}$ over a set of variables (possibly) linked by a set of logical constraints, we look for a directed acyclic l-graph G describing all the statements T in $\mathcal{M}$ and localizing the set of variables involved in some logical constraint. But, generally, it is not always feasible to have such graph G (i.e. describing all the independence statements) for a given $\mathcal{M}$ as shown by the following example.

Example 1. Let (X_1, X_2, X_3, X_4) be a random vector such that the range of X_i is $\{0,1\}$, let us denote $A_i = (X_i = 1)$ (so $A_i^c = (X_i = 0)$), and suppose that $A_1 \subset A_2$.

Consider the following coherent conditional probability

$P(A_1A_2) = \frac{1}{5}$, $P(A_1^cA_2) = \frac{3}{10}$, $P(A_1^cA_2^c) = \frac{1}{2}$,

$P(A_3A_4|A_1A_2) = P(A_3A_4|A_1^cA_2) = P(A_3A_4^c|A_1A_2) = P(A_3A_4^c|A_1^cA_2) = 0$,

$P(A_3^cA_4|A_1A_2) = \frac{2}{5} = P(A_3^cA_4|A_1^cA_2)$,

$P(A_3^cA_4^c|A_1A_2) = \frac{3}{5} = P(A_3^cA_4^c|A_1^cA_2)$,

$P(A_4|A_2A_3) = \frac{2}{5}$, $P(A_4|A_2^cA_3) = \frac{3}{20}$, $P(A_2|A_3) = \frac{1}{5}$,

$P(A_1|A_2A_3A_4) = \frac{1}{2}$, $P(A_1|A_2A_3A_4^c) = \frac{2}{5}$.

Since $P(A_1|A_2) = \frac{2}{5}$, it follows from condition (b) 3. of Theorem 2 the validity of the statements $A_3A_4 \perp\!\!\!\perp_{cs} A_1|A_2$ and $A_3A_4^c \perp\!\!\!\perp_{cs} A_1|A_2$; moreover from condition (a) of the same theorem it follows that also $A_3^cA_4 \perp\!\!\!\perp_{cs} A_1|A_2$ and $A_3^cA_4^c \perp\!\!\!\perp_{cs} A_1|A_2$ hold, so we have (by Definition 3 and Definition 4) that $(X_3, X_4) \perp\!\!\!\perp_{cs} X_1|X_2$.

While, the statement $X_1 \perp\!\!\!\perp_{cs} (X_3, X_4)|X_2$ does not hold under P, in fact we have $P(A_1|A_2A_3A_4) = \frac{1}{2} \neq P(A_1|A_2)$.

The validity of the two conditional independence statements $X_3 \perp\!\!\!\perp_{cs} X_4|X_2$ and $X_4 \perp\!\!\!\perp_{cs} X_3|X_2$ follows from these equalities $P(A_3|A_2A_4) = 0 = P(A_3|A_2A_4^c)$ and $P(A_4|A_2) = 0.4 = P(A_4|A_2A_3) = P(A_4|A_2A_3^c)$.

Note that $P(A_3|A_4) = 0 = P(A_3|A_4^c)$ and $P(A_4) = 0.2 = P(A_4|A_3) = P(A_4|A_3^c)$, so $X_4 \perp\!\!\!\perp_{cs} X_3$ and its symmetric statement hold under P.

Therefore, the independence model $\mathcal{M}_P$ (which has a-graphoid structure) contains the statements $(X_3, X_4) \perp\!\!\!\perp_{cs} X_1|X_2$, $X_3 \perp\!\!\!\perp_{cs} X_4|X_2$, $X_3 \perp\!\!\!\perp_{cs} X_4|(X_1, X_2)$; $X_4 \perp\!\!\!\perp_{cs} X_3|X_2$, $X_4 \perp\!\!\!\perp_{cs} X_3|(X_1, X_2)$, $X_3 \perp\!\!\!\perp_{cs} X_4$, $X_4 \perp\!\!\!\perp_{cs} X_3$.

Note that $\mathcal{M}_P$ is not completely representable by a directed acyclic l-graph.

Hence, we need to introduce, analogously as in [14], the notion of I-map.

Definition 7 *A directed acyclic l-graph G is an I-map for a given independence model $\mathcal{M}$ iff every independence statement represented by means of L-separation criterion in G is also in $\mathcal{M}$.*

Thus an I-map G for $\mathcal{M}$ may not represent every statement of $\mathcal{M}$, but the ones it represents are actually in $\mathcal{M}$, it means that the set $\mathcal{M}_G$ of statements described by G is contained in $\mathcal{M}$.

An I-map G for $\mathcal{M}$ is said *minimal* if removing any arrow from the l-graph G the obtained l-graph will no longer be an I-map for $\mathcal{M}$.

Given an independence model $\mathcal{M}$ over a random vector $(X_1,...,X_n)$, let $\pi = (\pi_1,...,\pi_n)$ be any ordering of the given variables, and, in addition, for any j, let $U_{\pi_j} = \{\pi_1,...,\pi_{j-1}\}$ be the set of indexes before π_j, and D_{π_j} the minimal subset of U_{π_j} such that $X_{\pi_j} \perp\!\!\!\perp_{cs} X_{R_{\pi_j}} | X_{D_{\pi_j}}$ where $R_{\pi_j} = U_{\pi_j} \setminus D_{\pi_j}$; moreover, let $W_{\pi_j} = \{v \in U_{\pi_j} : v \in D_{\pi_k} \cap D_{\pi_i}, i \neq k, i \leq j, k \leq j\}$ and S_{π_j} the maximal subset of U_{π_j} such that $X_{S_{\pi_j}} \perp\!\!\!\perp_{cs} X_{\pi_j} | X_{W_{\pi_j}}$.

The subset $\Theta_\pi = \{X_{\pi_j} \perp\!\!\!\perp_{cs} X_{R_{\pi_j}} | X_{D_{\pi_j}}, X_{S_{\pi_j}} \perp\!\!\!\perp_{cs} X_{\pi_j} | X_{W_{\pi_j}} : j = 1,...n\}$ is said the *basic list* of $\mathcal{M}$ relative to π. From the basic list Θ_π and the set of logical components $\mathcal{B}$, a directed acyclic l-graph G (related to π) is obtained by drawing the boxes $B \in \mathcal{B}$ and designating D_{π_j} as parents of vertex π_j (for any vertex $v \in D_{\pi_j}$, an arrow goes from v to π_j), moreover, for any vertex $\pi_i \in U_{\pi_j} \setminus S_{\pi_j}$ such that $\pi_i \in ds(w)$, with $w \in W_{\pi_j}$, but $\pi_i \notin an(\pi_j)$ draw an arrow from π_i to π_j.

This construction of G from the basic list differs from the classic construction given for directed acyclic graphs with d-separation [14] essentially for the second part, which is useful to avoid the introduction of symmetric statements not in the given independence model. For example, consider the independence model $\mathcal{M} = \{X_1 \perp\!\!\!\perp_{cs} X_3 | X_2\}$ and considering the ordering $\pi = (2,3,1)$, the related directed acyclic l-graph is obtained following these steps: draw an arrow from 2 to 3, then consider the vertex 1 and draw an arrow from 2 to 1; now since $3 \in ds(2)$ (i.e. $D_3 = \{2\}$), but $3 \notin an(1)$ and, since the statement $X_3 \perp\!\!\!\perp_{cs} X_1 | X_2$ is not in $\mathcal{M}$, we must draw an arrow from 3 to 1.

Now, we must prove that such directed acyclic l-graph obtained from the basic list Θ_π is an I-map for $\mathcal{M}$.

Theorem 4 *Let $\mathcal{M}$ be an independence model over a set of random variables linked by a set of logical constraints. Given an ordering π on the random variables, if $\mathcal{M}$ is an a-graphoid, then the directed acyclic l-graph G generated by the basic list Θ_π is an I-map for $\mathcal{M}$.*

Proof: For an a-graphoid of one variable it is obvious that the directed acyclic l-graph is an I-map. Suppose for a-graphoid structure with less than k variables that the directed acyclic l-graph is an I-map.

Let $\mathcal{M}$ be an independence model under k variables. Given an ordering π on the variables, let X_n be the last variable according to π (n denotes the vertex in G associated to X_n), $\mathcal{M}'$ the a-graphoid formed by removing all the independence statements involving X_n from $\mathcal{M}$ and G' the directed acyclic l-graph formed by removing n and all the arrows going to n (they cannot depart from n because is the last vertex) in G.

Since X_n is the last variable in the ordering π, it cannot appear in any set of parents D_{π_j} (with $j < k$), and the basic list $\Theta' = \Theta \setminus \{X_n \perp\!\!\!\perp_{cs} X_{R_n} | X_{D_n}\}$ generates $\mathcal{G}'$. Since $\mathcal{M}'$ has $k-1$ variables, $\mathcal{G}'$ is an I-map of it.

G is an I-map of $\mathcal{M}$ iff the set $\mathcal{M}_G$ of the independence statements represented in G by L-separation criterion is also in $\mathcal{M}$.

If X_n does not appear in T, then, being $T = (X_I \perp\!\!\!\perp_{cs} X_J | X_K) \in \mathcal{M}_G$, T must be represented also in $\mathcal{G}'$, if it were not, then there would be a path in $\mathcal{G}'$ from I to J that is not blocked (according to L-separation) by K. But then it must be not blocked also in G, since the addition of a vertex and some arrows going to the new vertex cannot block a path. Since $\mathcal{G}'$ is an I-map for of $\mathcal{M}'$, T must be an element of it, but $\mathcal{M}' \subset \mathcal{M}$, so $T \in \mathcal{M}$.

Otherwise (if X_n appears in T), T falls into one of the following three situations:

1. suppose that $T = ((X_I, X_n) \perp\!\!\!\perp_{cs} X_J | X_K) \in \mathcal{M}_G$, let $X_n \perp\!\!\!\perp_{cs} X_{R_n} | X_{D_n} \in \mathcal{M}$ (by construction). Obviously J and D_n have no vertices in common, otherwise we would have a path from a vertex in $j \in J \cap D_n$ pointing to n, so by L-separation n would not be separated from J given K in G.

   Since there is an arrow from every vertex in D_n to n and every path from n to J is blocked by K in G, then every path from D_n to J must be blocked by K in G. Therefore, every path from both D_n and I to J are blocked by K in G. Now, if there is a logical component $B \in \mathcal{B}$ such that $B \subseteq D_n \cup I \cup J \cup K$ and both $B \cap (D_n \cup I)$ and $B \cap J$ are not empty, then remove a suitable vertex in B from D_n, w.l.g. Hence, the statement $(X_I, X_{D_n}) \perp\!\!\!\perp_{cs} X_J | X_K$ belongs to $\mathcal{M}_G$. This statement does not contain the variable X_n, hence, being G' an I-map for $\mathcal{M}' \subset \mathcal{M}$, then $(X_I, X_{D_n}) \perp\!\!\!\perp_{cs} X_J | X_K \in \mathcal{M}$.

   Since $\mathcal{M}$ is closed under a-graphoid properties, (by weak union property) $X_n \perp\!\!\!\perp_{cs} X_J | (X_I, X_{D_n}, X_K) \in \mathcal{M}$ and it follows $(X_I, X_{D_n}, X_n) \perp\!\!\!\perp_{cs} X_J | X_K \in \mathcal{M}$ (using reverse contraction property), so $(X_I, X_n) \perp\!\!\!\perp_{cs} X_J | X_K \in \mathcal{M}$ by decomposition property.

2. suppose that $T = (X_I \perp\!\!\!\perp_{cs} (X_J, X_n) | X_K) \in \mathcal{M}_G$, it means, by definition of L-separation and from the assumption that n is the last vertex in the ordering, that every path going from I to $J \cup n$ is L-separated by K. Therefore, if there is no path as in condition 1. of Definition 5, then in the remaining two cases, also the statement $T_1 = ((X_J, X_n) \perp\!\!\!\perp_{cs} X_I | X_K) \in \mathcal{M}_G$, so the proof goes in the same line of that in step 1.

   Otherwise, (if there is a path as in condition 1 of Definition 5), then $I \not\subseteq an(n)$. Therefore, there is a subset $W_n \subseteq U_n$ such that every path between n and $I \cup K$ is blocked by W_n. Note that, $W_n = W^1 \cup W^2$ (W^1 or W^2 can be empty) with $W^2 \subseteq D_n$ and $W^1 \subseteq an(D_n)$. Moreover, let $J = J^1 \cup J^2 \cup J^3$ (J^1 or J^2 or J^3 can be empty) with $J^1 \subseteq ds(W) \cap an(K)$, while $J^2 \subseteq W$ and $J^3 = J \setminus (J^1 \cup J^2)$, so for any $j \in J^3$ one has that either $j \in an(W)$ or $j \in ds(W) \cap an(n)$.
   By construction, one has that every path between $n \cup J^3$ and $I \cup K \cup J^1$ is

blocked by W_n. Hence, one has that $(X_I, X_K, X_{J^1}) \perp\!\!\!\perp_{cs} (X_n, X_{J^3})|X_{W_n}$ and its symmetric statement belong to $\mathcal{M}$.

Therefore, one has $X_I \perp\!\!\!\perp_{cs} (X_n, X_{J^3})|(X_{W_n}, X_K, X_{J^1}) \in \mathcal{M}$ by weak union property. Since also $T_2 = (X_I \perp\!\!\!\perp_{cs} (X_{W_n}, X_{J^1})|X_K) \in \mathcal{M}_G$ and since that statement T_2 does not involve n, $T_2 \in \mathcal{M}$, so the statement $X_I \perp\!\!\!\perp_{cs} (X_n, X_{W_n}, X_{J_1}, X_{J_3})|X_K)$ belong to $\mathcal{M}$ (by contraction property), and it follows that $X_I \perp\!\!\!\perp_{cs} (X_n, X_J)|X_K$ belongs to $\mathcal{M}$ (by reverse decomposition).

3. suppose that $T = (X_I \perp\!\!\!\perp_{cs} X_J|(X_K, X_n)) \in \mathcal{M}_G$. It must be the case that I is L-separated by J given K in G for if it were not, then there would be a path from some vertex in I to some vertex in J not passing trough K. But I is separated by J given n and K, so this path would pass through n; but n is the last vertex in the ordering, so all arrows go on it. Hence, it cannot block any unblocked path, and so $T_1 = (X_I \perp\!\!\!\perp_{cs} X_J|X_K) \in \mathcal{M}_G$.

   The statements T_1 and T imply that either $(X_I, X_n) \perp\!\!\!\perp_{cs} X_J|X_K$ or $X_I \perp\!\!\!\perp_{cs} (X_J, X_n)|X_K$ holds in G: in fact, if both I and J are connected to n, since n is the last vertex (from n an arrow cannot leave), then there is a directed path from I to n and another from J to n, so that one would get $X_I \perp\!\!\!\perp_{cs} X_J|(X_K, X_n) \notin \mathcal{M}_G$. So, the conclusion follows by step 1 and 2. □

Example 1 (continued) – The following pictures show the minimal I-map obtained by means of the proposed procedure for two possible orderings: $(1,2,3,4)$ on the left-side and $(3,4,1,2)$ on the right-side

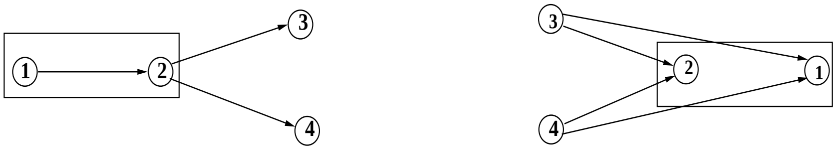


Figure 3: Two possible I-Maps for the independence model $\mathcal{M}_P$ of Example 1

Actually, the picture in the left-side represents the independence statements $(X_3, X_4) \perp\!\!\!\perp_{cs} X_1|X_2$, $X_3 \perp\!\!\!\perp_{cs} X_4|X_2$, $X_4 \perp\!\!\!\perp_{cs} X_3|X_2$ and those implied by a-graphoid properties; while that one on the right-side describes the statement $X_3 \perp\!\!\!\perp_{cs} X_4$ and its symmetric one. Note that these two graphs actually are minimal I-maps; in fact removing any arrow from them, we may read independence statements not in $\mathcal{M}_P$. The block $B = \{1,2\}$ localizes the logical constraint $A_1 \subset A_2$.

If for a given independence model over n variables there exists a perfect map G, then (at least) one of $n!$ orderings among the variables will generate the l-graph G. More precisely, such orderings, which give rise to G, are all the orderings compatible with the partial order induced by G.

6 Conclusions

The L-separation criterion for directed acyclic graphs has been recalled together with its main properties. This is very useful for effective description of independence models induced by different uncertainty measures [1, 2, 4, 5, 6, 13, 16, 18, 19]. In fact, these models cannot be represented efficiently by the well-known graphical models [12, 14], because the related separation criteria satisfy the symmetry property.

In this paper, we have considered the L-separation criterion introduced in [16], which satisfies asymmetric graphoid properties. We have shown that for some independence models there is not a perfect map even using L-separation criterion.

Therefore, the notion of minimal I-map has been redefined in this context and we have shown how to build it given an ordering on the variables. In addition, we have proved that for any ordering on the variables there is a minimal I-map for a given independence model obeying to asymmetric graphoid properties.

References

[1] B. Bouchon-Meunier, G. Coletti, C. Marsala. Independence and Possibilistic Conditioning. *Annals of Mathematics and Artificial Intelligence*, 35:107–124, 2002.

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

[3] G. Coletti. Coherent numerical and ordinal probabilistic assessments. *IEEE Trans. on Systems, Man, and Cybernetics*, 24(12):1747-1754, 1994.

[4] G. Coletti, R. Scozzafava. Zero probabilities in stochastic independence. In *Information, Uncertainty, Fusion* (Eds. B. Bouchon-Meunier, R. R. Yager, and L. A. Zadeh), Kluwer, Dordrecht 185-196, 2000. (Selected papers from IPMU ’98).

[5] G. Coletti, R. Scozzafava. Stochastic Independence in a Coherent Setting. *Annals of Mathematics and Artificial Intelligence*, 35:151–176, 2002.

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

[7] F.T. de Dombal, F. Gremy. *Decision Making and Medical Care.* North Holland, 1976.

[8] B. de Finetti. Sull’impostazione assiomatica del calcolo delle probabilitá. *Annali dell’Universitá di Trieste*, 19:3-55, 1949. (Eng. transl.: Ch. 5 in Probability, Induction, Statistics, London: Wiley, 1972).

[9] L.E. Dubins. Finitely additive conditional probabilities, conglomerability and disintegration. *Annals of Probability*, 3:89-99, 1975.

[10] D. Geiger, T. Verma, J. Pearl. Identifying Independence in Bayesian Networks. *Networks*, 20:507-534, 1990.

[11] J.R. Hill. Comment on “Graphical models”. *Statistical Science*, 8:258-261, 1993.

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

[13] S.Moral. Epistemic irrelevance on sets of desiderable gambles. In *Proc. of the Fird Int. Symposium on Imprecise Probabilities and Their Applications* (Eds. G. de Cooman, T.L. Fine, T. Seidenfeld), 247-254, 2001.

[14] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.

[15] M. Studeny, R.R. Bouckaert. On chain graph models for description of conditional independence structures. *The Annals of Statistics*, 26(4):1434-1495, 1998.

[16] B. Vantaggi. Conditional Independence in a Coherent Finite Setting. *Annals of Mathematic and Artificial Intelligence*, 32:287-314, 2001.

[17] B. Vantaggi. The L-separation criterion for description of cs-independence models. *International Journal of Approximate Reasoning*, 29:291-316, 2002.

[18] B. Vantaggi. Graphical models for conditional independence structures. In *Proc. of the Fird Int. Symposium on Imprecise Probabilities and Their Applications* (Eds. G. de Cooman, T.L. Fine, T. Seidenfeld), 332-341, 2001.

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

Barbara Vantaggi is with the Department of Metodi e Modelli Matematici per le Scienze Applicate at Universitá “La Sapienza” di Roma, Italy. E-mail: vantaggi@dmmm.uniroma1.it

Design of Iterative Proportional Fitting Procedure for Possibility Distributions[*]

JIŘINA VEJNAROVÁ
Laboratory for Intelligent Systems, Prague, Czech Republic

Abstract

We design an iterative proportional fitting procedure (parameterized by a continuous t-norm) for computation of multidimensional possibility distributions from its marginals, and discuss its basic properties.

Keywords

multidimensional possibility distributions, marginal problem, triangular norm, iterative proportional fitting procedure

1 Introduction

The complexity of practical problems that are of primary interest in the field of artificial intelligence usually results in the necessity to construct models with the aid of a great number of variables: more precisely, hundreds or thousands rather than tens. However, distributions of such dimensionality are usually not available; the global knowledge (joint distribution) must be integrated on the basis of its local pieces (marginal distributions). This problem type is often referred to as a marginal problem. More precisely, the marginal problem addresses the question of whether or not a common extension exists for a given set of marginal distributions.

In [14] we introduced a possibilistic marginal problem and found necessary and sufficient conditions, respectively. This contribution is a natural continuation of our work — it tries to solve a practical problem: how to compute the values of an extension. Its aim is to introduce a possibilistic version of Iterative Proportional Fitting Procedure and to discuss its basic properties.

Iterative Proportional Fitting Procedure (IPFP) was originally designed by Deming and Stephan [3] in 1940 for adjustment of frequencies in contingency tables. Later, IPFP was applied to several problems in different domains; e.g. for

[*] This work has been partially supported by the GAČR Grant No. 201/02/1269 and the GA AV ČR Grant No. A1075104.

maximum likelihood estimate in a hierarchical model, or for computation of values of joint probability distributions in a probabilistic expert system [6] (for other applications see [9]).

This contribution is organized as follows. First an overview, followed by the basic notions (Section 2); then in Section 3 we briefly recall a possibilistic marginal problem, introduce possibilistic IPFP and demonstrate, on a simple example, how its computations are performed. In Section 4 we find a sufficient condition for its convergence and present two counterexamples.

2 Basic Notions

The purpose of this section is to give, as briefly as possible, an overview of basic notions of De Cooman’s measure-theoretical approach to possibility theory [2], necessary for understanding the paper. We will start with the notion of a triangular norm, since most notions in this paper are parameterized by it.

Triangular Norms.

