*Proceedings of the*

# 9th WORKSHOP ON

# UNCERTAINTY PROCESSING

**Mariánské Lázně**

**12–15 September 2012**

## Programme Committee Board

**Nihat Ay**, *Germany*
**Gernot D. Kleiter**, *Austria*
**Vilém Novák**, *Czech Republic*
**Romano Scozzafava**, *Italy*
**Prakash P. Shenoy**, *U.S.A.*
**Milan Studený** (chair), *Czech Republic*

## Organizing Committee Board

**Vladislav Bína**
**Radim Jiroušek** (chair)
**Václav Kratochvíl**
**Milan Studený**
**Lucie Váchová**
**Jiří Vomlel**

## Proceedings Editors

**Tomáš Kroupa**
**Jiřina Vejnarová**

**Cover design: CANDY BOX s.r.o.**

**ISBN 978-80-245-1885-5**

# 9$^{\text{th}}$ WORKSHOP ON UNCERTAINTY PROCESSING

organized by

**Faculty of Management**
*University of Economics*

**Institute of Information Theory and Automation**
*Academy of Sciences of the Czech Republic*

**Centre of Excellence IT4Innovations**
*Division of the University of Ostrava, IRAFM*

under the auspices of

**Zdeněk Král**
*The Mayor of Mariánské Lázně*

**Mariánské Lázně, 12–15 September 2012**

# Foreword

It's my pleasure to welcome you to Mariánské Lázně on the occasion of

the 9th *Workshop on Uncertainty Processing* (*WUPES 2012*),

to be held from September 12 to September 15, 2012. This traditional international workshop, organized in the Czech Republic every three years since 1988, is devoted to the (mathematical) methods for representing, managing and exploiting uncertain knowledge for (computer-aided) decision making. There are other conferences devoted to this topic, but this workshop is special in some aspects. First, its aim is to foster creative intellectual activity and exchange the ideas in an informal atmosphere. For this reason, the tradition is to limit the number of participants. Second, the workshop is typically held in some (small) quite place, so that the participants are not disturbed in their discussion and, also have a chance to enjoy the beauty of the country. Third, there is a group of traditional participants and special topics to be discussed at the workshop.

This booklet consists of the papers accepted to WUPES 2012. The contributions were chosen by the Programme Committee on the basis of two-page abstracts. The final submissions were then processed by the editors of the Proceedings. Because the proceedings papers are considered to be preliminary versions of future journal papers, they have not been reviewed. A carefully selected subset of proceedings papers is tratiditionally published in a special journal issue after the regular review procedure.

The proceedings of WUPES 2012 contain 23 contributions and the organizers expect about 30 participants. Besides traditional themes, like the coherence theory, Bayesian networks, the possibility theory, belief functions and conditional independence, further topics emerged; namely, fuzzy logic, the entropy and algebraic (methods in) statistics.

The venue of the workshop, Mariánské Lázně, also known under German name *Marienbad*, is the second biggest spa town in the Czech Republic. It is placed in the western part of the Czech Republic and surrounded by green hills. There is about 100 springs of mineral water in the neighborhood of this town, built about two hundred years ago. I hope that the participants of WUPES 2012 will enjoy this mosaic of parks and yellow-and-white houses. The organizers of the workshop are indebted to the town of Mariánské Lázně for providing the lecture hall and for the help with the local organization.

Let me conclude by expressing my thanks to all my colleagues and friends for their commitment to prepare this event: the members of the Organizing Committee, the members of the Programme Committee, the editors of the Proceedings, the local people in the venue, and to the sponsoring organizations.
I wish you a nice stay in Mariánské Lázně.

Milan Studený

# Table of Contents

# The Irrelevant Information Principle for Collective Probabilistic Reasoning

**Martin Adamčík and George Wilmers**
School of Mathematics
The University of Manchester
martin.adamcik@manchester.ac.uk, george.wilmers@gmail.com

**Abstract**

Within the framework of discrete probabilistic uncertain reasoning a large literature exists justifying the maximum entropy inference process, **ME**, as being optimal in the context of a single agent whose subjective probabilistic knowledge base is consistent. In [9] Paris and Vencovská, extending the work of Johnson and Shore [6], completely characterised the **ME** inference process by an attractive set of axioms which an inference process should satisfy, thus providing a quite different justification for **ME** from that of the more traditional possible worlds or information theoretic arguments whose origins go back to nineteenth century statistical mechanics as in [8] or [5].

More recently the second author in [10] and [11] extended the Paris-Vencovská axiomatic approach to inference processes to the context of several agents whose subjective probabilistic knowledge bases, while individually consistent, may be collectively inconsistent. In particular he defines a "social entropy process", **SEP**, which is a natural extension of the single agent **ME**. However, while **SEP** is known to possess many attractive properties, these are almost certainly insufficient to uniquely characterise **SEP**. It is therefore of particular interest to study those Paris-Vencovská principles valid for **ME** whose immediate generalisations to the multiagent case are not satisfied by **SEP**. One of these principles is the Irrelevant Information Principle, a principle which very few inference processes satisfy even in single agent context. In this paper we will investigate whether **SEP** can satisfy an interesting modified generalisation of this principle.

## 1 Motivation

In this paper we consider the following fundamental problem of discrete multi-agent probabilistic uncertain reasoning. We are interested in finding a general procedure which, given a finite set of agents, each possessing a subjective probabilistic knowledge base over a finite space of possible events, yields a single probability function or *social probability function* defined over that space of events, which optimally represents the joint knowledge of all the agents, and such that that general procedure satisfies some natural criteria derived from logical or rational considerations.

There are several initial assumptions we want to make. Firstly we assume that the probabilistic knowledge of each *particular* expert is consistent with the laws of probability. Secondly all agents are assumed to have equal status, and the final social probability function should not depend on the order in which the agents' knowledge bases are considered.

We illustrate the motivation behind this idea by a toy two-agent example.

*Imagine that two safety experts are dealing with a fault in a chemical factory producing nitrogen fertilizers. There is a problem with ammonia supply. Ammonia is stored in a tank connected to the rest of the factory by a valve which is operated by an electric circuit.*

*The first expert believes that there is a* 40% *chance of a mechanical fault on the valve. The second expert comes up with a different opinion that there is a* 80% *chance that there is a mechanical problem on that valve. Moreover, the first safety expert thinks that there is a* 70% *chance that there is a malfunction of the electric circuit. We suppose that both experts have no other knowledge related to this problem.*

The joint beliefs (knowledge) of the two experts are inconsistent in this case. In practice, knowledge is usually incomplete and offers a lot of uncertainty; the first expert in above example has no knowledge about, for instance, the conditional probability that there is a fault on the the valve given that there is a fault on the electric circuit. The situation becomes more complicated once the second agent is considered whose knowledge is inconsistent with the knowledge of the first agent. Altogether we can ask the following question:

**Question.** *How should a rational adjudicator whose only knowledge consists of what is related to him by the two experts above, evaluate the probability that both the valve and the electric circuit are faulty, based only on the experts' subjective knowledge specified above and without any other assumptions?*

Assuming, as we do in this paper, that each agent's uncertain knowledge can be represented within the framework of probability theory, we can describe the knowledge of each expert by a set of possible probability distributions over four possible mutually exclusive cases: (1) a fault on the valve and no fault on the electric circuit, (2) a fault on the valve and a fault on the electric circuit, (3) no fault on valve and a fault on the electric circuit and (4) no faults on the valve or on the electric circuit (i.e. in this case there is a problem with something else). We can denote the corresponding probabilities that (1),(2),(3) and (4) is true by real numbers $w_1$, $w_2$, $w_3$ and $w_4$ from the interval $[0,1]$ which sum to 1. Based on the knowledge of the first expert $w_1 + w_2 = 0.4$ and $w_1 + w_3 = 0.7$. Any probability function $(x, 0.4 - x, 0.7 - x, x - 0.1)$, where $x \in [0.1, 0.4]$, is consistent with the knowledge of the first expert. Similarly, the second expert admits any $(x, 0.8-x, y, 0.2-y)$ where $x \in [0, 0.8]$ and $y \in [0, 0.2]$. This representation of the knowledge of the experts naturally abstracts from the complex nature of the actual problem. However we are not interested here in the particular manner in which this abstraction from the infinite complexity of a real world problem has been accomplished. Instead we will focus on the following narrower, abstract, but more clearly defined question:

**Question.** *Given two (or more) sets of probability functions corresponding to the knowledge bases of corresponding experts as in the above example, which single probability function best represents the combined probabilistic knowledge of the experts?*

Naturally, we would like to find a general procedure doing this for any knowledge bases which satisfies some natural principles. We will formalize this idea in a general setting in the next section.

# 2 Formalization

Let $L = \{a_1 \dots a_h\}$ be a finite propositional language where $a_1, \dots, a_h$ are propositional variables. In our example $n = 2$, $a_1$ stands for sentence "a fault on the valve" and $a_2$ stands for sentence "a fault on the electric circuit". By the disjunctive normal form theorem any $L$-sentence can by expressed as a disjunction of atomic sentences (atoms) and we will denote a maximal set of logically inequivalent atoms $\{\alpha_1, \dots, \alpha_J\}$, where $J = 2^h$, by $\mathrm{At}(L)$. The atoms of $\mathrm{At}(L)$ are thus mutually exclusive and exhaustive.

A probability function $\mathbf{w}$ over $L$ is defined by a function $\mathbf{w} : \mathrm{At}(L) \to [0, 1]$ such that $\sum_{j=1}^{J} \mathbf{w}(\alpha_j) = 1$. A value of $\mathbf{w}$ on any $L$-sentence $\varphi$ may then be defined by setting

$$\mathbf{w}(\varphi) = \sum_{\alpha_j \models \varphi} \mathbf{w}(\alpha_j).$$

We will denote the set of all probability functions over $L$ by $\mathbb{D}^L$. For the sake of simplicity we will often write $w_j$ instead of $\mathbf{w}(\alpha_j)$, but note this has a sense only for atomic sentences. Given a probability function $\mathbf{w} \in \mathbb{D}^L$, a conditional probability is defined by Bayes's formula

$$\mathbf{w}(\varphi|\psi) = \frac{\mathbf{w}(\varphi \wedge \psi)}{\mathbf{w}(\psi)}$$

for any $L$-sentence $\varphi$ and any $L$-sentence $\psi$ such that $\mathbf{w}(\psi) \neq 0$ and is left undefined otherwise.

Now consider two distinct propositional languages $L_1 = \{a_1, \dots, a_{h_1}\}$ and $L_2 = \{b_1, \dots, b_{h_2}\}$. Let $\mathrm{At}(L_1) = \{\alpha_1, \dots, \alpha_J\}$ and $\mathrm{At}(L_2) = \{\beta_1, \dots, \beta_I\}$. Then every atom of the joint language $L_1 \cup L_2$ can be written uniquely (up to logical equivalence) as $\alpha_j \wedge \beta_i$ for precisely one $1 \leq j \leq J$ and precisely one $1 \leq i \leq I$. With only a slight abuse of notation, for an $L_1 \cup L_2$-probability function $\mathbf{r}$ we will often write $r_{ji}$ instead of $\mathbf{r}(\alpha_j \wedge \beta_i)$, in a similar way as for an $L_1$-probability function $\mathbf{v}$ we write $v_j$ instead of $\mathbf{v}(\alpha_j)$.

Now notice that $\models \alpha_j \leftrightarrow \bigvee_{i=1}^{I} \alpha_j \wedge \beta_i$. Therefore, the marginal probability function whose $j$-th value is given by $\sum_{i=1}^{I} r_{ji}$ is the projection of an $L_1 \cup L_2$-probability function $\mathbf{r}$ to the language $L_1$. We will denote it by $\mathbf{r}|_{L_1}$. Similarly if $\Delta$ is a set of $L_1 \cup L_2$-probability functions, we denote the set $\{v|_{L_1} \ : \ v \in \Delta\}$ by $\Delta|_{L_1}$. Also if $\mathbf{v}$ is an

$L_1$-probability function and $\mathbf{w}$ is an $L_2$-probability function then $\mathbf{v}\cdot\mathbf{w}$ defined by $\mathbf{v}\cdot\mathbf{w}(\alpha_j\wedge\beta_i)=v_jw_i$ is an $L_1\cup L_2$-probability function such that $(\mathbf{v}\cdot\mathbf{w})|_{L_1}=\mathbf{v}$.

A (probabilistic) *knowledge base* $\mathbf{K}$ over $L$ is a set of constraints on probability functions over $L$ such that the set of all probability functions satisfying the constraints in $\mathbf{K}$ forms a nonempty closed convex subset $V_{\mathbf{K}}$ of $\mathbb{D}^L$. $V_{\mathbf{K}}$ may be thought of as the set of possible probability functions of a particular agent which are consistent with her subjective probabilistic knowledge base $\mathbf{K}$. In the sequel we shall loosely identify $\mathbf{K}$ with $V_{\mathbf{K}}$, and may also refer to such a $V_{\mathbf{K}}$ as a knowledge base. Note that the non-emptiness of $V_{\mathbf{K}}$ corresponds to the assumption that $\mathbf{K}$ is consistent, while if $\mathbf{K}$ and $\mathbf{F}$ are knowledge bases then the knowledge base $\mathbf{K}\cup\mathbf{F}$ corresponds to $V_{\mathbf{K}\cup\mathbf{F}}=V_{\mathbf{K}}\cap V_{\mathbf{F}}$. The set of all knowledge bases $V_{\mathbf{K}}$ over $L$ is denoted by $CL$.

In the toy example, the knowledge of the first expert can be represented by the knowledge base $\mathbf{K}$ which consist of a set of linear constraints on a probability function $\mathbf{w}=(w_1,w_2,w_3,w_4)$ defined over the atomic sentences $a_1\wedge a_2$, $a_1\wedge\neg a_2$, $\neg a_1\wedge a_2$ and $\neg a_1\wedge\neg a_2$. Then $\mathbf{K}=\{w_1+w_2=0.4,\ w_1+w_3=0.7\}$ and $V_{\mathbf{K}}=\{(x,0.4-x,0.7-x,x-0.1):\,x\in[0.1,0.4]\}$.

Given $\mathbf{K}\in CL_1$ note that the underlying language $L_1$ is implicitly understood in the notation $V_{\mathbf{K}}$ which should more properly be denoted $V_{\mathbf{K}}^{L_1}$. Thus if $L_1\subset L$ then $\mathbf{K}$ is also in $CL$ and $V_{\mathbf{K}}^{L}=\{\mathbf{w}\in\mathbb{D}^L:\,\mathbf{w}|_{L_1}\in V_{\mathbf{K}}^{L_1}\}$. For simplicity we shall normally just write $V_{\mathbf{K}}$ when the appropriate language is understood.

We now define the central notion which maps any given sequence of knowledge bases to a single probability function termed the *social probability function* for that sequence. A *social inference process* $\mathcal{S}$ defines for each $L$ and $n\geq 1$ a function

$$\mathcal{S}_L:\underbrace{CL\times\ldots\times CL}_{n}\to\mathbb{D}^L.$$

The number $n$ here intuitively represents the number of distinct agents or distinct sources of information.

The restricted notion $\mathcal{S}$ (or $\mathcal{S}_L$) in the case of a single knowledge base or agent, i.e. when $n=1$, is simply called an *inference process* and such inference processes have been extensively studied by Paris, Vencovská and others ([6], [8], [9], [5] or [4]).

As was noted above, a consistent knowledge base $\mathbf{K}$ yields a set of possible probability functions $V_{\mathbf{K}}$ consistent with $\mathbf{K}$. In the case of single agent with knowledge base $\mathbf{K}$ there are several possible procedures to choose a specific probability function from $V_{\mathbf{K}}$. However by the traditional possible worlds modeling or information theoretic arguments whose origins go back to nineteenth century statistical mechanics as in [8] or [5], the maximum entropy inference process $\mathbf{ME}$ has been justified as being optimal, where $\mathbf{ME}_L(\mathbf{K})$ is defined as that unique probability function $\mathbf{w}$ in $V_{\mathbf{K}}$ which maximizes the Shannon entropy $E(\mathbf{w})$ of $\mathbf{w}$ given by

$$E(\mathbf{w})=-\sum_{j=1}^{J}\mathbf{w}(\alpha_j)\log\mathbf{w}(\alpha_j).$$

$E$ is a strictly concave function and therefore it attains a unique maximum over any nonempty closed convex region $V_{\mathbf{K}}$ of $\mathbb{D}^L$.

A quite different justification for **ME** to the traditional ones was described in [6] by Johnson and Shore. Their work was developed by Paris and Vencovská in [9] where they showed that a list of principles based on symmetry and consistency uniquely characterises **ME**. It seems fruitful to look at the axiomatic approach also in the more general context of a social inference process. Accordingly we may ask:

*What general principles should a social inference process $\mathcal{S}$ satisfy in order to ensure that for given knowledge bases, and in the absence of any other information, $\mathcal{S}$ chooses a social probability function according to rational criteria?*

We might hope that ultimately such a set of rational principles may determine uniquely a particular social inference process $\mathcal{S}$.

## 3 Language Invariance and Irrelevant Information

In this section we examine how certain fundamental invariance principles formulated by Paris and Vencovská for an inference process (see [7]) can be extended to the notion of a social inference process.

An obvious question we need to ask regarding social inference processes is whether they depend on the choice of a particular propositional language $L = \{a_1, \ldots, a_h\}$. For fixed $\mathcal{S}$, $L$, $\varphi \in SL$ and $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ consider $\mathcal{S}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)(\varphi)$. It would seem to be irrational to change this value if $L$ is extended by a set of propositional variables $\{b_1, \ldots, b_k\}$, all distinct from the variables of $L$, provided that we have not supplied any new knowledge. Following [7] we will formulate this as the following principle:

**LI [Language Invariance Principle].** *A social inference process $\mathcal{S}$ satisfies language invariance if whenever $L_1$ and $L_2$ are languages with $L_1 \subseteq L_2$ and $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL_1$, then*

$$\mathcal{S}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)(\varphi) = \mathcal{S}_{L_2}(\mathbf{K}_1, \ldots, \mathbf{K}_n)(\varphi)$$

*for any $L_1$-sentence $\varphi$.*

Following [7] we can also ask a different question. What will happen if alongside the new propositional variables, new knowledge concerning these variables is also provided which contains no reference to the old variables. Again, it would seem to be rational that the value of a social inference process on a sentence that is formulated in original language should not change. This leads us to

**IIP [The Irrelevant Information Principle].** *Let $L = L_1 \cup L_2$ where $L_1$ and $L_2$ are disjoint propositional languages, and let $\mathbf{K}_1, \ldots, \mathbf{K}_n$ and $\mathbf{F}_1, \ldots, \mathbf{F}_n$ be knowledge bases formulated for the languages $L_1$ and $L_2$ respectively. Then for any $L_1$-sentence $\varphi$*

$$\mathcal{S}_L(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)(\varphi) = \mathcal{S}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)(\varphi).$$

In the case when $n = 1$ this principle plays a crucial role in the characterisation of **ME** in [9]. Nevertheless, despite its intuitive plausibility this principle is in fact very hard to satisfy; indeed although **ME** satisfies this principle, almost all other commonly used (single agent) inference processes do not do so (see [7] and [4] for details).

**IIP** appears even harder for a social inference processes to satisfy. However, in this multi-agent case we might argue that this principle is just too strong. If knowledge provided by agents for the language $L_2$ is inconsistent then the addition of such new knowledge may provide us with more information on how strongly the agents disagree, which in turn may affect our evaluation of the knowledge concerning $L_1$. However, if the new knowledge does not change the level of disagreement as is the case when the new knowledge of all the agents is jointly consistent, then the principle of irrelevant information is arguably more justified. Accordingly we formulate:

**CIIP [The Consistent Irrelevant Information Principle].** *Let* $L = L_1 \cup L_2$ *where* $L_1$ *and* $L_2$ *are disjoint propositional languages. Let* $\mathbf{K}_1, \ldots, \mathbf{K}_n$ *and* $\mathbf{F}_1, \ldots, \mathbf{F}_n$ *be knowledge bases formulated for the languages* $L_1$ *and* $L_2$ *respectively, and suppose that* $\mathbf{F}_1, \ldots, \mathbf{F}_n$ *are jointly consistent. Then for any* $L_1$*-sentence* $\varphi$

$$\mathcal{S}_L(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)(\varphi) = \mathcal{S}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)(\varphi).$$

Assuming **LI** this last equation is equivalent to

$$\mathcal{S}_L(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)(\varphi) = \mathcal{S}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)(\varphi).$$

For instance, in the toy example of section 1 the information of both experts about a fault on the electric circuit is both consistent and *a priori* irrelevant to the probability that there is a fault on the valve. Hence if we want to know only the probability that there is a fault on the valve, then applying the **CIIP** we need consider only the fact that the first expert states that this probability is 40% and the second states that this probability is 80%.

# 4 The Social Entropy Process

In this section we define a particular social inference process formulated by the second author in [10] and [11]. The Social Entropy Process **SEP**, is defined by the following two stage process. At the *first stage* we define the set $\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ as those probability functions $\mathbf{v}$ which globally minimise the sum of Kullback-Leibler divergences (cross-entropies)

$$\sum_{i=1}^{n} \mathrm{CE}(\mathbf{v}, \mathbf{w}^{(i)}) = \sum_{k=1}^{n} \sum_{j=1}^{J} v_j \log \frac{v_j}{w_j^{(k)}} \tag{1}$$

subject only to the conditions that $\mathbf{w}^{(1)} \in V_{\mathbf{K}_1}, \ldots, \mathbf{w}^{(n)} \in V_{\mathbf{K}_n}$, where

$$v_j \log \frac{v_j}{w_j^{(k)}} = \begin{cases} 0 & \text{if } v_j = 0 \text{ and } w_j^{(k)} = 0, \\ \infty & \text{if } v_j \neq 0 \text{ and } w_j^{(k)} = 0. \end{cases}$$

Recall that $v_j$ and $w_j^{(k)}$ stand for $\mathbf{v}(\alpha_j)$ and $\mathbf{w}^{(i)}(\alpha_j)$ respectively, where $\alpha_j$ is an atom and there are $J$ (logically inequivalent) atoms in $\mathrm{At}(L)$.

It is not difficult to see (see [11]) that $\Delta_L(\mathbf{K}_1,\dots,\mathbf{K}_n)$ is nonempty if *there is some atom $\alpha_j$ such that for no $i$ is it the case that for all $\mathbf{w}\in V_{\mathbf{K}_i}$ $\mathbf{w}(\alpha_j)=0$.* Under this condition $\Delta_L(\mathbf{K}_1,\dots,\mathbf{K}_n)$ is well-defined. From now on we shall consider only $n$-tuples of knowledge bases $\mathbf{K}_1,\dots,\mathbf{K}_n$ which satisfy this condition. Note that the definition of a social inference process is not much restricted by such an assumption.

In [11] it is proved that $\Delta_L(\mathbf{K}_1,\dots,\mathbf{K}_n)$ is also a closed convex region of $\mathbb{D}^L$ and therefore there is a unique probability function in $\Delta_L(\mathbf{K}_1,\dots,\mathbf{K}_n)$ having maximal entropy, and we will denote this function by $\mathbf{ME}_L(\Delta_L(\mathbf{K}_1,\dots,\mathbf{K}_n))$. Therefore, at the *second stage* of the definition we set $\mathbf{SEP}_L(\mathbf{K}_1,\dots,\mathbf{K}_n)=\mathbf{ME}_L(\Delta_L(\mathbf{K}_1,\dots,\mathbf{K}_n))$. It is clear that $\mathbf{SEP}_L$ coincides with $\mathbf{ME}_L$ in the case when $n=1$ and, it is straightforward to show that **SEP** is language invariant.

The set $\Delta_L(\mathbf{K}_1,\dots,\mathbf{K}_n)$ is often a singleton and in that case the second stage is essentially redundant. For instance, this happens whenever $V_{\mathbf{K}_k}$ is a singleton for some $k$. The function which maps $\mathbf{K}_1,\dots,\mathbf{K}_n$ to $\Delta_L(\mathbf{K}_1,\dots,\mathbf{K}_n)$ is therefore called the *weak social entropy process* and is denoted by $\mathbf{WSEP}(\mathbf{K}_1,\dots,\mathbf{K}_n)$.

For any $\mathbf{v}\in\Delta_L(\mathbf{K}_1,\dots,\mathbf{K}_n)$ there is an $n$-tuple $\mathbf{w}^{(1)}\in V_{\mathbf{K}_1},\dots,\mathbf{w}^{(n)}\in V_{\mathbf{K}_n}$ minimizing $\sum_{k=1}^{n}\mathrm{CE}(\mathbf{v},\mathbf{w}^{(k)})$ defined in (1). We will denote the set of all such $n$-tuples by $\Gamma_L(\mathbf{K}_1,\dots,\mathbf{K}_n)$.

**Lemma 4.1.** *The following are equivalent:*

*(i) The probability functions $\mathbf{v}$, $\mathbf{w}^{(1)},\dots,\mathbf{w}^{(n)}$ minimize (1) subject only to $\mathbf{w}^{(1)}\in V_{\mathbf{K}_1},\dots,\mathbf{w}^{(n)}\in V_{\mathbf{K}_n}$.*

*(ii) $\mathbf{w}^{(1)},\dots,\mathbf{w}^{(n)}$ maximize $\sum_{j=1}^{J}(\prod_{k=1}^{n}w_j^{(k)})^{\frac{1}{n}}$, subject only to $\mathbf{w}^{(1)}\in V_{\mathbf{K}_1},\dots,$ $\mathbf{w}^{(n)}\in V_{\mathbf{K}_n}$, and $v_j=\frac{(\prod_{k=1}^{n}w_j^{(k)})^{\frac{1}{n}}}{\sum_{j=1}^{J}(\prod_{k=1}^{n}w_j^{(k)})^{\frac{1}{n}}}$ for all $j=1,\dots,J$.*

For a proof see [11]. We will define the maximal value of $\sum_{j=1}^{J}(\prod_{k=1}^{n}w_j^{(k)})^{\frac{1}{n}}$ subject to $\mathbf{w}^{(1)}\in V_{\mathbf{K}_1},\dots,\mathbf{w}^{(n)}\in V_{\mathbf{K}_n}$ to be $\mathrm{M}_L(\mathbf{K}_1,\dots,\mathbf{K}_n)$.

The lemma above implies that $\mathbf{SEP}_L$ coincides with the logarithmic (or "normalised geometric mean") pooling operator of decision theory (cf. [2]) in the very special case when each $V_{\mathbf{K}_k}$ defines a single probability function.