A *triangular norm* (or a *t-norm*) T is an isotonic, associative and commutative binary operator on $[0,1]$ (i.e. $T : [0,1]^2 \to [0,1]$) satisfying the boundary condition: for any $x \in [0,1]$

$$T(1,x) = x.$$

Let $x,y \in [0,1]$ and T be a t-norm. We will call an element $z \in [0,1]$ T*-inverse* of x w.r.t. y if

$$T(z,x) = T(x,z) = y. \tag{1}$$

It is obvious that if $x \leq y$ then there are no T-inverses of x w.r.t. y. The T*-residual* $y \triangle_T x$ of y by x is defined as

$$y \triangle_T x = \sup\{z \in [0,1] : T(z,x) \leq y\}.$$

A t-norm T is called *continuous* if T is a continuous function. Within this paper, we will only deal with continuous t-norms, since for continuous t-norms $y \triangle_T x$ is the greatest solution of the equation (1) in z (if it exists).

Example 1 The most important examples of continuous t-norms are:

(i) *Gödel’s t-norm:* $T_G(x,y) = \min(x,y)$;

(ii) *product t-norm:* $T_\Pi(x,y) = x \cdot y$;

(iii) *Lukasziewicz’s t-norm:* $T_L(x,y) = \max(0, x+y-1)$;

and the corresponding residuals for $x > y$ (otherwise $y \triangle_T x = 1$ for any t-norm):

(i) $y \triangle_{T_G} x = y$;

(ii) $y \triangle_{T_\Pi} x = \frac{y}{x}$;

(iii) $y \triangle_{T_L} x = y - x + 1$.

Because of its associativity, any t-norm T can be extended to an n-ary operator $T^n : [0,1]^n \to [0,1]$, namely in the following way

$$\begin{aligned} T^2(a_1,a_2) &= T(a_1,a_2), \\ T^n(a_1,\ldots,a_n) &= T(T^{n-1}(a_1,\ldots a_{n-1}),a_n), \end{aligned}$$

for $n \geq 3$.

Possibility Measures and Distributions.

Let $\mathbf{X}$ be a finite set called *universe of discourse* which is supposed to contain at least two elements. A *possibility measure* Π is a mapping from the power set $\mathcal{P}(\mathbf{X})$ of $\mathbf{X}$ to the real unit interval $[0,1]$ satisfying the following requirement: for any family $\{A_j, j \in J\}$ of elements of $\mathcal{P}(\mathbf{X})$

$$\Pi(\bigcup_{j\in J} A_j) = \max_{j \in J} \Pi(A_j)^1.$$

Π is called *normal* if $\Pi(\mathbf{X}) = 1$. Within this paper we will always assume that Π is normal.

For any Π there exists a mapping $\pi : \mathbf{X} \to [0,1]$, called a *distribution* of Π, such that for any $A \in \mathcal{P}(\mathbf{X})$, $\Pi(A) = \max_{x \in A} \pi(x)$. This function is a possibilistic counterpart of a density function in probability theory. In the remaining part of this contribution we will deal with distributions rather than with measures.

Let $\mathbf{X}_1$ and $\mathbf{X}_2$ denote two finite universes of discourse provided by possibility measures Π_1 and Π_2 (with distributions π_1 and π_2), respectively. The possibility distribution π on $\mathbf{X}_1 \times \mathbf{X}_2$ is called *T-product possibility distribution* of π_1 and π_2 if for any $(x_1,x_2) \in \mathbf{X}_1 \times \mathbf{X}_2$

$$\pi(x_1,x_2) = T(\pi_1(x_1), \pi_2(x_2)). \tag{2}$$

Considering an arbitrary possibility distribution π defined on a product universe of discourse $\mathbf{X} \times \mathbf{Y}$, its *marginal possibility distribution* on $\mathbf{X}$ is defined by the equality

$$\pi_X(x) = \max_{y \in \mathbf{Y}} \pi(x,y) \tag{3}$$

for any $x \in \mathbf{X}$.

Conditioning.

Let T be a t-norm on $[0,1]$. For any possibility measure Π on $\mathbf{X}$ with distribution π, we define the following binary relation on the set $\mathcal{G}(\mathbf{X}) = \{h : \mathbf{X} \longrightarrow [0,1]\}$ of all fuzzy variables on $\mathbf{X}$: For h_1 and h_2 in $\mathcal{G}(\mathbf{X})$ we say that h_1 and h_2 are (Π,T)*-equal almost everywhere* (and write $h_1 \overset{(\Pi,T)}{=} h_2$) if for any $x \in X$

$$T(h_1(x), \pi(x)) = T(h_2(x), \pi(x)).$$

[1] max must be substituted by sup if $\mathbf{X}$ is not finite.

This notion is very important for the definition of *conditional possibility distribution*, which is defined (in accordance with [2]) as *any* solution of the equation

$$\pi_{XY}(x,y) = T(\pi_Y(y), \pi_{X|_T Y}(x|_T y)), \tag{4}$$

for any $(x,y) \in \mathbf{X} \times \mathbf{Y}$. Continuity of a t-norm T guarantees the existence of a solution of this equation. This solution is not unique (in general), but the ambiguity vanishes when almost-everywhere equality is considered. We are able to obtain a representative of these conditional possibility distributions (if T is a continuous t-norm) by taking the residual $\pi_{XY}(x,\cdot) \triangle_T \pi_Y(\cdot)$ since

$$\pi_{X|_T Y}(x|_T \cdot) \overset{(\Pi_{\underline{Y}},T)}{=} \pi_{XY}(x,\cdot) \triangle_T \pi_Y(\cdot). \tag{5}$$

This way of conditioning brings a unifying view on several conditioning rules [4, 5, 7], i.e., its importance from the theoretical viewpoint is obvious. On the other hand, its practical meaning is not so substantial. Although De Cooman [2] claims that conditional distributions are never used *per se*, there exist situations in which it is necessary to be careful to choose an appropriate representative of the set of solutions (cf. Example 5 in [14]). Therefore, in this contribution we also use residuals rather than general conditionals.

Independence.

Two variables X and Y (taking their values in $\mathbf{X}$ and $\mathbf{Y}$, respectively) are *possibilistically T-independent* [2] if for any $F_X \in X^{-1}(\mathcal{P}(\mathbf{X}))$, $F_Y \in Y^{-1}(\mathcal{P}(\mathbf{Y}))$,

$$\begin{aligned}
\Pi(F_X \cap F_Y) &= T(\Pi(F_X), \Pi(F_Y)),\\
\Pi(F_X \cap F_Y^C) &= T(\Pi(F_X), \Pi(F_Y^C)),\\
\Pi(F_X^C \cap F_Y) &= T(\Pi(F_X^C), \Pi(F_Y)),\\
\Pi(F_X^C \cap F_Y^C) &= T(\Pi(F_X^C), \Pi(F_Y^C)),
\end{aligned}$$

where A^C denotes the complement of A.

From this definition it immediately follows that the independence notion is parameterized by T. More specifically, it means that if X and Y are independent with respect to Gödel’s t-norm, they need not be, for example, independent with respect to product t-norm. This fact is reflected in most definitions and assertions that follow.

In [11] we generalized this notion and in the following way: Given a possibility measure Π on $\mathbf{X} \times \mathbf{Y} \times \mathbf{Z}$ with the respective distribution $\pi(x,y,z)$, variables X and Y are *possibilistically conditionally T-independent*[2] given Z (in symbols $I_T(X,Y|Z)$) if, for any pair$(x,y) \in \mathbf{X} \times \mathbf{Y}$,

$$\pi_{XY|_T Z}(x,y|_T \cdot) \quad \overset{(\Pi_{\underline{Z}},T)}{=} \quad T(\pi_{X|_T Z}(x|_T \cdot), \pi_{Y|_T Z}(y|_T \cdot)). \tag{6}$$

[2] Let us note that a similar definition of conditional independence can be found in [8].

Let us stress again that we do not deal with the pointwise equality but with the *almost everywhere equality*. This definition unifies, in a sense, several notions of conditional noninteractivity and that of conditional independence (for more details see [12]). Although it may seem to be controversial from the epistemic point of view [1], it is very suitable for our purpose, since it is closely connected (for more details see [13]) with a principal notion of multidimensional models — the notion of factorization.

We will say that a possibility distribution π *factorizes*[3] with respect to a system $\mathcal{A}$ and a t-norm T, if, for all complete subsets $A \in \mathcal{A}$, there exist fuzzy variables f_A of x_A such that π has the form

$$\pi(x) = T^{|\mathcal{A}|}(f_{A_1}(x_{A_1}), \ldots, f_{A_{|\mathcal{A}|}}(x_{A_{|\mathcal{A}|}})). \tag{7}$$

The functions f_A are not uniquely determined (in general), since they can be "multiplied" in several ways, cf. Example 14 in [13].

3 Iterative Proportional Fitting Procedure

In this section we define (in the most general way) an iterative proportional fitting procedure for possibility distributions and show, on a simple example, how it works.

Before doing that, let us recall what is possiblistic marginal problem.

Possibilistic Marginal Problem.

Let us assume that $\mathbf{X}_i$, $i \in N$, $1 \leq |N| < \infty$ are finite universes of discourse, $\mathcal{K}$ is a system of nonempty subsets of N and $\mathcal{S} = \{\pi_K, K \in \mathcal{K}\}$ is a family of possibility distributions, where each π_K is a distribution on a product space

$$\mathbf{X}_K = \bigtimes_{i \in K} \mathbf{X}_i.$$

The problem we are interested in is the existence of an *extension*, i.e., a distribution π on

$$\mathbf{X} = \bigtimes_{i \in N} \mathbf{X}_i.$$

whose marginals are distributions from $\mathcal{S}$; or, more generally, the set

$$\mathcal{P} = \{\pi(x) : \pi(x_K) = \pi_K(x_K), K \in \mathcal{K}\}$$

is of interest.

The necessary condition (but not sufficient, as shown in [14]) for the existence of an extension is the pairwise projectivity of distributions from $\mathcal{S}$. Let us recall

[3] Factorization is usually defined with respect to a graph, but this definition is more appropriate for the purpose of this contribution.

that two possibility distributions π_I and π_J are *projective* if they have common marginals, i.e. if

$$\pi_I(x_{I\cap J}) = \pi_J(x_{I\cap J}).$$

Since IPFP is able to solve a marginal problem (if a solution exists) within a probabilistic setting, it seems to be useful to design an analogous procedure for possibility distributions.

Design of Iterative Proportional Fitting Procedure.

Let $\mathcal{S} = \{\pi_i, i = 1, \ldots m\}$ be a sequence of low-dimensional normal possibility distributions, which will be referred to as an *input sequence*. Let

$$\rho_{(0)} \in \mathcal{R} = \{\rho : \mathbf{X} \longrightarrow [0,1]; \max_{x\in\mathbf{X}} \rho(x) = 1\}$$

be an *initial possibility distribution*.

The *iterative proportional fitting procedure with respect to a t-norm T* (IPFP(T)) is a computational process defined for $x \in \mathbf{X}$ and for $j = 1, 2, \ldots$ and $k = (((j-1) \bmod m) + 1)$ by the following formula:

$$\rho_{(j)}(x) \quad = \quad T(\rho_{(j-1)}(x) \triangle_T \rho_{(j-1)}(x_{K_k}), \pi_k(x_{K_k})). \tag{8}$$

Formula (8) has the following meaning: at every step j we udate distribution $\rho_{(j-1)}$ simply by “multiplying” the marginal π_k, $k = (((j-1) \bmod m) + 1)$ by the residual of $\rho_{(j-1)}$ in order to obtain distribution $\rho_{(j)}$ such that

$$\rho_{(j)}(x_{K_k}) = \pi_k(x_{K_k}).$$

It is completely analogous to probability theory, where (8) has form

$$Q_{(j)}(x) = P_k(x_{K_k}) \frac{Q_{(j-1)}(x)}{Q_{(j-1)}(x_{K_k})},$$

which is a generalization of the original procedure by Deming and Stephan [3].

Example.

The following simple example illustrates how the computations of IPFP(T) are performed.

Example 2 Let X_1, X_2 and X_3 be three binary variables with values in $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$, respectively ($\mathbf{X}_1 = \mathbf{X}_2 = \mathbf{X}_3 = \{0,1\}$), and let the input sequence consist of two possibility distributions $\pi_1(x_1,x_2)$ and $\pi_2(x_2,x_3)$ on $\mathbf{X}_{\{1,2\}}$ and $\mathbf{X}_{\{2,3\}}$, respectively.

- The initial distribution $\rho_{(0)} \in \mathcal{R}$ is the least informative distribution on $\mathbf{X}_{\{1,2,3\}}$, i.e. $\rho_{(0)} \equiv 1$ (initial and input distributions can be found at Figure 1).

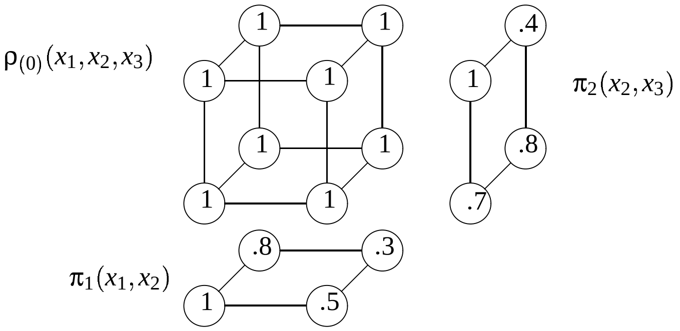


Figure 1: Initial and input distributions of IPFP(T)

- The operation of *fitting* the first input distribution $\pi_1(x_1,x_2)$ brings joint possibility distribution $\rho_{(1)}$ such that

$$\rho_{(1)}(x_1,x_2) = \pi_1(x_1,x_2),$$

as can be seen from Figure 2.

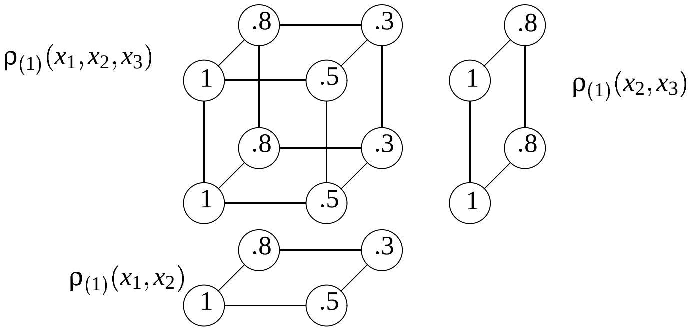


Figure 2: Joint distribution $\rho_{(1)}$ and its marginals after fitting π_1

- *Fitting* the second input distribution $\pi_2(x_2,x_3)$ gives the joint possibility distribution

$$\rho_2(x_1,x_2,x_3) = T(\pi_2(x_2,x_3),\rho_{(1)}(x_1,x_2,x_3) \triangle_T \rho_{(1)}(x_2,x_3))$$

with the property $\rho_{(2)}(x_2,x_3) = \pi_2(x_2,x_3)$ (cf. Figure 3).

From Figure 3 one can see that due to the projectivity of π_1 and π_2, $\rho_{(2)}$ preserves its marginal from previous step, i.e.

$$\rho_{(2)}(x_1,x_2) = \rho_{(1)}(x_1,x_2) = \pi_1(x_1,x_2).$$

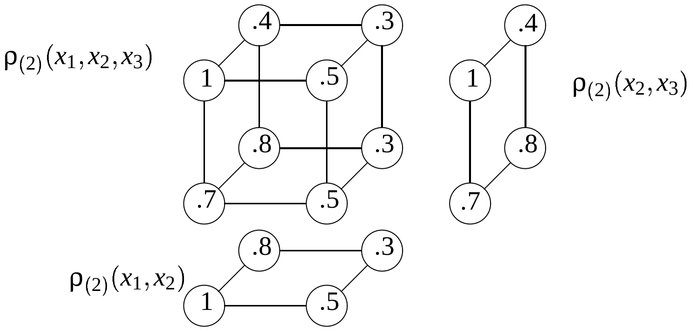


Figure 3: Joint distribution $\rho_{(2)}$ (with respect to Gödel’s t-norm) and its marginals after fitting π_1 and π_2

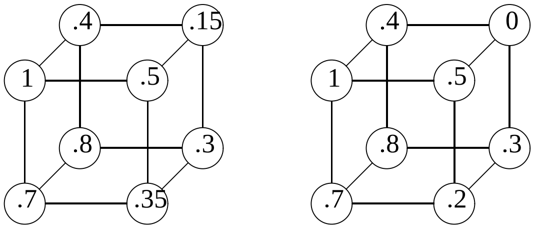


Figure 4: Joint distribution $\rho_{(2)}$ (with respect to product and Lukasziewicz’ t-norms, respectively) after fitting π_1 and π_2

It is evident that there is no reason to fit $\pi_1(x_1,x_2)$ again, since it cannot bring any change to $\rho_{(2)}$.

From this simple example, one can conclude that if the input set consists of two projective possibility distributions and the initial possibility distribution is $\rho_{(0)} \equiv 1$, IPFP(T) stops after one cycle for any continuous t- norm T. Nevertheless, the resulting distribution depends on the choice of the t-norm, which can be seen from Figures 3 and 4.

4 On Convergence of Possibilistic IPFP

In this section we will generalize the observation from the end of the foregoing section and find a sufficient condition for the convergence of possibilistic IPFP. Before doing that, let us briefly recall the notions of operators of composition of possibility distributions (introduced in [10]), which seem to be a useful technical tool for proofs.

Operators of Composition.

Considering a continuous t-norm T, two subsets K_1, K_2 of N and two normal possibility distributions $\pi_1(x_{K_1})$ and $\pi_2(x_{K_2})$,[4] we define the *operator of right composition* of these possibilistic distributions by the expression

$$\pi_1(x_{K_1}) \triangleright_T \pi_2(x_{K_2}) = T\left(\pi_1(x_{K_1}), \pi_2(x_{K_2}) \bigtriangleup_T \pi_2(x_{K_1\cap K_2})\right);$$

analogously the *operator of left composition* is defined by the expression

$$\pi_1(x_{K_1}) \triangleleft_T \pi_2(x_{K_2}) = T\left(\pi_1(x_{K_1}) \bigtriangleup_T \pi_1(x_{K_1\cap K_2}), \pi_2(x_{K_2})\right).$$

If $K_1 \cap K_2 = \emptyset$ then obviously

$$\pi_1(x_{K_1}) \triangleright_T \pi_2(x_{K_2}) = \pi_1(x_{K_1}) \triangleleft_T \pi_2(x_{K_2}) = T\left(\pi_1(x_{K_1}), \pi_2(x_{K_2})\right),$$

which means that the operators of composition generalize, in a sense, T- product possibility distributions defined by (2).

It is evident that both $\pi_1 \triangleright_T \pi_2$ and $\pi_1 \triangleleft_T \pi_2$ are (generally different) possibility distributions of variables $(X_i)_{i\in K_1\cup K_2}$. In fact, the first one is an extension of π_1, while the second of π_2, in a special case of both, as the following lemma suggests.

Lemma 1 *Consider two distributions $\pi_1(x_{K_1})$ and $\pi_2(x_{K_2})$. Then*

$$(\pi_1 \triangleright_T \pi_2)(x_{K_1\cup K_2}) = (\pi_1 \triangleleft_T \pi_2)(x_{K_1\cup K_2})$$

for any continuous t-norm T if and only if $\pi_1$and π_2 are projective.

[4] Let us stress that for the definition of these operators we do not require projectivity of distributions π_1 and π_2.

The following lemma (proven in [14]) expresses the relationship between the operators of composition and conditional T-independence.

Lemma 2 *Let T be a continuous t-norm and π_1 and π_2 be projective possibility distributions on $\mathbf{X}_{K_1}$ and $\mathbf{X}_{K_2}$, respectively. Then the distribution π of $X_{K_1 \cup K_2}$*

$$\pi(x_{K_1 \cup K_2}) = \pi_1\,(x_{K_1}) \triangleright_T \pi_2\,(x_{K_2}) = \pi_1\,(x_{K_1}) \triangleleft_T \pi_2\,(x_{K_2})$$

if and only if $X_{K_1 \setminus K_2}$ and $X_{K_2 \setminus K_1}$ are conditionally independent, given $X_{K_1 \cap K_2}$.

Perfect Sequences. Now, we will recall how to apply the operators iteratively. Consider a sequence of distributions $\pi_1(x_{K_1}), \pi_2(x_{K_2}), \ldots, \pi_m(x_{K_m})$ and the expression

$$\pi_1 \triangleright_T \pi_2 \triangleright_T \ldots \triangleright_T \pi_m.$$

Before presenting its properties, let us note that in the part that follows, we always apply the operators from left to right, i.e.,

$$\pi_1 \triangleright_T \pi_2 \triangleright_T \pi_3 \triangleright_T \ldots \triangleright_T \pi_m = (\ldots((\pi_1 \triangleright_T \pi_2) \triangleright_T \pi_3) \triangleright_T \ldots \triangleright_T \pi_m).$$

This expression defines a multidimensional distribution on $\mathbf{X}_{K_1 \cup \ldots \cup K_m}$. Therefore, for any permutation $i_1, i_2, \ldots, i_m$ of indices $1, \ldots, m$ the expression

$$\pi_{i_1} \triangleright_T \pi_{i_2} \triangleright_T \ldots \triangleright_T \pi_{i_m}$$

determines a distribution on the same universe of discourse. However, for different permutations these distributions can differ from one another. Some of them seem to possess the most advantageous properties.

An ordered sequence of possibility distributions $\pi_1, \pi_2, \ldots, \pi_m$ is said to be *T-perfect* if for any $j = 2, \ldots, m$

$$\pi_1 \triangleright_T \cdots \triangleright_T \pi_j = \pi_1 \triangleleft_T \cdots \triangleleft_T \pi_j.$$

The notion of T-perfectness suggests that a sequence perfect with respect to one t-norm needn't be perfect with respect to another t-norm, similarly to (conditional) T-independence.

Let us present two assertions, which will be used later.

Lemma 3 *Let T be a continuous t-norm. The sequence $\pi_1, \pi_2, \ldots, \pi_m$ is T-perfect, if and only if the pairs of distributions $(\pi_1 \triangleright_T \cdots \triangleright_T \pi_{k-1})$ and π_k are projective for all $k = 2, 3, \ldots, m$.*

Theorem 1 *The sequence $\pi_1, \pi_2, \ldots, \pi_m$ is T-perfect if and only if all the distributions $\pi_1, \pi_2, \ldots, \pi_m$ are marginal to distribution $\pi_1 \triangleright_T \pi_2 \triangleright_T \ldots \triangleright \pi_m$.*

Now, let us recall the notion of *running intersection property* (RIP) and the related results from [14]. A sequence of sets $K_1, K_2, \ldots, K_n$ is said to meet RIP if

$$\forall i = 2, \ldots, n \;\; \exists j (1 \leq j < i) \qquad (K_i \cap (K_1 \cup \ldots \cup K_{i-1})) \subseteq K_j.$$

Lemma 4 *If $\pi_1, \pi_2, \ldots, \pi_m$ is a sequence of pairwise projective low-dimensional distributions such that $K_1, \ldots, K_m$ meets RIP, then this sequence is T-perfect for any continuous t-norm T.*

Convergence of IPFP(T).

Theorem 2 *If there is an ordering $\pi_1, \ldots, \pi_m$ of possibility distributions from $\mathcal{S}$ such that $\pi_1, \ldots, \pi_m$ form a T-perfect sequence for some continuous t-norm T and $\rho_{(0)} \equiv 1$, then IPFP(T) converges in one cycle. Furthermore, distribution $\rho_{(m)}$ factorizes with respect to $\mathcal{K}$ and T.*

Proof. First, let us note that (8) for $\pi_1, \ldots, \pi_m$ can be rewritten using an operator of left composition, i.e.,

$$\begin{aligned} \rho_{(j)}(x) &= T(\rho_{(j-1)}(x) \triangle_T \rho_{(j-1)}(x_{K_k}), \pi_k(x_{K_k})) \\ &= \rho_{(j-1)}(x) \lhd_T \pi_k(x_{K_k}) \end{aligned}$$

for any $j = 1, \ldots$; especially for $j = 1, \ldots, n$ (which means that $k = j$) we obtain

$$\begin{aligned} \rho_{(j)}(x) &= \rho_{(j-1)}(x) \lhd_T \pi_j(x_{K_j}) \\ &= (\rho_{(j-2)}(x) \lhd_T \pi_{j-1}(x_{K_{j-1}})) \lhd_T \pi_j(x_{K_j}) \\ &\quad \ldots \\ &= (\ldots(\rho_{(0)}(x) \lhd_T \pi_1(x_{K_1})) \lhd_T \ldots \lhd_T \pi_{j-1}(x_{K_{j-1}})) \lhd_T \pi_j(x_{K_j}) \\ &= (\ldots T(\rho_{(0)}(x_{N \setminus \cup_{k=1}^{j} K_k}), \pi_1(x_{K_1})) \lhd_T \ldots \lhd_T \pi_{j-1}(x_{K_{j-1}})) \lhd_T \pi_j(x_{K_j}), \end{aligned}$$

since $\rho_{(0)} \equiv 1$. In particular we have

$$\rho_{(m)}(x) = (\ldots(\pi_1(x_{K_1}) \lhd_T \pi_2(x_{K_2}) \ldots \lhd_T \pi_{m-1}(x_{K_{m-1}})) \lhd_T \pi_m(x_{K_m}). \tag{9}$$

Since $\pi_1, \ldots, \pi_m$ is a T-perfect sequence of possibility distributions, every π_k is a marginal to the distribution on the right-hand side of (9). Therefore,

$$\rho_{(m)}(x_{K_k}) = \pi_k(x_{K_k})$$

for all $k = 1, \ldots, m$, which implies

$$\rho_{(j)}(x_{K_k}) = \rho_{(m)}(x_{K_k})$$

for any $j = m+1, \ldots$. To prove factorization it is enough to find fuzzy variables $f_{K_1}, \ldots, f_{K_m}$ such that

$$\rho_{(m)}(x) = T^m(f_{K_1}(x_{K_1}), \ldots, f_{K_m}(x_{K_m})).$$

But, due to T-perfectness of $\pi_1, \ldots, \pi_m$

$$\rho_{(m)}(x) = \pi_1(x_{K_1}) \triangleright_T \pi_2(x_{K_2}) \triangleright_T \ldots \triangleright \pi_m(x_{K_m}),$$

which can be rewritten in the form

$$\rho_{(m)}(x) = T^m(\pi_1(x_{K_1}), \pi_2(x_{K_2}) \triangle_T \pi_2(x_{K_2 \cap K_1}), \ldots \\ \ldots, \pi_m(x_{K_m}) \triangle_T \pi_m(x_{K_m \cap (K_1 \cup \ldots \cup K_{m-1})})),$$

which concludes the proof.

First, let us stress that perfectness with respect to a t-norm implies convergence with respect to the same t-norm (and *not with respect to any*) as can be seen from the following simple example.

Example 3 Let X_1, X_2 and X_3 be three binary variable as in Example 2 and π_1, π_2 and π_3 on $\mathbf{X}_{\{1,2\}}, \mathbf{X}_{\{2,3\}}$ and $\mathbf{X}_{\{1,3\}}$ be defined by Table 1.

π_1	X_2	0	1
	$X_1 = 0$	1	.8
	$X_1 = 1$	.6	.4

π_2	X_3	0	1
	$X_2 = 0$	1	.5
	$X_2 = 1$	.3	.8

π_3	X_3	0	1
	$X_1 = 0$	1	.8
	$X_1 = 1$	.6	.5

Table 1: Distributions forming min-perfect sequence

$\rho_{(j)}$	j						
(x_1, x_2, x_3)	0	1	2	3	4	5	6
$(0,0,0)$	1	1	1	1	1	1	1
$(0,0,1)$	1	1	.5	.5	.5	.5	.5
$(0,1,0)$	1	.8	.8	.8	.8	.8	.8
$(0,1,1)$	1	.8	.8	.8	.8	.8	.8
$(1,0,0)$	1	.6	.6	.6	.6	.6	.6
$(1,0,1)$	1	.6	.5	.5	.5	.5	.5
$(1,1,0)$	1	.4	.3	.3	.3	.3	.3
$(1,1,1)$	1	.4	.4	.4	.4	.4	.4

Table 2: Convergence of IPFP with respect to Gödel’s t-norm

Sequence π_1, π_2, π_3 is min-perfect (due to Lemma 3), since $\pi_1(x_2) = \pi_2(x_2)$ and $(\pi_1 \triangleright_{T_G} \pi_2)(x_1, x_3) = \pi_3(x_1, x_3)$. Starting from $\rho_{(0)} \equiv 1$, IPFP(T_G) converges after one cycle as can be seen from Table 2 while IPFP(T_Π) and IPFP(T_L) converge after four and five cycles, respectively (cf. Tables 3 and 4).

Corollary 1 *If there is a permutation $K_{i_1}, \ldots, K_{i_n}$ of sets from $\mathcal{K}$ such that $K_{i_1}, \ldots, K_{i_n}$ meets RIP, $\{\pi_{i_1}, \ldots, \pi_{i_n}\}$ is an input sequence of pairwise projective possibility distributions and $\rho_{(0)} \equiv 1$ then IPFP(T) converges in one cycle for any continuous t-norm T and $\rho_{(m)}$ factorizes with respect to the corresponding t-norm T.*

$\rho_{(j)}$	j								
(x_1,x_2,x_3)	0	1	2	3	4, 5	6	7, 8	9	$10,11,12$
$(0,0,0)$	1	1	1	1	1	1	1	1	1
$(0,0,1)$	1	1	.5	.5	.5	.5	.5	.5	.5
$(0,1,0)$	1	.8	.3	.3	.3	.3	.3	.3	.3
$(0,1,1)$	1	.8	.8	.8	.8	.8	.8	.8	.8
$(1,0,0)$	1	.6	.6	.6	.6	.6	.6	.6	.6
$(1,0,1)$	1	.6	.3	.375	.375	.46875	.46875	.5	.5
$(1,1,0)$	1	.4	.15	.15	.12	.096	.096	.096	.09
$(1,1,1)$	1	.4	.4	.5	.4	.5	.4	.427	.4

Table 3: Convergence of IPFP with respect to product *t*-norm

$\rho_{(j)}$	j										
(x_1,x_2,x_3)	0	1	2	3	4, 5	6	7, 8	9	10, 11	12	13, 14, 15
(0,0,0)	1	1	1	1	1	1	1	1	1	1	1
(0,0,1)	1	1	.5	.5	.5	.5	.5	.5	.5	.5	.5
(0,1,0)	1	.8	.3	.3	.3	.3	.3	.3	.3	.3	.3
(0,1,1)	1	.8	.8	.8	.8	.8	.8	.8	.8	.8	.8
(1,0,0)	1	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6
(1,0,1)	1	.6	.1	.2	.2	.3	.3	.4	.4	.5	.5
(1,1,0)	1	.4	.0	.0	.0	.0	.0	.0	.0	.0	.0
(1,1,1)	1	.4	.4	.5	.4	.5	.4	.5	.4	.5	.4

Table 4: Convergence of IPFP with respect to Lukasziewicz' *t*-norm

Proof follows directly from Theorem 2 and Lemma 4.

Let us also mention that $\rho_{(0)} \equiv 1$ is not only a technical requirement that makes the proof of Theorem 2 so simple; it may be substantial for convergence as can be seen from the following example.

Example 4 Let $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ and $\pi_1(x_1,x_2)$, $\pi_2(x_2,x_3)$ be as in Example 2 and $\rho_{(0)}$ be defined as follows:

$$\rho_{(0)}(0,0,0) = \rho_{(0)}(0,1,1) = \rho_{(0)}(1,0,1) = \rho_{(0)}(1,1,0) = 1,$$

values of remaining combinations being equal to $\alpha \in [0,1]$. The convergence depends on the value of α — the results of our experiments can be found in Table 5.

The reason for this behaviour lies in the tendency of IPF procedure to find a distribution with given marginals which, moreover, factorizes with respect to the system $\mathcal{K}$ and is "as close as possible" to $\rho_{(0)}$. It is evident that $\rho_{(0)} \equiv 1$ factorizes with respect to *any* system of cliques. Therefore, it is the "safe", although perhaps

α	Convergence of IPFP(T)		
	T_G	T_Π	T_L
1	1	1	1
.5	2	2	2
.1	cycles	4	3
0	cycles	cycles	3

Table 5: Convergence of IPFP(T) depends on α

not always an optimal, initial distribution. The more “distant” the structure of the starting distribution is from factorization with respect to $\mathcal{K}$ and T, the more problematic the convergence of IPFP(T) is.

5 Conclusions

We introduced a possibilistic version of IPF procedure with the aim of using it as a tool for marginal problem solving. This procedure is parameterized by a continuous t-norm and its behaviour (convergence) is strongly dependent on it. Another important finding is that convergence of IPFP(T) substantially depends on the choice of an input distribution.

Nevertheless, there are still many problems that remain to be solved. The most important is the proof of the convergence of IPFP(T) in a general case. Another question is whether the resulting distribution is independent of the ordering of input distributions. We should also study the behaviour of IPFP(T) in inconsistent cases.

Acknowledgements

The author wishes to express her gratitude to Jiří Vomlel for writing the program in Mathematica and for his invaluable help when performing the experiments.

References

[1] B. Bouchon-Meunier, G. Coletti and C. Marsala, Independence and possibilistic conditioning *Annals of Mathematics and Artificial Intelligence,* **35** (2002), pp. 107-123.

[2] G. de Cooman, Possibility theory I – III. *Int. J. General Systems* **25** (1997), pp. 291–371.

[3] W.E. Deming and F.F. Stephan, On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* **11** (1940), pp. 427–444.

[4] A. Dempster, Upper and lower probabilities induced by multivalued mappings. *Ann. Math. Statist.* **38** (1967), pp. 325–339.

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

[6] P. Hájek, T. Havránek and R. Jiroušek, *Uncertain information processing in expert systems.* CRC Press, Boca Raton, Ann Arbor, London, Tokyo, 1992.

[7] E. Hisdal, Conditional possibilities independence and noninteraction, *Fuzzy Sets and Systems,* **1** (1978), pp. 299–309.

[8] H. Janssen, G. de Cooman and E. E. Kerre, First results for a mathematical theory of possibilistic Markov processes, *Proceedings of IPMU'96, volume III (Information Processing and Management of Uncertainty in Knowledge-Based Systems),* Granada, Spain, 1996, pp. 1425–1431. (1988),

[9] L. Rüschendorf, Convergence of the iterative proportional fitting procedure. *Ann. Statist.* **24** (1995), pp. 1160–1174.

[10] J. Vejnarová, Composition of possibility measures on finite spaces: preliminary results. In: *Proceedings of IPMU'98,* Paris, 1998, pp. 25–30.

[11] J. Vejnarová, Possibilistic independence and operators of composition of possibility measures. In: M. Hušková, J. Á. Víšek, P. Lachout (eds.) *Prague Stochastics'98,* JČMF, 1998, pp. 575–580.

[12] J. Vejnarová, Conditional independence relations in possibility theory. *Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems* **8** (2000), pp. 253–269.

[13] J. Vejnarová, Markov properties and factorization of possibility distributions. *Annals of Mathematics and Artificial Intelligence,* **35** (2002), pp. 357-377.

[14] J. Vejnarová, A partial solution of possibilistic marginal problem. In: G. de Cooman, T. L. Fine and T. Seidenfeld (eds.), *ISIPTA'01 (2nd Int. Symposium on Imprecise Probabilities and Their Applications),* Shaker Publishing BV, Maastricht, 2001, pp. 342–351.

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

Jiřina Vejnarová is with the Laboratory for Intelligent Systems of the University of Economics, Prague, Czech Republic. She is also a senior research fellow of the Institute for Information Theory and Automation of the Academy of Sciences of the Czech Republic. E-mail: vejnar@vse.cz

Bi-elastic Neighbourhood Models

A. WALLNER
Ludwig-Maximilians-University Munich, Germany

Abstract

We extend Buja's concept of "pseudo-capacities", which comprises the neighbourhood models for classical probabilities commonly used in robust statistics. Although systematically developing various directions for generalizing that model, we especially show that robust statistics can be freed from the severe restriction to 2-monotone capacities by employing the more natural framework of coherent or F-probabilities. Our main new tool for doing this is to use bi-elastic instead of convex functions.

Keywords

interval probability, robust statistics, neighbourhood models, distorted probability, pseudo-capacity, convex and bi-elastic functions

1 Introduction

The major concept in robust statistics for "robustifying" statements concerning classical distributions is to construct *neighbourhoods* of precise probabilities, which are called *central distributions* in this context. There is a famous method, due to Buja, accommodating, up to now, many of the corresponding neighbourhood models: Let p be some fixed classical probability, let $f\colon [0;1] \to [0;1]$ be a function with $f(0) = 0$ and $f(1) = 1$, and define

$$L = f \circ p. \tag{1}$$

By Denneberg (see [4], p. 17), a set function L constructed like this, is called a *distorted probability*, if f is increasing. In case $f(x) \leq x$, $\forall x \in [0;1]$, L can be seen as the *lower bound* of an *interval probability*, which creates a neighbourhood of p in the sense that $L(A) \leq p(A) \leq U(A) := 1 - L(\neg A)$ for all events A.

Now in robust statistics the standard requirement concerning f is to be indeed *convex*. We suspect that nobody knows a reasonable philosophical argument, why this strong assumption is made. Instead it seems to have mere mathematical origins: "Only if f is convex, then L becomes an algebraic pushover." We want to convince the reader that not even this technical argument is true. Strictly speaking, the word "Only" should be replaced by "Not only".

If f is convex, then by Buja (cf. [3]) a set function L constructed in accordance with (1) is called a *pseudo-capacity*.[1] Now every pseudo-capacity is a *2-monotone* set function (see Theorem 1, model 5, and also [4], p. 17), and this fact seems to be the technical advantage. But from a philosophical point of view there are no visible reasons to restrict the frameworks of interval probability as well as of robust statistics to 2-monotonicity. Instead it is more natural to consider the wider class of Walley's *coherent probabilities* (cf. [8]), which are closely related to *F-probabilities* in the sense of Weichselberger (see [10] or [11]).

We will show that the formulation (1) is also useable for constructing the lower bound L of an F-probability, which is not necessarily 2-monotone. For this we have to weaken the condition of convexity for f and replace it by a new assumption: *bi-elasticity*.

Just as there exist 2-monotone set functions L, which cannot be described by (1) using convex functions f, we, of course, are not able to produce the whole class of F-probabilities by only employing the definition (1), letting p vary over all classical probabilities and f vary over all bi-elastic functions. But we will explain that bi-elasticity is exactly the *appropriate* requirement *when* defining F-probabilities via (1) (see Section 6). Moreover, from an algebraical point of view the generated subclass of F-probabilities is as easy manageable as the corresponding subclass of 2-monotone set functions, i.e. the class of pseudo-capacities.

In Section 2 we introduce the notion of bi-elasticity. In Section 3 a language for interval probability is fixed: As far as needed, we outline Weichselberger's formal and methodological framework. But this should be no restriction. Since, in particular, σ-additivity (instead of additivity) of classical probabilities does not play any role, the concepts developed could also be applied to other theories of imprecise probabilities, especially to Walley's theory. In Section 4 we go into the details of the convex and bi-elastic neighbourhood models described above, resulting in Theorem 1. There we, in fact, will not use the phrasing of equation (1): Since sometimes it is necessary to apply methods of robustness to interval probability itself[2] and, anyway, it is a natural mathematical task to look for closure properties, we consider the more generalized form

$$L = f \circ L_0, \tag{2}$$

where L_0 is the lower bound of some given *interval-valued central distribution*. Learning from Theorem 1, we also deal with a modified version of it, which is stated in Theorem 2. Its formulation serves, in essence, as a motivation for Section 5, i.e. for Theorems 3 and 4, which significantly generalize the neighbourhood models developed before. Section 6 is reserved for concluding remarks.

To give reasons for the successive steps, the structure of this technical paper is rather heuristic. Hence the proofs are postponed repeatedly — until the proof of the last theorem.

[1] See [2] for more detailed information.

[2] See [1], pp. 229ff, for a discussion of this topic.

2 Convex and Bi-elastic Functions

What is *bi-elasticity*? Suppose we concentrate on a function $f\colon [0;1] \to [0;1]$ with $f(0) = 0$ and $f(1) = 1$ and imagine that three points of f's graph, namely $(x, f(x))$, $(y, f(y))$, and $(z, f(z))$ with $0 \le x < y < z \le 1$, are *standing in convex position*. Obviously, by this terminology we mean that the following local comparison of quotients of differences is valid:

$$\frac{f(y) - f(x)}{y - x} \le \frac{f(z) - f(y)}{z - y}. \tag{3}$$

If this inequality is globally true, i.e. *for all* such x, y, z, we usually say that f is *convex*. Now fix $x = 0$, and let just y and z vary. Then it is easily seen that we get equivalently

$$\frac{f(y)}{y} \le \frac{f(z)}{z}, \quad \forall y, z \text{ with } 0 < y \le z \le 1, \tag{4}$$

i.e. that the *average of f is increasing*. In economic sciences this behaviour of f is called *elastic* (e.g. see [5]).

So, what's *bi*-elasticity? For this new concept (introduced in [9], Chapter 6), let first f be elastic, and secondly set $z = 1$ in (3) and let x and y vary. After simple transformations we get

$$\frac{1 - f(x)}{1 - x} \le \frac{1 - f(y)}{1 - y}, \quad \forall x, y \text{ with } 0 \le x \le y < 1, \tag{5}$$

as an equivalent form, which, in turn, is equivalent to

$$\frac{1 - f(1 - y)}{y} \le \frac{1 - f(1 - x)}{x}, \quad \forall x, y \text{ with } 0 < x \le y \le 1. \tag{6}$$

Thus, additionally, the *conjugate function of f*, i.e. $x \mapsto 1 - f(1 - x)$, has to have *decreasing average*. We summarize:

Definition 1 Let $f\colon [0;1] \to [0;1]$ with $f(0) = 0$ and $f(1) = 1$. Then f is called

1. *convex*, if (3) holds for all x, y, z with $0 \le x < y < z \le 1$,
2. *bi-elastic*, if (4) and (6) are valid. □

Corollary 1 *Let $f\colon [0;1] \to [0;1]$ with $f(0) = 0$ and $f(1) = 1$.*

1. *f is convex iff $f((1-\lambda)x + \lambda y) \le (1-\lambda) f(x) + \lambda f(y), \forall x, y, \lambda \in [0;1]$.*
2. *f is bi-elastic iff $f(\lambda x) \le \lambda f(x)$ and $\lambda(1 - f(1-x)) \le 1 - f(1 - \lambda x), \forall x, \lambda \in [0;1]$.*
3. *If f is convex, then f is bi-elastic.*
4. *If f is bi-elastic, then $f(x) \le x, \forall x \in [0;1]$.*[3] □

Proof. 1.) and 2.) can be shown straightforwardly. For 3.) see above, for 4.) put $y = x$ and $z = 1$ in (4). □

[3] Moreover, every bi-elastic function is monotone in $[0;1]$ and continuous in $[0;1[$.

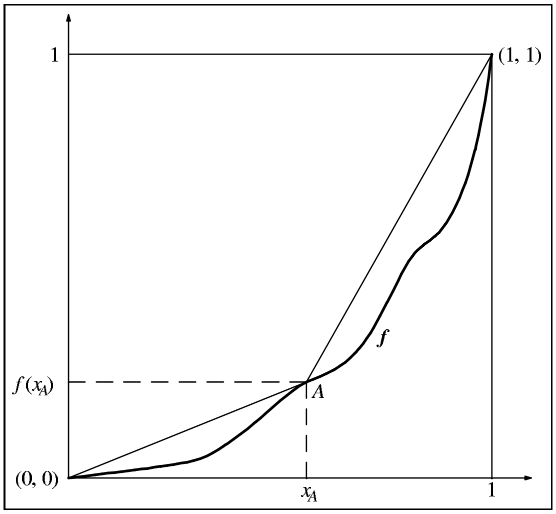


Figure 1: An example of a bi-elastic function f. Bi-elasticity of f can be described equivalently as follows: For each point $A = (x_A, f(x_A))$ on the graph of f, the graph of f between 0 and x_A nowhere is lying above the line between $(0,0)$ and A, and between x_A and 1 it nowhere is lying above the line between A and $(1,1)$.

3 Basic Definitions of Interval Probability according to Weichselberger

Here we report the main concepts of Weichselberger's theory of interval probability (see [10] or [11]), adding some slight modifications. For the following let Ω be a fixed *sample space* and $\mathcal{A}$ a fixed σ*-algebra* over Ω. Hence $(\Omega; \mathcal{A})$ is fixed *measurable space*.