In addition to the above pleasing properties, **SEP** satisfies a set of natural principles listed in [10] and [11] similar to those shown to be satisfied by **ME** in [9]. However these are almost certainly not sufficient to characterise **SEP** in the manner in which **ME** was characterised in [9].

Furthermore, although **SEP** is language invariant, it does not satisfy the Irrelevant Information Principle **IIP**. A simple counterexample is provided by the following[1]. Let $L_1=\{p\}$, $L_2=\{q\}$ and $L=L_1\cup L_2$. Knowledge bases $\mathbf{K}_1=\{\mathbf{w}(p)=0.2\}$, $\mathbf{F}_1=\{\mathbf{w}(q)=0.9\}$, $\mathbf{K}_2=\{\mathbf{w}(p)=0.4\}$, $\mathbf{F}_2=\{\mathbf{w}(q)=0\}$. There is only one $L$-probability function $\mathbf{w}^{(2)}\in V_{\mathbf{K}_2\cup\mathbf{F}_2}$: $(0,0.4,0,0.6)$. Hence

$$\mathrm{M}_L(x)=\sqrt{0.4(0.2-x)}+\sqrt{0.6(-0.1+x)},$$

[1] A counterexample to **IIP** for **SEP** was first found by Soroush Rafiee Rad (private communication, 2010).

which is maximal for $x = 0.16$.

$$\mathbf{SEP}(\mathbf{K}_1 \cup \mathbf{F}_1, \mathbf{K}_2 \cup \mathbf{F}_2)(p) = \frac{\sqrt{0.4(0.2-0.16)}}{\sqrt{0.4(0.2-0.16)} + \sqrt{0.6(-0.1+0.16)}} = 0.4 \neq$$
$$\neq \frac{\sqrt{2}}{\sqrt{2}+2\sqrt{3}} = \frac{\sqrt{0.08}}{\sqrt{0.08}+\sqrt{0.48}} = \mathbf{SEP}(\mathbf{K}_1, \mathbf{K}_2)(p).$$

Since **IIP** in its single agent variant played a crucial role in the characterisation of **ME** this failure could be interpreted as a significant criticism of **SEP**. However, while this principle may be too strong in the multi-agent case, note that the weaker **CIIP** principle may still be regarded as a natural generalization of the single agent **IIP** since it reduces to **IIP** for the case $n = 1$.

We say that **WSEP** satisfies **CIIP** if whenever $L = L_1 \cup L_2$ where $L_1$ and $L_2$ are disjoint propositional languages and $\mathbf{K}_1, \dots, \mathbf{K}_n$ and $\mathbf{F}_1, \dots, \mathbf{F}_n$ are knowledge bases formulated for the languages $L_1$ and $L_2$ respectively such that $\mathbf{F}_1, \dots, \mathbf{F}_n$ are jointly consistent, then

$$\mathbf{WSEP}_L(\mathbf{K}_1 \cup \mathbf{F}_1, \dots, \mathbf{K}_n \cup \mathbf{F}_n)|_{L_1} = \mathbf{WSEP}_{L_1}(\mathbf{K}_1, \dots, \mathbf{K}_n).$$

We prove that **WSEP** satisfies **CIIP** in the following section. However except in the cases when $\Delta_{L_1}(\mathbf{K}_1, \dots, \mathbf{K}_n)$ is a singleton the question whether **SEP** also satisfies **CIIP** remains open.

# 5 WSEP satisfies CIIP

In what follows we will fix two distinct propositional languages $L_1 = \{a_1, \dots, a_{h_1}\}$ and $L_2 = \{b_1, \dots, b_{h_2}\}$. Let $L = L_1 \cup L_2$ and let $\mathrm{At}(L_1) = \{\alpha_1, \dots, \alpha_J\}$ and $\mathrm{At}(L_2) = \{\beta_1, \dots, \beta_I\}$.

For $\mathbf{r} \in SL$, to simplify the notation we will often denote $\mathbf{r}|_{L_1}(\alpha_j)$ by $r_{j\cdot}$. We will also denote the conditional probability function $\mathbf{r}(\beta_i|\alpha_j)$ by $r_{i|j}$. It follows that $r_{ji} = r_{j\cdot} r_{i|j}$, i.e. the value $r_{ji}$ can be computed as the product of the projection of $\mathbf{r}$ to $L_1$ on the $L_1$-atom $\alpha_j$ and the conditional probability $\mathbf{r}(\beta_i|\alpha_j)$.

**Lemma 5.1.** *Let $w_j^{(k)} \geq 0$ be real numbers for all $1 \leq j \leq J$ and $1 \leq k \leq n$ where $k, j, J, n \in \mathbb{N}$. Then*

$$\sum_{j=1}^{J} (\prod_{k=1}^{n} w_j^{(k)})^{\frac{1}{n}} \leq (\prod_{k=1}^{n} \sum_{j=1}^{J} w_j^{(k)})^{\frac{1}{n}}. \tag{2}$$

*Equality holds if and only if either there are real constants $l^{(1)} > 0, \dots, l^{(n)} > 0$ such that $l^{(1)}(w_1^{(1)}, \dots, w_J^{(1)}) = l^{(2)}(w_1^{(2)}, \dots, w_J^{(2)}) = \dots = l^{(n)}(w_1^{(n)}, \dots, w_J^{(n)})$ or $\sum_{j=1}^{J} w_j^{(k)} = 0$ for some $k$.*

This lemma is Hölder's inequality, see [3], and it will be very useful in the following proof.

**Lemma 5.2.** *Let* $\mathbf{K}_1,\ldots,\mathbf{K}_n \in CL_1$, $\mathbf{F}_1,\ldots,\mathbf{F}_n \in CL_2$ *be such that* $\mathbf{F}_1,\ldots,\mathbf{F}_n$ *are jointly consistent.*

*(a) If* $\mathbf{v} \in \Delta_{L_1}(\mathbf{K}_1,\ldots,\mathbf{K}_n)$ *and* $\mathbf{t}$ *is an* $L_2$*-probability function such that* $\mathbf{t} \in \bigcap_{i=1}^n V_{\mathbf{F}_i}$ *then* $\mathbf{v}\cdot\mathbf{t} \in \Delta_{L_1\cup L_2}(\mathbf{K}_1\cup\mathbf{F}_1,\ldots,\mathbf{K}_n\cup\mathbf{F}_n)$. *In particular* $\mathbf{F}_1,\ldots,\mathbf{F}_n$ *could be empty in which case* $\mathbf{t}$ *can be arbitrary.*

*(b) Let* $\mathbf{r} \in \Delta_{L_1\cup L_2}(\mathbf{K}_1\cup\mathbf{F}_1,\ldots,\mathbf{K}_n\cup\mathbf{F}_n)$. *Then* $\mathbf{r}|_{L_1} \in \Delta_{L_1}(\mathbf{K}_1,\ldots,\mathbf{K}_n)$. *Moreover* $\mathrm{M}_{L_1\cup L_2}(\mathbf{K}_1\cup\mathbf{F}_1,\ldots,\mathbf{K}_n\cup\mathbf{F}_n) = \mathrm{M}_{L_1}(\mathbf{K}_1,\ldots,\mathbf{K}_n)$.

*Proof.* For a given $\mathbf{v} \in \Delta_{L_1}(\mathbf{K}_1,\ldots,\mathbf{K}_n)$ let $(\mathbf{p}^{(1)},\ldots,\mathbf{p}^{(n)}) \in \Gamma_{L_1}(\mathbf{K}_1,\ldots,\mathbf{K}_n)$ be such that $v_j = \frac{(\prod_{k=1}^n p_j^{(k)})^{\frac{1}{n}}}{\mathrm{M}_{L_1}(\mathbf{K}_1,\ldots,\mathbf{K}_n)}$. Note that $\mathrm{M}_{L_1}(\mathbf{K}_1,\ldots,\mathbf{K}_n) = \sum_{j=1}^J (\prod_{k=1}^n p_j^{(k)})^{\frac{1}{n}}$.

For a given $\mathbf{r} \in \Delta_{L_1\cup L_2}(\mathbf{K}_1\cup\mathbf{F}_1,\ldots,\mathbf{K}_n\cup\mathbf{F}_n)$ let

$$(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)}) \in \Gamma_{L_1\cup L_2}(\mathbf{K}_1\cup\mathbf{F}_1,\ldots,\mathbf{K}_n\cup\mathbf{F}_n)$$

be such that $r_{ji} = \frac{(\prod_{k=1}^n w_{ji}^{(k)})^{\frac{1}{n}}}{\mathrm{M}_{L_1\cup L_2}(\mathbf{K}_1\cup\mathbf{F}_1,\ldots,\mathbf{K}_n\cup\mathbf{F}_n)}$.

Let us consider probability functions $\mathbf{w}^{(1)}|_{L_1},\ldots,\mathbf{w}^{(n)}|_{L_1}$. Denote

$$M = \sum_{j=1}^J (\prod_{k=1}^n w_{j\cdot}^{(k)})^{\frac{1}{n}}.$$

Then $M \le \mathrm{M}_{L_1}(\mathbf{K}_1,\ldots,\mathbf{K}_n)$ since $\mathrm{M}_{L_1}(\mathbf{K}_1,\ldots,\mathbf{K}_n)$ is maximal. But by the lemma 5.1 also $\mathrm{M}_{L_1\cup L_2}(\mathbf{K}_1\cup\mathbf{F}_1,\ldots,\mathbf{K}_n\cup\mathbf{F}_n) \le M$, hence

$$\mathrm{M}_{L_1\cup L_2}(\mathbf{K}_1\cup\mathbf{F}_1,\ldots,\mathbf{K}_n\cup\mathbf{F}_n) \le \mathrm{M}_{L_1}(\mathbf{K}_1,\ldots,\mathbf{K}_n). \tag{3}$$

(a) Let $\mathbf{t} \in \bigcap_i V_{\mathbf{F}_i}$. We are going to prove that

$$(\mathbf{p}^{(1)}\cdot\mathbf{t},\ldots,\mathbf{p}^{(n)}\cdot\mathbf{t}) \in \Gamma_{L_1\cup L_2}(\mathbf{K}_1\cup\mathbf{F}_1,\ldots,\mathbf{K}_n\cup\mathbf{F}_n). \tag{4}$$

It is easy to see that $\mathbf{p}^{(1)}\cdot\mathbf{t},\ldots,\mathbf{p}^{(n)}\cdot\mathbf{t}$ satisfy $\mathbf{K}_1\cup\mathbf{F}_1,\ldots,\mathbf{K}_n\cup\mathbf{F}_n$ respectively. Moreover,

$$\sum_{j=1,\ldots,J,i=1,\ldots,I} (\prod_{k=1}^n p_j^{(k)} t_i)^{\frac{1}{n}} = \sum_{j=1,\ldots,J,i=1,\ldots,I} (\prod_{k=1}^n p_j^{(k)})^{\frac{1}{n}} t_i = \mathrm{M}_{L_1}(\mathbf{K}_1,\ldots,\mathbf{K}_n),$$

since $\sum_{i=1}^I t_i = 1$. But from (3) we already know that $\mathrm{M}_{L_1\cup L_2}(\mathbf{K}_1\cup\mathbf{F}_1,\ldots,\mathbf{K}_n\cup\mathbf{F}_n) \le \mathrm{M}_{L_1}(\mathbf{K}_1,\ldots,\mathbf{K}_n)$ hence (4) is proved.

(b) By the maximality of $\mathrm{M}_{L_1\cup L_2}(\mathbf{K}_1\cup\mathbf{F}_1,\ldots,\mathbf{K}_n\cup\mathbf{F}_n)$ and by (3) we have

$$\mathrm{M}_{L_1\cup L_2}(\mathbf{K}_1\cup\mathbf{F}_1,\ldots,\mathbf{K}_n\cup\mathbf{F}_n) = M = \mathrm{M}_{L_1}(\mathbf{K}_1,\ldots,\mathbf{K}_n) \tag{5}$$

hence

$$\sum_{j=1,\ldots,J,i=1,\ldots,I} (\prod_{k=1}^n w_{ji}^{(k)})^{\frac{1}{n}} = \sum_{j=1}^J (\prod_{k=1}^n \sum_{i=1}^I w_{ji}^{(k)})^{\frac{1}{n}}.$$

By lemma 5.1 this equality could only occur if for each $j$ there are real constants $l_j^{(1)} > 0, \dots, l_j^{(n)} > 0$ such that the proportionality

$$l_j^{(1)}(w_{j1}^{(1)}, \dots, w_{jI}^{(1)}) = l_j^{(2)}(w_{j1}^{(2)}, \dots, w_{jI}^{(2)}) = \ldots = l_j^{(n)}(w_{j1}^{(n)}, \dots, w_{jI}^{(n)})$$

holds, or $w_{j\cdot}^{(k)} = \sum_{i=1}^{I} w_{ji}^{(k)} = 0$ holds for some $k$.

Let us consider coefficient $j$ to be fixed. If $w_{j\cdot}^{(k)} = 0$ for every $k$ let $\mathbf{q}_{\cdot|j}$ be an arbitrary $L_2$-probability function with value on $i$-th atom denoted as $q_{i|j}$. Otherwise for $\bar{k}$ such that $w_{j\cdot}^{(\bar{k})} \neq 0$ let us define

$$q_{i|j} = \frac{w_{ji}^{(\bar{k})}}{w_{j\cdot}^{(\bar{k})}}.$$

Obviously,

$$\sum_{i=1}^{I} q_{i|j} = \sum_{i=1}^{I} \frac{w_{ji}^{(\bar{k})}}{\sum_{i=1}^{I} w_{ji}^{(\bar{k})}} = 1$$

and hence $\mathbf{q}_{\cdot|j}$ is a well defined $L_2$-probability function. Notice that thanks to proportionality the definition does not depend on the choice of $\bar{k}$:

$$\frac{l_j^{(\bar{k})} w_{ji}^{(\bar{k})}}{l_j^{(\bar{k})} \sum_{i=1}^{I} w_{ji}^{(\bar{k})}} = \frac{l_j^{(k)} w_{ji}^{(k)}}{l_j^{(k)} \sum_{i=1}^{I} w_{ji}^{(k)}}.$$

In other words

$$w_{ji}^{(k)} = w_{j\cdot}^{(k)} q_{i|j}. \tag{6}$$

By (5) the projections to $L_1$ satisfy

$$(w^{(1)}|_{L_1}, \dots, w^{(n)}|_{L_1}) \in \Gamma_{L_1}(\mathbf{K}_1, \dots, \mathbf{K}_n).$$

Then for $L_1$-probability function $\mathbf{v}$ defined by $v_j = \frac{(\prod_{k=1}^{n} w_{j\cdot}^{(k)})^{\frac{1}{n}}}{\sum_{j=1}^{J}(\prod_{k=1}^{n} w_{j\cdot}^{(k)})^{\frac{1}{n}}}$ we have that $\mathbf{v} \in \Delta_{L_1}(\mathbf{K}_1, \dots, \mathbf{K}_n)$.

Moreover,

$$r_{ji} = \frac{(\prod_{k=1}^{n} w_{ji}^{(k)})^{\frac{1}{n}}}{\sum_{j=1}^{J}\sum_{i=1}^{I}(\prod_{k=1}^{n} w_{ji}^{(k)})^{\frac{1}{n}}} = \frac{(\prod_{k=1}^{n} w_{j\cdot}^{(k)} q_{i|j})^{\frac{1}{n}}}{\sum_{j=1}^{J}\sum_{i=1}^{I}(\prod_{k=1}^{n} w_{j\cdot}^{(k)} q_{i|j})^{\frac{1}{n}}} = v_j q_{i|j},$$

where $r_{j\cdot} = \sum_i v_j q_{i|j} = v_j$ and $r_{i|j} = \frac{r_{ji}}{r_{j\cdot}} = \frac{v_j q_{i|j}}{r_{j\cdot}} = q_{i|j}$ which gives us the required result that $\mathbf{r}|_{L_1} \in \Delta_{L_1}(\mathbf{K}_1, \dots, \mathbf{K}_n)$. □

**Theorem 5.3. WSEP** *satisfies* **CIIP**.

This follows at once from lemma 5.2.

**Theorem 5.4.** **SEP** *satisfies the* **CIIP** *in the special case when there is only one probability function in* $\Delta_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$*, say* $\Delta_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \{\mathbf{w}\}$*. Note that by theorem 3.8 in [11] this holds whenever at least one of the agents has a knowledge base which fixes a probability function for* $L_1$*.*

*Proof.* By lemma 5.2 (b) clearly

$$\mathbf{SEP}_{L_1 \cup L_2}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)|_{L_1} = \mathbf{r}|_{L_1} = \mathbf{w} = \mathbf{SEP}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

□

## 6 Conclusion

In this paper we have sought to investigate the Irrelevant Information Principle in the context of multi-agent uncertain reasoning. While this principle plays a crucial role in an axiomatic characterization of **ME** given in [9], we have argued that the most obvious generalization of the Irrelevant Information Principle to the multi-agent context may be too strong. We have proposed an alternative generalization called the Consistent Irrelevant Information Principle for a social inference process (**CIIP**). We have described the promising social inference process **SEP** first formulated in [10] and its weaker counterpart **WSEP**. We have shown that **WSEP** satisfies **CIIP** and that **SEP** satisfies **CIIP** in many cases. The question as to whether **SEP** satisfies **CIIP** remains open.

The authors are very grateful to Dr Alena Vencovská for spotting a mistake in the original proof of lemma 5.2. The first author is supported by Marie Curie Early Stage Researcher Grant (MALOA - from Mathematical Logic to Applications) financed by the European Commission.

## References

[1] Ch. Genest and C. G. Wagner, (1987) *Further evidence against independence preservation in expert judgement synthesis*, Aequationes Mathematicae, 32

[2] S. French, *Group Consensus Probability Distributions: A Critical Survey*, in J. M. Bernardo, M. H. De Groot, D. V. Lindley, and A. F. M. Smith (Eds.), Bayesian Statistics, Elsevier, North Holland, 1985, pp. 183-201.

[3] G. H. Hardy, J. E. Littlewood, G. Plya (1934), *Inequalities*, Cambridge University Press

[4] P. Hawes, *An Investigation of Properties of Some Inference Processes*, PhD Thesis, Manchester University, MIMS eprints, 2007,

[5] E. T. Jaynes, 1979 *Where do we Stand on Maximum Entropy?* in "The Maximum Entropy Formalism", R. D. Levine and M. Tribus (eds.), M.I.T. Press, Cambridge, MA.

[6] R. W. Johnson and J. E. Shore, 1980 *Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy*, IEEE Transactions on Information Theory, IT-26(1)pp. 26-37

[7] J. B. Paris, (1994) *The uncertain reasoner companion*, Cambridge University Press, Cambridge

[8] J. B. Paris and A. Vencovská, (1989) *On the Applicability of Maximum Entropy to Inexact Reasoning*, International Journal of Approximate Reasoning, 3

[9] J. B. Paris and A. Vencovská, (1990) *A Note on the Inevitability of Maximum Entropy*, International Journal of Approximate Reasoning, 4, pp. 183-224

[10] G. Wilmers (2010), *The Social Entropy Process: Axiomatising the Aggregation of Probabilistic Beliefs* , Probability, Uncertainty and Rationality, pp 87-104, edited by Hykel Hosni and Franco Montagna, 10 CRM series, Scuola Normale Superiore, Pisa

[11] G. Wilmers (2011), *Generalising the Maximum Entropy Inference Process to the Aggregation of Probabilistic Beliefs*, available from http://manchester.academia.edu/GeorgeWilmers/Papers

# Preliminary Steps in Uninformed Approach to the Extraction of Context in Multilingual Corpora

**Vladislav Bína and Jiří Přibil**
Faculty of Management
University of Economics in Prague
Jarošovská 1117/II, 37701 J. Hradec
{bina,pribil}@fm.vse.cz

**Abstract**

At the present era, the most of the facets of human life are exposed to an informational explosion. This assertion is valid especially for the fields of business and managerial decision making. The immense unstructured information sources indispensably demand automatic methods for extraction of the context and efficient handling of the texts. The requirement of automaticity implies the necessity to develop uninformed approaches in text mining (more profound consideration can be found in [4]).

This contribution thus belongs to the area of text mining and ranks among uninformed tools for the analysis of text corpora, namely for the automatic classification of documents into groups according to a language based on non-informative words. The other important feature is the removal of related words (or "desynonymization") of given text.

## 1 Introduction

First of all, let us specify the prerequisites of the approach. The presented method assumes that alphabetic scripts are used in the considered texts, requires a sufficient size of the particular documents (it means at least thousands of words, for an example see word counts in Table 1) and expects their monolinguality (the text corpus itself is assumed to be multilingual). The corpus is dynamic which means that we build it up by sequential addition of considered documents together with information about the membership in monolingual group. Based on these requirements, the first step of addition to the corpus consist of preprocessing operations. We start with a search of non-informative words based on frequency analysis of the considered text. This step corresponds to stopwords as one of the features used for language identification (see e.g. [3]).

As the first approximation, we assume that a significant portion of non-informative words must belong to the most frequent words in the document. Such words are then

Table 1: Word counts, languages and values of threshold for small Gutenberg corpus.

| Title of ebook | Word count | Language | Criterion |
|---|---|---|---|
| A Vuela Pluma | 83334 | spanish | 0.00122 |
| Reise in die Aequinoctial | 106053 | german | 0.00117 |
| At Sundown | 5904 | english | 0.00126 |
| Autour de la Lune | 59022 | french | 0.00120 |
| Tre Racconti Sentimentali | 28039 | italian | 0.00040 |
| Briefe aus dem Gefängnis | 13347 | german | 0.00090 |
| Cidades e Paizagens | 17342 | portuguese | 0.00067 |
| Die Geschwister | 5368 | german | 0.00083 |
| Hendes Højhed | 17203 | danish | 0.00168 |
| Judith | 23263 | dutch | 0.00161 |
| King Henry the Eighth | 26985 | english | 0.00059 |
| La princesse de Cleves | 64248 | french | 0.00119 |
| Marta y María | 90234 | spanish | 0.00039 |
| Nature and Culture | 47055 | english | 0.00047 |
| Nervosos, Lymphaticos ... | 38841 | portuguese | 0.00055 |
| Octavia | 12216 | portuguese | 0.00057 |
| Principles of Orchestration | 42776 | english | 0.00034 |
| Rautatie | 22918 | finnish | 0.00202 |
| Suicida | 5271 | portuguese | 0.00054 |
| Wilde Bob | 57225 | dutch | 0.00226 |

compared to the most frequent words of particular monolingual groups in corpus. When sufficient proportion of same words is detected, the document is classified in the corresponding language group.

The subsequent step is the “quasi-lemmatization” or “desynonymization” of the text based on the employment of modified (weighted) Levenshtein distance measure, which takes into account possible deletions, insertions or substitutions in two compared sequences together with a possibility to use the additional operation of transposition (see e.g. [6]). The procedure of “quasi-lemmatization” itself is based on sequential addition of the actually processed text into (already) quasi-lemmatized corpus. However, this approach demands comparison of all informative words in actually analyzed text with all words in the corpus, which in case of larger corpora appears to be computationally infeasible. Therefore, a heuristics based on proportions of the same letters can be applied in order to select rather similar words further examined using modified Levenshtein measure of distance.

## 2 Preprocessing

As the very first step in the processing of the text document we perform two usual linearization (document content filtration) steps - (a) markup and format removal and (b) tokenization (lower case conversion, removal of the digits, special symbols, punctuation, etc). Another reasonable preliminary step is a removal of very short (uninformative) words.

## 3 Determination of stopwords

As a stopword we consider commonly used words not important for the content of document. In the classical concept of text minning the stopwords are given in a list separately for each language and the stopword lists slightly differ in various approaches and in particular software tools. Such stopwords list need to be defined manually which is no use in an uninformed approach.

This step is based on the construction of term-document matrix containing frequencies of particular terms in the whole corpus of documents. A term is considered to be *stopword* if its occurrence expressed as percentage is above some threshold[1]. The question arises: How to determine the value of threshold? A possible answer lies in the computational experiments.

Let us consider the following example, a small document corpus containing twenty ebooks (see first column of Table 1) from Project Gutenberg web pages [2]. It should be stressed that the languages of particular books mentioned in the table are presented only for sake of clarity; the stopword search and consequent steps are still uninformed.

As we already mentioned as a stopword we consider any term appearing in the document with sufficient frequency. Which value of relative frequency can serve as a threshold? To find a suitable value we perform a computations with different setting and evaluate some suitable criterion allowing a reasonable choice based on the knowledge of document's language and a corresponding list of stopwords in R text mining package [5].

The determination of threshold is based on comparison of the frequency based stopword list and known official list of stopwords. Such criterion should obey the following requirements. The optimal value is achieved when the highest proportion of true stopwords (from a `tm` list) is found, but it is also crucial to keep the number of false stopwords limited. Therefore we propose a criterion in the form of proportion product

$$Q = \frac{\hat{n}_t}{\hat{N}} \cdot \frac{\hat{n}_t}{N_t},$$

where $N_t$ means the number of all true stopwords from `tm` package, $\hat{N}$ means the number of all stowords found by the frequency based approach and $\hat{n}_t$ is the size of intersection of both groups, i.e. the number of all true stopwords found by frequency based approach. As a result, a value of threshold is chosen resulting into the highest criterion value.

To illustrate the typical shape of criterion curve we draw corresponding graphs for the first occurrence of each language in our example (see Figure 1).

For our corpus of twenty documents we found the following values of threshold. Average value of the twenty threshold values is 0.00099, hence the choice of value 0.001 appears to be reasonable as an estimate relative frequency separating the stopwords from the other terms.

[1] An alternative approach can avoid definition of any threshold but can simply consider as stopwords, let us say, one hundred most frequent words in the document.