Definition 2 A set function p: $\mathcal{A} \to [0; 1]$ is called a *K-function* (classical probability) on $(\Omega; \mathcal{A})$, if it satisfies the axioms of Kolmogorov. The set of all K-functions on $(\Omega; \mathcal{A})$ is denoted by $\mathcal{K}(\Omega; \mathcal{A})$. □

Definition 3

1. A triple $O = (\Omega; \mathcal{A}; L)$ is called an *adjusted O-field*, if L: $\mathcal{A} \to [0; 1]$ is a set function, which is *normed*, i.e. $L(\emptyset) = 0$ and $L(\Omega) = 1$. The set $\mathcal{M}(O) = \{p \in \mathcal{K}(\Omega; \mathcal{A}) \mid L(A) \leq p(A), \forall A \in \mathcal{A}\}$ is called the *structure of O*.

2. An adjusted O-field $\mathcal{R}$ is called an *adjusted R-(probability) field*, if $\mathcal{M}(\mathcal{R}) \neq \emptyset$.

3. An adjusted R-field $\mathcal{F} = (\Omega; \mathcal{A}; L)$ is called an *F-(probability) field*, if it satisfies the axiom $L(A) = \inf_{p \in \mathcal{M}(\mathcal{F})} p(A), \forall A \in \mathcal{A}$.[4]

4. An adjusted R-field $\mathcal{F} = (\Omega; \mathcal{A}; L)$ is called an F_0*-(probability) field*, if it satisfies the axiom $L(A) = \min_{p \in \mathcal{M}(\mathcal{F})} p(A), \forall A \in \mathcal{A}$.

5. An adjusted O-field $(\Omega; \mathcal{A}; L)$ is called a *CA-field*, if L is *2-monotone*, i.e. $L(A) + L(B) \leq L(A \cup B) + L(A \cap B), \forall A, B \in \mathcal{A}$.

6. A CA-field is called a *C-(probability) field*, if it is an F-field.

7. A CA-field is called a C_0*-(probability) field*, if it is an F_0-field.

8. A triple $(\Omega; \mathcal{A}; p)$ is called a *K-(probability) field*, if p is a K-function. □

Since $(\Omega; \mathcal{A})$ is fixed, every adjusted O-field $O = (\Omega; \mathcal{A}; L)$ is determined by the “lower bound” L. Subsequently we always “associate” the “upper bound” U of O via *conjugation* of L, i.e. $U(.) = 1 - L(\neg.)$.

Some comments on Definition 3 are useful:

- Weichselberger’s original definition of an *R-field* is that of a quadruple $\mathcal{R} = (\Omega; \mathcal{A}; L, U)$ having a non-empty structure $\mathcal{M}(\mathcal{R}) = \{p \in \mathcal{K}(\Omega; \mathcal{A}) \mid L(A) \leq p(A) \leq U(A),\ \forall A \in \mathcal{A}\}$. In this setting neither L is normed necessarily, nor L and U have to be conjugate, what both is not appropriate for our purposes.

- In [1], Corollary 2.13, it is shown that every *continuous* F-field is an F_0-field. (Hence, in particular, every F-field on a finite measurable space has the F_0-property.) Since, on the one hand, we don’t want to discuss topological features here, but, on the other hand, intend to deal with closure properties concerning F-fields as well as F_0-fields, we distinguish both cases by introducing these two terms.

- It is known that every CA-field is a C_0-field, and hence a C-field, in case the sample space Ω is finite. For the general case, usually additional topological assumptions are made to enforce the F-(or F_0-)property, in particular, for defining *2-monotone capacities* (cf. [7]). But, as mentioned above, we want to abstain from topological aspects here. So the CA-property, i.e., essentially, the 2-monotonicity of the lower bound, should be considered as the extracted pure algebraic part of the definition of C-(or C_0-)fields. There are some closure properties, we want to emphasize later, only concerning this algebraic part. Therefore the definitions of CA-, C-, and C_0-fields are organized as stated.

[4] The definitions of adjusted R-fields and of F-fields are closely related to Walley’s *avoiding sure loss* and *coherence* respectively (cf. [8]).

For later use we record the following corollary, which can be proven straightforwardly.

Corollary 2

1. *If* $\mathcal{O} = (\Omega;\mathcal{A};L)$ *is an adjusted O-field and* $U(.) = 1 - L(\neg .)$, *then* $\mathcal{M}(\mathcal{O}) = \{p \in \mathcal{K}(\Omega;\mathcal{A}) \mid p(A) \le U(A),\ \forall A \in \mathcal{A}\}$.
2. *If* $(\Omega;\mathcal{A};L)$ *is an F- or a CA-field, then* L *and its conjugate* U *are* monotone, *i.e., for* $\Psi \in \{L,U\}$ *we have* $\forall A,B \in \mathcal{A}$*:* $A \subseteq B \Longrightarrow \Psi(A) \le \Psi(B)$.
3. *If* $\mathcal{O}_1 = (\Omega;\mathcal{A};L_1)$ *and* $\mathcal{O}_2 = (\Omega;\mathcal{A};L_2)$ *are adjusted O-fields, then*

$$L_1(.) \le L_2(.) \Longrightarrow \mathcal{M}(\mathcal{O}_2) \subseteq \mathcal{M}(\mathcal{O}_1). \qquad \square$$

As a mnemonic device concerning the definitions above, we get the following clear picture:

$$\begin{tikzcd}
& \text{CA-f.} \wedge \text{F}_0\text{-f.} \arrow[r, Rightarrow] \arrow[d, Leftrightarrow] & \text{F}_0\text{-field} \arrow[d, Rightarrow] \\
\text{K-field} \arrow[r, Rightarrow] & \text{C}_0\text{-field} \arrow[d, Rightarrow] & \text{F-field} \arrow[r, Rightarrow] & \text{adj. R-field} \arrow[r, Rightarrow] & \text{adj. O-field.} \\
& \text{C-field} \arrow[r, Leftrightarrow] & \text{CA-f.} \wedge \text{F-f.} \arrow[u, Rightarrow]
\end{tikzcd}$$

4 Convex and Bi-elastic Neighbourhood Models

For constructing *neighbourhoods* of classical probabilities, in robust statistics mainly metrics are used to define appropriate topologies over the space $\mathcal{K}(\Omega;\mathcal{A})$ (e.g. see [6]). Here we do not rely on the term “neighbourhood” in some topological sense, and that is why we give the trivial

Definition 4 For adjusted O-fields $\mathcal{O}_0 = (\Omega;\mathcal{A};L_0)$, $\mathcal{O} = (\Omega;\mathcal{A};L)$ and a K-function p, we say that

- $\mathcal{O}$ is a *neighbourhood of* $\mathcal{O}_0$, if $L(.) \le L_0(.)$,
- $\mathcal{O}$ is a *neighbourhood of* p, if $\mathcal{O}$ is a neighbourhood of $(\Omega;\mathcal{A};p)$. □

Therefore, $\mathcal{O}$ is a neighbourhood of the K-function p iff simply p is an element of the structure of $\mathcal{O}$, and hence $\mathcal{O}$ is an adjusted R-field. In general, we have $\mathcal{M}(\mathcal{O}_0) \subseteq \mathcal{M}(\mathcal{O})$, if $\mathcal{O}$ is a neighbourhood of $\mathcal{O}_0$ (cf. Corollary 2, 3.)).

Now we come to a first category of neighbourhood models motivated in Section 1. Inspired by the notions of *pseudo-capacities* (the starting point of our developments), *bi-elastic functions* (generalizing convex functions), and *interval-valued central distributions* (including precise central distributions as a specific case), we get

Theorem 1 ***(First class of neighbourhood models)*** *Let $L_0\colon \mathcal{A} \to [0;1]$ be a set function, $f\colon [0;1] \to [0;1]$ a function with $f(0) = 0$ and $f(1) = 1$, $L = f \circ L_0$, $\mathcal{O}_0 = (\Omega; \mathcal{A}; L_0)$, and $\mathcal{O} = (\Omega; \mathcal{A}; L)$. Then we have:*[5]

1. *If $\mathcal{O}_0$ is an adjusted O-field, then so is $\mathcal{O}$.*
2. *If $f(x) \leq x$, $\forall x \in [0;1]$, and $\mathcal{O}_0$ is an adjusted R-field, then so is $\mathcal{O}$.*
3. *If f is bi-elastic, and $\mathcal{O}_0$ is an F-field, then so is $\mathcal{O}$.*
4. *If f is bi-elastic, and $\mathcal{O}_0$ is an F_0-field, then so is $\mathcal{O}$.*
5. *If f is convex, and $\mathcal{O}_0$ is a CA-field, then so is $\mathcal{O}$.*
6. *If f is convex, and $\mathcal{O}_0$ is a C-field, then so is $\mathcal{O}$.*
7. *If f is convex, and $\mathcal{O}_0$ is a C_0-field, then so is $\mathcal{O}$.*

Moreover, in the cases 2.)–7.) $\mathcal{O}$ is a neighbourhood of $\mathcal{O}_0$. □

Proof. 1.) and 2.) are obvious. For 3.)–7.) see Theorem 2 below.[6] The “Moreover”-statement follows from Corollary 1, 3.) and 4.). □

From now on we concentrate on the most interesting cases, namely F-, F_0-, CA-, C-, and C_0-fields. Our goal is to generalize models 3–7 of Theorem 1 in two steps, which leads to Theorems 2, 3, and 4.

The *first step* is just a small one and is based on an elementary observation. Let us for the moment consider model 5 of Theorem 1: In order to maintain the 2-monotonicity, we, in essence, made two assumptions: the definition of L, i.e. $L = f \circ L_0$, and the convexity of f. By Definition 1, 1.), this implies

$$\frac{L(B) - L(A)}{L_0(B) - L_0(A)} \leq \frac{L(C) - L(B)}{L_0(C) - L_0(B)}, \tag{7}$$

for all $A, B, C \in \mathcal{A}$ with $L_0(A) < L_0(B) < L_0(C)$. Now it is natural to suspect that it doesn't matter, how f is defined on $[0;1] \setminus \{L_0(A) \mid A \in \mathcal{A}\}$. It should be sufficient for our CA-model to presuppose the inequalities (7). Similarly, we expect that models 3 and 4 of Theorem 1 could be modified analogously: The corresponding inequalities given by bi-elasticity are (cf. (4) and (5))

$$\frac{L(A)}{L_0(A)} = \frac{f(L_0(A))}{L_0(A)} \leq \frac{f(L_0(B))}{L_0(B)} = \frac{L(B)}{L_0(B)},$$

for all $A, B \in \mathcal{A}$ with $0 < L_0(A) \leq L_0(B)$, and, additionally, using $U_0(.) = 1 - L_0(\neg .)$ and $U(.) = 1 - L(\neg .)$,

$$\frac{U(B)}{U_0(B)} = \frac{1 - L(\neg B)}{1 - L_0(\neg B)} = \frac{1 - f(L_0(\neg B))}{1 - L_0(\neg B)} \leq \frac{1 - f(L_0(\neg A))}{1 - L_0(\neg A)} = \frac{1 - L(\neg A)}{1 - L_0(\neg A)} = \frac{U(A)}{U_0(A)},$$

for all $A, B \in \mathcal{A}$ with $L_0(\neg B) \leq L_0(\neg A) < 1$, i.e., equivalently, $0 < U_0(A) \leq U_0(B)$.

These considerations are summed up in

[5] Models 5–7 reflect the concept of pseudo-capacities, in case of a precise central distribution $\mathcal{O}_0$.

[6] For the moment, we can say that 6.) is a consequence of 3.) and 5.), and 7.) is a consequence of 4.) and 5.), since convexity implies bi-elasticity (cf. Corollary 1, 3.)).

Theorem 2 ***(Second class of neighbourhood models)*** *Let $\mathcal{O}_0 = (\Omega; \mathcal{A}; L_0)$ and $\mathcal{O} = (\Omega; \mathcal{A}; L)$ be adjusted O-fields, $U_0(.) = 1 - L_0(\neg.)$, and $U(.) = 1 - L(\neg.)$.*

1. *Suppose that $\mathcal{O}_0$ is an F-field and that the following two conditions hold:*

   (a) $$L(A) \cdot L_0(B) \leq L_0(A) \cdot L(B), \quad \forall A, B \in \mathcal{A} \text{ with } L_0(A) \leq L_0(B); \tag{8}$$

   (b) $$U_0(A) \cdot U(B) \leq U(A) \cdot U_0(B), \quad \forall A, B \in \mathcal{A} \text{ with } U_0(A) \leq U_0(B). \tag{9}$$

   Then $\mathcal{O}$ is an F-field, too.

2. *Suppose that $\mathcal{O}_0$ is an F_0-field and that conditions (8) and (9) hold. Then $\mathcal{O}$ is an F_0-field, too.*

3. *Suppose that $\mathcal{O}_0$ is a CA-field and that the following condition holds:*[7]

   $$\begin{gathered}(L(B) - L(A)) \cdot (L_0(C) - L_0(B)) \leq (L_0(B) - L_0(A)) \cdot (L(C) - L(B)), \\ \forall A, B, C \in \mathcal{A} \text{ with } L_0(A) \leq L_0(B) \leq L_0(C).\end{gathered} \tag{10}$$

   Then $\mathcal{O}$ is a CA-field, too.

4. *Suppose that $\mathcal{O}_0$ is a C-field and that condition (10) holds. Then $\mathcal{O}$ is a C-field, too.*

5. *Suppose that $\mathcal{O}_0$ is a C_0-field and that condition (10) holds. Then $\mathcal{O}$ is a C_0-field, too.*

Moreover, in all five cases we have: $\mathcal{O}$ is a neighbourhood of $\mathcal{O}_0$, and the "functional connection"

$$\forall A, B \in \mathcal{A}: \quad L_0(A) = L_0(B) \Longrightarrow L(A) = L(B) \tag{11}$$

between L_0 and L is valid. □

Proof. It is straightforward that from condition (10) we can derive conditions (8) and (9) (for (8) put $A = \emptyset$ in (10), for (9) set $C = \Omega$ in (10)[8]). Hence, on the one hand, 4.) is a direct consequence of 1.) and 3.), and 5.) is a consequence of 2.) and 3.). On the other hand, the "Moreover"-statement can be deduced from (8): By putting $B = \Omega$, we get

$$L(A) \leq L_0(A), \quad \forall A \in \mathcal{A}, \tag{12}$$

which is the statement that $\mathcal{O}$ is a neighbourhood of $\mathcal{O}_0$. To prove (11), let $L_0(A) = L_0(B)$. By (12), we can assume $L_0(B) > 0$. But then, two applications of (8) lead to $L(A) \cdot L_0(B) = L_0(A) \cdot L(B) = L_0(B) \cdot L(B)$, thus $L(A) = L(B)$.

Summarizing, we have shown all parts of Theorem 2 — with the exception of its heart: statements 1.), 2.), and 3.). For this we refer to Theorem 3 below, since in the situations of 1.), 2.), and 3.) the set functions L_0 and U_0 are monotone (cf. Corollary 2, 2.)). To be complete, we have to prove the additional premise (15) in Theorem 3. But this is an easy result of (11) (just set $B = \Omega$), which is proved already. □

[7] In (10) it's not sufficient to use quotients as above, excluding the possibility that the denominator is 0.

[8] We can argue in a manner similar to the proof sketch, given at the beginning of Section 2, where we deduced bi-elasticity from convexity.

Clearly, the most important case of Theorem 2 is that the central distribution $\mathcal{O}_0$ is some K-field $(\Omega;\mathcal{A};p_0)$. For this we give an example, which — historically — led to all generalized neighbourhood models presented here.

Example 1 Let $(\Omega;\mathcal{A}) = (\Omega_k;\mathcal{P}(\Omega_k))$ be a finite measurable space, where $|\Omega_k| = k \in \mathbb{N}$ and $\mathcal{P}(\Omega_k)$ is the power set of Ω_k. We consider the consequences of Theorem 2 for the case $\mathcal{O}_0 = (\Omega_k;\mathcal{P}(\Omega_k);p_0^k)$, in which p_0^k is the *classical uniform probability* on $(\Omega_k;\mathcal{P}(\Omega_k))$, i.e. $p_0^k(A) = \frac{|A|}{k}$, $\forall A \subseteq \Omega_k$. Let $\mathcal{O} = (\Omega_k;\mathcal{P}(\Omega_k);L)$ be some adjusted O-field. From (11) we conclude

$$\forall A,B \subseteq \Omega_k : \quad |A| = |B| \Longrightarrow L(A) = L(B), \tag{13}$$

which means that the only possibility in generating $\mathcal{O}$ as a neighbourhood of p_0^k with the methods of Theorem 2, we have to restrict ourselves to *uniform interval probability*. Hence we assume (13) and write for $i = 0,\dots,k$: $L^{(i)} = L(A)$, if $i = |A|$ for some $A \subseteq \Omega_k$, and consistently $U^{(i)} = 1 - L^{(k-i)}$. Additionally, we concentrate on considering models 1 and 2 of Theorem 2, the F- and the F$_0$-model, which are the same, since Ω_k is finite. Conditions (8) and (9) are equivalent to the chain

$$\frac{L^{(1)}}{1} \le \frac{L^{(2)}}{2} \le \cdots \le \frac{L^{(k-1)}}{k-1} \le \frac{1}{k} \le \frac{U^{(k-1)}}{k-1} \le \cdots \le \frac{U^{(2)}}{2} \le \frac{U^{(1)}}{1}. \tag{14}$$

Therefore, model 1 of Theorem 2 says: Every adjusted uniform O-field $\mathcal{O} = (\Omega_k;\mathcal{P}(\Omega_k);L)$ is an F-field — an "*uniform F-field*" —, if it obeys the chain (14). In [11], Lemma 4.3.5, it is shown over and above that, that (14) is also necessary for $\mathcal{O}$ to be an uniform F-field on $(\Omega_k;\mathcal{P}(\Omega_k))$. □

Theorem 2 is only a very slight generalization of the models 3–7 in Theorem 1. For example, if condition (10) holds, it always is possible to construct a convex function f: $[0;1] \to [0;1]$ with $f(0) = 0$ and $f(1) = 1$ such that $L = f \circ L_0$. Similarly for (8) and (9) on the one hand and bi-elastic functions defined on $[0;1]$ on the other hand.

Theorem 2 should rather be seen as a motivation for Theorem 3 given in the next section.

5 Generalized Convex and Bi-elastic Neighbourhood Models

The inequalities, working as premises in conditions (8), (9), and (10) do not seem to be very natural. For example, in (8) it would be nice to replace "$L_0(A) \le L_0(B)$" by "$A \subseteq B$", since then, e.g., we would have a connection to conditional interval probability (see Section 6).

Let us formulate this big *second step* of generalizing the neighbourhood models, fundamentally first presented in [9], Chapter 6:

Theorem 3 ***(Third class of neighbourhood models, part 1)*** *Let $\mathcal{O}_0 = (\Omega; \mathcal{A}; L_0)$ and $\mathcal{O} = (\Omega; \mathcal{A}; L)$ be adjusted O-fields, $U_0(.) = 1 - L_0(\neg .)$, and $U(.) = 1 - L(\neg .)$. Assume additionally that we have*[9]

$$\forall A \in \mathcal{A}: \quad L_0(A) = 1 \Longrightarrow L(A) = 1. \tag{15}$$

1. *Suppose that $\mathcal{O}_0$ is an F-field and that the following two conditions hold:*

   $$\text{(a)}\ L(A) \cdot L_0(B) \leq L_0(A) \cdot L(B), \quad \forall A, B \in \mathcal{A} \text{ with } A \subseteq B; \tag{16}$$

   $$\text{(b)}\ U_0(A) \cdot U(B) \leq U(A) \cdot U_0(B), \quad \forall A, B \in \mathcal{A} \text{ with } A \subseteq B. \tag{17}$$

   Then $\mathcal{O}$ is an F-field, too.

2. *Suppose that $\mathcal{O}_0$ is an F_0-field and that conditions (16) and (17) hold. Then $\mathcal{O}$ is an F_0-field, too.*

3. *Suppose that $\mathcal{O}_0$ is a CA-field and that the following condition holds:*

   $$\begin{gathered}(L(B) - L(A)) \cdot (L_0(C) - L_0(B)) \leq (L_0(B) - L_0(A)) \cdot (L(C) - L(B)), \\ \forall A, B, C \in \mathcal{A} \text{ with } A \subseteq B \subseteq C.\end{gathered} \tag{18}$$

   Then $\mathcal{O}$ is a CA-field, too.

4. *Suppose that $\mathcal{O}_0$ is a C-field and that condition (18) holds. Then $\mathcal{O}$ is a C-field, too.*

5. *Suppose that $\mathcal{O}_0$ is a C_0-field and that condition (18) holds. Then $\mathcal{O}$ is a C_0-field, too.*

Moreover, in all five cases we have: $\mathcal{O}$ is a neighbourhood of $\mathcal{O}_0$, and the "functional connection"

$$\forall A, B \in \mathcal{A}: \quad A \subseteq B \wedge L_0(A) = L_0(B) \Longrightarrow L(A) = L(B) \tag{19}$$

between L_0 and L is valid. □

Proof. Let $\mathcal{O}_0$ and $\mathcal{O}$ be adjusted O-fields as denoted. First we prove:

1. (16) $\Longrightarrow$ $L(A) \leq L_0(A)$, $\forall A \in \mathcal{A}$.
2. (16) $\Longrightarrow$ $\big(\forall A, B \in \mathcal{A}: A \subseteq B \wedge L_0(A) \leq L_0(B) \Longrightarrow L(A) \leq L(B)\big)$.
3. (15) $\wedge$ (16) $\wedge$ (17) $\Longrightarrow$ (19).

[9] It can easily be seen that the models don't work, if we drop this additional condition. (15) is equivalent with $\forall A \in \mathcal{A}: U_0(A) = 0 \Rightarrow U(A) = 0$, and hence with

$$\forall A \in \mathcal{A}: \ (\forall p_0 \in \mathcal{M}(\mathcal{O}_0).\ p_0(A) = 0) \Longrightarrow (\forall p \in \mathcal{M}(\mathcal{O}).\ p(A) = 0),$$

which means that $\mathcal{O}$ is *absolutely continuous with respect to* $\mathcal{O}_0$.

For a), just let $B = \Omega$ in (16). For b) assume $A \subseteq B$ and $L_0(A) \leq L_0(B)$, where by a) w.l.o.g. $L_0(A) > 0$. Together with (16) we get $L(A) \cdot L_0(A) \leq L(A) \cdot L_0(B) \leq L_0(A) \cdot L(B)$, hence $L(A) \leq L(B)$. For c) suppose (15), (16), (17), $A \subseteq B$, and $L_0(A) = L_0(B)$, thus also $\neg B \subseteq \neg A$ and $U_0(\neg A) = U_0(\neg B)$. By (15), w.l.o.g. we can assume $L_0(B) < 1$, hence $U_0(\neg B) > 0$. Now (17) gives $U_0(\neg B) \cdot U(\neg A) \leq U(\neg B) \cdot U_0(\neg A) = U(\neg B) \cdot U_0(\neg B)$, thus $U(\neg A) \leq U(\neg B)$, i.e. $L(B) \leq L(A)$. Together with b) we infer $L(A) = L(B)$.

Now we give the proof of Theorem 3. First it can easily be seen that (18) implies (16) and (17) (let $A = \emptyset$ or $C = \Omega$ in (18)). Therefore, on the one hand, the “Moreover”-statement is a trivial conclusion of a) and c), and, on the other hand, 4.) is a consequence of 1.) and 3.), and 5.) is a consequence of 2.) and 3.).

For 1.) and 2.) we refer to Theorem 4 (see below).

So here we just have to prove 3.), i.e., we have to show that condition (18) transfers 2-monotonicity from L_0 to L. For this, let (18) be valid and L_0 be 2-monotone. Then, according to Corollary 2, 2.), L_0 is monotone 2, and by b) we also infer the monotonicity of L. Now let $A, B \in \mathcal{A}$ be given. We have to show that

$$L(A) + L(B) \leq L(A \cup B) + L(A \cap B). \tag{20}$$

If $L_0(A \cap B) = L_0(A)$, then by (19) $L(A \cap B) = L(A)$, hence (20) follows from the monotonicity of L. Thus we assume $L_0(A \cap B) < L_0(A)$ and, symmetrically, $L_0(A \cap B) < L_0(B)$. But then, by the 2-monotonicity of L_0 we have $L_0(A) < L_0(A \cup B)$ and $L_0(B) < L_0(A \cup B)$. Together with (18), we infer for $X \in \{A, B\}$:

$$0 \leq x(X) := \frac{L(X) - L(A \cap B)}{L_0(X) - L_0(A \cap B)} \leq \frac{L(A \cup B) - L(X)}{L_0(A \cup B) - L_0(X)} =: y(X).$$

Now, by symmetric reasons, we suppose that $y(A) \leq y(B)$, hence $x(A) \leq y(B)$. Finally, the 2-monotonicity of L_0 leads to $L(A) - L(A \cap B) = x(A) \cdot (L_0(A) - L_0(A \cap B)) \leq x(A) \cdot (L_0(A \cup B) - L_0(B)) \leq y(B) \cdot (L_0(A \cup B) - L_0(B)) = L(A \cup B) - L(B)$, thus (20) holds. □

The proof of Theorem 3 is not complete, because models 1 and 2 are waiting for verification. The reason for this is that we want to emphasize that these models are, in fact, *local* models with respect to the F- and F$_0$-property respectively.[10] This is the content of the following Theorem 4, the last one in the sequence of theorems.

Definition 5 Let $A \in \mathcal{A}$ be fixed. An adjusted R-field $\mathcal{R} = (\Omega; \mathcal{A}; L)$ is called

1. an *F(A)-field*, if it satisfies the axiom $L(A) = \inf_{p \in \mathcal{M}(\mathcal{R})} p(A)$,
2. an *$F_0(A)$-field*, if it satisfies the axiom $L(A) = \min_{p \in \mathcal{M}(\mathcal{R})} p(A)$. □

The trivial connection with Definition 3, 3.) and 4.), is given by

Corollary 3 *Let $\mathcal{R} = (\Omega; \mathcal{A}; L)$ be an adjusted R-field. Then we have:*

1. *$\mathcal{R}$ is an F-field iff for all $A \in \mathcal{A}$, $\mathcal{R}$ is an F(A)-field.*
2. *$\mathcal{R}$ is an F$_0$-field iff for all $A \in \mathcal{A}$, $\mathcal{R}$ is an $F_0(A)$-field.* □

[10] This also is true for the F- and F$_0$-models in Theorems 1 and 2.

Theorem 4 ***(Third class of neighbourhood models, part 2)*** *Let $A \in \mathcal{A}$ be fixed. Let $O_0 = (\Omega; \mathcal{A}; L_0)$ and $O = (\Omega; \mathcal{A}; L)$ be adjusted O-fields, $U_0(.) = 1 - L_0(\neg.)$, and $U(.) = 1 - L(\neg.)$. Assume that (15), (16), and (17) hold. Then we have:*

1. *If O_0 is an F(A)-field, then so is O.*
2. *If O_0 is an $F_0(A)$-field, then so is O.*

Moreover, in both cases O is a neighbourhood of O_0, and the "functional connection" (19) between L_0 and L is valid. □

Proof. The "Moreover"-statement can be shown like a) and c) in the proof of Theorem 3. So we only have to prove statements 1.) and 2.), where we restrict ourselves to model 1.[11] For this let all the corresponding premises be given, especially let $A \in \mathcal{A}$ be fixed and O_0 be an F(A)-field. Since O is a neighbourhood of O_0, we have

$$L(.) \le L_0(.) \text{ and } U_0(.) \le U(.), \quad \text{and thus} \quad \mathcal{M}(O_0) \subseteq \mathcal{M}(O). \tag{21}$$

(Hence the R-property moves from O_0 to O.) Now we concentrate on proving the F(A)-property of O, where by (21) w.l.o.g. we assume $L(A) < L_0(A)$. Together with (15) we infer

$$U(\neg A) > U_0(\neg A) > 0. \tag{22}$$

Let $\varepsilon > 0$, w.l.o.g.

$$\varepsilon < U(\neg A) - U_0(\neg A). \tag{23}$$

We have to show that there exists $p \in \mathcal{M}(O)$ with $p(A) \le L(A) + \varepsilon$. Define

$$\delta = \frac{U_0(\neg A)}{U(\neg A)} \cdot \varepsilon. \tag{24}$$

Then, by (22), $\delta > 0$. Since O_0 is an F(A)-field, there is

$$p_0 \in \mathcal{M}(O_0) \quad \text{with} \quad p_0(A) \le L_0(A) + \delta. \tag{25}$$

Together with (22), (23), and (24) we get by easy calculations

$$0 < U_0(\neg A) - \delta \le p_0(\neg A) \le U_0(\neg A) < U(\neg A) - \varepsilon \le 1. \tag{26}$$

Therefore

$$1 \le \frac{U(\neg A) - \varepsilon}{p_0(\neg A)} \le \frac{U(\neg A) - \varepsilon}{U_0(\neg A) - \delta} = \frac{U(\neg A)}{U_0(\neg A)}, \tag{27}$$

where (24) is used for the equality. In addition, (26) implies that $p_0(A)$ and $p_0(\neg A)$ have positive values, and hence it is possible to define the classical conditional probabilities

$$p_0(. \mid A) = \frac{p_0(A \cap .)}{p_0(A)} \quad \text{and} \quad p_0(. \mid \neg A) = \frac{p_0(\neg A \cap .)}{p_0(\neg A)}.$$

Now we let

$$p(.) = (L(A) + \varepsilon) \cdot p_0(. \mid A) + (U(\neg A) - \varepsilon) \cdot p_0(. \mid \neg A), \tag{28}$$

which (using (26)) is a convex combination of $p_0(. \mid A)$ and $p_0(. \mid \neg A)$. Hence p is a well-defined K-function on $(\Omega; \mathcal{A})$. Moreover, we have $p(A) = L(A) + \varepsilon$.

To verify that p is an element of the structure of O, let $B \in \mathcal{A}$. We have to prove that $p(B) \ge L(B)$, where w.l.o.g. $L(B) > 0$. But then, by (21) we also have $L_0(B) > 0$. In addition, $L_0(A \cup B) > 0$.[12] From (16) we derive $\frac{L(A \cup B)}{L_0(A \cup B)} \ge \frac{L(B)}{L_0(B)}$, thus with (25),

[11] Modifying the following arguments by setting $\varepsilon = \delta = 0$, we also get a proof of model 2.

[12] Note that we are not able to infer this inequality from $L_0(B) > 0$ by monotonicity of L_0, since

$$p_0(B)\cdot\frac{L(A\cup B)}{L_0(A\cup B)} \;\geq\; L(B). \tag{29}$$

Furthermore, using the abbreviation

$$\Delta \;=\; U_0(\neg A\cap\neg B)-p_0(\neg A\cap\neg B) \;=\; p_0(A\cup B)-L_0(A\cup B), \tag{30}$$

we get the following inequalities, where (31) follows from (27) and (17), (32) is a consequence of (25), and (33) is implied by (27) and (21):

$$U_0(\neg A\cap\neg B)\cdot\frac{U(\neg A)-\varepsilon}{p_0(\neg A)} \quad\leq\quad U(\neg A\cap\neg B)\,, \tag{31}$$

$$\Delta \quad\geq\quad 0\,, \tag{32}$$

$$\frac{U(\neg A)-\varepsilon}{p_0(\neg A)} \quad\geq\quad \frac{L(A\cup B)}{L_0(A\cup B)}\,. \tag{33}$$

If $\frac{L(A)+\varepsilon}{p_0(A)} > \frac{L(A\cup B)}{L_0(A\cup B)}$,[13] we calculate

$$\begin{aligned}
p(B) &\overset{(28)}{=} p_0(A\cap B)\cdot\frac{L(A)+\varepsilon}{p_0(A)}+p_0(\neg A\cap B)\cdot\frac{U(\neg A)-\varepsilon}{p_0(\neg A)}\\
&\overset{(33)}{\geq} p_0(A\cap B)\cdot\frac{L(A\cup B)}{L_0(A\cup B)}+p_0(\neg A\cap B)\cdot\frac{L(A\cup B)}{L_0(A\cup B)}\\
&= p_0(B)\cdot\frac{L(A\cup B)}{L_0(A\cup B)} \quad\overset{(29)}{\geq}\quad L(B).
\end{aligned}$$

Therefore, we can assume

$$\frac{L(A)+\varepsilon}{p_0(A)} \;\leq\; \frac{L(A\cup B)}{L_0(A\cup B)}. \tag{34}$$

Now we receive

$$\begin{aligned}
p(B) &= 1-p(\neg B)\\
&\overset{(28)}{=} 1-(L(A)+\varepsilon)\cdot\frac{p_0(A\cap\neg B)}{p_0(A)}-(U(\neg A)-\varepsilon)\cdot\frac{p_0(\neg A\cap\neg B)}{p_0(\neg A)}\\
&\overset{(30)}{=} 1-U_0(\neg A\cap\neg B)\cdot\frac{U(\neg A)-\varepsilon}{p_0(\neg A)}+\Delta\cdot\frac{U(\neg A)-\varepsilon}{p_0(\neg A)}-p_0(A\cap\neg B)\cdot\frac{L(A)+\varepsilon}{p_0(A)}\\
&\overset{(31)\text{–}(34)}{\geq} 1-U(\neg A\cap\neg B)+\Delta\cdot\frac{L(A\cup B)}{L_0(A\cup B)}-p_0(A\cap\neg B)\cdot\frac{L(A\cup B)}{L_0(A\cup B)}\\
&\overset{(30)}{=} p_0(B)\cdot\frac{L(A\cup B)}{L_0(A\cup B)} \quad\overset{(29)}{\geq}\quad L(B).
\end{aligned}$$

Hence Theorem 4 is proven. □

we did not presuppose this monotonicity. But we can argue as follows: Assume $L_0(A\cup B)=0$. Then $U_0(\neg A\cap\neg B)=1$, thus by (21), $U(\neg A\cap\neg B)=1$. Using (17), we get $U(\neg A)=U_0(\neg A\cap\neg B)\cdot U(\neg A)\leq U(\neg A\cap\neg B)\cdot U_0(\neg A)=U_0(\neg A)$, contradicting (22).

[13] Concerning the modified proof of model 2 mentioned above, note that due to (16) this case does not occur, if $\varepsilon=0$.

6 Concluding Remarks

To start with a topic raised in Section 1, consider again equation (2), that is $L = f \circ L_0$, with the standard assumption that $f\colon [0;1] \to [0;1]$ is a function with $f(0) = 0$ and $f(1) = 1$. Already in [2], Proposition 5.2, it is shown that via (2) every *convex* f transfers any given F-field $\mathcal{O}_0 = (\Omega; \mathcal{A}; L_0)$ to a neighbourhood $\mathcal{O} = (\Omega; \mathcal{A}; L)$, which is an F-field too. But in a strict sense, this neighbourhood model is not “appropriate”, since by Theorem 1, model 3, there is a weaker condition on f doing the same — namely the condition of bi-elasticity. The question arises, whether this requirement is “appropriate” instead. Indeed, bi-elasticity is even the weakest assumption on f ensuring that via (1), i.e. $L = f \circ p$, every K-function $p \in \mathcal{K}(\Omega; \mathcal{A})$ is transfered to an F-neighbourhood $\mathcal{O} = (\Omega; \mathcal{A}; L)$, if we are allowed to vary the underlying measurable space $(\Omega; \mathcal{A})$. For this, there is a quick argument, if additionally it is assumed that our functions f are continuous on $]0;1[$. In this case it is even sufficient to consider all *finite* measurable spaces and, for each of them, only *one* central distribution p “testing” equation (1):

Let f be fixed, being continuous on $]0;1[$ and having the above-mentioned property of generating F-neighbourhoods via (1). We restrict ourselves in deriving condition (4), where, by continuity, it is possible to assume that in there y and z are rational numbers: $y = \frac{i}{k}$ and $z = \frac{j}{k}$ for $0 < i \leq j \leq k$. Now, for this $k \in \mathbb{N}$, we walk up to the finite measurable space $(\Omega_k; \mathcal{P}(\Omega_k))$ and employ the corresponding classical uniform probability $p = p_0^k$ as central distribution, generating via (1) an F-neighbourhood $\mathcal{O} = (\Omega_k; \mathcal{P}(\Omega_k); L)$ (cf. Example 1). Hence $\mathcal{O}$ is an uniform F-field on $(\Omega_k; \mathcal{P}(\Omega_k))$, which — according to the last sentence in Example 1 — obeys the chain (14), especially its left part. Finally, an easy transformation leads to the desired inequality in (4).

Apart from this — last — positive result given here, many questions concerning the role of bi-elasticity within the theory of interval probability remain open. For example, the *concept of conditional interval probability* is still debated (see [10] or [12] and the references therein). In particular, Weichselberger’s notion of the *canonical concept* has the disadvantage that some constructions are not closed w.r.t. the F-property. Using the corresponding notation $\Psi(A \mid B) = \frac{\Psi(A)}{\Psi(B)}$ for every $A, B \in \mathcal{A}$ with $A \subseteq B$ such that $\Psi(B) \neq 0$, where Ψ can be a K-function as well as the lower or the upper bound of a probability field, it is possible to rephrase sensitively Theorem 3, models 1 and 2, and Theorem 4. Considering the outcome, it perhaps is feasible to modify these theorems in a way, which is profitable for a better understanding of the phenomenon of conditional interval probability.

Summarizing, the results presented in this article can be seen as the formal basis for joining together robust statistics and interval probability in its most expressive form, i.e., the concept of coherent or F-probabilities. The systematic development of distorted probabilities should be able to initiate a variety of applications in robust statistics and beyond.

Acknowledgements

I thank Thomas Augustin and three anonymous referees for their valuable remarks.

References

[1] T. Augustin. *Optimale Tests bei Intervallwahrscheinlichkeit*. Vandenhoeck & Ruprecht, Göttingen, 1998.

[2] T. Augustin. Neyman-Pearson testing under interval probability by globally least favorable pairs — Reviewing Huber-Strassen theory and extending it to general interval probability. *Journal of Statistical Planning and Inference*, 105: 149–173, 2002.

[3] A. Buja. On the Huber-Strassen theorem. *Probability Theory and Related Fields*, 73: 149–152, 1986.

[4] D. Denneberg. *Non-Additive Measure and Integral*. Kluwer, Dordrecht, 1994.

[5] D. Greenwald. *The McGraw-Hill Dictionary of Modern Economics*. McGraw-Hill Book Company, New York, 1973.

[6] P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.

[7] P.J. Huber and V. Strassen. Minimax tests and the Neyman-Pearson lemma for capacities. *Annals of Statistics*, 1: 251–263, 1973. Correction: 2: 223–224, 1974.

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

[9] A. Wallner. *Beiträge zur Theorie der Intervallwahrscheinlichkeit — Der Blick über Kolmogorov und Choquet hinaus*. Dr. Kovač, Hamburg, 2002.

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

[11] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I — Intervallwahrscheinlichkeit als umfassendes Konzept*, in cooperation with T. Augustin and A. Wallner, Physica, Heidelberg, 2001.

[12] K. Weichselberger and T. Augustin. *On the Symbiosis of Two Concepts of Conditional Interval Probability*. `www.stat.uni-muenchen.de/~thomas/weichselberger-augustin-isipta03long.pdf`, 2003.

Anton Wallner is with the Department of Statistics, Ludwig-Maximilians-University Munich, Ludwigstr. 33, 80539 München, Germany. E-mail: toni@stat.uni-muenchen.de

On the Symbiosis of Two Concepts of Conditional Interval Probability[*]

KURT WEICHSELBERGER
Ludwig-Maximilians-University Munich, Germany

THOMAS AUGUSTIN
Ludwig-Maximilians-University Munich, Germany

Abstract

This paper argues in favor of the thesis that two different concepts of conditional interval probability are needed, in order to serve the huge variety of tasks conditional probability has in the classical setting of precise probabilities. We compare the commonly used intuitive concept of conditional interval probability with the canonical concept, and see, in particular, that the canonical concept is the appropriate one to generalize the idea of transition kernels to interval probability: only the canonical concept allows reconstruction of the original interval probability from the marginals and conditionals, as well as the powerful formulation of Bayes Theorem.

Keywords

conditional interval probability, intuitive concept of conditional interval probability, canonical concept of conditional interval probability, conditioning, updating, theorem of total probability, Markov chains, Bayes theorem, decision theory

1 Introduction

In the last years a comprehensive theory of interval probability has been developed which systematically generalizes Kolmogorov's axiomatic approach to classical probability. Just as in Kolmogorov's approach, the basic axioms have to be supplemented by appropriate concepts of independence and by a definition of conditional probability.

The goal of the theory of interval probability is not only the creation of methods for dealing with imprecise probability but also a systematic one: the establishment of a body of definitions and results comparable to the analogous elements of

[*] We thank Anton Wallner for his valuable remarks and his support.

the classical theory with respect to rigidity and efficiency but with a much wider field of appropriate application.

While the system of axioms describing the properties of probability assignments is thoroughly discussed in [27] (see also [25] and [26]), the necessary supplements concerning independence and conditional probability are not included in that volume. A report summarizing basic aspects results in the statement that there is need for two different definitions of conditional probability associated with different roles in employing interval probability: the intuitive concept of conditional probability and the canonical concept of conditional probability ([26]).

The intuitive concept of conditional probability is widely used as the only generalization of classical conditional probability to imprecise probability in general (for a recent study in the context of numerical possibility theory, see [8], section 6). This way of generalizing conditional probability was rigorously justified by Walley [22], who derived it from coherence considerations between gambles and contingent gambles. It is almost exclusively used in statistical inference with imprecise probabilities: In particular, it underlies Walley's imprecise Dirichlet model (cf. [23], see also, e.g., [3] and [31]), and it is often understood as self-evident in robust Bayesian inference (e.g. [24], [18]).

Mainly in the area of artificial intelligence, often another definition of conditional interval probability is applied. It dates back to Dempster [10] and his proposed method of statistical inference. Since Shafer [19] it is often used isolated from its original motivation as *Dempster's rule of conditioning*. It has experienced many modifications, see [30] for a comparison of different proposals.

Only very few authors have argued in favor of a symbiosis of different concepts of conditional probabilities. Dubois and Prade [11] use the intuitive concept for what they call 'focusing', and Dempster's rule for 'conditioning'. Halpern and Fagin [12] stress that there are different ways to understand belief functions, a fact which naturally leads to different concepts of conditional probability.

Weichselberger argues that the intuitive concept has to be supplemented by the canonical concept, which in rare situations produces the same result as the concept proposed by Dempster. Although in many situations the canonical concept does not qualify for being interpreted as an assignment of interval probability itself, it serves as the inevitable bearer of information for solving important problems. This is not surprising, since even in classical theory conditional probability has two different roles: sometimes as an information of its own value, but in other cases as a tool allowing the derivation of other quantities. In the theory of interval probability canonical conditional probability and canonical conditional expectation can be used for such purposes, irrespective of their qualification as interval probability or as interval expectation. The relation between the two concepts of conditional interval probability with respect to different situations is the subject of the present article. After introducing basic requirements in Section 2, in Sections 3 to 6 we compare the consequences of the employment of each concept with respect to some relevant aspect of conditioning. Section 7 contains the conclusions which can be drawn.

2 Basic Concepts

Every probability measure in the usual sense, i.e. every set function $p(.)$ satisfying Kolmogorov’s axioms, is called a *classical probability*. The set of all classical probabilities on a measurable space $(\Omega;\mathcal{A})$ will be denoted by $\mathcal{K}(\Omega;\mathcal{A})$. According to [27] axioms for interval probability $P(.) = [L(.),U(.)]$ can be obtained by describing the relation between the non-additive set functions $L(.)$ and $U(.)$ and the set of classical probabilities being in accordance with them. Set functions $P(.)\colon \mathcal{A} \to \mathcal{Z}_0 := \{[L;U] \,|\, 0 \leq L \leq U \leq 1\}$, $A \mapsto P(A) = [L(A);U(A)]$, with $\mathcal{M} := \{p(.) \in \mathcal{K}(\Omega;\mathcal{A}) \,|\, L(A) \leq p(A) \leq U(A), \forall A \in \mathcal{A}\} \neq \emptyset$ are called *R-probability* with *structure* $\mathcal{M}$. If additionally $\inf_{p(.)\in\mathcal{M}} p(A) = L(A)$, and $\sup_{p(.)\in\mathcal{M}} p(A) = U(A)$, $\forall A \in \mathcal{A}$, hold, then $P(.)$ is *F-probability*. (With allowance to the different attitudes of Kolmogorov and de Finetti towards σ-additivity R-probability materially corresponds to a probability assignment ‘avoiding sure loss’ described by interval limits and F-probability to a ‘coherent’ assignment by interval limits.) The triple $\mathcal{F} = (\Omega;\mathcal{A};L(.))$ is called an *F-probability field*.