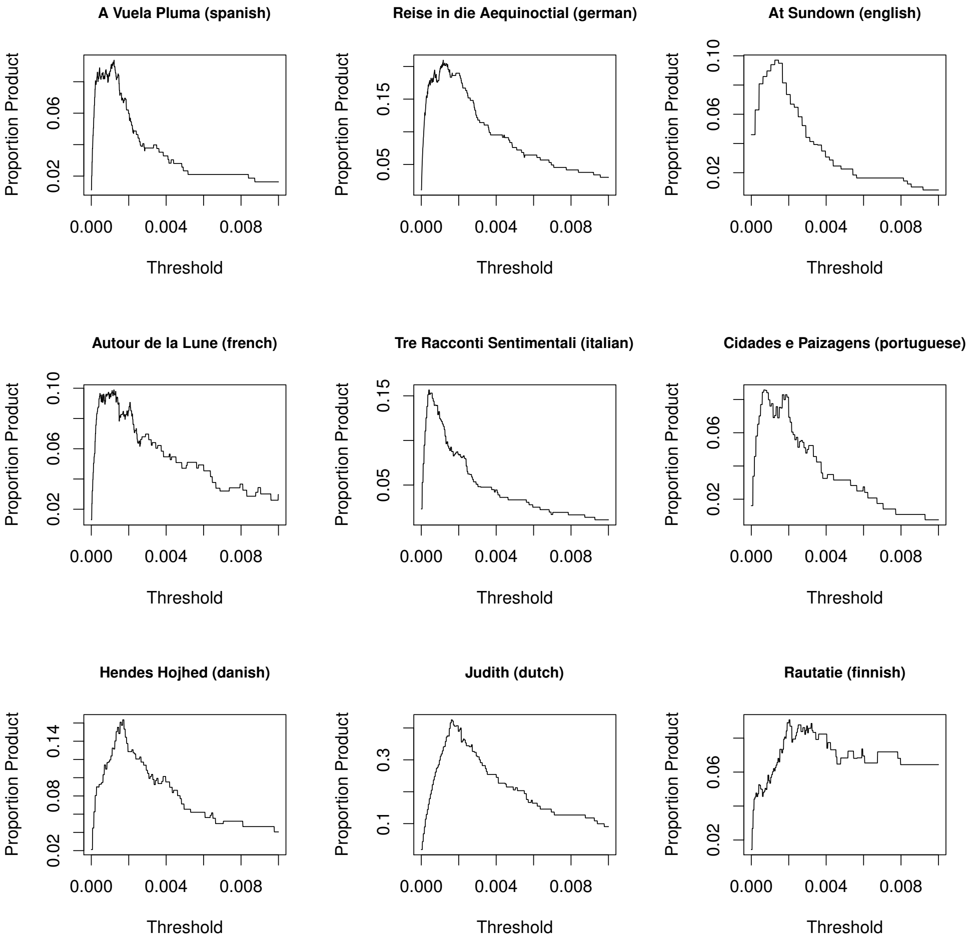


Figure 1: Sensitivity of stopword search on value of threshold for selected documents.

# 4 Language categorization

In the preceding section we described a frequency based approach to the identification of stopwords in a monolingual document. Now on the basis of this capability we perform the categorization of languages in the documents incorporated to the corpus.

As we mentioned above the corpus is dynamic. It means that we begin with a first document where the stopwords are identified. This document is labeled as language number 1. During the process of addition of any subsequent document to a corpus we find its stopwords and classify as one of the preceding languages in the case when sufficient proportion of stopwords is common. In the opposite case we assign next free number as a label of still unnoticed language.

In the experiments with a small Gutenberg Project corpus we found out that reasonable proportions of common stopwords in documents writen in same languages appear to lay somewhere between 20 and 30 percents. When the smaller proportion is chosen (in our case less than 18 percent), the method fails to recognize between spanish and portuguese texts. On contrary, when we expect proportion of common stopwords as high as 35 percent, the approach begin to create artificial languages in case of documents somehow differing. In our case the document "Principles of Orchestration" began to differ.

A useful modification of the method can be proposed. It is based on collecting of the documents with same language and determination of the stopwords in whole monolingual subcorpus. Of course, at this point it is reasonable to develop a suitable heuristics labouring under the knowledge of stopword frequences in each document without necessity to recompute repeatedly the stopwords in whole monolingual subcorpus.

Naturally, the setting of parameters requires more profound analysis and use of larger corpora but the above stated values appear to be a reasonable starting point.

Let us have a look at the result of language categorization in case of setting the common proportion of stopwords to the value 0.2 (see Table 2). We can see that all languages in the corpus were successfully recognized and on the other hand all document written in the same language were classified into the same group.

As an example of some source of problem in case of technical and other specific text in Table 3 we present the first fifty stopwords (in alphabetical ordering) from the document "Principles of Orchestration".

The attentive reader surely noticed that a significant proportion of stopwords in this document is connected with the musical terminology which can be easily considered as a separate language. Alternatively - with an appropriate setting of parameter - the methodology can be employed to categorize well distinguished areas of interest.

# 5 Quasi-lemmatization

As we can observe in Table 3, many stopwords appear in different forms, in English usually in singular and plural (in our case e.g. couples if words: case and cases or flute and flutes). However, this is a general problem arising not only in the procedure of stopword removal. Moreover, this issue is even more substantial in slavic (and many

Table 2: The third column presents language categorization in case of 20 percent of common stopwords.

| Title of ebook | Language | Category |
|---|---|---|
| A Vuela Pluma | spanish | 1 |
| Reise in die Aequinoctial | german | 2 |
| At Sundown | english | 3 |
| Autour de la Lune | french | 4 |
| Tre Racconti Sentimentali | italian | 5 |
| Briefe aus dem Gefängnis | german | 2 |
| Cidades e Paizagens | portuguese | 6 |
| Die Geschwister | german | 2 |
| Hendes Højhed | danish | 7 |
| Judith | dutch | 8 |
| King Henry the Eighth | english | 3 |
| La princesse de Cleves | french | 4 |
| Marta y María | spanish | 1 |
| Nature and Culture | english | 3 |
| Nervosos, Lymphaticos ... | portuguese | 6 |
| Octavia | portuguese | 6 |
| Principles of Orchestration | english | 3 |
| Rautatie | finnish | 9 |
| Suicida | portuguese | 6 |
| Wilde Bob | dutch | 8 |

Table 3: The first fifty stopwords in the document "Principles of Orchestration".

| | | | | | |
|---|---|---|---|---|---|
| above | act | all | also | and | another |
| any | are | balance | bar | bass | basses |
| bassoon | bassoons | before | being | between | brass |
| bride | but | can | cantabile | case | cases |
| cellos | certain | clarinet | clarinets | cockerel | colour |
| combination | composer | different | distribution | divided | double |
| doubled | doubling | each | effect | employed | eng |
| etc | example | examples | expression | fag | first |
| flute | flutes | | | | |

other) languages, where grammatical cases of the same word differ in a suffix.

This problem is usually solved by language-specific tools like stemmers or thesauri. However, this is not feasible in case of an uninformed approach. We are facing the need to develop language-independent method for searching the similar words which is called *quasi-lemmatization*. As a reasonable choice appears to be the similarity measurement based on the well known string metric *Levenshtein distance*. This measure has interesting features, it is typically used in cases when small number of differences is expected. The operations considered by this measure are addition, removal and substitution of single letters [6]. Sometimes an additional operation of transposition is also involved (see [1]), but this appears even more computationally demanding, therefore we use the classical variant.

Since mutual comparison of all pairs of words using Levenshtein distance is not computationally feasible, we employ a heuristic method based on significant proportion of common letters in the pair of compared words. Whenever this heuristic detects sufficient similarity, we refine the recognition of similar words employing the Levenshtein distance.

## 5.1 Mutual comparison - a heuristic

As we already stressed, the mutual comparison performed for each pair of words needs to be simple and fast. We propose a method based on computing of letter frequencies in both words and their subtraction. This guarantees that all letters with same number of occurrences in both words cancel out and as nonzero term remain only letters differing in their counts. If we then sum up the absolute values of these differences we obtain a criterion usually giving high numbers for different words and small values for similar ones

$$T_H = \sum_i |n_{1i} - n_{2i}| \,.$$

Index $i$ stands for a summation over all words appearing in both words, $n_{1i}$ is a frequency of particular letters in the first (and $n_{2i}$ in the second) word.

It is reasonable to use this criterion in a relative manner, namely to compare its value with a reasonable multiple of average length of both words. E.g. we consider two words to be similar if

$$C_H \cdot T_H = 2.5 T_H < \frac{N_1 + N_2}{2},$$

where $C_H$ is a multiplying coefficient with reasonable values between 2 and 3 (in examples 2.5 is used) and $N_1$ resp. $N_2$ stand for lengthes of both words.

## 5.2 Mutual comparison - Levenshtein distance

A heuristic presented in previous part provide some candidates for similar words. More sophisticated analysis can be performed using Levenshtein string metric [6]. This well known refines the process and identifies similar pairs. Again, this measure of string distance can be applied relatively, which means in comparison to an average length of the two words allowing more differences in longer words. Moreover, if the proportion is smaller than some constant (we used value 0.4), pair of words is considered to be similar.

Now among the pair of similar words one representant is chosen and the other variant is removed from the corpus (and substituted by the chosen word). This choice can be based on two alternative principles:

- choice according to the length - the shorter word is chosen assuming that it represents more fundamental variant,
- choice according to the frequency - more frequent variant is chosen representing the more obvious word.

We employ the first of preceding two variants.

Let us have a look at the example of comparison results. All pairs of words in a sample listing bellow passed through the heuristic criterion from Subsection 5.1 with result of similarity. However, the pairs in parentheses (and denoted by the word `rejected`) were considered by Levenshtein string metric (see Subsection 5.2) as different.

```
gebirgsart <- urgebirgsarten
gebirgsbevölkerung -> gebirgsvölker
    (rejected:  gebirgsbildungen, uebergangsbildung)
gebirgsbildungen <- urgebirgsbildungen
    (rejected:  gebirgskette, gesteigert)
gebirgskette <- kalkgebirgskette
    (rejected:  gebirgskette, steigerte)
    (rejected:  gebirgskette, steigerten)
    (rejected:  gebirgskette, bersteigt)
    (rejected:  gebirgskette, berstiege)
gebirgskette <- urgebirgsketten
gebirgsland <- gebirgslande
gebirgsland <- gebirgswand
gebirgsschichten -> gebirgsstrichen
    (rejected:  gebirgsspalten, geistesanlage)
    (rejected:  gebirgsspalten, granitberges)
```

The changes accepted by Levenshtein distance measure (not in parentheses) contain in the listing also hinted direction of a change. It is apparent that sometimes a problematic change can occur as for example in case of the two last accepted pairs.

The result of such “quasi-lemmatization” strongly depends on language, type of text and setting of constants but usually leads to a circa fifty percent reduction of the word count.

## 6 Conclusions

Authors bring fundamental ideas of preliminary steps necessary to design an efficient method for text mining in case of uninformed approach. It is apparent that the presented techniques are dependent on several constants with seemingly arbitrarily specified values. These values are not necessarily the optimal ones but were estimated during several testing runs and - in case of threshold for stopword search - set up using a criterion based on real and frequency based stopwords' proportions.

The basic preliminary steps thus consist of stopword search, language classification based on stopwords and so called quasi-lemmatization allowing the significant reduction of terms contained in the corpus.

**Acknowledgement:**

The research was partially supported by GA ČR under grant no. 403/12/2175 and University of Economics in Prague under project no. F6/15/2012.

## References

[1] Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. Commun. *ACM* **7**(3), pp. 171-176.

[2] Project Gutenberg. (2012). Project Gutenberg Association. `http://www.gutenberg.org/`

[3] Prager, J.M. (1999), Linguini: Language Identification for Multilingual Documents. In *Proceedings of the Thirty-Second Annual Hawaii International Conference on System Sciences, vol. 2*, IEEE Computer Society. pp. 2035.

[4] Přibil, J., Kincl, T., Bína, V., Novák, M. (2011). Language Independent System for Document Context Extraction. In *Proceedings of the World Congress on Engineering and Computer Science 2011, vol. 1*, WCECS 2011, October 19-21, 2011, San Francisco, USA. pp. 51-55.

[5] Feinerer, I., Hornik, K., Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software* **25**(5). `http://www.jstatsoft.org/v25/i05/`

[6] Ziolko, M. et al. (2010). Modified weighted Levenshtein distance in automatic speech recognition. In *Proceedings of the sixteen national conference on Applications of mathematics in biology and medicine*, Krynica, Poland, pp. 116-120.

# A Dynamic Conditional Decision Model

**Andrea Capotorti and Giulianella Coletti**
Dept. of Matematica e Informatica
University of Perugia
via Vanvitelli, 1, 06123 Perugia, Italy
{capot,coletti}@dmi.unipg.it

**Barbara Vantaggi**
Dept. of S.B.A.I.
"Sapienza" University of Rome
via Scarpa 46a, 00185 Rome, Italy
barbara.vantaggi@sbai.uniroma1.it

**Abstract**

We intend to provide a characterization of decision models based on partial information and on choices among conditional acts. Proper representability of choices in such frameworks can be performed locally and globally: we deal with local representability. In particular we study partial conditional ordinal relations among events and complete conditional preferences on acts.

## 1 Motivations

In a comparative framework it is natural to require that every possible event $E$ of a family $\mathcal{E}$ is strictly more likely then the impossible event $\phi$.

Nevertheless we know that, whenever $\mathcal{E}$ is an uncountable set, most events must have zero or one probability, hence, to have a comparative degree of belief representable by a probability, we can only require that $\phi \preccurlyeq E$ and $\phi \prec \Omega$. For this reason Savage introduced the concept of null events through the condition that any two acts restricted to one of such event are indifferent (and so indistinguishable). This is actually caused by chosen numerical framework (real valued probability) which does not allow to give positive degree of belief to all events.

The same problem can be present also in a dynamic situation in which we start from a possibly finite set of random quantities (acts) and by considering as initial set of events (state of world) only those induced by the random quantities at hand. Even when we require all these events $E_i$ to be not-null, i.e. $\phi \prec E_i$, and the relation among acts is representable by a (positive) probability of the induced comparative probability, this does not ensure that the comparative probability can be extended in a way that it is representable by a positive probability, as the following example shows.

**Example 1.** *Let $\vee$, $\wedge$ and $(\cdot)^c$ denote the usual logical disjunction, conjunction and negation operator. Let $\mathcal{E} = \{A, B, C\}$ with $A \wedge B \wedge C = \phi$. Consequently the algebra spanned by $\mathcal{E}$ has 7 atoms and it is immediate to see that the comparative degree of belief*

$$\phi \prec A \prec B \prec C \quad A \vee B \sim A \vee B \vee C \tag{1}$$

*is representable by a probability but that necessarily it will result $A^c \wedge B^c \wedge C \sim \phi$.*

In order to have $\phi \prec A$, for any potential event $A$, it is necessary to resume different representability: either through a probability with values on a non-archimedian field $\mathbb{R}^* \supsetneq \mathbb{R}$ or a local one using conditional probabilities (both in $\mathbb{R}^*$ or in $\mathbb{R}$). In both cases we will be able to distinguish "zero" events in different levels of zeroes.

In literature several attempts to generalize Savage axioms to admit *negligible even possible* events, partial assessments and dynamic decision have been proposed, see e.g. [2, 10, 8]. Dynamism can be reduced to reason about classes of preferences, each relation being conditioned to a specific information. This can be formalized through conditional preference relations (see again [8]). In this contribution we bring to the light peculiarities of conditional preference relations with specific representability requirements. We study the representability problem of partial conditional preference relations which allow negligible events by referring either to non-archimedean conditional probabilities or to real valued conditional probabilities in the sense of de Finetti [7].

## 2 Preliminaries

As already mentioned, a preference relation among acts induces a comparative probability among events. Acts are seen here as simple random quantities with possible consequences on a set $\mathcal{X}$. Differently from usual approaches, we think the state space $\Omega$ implicitly described by the values expressed by acts. In particular, a binary act $f$ expressing only two values $x_1$ and $x_2$ implicitly define an event $A$ as all real situations letting $f$ to take value $x_1$ and its contrary event $A^c$ as all real situations letting $f$ to take value $x_2$. Constat act, i.e. those that takes a unique value $x \in \mathcal{X}$ whatever it happens, implicitly define the sure event $\Omega$ and its contrary $\phi$, the impossible event. A preference relation $\preccurlyeq$ among an arbitrary set of acts hence induces a preference relation, that with an abuse of notation we continue to indicate with $\preccurlyeq$, among an arbitrary set of events $\mathcal{E}$. As usual we denote with $\prec$ and $\sim$ the asymmetrical and symmetrical parts of $\preccurlyeq$, respectively.

Let us hence focus on a comparative probability on an arbitrary set of events $\mathcal{E}$ which is a binary relation $\preccurlyeq$ expressing evaluations of the type "no more probable than". Axioms for a comparative probability are:

**(1)** for any $E \in \mathcal{E}$ we have $\phi \preccurlyeq E$ and the not-triviality requirement $\phi \prec \Omega$;

**(2)** $\preccurlyeq$ is a weak order;

**(P)** If $A, B, C, A \vee C, B \vee C \in \mathcal{E}$ are such that $A \wedge C = B \wedge C = \phi$ then

$$A \preccurlyeq B \iff A \vee C \preccurlyeq B \vee C \quad .$$

De Finetti instead of (1) required (1') $\phi \prec E$. We will say a comparative probability to be *positive* if it verifies (1'). A comparative probability on $\mathcal{E}$ is said to be *coherent* if for any finite sub-family $\mathcal{F} \subseteq \mathcal{E}$ and for every $\lambda_i > 0$ and $E_i, F_i \in \mathcal{F}$ such that $E_i \preccurlyeq F_i$ it results

$$\sup \sum_i \lambda_i(|F_i| - |E_i|) \leq 0 \Rightarrow E_i \sim F_i \forall i \tag{2}$$

where the supremum is taken over all possible values derived from the indicator functions $|\cdot|$ of the events.

Coherence implies that for any event $E \in \mathcal{E}$ we have $\phi \preccurlyeq E$, reflexivity and transitivity of $\preccurlyeq$ and axiom (P). On the contrary it does not imply neither (1') nor the non-triviality $\phi \prec \Omega$ conditions.

## 3 Representability

In order to give a representability result based on conditional probabilities taking values in a non-archimedean field, we refer to [12] for the main notions on hypereals and we denote by $[0,1]^*$ a non-archimedean extension of the real unit interval $[0,1]$.

**Theorem 1.** *Let $\mathcal{E}$ be a finite set of events and let $\mathcal{C}_{\mathcal{E}}$ be the set of atoms generated by $\mathcal{E}$. For a positive comparative probability the following statements are equivalent:*

**i)** $(\mathcal{E}, \preccurlyeq) \cup (\phi \prec C_k)_{C_k \in \mathcal{C}_{\mathcal{E}}}$ *is a positive coherent comparative probability;*

**ii)** *there exists a probability function $p :< \mathcal{E} >\rightarrow [0.1]$ strictly positive which represents $\preccurlyeq_{|\mathcal{E} \cup \mathcal{C}_{\mathcal{E}}}$;*

**iii)** *there exists a non-archimedean probability function $p^* :< \mathcal{E} >\rightarrow [0.1]^*$ strictly positive which represents $\preccurlyeq_{|\mathcal{E} \cup \mathcal{C}_{\mathcal{E}}}$.*

Proof: Equivalence between i) and ii) is the usual representation theorem of coherent comparative probabilities (see e.g. [4]). Equivalence between ii) and iii) directly derives from the transfer principle between $[0,1]$ and $[0,1]^*$ (see e.g. [12]), that holds since the finiteness of $\mathcal{E}$. Strict positivity of $p$ derives from the fact that the represented comparative probability is positive also on atoms.

□

A coherent comparative probability on $\mathcal{E}$ can be extended to a coherent comparative probability on $\mathcal{E}' \supset \mathcal{E}$, but a positive coherent comparative probability is not necessarily extendible to positive one. In particular positivity requirement cannot hold for some atom (see again Example 1). Nevertheless it is important to note that, given a positive coherent comparative probability on a countable algebra, representable by a positive probability, there is a positive coherent extension on any countable superalgebra, representable by a positive probability:

**Theorem 2.** *Let $\mathcal{E}, \mathcal{A}$ be two countable algebras such that $\mathcal{E} \subseteq \mathcal{A}$, given a positive comparative probability $\preccurlyeq$, which is representable by a positive probability, there is at least a positive comparative probability $\preccurlyeq'$ extending $\preccurlyeq$, which is representable by a positive probability.*

Proof: By hypothesis $\preccurlyeq$ representable by a positive probability $p$ on $\mathcal{E}$. Let $\mathcal{C}_{\mathcal{E}}$ be the set of atoms of $\mathcal{E}$, then $p(C) > 0$ for any $C \in \mathcal{C}_{\mathcal{E}}$. Moreover, let $\mathcal{C}_{\mathcal{A}}$ be the set of atoms of $\mathcal{A}$, then for any $K \in \mathcal{C}_{\mathcal{A}}$ there is a unique $C \in \mathcal{C}_{\mathcal{E}}$ such that $K \subseteq C$. Given any $C \in \mathcal{C}_{\mathcal{E}}$ consider the set $\mathcal{K}_C = \{K \in \mathcal{C}_{\mathcal{A}} : K \subseteq C\}$, it follows that the sets $\mathcal{K}_C$ are a partition of $\mathcal{C}_{\mathcal{A}}$. Furthermore, consider a function $p'$ on $\mathcal{C}_{\mathcal{A}}$ defined in such a way that $p'(K) = \frac{p(C)}{\sharp(\mathcal{K}_C)}$ for any $K \subseteq\in \mathcal{K}_C$ with $\mathcal{K}_C$ finite and $p'(K_n) = \frac{p(C)}{2^n}$ for any $K_n \in \mathcal{K}_C$ with $\mathcal{K}_C$ countable (but not finite).

From strict positivity of $p$ on $\mathcal{C}_{\mathcal{E}}$, it follows the strict positivity of $p'$ on $\mathcal{C}_{\mathcal{A}}$. Moreover,

$$p(C) = \sum_{K \in \mathcal{K}_C} p'(K)$$

even when $C$ is obtained as a countable (but not finite) logical sum of atoms in $\mathcal{C}_{\mathcal{A}}$ and

$$\sum_{K \in \mathcal{C}_{\mathcal{A}}} p'(K) = \sum_{C \in \mathcal{C}_{\mathcal{E}}} \sum_{K \in \mathcal{K}_C} p'(K) = \sum_{C \in \mathcal{C}_{\mathcal{E}}} p(C)$$

that is not necessarily 1 when $\mathcal{E}$ is not finite and $p$ is not $\sigma$-additive.

Then, for any $A \in \mathcal{A}$ the function $p'$ on $\mathcal{C}_{\mathcal{A}}$ can be extended on $\mathcal{A}$ as follows: let $B \in \mathcal{E}$ be the greatest event, with respect to logical sum, contained in $A$ and let $K_A^B = \bigcup_{K \subseteq A \wedge B^c, K \in \mathcal{C}_{\mathcal{A}}} K$, then $A = B \vee K_A^B$, define $p'(A) = p(B) + \sum_{K \subseteq A \wedge B^c, K \in \mathcal{C}_{\mathcal{A}}} p'(K)$. Notice that when $A \in \mathcal{E}$ $(B = A)$ $p'(A) = p(A)$; moreover if $p$ is $\sigma$-additive, then $p'$ is obtained by $\sigma$-additivity from $p'$ on $\mathcal{C}_{\mathcal{A}}$ and so $p'$ is a $\sigma$-additive probability on $\mathcal{A}$.
We need to prove that even when $p$ is just finite additive, also $p'$ is a finite additive probability. For any set of pairwise incompatible events $A_1, ..., A_n \in \mathcal{A}$, there are the corresponding maximal (with respect to logical sum) events $B_1, ..., B_n, B \in \mathcal{E}$ contained in $A_1, ..., A_n, \vee_{i=1}^n A_i$, respectively, and the events $K_{A_i}^{B_i} = \bigcup_{K \subseteq A_i \wedge B_i^c, K \in \mathcal{C}_{\mathcal{A}}} K$ for $i = 1, ..., n$ and $K_A^B = \bigcup_{K \subseteq A \wedge B^c, K \in \mathcal{C}_{\mathcal{A}}} K$ with $\vee_{i=1}^n B_i \subseteq B$, $K_A^B \subseteq \vee_{i=1}^n K_{A_i}^{B_i}$. Then, $B = \vee_{i=1}^n (B_i \vee (K_{A_i}^{B_i} \wedge B))$ and $K_A^B = \vee_{i=1}^n (K_A^B \wedge K_{A_i}^{B_i})$, so

$$p'(A) = p'(\vee_{i=1}^n A_i) = p'(B) + p'(K_A^B) =$$

$$p'(\vee_{i=1}^n B_i) + p'(\vee_{i=1}^n (K_{A_i}^{B_i} \wedge B)) + p'(\vee_{i=1}^n (K_{A_i}^{B_i} \wedge K_A^B)) =$$

$$\sum_{i=1}^n p'(B_i) + \sum_{i=1}^n p'(K_{A_i}^{B_i} \wedge B) + \sum_{i=1}^n p'(K_{A_i}^{B_i} \wedge K_A^B) = \sum_{i=1}^n p'(A_i).$$

The probability $p'$ on $\mathcal{A}$ induces a positive comparative probability $\preccurlyeq'$ that extends $\preccurlyeq$ on $\mathcal{E}$.

□

The following example shows a case of the previous result:

**Example 2.** *Let $\mathcal{E}$ be an algebra generated by the set $\{C_1, ..., C_4\}$ of atoms with $C_i =$ "the number $i$ is drawn", for $i = 1, 2, 3$ and $C_4 =$ "a number greater or equal to 4 is drawn". We define $\preccurlyeq$ on $\mathcal{E}$ as induced by the probability on $\mathcal{E}$ such that $p(C_i) = 1/2^{1+i}$ for $i = 1, 2, 3$ and $p(C_4) = 9/16$. If we extend the relation on the algebra $\mathcal{A}$ of finite and co-finite subsets of the natural numbers, we could take that one generated by $p'(n) = 1/2^{1+n}$ for any $n \in \mathbb{N}$ and $p'(K) = 1 - p'(K^c)$ for any $K \subset \mathbb{N}$ co-finite .*

*Note that such $p'$ is a finitely additive but not $\sigma$-additive probability: in fact the sum over all the atoms $n \in \mathbb{N}$ is $1/2$.*

Coherence characterization can be maintained to sets of events $\mathcal{E}$ with arbitrary cardinality only for non-archimedean representability. In fact the following Theorem holds:

**Theorem 3.** *Let $\mathcal{E}$ be a set of events and $\preccurlyeq$ a positive comparative probability on $\mathcal{E}$, then the following are equivalent:*

**i)** *for every finite subset $\mathcal{F} \subseteq \mathcal{E}$ $(\mathcal{F}, \preccurlyeq_{|\mathcal{F}}) \cup (\phi \prec C_k)_{C_k \in \mathcal{C}_\mathcal{F}}$ is coherent;*

**ii)** *there exists a non-archimedean probability function $p^* :< \mathcal{E} >\rightarrow [0.1]^*$ strictly positive which represents $\preccurlyeq$.*

Proof: From Theorem 1 we have that for any finite $\mathcal{F} \subset \mathcal{E}$ there exists a strictly positive $p$ defined on the algebra generated by $\mathcal{E}$ which represents $\preccurlyeq$ restricted to $\mathcal{F} \cup \mathcal{C}_\mathcal{F}$. As already proved in [11][Th.5.1], there exists a strictly positive non-archimedean probability $p^*$ defined on the whole algebra $< \mathcal{E} >$ that represents $\preccurlyeq$. The explicit proof of strict positivity of $p^*$ derives directly from representability of a positive comparative probability (see again last rows of the proof of Th.5.1 in [11]).

□

An example of positive comparative probability, coherent on any finite set that is not representable by a strictly positive probability is the following:

**Example 3.** *Let $\mathcal{A}$ be the algebra of finite and co-finite subsets of $\mathbb{N}$ and $\preccurlyeq$ induced by cardinalities, i.e.:*

$$A \preccurlyeq B \Leftrightarrow \begin{cases} \sharp(A) \leq \sharp(B) & \text{if } A \text{ is finite} \\ \sharp(B^c) \leq \sharp(A^c) & \text{if } B \text{ is co-finite} \end{cases} .$$

*It is representable through the non-archimedean probability generated by $p^*(n) = \epsilon$, with $\epsilon$ any infinitesimal of $[0, 1]^*$, if $n \in \mathbb{N}$ and $p^*(B) = 1 - \sharp(B^c)\epsilon$ if $B$ is a co-finite. While it can be only weakly represented by a real valued probability since $n \sim m$ for all $n, m \in \mathbb{N}$ implies inevitably $p(n) = p(m) = 0$ and consequently $p(A) = p(B) = 0$ even if $A$ and $B$ are finite but with different cardinalities.*

# 4 Reference dependent comparative probabilities

As already stated in the motivations, we want explicitly face a dynamic context, i.e. situations where the Decision Maker has to express preferences conditioned to different information scenarios, i.e. to different events $H$ varying in a arbitrary set of alternatives $\mathcal{H}$. As before let us firstly focus on preferences reflected among conditional events.

Let $\mathcal{L}$ be a set of conditional events $\mathcal{L} = \{E_i|H_i\}_{i\in I}$ with the requirement that if $E_i|H_i \in \mathcal{L}$ then $\phi|H_i \in \mathcal{L}$. Denote with $\mathcal{H} = \{H_i : E_i|H_i \in \mathcal{L}\}$ the set of conditioning events and with $\mathcal{E} = \{E_i : E_i|H_i \in \mathcal{L}\}$ the set of the conditioned ones. Let $\mathcal{A}$ be the algebra spanned by $\mathcal{E} \cup \mathcal{H}$. In the following $\preccurlyeq = \bigcup\limits_{H\in\mathcal{H}} \{\preccurlyeq_H\}$ will be a partial binary relation defined for the couples of conditional events $E|H, F|H$ in $\mathcal{L}$ conditioned to the same event $H \in \mathcal{H}$.

In such a context it is natural to search for representability of $\preccurlyeq$ through conditional measures, taking in particular consideration negligible events even as conditioning ones. Anyhow, since we have seen that a strictly positive non-archimedean representability is permitted also with the presence of negligible events, if such measures are allowed by the Decision Maker representability can be guaranteed by rationality of a simpler unconditional relation derived from $\preccurlyeq$.This can be obtained as the following *projection* of $\preccurlyeq$:

**Definition 1.** *Let $\mathcal{L}^* \subset \mathcal{A}$ the set of events $\{E \wedge H : E|H \in \mathcal{L}\}$ and $\preccurlyeq^*$ the partial relation in $\mathcal{L}^*$ defined through*

$$E \wedge H \preccurlyeq^* F \wedge H \Leftrightarrow E|H \preccurlyeq F|H \quad . \tag{3}$$

Let us show how such projection $\preccurlyeq^*$ suffices to guarantee a non-archimedean representability of the original conditional preference relation:

**Theorem 4.** *Let $\preccurlyeq^*$ be a partial relation defined as in in Definition 1. Then the following statements are equivalent:*

**i)** *For any finite subset $\mathcal{F} \subseteq \mathcal{L}$ the relation $\preccurlyeq'$ defined on $\mathcal{F}' = \mathcal{F}^* \cup \mathcal{C}_{\mathcal{F}^*}$ by taking $(\mathcal{F}^*, \preccurlyeq^*) \cup \{\phi \prec \mathcal{C}_k\}_{\mathcal{C}_k \in \mathcal{C}_{\mathcal{F}^*}}$ is coherent;*

**ii)** *There exists a strictly positive non-archimedean probability $p^* : \mathcal{A} \times \mathcal{A}^0 \to [0,1]^*$ that represents $\preccurlyeq$ in $\mathcal{L}$.*

Proof: i) implies ii) since from previous Theorem 2 there exists a $p^* : \mathcal{A} \to [0,1]^*$ strictly positive that represents $\preccurlyeq^*$ in $\mathcal{L}^*$. Hence

$$E|H \preccurlyeq F|H \Leftrightarrow p^*(E \wedge H) \leq p^*(F \wedge H) \quad . \tag{4}$$

Since $p^*$ is strictly positive in $\mathcal{A}$, then $p^*(H) > 0$ and hence we have

$$p^*(E|H) = \frac{p^*(E \wedge H)}{p^*(H)} \leq p^*(F|H) = \frac{p^*(F \wedge H)}{p^*(H)} \quad . \tag{5}$$

The proof of the implication ii) ⇒ i) goes straightforward: in fact if $p^*(\cdot|\cdot)$ represents $\preccurlyeq$ and is strictly positive, then

$$p^*(E|H) \leq p^*(F|H) \Leftrightarrow p^*(E \wedge H) \leq p^*(F \wedge H) \tag{6}$$

and hence $\preccurlyeq^*$ must be coherent, this means, from previous Th.2, that i) must hold.

□

It is known, see e.g. [9], that from any $p^*(E|H) = \frac{p^*(EH)}{p^*(H)}$ with $p^*(\cdot) > 0$ it is possible to obtain a real valued full conditional probability on $\mathcal{A} \times \mathcal{A}^0$ by taking $p(E|H) = Re[p^*(E|H)]$ (with $Re[\cdot]$ the *real part* function). Anyhow relations induced on $\mathcal{A} \times \mathcal{A}^0$ from a non-archimedean conditional probability $p^*(\cdot|\cdot)$ are not the same of those by the corresponding $p(\cdot|\cdot)$, even if we limit ourselves to the $\preccurlyeq_H$, i.e. to compare events conditioned to the same reference events $H$.

We show how, for real valued probabilities, we can preserve the feature of distinguishing the different layers of admissibility among different scenarios. The following definition generalize the coherence condition for a dynamical setting:

**Definition 2.** *The partial relation $\preccurlyeq$ on $\mathcal{L}$ is* conditionally *coherent if for all $E_i|H_i \preccurlyeq F_i|H_i$ there exists $\delta_i \in [0,1]$, with $\delta_i > 0$ whenever $E_i|H_i \prec F_i|H_i$, such that for every $n \in \mathbb{N}$, $\lambda_i > 0$ and $E_i|H_i \preccurlyeq F_i|H_i$, $i = 1, \ldots, n$, we have*

$$\sup_{H^0} \sum_{i=1}^{n} \lambda_i(|F_i| - |E_i| - \delta_i)|H_i| \geq 0 \tag{7}$$

*with $H^0 = \bigcup_{i=1}^{n} H_i$.*

Note that if there is a single conditioning event, i.e. $\mathcal{H} = \{H\}$, then Definition 2 coincides with the so called *strong coherence* condition (sc) in [4, 6].

The following theorem shows that the previous rationality requirement is what is needed to have the representability of the preference relation through conditional probabilities:

**Theorem 5.** *Let $\preccurlyeq$ be a partial binary relation on $\mathcal{L}$. The following statements are equivalent:*

**i)** *$\preccurlyeq$ is conditionally coherent;*

**ii)** *there exists a coherent real valued conditional probability $p : \mathcal{L} \to [0,1]$ that represents $\preccurlyeq$.*

Proof: This proof is a particular case of the more general one already proved in [6]. In fact, for more general comparative conditional assessments where comparisons can be made also among different conditioning events, the following coherence condition (ccp) has been proved to be equivalent to the representability through a conditional probability:

**(ccp)** for all $E_i|H_i \preccurlyeq F_i|K_i$ there exists $\alpha_i, \beta_i \in [0,1]$, with $\alpha_i \leq \beta_i$ and $\alpha_i < \beta_i$ whenever $E_i|H_i \prec F_i|K_i$, such that for every $n \in \mathbb{N}$ and $\lambda_i, \lambda'_i \geq 0$ for every $E_i|H_i \preccurlyeq F_i|K_i$ we have

$$\sup_{H^0} \sum_{i=1}^{n} [\lambda_i(|F_i \wedge K_i| - \beta_i|K_i|) + \lambda'_i(\alpha_i|H_i| - |E_i \wedge H_i|)] \geq 0$$

with $H^0$ union of the conditioning events whose corresponding $\lambda_i$ or $\lambda'_i$ is positive.

Now we are dealing with a relation $\preccurlyeq$ that compares only events conditioned to the same event $H$. Hence if (ccp) holds then (7) is obtained by taking $\lambda_i = \lambda'_i$ and $\delta_i = \beta_i - \alpha_i$. Vice versa, suppose (7) holds. We will show that every single therm of the summation in (ccp) can be obtained and the non-negativity of the supremum maintained. Without loss of generality take $\lambda_i \leq \lambda'_i$ (the opposite works symmetrically), then the single therm in (7) is of the form

$$\lambda_i|F_i \wedge H_i| - \lambda_i|E_i \wedge H_i| - \lambda_i\delta_i|H_i| \tag{8}$$

for some specific $\delta_i \geq 0$. Moreover, since $\phi|H_i \preccurlyeq F_i|H_i$, there will exists a $\gamma_i \geq 0$ such that

$$(\lambda'_i - \lambda_i)|F_i \wedge H_i| - (\lambda'_i - \lambda_i)\gamma_i|H_i| \tag{9}$$

has a non-negative supremum over $H_i$. By adding (9) to (8) non-negativity of the supremum is maintained and a therm of those of (ccp) is obtained by taking $\alpha_i = \delta_i$ and $\beta_i = \delta_i + \gamma_i$. □

Note that if $\mathcal{E} = \mathcal{H} = \mathcal{A}$ is a *finite* algebra, then $\preccurlyeq$ defined on $\mathcal{A} \times \mathcal{A}^0$ and positive has a projection $\preccurlyeq^*$ as in Definition 1 coherent if and only if it is conditionally coherent. In fact, if $\preccurlyeq$ has a projection $\preccurlyeq^*$ coherent and since it contains $\preccurlyeq_\Omega$, from Theorem 1 $\preccurlyeq^*$ is strictly positive, coherent and complete on $\mathcal{A}$ if and only if it is representable through a strictly positive real valued (or equivalently non-archimedean valued) probability $p$ ($p^*$) hence for any couple $E_i|H_i \preccurlyeq F_i|H_i$ it holds $\frac{p^*(E_i \wedge H_i)}{p^*(Hi)} \leq \frac{p^*(F_i \wedge H_i)}{p^*(Hi)}$ as well as $\frac{p(E_i \wedge H_i)}{p(Hi)} \leq \frac{p(F_i \wedge H_i)}{p(Hi)}$. So we have the representability through a conditional probability that, by the previous Theorem5, is equivalent to the conditional coherence.

On the other hand, in the *infinite* case we have that a positive $\preccurlyeq$ can admit a representation through a non-archimedean probability $p^*$ in $[0,1]^*$, and hence it is coherent, but it can happen that it is not representable through a real valued probability $p$ in $[0,1]$. In fact $p$ comes from $p^*$ as $Re[p^*]$ and it could only *almost represent* and not represent $\preccurlyeq$. On the other hand, as already stated in the motivations, if we want to represent $\preccurlyeq$ on $\mathcal{A} \times \mathcal{A}^0$ through a real valued conditional probability $p(\cdot|\cdot)$, it is impossible to require all $E_i|H_i$ being more probable than $\phi$, since this surely cannot hold for $H_i = \Omega$. Consequently a binary relation $\preccurlyeq$ representable by a real valued conditional probability $p$ is necessarily less fine of an other relation representable by a non-archimedean conditional probability, except of course the finite case as proved earlier.

We have to stress that all previous coherence conditions do not ensure uniqueness of the representing probabilities. This reflects on the original relation $\preccurlyeq$ among acts, obtaining representability only by a family of conditional expected utilities, even if the set of acts is complete and the induced events form a $\sigma$-algebra. To have uniqueness we need of a condition similar to that of *fineness* and *tightness* used by Savage and in the following section we will provide it but paying attention to allow negligible conditioning events.

## 5 Rationality of conditional preferences among acts

Starting with a partial conditional preference there could be representability problem, not only about uniqueness of the conditional probability but also about the existence of utility functions. Anyhow, if the consequence space $\mathcal{X}$ is a subset of the reals, so that acts are simple random variables, comparisons could be represented by conditional previsions (expectations). In this case a rational condition has been already introduced in [14] and it is a generalization of the conditional coherence Definition 2 with simple random variables $X_i$ and $Y_i$ replacing events $E_i$ and $F_i$.

To avoid representability troubles, let us consider the set of acts $\mathcal{S}$ be rich enough to induce the set of derived events $\mathcal{A}$ to be a $\sigma$-algebra. Moreover, we limit to consider settings where the different scenarios $\mathcal{H}$ constitutes a countable additive class of events in $\mathcal{A} \setminus \{\phi\}$. Such limitation is quite usual and generally accepted in practical applications.

Hence we deal with a conditional comparative relation $\preccurlyeq = \bigcup\limits_{H\in\mathcal{H}} \{\preccurlyeq_H\}$, with $\preccurlyeq_H$ defined for every couple of single acts $f, g : \mathcal{A} \to \mathcal{X}$ and thought on the hypothesis that a specific scenario $H \in \mathcal{H}$ occurs.

In the description we will mainly follow the approach of [8], with the main difference that we allow for negligible conditioning events. In the following, constant acts $f \in \mathcal{S}$ will be identified with their unique consequence $x \in \mathcal{X}$. Consequently the comparative relation $\preccurlyeq$ induces also a comparative relation on $\mathcal{X}$ that we will continue to denote with the same symbol, as we made for events. For a generic event $A \in \mathcal{A}$ and acts $f, g \in \mathcal{S}$ we denote with

$$fAg = \begin{cases} f(\omega) & \omega \in A \\ g(\omega) & \omega \in A^c \end{cases}$$

the act that coincides with $f$ if $A$ occurs and with $g$ otherwise.

Rationality of a conditional preference relation $\preccurlyeq = \bigcup_{H\in\mathcal{H}} \preccurlyeq_H$ usually realizes with the existence of a utility function $u$ on the consequences space $\mathcal{X}$ and of at least a conditional probability $p(\cdot|\cdot)$ such that

$$f \preccurlyeq_A g \Leftrightarrow E_A(u(f)) \leq E_A(u(g))$$

with $E_A(\cdot)$ conditional expectation (i.e. expectation computed w.r.t. $p(\cdot|\cdot)$).

We need now to introduced conditioning events equivalent to the impossible one but whose conditioning is not trivial, hence *not null* even *negligible*. Negligibility can be formalized, by borrowing the similar definition given in [10], with the following definition:

**Definition 3.** *An event $H \in \mathcal{H}$ is* negligible *with respect to an other event $K \in \mathcal{H}$, with $K \supset H$, if $\forall f, g, h, h' \in \mathcal{S}$*

$$fH^ch \preccurlyeq_K gH^ch' \Leftrightarrow f \preccurlyeq_K g \quad .$$

It is easy to show that this definition capture the full meaning of negligibility. In fact the following lemma holds:

**Lemma 1.** *If $H$ is negligible w.r.t. $K$ then $\phi \sim_K H$.*

Proof: The impossible event $\phi$ can be identified with the binary act $0H^c0$, while $H$ with the binary act $0H^c1$. Since the relation $0 \sim_K 0$ always holds, then negligibility of $H$ w.r.t. $K$ implies $0H^c0 \sim_K 0H^c1$ that coincides with the thesis $\phi \sim_K H$.

□

Negligible events can be organized in a hierarchy, forming the so called different *zero layers*:

**Definition 4.** $H^0 = \Omega \quad , \quad H^\alpha = \bigvee_{\phi \sim_{H(\alpha-1)} H} H$ *for* $\alpha \geq 1$.

Note that such $H^\alpha$ are maximal events in $\mathcal{H}$ such that $\phi \sim_{H^{(\alpha-1)}} H^\alpha$, while not-negligible events $K$ w.r.t. a zero-layer $H^\alpha$ can be identified by those elements of $\mathcal{H}$ such that $\phi \prec_{H^\alpha} K$.

Let us introduce axioms that we bring us to the last representability result:

**Ax1** For each $H \in \mathcal{H}$ $\preccurlyeq_H$ is a complete and transitive binary relation on $\mathcal{S}$;

**Ax2** For each $K \in \mathcal{H}$:
$g \preccurlyeq_K f \Rightarrow g \preccurlyeq_\Omega fKg$
if $\phi \prec_{H^\alpha} K$ then $g \prec_K f \Rightarrow g \prec_{H^\alpha} fKg$;

**Ax3** For each $H \in \mathcal{H}$ and $x, y \in \mathcal{X}$
$y \preccurlyeq_\Omega x \Leftrightarrow y \preccurlyeq_H x$;

**Ax4** For every $x, x', y, y' \in \mathcal{X}$
if $y \prec_\Omega x$ and $y' \prec_\Omega x'$ then for any $A, B \in \mathcal{A}$
$xBy \preccurlyeq_{H^\alpha} xAy \Leftrightarrow x'By' \preccurlyeq_{H^\alpha} x'Ay' \quad \forall \alpha \geq 0$;

**Ax5** $\exists x, y \in \mathcal{X}$ s.t. $y \prec_\Omega x$;

**Ax6** For every $x \in \mathcal{X}$ and $f, g \in \mathcal{S}$ with $g \prec_{H^\alpha} f$ there exists a finite partition $\mathcal{H}_\alpha$ of $H^\alpha$ s.t. for every $H \in \mathcal{H}_\alpha$ we have:

**i)** $g \prec_{H^\alpha} xHf$;

**ii)** $xHg \prec_{H^\alpha} f$.

**Ax7** For every $H \in \mathcal{H}$ and $f, g \in \mathcal{S}$
$f(\omega) = g(\omega)\ \forall\{\omega\} \subset H$ then $f \sim_H g$.

**Theorem 6.** *If the set $\mathcal{H}$ is finite the following conditions are equivalent:*

**(i)** $\preccurlyeq = \bigcup_{H \in \mathcal{H}} \{\preccurlyeq_H\}$ *is representable through a conditional expected utility;*

**(ii)** $\preccurlyeq = \bigcup_{H \in \mathcal{H}} \{\preccurlyeq_H\}$ *satisfies axioms Ax1–Ax7.*

Proof:The proof that $(i) \Rightarrow (ii)$ is straightforward.

We now prove $(ii) \Rightarrow (i)$. By hypothesis $\preccurlyeq_\Omega$ satisfies Savage's axioms and so it is representable by an expected utility EU. Then, there is a unique (up to a positive affine transformation) utility on $\mathcal{X}$ and a unique probability on $\mathcal{A}$.

For any $H \in \mathcal{H}$ such that $\emptyset \prec_\Omega H$, the relation $\preceq_H$ is representable by a $E_H$ with

$$E_H(f) = \frac{E(fI_H)}{p(H)}$$

with $I_H$ the binary act $1H0$. Now consider the maximal negligible event $H^1 \in \mathcal{H}$ with respect to $\Omega$. This event $H^1$ plays the same role that $\Omega$ in the previous step and, since $\preceq_{H^1}$ satisfies the Savage's axioms it is representable by an expected utility EU where the utility is the same that in the previous step and the probability $p_1 = p_{H^1}$ with $p_1(H^1) = 1$.

The procedure continues considering events $H^\alpha$ (with $\alpha$ less or equal to the cardinality of $\mathcal{H}$) and $p_\alpha$ as in the previous step.

Note that the probabilities $p_\alpha$ are the agreeing class of a conditional probability $p'$ on $\mathcal{A} \times \mathcal{H}$ (see [5]):

for any $E|H \in \mathcal{A} \times \mathcal{H}$

$$p'(E|H) = \frac{p_\alpha(E \wedge H)}{p_\alpha(H)}$$

with $p_\alpha(H) = p'(H|H^\alpha)$ and $\alpha$ the maximum index such that $H \subseteq H^\alpha$.

□

## References

[1] S. Bernardi, G. Coletti, A Rational Conditional Utility Model in a Coherent Framework, Lecture Notes in Computer Science, 2001, Volume 2143, Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Pages 108-119.

[2] L. Blume, A. Brandenburger, E. Dekel, Lexicographic Probabilities and Choice Under Uncertainty, Econometrica, 59(1), 61–79, 1991.

[3] L. Blume, D. Easley, J.Y. Halpern, Constructive Decision Theory, `arXiv:0906.4316v1`, 2009.

[4] G. Coletti, Coherent qualitative probability, Journal of Mathematical Psychology, 34, 297-310, 1990.

[5] G. Coletti, Coherent numerical and ordinal Probabilistic assessments, IEEE Trasactions on Systems, Man, and Cybernetics 24(12), 1747-1754, 1994.

[6] G. Coletti, B. Vantaggi, Representability of Ordinal Relations on a Set of Conditional Events, *Theory and Decision*, 60, 137 – 174, 2006.

[7] B. de Finetti, Teoria della Probabilit'a. Torino: Einaudi, 1970 (Engl. transl. Theory of probability, London: Wiley & Sons, 1974).

[8] P. Ghirardato, Revisiting Savage in a Conditional World, Economic Therory, 20, 83–92, 2002.

[9] P.H. Krauss, Representation of Conditional Probability Measures on Boolean Algebras, Acta Mathematica Academiae Sceintiarum Hungaricae, 19 (3–4), 229–241, 1968.

[10] D.Lehmann, Generalized Qualitative Probability: Savage Revisited, *Proc. UAI'96*, 381–388, 1996.

[11] L. Narens, Minimal Conditions for Additive Conjoint Measurement and Qualitative Probability, Journal of Mathematical Psychology, 11,404–430, 1974.

[12] A. Robinson, Non-Standard Analysis. Amsterdam: North Holland, 1966.

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

[14] B. Vantaggi, Incomplete preferences on conditional random quantities: representability by conditional previsions, Mathematical Social Sciences 60, 104–112, 2010.

# A Syntactical Comparison Between Probabilistic and Possibilistic Likelihood

**Giulianella Coletti and Davide Petturiti**
Dept. of Matematica e Informatica
University of Perugia
via Vanvitelli 1, 06123 Perugia, Italy
{coletti,davide.petturiti}@dmi.unipg.it

**Barbara Vantaggi**
Dept. of S.B.A.I.
"La Sapienza" University of Rome
via Scarpa 16, 00161 Rome, Italy
barbara.vantaggi@sbai.uniroma1.it

**Abstract**

We study likelihood functions as point functions and then as set functions. In particular, we characterize likelihood functions consistent with a conditional possibility and we provide a comparison with probabilistic likelihood functions from a formal point of view.

## 1 Introduction

In some applications the available information is related to different sets of events moreover different uncertainty frameworks are involved. Thus there is a growing interest to study more flexible modeling able to manage uncertainty, vagueness and partial information.

Here we focus on inferential processes in which the available information is expressed by probability and possibility (for analogous issues see [10, 8, 14, 13]).

An ensuing problem is to maintain coherence with respect to a (probabilistic or possibilistic) framework by using all the available information, so we need to check when a probabilistic likelihood has the same properties of a possibilistic one. For this aim we study the extension of a possibilistic likelihood function to events of a partition less fine than that in which it is defined.

By referring to conditional probabilities [6] and $T$-conditional possibilities (where $T$ stands for any triangular norm) [1], a (probabilistic or possibilistic) likelihood function can be regarded as an assessment on a class of conditional events $\{E|H_i\}$, with $\{H_i\}$ a finite partition, satisfying a trivial condition. We characterize its coherence and the coherent extensions on the conditional events $\{E|K\}$, with $K$ belonging to the additive set $\mathcal{H}$ spanned by the $H_i$'s.

This will provide a comparison of properties characterizing probabilistic and possibilistic likelihood functions both as point and set functions.

Our analysis shows that there is no particular property distinguishing probabilistic point likelihood from possibilistic one, i.e., the same assessment is always coherent in both frameworks.

On the other hand, for what concerns aggregated likelihood an interesting difference is that in probabilistic setting no kind of monotonicity is required for the coherence, while in the possibilistic one there is a local form of monotonicity (i.e., the possibilistic aggregated likelihood is monotone on the elements of a suitable partition $\mathbf{H}$ of the additive set $\mathcal{H}$).

## 2 Coherent conditional possibility assessments

Coherence is well known in probability theory [6], moreover this notion has been studied also in other frameworks such as possibility theory (see [4]) by referring to $T$-conditional possibilities (with $T$ a triangular norm) [1, 4].

**Definition 1.** *Let $T$ be any $t$-norm. A function $\Pi : \mathcal{B} \times \mathcal{H} \to [0,1]$ is a $T$-conditional possibility if it satisfies the following properties:*

*(i)* $\Pi(E|H) = \Pi(E \wedge H|H)$*, for every* $E \in \mathcal{B}$ *and* $H \in \mathcal{H}$*;*

*(ii)* $\Pi(\cdot|H)$ *is a possibility on* $\mathcal{B}$*, for any* $H \in \mathcal{H}$*;*

*(iii)* $\Pi(E \wedge F|H) = T(\Pi(E|H), \Pi(F|E \wedge H))$*, for any* $H, E \wedge H \in \mathcal{H}$ *and* $E, F \in \mathcal{B}$*.*

This definition generalize some other definitions present in literature (see, for instance, [5, 7, 9, 11, 12]).

In the following we deal with strict $t$-norms or minimum.

In analogy with conditional probability, an assessment $\Pi$ on an arbitrary set $\mathcal{G}$ of conditional events is a coherent $T$-conditional possibility if (and only if) $\Pi$ is a restriction of a $T$-conditional possibility (in the sense of Definition 1) defined on $\mathcal{B} \times \mathcal{H} \supseteq \mathcal{G}$.

We recall a characterization of a finite coherent $T$-conditional possibility assessment given in [4].