A non-empty subset $\mathcal{V}$ of $\mathcal{M}$ is called a *prestructure* of $\mathcal{F} = (\Omega;\mathcal{A};L(.))$ if the following equations hold: $\inf_{p(.)\in\mathcal{V}} p(A) = L(A)$, $\sup_{p(.)\in\mathcal{V}} p(A) = U(A)$, $\forall A \in \mathcal{A}$. The concept of independence[1] is introduced by

Definition 1 *Let $\mathcal{F} = (\Omega;\mathcal{A};L(.))$ be an F-probability field with structure $\mathcal{M}$ and let $\mathcal{C}_i$, $i = 1,2$, be partitions of Ω. Then $\mathcal{C}_1$ and $\mathcal{C}_2$ are* mutually independent, *if the set $\mathcal{M}_I = \{p(.) \in \mathcal{M} \,|\, p(A_1 \cap A_2) = p(A_1)\cdot p(A_2), \forall A_i \in \mathcal{C}_i, i = 1,2\}$ serves as a prestructure of the field $\mathcal{F}$.* □

In [26] this definition is illustrated in the case of a fourfold-table. As also mentioned there, apart from cases for which at least one of the marginal probabilities is a classical probability, the structure $\mathcal{M}$ will always contain classical probabilities with some dependence of $\mathcal{C}_1$ and $\mathcal{C}_2$ in the classical sense.

The classical concept of conditional probability can be generalized to interval probability in two different ways, generating on the one hand the intuitive concept, on the other hand the canonical concept of conditional probability, two concepts with different properties in many respects.

[1] The question how to generalize the notion of independence has received considerable attention (see, e.g., [4] for a survey and [13] for a comprehensive treatment in the context of random sets). Recently, in particular the concepts of epistemic irrelevance and independence, introduced by Walley [22], have been investigated in detail (see, among others, [7], [14], [16], [17], [21], [20]).

In the context studied here the most natural definition is to call two partitions independent if the structure of the underlying F-probability field is generated by the set of independent classical probabilities (cf. [25], [26]). This way of defining independence corresponds to the notion of strong extension ([4], [6]).

Definition 2 *Let $\mathcal{F} = (\Omega; \mathcal{A}; L(.))$ be an F-probability field with structure $\mathcal{M}$, and $\mathcal{C}$ be a partition of Ω where $U(C) > 0$, $\forall C \in \mathcal{C}$. With $\mathcal{M}_C := \{p(.) \in \mathcal{M} \mid p(C) > 0\}$ the* intuitive concept of conditional probability *is given by defining $iP_C(A \mid C) = [iL_C(A \mid C); iU_C(A \mid C)]$, where*

$$iL_C(A \mid C) = \inf_{p(.)\in\mathcal{M}_C} \frac{p(A\cap C)}{p(C)},\, iU_C(A \mid C) = \sup_{p(.)\in\mathcal{M}_C} \frac{p(A\cap C)}{p(C)}, \forall A \in \mathcal{A}, \forall C \in \mathcal{C}.$$

□

It can be demonstrated that this definition generates a conditional F-field for every C with $U(C) > 0$. The motivation of employing the intuitive concept is straightforward: As long as $L(C) > 0$ it may be understood as the transition from the structure $\mathcal{M}$ to the structure $i\mathcal{M}_C(. \mid C) = \{p(. \mid C) \mid p(.) \in \mathcal{M}\}$, which consists exactly of all classical conditional probabilities corresponding to elements of the original structure $\mathcal{M}$. The conditional interval expectation of any gain function $G(.)$ — defined for all elements E of Ω — therefore is calculated as $i\mathbb{E}(G(.) \mid C) = [i\mathbb{L}(G(.) \mid C); i\mathbb{U}(G(.) \mid C)] = \{E_p(G(.)) \mid p(.) \in i\mathcal{M}_C(. \mid C)\}$.

Weichselberger ([25], [26]) argues that the intuitive concept has to be supplemented by a concept, which is derived from a canon of desirable properties, and therefore is called the *canonical concept of conditional interval probability*.

Definition 3 *Let $\mathcal{F} = (\Omega; \mathcal{A}; L(.))$ be an F-probability field and $\mathcal{C}$ be a partition of Ω where $L(C) > 0$, $\forall C \in \mathcal{C}$. The canonical concept of conditional probability is given by $L_C(A \mid C) := \frac{L(A\cap C)}{L(C)}$ and $U_C(A \mid C) := \frac{U(A\cap C)}{U(C)}$, $\forall A \in \mathcal{A}$, $\forall C \in \mathcal{C}$.*

The canonical concept of conditional expectation of the gain function $G(.)$ for each $C \in \mathcal{C}$ with $L(C) > 0$ is defined as $\mathbb{E}[G(.) \mid C] := [\mathbb{L}(G(.) \mid C); \mathbb{U}(G(.) \mid C)]$ with $\mathbb{L}(G(.) \mid C) := \frac{\mathbb{L}(G(.)\cap C)}{L(C)}$, $\mathbb{U}(G(.) \mid C) := \frac{\mathbb{U}(G(.)\cap C)}{U(C)}$ and $\mathbb{L}(G(.) \cap C) = \inf_{p\in\mathcal{M}} \sum_{E\subseteq A} G(E) \cdot p(E)$, $\mathbb{U}(G(.) \cap C) = \sup_{p\in\mathcal{M}} \sum_{E\subseteq A} G(E) \cdot p(E)$. □

Three simple examples demonstrate the different types of conditional probability according to the canonical concept. Each is constructed from an F-probability field on $\Omega = E_1 \cup E_2 \cup E_3$ with $\mathcal{A} = \mathcal{P}(\Omega)$, and the same partition $\mathcal{C} = (C_1, C_2)$ with $C_1 = E_1 \cup E_2$, $C_2 = E_3$ is considered. Example 1 describes a constellation $(\mathcal{F}; \mathcal{C})$ with conditional F-probability according to the canonical concept.

Example 1 *An F-probability field $\mathcal{F}^{(1)}$ is given by*

$$\begin{array}{ll} P(E_1) = [0.10; 0.30] & \quad P(E_1 \cup E_2) = [0.40; 0.60] \\ P(E_2) = [0.20; 0.45] & \quad P(E_1 \cup E_3) = [0.55; 0.80] \\ P(E_3) = [0.40; 0.60] & \quad P(E_2 \cup E_3) = [0.70; 0.90] \end{array}$$

Because of $L(C_1) = 0.40$, $U(C_1) = 0.60$ the conditional probability according to *the canonical concept is given by*

$$\begin{array}{ll} P_{\mathcal{C}}(E_1 \mid C_1) = [0.25;0.50] & P_{\mathcal{C}}(E_1 \mid C_2) = [0] \\ P_{\mathcal{C}}(E_2 \mid C_1) = [0.50;0.75] & P_{\mathcal{C}}(E_2 \mid C_2) = [0] \\ P_{\mathcal{C}}(E_3 \mid C_1) = [0] & P_{\mathcal{C}}(E_3 \mid C_2) = [1] \end{array}$$

It is easily seen that in this case both conditional probability fields are F-probability. In addition the results may be compared with those from applying the intuitive concept:

$$\begin{array}{ll} iP_{\mathcal{C}}(E_1 \mid C_1) = [0.182;0.600] & iP_{\mathcal{C}}(E_1 \mid C_2) = [0] \\ iP_{\mathcal{C}}(E_2 \mid C_1) = [0.400;0.818] & iP_{\mathcal{C}}(E_2 \mid C_2) = [0] \\ iP_{\mathcal{C}}(E_3 \mid C_1) = [0] & iP_{\mathcal{C}}(E_3 \mid C_2) = [1] \end{array}$$ □

It can be shown that interval limits resulting from the intuitive concept cannot be narrower than those arising from the canonical one (cf., e.g., [22], p. 301). In general for all $C \in \mathcal{C}$, $iP_{\mathcal{C}}(E_i \mid C) \supsetneqq P_{\mathcal{C}}(E_i \mid C)$ holds. Both concepts coincide if the marginals consist of classical probabilities.[2] Example 2 shows a constellation $(\mathcal{F};\mathcal{C})$ for which the conditional probability according to the canonical concept possesses R-quality but not F-quality.

Example 2 *The F-field $\mathcal{F}^{(2)}$ is given by*

$$\begin{array}{ll} P(E_1) = [0.10;0.25] & P(E_1 \cup E_2) = [0.40;0.60] \\ P(E_2) = [0.20;0.40] & P(E_1 \cup E_3) = [0.60;0.80] \\ P(E_3) = [0.40;0.60] & P(E_2 \cup E_3) = [0.75;0.90] \end{array}$$

The conditional probability according to the canonical concept now reads as follows:

$$\begin{array}{ll} P_{\mathcal{C}}(E_1 \mid C_1) = [0.250;0.417] & P_{\mathcal{C}}(E_1 \mid C_2) = [0] \\ P_{\mathcal{C}}(E_2 \mid C_1) = [0.500;0.667] & P_{\mathcal{C}}(E_2 \mid C_2) = [0] \\ P_{\mathcal{C}}(E_3 \mid C_1) = [0] & P_{\mathcal{C}}(E_3 \mid C_2) = [1] \end{array}$$

The fact, that $L_{\mathcal{C}}(E_1 \mid C_1) + U_{\mathcal{C}}(E_2 \mid C_1) \neq 1$ and $L_{\mathcal{C}}(E_2 \mid C_1) + U_{\mathcal{C}}(E_1 \mid C_1) \neq 1$ makes it clear, that $P_{\mathcal{C}}(. \mid C_1)$ is not F-probability. On the other hand the assignment $p_{\mathcal{C}}(E_1 \mid C_1) = 0.4$, $p_{\mathcal{C}}(E_2 \mid C_1) = 0.6$, $p_{\mathcal{C}}(E_3 \mid C_1) = 0.0$ is an element of the structure of this field: The canonical concept here produces an R-field, but not an F-field. Again the intuitive concept produces an F-probability-field with wider interval limits: $iP_{\mathcal{C}}(E_1 \mid C_1) = [0.200;0.556]$ and $iP_{\mathcal{C}}(E_2 \mid C_1) = [0.444;0.800]$ completed by the same trivial interval limits as in Example 1. □

Example 3 describes a constellation for which canonical conditional probability has not even R-quality, since it contains intervals for which $L > U$ holds.

[2] For an attractive example for this special situation see the nonparametric predictive inference discussed in [2].

Example 3 *The F-field $\mathcal{F}^{(3)}$ is given through*

$$\begin{array}{ll} P(E_1) = [0.16; 0.18] & P(E_1 \cup E_2) = [0.40; 0.60] \\ P(E_2) = [0.22; 0.42] & P(E_1 \cup E_3) = [0.58; 0.78] \\ P(E_3) = [0.40; 0.60] & P(E_2 \cup E_3) = [0.82; 0.84] \end{array}$$

The canonical concept produces: $P_C(E_1 \mid C_1) = [0.40; 0.30]$ and $P_C(E_2 \mid C_1) = [0.55; 0.70]$ and the same trivial interval limits as the foregoing examples. Since $L_C(E_1 \mid C_1) > U_C(E_1 \mid C_1)$, it is impossible to find K-functions in accordance with the interval limits: A structure does not exist. Concerning the intuitive concept, there are no problems: $iP_C(E_1 \mid C_1) = [0.276; 0.450]$ and $iP_C(E_2 \mid C_1) = [0.550; 0.724]$. □

It is obvious that in a case like this the outcome of the canonical concept cannot be interpreted as interval probability in the usual sense. In order to allow the employment of the word *probability*, the usage of this expression has to be extended.

Definition 4 *Given a sample space Ω and a σ-field $\mathcal{A}$ of random events in Ω, $P(A) = [L(A); U(A)]$, $\forall A \in \mathcal{A}$, is named* O-probability, *if* $0 \leq L(A), U(A) \leq 1$, *$\forall A \in \mathcal{A}$.* □

$P(A), A \in \mathcal{A}$, need not be intervals, $L(A)$ may be larger than $U(A)$. It will be shown in the following sections that the canonical concept produces results which are bearers of important information, even they do not qualify for being interpreted as R-probability or as interval expectation.

3 Independence and Conditional Probability

In the classical theory mutual independence of two partitions $\mathcal{C}_1$ and $\mathcal{C}_2$ can be characterized by

$$p(A_1 \mid A_2) = p(A_1), \quad \forall A_1 \in \mathcal{C}_1, \forall A_2 \in \mathcal{C}_2 : p(A_2) \neq 0. \tag{1}$$

If the intuitive concept of conditional interval probability and the definition of independence along the lines of Definition 1 are applied, this appealing property does not hold in general. This fact led to the introduction of the notions of epistemic irrelevance and epistemic independence (see the references in footnote 1), which use variants of (1) to define independence.

In contrast, (1) extends to interval probability if the canonical concept of conditional probability is employed (and $L(A_2) \neq 0$). According to Definition 1 $L(A_1 \cap A_2) = L(A_1) \cdot L(A_2)$ and $U(A_1 \cap A_2) = U(A_1) \cdot U(A_2)$ hold. This leads to

Theorem 1 *If $\mathcal{F} = (\Omega; \mathcal{A}; L(.))$ is an F-probability field and $\mathcal{C}_1$, $\mathcal{C}_2 \subseteq \mathcal{A}$ are partitions of Ω, the statements a) and b) are equivalent:*

1. *$P_{\mathcal{C}_2}(A_1 \mid A_2) = P(A_1), \quad \forall A_1 \in \mathcal{C}_1, A_2 \in \mathcal{C}_2 : L(A_2) \neq 0.$*
2. *$\mathcal{C}_1$ and $\mathcal{C}_2$ are mutually independent.* □

It is, therefore, in this case guaranteed that the canonical concept produces conditional F-probability fields. On the other hand: Since the interval limits of the intuitive concept are generally wider than that of the canonical one, in the case of mutual independence $iP_{\mathcal{C}_2}(A_1|A_2) \supsetneqq P(A_1)$ must be expected. Employing the model of double-dichotomy this phenomenon is demonstrated in Example 4.

Example 4 *Let $\mathcal{F} = (\Omega; \mathcal{A}; L(.))$ be an F-probability field with $\Omega_4 = E_1 \cup E_2 \cup E_3 \cup E_4$, $\mathcal{A} = \mathcal{P}(\Omega_4)$ and*

$$\begin{array}{ll} P(E_1) = [0.08; 0.21] & P(E_1 \cup E_2) = [0.20; 0.30] \\ P(E_2) = [0.06; 0.18] & P(E_1 \cup E_3) = [0.40; 0.70] \\ P(E_3) = [0.28; 0.49] & P(E_1 \cup E_4) = [0.33; 0.66] \\ P(E_4) = [0.21; 0.48] & \end{array}$$

(The remaining components of this F-field follow from $L(A) + U(\neg A) = 1, \forall A \in \mathcal{A}$.) Let two partitions be given by $\mathcal{C}_1 = (A_1, \neg A_1)$, where $A_1 = (E_1 \cup E_2)$, and $\mathcal{C}_2 = (A_2, \neg A_2)$, where $A_2 = (E_1 \cup E_3)$. By means of a four-fold table independence of $\mathcal{C}_1$ and $\mathcal{C}_2$ is directly controlled:

$P(E_1)$=[0.08; 0.21]	$P(E_1)$=[0.06; 0.18]	$P(E_1 \cup E_2)$=[0.20; 0.30]
$P(E_3)$=[0.28; 0.49]	$P(E_4)$=[0.21; 0.48]	$P(E_3 \cup E_4)$=[0.70; 0.80]
$P(E_1 \cup E_3)$=[0.40; 0.70]	$P(E_2 \cup E_4)$=[0.30; 0.60]	$P(\Omega_4)$=[1]

$$L(E_1 \cup E_4) = \max(L(E_1) + L(E_4), 1 - U(E_2) - U(E_3)) = 0.33$$
$$U(E_1 \cup E_4) = \min(U(E_1) + U(E_4), 1 - L(E_2) - L(E_3)) = 0.66.$$

The canonical concept of conditional probability produces:

$$L_{\mathcal{C}_2}(A_1 \mid A_2) = L_{\mathcal{C}_2}(E_1 \cup E_2 \mid E_1 \cup E_3) = \frac{L(E_1)}{L(E_1 \cup E_3)} = \frac{0.08}{0.40} = 0.20 = L(E_1 \cup E_2)$$
$$U_{\mathcal{C}_2}(A_1 \mid A_2) = U_{\mathcal{C}_2}(E_1 \cup E_2 \mid E_1 \cup E_3) = \frac{U(E_1)}{U(E_1 \cup E_3)} = \frac{0.21}{0.70} = 0.30 = U(E_1 \cup E_2).$$

The intuitive concept leads to

$$iL_{\mathcal{C}_2}(A_1 \mid A_2) = \inf_{\mathcal{M}} p(E_1 \mid E_1 \cup E_3) = \frac{0.08}{0.08 + 0.49} = 0.140 < 0.20$$
$$iU_{\mathcal{C}_2}(A_1 \mid A_2) = \sup_{\mathcal{M}} p(E_1 \mid E_1 \cup E_3) = \frac{0.21}{0.21 + 0.28} = 0.429 > 0.30.$$
□

The conclusion from these results is evident: If it is of importance, that in the case of mutual independence conditional and marginal probability are equal, then the canonical concept of conditional probability must be employed.

4 Updating with Conditional Probability

The essential aspects concerning updating by means of conditional interval probability already become clear in the simple case of two states, A_1 and A_2, and two (or later, three) possible diagnoses, B_1 and B_2 (and later, B_3). If the overall probability is given by an F-field $\mathcal{F} = (\Omega; \mathcal{P}(\Omega); L(.))$ with $|\Omega| = 4$, one has

$$\begin{array}{cc|c}
P(A_1 \cap B_1) = [L_{11}; U_{11}] & P(A_1 \cap B_2) = [L_{12}; U_{12}] & P(A_1) = [L_{1.}; U_{1.}] \\
P(A_2 \cap B_1) = [L_{21}; U_{21}] & P(A_2 \cap B_2) = [L_{22}; U_{22}] & P(A_2) = [L_{2.}; U_{2.}] \\
\hline
P(B_1) = [L_{.1}; U_{.1}] & P(B_2) = [L_{.2}; U_{.2}] & P(\Omega_4) = [1]
\end{array}$$

While the prior probability of state A_1 is given by $P(A_1)$, updating in case of diagnosis B_1 produces the conditional probability of $(A_1 \cap B_1)$ given B_1. If more than two diagnoses are possible, it is important to ensure that the process of updating is associative: Does stepwise learning lead to the same result as instantaneous learning? In the case of classical probability the answer is affirmative.

An F-probability field $\mathcal{F} = (\Omega; \mathcal{P}(\Omega); L(.))$ with $|\Omega| = 6$ is given by:

$$\begin{array}{cc|c|c|c}
P_{11} & P_{12} & P_{1S} & P_{13} & P_{1.} \\
P_{21} & P_{22} & P_{2S} & P_{23} & P_{2.} \\
\hline
P_{.1} & P_{.2} & P_{.S} & P_{.3} & [1]
\end{array}
\quad \text{where} \quad
\begin{array}{l}
P_{ij} := P(A_i \cap B_j) \\
P_{i.} := P(A_i),\ i = 1, 2 \\
P_{.j} := P(B_j),\ j = 1, 2, 3 \\
P_{iS} := P(A_i \cap (B_1 \cup B_2)) \\
P_{.S} := P(B_1 \cup B_2)
\end{array}$$

and in an analogous way for L and U.

Let instantaneous learning immediately transfer the information from Ω to B_1, while stepwise learning leads from Ω to $B_1 \cup B_2$ and from there to B_1. A method of updating can only be accepted, if the final result is equal in both cases.

For the intuitive concept it is sufficient to remember that for classical probability the equation $p(A \mid B_1) = \frac{p(A \cap B_1 \mid B_1 \cup B_2)}{p(B_1 \mid B_1 \cup B_2)}$ is valid. This is not only the reason, why associativity holds for updating with the classical conditional probability; it also means that $i\mathcal{M}(. \mid B_1) = \left\{ p(. \mid B_1) \mid p(.) \in i\mathcal{M}(. \mid B_1 \cup B_2) \right\}$ must be true and updating with the intuitive concept of conditional probability produces the same results for instantaneous and for stepwise learning.

With respect to the canonical concept the first step of information (“$B_1 \cup B_2$”) produces the conditional probability field with the following interval limits:

$$\begin{array}{cc|c}
\left[\frac{L_{11}}{L_{.S}}; \frac{U_{11}}{U_{.S}}\right] & \left[\frac{L_{12}}{L_{.S}}; \frac{U_{12}}{U_{.S}}\right] & \left[\frac{L_{1S}}{L_{.S}}; \frac{U_{1S}}{U_{.S}}\right] \\
\left[\frac{L_{21}}{L_{.S}}; \frac{U_{21}}{U_{.S}}\right] & \left[\frac{L_{22}}{L_{.S}}; \frac{U_{22}}{U_{.S}}\right] & \left[\frac{L_{2S}}{L_{.S}}; \frac{U_{2S}}{U_{.S}}\right] \\
\hline
\left[\frac{L_{.1}}{L_{.S}}; \frac{U_{.1}}{U_{.S}}\right] & \left[\frac{L_{.2}}{L_{.S}}; \frac{U_{.2}}{U_{.S}}\right] & [1]
\end{array}$$

The second step of information (“B_1”) leads to the interval limits $\frac{L_{11}}{L_{.S}} : \frac{L_{.1}}{L_{.S}} = \frac{L_{11}}{L_{.1}}$, $\frac{U_{11}}{U_{.S}} : \frac{U_{.1}}{U_{.S}} = \frac{U_{11}}{U_{.1}}$, which are the same as if the information “B_1” had been given at once. Therefore the canonical concept satisfies the necessary condition for reasonable updating as well.

It may be concluded that in principle each of the two concepts can be employed for updating. Since the intuitive concept guarantees the F-property of the outcome it should be preferred under usual circumstances.

5 Transfer of Information

The idea of conditional probability often is employed in designing new models, combining marginal probability derived from one source of information, with conditional probability gained from another source. In particular the theory of Markov chains relies on this principle: The dynamic evolution is completely described by specifying an initial distribution and a matrix of transition probabilities, consisting of the conditional probabilities to reach a state i given state j.