**Theorem 1.** *Let $\mathcal{G} = \{E_1|H_1, \dots, E_n|H_n\}$ be an arbitrary set of conditional events, and $\mathcal{C}_0$ and $\mathcal{B}$ denote the set of atoms and the algebra spanned by $\{E_1, H_1, \dots, E_n, H_n\}$, respectively.*
*For a real function $\Pi : \mathcal{G} \to [0,1]$, the following statements are equivalent:*

*a)* $\Pi$ *is a coherent $T$-conditional possibility assessment on* $\mathcal{E}$*;*

*b) there exists a sequence of compatible systems $\mathcal{S}^{\Pi}_{\alpha}$ ($\alpha = 0, \dots, k$), with unknowns $x_r^{\alpha} \geq 0$ for $C_r \in \mathcal{C}_{\alpha}$,*

$$\mathcal{S}^{\Pi}_{\alpha} = \begin{cases} \max\limits_{C_r \subseteq E_i \wedge H_i} x_r^{\alpha} = T\left(\Pi(E_i|H_i), \max\limits_{C_r \subseteq H_i} x_r^{\alpha}\right) & \max\limits_{C_r \subseteq H_i} \mathbf{x}_r^{\alpha-1} < 1 \\ x_r^{\alpha} \geq \mathbf{x}_r^{\alpha-1} & \textit{if } C_r \in \mathcal{C}_{\alpha} \\ \mathbf{x}_r^{\alpha-1} = T\left(x_r^{\alpha}, \max\limits_{C_j \in \mathcal{C}_{\alpha}} \mathbf{x}_j^{\alpha-1}\right) & \textit{if } C_r \in \mathcal{C}_{\alpha} \\ \max\limits_{C_r \in \mathcal{C}_{\alpha}} x_r = 1 & \end{cases}$$

*with $\alpha = 0, \ldots, k$, where $\mathbf{x}^\alpha$ (with $r$-th component $\mathbf{x}_r^\alpha$) is the solution of $\mathcal{S}_\alpha^\Pi$ and $\mathcal{C}_\alpha = \{C_r \in \mathcal{C}_{\alpha-1} : \mathbf{x}_r^{\alpha-1} < 1\}$, moreover $\mathbf{x}_r^{-1} = 0$ for any $C_r \in \mathcal{C}_0$.*

**Remark**: Any sequence of solutions of $\{\mathcal{S}_\alpha^\Pi\}$ is related to a class $\mathcal{P} = \{\Pi_0, \ldots, \Pi_k\}$ of possibilities agreeing with the given assessment. The class is not necessarily unique and as shown in [4] any class $\mathcal{P}$ induces a unique full $T$-conditional possibility $\Pi(\cdot|\cdot)$ extending the given assessment.

We recall that a coherent (conditional) probability or possibility assessment can be extended to any new (conditional) event and the coherent extension in both cases lays on a closed interval ([3, 4]).

## 3 Likelihood: as point function and set function

This section is devoted to a comparative analysis of likelihood in probabilistic and possibilistic framework.

**Theorem 2.** *Let $\mathcal{L} = \{H_1, \ldots, H_n\}$ be a finite partition of $\Omega$ and $E$ an event. For every function $f : \{E\} \times \mathcal{L} \to [0,1]$ satisfying condition*

*(L1) $f(E|H_i) = 0$ if $E \wedge H_i = \emptyset$ and $f(E|H_i) = 1$ if $H_i \subseteq E$*

*the following statements hold:*

*i) $f$ is a coherent conditional probability;*

*ii) $f$ is a coherent $T$-conditional possibility (for every t-norm $T$).*

*Proof.* Condition *i)* has been proved in [3]. We prove *ii)*. From condition *b)* of Theorem 1, by the incompatibility of the events $H_i$, the equations in system $\mathcal{S}_0^\Pi$ have different unknowns (each of them is linked only with the last equation), and so the system $\mathcal{S}_0^\Pi$ admits a solution assigning possibility 1 to each conditioning event $H_i$. Then, the assessment $f$ is a coherent $T$-conditional possibility. □

The above result shows a common feature between probabilistic and possibilistic (point) likelihood, so this allows to regard, from a syntactical point of view, a probabilistic likelihood as a possibilistic one and vice versa. Moreover, it emphasizes that no significant property characterizes likelihood as point function, so in the sequel we call *likelihood function* any function $f : \{E\} \times \mathcal{L} \to [0,1]$, with $\mathcal{L} = \{H_1, \ldots, H_n\}$ a finite partition of $\Omega$, satisfying condition *(L1)*.

Our aim now is to study the properties of *aggregated likelihood functions*, that is all the coherent extensions $g$ of the assessment $\{f(E|H_i) : H_i \in \mathcal{L}\}$ to the events $E|K$, with $K$ belonging to the additive set $\mathcal{H} = \langle \mathcal{L} \rangle \setminus \{\emptyset\}$ spanned by $\mathcal{L}$.

The interest derives by inferential problems in which the available information consists of a (probabilistic or possibilistic) "prior" on a partition $\{K_j\}$ and a likelihood related to the events of another partition refining the previous one.

In what follows $g : \{E\} \times \mathcal{H} \to [0,1]$ denotes a function such that the restriction $g_{|\{E\}\times\mathcal{L}}$ of $g$ to $\{E\} \times \mathcal{L}$ coincides with $f$, while $\mathcal{B} = \langle \{E\} \cup \mathcal{H} \rangle$ is the Boolean algebra generated by $\{E\} \cup \mathcal{H}$.

**Theorem 3.** *If $g$ is either a coherent conditional probability or a coherent $T$-conditional possibility, then the following condition holds for every $K \in \mathcal{H}$:*

*(L2)* $\min_{H_i \subseteq K} f(E|H_i) \leq g(E|K) \leq \max_{H_i \subseteq K} f(E|H_i)$.

*Proof.* For coherent conditional probability, condition *(L2)* is proved in [3].

Let $g(E|\cdot)$ be a coherent $T$-conditional possibility assessment, then there is an extension $\Pi(\cdot|\cdot)$ on $\mathcal{B} \times \mathcal{H}$. For every $K \in \mathcal{H}$ it is

$$\Pi(E|K) = \max_{H_i \subseteq K} T(\Pi(E|H_i), \Pi(H_i|K)) \leq \max_{H_i \subseteq K} \Pi(E|H_i).$$

Moreover, taking $\beta = \min_{H_i \subseteq K} \Pi(E|H_i)$

$$\begin{aligned} \Pi(E|K) &= \max_{H_i \subseteq K} T\left(\Pi(E|H_i), \Pi(H_i|K)\right) \geq \max_{H_i \subseteq K} T\left(\beta, \Pi(H_i|K)\right) \\ &= T\left(\beta, \max_{H_i \subseteq K} \Pi(H_i|K)\right) = T\left(\beta, 1\right) = \beta\,. \end{aligned}$$

□

Condition *(L2)* implies that both probabilistic and possibilistic aggregated likelihood are monotone, with respect to $\subseteq$, only if the extension is obtained, for every $K$, as $\max_{H_i \subseteq K} f(E|H_i)$. In this case the aggregated likelihood is a capacity.

Similarly they are anti-monotone if and only if their extensions are obtained as $\min_{H_i \subseteq K} f(E|H_i)$.

The next Theorem 4 assures the coherence of the extensions, obtained by using operators max and min, in both probabilistic and possibilistic frameworks.

**Theorem 4.** *If for all $K_1, K_2 \in \mathcal{H}$*

$$g(E|K_1 \vee K_2) = \max\{g(E|K_1), g(E|K_2)\} \tag{1}$$

$$g^*(E|K_1 \vee K_2) = \min\{g^*(E|K_1), g^*(E|K_2)\} \tag{2}$$

*then the following statements hold:*

*i) $g$ and $g^*$ are coherent conditional probabilities;*

*ii) $g$ and $g^*$ are coherent $T$-conditional possibilities (for every $t$-norm $T$).*

*Proof.* The proof of *i)* is in [3]. Concerning *ii)* to prove the statement for $g$ consider as a solution of system $\mathcal{S}_0^{\Pi}$ in Theorem 1 the possibility assigning $\Pi_0(E \wedge H_i) = f(E|H_i)$ and $\Pi_0(H_i) = 1$, for any $H_i \in \mathcal{L}$. Regarding $g^*$ suppose without loss of generality $f(E|H_1) \leq \cdots \leq f(E|H_n)$. Then it is sufficient to consider the class $\mathcal{P} = \{\Pi_0, \ldots, \Pi_{n-1}\}$ of possibilities on $\mathcal{B}$ such that $\Pi_{i-1}(E \wedge H_i) = f(E|H_i)$, $\Pi_{i-1}(H_i) = 1$ and $\Pi_{i-1}(E) = 0$ for all $E \in \mathcal{B}$ such that $E \wedge H_i = \emptyset$, with $i = 1, \ldots, n$. □

**Remark**: Theorem 4 states that $g(E|\cdot)$ is a coherent $T$-conditional possibility, but it is not necessarily a normalized possibility even if

$$g(E|K_1 \vee K_2) = \max\{g(E|K_1), g(E|K_2)\}$$

for every $K_1, K_2 \in \mathcal{H}$ since $g(E|\Omega)$ can be strictly less than 1.

Actually, $g(E|\Omega)$ is 1 if and only if there is an event $H_i$ with $f(E|H_i) = 1$: this condition could seem natural since it claims the existence of an event $H_i$ supporting the evidence $E$.

It is easy to see from characterization theorems, that, if we extend a likelihood function $f$ only to one new event $K$ obtained by disjunction of some elements of $\mathcal{L}$, then condition *(L2)* is not only necessary but also sufficient for the coherence of the extension with both conditional probability and possibility. Nevertheless if we extend to more than one event of the additive set $\mathcal{H}$, then condition *(L2)* is no more sufficient:

**Example 1.** *Consider the partition $\mathcal{L} = \{H_1, H_2, H_3\}$ together with the logically independent event $E$ and the likelihood assessment*

$$f(E|H_1) = \frac{4}{5},\ f(E|H_2) = \frac{2}{5},\ f(E|H_3) = \frac{1}{5}.$$

*Let $C_i = E \wedge H_i$, $C_{i+3} = E^c \wedge H_i$, $i = 1, 2, 3$ be the atoms spanned by $\{E\} \cup \mathcal{L}$. The following extension $g$ of $f$*

$$g(E|H_1 \vee H_2) = \frac{3}{5},\ g(E|H_1 \vee H_3) = \frac{1}{2},\ g(E|H_2 \vee H_3) = \frac{3}{10},\ g(E|\Omega) = \frac{2}{5},$$

*is obtained taking the median of the relevant $f(E|H_i)$'s (for an even number of $H_i$'s, the arithmetic mean of the two median values is taken).*

*Clearly, $g$ satisfies condition (L2), anyway it is not coherent with a conditional probability since the following system with unknowns $x_r^0 \geq 0$, $r = 1, \dots, 6$,*

$$\mathcal{S}_0^P = \begin{cases} x_1^0 = \frac{4}{5}(x_1^0 + x_4^0) \\ x_2^0 = \frac{2}{5}(x_2^0 + x_5^0) \\ x_3^0 = \frac{1}{5}(x_3^0 + x_6^0) \\ x_1^0 + x_2^0 = \frac{3}{5}(x_1^0 + x_2^0 + x_4^0 + x_5^0) \\ x_1^0 + x_3^0 = \frac{1}{2}(x_1^0 + x_3^0 + x_4^0 + x_6^0) \\ x_2^0 + x_3^0 = \frac{3}{10}(x_2^0 + x_3^0 + x_5^0 + x_6^0) \\ x_1^0 + x_2^0 + x_3^0 = \frac{2}{5}(x_1^0 + x_2^0 + x_3^0 + x_4^0 + x_5^0 + x_6^0) \\ x_1^0 + x_2^0 + x_3^0 + x_4^0 + x_5^0 + x_6^0 = 1 \end{cases}$$

*is not compatible. Indeed, the subsystem obtained removing the seventh equation has the unique solution $x_1^0 = x_6^0 = \frac{4}{15}$, $x_2^0 = \frac{2}{15}$, $x_3^0 = x_4^0 = \frac{1}{15}$, $x_5^0 = \frac{1}{2}$, which contradicts the seventh equation.*

*Moreover, $g$ is not coherent with a $T$-conditional possibility (with $T$ an arbitrary* *t-norm): the following system with unknowns* $x_r^0 \geq 0$, $r = 1, \dots, 6$,

$$\mathcal{S}_0^{\Pi} = \begin{cases} y_1^0 = T\left(\frac{4}{5}, \max\{x_1^0, x_4^0\}\right) \\ x_2^0 = T\left(\frac{2}{5}, \max\{x_2^0, x_5^0\}\right) \\ x_3^0 = T\left(\frac{1}{5}, \max\{x_3^0, x_6^0\}\right) \\ \max\{x_1^0, x_2^0\} = T\left(\frac{3}{5}, \max\{x_1^0, x_2^0, x_4^0, x_5^0\}\right) \\ \max\{x_1^0, x_3^0\} = T\left(\frac{1}{2}, \max\{x_1^0, x_3^0, x_4^0, x_6^0\}\right) \\ \max\{x_2^0, x_3^0\} = T\left(\frac{3}{10}, \max\{x_2^0, x_3^0, x_5^0, x_6^0\}\right) \\ \max\{x_1^0, x_2^0, x_3^0\} = T\left(\frac{2}{5}, \max\{x_1^0, x_2^0, x_3^0, x_4^0, x_5^0, x_6^0\}\right) \\ \max\{x_1^0, x_2^0, x_3^0, x_4^0, x_5^0, x_6^0\} = 1 \end{cases}$$

*is not compatible. In fact, only the unknowns* $x_4^0$, $x_5^0$ *and* $x_6^0$ *can assume value* 1. *Assigning* $x_4^0 = 1$, *the first equation implies* $x_1^0 = \frac{4}{5}$, *but this contradicts the fourth equation. Analogously, putting* $x_5^0 = 1$, *from the second equation it follows* $x_2^0 = \frac{2}{5}$, *but this contradicts the sixth equation. By assigning* $x_6^0 = 1$, *the third and fifth equations imply* $x_3^0 = \frac{1}{5}$ *and* $x_1^0 = \frac{1}{2}$, *which contradicts the seventh equation.*

Condition *(L2)* is sufficient for the coherence of extensions when also the new events are mutually incompatible and exhaustive. For any set of events $\mathcal{E}$ we denote by $\langle \mathcal{E} \rangle$ the algebra spanned by $\mathcal{E}$.

**Theorem 5.** *Let* $f : \{E\} \times \mathcal{L} \to [0,1]$ *be a likelihood function and* $\mathcal{K} = \{K_h\}$, $(h = 1, \dots, r)$ *a partition of* $\Omega$ *whose elements are contained in* $\langle \mathcal{L} \rangle$. *Then, for the function* $g : \{E\} \times (\mathcal{L} \cup \mathcal{K}) \to [0,1]$ *extending* $f$, *the following statements hold:*

*i)* *g is a coherent conditional probability if and only if condition (L2) is satisfied for every* $K_h \in \mathcal{K}$;

*ii)* *g is a coherent* $T$*-conditional possibility (for every t-norm* $T$*) if and only if condition (L2) is satisfied for every* $K_h \in \mathcal{K}$.

*Proof.* The proof of both conditions *i)* and *ii)* is based on the relevant characterization in terms of solvability of the class of systems. In this case the sets of atoms related to the events $K_h$ are disjoint and so the relevant equations are independent. Then every system has a solution if and only if each equation has solution. Therefore every system has solution if and only if *(L2)* holds. □

## 3.1 Scale monotonicity

In order to compare possibilistic and probabilistic aggregated likelihood functions we introduce the notion of *scale* and then a relevant local form of monotonicity.

**Definition 2.** *Let* $\mathcal{L} = \{H_1, \dots, H_n\}$ *be a finite partition of* $\Omega$ *and* $\mathcal{H} = \langle \mathcal{L} \rangle \setminus \{\emptyset\}$ *the additive set spanned by* $\mathcal{L}$. *A scale of* $\mathcal{H}$ *is every partition* $\mathbf{H} = \{\mathcal{H}^0, \dots, \mathcal{H}^k\}$ *of* $\mathcal{H}$, *such that each* $\mathcal{H}^\alpha$ $(\alpha = 0, \dots, k)$ *is an additive set containing at least an event* $H_i \in \mathcal{L}$ *and every* $H \supseteq H_i$, *with* $H \notin \mathcal{H}^\gamma$ *with* $\gamma < \alpha$.

The next result shows that every $T$-conditional possibility $\Pi(\cdot|\cdot)$ on $\mathcal{B} \times \mathcal{H}$ induces a suitable scale of $\mathcal{H}$.

**Lemma 1.** *Let $\mathcal{L} = \{H_1, \dots, H_n\}$ be a finite partition of $\Omega$, $\mathcal{H} = \langle\mathcal{L}\rangle \setminus \{\emptyset\}$ the additive set spanned by $\mathcal{L}$, $\mathcal{B} = \langle\{E\} \cup \mathcal{H}\rangle$ and $\Pi : \mathcal{B} \times \mathcal{H} \to [0,1]$ a $T$-conditional possibility. Then, there exists a scale $\mathbf{H} = \{\mathcal{H}^0, \dots, \mathcal{H}^k\}$ of $\mathcal{H}$ such that for $\alpha = 0, \dots, k$, $\Pi(H|H_0^\alpha) = 1$ for every $H \in \mathcal{H}^\alpha$, where $H_0^\alpha = \bigvee\limits_{H_i \in \bigcup_{\gamma \geq \alpha} \mathcal{H}^\gamma} H_i$.*

*Proof.* Define $H_0^0 = \Omega$ and $\mathcal{H}^0 = \{H \in \mathcal{H} \ : \ \Pi(H|H_0^0) = 1\}$. For $\alpha = 1, \dots, k$, put $H_0^\alpha = \bigvee_{H_i \notin \bigcup_{\gamma < \alpha} \mathcal{H}^\gamma} H_i$ and $\mathcal{H}^\alpha = \{H \in \mathcal{H} \ : \ \Pi(H|H_0^\alpha) = 1\}$. The class $\mathbf{H} = \{\mathcal{H}^0, \dots, \mathcal{H}^k\}$ is a partition of $\mathcal{H}$ and each $\mathcal{H}^\alpha$ is an additive set, and since $\Pi(\cdot|H_0^\alpha)$ is a possibility, there exists $H_i \in \mathcal{L}$ such that $\Pi(H_i|H_0^\alpha) = 1 = \Pi(H|H_0^\alpha)$ for any $H \supseteq H_i$. Therefore $\mathbf{H}$ is a scale of $\mathcal{H}$. □

**Definition 3.** *Let $\mathcal{L} = \{H_1, \dots, H_n\}$ be a finite partition of $\Omega$ and $\mathcal{H} = \langle\mathcal{L}\rangle \setminus \{\emptyset\}$. A function $\varphi : \mathcal{H} \to [0,1]$ is said to be $\mathbf{H}$-scale monotone with respect to the scale $\mathbf{H} = \{\mathcal{H}^0, \dots, \mathcal{H}^k\}$ of $\mathcal{H}$ if $\varphi_{|\mathcal{H}^\alpha}$ is monotone with respect to implication $\subseteq$ for $\alpha = 0, \dots, k$.*

**Example 2.** *Let $\mathcal{L} = \{H_1, H_2, H_3\}$ be a partition of the sure event and denote by $\mathcal{H}$ the additive set spanned by $\mathcal{L}$. It is easy to prove that the set $\mathbf{H} = \{\mathcal{H}^0, \mathcal{H}^1\}$ with $\mathcal{H}^0 = \{H_2, H_1 \vee H_2, H_2 \vee H_3, \Omega\}$ and $\mathcal{H}^1 = \{H_1, H_3, H_1 \vee H_3\}$ is a scale.*

*Consider now the assessment $\varphi(H_1) = 0.3$, $\varphi(H_2) = 0.5$, $\varphi(H_3) = 0.8$, $\varphi(H_1 \vee H_2) = 0.5$, $\varphi(H_1 \vee H_3) = 0.8$, $\varphi(H_2 \vee H_3) = 0.7$ and $\varphi(\Omega) = 0.75$.*

*For every $K \in \mathcal{H}$, condition (L2) holds and $\varphi(\cdot)$ is $\mathbf{H}$-scale monotone.*

The following theorem shows that any coherent $T$-conditional possibility $g(E|\cdot)$ is $\mathbf{H}$-scale monotone with respect to some scale $\mathbf{H}$ of $\mathcal{H}$.

**Theorem 6.** *If $g$ is a coherent $T$-conditional possibility on $\{E\} \times \mathcal{H}$, then there exists at least one scale $\mathbf{H} = \{\mathcal{H}^0, \dots, \mathcal{H}^k\}$ of $\mathcal{H}$ such that $g(E|\cdot)$ is $\mathbf{H}$-scale monotone.*

*Proof.* Let $\Pi(\cdot|\cdot)$ be a $T$-conditional possibility extending $g(E|\cdot)$ on $\mathcal{B} \times \mathcal{H}$. Consider the scale $\mathbf{H} = \{\mathcal{H}^0, \dots, \mathcal{H}^k\}$ of $\mathcal{H}$ satisfying condition of Lemma 1, for all $\alpha = 0, \dots, k$, if $H, H \vee K \in \mathcal{H}^\alpha$, then $\Pi(H|H_0^\alpha) = \Pi(H \vee K|H_0^\alpha) = \Pi(H|H \vee K) = 1$. From that

$$\begin{aligned}\Pi(E|H \vee K) &= \max\{T(\Pi(H|H \vee K), \Pi(E|H)), T(\Pi(K|H \vee K), \Pi(E|K))\} \\ &= \max\{\Pi(E|H), T(\Pi(K|H \vee K), \Pi(E|K))\} \geq \Pi(E|H).\end{aligned}$$

□

Notice that for coherent conditional probabilities a similar result does not hold as the following example shows.

**Example 3.** *Let $\mathcal{L} = \{H_1, H_2, H_3\}$ be a partition of the sure event, $E$ an event logically independent of the events in $\mathcal{L}$ and consider the likelihood assessment: $P(E|H_1) = \frac{1}{2}$, $P(E|H_2) = \frac{1}{4}$ and $P(E|H_3) = \frac{1}{8}$.*

*A coherent extension of the above likelihood (obtained by giving $P(H_i) = \frac{1}{3}$, for $i = 1, 2, 3$) is*

$$P(E|H_1 \vee H_2) = \frac{3}{8}, P(E|H_1 \vee H_3) = \frac{5}{16}, P(E|H_2 \vee H_3) = \frac{3}{16}, P(E|\Omega) = \frac{7}{24}.$$

*This aggregated likelihood is not scale monotone. In fact, the set $\mathcal{H}^0$ would contain an element of $\mathcal{L}$ and all its supersets and this is not possible since $P(E|H_1) > P(E|H_1 \vee H_2)$, $P(E|H_2) > P(E|H_2 \vee H_3)$ and $P(E|H_1 \vee H_3) > P(E|H_1 \vee H_2 \vee H_3)$.*

This example highlights a first difference between the possibilistic and the probabilistic framework. The next one shows that the existence of a scale $\mathbf{H}$ and the relevant $\mathbf{H}$-scale monotonicity is not sufficiemt to characterize possibilistic aggregated likelihood as a coherent extension of a point likelihood, even when $(L2)$ holds.

**Example 4.** *Consider the scale $\mathbf{H}$ and the function $\varphi$ of Example 2. Let $E$ be an event logically independent of the events of the partition $\mathcal{L}$ and let $C_i = E \wedge H_i$, $C_{i+3} = E^c \wedge H_i$ $(i = 1, 2, 3)$ be the atoms spanned by $E$ and the $H_i$'s. In order to show that $\varphi$ is not a coherent conditional possibility assessment, consider the following system with unknowns $x_r^0 \geq 0$ for $r = 1, \dots, 6$:*

$$\mathcal{S}_0^\Pi = \begin{cases} x_1^0 = \min\{0.3, \max\{x_1^0, x_4^0\}\} \\ x_2^0 = \min\{0.5, \max\{x_2^0, x_5^0\}\} \\ x_3^0 = \min\{0.8, \max\{x_3^0, x_6^0\}\} \\ \max\{x_1^0, x_2^0\} = \min\{0.5, \max\{x_1^0, x_2^0, x_4^0, x_5^0\}\} \\ \max\{x_1^0, x_3^0\} = \min\{0.8, \max\{x_1^0, x_3^0, x_4^0, x_6^0\}\} \\ \max\{x_2^0, x_3^0\} = \min\{0.7, \max\{x_2^0, x_3^0, x_5^0, x_6^0\}\} \\ \max\{x_1^0, x_2^0, x_3^0\} = \min\{0.75, \max\{x_1^0, x_2^0, x_3^0, x_4^0, x_5^0, x_6^0\}\} \\ \max\{x_1^0, x_2^0, x_3^0, x_4^0, x_5^0, x_6^0\} = 1 \end{cases}$$

*System $\mathcal{S}_0^\Pi$ has no solution: in fact, only $x_4^0$, $x_5^0$ and $x_6^0$ can assume value $1$, but the seventh equation forces to be $x_4^0 < 1$ and $x_6^0 < 1$ in the fifth one, while the sixth equation implies $x_5^0 < 1$ in the seventh.*

*Similarly, it is possible to prove that $\varphi$ is not a coherent $T$-conditional possibility, for any strict t-norm $T$.*

Given a $\mathbf{H}$-scale monotone function $g(E|\cdot)$ with respect to a scale $\mathbf{H}$ of $\mathcal{H}$, it is always possible to derive two sequences of elements of $\mathcal{L}$, useful to characterize a coherent $T$-conditional possibility assessment. In fact by them we can establish whether the scale $\mathbf{H}$ is induced by a $T$-conditional possibility $\Pi$ extending $g$ (and so satisfying condition of Lemma 1).

**Definition 4.** *Assume $g(E|\cdot)$ is $\mathbf{H}$-scale monotone with respect to the scale $\mathbf{H} = \{\mathcal{H}^0, \dots, \mathcal{H}^k\}$ of $\mathcal{H}$. Define the sequences $\mathbf{K} = \{\mathcal{K}^0, \dots, \mathcal{K}^k\}$ and $\overline{\mathbf{K}} = \{\overline{\mathscr{K}}^0, \dots, \overline{\mathscr{K}}^k\}$ generated by $\mathbf{H}$ as follows. For $\alpha = 0, \dots, k$ put:*

- $\mathcal{K}^\alpha = \{H_i \in \mathcal{L} \setminus \mathcal{H}^\alpha \ : \ H_i \subseteq H \in \mathcal{H}^\alpha, \ g(E|H) > \max\limits_{\substack{H_j \subseteq H \\ j \neq i}} f(E|H_j)\}$;

- $\overline{\mathscr{K}}^\alpha = \bigcup_{\beta=0}^{\alpha} \mathcal{K}^\beta \setminus \bigcup_{\beta=0}^{\alpha} \mathcal{H}^\beta$.

**Remark**: If $g(E|\cdot)$ is a coherent $T$-conditional possibility, by Theorem 3 in $\mathcal{K}^\alpha$ there are events $H_i \in \mathcal{H}^\beta$ $(\beta > \alpha)$ such that $f(E|H_i) \geq g(E|H)$, where $H \supseteq H_i$ and $H \in \mathcal{H}^\alpha$.

In particular, for a strict $t$-norm $T$, $\mathcal{K}^\alpha$ contains $H_i$ with $H_i \subseteq H \in \mathcal{H}^\alpha$ such that $f(E|H_i) > g(E|H)$. In fact, $f(E|H_i) \geq g(E|H)$ and when $f(E|H_i) = g(E|H)$ one would have $\Pi(H_i|H_0^\alpha) = \Pi(H|H_0^\alpha) = 1$ and this is not possible.

When the minimum $t$-norm is considered, for $H_i \in \mathcal{K}^\alpha$, one has $\Pi(H_i|H) = \Pi(H_i|H_0^\alpha) \geq g(E|H)$ and the equality must hold for $f(E|H_i) > g(E|H)$.

Note that, for $\alpha = 0, \dots, k$, $\overline{\mathscr{K}}^\alpha$ contains all the elements $H_i$ of some $\mathcal{K}^\beta$ $(\beta \leq \alpha)$ such that $0 < \Pi(H_i|H_0^\alpha) < 1$.

In order to deepen the similarities between conditional probability and $T$-conditional possibility we present the following result related to the sequences $\mathbf{K}$ and $\overline{\mathbf{K}}$. For this aim we introduce the following condition:

*(S1)* for every $\mathcal{H}^\alpha \in \mathbf{H}$, if $K_1, K_2 \in \mathcal{H}^\alpha$, then

$$g(E|K_1 \vee K_2) = \max\{g(E|K_1), g(E|K_2)\}.$$

**Theorem 7.** *If $g(E|\cdot)$ is $\mathbf{H}$-scale monotone with respect to a scale $\mathbf{H} = \{\mathcal{H}^0, \dots, \mathcal{H}^k\}$ of $\mathcal{H}$, satisfies property (S1), and the associated $\overline{\mathbf{K}} = \{\overline{\mathscr{K}}^0, \dots, \overline{\mathscr{K}}^k\}$ is such that $\overline{\mathscr{K}}^\alpha = \emptyset$ for every $\alpha = 0, \dots, k$, then the following statements hold:*

*i) $g$ is a coherent conditional probability;*

*ii) $g$ is a coherent $T$-conditional possibility (for every $t$-norm $T$).*

*Proof.* Requirements on $g$ imply the validity of condition *(L2)*. Define $H_0^0 = \Omega$ and $H_0^\alpha = \bigvee_{H_i \in \bigcup_{\gamma \geq \alpha} \mathcal{H}^\gamma} H_i$ for $\alpha = 1, \dots, k$.

*i)* For $\alpha = 0, \dots, k$, if $\mathcal{L} \cap \mathcal{H}^\alpha = \{H_{i_1}, \dots, H_{i_t}\}$, without loss of generality we can suppose $f(E|H_{i_1}) \geq \ldots \geq f(E|H_{i_t})$. Put $H_{0,1}^\alpha = H_0^\alpha$ and for $j = 1, \dots, t-1$, $H_{0,j+1}^\alpha = \bigvee_{l=j+1}^t H_{i_l} \vee H_0^{\alpha+1}$. Define $P(\cdot|H_{0,j}^\alpha)$ as:

- $P(E \wedge H_{i_j}|H_{0,j}^\alpha) = f(E|H_{i_j})$;
- $P(H_{i_j}|H_{0,j}^\alpha) = 1$;
- $P(E^c \wedge H_{i_j}|H_{0,j}^\alpha) = 1 - f(E|H_{i_j})$ if $E^c \wedge H_{i_j} \neq \emptyset$;
- $P(C|H_{0,j}^\alpha) = 0$ for every atom $C \nsubseteq H_{i_j}$ in $\mathcal{B}$.

It is straightforwardly verified that probabilities $P(\cdot|H_{0,j}^\alpha)$ for increasing $\alpha$ and $j$ determine a conditional probability on $\mathcal{B} \times \mathcal{H}$ extending $g(E|\cdot)$.

*ii)* For $\alpha = 0, \dots, k$, for every $H_i \in \mathcal{H}^\alpha$, define $\Pi(\cdot|H_0^\alpha)$ as:

- $\Pi(E \wedge H_i|H_0^\alpha) = f(E|H_i)$;
- $\Pi(H_i|H_0^\alpha) = 1$;
- $\Pi(E^c \wedge H_i|H_0^\alpha) = 1$ if $E^c \wedge H_i \neq \emptyset$;
- $\Pi(C|H_0^\alpha) = 0$ for every atom $C \nsubseteq \bigvee_{H_i \in \mathcal{H}^\alpha} H_i$ in $\mathcal{B}$.

Since $g$ satisfies *(S1)*, it follows that $\Pi(\cdot|H_0^\alpha)$ is a possibility. Theorem 1 implies that the class of possibilities $\mathcal{P} = \{\Pi_\alpha\}$ with $\Pi_\alpha(\cdot) = \Pi(\cdot|H_0^\alpha)$ determines a $T$-conditional possibility on $\mathcal{B} \times \mathcal{H}$ extending $g(E|\cdot)$. □

**Remark**: Theorem 4 assures extendability of an aggregated likelihood assessment either as a $T$-conditional possibility or as a conditional probability. Theorem 7 covers as particular cases both the monotone and the anti-monotone aggregation. Moreover, it generalizes Theorem 4 by considering also some neither monotone nor anti-monotone cases.

The following example shows an aggregated likelihood assessment satisfying condition expressed in Theorem 7.

**Example 5.** *Consider the partition $\mathcal{L} = \{H_1, H_2, H_3, H_4\}$ and an event $E$ with $H_1 \subseteq E$ and $E \wedge H_4 = \emptyset$.*

*Given the likelihood assessment*

$$f(E|H_1) = 1,\ f(E|H_2) = \frac{1}{2},\ f(E|H_3) = \frac{4}{5},\ f(E|H_4) = 0,$$

*consider the following aggregated likelihood assessment:*

- $g(E|H) = \frac{4}{5}$ *for all $H \in \mathcal{H}$ such that $H \supseteq H_3$;*
- $g(E|H) = 1$ *for all $H \in \mathcal{H}$ such that $H \supseteq H_1$ and $H \not\supseteq H_3$;*
- $g(E|H) = \frac{1}{2}$ *for all $H \in \mathcal{H}$ such that $H \supseteq H_2$ and $H \not\supseteq H_1$ and $H \not\supseteq H_3$;*
- $g(E|H_4) = 0$.

*Note that $g$ extends $f$ and fulfills condition (L2). Moreover, $g(E|\cdot)$ is neither monotone nor anti-monotone with respect to implication. Indeed, $H_1 \subseteq H_1 \vee H_3$ but $g(E|H_1) = 1 > \frac{4}{5} = g(E|H_1 \vee H_3)$, and also $H_2 \subseteq H_2 \vee H_3$ but $g(E|H_2) = \frac{1}{2} < \frac{4}{5} = g(E|H_2 \vee H_3)$.*

*This assessment is* **H***-scale monotone with respect to the scale $\mathbf{H} = \{\mathcal{H}^0, \mathcal{H}^1\}$ of $\mathcal{H}$ defined as follows.*

| $\mathbf{H}$ | $\mathbf{K}$ | $\overline{\mathbf{K}}$ |
|---|---|---|
| $\mathcal{H}^0 = \{H_3, H_1 \vee H_3, H_2 \vee H_3, H_3 \vee H_4, \ldots\}$ | $\mathcal{K}^0 = \emptyset$ | $\overline{\mathscr{K}}^0 = \emptyset$ |
| $\mathcal{H}^1 = \{H_1, H_2, H_4, H_1 \vee H_2, H_1 \vee H_4, H_2 \vee H_4, \ldots\}$ | $\mathcal{K}^1 = \emptyset$ | $\overline{\mathscr{K}}^1 = \emptyset$ |

*Being $\overline{\mathscr{K}}^0 = \overline{\mathscr{K}}^1 = \emptyset$, Theorem 7 implies $g(E|\cdot)$ is simultaneously a coherent $T$-conditional possibility and a coherent conditional probability.*

Notice that Theorem 7 provides only a sufficient condition as shown by the following example.

**Example 6.** *Let $\mathcal{L} = \{H_1, H_2, H_3\}$ be a partition and $E$ an event logically independent of $\mathcal{L}$. The generated atoms are $C_i = E \wedge H_i$ and $C_{i+3} = E^c \wedge H_i$, for $i = 1, 2, 3$.*

*Consider the likelihood assessment*

$$f(E|H_1) = \frac{1}{5},\ f(E|H_2) = \frac{1}{2},\ f(E|H_3) = \frac{4}{5}.$$

*The following aggregated likelihood assessment $g$ extending $f$ as:*

- $g(E|H_1 \vee H_2) = g(E|H_1 \vee H_3) = g(E|\Omega) = \frac{1}{5}$*;*
- $g(E|H_2 \vee H_3) = \frac{3}{5}$*;*

*satisfies condition (L2).*

*The only scale of* $\mathcal{H}$ *with respect to* $g(E|\cdot)$ *is* $\mathbf{H}$*-scale monotone is* $\mathbf{H} = \{\mathcal{H}^0, \mathcal{H}^1, \mathcal{H}^2\}$ *defined as follows.*

| $\mathbf{H}$ | $\mathbf{K}$ | $\overline{\mathbf{K}}$ |
|---|---|---|
| $\mathcal{H}^0 = \{H_1, H_1 \vee H_2, H_1 \vee H_3, \Omega\}$ | $\mathcal{K}^0 = \emptyset$ | $\overline{\mathscr{K}}^0 = \emptyset$ |
| $\mathcal{H}^1 = \{H_2, H_2 \vee H_3\}$ | $\mathcal{K}^1 = \{H_3\}$ | $\overline{\mathscr{K}}^1 = \{H_3\}$ |
| $\mathcal{H}^2 = \{H_3\}$ | $\mathcal{K}^2 = \emptyset$ | $\overline{\mathscr{K}}^2 = \emptyset$ |

*Being* $\overline{\mathscr{K}}^1 \neq \emptyset$*, Theorem 7 cannot be applied. Anyway,* $g(E|\cdot)$ *can be extended on* $\mathcal{B} \times \mathcal{H}$ *either as a* $T$*-conditional possibility (with* $T = \min$ *or strict t-norm) or as a conditional probability.*

*Indeed, a* min*-conditional possibility* $\Pi(\cdot|\cdot)$ *extending* $g$ *on* $\mathcal{B} \times \mathcal{H}$ *is induced by the following class of unconditional possibilities on* $\mathcal{B}$

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ |
|---|---|---|---|---|---|---|
| $\Pi(\cdot|H_0^0)$ | $\frac{1}{5}$ | $0$ | $0$ | $1$ | $0$ | $0$ |
| $\Pi(\cdot|H_0^1)$ | $0$ | $\frac{1}{2}$ | $\frac{3}{5}$ | $0$ | $1$ | $\frac{3}{5}$ |
| $\Pi(\cdot|H_0^2)$ | $0$ | $0$ | $\frac{4}{5}$ | $0$ | $0$ | $1$ |

*where* $H_0^0 = \Omega$*,* $H_0^1 = H_2 \vee H_3$ *and* $H_0^3 = H_3$*.*

*Finally, a conditional probability* $P(\cdot|\cdot)$ *extending* $g$ *on* $\mathcal{B} \times \mathcal{H}$ *is induced by the following class of unconditional probabilities on* $\mathcal{B}$

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ |
|---|---|---|---|---|---|---|
| $P(\cdot|H_0^0)$ | $\frac{1}{5}$ | $0$ | $0$ | $\frac{4}{5}$ | $0$ | $0$ |
| $P(\cdot|H_0^1)$ | $0$ | $\frac{1}{3}$ | $\frac{4}{15}$ | $0$ | $\frac{1}{3}$ | $\frac{1}{15}$ |

*where* $H_0^0 = \Omega$*,* $H_0^1 = H_2 \vee H_3$*.*

# References

[1] B. Bouchon-Meunier, G. Coletti, C. Marsala, Independence and Possibilistic Conditioning. *Annals of Mathematics and Artificial Intelligence*, 35 pp. 107–123, 2002.

[2] G. Coletti, O. Gervasi, S. Tasso, B. Vantaggi, Generalized Bayesian inference in a fuzzy context: From theory to a virtual reality application. *Computational Statistics & Data Analysis*, 56, pp. 967–980, 2012.

[3] G. Coletti, R. Scozzafava, *Probabilistic Logic in a Coherent Setting.* Kluwer Academic Publisher Dordrecht/Boston/London, 2002.

[4] G. Coletti, B. Vantaggi, T-conditional possibilities: Coherence and inference. *Fuzzy Sets and Systems*, 160(3), pp. 306–324, 2009.

[5] G. de Cooman, Possibility theory II: Conditional Possibility. *Int. J. General Systems*, 25, pp. 325-351, 1997.

[6] B. de Finetti, Sul significato soggettivo della probabilità, *Fundamenta Mathematicae*, 17, pp. 298–329, 1931 (Engl. transl. in: Induction and Probability – Eds. P. Monari, D. Cocchi, CLUEB, Bologna: 291–321, 1993).

[7] D. Dubois, L. Fariñas del Cerro, A. Herzig, H. Prade, An ordinal view of independence with application to plausible reasoning. *Uncertainty in Artificial Intelligence*, pp. 195-203, 1994.

[8] D. Dubois, H. Prade, Fuzzy sets and statistical data, *European Journal of Operational Research*, 25, pp. 345–356, 1986.

[9] E. Hisdal, Conditional possibilities independence and noninteraction. *Fuzzy Sets and Systems*, 1, pp. 283-297, 1978.

[10] W. Näther, On possibilistic inference, *Fuzzy Sets and Systems*, 6, pp. 327–337, 1990.

[11] J. Vejnarová, Conditional independence relations in possibility theory. *Int. J. Uncert., Fuzziness and Knowledge-Based Systems*, 8, pp. 253-269, 2000.

[12] L.A. Zadeh, Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1, pp. 3-28, 1978.

[13] L.A. Zadeh, Toward a generalized theory of uncertainty (GTU)–an outline, *Information Sciences*, 172, pp. 1–40 2005.

[14] L.A. Zadeh, Toward a perception-based theory of probabilistic reasoning with imprecise probabilities, *Journal of Statistical Planning and Inference*, 105(1), pp. 233–264, 2002.

# Introduction to Algebra of Belief Functions on Three-element Frame of Discernment — A General Case

**Milan Daniel**[*]
Institute of Computer Science
Academy of Sciences of the Czech Republic
milan.daniel@cs.cas.cz

**Abstract**

This contribution presents the second part of the introductive study of algebraic structure of belief functions (BFs) on 3-element frame of discernment.

Algebraic method by Hájek & Valdés for BFs on 2-element frames is generalized to larger frame of discernment. Due to complexity of the algebraic structure, the study is divided into 2 parts, the present one is devoted to a case of general BFs.

The definition of Dempster's semigroup (an algebraic structure) of BFs on 3-element frame is recalled from the first part of the study. Results related to Bayesian and quasi Bayesian BFs from the first part are also briefly recalled. Further substructures related to another subsets of general BFs are described and analyzed (including idempotents, simple complementary BFs, generalizations of subsemigroups $S_i$'s) and subalgebras isomorphic to Dempster's semigroup on 2-element frame of discernment.

Ideas and open problems for future research are presented.

## 1 Introduction

Belief functions (BFs) are one of the widely used formalisms for uncertainty representation and processing that enable representation of incomplete and uncertain knowledge, belief updating, and combination of evidence. They were originally introduced as a principal notion of the Dempster-Shafer Theory or the Mathematical Theory of Evidence [18].

When combining belief functions by the conjunctive rules of combination, conflicts often appear, which are assigned to $\emptyset$ by un-normalized conjunctive rule $\bigcirc\!\!\!\!\cap$ or normalized by Dempster's rule of combination $\oplus$. Combination of conflicting BFs and interpretation of conflicts is often questionable in real applications, thus a series of

[*] This research is supported by the grant P202/10/1826 of the Grant Agency of the Czech Republic. Partial institutional support RVO: 67985807 is also acknowledged.

alternative combination rules was suggested and a series of papers on conflicting belief functions was published, e.g. [1, 6, 9, 11, 17, 19].

A need of algebraic analysis of belief functions on frames of discernment with more then two elements comes from our previous study of conflicting belief functions (a decomposition of BFs into their non-conflicting and conflicting parts, requires a generalization of Hájek-Valdés operation "*minus*") [12] which was motivated by series of papers on conflicting belief functions [1, 6, 9, 17, 19]. Inspired by this demand we start with a generalization of the algebraic analysis of BFs in this study. Due to exponential growth of a complexity of a structure of belief functions with respect to the size of corresponding frame of discernment, we are starting on 3-element frame.

Method by Hájek & Valdés for BFs on 2-element frames [15, 16, 20] is generalized to larger frame of discernment here. Due to complexity of the algebraic structure, the study is divided into 2 parts; the first one [13] is devoted to the special simplified case of quasi Bayesian BFs (i.e., to the case of very simple BFs, which are analogy of non-normalized probability); this is the second part, which is devoted to general BFs.

This part starts with brief recalling of Hájek-Valdés definition of Dempster's semigroup (an algebraic structure) of BFs on 2-element frame and also of the recent definition on 3-element frame from [13] (Section 2). Further a study of general subalgebras of Dempster's semigroup follows (Section 3); it includes idempotents, (simple) complementary BFs, brief overview of the principal results on Bayesian and quasi Bayesian BFs, a generalization of subalgebra $S$ and that of $S_i's$, and substructures isomorphic to Dempster's semigroup on 2-element frame.

Ideas and open problems for future research are presented in Section 4.

# 2 Preliminaries

## 2.1 General Primer on Belief Functions

We assume classic definitions of basic notions from theory of *belief functions* [18] on finite frames of discernment $\Omega_n = \{\omega_1, \omega_2, ..., \omega_n\}$. A *basic belief assignment (bba)* is a mapping $m : \mathcal{P}(\Omega) \longrightarrow [0,1]$ such that $\sum_{A\subseteq\Omega} m(A) = 1$; the values of the bba are called *basic belief masses (bbm)*. $m(\emptyset) = 0$ is usually assumed. A *belief function (BF)* is a mapping $Bel : \mathcal{P}(\Omega) \longrightarrow [0,1]$, $Bel(A) = \sum_{\emptyset\neq X\subseteq A} m(X)$. A *plausibility function* $Pl(A) = \sum_{\emptyset\neq A\cap X} m(X)$. There is a unique correspondence among $m$ and corresponding $Bel$ and $Pl$ thus we often speak about $m$ as about belief function.

A *focal element* is a subset $X$ of the frame of discernment, such that $m(X) > 0$. If all the focal elements are *singletons* (i.e., one-element subsets of $\Omega$), then we speak about *Bayesian belief function* (BBF); if all the focal elements are either singletons or whole $\Omega$ (i.e. $|X| = 1$ or $|X| = |\Omega|$), then we speak about *quasi Bayesian belief function* (qBBF); if all focal elements are nested, we speak about *consonant belief function*; if all focal elements have non-empty intersection, we speak about *consistent belief function*.

Let us recall $U_n$ the *uniform Bayesian belief function* [9], i.e., the uniform probability distribution on $\Omega_n$, and *normalized plausibility of singletons* of $Bel$: the BBF $Pl\_P(Bel)$ such, that $(Pl\_P(Bel))(\omega_i) = \frac{Pl(\{\omega_i\})}{\sum_{\omega\in\Omega} Pl(\{\omega\})}$ [2, 8]. An *indecisive BF* is a

BF, which does not prefer any $\omega_i \in \Omega_n$, i.e., BF which gives no decisional support for any $\omega_i$, i.e., BF such that $h(Bel) = Bel \oplus U_n = U_n$, i.e., $Pl(\{\omega_i\}) = const.$, i.e., $(Pl\_P(Bel))(\{\omega_i\}) = \frac{1}{n}$, [10].

Let us define *exclusive BF* as a BF such that $Pl(X) = 0$ for some $\emptyset \neq X \subset \Omega$; BF is non-exclusive otherwise. *(Simple) complementary BF* has up to two focal elements $\emptyset \neq X \subset \Omega$ and $\Omega \setminus X$. *(Simple) quasi complementary BF* has up to 3 focal elements $\emptyset \neq X \subset \Omega$, $\Omega \setminus X$ and $\Omega$.

## 2.2 Belief Functions on 2-Element Frame of Discernment; Dempster's Semigroup

Let us suppose, that the reader is slightly familiar with basic algebraic notions like *a semigroup* (an algebraic structure with an associative binary operation), *a group* (a structure with an associative binary operation, with a unary operation of inverse, and with a neutral element), *a neutral element n* $(n * x = x)$, *an absorbing element a* $(a * x = a)$, *an idempotent* i $(i * i = i)$, *a homomorphism f* $(f(x * y) = f(x) * f(y))$, etc. (Otherwise, see e.g., [4, 7, 15, 16].)

We represent belief functions on $\Omega_2 = \{\omega_1, \omega_2\}$ by Dempster's pairs $(d_1, d_2) = (m(\{\omega_1\}), m(\{\omega_2\}))$ as $m(\{\omega_1, \omega_2\}) = 1 - d_1 - d_2$, see Figure 1, and analogously BFs on $\Omega_3 = \{\omega_1, \omega_2, \omega_3\}$ by *Dempster's 6-tuples* or *d-6-tuples*[1] $(d_1, d_2, d_3, d_{12}, d_{13}, d_{23}) = (m(\{\omega_1\}), m(\{\omega_2\}), m(\{\omega_3\}), m(\{\omega_1, \omega_2\})\ m(\{\omega_1, \omega_3\}), m(\{\omega_2, \omega_3\}))$, where $0 \leq d_i, d_{ij} \leq 1$ and $\sum_{i=1}^{3} d_i + \sum_{ij=12}^{23} d_{ij} \leq 1$. We can represent $d$-6-tuples by a six-dimensional 'triangle', see Figure 2.

*Extremal d-pairs* are the pairs corresponding to BFs for which either $m(\{\omega_1\}) = 1$ or $m(\{\omega_2\}) = 1$, i.e., $\perp = (0, 1)$ and $\top = (1, 0)$. The set of all non-extremal d-pairs is denoted as $D_0$; the set of all non-extremal *Bayesian d-pairs* (i.e., $d_1 + d_2 = 1$) is denoted as $G$; the set of d-pairs such that $a = b$ is denoted as $S$ (set of indecisive[2] d-pairs), the set where $b = 0$ as $S_1$, and analogically, the set where $a = 0$ as $S_2$ (simple support BFs). Vacuous BF is denoted as $0 = (0, 0)$.

*Dempster's (conjunctive) rule of combination* $\oplus$ is given as $(m_1 \oplus m_2)(A) = \sum_{X \cap Y = A} K m_1(X) m_2(Y)$ for $A \neq \emptyset$, where $K = \frac{1}{1-\kappa}$, $\kappa = \sum_{X \cap Y = \emptyset} m_1(X) m_2(Y)$, and $(m_1 \oplus m_2)(\emptyset) = 0$, see [18]; on $\Omega_2$ we have: $(a, b) \oplus (c, d) = (1 - \frac{(1-a)(1-c)}{1-(ad+bc)},\ 1 - \frac{(1-b)(1-d)}{1-(ad+bc)})$ [15].

The *(conjunctive) Dempster's semigroup* $\mathbf{D}_0 = (D_0, \oplus, 0, 0')$ is the set of non-extremal $d$-pairs $D_0 = \{(d_1, d_2) \,|\, 0 \leq d_1, d_2 \leq 1, d_1 + d_2 \leq 1\} \setminus \{\perp, \top\}$ endowed with the binary operation $\oplus$ (i.e. with the Dempster's rule) and two distinguished elements $0 = (0, 0)$ and $0' = (\frac{1}{2}, \frac{1}{2})$.

In $D_0$ it is defined further: $-(a, b) = (b, a)$, $h(a, b) = (a, b) \oplus 0' = (\frac{1-b}{2-a-b}, \frac{1-a}{2-a-b})$, $h_1(a, b) = \frac{1-b}{2-a-b}$, $f(a, b) = (a, b) \oplus (b, a) = (\frac{a+b-a^2-b^2-ab}{1-a^2-b^2}, \frac{a+b-a^2-b^2-ab}{1-a^2-b^2})$; $(a, b) \leq (c, d)$ iff $[h_1(a, b) < h_1(c, d)$ or $h_1(a, b) = h_1(c, d)$ and $a \leq c]$ [3].

The principal properties of $\mathbf{D}_0$ are summarized by the following theorem:

[1] For simplicity of expressions, we speak often simply on 6-tuples only.

[2] BFs $(a, a)$ from $S$ are called *indifferent* BFs by Haenni [14].

[3] Note, that $h(a, b)$ is an abbreviation for $h((a, b))$, similarly for $h_1(a, b)$ and $f(a, b)$.

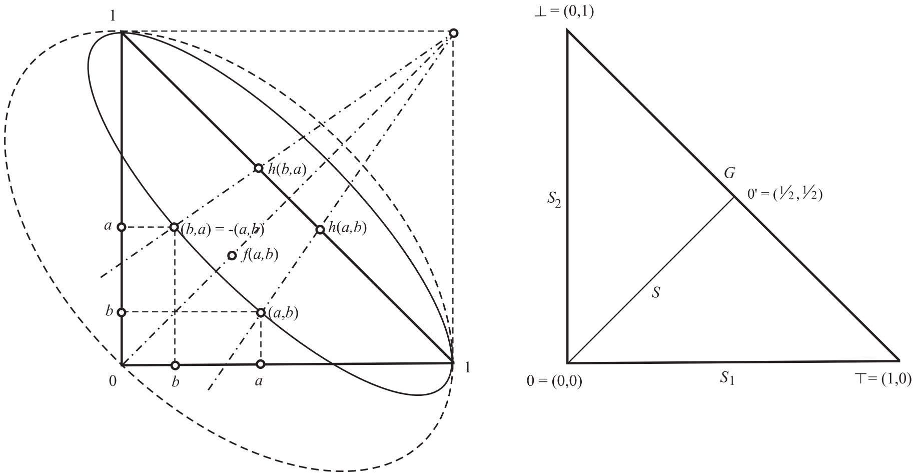


Figure 1: Dempster's semigroup $D_0$. Homomorphism $h$ is in this representation a projection to group $G$ along the straight lines running through the point $(1, 1)$. All the Dempster's pairs lying on the same ellipse are mapped by homomorphism $f$ to the same $d$-pair in semigroup S.