A necessary condition for the qualification of any concept of conditional probability with respect to such transfer obviously is the possibility to reconstruct an F-probability field by means of marginal probability and conditional probability. It was demonstrated in [26], that this reconstruction need not be possible if the intuitive concept is employed: different F-fields may be equal with respect to marginal probability and to intuitive conditional probability for a certain partition. This phenomenon is quite common for the intuitive concept: There are very rare borderline cases where it is possible to determine an F-field uniquely by means of marginal probability and the respective intuitive conditional probability.

On the other hand, reconstruction of an F-probability field using the marginal probability of a partition together with the canonical conditional probability is practicable, if a so called *laminar constellation* in the following sense is given.

Definition 5 *i)* $(\mathcal{A}_L, \mathcal{A}_U)$ *is named a* support *of the F-field* $\mathcal{F} = (\Omega; \mathcal{A}; L(.))$ *with structure* $\mathcal{M}$*, if the set of equations:* $L(A) \leq p(A), \forall A \in \mathcal{A}_L$*, and* $p(A) \leq U(A), \forall A \in \mathcal{A}_U$*, is sufficient to determine* $\mathcal{M}$*.*

ii) A constellation $(\mathcal{F}, \mathcal{C})$*, consisting of an F-field* $\mathcal{F} = (\Omega; \mathcal{A}; L(.))$ *and a partition* $\mathcal{C}$ *of* Ω*, is named a* laminar constellation, *if there exists a support* $(\mathcal{A}_L, \mathcal{A}_U)$ *of* $\mathcal{F}$*, so that for each* $A \in \mathcal{A}_L \cup \mathcal{A}_U$ *one of the two following conditions is satisfied:*

1. $\exists C_{(1)}, \ldots, C_{(q)} \in \mathcal{C} : A = \bigcup_{i=1}^{q} C_{(i)}$.
2. $\exists C \in \mathcal{C} : A \subset C$. □

This definition characterizes constellations, where all information about the structure $\mathcal{M}$ — and therefore about $\mathcal{F}$ itself — is contained only in the marginal probability on $\mathcal{C}$ or in events which are subsets of single elements of the partition.

If laminarity of the constellation is given, reconstruction of the original F-field by means of marginal probability of $\mathcal{C}$ and the canonical conditional probabilities for all $C \in \mathcal{C}$ is possible, irrespective of the quality of the canonical conditional probabilities, since for each A satisfying condition a) the interval limits are determined by the marginal probability and for each A satisfying condition b) the interval limits are to be reconstructed by $L(A) = L_{\mathcal{C}}(A \mid C) \cdot L(C) = \frac{L(A)}{L(C)} \cdot L(C)$ and $U(A) = U_{\mathcal{C}}(A \mid C) \cdot U(C) = \frac{U(A)}{U(C)} \cdot U(C)$.

The reconstruction of an F-field using conditional O-probability is demonstrated in Example 5 for a sample space of size 3.

Example 5 *Let an F-field $\mathcal{F} = (\Omega_3; \mathcal{P}(\Omega_3); L(.))$ be given by:*

$$P(E_1) = [0.16; 0.21] \quad P(E_2) = [0.22; 0.42] \quad P(E_3) = [0.40; 0.60].$$

The partition $\mathcal{C} = (C_1, C_2)$ with $C_1 = E_1 \cup E_2$ and $C_2 = E_3$ leads to $P(C_1) =$ $[0.40; 0.60]$, $P(C_2) = [0.40; 0.60]$. It is obvious, that this is a laminar constellation: E_1 and E_2 obey condition b), E_3 satisfies condition a). The interval limits of conditional probability according to the canonical concept are:

$$\begin{array}{ll}
L_{\mathcal{C}}(E_1 \mid C_1) = 0.40 & U_{\mathcal{C}}(E_1 \mid C_1) = 0.35 \\
L_{\mathcal{C}}(E_2 \mid C_1) = 0.55 & U_{\mathcal{C}}(E_2 \mid C_1) = 0.70 \\
L_{\mathcal{C}}(E_3 \mid C_2) = 1 & U_{\mathcal{C}}(E_3 \mid C_2) = 1 \,.
\end{array}$$

Therefore $P_{\mathcal{C}}(E_1 \mid C_1) = [0.40; 0.35]$, $P_{\mathcal{C}}(E_2 \mid C_1) = [0.55; 0.70]$ is an assignment which can not be interpreted as a generalization of classical probability, but it is useful for reconstructing $\mathcal{F}$:

$$\begin{aligned}
L(E_1) &= L_{\mathcal{C}}(E_1 \mid C_1) \cdot L(C_1) = 0.40 \cdot 0.40 = 0.16 \\
U(E_1) &= U_{\mathcal{C}}(E_1 \mid C_1) \cdot U(C_1) = 0.35 \cdot 0.60 = 0.21 \\
L(E_2) &= L_{\mathcal{C}}(E_2 \mid C_1) \cdot L(C_1) = 0.55 \cdot 0.40 = 0.22 \\
U(E_2) &= U_{\mathcal{C}}(E_2 \mid C_1) \cdot U(C_1) = 0.70 \cdot 0.60 = 0.42 \\
L(E_3) &= L_{\mathcal{C}}(E_3 \mid C_2) \cdot L(C_2) = 1 \cdot 0.40 = 0.40 \\
U(E_3) &= U_{\mathcal{C}}(E_3 \mid C_2) \cdot U(C_1) = 1 \cdot 0.60 = 0.60 \,.
\end{aligned}$$

Because of the laminarity of the constellation $(\mathcal{F}, \mathcal{C})$ these interval limits are sufficient to reconstruct $\mathcal{F}$. □

The Theorem of Total Probability can be formulated as

Corollary 1 *If $(\mathcal{F}; \mathcal{C})$ is a laminar constellation, the F-field $\mathcal{F}$ is uniquely determined by the marginal probability field for $\mathcal{C}$ and by the canonical conditional probability fields resulting for each $C \in \mathcal{C}$, irrespective of the F- or R- or O-quality of the conditional fields.* □

This result allows interpretations:

1. If the conditional probabilities do not possess F-(R-)quality, there *cannot be any set* of conditional F-(R-)probabilities which allows to reconstruct the given F-field through the given marginal F-probability.

2. If the process of matching a given marginal F-probability with given canonical conditionals results in a field not possessing F-(R-)quality, it is *impossible to find* an F-(R-)field with this marginal and with these conditionals.

Therefore, the results of transfer of information from one model to another to some extent can be foreseen:

1. If $P_C(.\mid .)$ describes an F-field, matching with marginal F-probability always produces an F-field.

2. If $P_C(.\mid .)$ describes an R-field, matching with marginal F-probability produces either an F-field or an R-field which does not possess the F-quality.

3. If $P_C(.\mid .)$ does not fit to an R-field, nothing can be predicted about quality of the outcome, if it is matched with marginal F-probability.

6 The Theorem of Bayes

The Theorem of Bayes is an important result of classical probability theory. While it is of highest significance for any subjectivistic school, even the objectivistic view sometimes finds conditions, under which it is legitimate to accept a certain prior information which is described by classical probability. On the other hand even the subjectivist cannot deny that in most practically relevant cases the choice of a particular classical prior is at least highly debatable.

Therefore this is a situation inviting to propose the employment of interval probability. If a successful transfer of the Theorem of Bayes into the theory of interval probability can be achieved, a strong argument favouring the efficiency of this theory is presented.[3] Ambiguity, however, — distinguishing interval and classical probability — does not obey to those laws which are the basis of the Theorem of Bayes in the classical theory. It should therefore not be expected that the roles of this theorem in classical probability and in generalized probability are the same. References to the obvious limitations for the efficiency of particular types of this theorem have been given only recently ([29], [1]).

In classical probability the Theorem of Bayes results from the properties of the concept of conditional probability. Therefore it has to be expected that in the

[3] The Theorem of Bayes and the problem of computing posterior probabilities or posterior expectations is a frequent subject in literature dealing with generalized probability: In his fundamental book [22], Walley derived the 'Generalized Bayes Rule', which also is used in the robust Bayesian approach (see, f.i., [5], [24], [15], [18]).

theory of interval probability the role of conditional probability — and especially of the concept employed — proves to be decisive. The transition from prior probability to posterior probability necessarily consists of two steps:

1. Derivation of an F-probability field for which the prior is marginal probability and the conditional probability is given.
2. Derivation of conditional probability relative to the actual observation.

For the first step the method to be applied in case of interval probability is obvious: Marginal probability and conditional probability (due to the canonical concept) have to be combined by means of the Cartesian product of two structures. This generates the product rules $L(A \cap Z) = L(A) \cdot L(A \mid Z)$ and $U(A \cap Z) = U(A) \cdot U(A \mid Z)$. In Example 6 this procedure is demonstrated introducing a special case of double-dichotomy which will be employed in all of the examples to come.

Example 6 *Let the F-field describing the probability for a dichotomy of states of nature be given by* $P(Z_1) = [0.2; 0.3]$, $P(Z_2) = [0.7; 0.8]$, *and the probability of the outcome of a certain trial in case of state* Z_1 *be given by the F-field* $P(A_1 \mid Z_1) = [0.6; 0.7]$, $P(A_2 \mid Z_1) = [0.3; 0.4]$ *in case of state* Z_2 *by the F-field* $P(A_1 \mid Z_2) = [0.1; 0.2]$, $P(A_2 \mid Z_2) = [0.8; 0.9]$. *Interpreting the first of the three fields as marginal probability and the two others as conditional probability according to the canonical concept one arrives at the following components of an F-field describing the combined probability of the states and outcomes:*

$P(A_1 \cap Z_1)=[0.12; 0.21]$	$P(A_2 \cap Z_1)=[0.06; 0.12]$	$P(Z_1)=[0.2; 0.3]$
$P(A_1 \cap Z_2)=[0.07; 0.16]$	$P(A_2 \cap Z_2)=[0.56; 0.72]$	$P(Z_2)=[0.7; 0.8]$
$P(A_1)$	$P(A_2)$	$P(\Omega_4)=[1]$

This is partial determinate F-probability and the process of normal completion has to be employed in order to calculate the components $P(A_1)$ *and* $P(A_2)$. *In the present situation the results are gained easily: Let* $p(Z_1) = a$ *be a K-function belonging to the structure of the prior probability,* $p(A_1 \mid Z_1) = b$ *and* $p(A_1 \mid Z_2) = c$ *be K-functions belonging to the structures of two marginal probabilities. Therefore:* $0.2 \leq a \leq 0.3$*;* $0.6 \leq b \leq 0.7$*;* $0.1 \leq c \leq 0.2$*. The possible values of* $a \cdot b$ *produce* $P(A_1 \cap Z_1)$*, those of* $(1-a) \cdot c$ *produce* $P(A_1 \cap Z_2)$ *and the values of* $a \cdot b + (1-a) \cdot c$ *produce* $P(A_1)$*. It is easily controlled, that* $a = 0.2$*,* $b = 0.6$*,* $c = 0.1$ *render the minimum of* $a \cdot b + (1-a) \cdot c$*, so that* $L(A_1) = 0.20$ *results, and* $a = 0.3$*,* $b = 0.7$*,* $c = 0.2$ *render* $U(A_1) = 0.35$*. Since an F-field possesses conjugate interval limits, one arrives for* $A_1 = \neg A_2$ *at* $L(A_2) = 0.65$*,* $U(A_2) = 0.80$*. The last line of the table above reads:*

$$P(A_1) = [0.20; 0.35] \quad P(A_2) = [0.65; 0.80] \quad P(\Omega_4) = [1].$$

The results of the procedure described are those components of the combined F-field which are relevant with respect to posterior probability. The components still *lacking would be calculated in an analogous manner, for instance* $P[(A_1 \cap Z_1) \cup (A_2 \cap Z_2)] = [0.76; 0.86]$. □

While the canonical concept is inevitable for step 1, there is a possibility to choose between the concepts as far as step 2 is concerned: the calculation of the posterior probability for each observation. The decision in favour of the intuitive concept is quite common and promises some remarkable advantages:

1. The F-quality of the posterior probability is guaranteed.
2. The structure of this F-field can be interpreted as the Cartesian product of the structures of the marginal probability and of the conditional F-probability belonging to the actual observation.

On the other hand, use of the canonical concept includes the risk that the outcome cannot be interpreted as a generalization of a classical probability, since the resulting intervals do not define a structure.

It is therefore advisable to calculate posterior probability by means of the intuitive concept, if this posterior constitutes the only and final goal of the analysis. However, in the following it will be demonstrated, that there are good reasons for the opposite decision, if the posterior probability is to be employed as a basis for further analysis. Two situations will be considered:

1. The posterior probability of one trial is used as prior probability for another trial which is independent from the first one.
2. The posterior probability is the basis of a decision between different actions.

As to the *first of the two aspects*: In classical theory it is seen as one of the most important merits attributed to the employment of Bayes’ theorem that the transition from the prior probability to the posterior is a definitive one: After the trial the posterior takes over the role of the prior. If a next trial is independent from the first one the posterior of the former trial, therefore, is the prior of the next. Obviously the following must be seen as a substantial criterion for a successful transfer of Bayes’ theorem to interval probability: The results have to be the same, whether two mutually independent trials are combined to one trial, or the posterior of the first one is used as prior for the second one. It can be shown that these requirements are met, provided that the Theorem of Bayes is executed by means of the canonical concept of conditional probability. For brevity the proof will be limited to the case of two states of nature and two possible observations.

Proposition 1 *Let* (Z_1, Z_2) *be a dichotomy of the states of nature with the prior F-probability given by* $P(Z_1) = [L; U]$.

A first trial with possible outcome A_1 *or* A_2 *is characterized by F-probabilities given by* $P(A_1 \mid Z_1) = [l_{11}; u_{11}]$, $P(A_1 \mid Z_2) = [l_{21}; u_{21}]$. *A second trial which* *is independent from the first one, has the outcomes B_1 and B_2. The ruling F-probabilities are given by $P(B_1 \mid Z_1) = [l_{12}; u_{12}]$, $P(B_1 \mid Z_2) = [l_{22}; u_{22}]$. If in the Theorem of Bayes the canonical concept of conditional probability is employed, a trial which originates from a combination of the observations A. and B. renders the same posterior probability as the procedure, in which the posterior probability of the first trial is taken as prior probability for the second one.* □

For the *proof* of this proposition it is sufficient to show that both procedures produce the same probability components $P(A_i \cap B_j \cap Z_r)$ and $P(A_i \cap B_j)$, since the final probability is derived from the interval limits of these components. The demonstration will be given for $P(A_1 \cap B_1 \cap Z_r)$, $r = 1, 2$, and $P(A_1 \cap B_1)$.

1. In case of a combined trial, because of mutual independence of the trials one arrives at $P(A_1 \cap B_1 \mid Z_1) = [l_{11} \cdot l_{12}; u_{11} \cdot u_{12}]$, $P(A_1 \cap B_1 \mid Z_2) = [l_{21} \cdot l_{22}; u_{21} \cdot u_{22}]$ and together with the marginal probability of the states of nature: $P(A_1 \cap B_1 \cap Z_1) = [l_{11} \cdot l_{12} \cdot L; u_{11} \cdot u_{12} \cdot U]$, $P(A_1 \cap B_1 \cap Z_2) = [l_{21} \cdot l_{22} \cdot (1-U); u_{21} \cdot u_{22} \cdot (1-L)]$. These two components are sufficient to calculate $P(A_1 \cap B_1)$.
2. If the first trial is executed separately, conditional probability and marginal probability produce $P(A_1 \cap Z_1) = [l_{11} \cdot L; u_{11} \cdot U]$, $P(A_1 \cap Z_2) = [l_{21} \cdot (1-U); u_{21} \cdot (1-L)]$. The component of the union of these events[4] is designated by $P(A_1) = [L_1; U_1]$. The posterior probability of the first trial in case of observation A_1 — which will be used as prior for the second trial — is defined by the canonical conditional probability as $P(Z_1 \mid A_1) = \left[\frac{l_{11} \cdot L}{L_1}; \frac{u_{11} \cdot U}{U_1}\right]$, $P(Z_2 \mid A_1) = \left[\frac{l_{21} \cdot (1-U)}{L_1}; \frac{u_{21} \cdot (1-L)}{U_1}\right]$. Hence, conditional to A_1 the probability-components for the observation B_1 of the second trial read as $P(B_1 \cap Z_1 \mid A_1) = \left[\frac{l_{12} \cdot l_{11} \cdot L}{L_1}; \frac{u_{12} \cdot u_{11} \cdot U}{U_1}\right]$, $P(B_1 \cap Z_2 \mid A_1) = \left[\frac{l_{22} \cdot l_{21} \cdot (1-U)}{L_1}; \frac{u_{22} \cdot u_{21} \cdot (1-L)}{U_1}\right]$. In order to arrive at the components of the events $A_1 \cap B_1 \cap Z_1$ and $A_1 \cap B_1 \cap Z_2$, canonical conditional and marginal probability must be combined:
$L(A_1 \cap B_1 \cap Z_1) = L(B_1 \cap Z_1 \mid A_1) \cdot L(A_1) = \frac{l_{11} \cdot l_{12} \cdot L}{L_1} \cdot L_1 = l_{11} \cdot l_{12} \cdot L$,
$U(A_1 \cap B_1 \cap Z_1) = U(B_1 \cap Z_1 \mid A_1) \cdot U(A_1) = \frac{u_{11} \cdot u_{12} \cdot U}{U_1} \cdot U_1 = u_{11} \cdot u_{12} \cdot U$
and corresponding procedures for Z_2. Both components are equal to those resulting from the combined trial and consequently as well $P(A_1 \cap B_1)$ as the canonical conditional probability are alike: Both methods produce the same posterior. □

In Example 7 this equivalence is demonstrated in the case of the F-probability field introduced in Example 6.

Example 7 *For the prior probability and the conditional probability of Example 6 the posterior probability — defined by the canonical concept results as*

$$\begin{aligned} P(Z_1 \mid A_1) &= \left[\tfrac{0.12}{0.20}; \tfrac{0.21}{0.35}\right] = [0.60; 0.60] \\ P(Z_2 \mid A_1) &= \left[\tfrac{0.07}{0.20}; \tfrac{0.16}{0.35}\right] = [0.35; 0.46]. \end{aligned}$$

These two components can be interpreted as R-probability, since $p(Z_1|A_1) = 0.60$, $p(Z_2 \mid A_1) = 0.40$, is a K-function in accordance with all

[4] The appropriate method of calculation is demonstrated in Example 6.

interval limits. It will be seen that despite the lack of F-quality this assignment can be used as a prior for a next trial. Let

$$\begin{array}{ll} P(B_1 \mid Z_1) = [0.7; 0.9] & P(B_2 \mid Z_1) = [0.1; 0.3] \\ P(B_1 \mid Z_2) = [0.2; 0.4] & P(B_2 \mid Z_2) = [0.6; 0.8]. \end{array}$$

Combined with the new prior produced by observation A_1:

$$\begin{array}{ll|l} P(B_1 \cap Z_1 \mid A_1){=}[0.42; 0.54] & P(B_2 \cap Z_1 \mid A_1){=}[0.06; 0.18] & P(Z_1 \mid A_1){=}[0.60; 0.60] \\ P(B_1 \cap Z_2 \mid A_1){=}[0.07; 0.184] & P(B_2 \cap Z_2 \mid A_1){=}[0.21; 0.37] & P(Z_2 \mid A_1){=}[0.35; 0.46] \end{array}$$

In order to calculate components of the absolute probability, the component $P(A_1) = [0.20; 0.35]$ according to Example 6 has to be multiplied — which is executed only for the events produced by observation B_1:

$$\begin{array}{ll} L(A_1 \cap B_1 \cap Z_1) = 0.42 \cdot 0.20 = 0.084 & U(A_1 \cap B_1 \cap Z_1) = 0.54 \cdot 0.35 = 0.189 \\ L(A_1 \cap B_1 \cap Z_2) = 0.07 \cdot 0.20 = 0.014 & U(A_1 \cap B_1 \cap Z_2) = 0.184 \cdot 0.35 = 0.064. \end{array}$$

If, on the other hand, the mutually independence trials were combined, the components of $A_1 \cap B_1$ would be:

$$\begin{array}{rcccl} P(A_1 \cap B_1 \mid Z_1) & = & [0.6 \cdot 0.7; 0.7 \cdot 0.9] & = & [0.42; 0.63] \\ P(A_1 \cap B_1 \mid Z_2) & = & [0.1 \cdot 0.2; 0.2 \cdot 0.4] & = & [0.02; 0.08]. \end{array}$$

With respect to the marginal probability $P(Z_1) = [0.2; 0.3]$, $P(Z_2) = [0.7; 0.8]$ the outcome of the combined trial is partially described by the components

$$\begin{array}{rcccl} P(A_1 \cap B_1 \cap Z_1) & = & [0.42 \cdot 0.2; 0.63 \cdot 0.3] & = & [0.084; 0.189] \\ P(A_1 \cap B_1 \cap Z_2) & = & [0.02 \cdot 0.7; 0.08 \cdot 0.8] & = & [0.014; 0.064] \end{array}$$

demonstrating the conformity of the two procedures with regard to probability of the observations. □

It should be noted that the procedure described in Proposition 1 and demonstrated in Example 7 is not a mere transfer of the procedures customary in classical theory. The posterior probability resulting from the first trial is conditional probability relative to the actual observation. Prior probability is always marginal probability, hence total probability, not a conditional one. Therefore total probability has to be reconstructed by means of the marginal component of the actual observation in the first trial. This step does not influence the result in classical theory — and is left out therefore — but it is inevitable for interval probability!

Concerning the *decision-theoretic approach* it has been shown recently ([1]) that with regard to the optimization of decisions in the general case of interval probability the Theorem of Bayes — at least as far as it employs the intuitive concept — does not render what its counterpart for classical probability renders: that the Bernoulli-optimal action with respect to the posterior probability generated by the actual observation produces the corresponding branch of the optimal decision function. Hence the so called ‘Main Theorem of Bayesian Decision Analysis’ does not hold for interval probability.

It can be demonstrated that in the general case of interval probability this phenomenon is inevitable — beyond all questions about the methodology of Bayes’ theorem. In classical theory the branch of a decision function attributed to a certain observation produces an expected gain not depending on the circumstances related to the other possible observations. Therefore this expectation can be compared directly with those of respective branches belonging to other — competing — decision functions, a task, which is achieved easily via the Theorem of Bayes.

In presence of ambiguity the situation is different: If the expected gain of a decision function is calculated, each of the partial sums generated by an observation can be influenced by circumstances which originally refer to any of the other possible observations.This is a rule of thumb for decision functions:

Classical probability — only the actual observation counts.

Interval probability — all possible observations count.