**Theorem 1** *(i) The Dempster's semigroup* $\mathbf{D}_0$ *with the relation* $\leq$ *is an ordered commutative (Abelian) semigroup with the neutral element* $0$*;* $0'$ *is the only non-zero idempotent of* $\mathbf{D}_0$*.*
*(ii)* $\mathbf{G} = (G, \oplus, -, 0', \leq)$ *is an ordered Abelian group, isomorphic to the additive group of reals with the usual ordering. Let us denote its negative and positive cones as* $G^{\leq 0'}$ *and* $G^{\geq 0'}$*.*
*(iii) The sets* $S, S_1, S_2$ *with the operation* $\oplus$ *and the ordering* $\leq$ *form ordered commutative semigroups with neutral element* $0$*; they are all isomorphic to the positive cone of the additive group of reals.*
*(iv) h is an ordered homomorphism:* $(D_0, \oplus, -, 0, 0', \leq) \longrightarrow (G, \oplus, -, 0', \leq)$*;* $h(Bel) = Bel \oplus 0' = Pl\_P(Bel)$*, i.e., the normalized plausibility of singletons probabilistic transformation.*
*(v)* $f$ *is a homomorphism:* $(D_0, \oplus, -, 0, 0') \longrightarrow (S, \oplus, -, 0)$*; (but, not an ordered one).*

For proofs see [15, 16, 20].

For other properties of $\mathbf{D}_0$ see [15, 16, 20] and further [3, 4, 5, 7, 12].

*Exclusive d-6-tuples* are 6-tuples representing BFs, such that $Pl(\{\omega_i\}) = 0$ for some $1 \leq i \leq 3$, i.e., $d$-6-tuples$(a, b, 0, 1{-}a{-}b, 0, 0)$, $(a, 0, b, 0, 1{-}a{-}b, 0)$, $(0, a, b, 0, 0, 1{-}a{-}b)$.

**Definition 1** *The* (conjunctive) Dempster's semigroup $\mathbf{D}_3 = (D_3, \oplus, 0, 0')$ *is the set* $D_3$ *of all non-exclusive Dempster's* $6$*-tuples, endowed with the binary operation* $\oplus$ *(i.e. with the Dempster's rule) and two distinguished elements* $0$ *and* $0' = U_3$*, where* $0 = 0_3 = (0, 0, ..., 0)$ *and* $0' = 0'_3 = U_3 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0)$*.*

There is homomorphism $h : \mathbf{D}_3 \longrightarrow \mathcal{BBF}_3 = \{Bel \in D_3 \,|\, Bel \text{ is BBF}\}$ defined by $h(Bel) = Bel \oplus U_3$, it holds that $h(Bel) = Pl\_P(Bel)$, see [10].

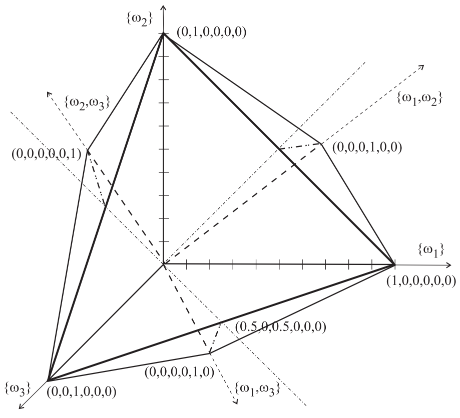


Figure 2: General BFs on 3-element frame $\Omega_3$.

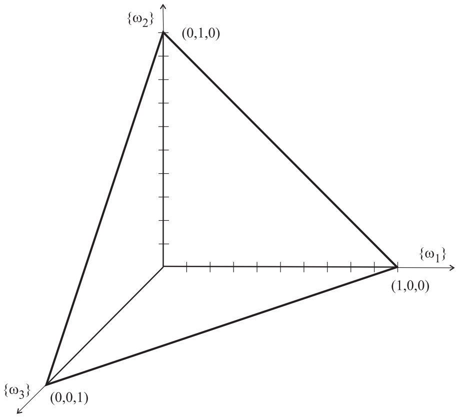


Figure 3: Quasi Bayesian BFs on 3-element frame $\Omega_3$.

### 2.3 The extended Dempster's semigroup

There are only single 2 extremal (categorical, exclusive) $d$-pairs on $\Omega_2$, thus the extension of $\mathbf{D}_0$ to $\mathbf{D}_0^+$, (where $D_0^+ = D_0 \cup \{\bot, \top\}$ and $\bot \oplus \top$ is undefined) is important for applications, but not interesting from the theoretical point of view. Extended Dempster's semigroup $\mathbf{D}_0^+$ has never been explicitly published; implicitly, it is the special case of the algebraic structure of extended Dempsteroid [16].

There are 6 categorical (exclusive) $d$-6-tuples in $\mathbf{D}_3^+$ (in the set of BFs defined over $\Omega_3$) and many general exclusive 6-tuples (BFs) in $\mathbf{D}_3^+$, thus the issue of extension of Dempster's semigroup to all BFs on $\Omega_3$ is more interesting and also more important than in the case of BFs on $\Omega_2$, because a complex structure of exclusive BFs is omitted in Dempster's semigroup of non-exclusive BFs, in the case of $\Omega_3$. Nevertheless, due to the extent of this study we are concentrating to the non-extended case only in this text.

## 3 Subalgebras of Dempster's semigroup

### 3.1 Idempotents of Dempster's semigroup

Let us start with investigation of the idempotents of Dempster's semigroup, which are trivial subalgebras of the entire Dempster's semigroups and also of other corresponding substructures of it.

There are two simple generalizations of idempotents 0 and $0'$ from $\mathbf{D}_0$. There is vacuous BF $0 = (0,0,0,0,0,0)$, which is neutral hence also idemponent in $\mathbf{D}_3$, and $0' = U_3 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0)$ which is idempotent in $\mathbf{D}_3$, both these idempotents are mentioned already in [13]. Note that neither indecisive BF $(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$ nor $(0,0,0,\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is idempotent w.r.t. $\oplus$.

There are simple generalizations of absorbing idempotents $\bot$ and $\top$: $1_1 = (1,0,0,0,0,0), 1_2 = (0,1,0,0,0,0), 1_3 = (0,0,1,0,0,0)$, all of them are exclusive absorbing idempotents, thus they are in $\mathbf{D}_3^+$; similarly for another generalizations of $\perp$ and $\top$: $1_{12} = (0,0,0,1,0,0), 1_{13} = (0,0,0,0,1,0), 1_{23} = (0,0,0,0,0,1)$, which are idempotents in $\mathbf{D}_3^+$, thus out of scope of this study.

There are another 3 idempotents in $\mathbf{D}_3$: $(\frac{1}{2},0,0,0,0,\frac{1}{2})$, $(0,\frac{1}{2},0,0,\frac{1}{2},0)$, and $(0,0,\frac{1}{2},\frac{1}{2},0,0)$.

To complete this subsection we have to mention transformation of $0'$ from $\mathbf{D}_0$ to three-element frame $\Omega_3$ $(\frac{1}{2},\frac{1}{2},0,0,0,0)$ and analogous $(\frac{1}{2},0,\frac{1}{2},0,0,0)$ and $(0,\frac{1}{2},\frac{1}{2},0,0,0)$ which are exclusive idempotents, thus in $\mathbf{D}_3^+$. We have to note here that 3 similar BFs $(0,0,0,\frac{1}{2},\frac{1}{2},0)$, $(0,0,0,\frac{1}{2},0,\frac{1}{2})$, and $(0,0,0,0,\frac{1}{2},\frac{1}{2})$, are non-exclusive, i.e., in $\mathbf{D}_3$ but they are not idempotents. Similarly 6 analogous non-exclusive BFs $(\frac{1}{2},0,0,\frac{1}{2},0,0)$, $(\frac{1}{2},0,0,0,\frac{1}{2},0)$, $(0,\frac{1}{2},0,\frac{1}{2},0,0)$, $(0,\frac{1}{2},0,0,0,\frac{1}{2})$, $(0,0,\frac{1}{2},0,\frac{1}{2},0)$ and $(0,0,\frac{1}{2},0,0,\frac{1}{2})$ are not idempotents.

Proofs of all the above statements are simple verifications of the properties.

Further it is possible to prove that there do not exist another idempotents w.r.t. $\oplus$ in $\mathbf{D}_3$ and $\mathbf{D}_3^+$.

We can summarize this subsection at it follows. There are just 5 idempotents w.r.t. $\oplus$ in $\mathbf{D}_3$: $0 = (0,0,0,0,0,0)$ which is neutral, and further $U_3$, $(\frac{1}{2},0,0,0,0,\frac{1}{2})$, $(0,\frac{1}{2},0,0,\frac{1}{2},0)$, and $(0,0,\frac{1}{2},\frac{1}{2},0,0)$.

There are just 9 other idempotents in $\mathbf{D}_3^+$, 3 absorbing ones: $1_1$, $1_2$, $1_3$, and 6 non-absorbing: $1_{12}$, $1_{13}$, $1_{23}$, $(\frac{1}{2},\frac{1}{2},0,0,0,0)$, $(\frac{1}{2},0,\frac{1}{2},0,0,0)$, and $(0,\frac{1}{2},\frac{1}{2},0,0,0)$.

We have to note that all idempontents form trivial subalgebras (subsemigroups, subgroups) of $\mathbf{D}_3$, e.g., $(\{U_3\},\oplus,Id,U_3)$ for $U_3$, $(\{(\frac{1}{2},0,0,0,0,\frac{1}{2},)\},\oplus,Id,(\frac{1}{2},0,0,0,0,\frac{1}{2}))$ for $(\frac{1}{2},0,0,0,0,\frac{1}{2})$, and similarly.

## 3.2 The subgroups/subalgebras of Bayesian belief functions

Bayesian BFs form a subset of quasi Bayesian BFs, thus investigation of their structure was included in [13]. Bayesian BFs are represented by the triangle with vertices $1_1 = (1,0,0,0,0,0), 1_2 = (0,1,0,0,0,0), 1_3 = (0,0,1,0,0,0)$. On the border of the triangle, there are exclusive BFs, thus only internal part of the triangle is relevant to this study. We will recall the principal results from [13] in the form of the following lemma here, and recall also Hypothesis 1.

**Lemma 1** *(i)* $G_3 = (\{(d_1,d_2,d_3,0,0,0)\,|\,d_i > 0, \sum_{i=1}^{3} d_i = 1\},\oplus,\text{"}-\text{"},U_3)$ *is a group, i.e., subgroup of* $\mathbf{D}_3$*; where* "$-$" *is given by* $-(d_1,d_2,1-(d_1+d_2),0,0,0) = (x_1,\frac{d_1}{d_2}x_1,\frac{d_1}{1-(d_1+d_2)}x_1,0,0,0;0)$*, and* $x_1 = 1/(1+\frac{d_1}{d_2}+\frac{d_1}{1-(d_1+d_2)})$*.*
*(ii)* $G_{2=3} = (\{(d_1,d_2,d_2,0,0,0;0)\},\oplus,\ minus_{2=3},U_3)$ *is subgroup of* $G_3$ *and of* $\mathbf{D}_3$*, where* $minus_{2=3}(d_1,d_2,d_2,0,0,0;0) = (\frac{1-d_1}{1+3d_1},\frac{2d_1}{1+3d_1},\frac{2d_1}{1+3d_1},0,0,0;0)$*. The same holds true also for analogous structures* $G_{1=3} = (\{(d_1,d_2,d_1,0,0,0;0)\},\oplus,\ minus_{1=3},0'_3)$ *and* $G_{1=2} = (\{(d_1,d_1,d_3,0,0,0;0)\},\oplus,\ minus_{1=2},U_3)$*.*

As we have 3 different non-ordered elements, without any priority, we have not any linear ordering of $G_3$ in general, thus neither any isomorphism to additive group of reals in general. This is a difference of $G_3$ subgroup of $\mathbf{D}_3$ from $G$ subgroup of $\mathbf{D}_0$.

On the other hand, we can define ordering analogous to that of $G \subset D_0$ on all three subgroups $G_{2=3}$, $G_{1=3}$, and $G_{1=2}$.

**Hypothesis 1** *Groups $G_{2=3}$, $G_{1=3}$, and $G_{1=2}$ from Lemma 1 are subgroups of $\mathbf{D}_3$ isomorphic to the additive group of reals.*

According to ordering defined on the subgroups, they are *o*-isomorphic to **Re** with usual or with inverse ordering.

## 3.3 The subgroups of (simple) complementary belief functions

Reassigning $d_2 + d_2$ of elements $G_{2=3}$ from singletons $\{\omega_2\}$ and $\{\omega_3\}$ to $\{\omega_2, \omega_3\}$ we obtain set of BFs $(d_1, 0, 0, 0, 0,\ 1 - d_1)$, similarly to $G_{2=3}$ there is subgroup $G_{1-23} = (\{(a, 0, 0, 0, 0, 1 - a) \in D_3\}, \oplus, minus_{1-23}, (\frac{1}{2}, 0, 0, 0, 0, \frac{1}{2}))$, where $minus_{1-23}(a, 0, 0, 0, 0, 1-a) = (1-a, 0, 0, 0, 0, a)$, and analogically subgroups $G_{2-13} = (\{(0, a, 0, 0, 1 - a, 0) \in D_3\}, \oplus,\ minus_{2-13}, (0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0))$ and $G_{3-12} = (\{(0, 0, a, 1 - a, 0, 0) \in D_3\}, \oplus, minus_{3-12}, (0, 0, \frac{1}{2}, \frac{1}{2}, 0, 0))$. There is simple isomorphism $z : G_{1-23} \longrightarrow G$, $z(a, 0, 0, 0, 0, 1 - a) = (a, 1 - a)$, and analogous isomorphisms for $G_{2-13}$ and $G_{3-12}$, hence all 3 subgroups $G_{1-23}$, $G_{2-13}$, and $G_{3-12}$ are isomorphic also to the additive group of reals. Hence we have proved the following lemma:

**Lemma 2** *Structure $G_{1-23} = (\{(a, 0, 0, 0, 0, 1-a) \in D_3\}, \oplus, minus_{1-23}, (\frac{1}{2}, 0, 0, 0, 0, \frac{1}{2}))$, where $minus_{1-23}(a, 0, 0, 0, 0, 1 - a) = (1 - a, 0, 0, 0, 0, a)$, is a subgroup of $\mathbf{D}_3$ which is isomorphic to the additive group of reals (o-isomorphic when appropriate ordering is defined on $G_{1-23}$).*

*The same holds true also for the structures $G_{2-13} = (\{(0, a, 0, 0, 1-a, 0) \in D_3\}, \oplus,$ $minus_{2-13}, (0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0))$ and $G_{3-12} = (\{(0, 0, a, 1 - a, 0, 0) \in D_3\}, \oplus,\ minus_{3-12},$ $(0, 0, \frac{1}{2}, \frac{1}{2}, 0, 0))$.*

Positive and negative cones of $G_{1-23}, G_{2-13}, G_{3-12}$ (with and without corresponding neutral elements) are subsemigroups of $G_{1-23}, G_{2-13}, G_{3-12}$ thus also subsemigroups of $\mathbf{D}_3$.

## 3.4 The subsemigroup of quasi Bayesian belief functions

The algebraic structure $\mathbf{D}_{3-0}$ of all non-exclusive quasi Bayesian belief functions $D_{3-0} = \{(a, b, c,\ 0, 0, 0);\ 0 \le a + b + c \le 1,\ 0 \le a, b, c,\ a{+}b{<}1, a{+}c{<}1, b{+}c{<}1\}$ was particularly studied in [13]. As entire $\mathbf{D}_{3-0}$ and all its subalgebras are substructures of $\mathbf{D}_3$, we can recall summary of [13] here:

**Theorem 2** *(i) Monoid $\mathbf{D}_{3-0} = (D_{3-0}, \oplus, 0, U_3)$ is subsemigroup of $\mathbf{D}_3$ with neutral element $0 = (0, 0, 0, 0, 0, 0)$ and the only other idempotent $0' = U_3 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0)$.*
*(ii) Subgroup of non-exclusive BBFs $G_3 = (\{(a, b, c, 0, 0, 0)|a + b + c = 1, 0 < a, b, c\}, \oplus, "-", U_3)$ and its subalgebras are subalgebras of $\mathbf{D}_{3-0}$, where $"-"$ is given by $-(d_1, d_2, 1 - (d_1 + d_2), 0, 0, 0) = (x_1, \frac{d_1}{d_2}x_1, \frac{d_1}{1-(d_1+d_2)}x_1, 0, 0, 0; 0)$, and $x_1 = 1/(1 + \frac{d_1}{d_2} + \frac{d_1}{1-(d_1+d_2)})$.*
*(iii) The set of non-exclusive BFs $S_0, S_1, S_2, S_3, S_{1-2}, S_{1-3}, S_{2-3}$ with the operation $\oplus$ and VBF $0$ form commutative semigroups with neutral element $0$ (monoids); they*

*are all isomorphic*[4] *to the positive cone of the additive group of reals* $\mathbf{Re}^{\geq 0}$ *(to* $\mathbf{Re}^{\geq 0+}$ *extended with* $\infty$ *in the case of* $S_0$*).*
*(iv) Subsemigroups* $\mathbf{D}_{1-2=3}$*,* $\mathbf{D}_{2-1=3}$ *and* $\mathbf{D}_{3-1=2}$ *(with their subalgebras* $S_i$*'s,* $G_{2=3}$*,*$G_{1=3}$ *and* $G_{1=2}$*) are subsemigroups (resp. subgroups in the case of* $G_i$*'s) of* $\mathbf{D}_{3-0}$ *(hence also of* $\mathbf{D}_3$*). Assuming validity of Hypothesis 1,* $\mathbf{D}_{1-2=3}$*,* $\mathbf{D}_{2-1=3}$ *and* $\mathbf{D}_{3-1=2}$ *are isomorphic to Dempster's semigroup* $\mathbf{D}_0$*.*
*(v) Semigroups of non-exclusive BFs* $\mathbf{D}_{1-2} = (\{(a,b,0,0,0,0) \in D_{3-0}\}, \oplus)$*,* $\mathbf{D}_{1-3} = (\{(a,0,c,0{,}0{,}0) \in D_{3-0}\}, \oplus)$*,* $\mathbf{D}_{2-3} = (\{(0,b,c,0{,}0{,}0) \in D_{3-0}\}, \oplus)$*, are subsemigroups of* $\mathbf{D}_{3-0}$ *and all three are isomorphic to* $\mathbf{D}_0$ *without set of BBFs* $G$*.*
*(vi) h is homomorphism:* $(D_{3-0}, \oplus, 0, U_3) \longrightarrow (G_3, \oplus, "-", U_3)$*;* $h(Bel) = Bel \oplus 0' = Pl\_P(Bel)$*, i.e., the normalized plausibility of singletons probabilistic transformation.*

Generalization of the operation $-$ and homomorphism $f$ from $\mathbf{D}_0$ to entire $\mathbf{D}_{3-0}$ is still under development.

Let us try to describe subalgebras of $\mathbf{D}_3$ in full generality in the next/following subsections.

### 3.5 Generalizations of subsemigroup $S$

There are 3 generalizations of subsemigroup $S$ in $\mathbf{D}_3$. At first $S_0 = (\{(a,a,a,0,0,0) \in D_{3-0}\}, \oplus)$ isomorphic to the extended positive cone $\mathbf{Re}_{\geq 0+}$ (already presented in the previous subsection), second is $S = (\{(a,a,a,b,b,b) \in D_3\}, \oplus)$, similarly with its subsemigroup $S_0$ with neutral idempotent 0 and absorbing one $U_3$. (note that set of BFs $\{(a,a,a,a,a,a) \in D_3\}$ is not closed under $\oplus$, thus it does not form a semigroup). The third is $S_{Pl} = \{(d_1, d_2, ..., d_{23}) \in D_3 \mid Pl((d_1, d_2, ..., d_{23}) = U_3\}, \oplus)$; closeness w.r.t. $\oplus$ follows commutation of $\oplus$ with $Pl$ [2], both $S$ and $S_0$ are subsemigroups of $S_{Pl}$.

### 3.6 Generalizations of subsemigroups $S_i$'s

BFs in $S_i$ in $\mathbf{D}_0$ are qBBFs, simple (support), consonant, consistent, etc., ... According to these properties there are many different generalizations of $S_i$'s. Quasi Bayesian subsemigroups were already mentioned: $S_1 = (\{(d_1,0,0,0,0,0) \in D_{3-0}\}, \oplus)$, $S_2 = (\{(0,d_2,0,0,0,0) \in D_{3-0}\}, \oplus)$, $S_3 = (\{(0,0,d_3,0,0,0) \in D_{3-0}\}, \oplus)$.

Simple (support) BFs contain also generalizations $S_{12} = (\{(0,0,0,d_{12},0,0) \in D_3\}, \oplus)$, $S_{13} = (\{(0,0,0,0,d_{13},0) \in D_3\}, \oplus)$, $S_{23} = (\{(0,0,0,0,0,d_{23}) \in D_3\}, \oplus)$, all of them are isomorphic to the positive cone of the additive group of reals via $S_1$ and trivial isomorphism $z(0,0,0,d_{12},0,0) = (d_{12},0)$ and analogous ones.

Consonant generalizations are $S_{1-12} = (\{(d_1,0,0,d_{12},0,0) \in D_3\}, \oplus)$, $S_{1-13} = (\{(d_1,0,0,0,d_{13},0) \in D_3\}, \oplus)$, both of them include subsemigroups $S_1$ including neutral element 0, $S_{1-12}$ has also subsemigroup $S_{12}$ while $S_{1-13}$ has also subsemigroup $S_{13}$. Analogously there are subsemigroups of consonant BFs $S_{2-12}, S_{2-23}$, $S_{3-13}, S_{3-23}$.

All focal elements of a consistent generalization contain one of $\omega_i's$, thus we have $(\{(d_1,0,0,d_{12},d_{13},0) \in D_3\}, \oplus)$, $(\{(0,d_2,0,d_{12},0,d_{23}) \in D_3\}, \oplus)$, etc.

[4] $o$-isomorphic as in the case of $\mathbf{D}_0$ in fact, see Theorem 1. There is no ordering of elements of $\Omega_3$, thus we are either not interesting in ordering of algebras $S_i$ in this text.

Another generalizations of $S_i$'s contain BFs assigning masses only to singletons upto one element of $\Omega_3$. Thus we have quasi Bayesian subsemigroups mentioned in Theorem 2 (v).

Another generalizations are such that the only focal element containing one of elements of $\Omega_3$ is entire $\Omega_3$ (i.e. nothing is assigned to any proper subset of $\Omega_3$ containing the given element). Thus we have subsemigroups $\{(d_1, d_2, 0, d_{12}, 0, 0) \in D_3, \oplus)\}$, $\{(d_1, 0, d_3, 0, d_{13}, 0) \in D_3, \oplus)\}$ and $\{(0, d_2, d_3, 0, 0, d_{23}) \in D_3, \oplus)\}$.

Another generalizations are such that $Bel$ of one of elements of $\Omega$ is zero[5] thus for $\omega_3$ we have $\{(d_1, d_2, 0, d_{12}, d_{13}, 0) \in D_3, \oplus)\}$, $\{(d_1, d_2, 0, d_{12}, 0, d_{23}) \in D_3, \oplus)\}$ and analogically $\{(d_1, 0, d_3, d_{12}, d_{13}, 0) \in D_3, \oplus)\}$, $\{(d_1, 0, d_3, 0, d_{13}, d_{23}) \in D_3, \oplus)\}$ for $\omega_2$, and $\{(0, d_2, d_3, d_{12}, 0, d_{23}) \in D_3, \oplus)\}$, $\{(0, d_2, d_3, 0, d_{13}, d_{23}) \in D_3, \oplus)\}$ for $\omega_1$.

## 3.7 Subalgebras isomorphic to Dempster's semigroup $\mathbf{D}_0$

We can mention subsemigroups of quasi Bayesian BFs from Theorem 2 (iv), isomorphicity of which depends on Hypothesis 1, and those from Theorem 2 (v) again, i.e. non-exclusive BFs from $\mathbf{D}_{1-2}$, $\mathbf{D}_{1-3}$, $\mathbf{D}_{2-3}$, $(\{(a, b, 0, 0, 0, 0)\}, \oplus)$, $(\{(a, 0, c, 0, 0, 0)\}, \oplus)$, $(\{(0, b, c, 0, 0, 0)\}, \oplus)$; note that $m(\Omega_3) > 0$ here.

Reassigning bbms of $\Omega_3$ to union of other focal elements, we obtain subsemigroups $(\{(a, b, 0, 1-a-b, 0, 0)\}, \oplus)$, $(\{(a, 0, c, 0, 1-a-c, 0)\}, \oplus)$, $(\{(0, b, c, 0, 0, 1-b-c)\}, \oplus)$ which contain only exclusive BFs, thus they are completely out of our present interest.

More interesting are subsemigroups $\mathbf{D}_{1-23} = (\{(d_1, 0, 0, 0, 0, d_{23}) \in D_3\}, \oplus)$, $\mathbf{D}_{2-13} = (\{(0, d_2, 0, 0, d_{13}, 0) \in D_3\}, \oplus)$, $\mathbf{D}_{3-12} = (\{(0, 0, d_3, d_{12}, 0, 0) \in D_3\}, \oplus)$. Let us turn our focus to $\mathbf{D}_{1-23} = (\{(d_1, 0, 0, 0, 0, d_{23})\}, \oplus)$ now. There are only two the exclusive BFs $(1, 0, 0, 0, 0, 0), (0, 0, 0, 0, 0, 1)$ in $\mathbf{D}^+_{1-23}$, $\mathbf{D}_{1-23}$ has subsemigroups $S_1$, $S_{23}$ both isomorphic to the positive cone $\mathbf{Re}^{\geq 0}$, there is another subsemigroup $S_{1-23} = (\{(d_1, 0, 0, 0, 0, d_1) \in D_3\}, \oplus)$, it is isomorphic to $\mathbf{Re}^{\geq 0+}$ using $S$ and simple isomorphism $z : S_{1-23} \longrightarrow S$, such that $z(d_1, 0, 0, 0, 0, d_1) = (d_1, d_1)$. There is subgroup $G_{1-23} = (\{(d_1, 0, 0, 0, 0, 1-d_1)\}, \oplus, minus_{1-23}, (\frac{1}{2}, 0, 0, 0, 0, \frac{1}{2}))$, where $minus_{1-23}$ $(d, 0, 0, 0, 0, 1-d) = (1-d, 0, 0, 0, 0, d)$. The mapping $z$ we can use also as isomorphism $G_{1-23} \longrightarrow G$, where $z(d, 0, 0, 0, 0, 1-d) = (d, 1-d)$, hence $G_{1-23}$ is consequently isomorphic to the additive group or reals $\mathbf{Re}$. We can use mapping $z$ also as more general isomorphism $z : \mathbf{D}_{1-23} \longrightarrow \mathbf{D}_0$, such that $z(d_1, 0, 0, 0, 0, d_{23}) = (d_1, d_{23})$. Thus $\mathbf{D}_{1-23}$ is subsemigroup of $\mathbf{D}_3$ which is isomorphic to $\mathbf{D}_0$ using simple isomorphism $z$ which preserves values, hence $\mathbf{D}_{1-23}$ is Dempster's subsemigroup of $\mathbf{D}_3$. Analogously for Dempster's subsemigroups $\mathbf{D}_{2-13}$ and $\mathbf{D}_{3-12}$.