Example 8, related to Examples 6 and 7, shows: If two gain functions differ only for observation A_2, nevertheless the contribution of observation A_1 to the interval expectation of the total gain may be influenced by this difference.

Example 8 *Z_1, Z_2 are two states of nature and A_1, A_2 are two possible observations, where the marginal probability $P(Z_1)$, $P(Z_2)$ and the canonical conditional probabilities $P(A_1 \mid Z_1)$, $P(A_2 \mid Z_1)$, $P(A_1 \mid Z_2)$, $P(A_2 \mid Z_2)$ are given in Example 1. Remember, that for K-functions of the respective structures*

$$p(Z_1) = a, \quad p(A_1 \mid Z_1) = b, \quad p(A_1 \mid Z_2) = c$$

the interval limits are given by

$$0.2 \leq a \leq 0.3, \quad 0.6 \leq b \leq 0.7, \quad 0.1 \leq c \leq 0.2.$$

The structure of the resulting F-field then consists of K-functions with the components given by

$p(A_1 \cap Z_1) = a \cdot b$	$p(A_2 \cap Z_1) = a \cdot (1-b)$	$p(Z_1) = a$
$p(A_1 \cap Z_2) = (1-a) \cdot c$	$p(A_2 \cap Z_2) = (1-a) \cdot (1-c)$	$p(Z_2) = 1-a$
$p(A_1) = a \cdot b + (1-a) \cdot c$	$p(A_2) = a \cdot (1-b) + (1-a) \cdot (1-c)$	$p(\Omega_4) = [1]$

producing the interval limits of these components as

$P(A_1 \cap Z_1) = [0.12;\, 0.21]$	$P(A_2 \cap Z_1) = [0.06;\, 0.12]$	$P(Z_1) = [0.2;\, 0.3]$
$P(A_1 \cap Z_2) = [0.07;\, 0.16]$	$P(A_2 \cap Z_2) = [0.56;\, 0.72]$	$P(Z_2) = [0.7;\, 0.8]$
$P(A_1) = [0.20;\, 0.35]$	$P(A_2) = [0.65;\, 0.80]$	$P(\Omega_4) = [1]$

A first decision function $D_1(.)$ is characterized by the following gains

$$\begin{array}{ll} D_1(A_1 \cap Z_1) = 4 & \qquad D_1(A_2 \cap Z_1) = 6 \\ D_1(A_1 \cap Z_2) = 8 & \qquad D_1(A_2 \cap Z_2) = 2. \end{array}$$

The expected gain $e(D_1(.))$ for a K-function described by a, b and c is given as

$$e(D_1(.)) = 4ab + 8(1-a)c + 6a(1-b) + 2(1-a)(1-c) = 2 + 2a(2-b-3c) + 6c.$$

Since $e(D_1(.))$ is minimal for $a = 0.2$, $b = 0.7$, $c = 0.1$ and $\mathbb{E}(D_1(.)) = [3.0; 3.68]$. *For every K-function, $e(D_1(.))$ can be divided into the two branches: $e(D_1(.)) = e(D_1(.) \cap A_1) + e(D_1(.) \cap A_2)$ where*

$$e(D_1(.) \cap A_1) = 4ab + 8(1-a)c, \; e(D_1(.) \cap A_2) = 6a(1-b) + 2(1-a)(1-c).$$

With respect to the roles of the two branches in determining $e(D_1(.))$ they have to be evaluated in the same way as $e(D_1(.))$ itself, i.e., using $a = 0.2$, $b = 0.7$, $c = 0.1$ to produce the two parts of $\mathbb{L}(D_1(.)) = 3.0$: $\mathbb{L}^*(D_1(.) \cap A_1) = 1.20$, $\mathbb{L}^*(D_1(.) \cap A_2) = 1.80$, and $a = 0.3$, $b = 0.6$, $c = 0.2$ to produce the respective parts of $\mathbb{U}(D_1(.)) = 3.68$: $\mathbb{U}^*(D_1(.) \cap A_1) = 1.84$, $\mathbb{U}^*(D_1(.) \cap A_2) = 1.84$. As far as comparisons with other decision functions are concerned, the branch of $D_1(.)$ determined by the observation A_1 therefore is represented by $[1.20; 1.84]$. *Now let a second decision function $D_2(.)$ be given by*

$$\begin{array}{ll} D_2(A_1 \cap Z_1) = 4 & \quad D_2(A_2 \cap Z_1) = 3 \\ D_2(A_1 \cap Z_2) = 8 & \quad D_2(A_2 \cap Z_2) = 2. \end{array}$$

This leads to

$$e(D_2(.)) = 4ab + 8(1-a)c + 3a(1-b) + 2(1-a)(1-c) = 2 + a(1+b-6c) + 6c$$

and this is minimal for $a = 0.2$, $b = 0.6$, $c = 0.1$ and maximal for $a = 0.3$, $b = 0.7$, $c = 0.2$, producing $\mathbb{E}(D_2(.)) = [2.8; 3.35]$. *If this interval expectation is divided into the two branches generated by the observation of A_1 and A_2, one arrives at*

$$e(D_2(.) \cap A_1) = 4ab + 8(1-a)c, \quad e(D_2(.) \cap A_2) = 3a(1-b) + 2(1-a)(1-c)$$

together with the results for $a = 0.2$, $b = 0.6$, $c = 0.1$: $\mathbb{L}^*(D_2(.) \cap A_1) = 1.12$, $\mathbb{L}^*(D_2(.) \cap A_2) = 1.68$, and for $a = 0.3$, $b = 0.7$, $c = 0.2$: $\mathbb{U}^*(D_2(.) \cap A_1) = 1.96$, $\mathbb{U}^*(D_2(.) \cap A_2) = 1.39$.

There are two striking findings:

1. $\mathbb{U}^*(D_2(.) \cap A_2) < \mathbb{L}^*(D_2(.) \cap A_2)$. *Obviously $\mathbb{L}^*(D_2(.) \cap A_2)$ and $\mathbb{U}^*(D_2(.) \cap A_2)$ may not be confounded with the lower and upper interval limits for the expectation of $D_2(.) \cap A_2$, which can be calculated as* $\mathbb{L}(D_2(.) \cap A_2) = 1.39$ *(produced by $a = 0.3$, $b = 0.7$, $c = 0.2$) and* $\mathbb{U}(D_2(.) \cap A_2) = 1.68$ *(produced by $a = 0.2$, $b = 0.6$, $c = 0.1$). In the case of decision function $D_2(.)$ therefore that constellation of K-functions, which leads to the maximal $e(D_2(.))$, results in the smallest possible value of $e(D_2(.) \cap A_2)$, and that constellation, which minimizes $e(D_2(.))$, happens to maximize the value of $e(D_2(.) \cap A_2)$.*

2. $\mathbb{L}^*(D_2(.) \cap A_1) \neq \mathbb{L}^*(D_1(.) \cap A_1)$ *and* $\mathbb{U}^*(D_2(.) \cap A_1) \neq \mathbb{U}^*(D_1(.) \cap A_1)$. *Both interval limits describing the contribution of branch* A_1 *to the expected gain are different for decision function* $D_1(.)$ *and decision function* $D_2(.)$ *— although all of the data describing branch* A_1 *are equal for both decision functions. The differences between the contributions of branch* A_1 *are caused by differences concerning the gains in case of observation* A_2. □

This phenomenon demonstrates the impossibility of qualifying the contribution of the branch attributed to the actual observation only by the circumstances of this observation without consideration of data related to other possible observations. In interval probability a decision function can only be judged or compared with others as a whole — not piecewise for each branch separately. Any kind of Theorem of Bayes, however, bases its calculation of the posterior probability only upon the circumstances of the actual observation — irrespective of the circumstances relating to other observations. Therefore no *posterior probability contains enough information* to qualify a branch of a decision function in comparison with the corresponding branches of competing decision functions.

The situation is different, if the problem considered is characterized by a very special type of gain function: Gains different from zero are supposed to be possible only if the actual observation is A_1. Therefore decision functions $D(. \cap .)$ are admissible for competition only if satisfying the requirements $D(A_i \cap Z_j) = 0,\ \forall i \neq 1, \forall j$. In this case the expected total gain and the expected gain for the branch A_1 are identical for every K-function: $e(D(.)) = e(D(.) \cap A_1)$. Consequently the following relations hold: $\mathbb{L}(D(.) \cap A_1) = \mathbb{L}(D(.))$ and $\mathbb{U}(D(.) \cap A_1) = \mathbb{U}(D(.))$. While at first this assumption seems to be very unrealistic, its systematic application to every actual observation A_i — instead of A_1 — generates a strategy which obviously is suboptimal in the general case, but may be understood as a kind of approximation to the optimal strategy: For each actual observation A_i that action $D(.) \cap A_i$ is chosen, which is best w.r.t. $[\mathbb{L}(D(.) \cap A_i); \mathbb{U}(D(.) \cap A_i)]$, irrespective of all observations which could have been made and the gains which would have been possible, if this observations had occured.

This strategy is much simpler than that founded on the complete decision function. It is an imitation of the proceeding in classical probability. In Example 9 it is demonstrated using the data of Example 8.

Example 9 *In the case of observation* A_1 *for the branch* $D_1(A_1 \cap Z_1) = D_2(A_1 \cap Z_1) = 4$, $D_1(A_1 \cap Z_1) = D_2(A_1 \cap Z_2) = 8$ *the decisive interval-expectation is given by*

$$\begin{aligned}\mathbb{L}(D_1(.) \cap A_1) = \mathbb{L}(D_2(.) \cap A_1) &= 1.12 \quad (a = 0.2, b = 0.6, c = 0.1)\\ \mathbb{U}(D_1(.) \cap A_1) = \mathbb{U}(D_2(.) \cap A_1) &= 1.96 \quad (a = 0.3, b = 0.7, c = 0.2).\end{aligned}$$

In case of observation A_2*: For* $D_1(A_2 \cap Z_1) = 6$, $D_1(A_2 \cap Z_2) = 2$ *one arrives at*

$$\begin{aligned}\mathbb{L}(D_1(.) \cap A_2) &= 1.64 \quad (a = 0.2, b = 0.7, c = 0.2)\\ \mathbb{U}(D_1(.) \cap A_2) &= 1.98 \quad (a = 0.3, b = 0.6, c = 0.1),\end{aligned}$$

for $D_2(A_2 \cap Z_1) = 3$, $D_2(A_2 \cap Z_2) = 2$*:*

$$\begin{aligned} \mathbb{L}(D_2(.) \cap A_2) &= 1.39 \quad (a = 0.3, b = 0.7, c = 0.2) \\ \mathbb{U}(D_2(.) \cap A_2) &= 1.68 \quad (a = 0.2, b = 0.6, c = 0.1). \end{aligned} \qquad \square$$

Two remarks are useful:

Expectations belonging to different observations are based on contradictory assumptions. Therefore they are not suitable for being combined.

Comparison of actions which are characterized by means of interval expectation depends on the attitude of the decision-maker towards ambiguity. It may be described by the choice of $\eta L(G) + (1 - \eta)U(G)$, $0 \leq \eta \leq 1$, as the decisive quantity. Since it can be understood, that the larger value of gain G always is preferred, η is interpreted as a measure of caution.

Because of the goal of this section it is asked whether a posterior probability generated by the Theorem of Bayes can be employed in calculating the expectation $[\mathbb{L}(D(.) \cap A_1); \mathbb{U}(D(.) \cap A_1)]$ produced by the actual observation A_1.

Using again the data of Example 9 it will be demonstrated in Example 10 that with respect to that type of Theorem of Bayes, which employs the intuitive concept of conditional probability, the answer to this question must be negative.

Example 10 *The intuitive conditional probability* $iP(Z_1 \mid A_1)$, $iP(Z_2 \mid A_1)$ *obviously is determined by* $iL(Z_1 \mid A_1) = \min\limits_{\mathcal{M}} \frac{ab}{ab+(1-a)c}$, $iU(Z_1 \mid A_1) = \max\limits_{\mathcal{M}} \frac{ab}{ab+(1-a)c}$ *with*

$$\mathcal{M} = \{p_{a,b,c}(.); 0.2 \leq a \leq 0.3; 0.6 \leq b \leq 0.7; 0.1 \leq c \leq 0.2\}.$$

It is easily seen, that the minimum is produced by $a = 0.2$, $b = 0.6$, $c = 0.2$ *and the maximum by* $a = 0.3$, $b = 0.7$, $c = 0.1$. *The resulting i-conditional F-probability field is given by* $iP(Z_1 \mid A_1) = [0.429; 0.750]$, $iP(Z_2 \mid A_1) = [0.250; 0.571]$. *The i-conditional expectation of the gain function produced by the decision function* $D(.) = D_1(.) = D_2(.)$ *with* $D(A_1 \cap Z_1) = 4$, $D(A_1 \cap Z_2) = 8$ *is determined by*

$$\begin{aligned} i\mathbb{L}(D(.) \mid A_1) &= 0.750 \cdot 4 + 0.250 \cdot 8 = 5 \\ i\mathbb{U}(D(.) \mid A_1) &= 0.429 \cdot 4 + 0.571 \cdot 8 = 7.429. \end{aligned}$$

To achieve the interval-expectation of $D(.) \cap A_1$, *conditional expectation must be combined with the corresponding component of marginal probability:* $P(A_1) = [0.20; 0.35]$. *Therefore* $i\mathbb{L}(D(.) \cap A_1) = 5 \cdot 0.20$, $i\mathbb{U}(D(.) \cap A_1) = 7.429 \cdot 0.35$ *and* $i\mathbb{E}(D(.) \cap A_1) = [1.00; 2.60]$ *instead of the true interval expectation, as calculated in Example 9:* $\mathbb{E}(D(.) \cap A_1) = [1.12; 1.96]$. *Like in other situations, employment of the intuitive concept generates a loss in sharpness of the result.* □

If, however, the canonical concept is applied, the conditional expectation of the gain produced by the decision function $D(.)$ for the observation A_1, due to the definition described in Section 2, reads as $\mathbb{E}(D(.) \mid A_1) = \left[\frac{\mathbb{L}(D(.) \cap A_1)}{L(A_1)}; \frac{\mathbb{U}(D(.) \cap A_1)}{U(A_1)}\right]$.

Combined with the component $[L(A_1); U(A_1)]$ of the marginal probability this conditional expectation produces $\mathbb{E}(D(.) \cap A_1) = [\mathbb{L}(D(.) \cap A_1); \mathbb{U}(D(.) \cap A_1)]$, due to the simplified optimal strategy, as it was described above (Example 11).

It is, therefore, justified to use the designation 'Interval Bayes-Strategy' for the method of selecting in case of observation A_1 that action which produces the largest expected gain — judged by means of the individual caution — with respect to the posterior probability generated by the Theorem of Bayes with the canonical concept of conditional probability.

Example 11 *Because of* $\mathbb{E}(D(.) \cap A_1) = [1.12; 1.96]$ *(Example 9) and* $P(A_1) = [0.20; 0.35]$ *(Example 8), the conditional expectation results as* $\mathbb{L}(D(.) \mid A_1) = \frac{1.12}{0.20}$, $\mathbb{U}(D(.) \mid A_1) = \frac{1.96}{0.35}$ *or* $\mathbb{E}(D(.) \mid A_1) = [5.60; 5.60]$. *The quality of the result is not affected by the fact, that this interval possesses length zero.* □

Hence, use of the canonical concept allows the Interval Bayes-Strategy, distinguished from the strategy based upon the optimal decision function only by neglecting any information concerning observations which did not occur. If the omission of such 'counterfactual information' is accepted on principle, the Interval Bayes-Strategy must be regarded as optimal.

7 Conclusions

This paper contributes to the question of defining conditional interval probability appropriately. A symbiosis of the intuitive and the canonical concept of conditional probability is proposed, resulting in recommendations which of the concepts should be used for what propose.

The results of Sections 3–6 can for short be interpreted to favour the employment of the intuitive concept in any situation where conditional probability is seen as a goal in itself, therefore in updating, whether it is achieved directly or by means of the Theorem of Bayes: the final result should be described by the intuitive concept of conditional probability.

The canonical concept proves to be superior always when conditional probability is used as a tool for further analysis. This applies to the transfer of information from one model to another and to the derivation of a posterior probability by means of the Theorem of Bayes, if this posterior is employed as prior for an independent trial, or as basis for decisions between possible actions. Additionally this concept produces a Theorem of Total Probability and the consistency with marginal probability in the case of independence as defined by strong extension.

While the intuitive concept guarantees that its outcome describing the final result of an analysis always can be interpreted as interval probability or as interval expectation, it is possible that the outcome of the canonical concept, which is employed as a tool for further calculations, does not possess the F- or even R-quality, resp., the quality of interval expectation, without loss of usefulness: an obvious analogy to the role of complex numbers in algebra.

References

[1] T. Augustin. On the suboptimality of the Generalized Bayes Rule and robust Bayesian procedures from the decision theoretic point of view — a cautionary note on updating imprecise priors. `www.stat.uni-muenchen.de/~thomas/robust-bayes-suboptimal.pdf`, 2003.

[2] T. Augustin and F. Coolen. Nonparametric predictive inference and interval probability. To appear in: *Journal of Statistical Planning and Inference*, 2003.

[3] J.M. Bernard. Non-parametric inference about an unknown mean using the imprecise Dirichlet model. *In:* [9], 40–50.

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

[5] F.G. Cozman. Computing posterior upper expectations. *In:* G. de Cooman, F.G. Cozman, S. Moral, and P. Walley (eds.), *ISIPTA ’99: Proceedings of the First International Symposium on Imprecise Probabilities and their Applications*. University of Ghent, 131–140, 1999.

[6] F.G. Cozman. Constructing sets of probability measures through Kuznetsov’s independence condition. *In:* [9], 104–111.

[7] F. Cozman and P. Walley. Graphoid properties of epistemic irrelevance and independence. *In:* [9], 112–121.

[8] G. de Cooman. Integration and conditioning in numerical possibility theory. *Annals of Mathematics and Artificial Intelligence*, 32: 87–123, 2001.

[9] G. de Cooman, T. Fine, S. Moral, and T. Seidenfeld (eds.). *ISIPTA01: Proceedings of the Second International Symposium on Imprecise Probabilities and their Applications*. Cornell University, Ithaca (N.Y.). Shaker, Maastricht, 2001.

[10] A.P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38: 325–339, 1967.

[11] D. Dubois and H. Prade. Focusing versus updating in belief function theory. *In:* R.R. Yager, M. Fedrizzi, and J. Kacprzyk (eds.), *Advances in the Dempster-Shafer Theory of Evidence*. Wiley, New York, 71–95, 1994.

[12] J.Y. Halpern and R. Fagin. Two views of belief: belief as generalized probability and belief as evidence. *Artificial Intelligence*, 54: 275–317, 1992.

[13] T. Fetz. *Sets of joint probability measures generated by random sets*. Doctoral Thesis, University of Innsbruck, 2002.

[14] V.P. Kuznetsov. *Interval Statistical Methods* (in Russian). Radio i Svyaz Publ., Moscow, 1991.

[15] M. Lavine. Sensitivity in Bayesian statistics, the prior and the likelihood. *Journal of the American Statistical Association*, 86 (414): 396–399, 1991.

[16] E. Miranda and G. de Cooman. Independent products of numerical possibility measures. *In:* [9], 237–246.

[17] S. Moral. Epistemic irrelevance on sets of desirable gambles. *In:* [9], 247–254.

[18] D. Rios Insua and F. Ruggeri (eds.). *Robust Bayesian Analysis*. Lecture Notes in Statistics 152, Springer, New York, 2000.

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

[20] B. Vantaggi. Graphical models for conditional independence structures. *In:* [9], 332–341.

[21] P. Vicig. Epistemic independence for imprecise probabilities. *International Journal of Approximate Reasoning*, 24: 235–250, 2000.

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

[23] P. Walley. Inferences from multinomial data: Learning from a bag of marbles (with discussion). *Journal of the Royal Statistical Society*, B 58: 3–57, 1996.

[24] L. Wasserman. Bayesian robustness. *In:* S. Kotz, C.B. Read, and D.L. Banks (eds.), *Encyclopedia of Statistical Sciences, Update Volume 1*. Wiley, New York, 45–51, 1997.

[25] K. Weichselberger. Axiomatic foundations of the theory of interval-probability. *In:* V. Mammitzsch and H. Schneeweiß (eds.), *Symposia Gaussiana, Proceedings of the 2nd Gauss-Symposium, Conference B*. De Gruyter, Berlin, 47–64, 1995.

[26] K. Weichselberger. The theory of interval probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24: 149–170, 2000.

[27] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I — Intervallwahrscheinlichkeit als umfassendes Konzept*, in cooperation with T. Augustin and A. Wallner. Physica, Heidelberg, 2001.

[28] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung II — Die Theorie von Intervallwahrscheinlichkeit*. In preparation.

[29] N. Wilson. Modified upper and lower probabilities based on imprecise likelihoods. *In:* [9],

[30] C. Yu and F. Arasta. On conditional belief functions. *International Journal of Approximate Reasoning*, 10: 155–172, 1994.

[31] M. Zaffalon. Statistical inference of the naive credal classifier. *In:* [9], 384–393.

K. Weichselberger and T. Augustin are with the Department of Statistics, Ludwig-Maximilians-University Munich, Ludwigstr. 33, 80539 München, Germany. E-mail: {weichsel, augustin}@stat.uni-muenchen.de

Author Index

Keyword Index

Page numbers refer to the first page of a paper containing the keyword or keyword phrase listed here.

Proceedings in Informatics

Carleton Scientific publishes the series *Proceedings in Informatics* as a contribution to the rapid dissemination of high quality results in all aspects and all areas of Informatics, including Computing, Communication, Discrete Mathematics, Algorithms, Complexity, and Networking.

It has as its mandate the publishing of proceedings of conferences, colloquia, workshops and symposia that focus on research themes of scientific importance and relevance. In particular, reflecting the continuous changes in the domain, it actively pursues those meetings devoted to new research topics, emerging areas of scientific investigation, and mutation and integration of existing fields.

The goal of this series is to provide an invaluable working tool for researchers active in (or entering) the specific area covered by each volume, for scientists who want to discover the current directions in Informatics, and for application designers and software engineers who want to keep up to date with the latest results and problems being investigated by the research community.

To achieve this goal and fulfill its mandate, Carleton Scientific has established an international editorial board of scholars working in the area of Informatics.

If you are interested in publishing the proceedings of your meeting in this series, please contact any of the members of the Editorial Board. For any other information contact: editor@carleton-scientific.com