## 3.8 Summary of subalgebras

Although our overview of subalgebras of $\mathbf{D}_3$ is not complete due to complexity of the structure, we can summarize properties of important subalgebras of $\mathbf{D}_3$ as it follows:

**Theorem 3** *(i) Dempster's semigroup* $\mathbf{D}_3 = (D_3, \oplus, 0, U_3)$ *of non-exclusive BFs on* $\Omega_3$ *is commutative semigroup with neutral element* $0 = (0, 0, 0, 0, 0, 0)$ *(i.e. it is monoid), and with four other idempotents* $0' = U_3 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0)$*,* $(\frac{1}{2}, 0, 0, 0, 0, \frac{1}{2})$*,*

[5] Note, that for any $\omega_i$ the set of all such BFs is not semigroup.

$(0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0)$*, and* $(0, 0, \frac{1}{2}, \frac{1}{2}, 0, 0)$ *(there are no other idempotents there).*
*(iia) Subgroup of non-exclusive BBFs* $G_3 = (\{(a, b, c, 0, 0, 0) | a + b + c = 1, 0 < a, b, c\}, \oplus, "-", U_3)$ *and its subalgebras are subalgebras of* $\mathbf{D}_3$*.*
*(iib)* $G_{1-23} = (\{(d, 0, 0, 0, 0, 1-d)\}, \oplus, minus_{1-23}, (\frac{1}{2}, 0, 0, 0, 0, \frac{1}{2}))$*, where* $minus_{1-23}(d, 0, 0, 0, 0, 1-d) = (1-d, 0, 0, 0, 0, d)$ *is commutative (Abelian) group isomorphic to the additive group of reals* $\mathbf{Re}$*. Similarly for* $G_{2-13}$ *and* $G_{3-12}$*.*
*(iic)* $G_{2=3} = (\{(d, \frac{1-d}{2}, \frac{1-d}{2}, 0, 0, 0)\}, \oplus, minus_{2=3}, 0'_3)$*, where* $minus_{2=3}(d, \frac{1-d}{2}, \frac{1-d}{2}, 0, 0, 0) = (\frac{1-d}{1+3d}, \frac{2d}{1+3d}, \frac{2d}{1+3d}, 0, 0, 0)$*, is subgroup of* $G_3$*. Analogously for subgroups* $G_{1=3}$ *and* $G_{1=2}$*; (existence of isomorphisms of* $G_{2=3}$*,* $G_{1=3}$*,* $G_{1=2}$ *onto the additive group of reals* $\mathbf{Re}$ *is assumed by Hypothesis 1, but not proved).*
*(iiia) Besides of subsemigroups* $S_0, S_1, S_2, S_3, S_{1-2}, S_{1-3}, S_{2-3}$ *isomorphic to the positive cone of additive group of reals* $\mathbf{Re}^{\geq 0}$ *[13], there are other subsemigroups* $S_{12} = (\{(0, 0, 0, d_{12}, 0, 0) \in D_3\}, \oplus)$*,* $S_{13}$*, and* $S_{23}$ *isomorphic to* $\mathbf{Re}^{\geq 0}$*.*
*(iiib) There is another important subsemigroups* $S = (\{(a, a, a, b, b, b) \in D_3\}, \oplus)$ *generalizing* $S$ *from* $\mathbf{D}_0$*. There are consonant generalizations of* $S_i$*'s* $S_{1-12} = (\{(d_1, 0, 0, d_{12}, 0, 0) \in D_3\}, \oplus)$*,* $S_{1-13}$*,* $S_{2-12}$*,* $S_{2-23}$*,* $S_{3-13}$*,*$S_{3-23}$*, and another generalizations of* $S_i$*'s.*
*(iv) There are subsemigroups* $\mathbf{D}_{1-2=3}$*,* $\mathbf{D}_{2-1=3}$ *and* $\mathbf{D}_{3-1=2}$ *isomorphic to* $\mathbf{D}_0$ *assuming Hypothesis 1. Further there are subsemigroups* $\mathbf{D}_{1-2}$*,* $\mathbf{D}_{1-3}$ *and* $\mathbf{D}_{2-3}$ *isomorphic to* $\mathbf{D}_0$ *without BBFs.*
*(v) There are subsemigroups* $\mathbf{D}_{1-23}$*,* $\mathbf{D}_{2-13}$ *and* $\mathbf{D}_{3-12}$ *isomorphic to* $\mathbf{D}_0$*.*
*(vi)* $h$ *is homomorphism:* $(D_3, \oplus, 0, U_3) \longrightarrow (G_3, \oplus, "-", U_3)$*;* $h(Bel) = Bel \oplus 0' = Pl\_P(Bel)$ *i.e., the normalized plausibility of singletons [10].*

## 4 Ideas for future research and open problems

The presented introductive study opens a lot of interesting problems related to algebraic properties of belief functions on 3-element frame of discernment. Let us mention some them:

- Elaboration of properties of $\mathbf{D}_{3-0}$; ideas related to operation "*minus*" mentioned in [13].
- Study of subalgebras of $\mathbf{D}_3^+$ containing all BFs (both exclusive and non-exclusive) should follow.
- The very interesting and important is a open problem of existence of operation "*minus*" in $\mathbf{D}_3$ ($\mathbf{D}_3^+$).
- Study of homomorphisms; generalization of homomorphism $f$ and of all other homomorphismsm from [12].

## 5 Conclusion

Dempster's semigroup of belief functions on 3-element frame of discernment was elaborated in this contribution. Its substructures related to important classes of general belief functions were described and analyzed.

A basis for a solution of the questions coming from research of conflicting belief functions (e.g. the question of an existence of a generalization of Hájek-Valdés operation "*minus*") was established.

One of corner-stones for further study of conflicts between belief functions and for better understanding of a notion of belief function in general was laid.

# References

[1] Ayoun A., Smets Ph. (2001), Data association in multi-target detection using the transferable belief model. *Int. Journal of Intelligent Systems* **16** (10): 1167–1182.

[2] Cobb B. R., Shenoy P. P. (2003), A Comparison of Methods for Transforming Belief Functions Models to Probability Models. In: Nielsen T. D., Zhang N. L. (eds.) ECSQARU 2003, LNAI 2711: pp 255–266. Springer.

[3] Daniel M. (1994) More on Automorphisms of Dempster's Semigroup. In: *Proceedings of the 3-rd Workshop in Uncertainty in Expert Systems; WUPES'94*, pp 54 – 69. University of Economics, Prague.

[4] Daniel M. (1995) Algebraic structures related to Dempster-Shafer theory. In: Bouchon-Meunier B., Yager R. R., Zadeh L. A. (eds.) *Advances in Intelligent Computing - IPMU'94*. LNCS 945, 51–61, Springer, Heidelberg.

[5] Daniel M. (1996) Algebraic Properties of Structures Related to Dempster-Shafer Theory. In: Bouchon-Meunier, B., Yager, R. R., Zadeh, L. A. (eds.) *Proceedings IPMU'96*, pp 441 – 446. Universidad de Granada, Granada.

[6] Daniel M. (2000), Distribution of Contradictive Belief Masses in Combination of Belief Functions. In: B. Bouchon-Meunier, R. R. Yager, L. A. Zadeh (eds.) *Information, Uncertainty and Fusion*, pp 431–446. Kluwer Acad. Publ., Boston.

[7] Daniel M. (2004), Algebraic Structures Related to the Combination of Belief Functions. *Scientiae Mathematicae Japonicae* **60**: 245–255. *Scientiae Mathematicae Japonicae Online* **10**.

[8] Daniel M. (2005), Probabilistic Transformations of Belief Functions. In: Godo, L. (ed.) *ECSQARU 2005*. LNAI, vol. 3571, pp 539–551, Springer.

[9] Daniel M. (2010), Conflicts within and between Belief Functions. In: E. Hüllermeier, et al. (eds.) *IPMU 2010*. LNAI, vol. 6178, pp 696–705. Springer, Heidelberg.

[10] Daniel M. (2011), Non-conflicting and Conflicting Parts of Belief Functions. In: Coolen, F., de Cooman, G., Fetz, T., Oberguggenberger, M. (eds.) *ISIPTA'11; Proceedings of the 7th ISIPTA*, pp 149–158. Studia Universitätsverlag, Innsbruck.

[11] Daniel M. (2011), *Conflicts of Belief Functions.* Technical report V-1108, ICS AS CR, Prague.

[12] Daniel M. (2011), Morphisms of Dempster's Semigroup: A Revision and Interpretation. In: Barták, R. (ed.) *Proc. of 14th Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty CJS 2011*, pp 26–34. Matfyzpress, Prague.

[13] Daniel M. (2012), Introduction to an Algebra of Belief Functions on Three-element Frame of Discernment — A Quasi Bayesian Case. In: S. Greco et al. (eds.) *Advances in Computational Intelligence, IPMU 2012, Part III.* CCIS, vol. 299, pp 532–542. Springer, Berlin Heidelberg.

[14] Haenni R. (2008), Aggregating Referee Scores: an Algebraic Approach. *COMSOC'08, 2nd Int.Workshop on Comput.Social Choice*, Liverpool, UK.

[15] Hájek P., Havránek T., Jiroušek R. (1992), *Uncertain Information Processing in Expert Systems.* CRC Press, Boca Raton, Florida.

[16] Hájek P., Valdés (1991), Generalized algebraic foundations of uncertainty processing in rule-based expert systems (dempsteroids). *Computers and Artificial Intelligence* **10** (1): 29–42.

[17] Liu W. (2006), Analysing the degree of conflict among belief functions. Artificial Intelligence **170**: 909–924.

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

[19] Smets Ph. (2007), Analyzing the combination of conflicting belief functions. *Information Fusion* **8**: 387–412.

[20] Valdés J. J. (1987), *Algebraic and logical foundations of uncertainty processing in rule-based expert systems of Artificial Intelligence.* PhD Thesis, Czechoslovak Academy of Sci., Prague.

# Markov Degree of the Three-state Toric Homogeneous Markov Chain Model

**David Haws**
Computational Genetics
IBM
dchaws@gmail.com

**Abraham Martín del Campo**
Department of Mathematics
Texas A&M University
asanchez@math.tamu.edu

**Akimichi Takemura**
Department of Mathematical Informatics
University of Tokyo
takemura@stat.t.u-tokyo.ac.jp

**Ruriko Yoshida**
Department of Statistics
University of Kentucky
ruriko.yoshida@uky.edu

**Abstract**

For the three-state *toric homogeneous Markov chain (THMC) model* without loops, initial parameters, and time $T$, the size of the design matrix is $6 \times 3 \cdot 2^{T-1}$. In this paper, we study the behavior of the model polytope, the convex hull of the columns of its design matrix, when the time $T \geq 3$ is arbitrarily large and we show that the polytope is defined by 24 facets, that do not depend on $T$. From this, we deduce that the toric ideal associated with the design matrix is generated by binomials of degree at most 6. Our proof is based on a result due to Sturmfels, who gave a bound on the degree of the generators for a toric ideal, provided the normality of the corresponding toric variety. In our setting, we established the normality of the toric variety associated to the THMC model by studying the geometric properties of the model polytope. Moreover, we give a complete description of the facets for arbitrary $T$.

## 1 Introduction

A discrete time Markov chain, $X_t$ for $t = 1, 2, \ldots$, is a stochastic process with the Markov property, that is $P(X_{t+1} = y | X_1 = x_1, \ldots, X_{t-1} = x_{t-1}, X_t = x) = P(X_{t+1} = y | X_t = x)$ for any states $x$, $y$. Discrete time Markov chains have applications in several fields, such as physics, chemistry, information sciences, economics, finances, mathematical biology, social sciences, and statistics [7]. In this paper, we consider a discrete time Markov chain $X_t$ over a set of states $[S] = \{1, \ldots, S\}$, with $t = 1, \ldots, T$ ($T \geq 3$), focusing on the case $S = 3$.

Discrete time Markov chains are often used in statistical models to fit the observed data from a random physical process. Sometimes, in order to simplify the model, it is convenient to consider time-homogeneous Markov chains, where the transition probabilities do not depend on the time, in other words, when

$$P(X_{t+1}{=}y|X_t{=}x) = P(X_2{=}y|X_1{=}x) \quad \forall\, x,\, y \in [S] \text{ and for any } t = 1\ldots, T.$$

In order for a statistical model to reflect the observed data, it is necessary to verify if the model fits the data via a goodness-of-fit test. For instance, for the time-homogeneous Markov chain model, it is necessary to test if the assumption of time-homogeneity fits the observed data. In this paper, we study some properties of these algebraic relations for the toric homogeneous Markov chain (THMC) model, which is a slight generalization of the time-homogeneous Markov chain model, and which we explain as follows.

Let $\mathbf{w} = s_1 \cdots s_T$ denote a word of length $T$ on states $[S]$. Let $p(\mathbf{w})$ denote the likelihood of observing the word $\mathbf{w}$. Since the time-homogeneous Markov chain model assumes that the transition probabilities do not depend on time, we can write the likelihood as the product of probabilities

$$p(\mathbf{w}) = \pi_{s_1} p_{s_1,s_2} \cdots p_{s_{T-1},s_T}, \tag{1.1}$$

where, $\pi_{s_i}$ indicates the initial distribution at the first state, and $p_{s_i,s_j}$ are the transition probabilities from state $s_i$ to $s_j$. In the usual time-homogeneous Markov chain model it is assumed that the row sums of the transition probabilities are equal to one: $\sum_{j=1}^{S} p_{i,j} = 1$, $\forall i \in [S]$. On the other hand, in the toric homogeneous Markov model (1.1), the parameters $p_{i,j}$ are free and the row sums of the transition probabilities are assumed to be different to one. This simplifies the model as we can disregard some information from it. In many cases the parameters $\pi_{s_1}$ for the initial distribution are known, or sometimes these parameters are all constant, namely $\pi_1 = \pi_2 = \cdots = \pi_S = c$; in this situation it is no longer necessary to take them in consideration for the model. Another simplification that arises from practice is when the only transition probabilities considered are those between two different states, i.e. when $p_{i,j} = 0$ whenever $i = j$; this situation is referred as the THMC model without self-loops.

The THMC model is a toric model defined by a linear map with a matrix $A$. This matrix is called the *design matrix* of a model. In the THMC model, the goodness-of-fit test is encoded by polynomial relations among the probabilities $p_{s_i,s_j}$. The set of all these algebraic relations defines the *toric ideal* associated to the design matrix of the model. The test of goodness-of-fit is summarized by a *Markov basis*, a generating set for the toric ideal associate with the design matrix.

In [10], the authors provided a full description of the Markov bases for the THMC model in two states (i.e. when $S = 2$) which does not depend on $T$, even though the toric ideal lie on a polynomial ring with $2^T$ indeterminates. Inspired by their work, we study the algebraic and polyhedral properties of the Markov bases of the three-state THMC model without initial parameters and without self-loops for time $T$. As a result, we showed that for arbitrarily large time $T \geq 3$, the *model polytope* –the convex hull of the columns of the design matrix– has only 24 facets, that do not depend on $T$ and we provide a complete description of these facets. In addition, by showing the normality of the polytope, we deduced that the Markov bases of the model consist of binomials of degree at most 6.

The outline of this paper is as follows. In Section 2, we recall some definitions from Markov bases theory. In Section 3, we explicitly describe the hyperplane representation of the model polytope for the three-state THMC model for any time $T \geq 3$. In Section 4, we show that the model polytope is normal for arbitrary $T \geq 3$, this is equivalent to show that the semigroup generated by the columns of the design matrix is integrally closed. Finally, using these results, we prove the bound on the degree of the Markov bases in Section 5; and we conclude that section with some observations based on the analysis of our computational experiments.

# 2 Notation

Let $\langle S\rangle^T$ be the set of all words of length $T$ on states $[S]$ such that every word has no self-loops; that is, if $\mathbf{w} = (s_1, \dots, s_T) \in \langle S\rangle^T$ then $s_i \neq s_{i+1}$ for $i = 1, \dots, T-1$. We define $\mathcal{P}^*(\langle S\rangle^T)$ to be the set of all multisets of words in $\langle S\rangle^T$.

Let $\mathbb{V}\left(\langle S\rangle^T\right)$ be the real vector space with basis $\langle S\rangle^T$ and note that $\mathbb{V}\left(\langle S\rangle^T\right) \cong \mathbb{R}^{S(S-1)^T}$. We recall some definitions from the book of Pachter and Sturmfels [6]. Let $A = (a_{ij})$ be a non-negative integer $d \times m$ matrix with the property that all column sums are equal:

$$\sum_{i=1}^{d} a_{i1} = \sum_{i=1}^{d} a_{i2} = \cdots = \sum_{i=1}^{d} a_{im}.$$

Write $A = [a_1\ a_2\ \cdots\ a_m]$ where $a_j$ are the column vectors of $A$ and define $\theta^{\mathbf{a}_j} = \prod_{i=1}^{d} \theta_i^{a_{ij}}$ for $j = 1, \dots, m$. The *toric model* of $A$ is the image of the orthant $\mathbb{R}^d_{\geq 0}$ under the map

$$f : \mathbb{R}^d \to \mathbb{R}^m, \quad \theta \mapsto \frac{1}{\sum_{j=1}^{m} \theta^{a_j}} \left(\theta^{a_1}, \dots, \theta^{a_m}\right).$$

Here we have $d$ parameters $\theta = (\theta_1, \dots, \theta_d)$ and a discrete state space of size $m$. In our setting, the discrete space will be the set of all possible words on $[S]$ of length $T$ without self-loops ($\langle S\rangle^T$) and we can think of $\theta_1, \dots, \theta_d$ as the probabilities $p_{1,2}, p_{1,3}, \dots, p_{S-1,S}$.

In this paper, we focus on the THMC model without initial parameters and with no self-loops in three states, (i.e., $S = 3$), which is parametrized by 6 positive real variables: $p_{12}$, $p_{13}$, $p_{21}$, $p_{23}$, $p_{31}$, $p_{3,2}$. Thus, the number of parameters is $d = 6$ and the size of the discrete space is $m = 6^{T-1}$, which is precisely the number of words in $\langle 3\rangle^T$. The model we study is thus the toric model represented by the $6 \times 6^{T-1}$ matrix $A^T$, which will be referred to as the *design matrix* for the model on 3 states with time $T$. The rows of $A^T$ are indexed by elements in $\langle 3\rangle^2$ and the columns are indexed by words in $\langle 3\rangle^T$. The entry of $A^T$ indexed by row $\sigma_1\sigma_2 \in \langle 3\rangle^2$, and column $\mathbf{w} = (s_1, \dots, s_T) \in \langle 3\rangle^T$ is equal to the cardinality of the set $\{\, i \in \{1, \dots, T-1\} \mid \sigma_1\sigma_2 = s_i s_{i+1} \,\}$.

**Example 2.1.** *Ordering $\langle 3\rangle^2$ and $\langle 3\rangle^T$ lexicographically, and letting $T = 4$, the matrix $A^4$ is:*

| | 1212 | 1213 | 1231 | 1232 | 1312 | 1313 | 1321 | 1323 | 2121 | 2123 | 2131 | 2132 | 2312 | 2313 | 2321 | 2323 | 3121 | 3123 | 3131 | 3132 | 3212 | 3213 | 3231 | 3232 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12 | 2 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 13 | 0 | 1 | 1 | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| 21 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 23 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| 31 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 2 | 1 | 1 | 0 |
| 32 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 2 |

## 2.1 Sufficient statistics, ideals, and Markov basis

Let $A^T$ be the design matrix for the THMC model without initial parameters and with no self-loops. The column of $A^T$ indexed by $\mathbf{w} \in \langle 3 \rangle^T$ is denoted by $\mathbf{a}^T_{\mathbf{w}}$. Thus, by extending linearly, the map $A^T : \mathbb{V}(\langle 3 \rangle^T) \to \mathbb{R}^6$ is well-defined.

Let $W = \{w_1, \ldots, w_N\} \in \mathcal{P}^*(\langle 3 \rangle^T)$ where we regard $W$ as observed data which can be summarized in the *data vector* $\mathbf{u} \in \mathbb{N}^{6^{T-1}}$. We index $\mathbf{u}$ by words in $\langle 3 \rangle^T$, so the coordinate representing for the word $\mathbf{w}$ in the vector $\mathbf{u}$ is denoted by $u_{\mathbf{w}}$, and its value is the number of words in $W$ equal to $\mathbf{w}$. Note since $A^T$ is linear then $A^T\mathbf{u}$ is well-defined. We also adopt the notation $A^T(W) := A^T\mathbf{u}$. For $W$ from $\mathcal{P}^*(\langle 3 \rangle^T)$, let $\mathbf{u}$ be its data vector, the *sufficient statistics* for the model are stored in the vector $A^T\mathbf{u}$. Often the data vector $\mathbf{u}$ is also referred to as a *contingency table*, in which case $A^T\mathbf{u}$ is referred to as the *marginals*.

The design matrix $A^T$ above defines a toric ideal which is of central interest in this paper, as their set of generators are in bijection with the Markov bases. The toric ideal $I_{A^T}$ is defined as the kernel of the homomorphism of polynomial rings $\psi : \mathbb{C}[\{\, p(\mathbf{w}) \mid \mathbf{w} \in \langle S \rangle^T \,\}] \to \mathbb{C}[\{\, p_{ij} \mid i, j \in [3],\, i \neq j \,\}]$ defined by $\psi(p(\mathbf{w})) = p_{s_1,s_2} \cdots p_{s_{T-1},s_T}$, where $\{\, p(\mathbf{w}) \mid \mathbf{w} \in \langle S \rangle^T \,\}$ is regarded as a set of indeterminates.

The set of all contingency tables (data vectors) satisfying a given set of marginals $\mathbf{b} \in \mathbb{Z}^d_{\geq 0}$ is called a *fiber* which we denote by $\mathcal{F}_{\mathbf{b}} = \{\, \mathbf{x} \in \mathbb{Z}^m_{\geq 0} \mid A^T(\mathbf{x}) = \mathbf{b} \,\}$. A *move* $\mathbf{z} \in \mathbb{Z}^m$ is an integer vector satisfying $A^T(\mathbf{z}) = 0$. A *Markov basis* for our model defined by the design matrix $A^T$ is defined as a finite set $\mathcal{Z}$ of moves satisfying that for all $\mathbf{b}$ and all pairs $\mathbf{x}, \mathbf{y} \in \mathcal{F}_{\mathbf{b}}$ there exists a sequence $\mathbf{z}_1, \ldots, \mathbf{z}_K \in \mathcal{Z}$ such that

$$\mathbf{y} = \mathbf{x} + \sum_{k=1}^{K} \mathbf{z}_k, \quad \text{with } \mathbf{x} + \sum_{k=1}^{l} \mathbf{z}_k \geq \mathbf{0}, \quad \text{for all } l = 1, \ldots, K.$$

A *minimal Markov basis* is a Markov basis which is minimal in terms of inclusion. See Diaconis and Sturmfels[1] for more details on Markov bases and their toric ideals.

## 2.2 State Graph

We give here a useful tool to visualize multisets of $\mathcal{P}^*(\langle 3 \rangle^T)$. Given any multiset $W \in \mathcal{P}^*(\langle 3 \rangle^T)$ we consider the directed multigraph called the *state graph* $G(W)$. The vertices of $G(W)$ are given by the states $[3]$ and the directed edges $i \to j$ are given by the transitions from state $i$ to $j$ in $\mathbf{w} \in W$. Thus, we regard $\mathbf{w} \in W$ as a path with $T-1$ edges (steps, transitions) in $G(W)$.

See Figure 1 for an example of the state graph $G(W)$ of the multiset $W = \{(12132), (12321)\}$ of paths with length 4. Notice that the state graph in Figure 1 is the same for another multiset of paths $\overline{W} = \{(13212), (21232)\}$.

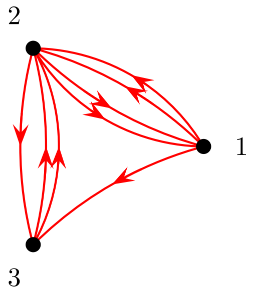


Figure 1: The state graph $G(W)$ of $W = \{(12132), (12321)\}$. Also the state graph $G(\overline{W})$ where $\overline{W} = \{(13212), (21232)\}$.

From the definition of state graph it is clear that it records the transitions in a given multiset of words and we state the following proposition.

**Proposition 2.2** (Proposition 2.1 in [4]). *Let $A$ be the design matrix for the THMC, and $W, \overline{W} \in \mathcal{P}^*(\langle S\rangle^T)$. Then $A(W) = A(\overline{W})$ if and only if $G(W) = G(\overline{W})$.*

Throughout this paper we alternate between terminology of the multisets of words $W$ and the graph it defines $G(W)$.

## 2.3 Semigroup and Smith Normal Form

Given an integer matrix $A \in \mathbb{Z}^{d\times m}$ we associate an integer lattice $\mathbb{Z}A = \{n_1\mathbf{a}_1 + \cdots + n_m\mathbf{a}_m \mid n_i \in \mathbb{Z}\}$. We can also associate the semigroup $\mathbb{N}A := \{n_1\mathbf{a}_1 + \cdots + n_m\mathbf{a}_m \mid n_i \in \mathbb{N}\}$. We say that the semigroup $\mathbb{N}A$ is *normal* when $\mathbf{x} \in \mathbb{N}A$ if and only if there exist $\mathbf{y} \in \mathbb{Z}^d$ and $\alpha \in \mathbb{R}^d_{\geq 0}$ such that $\mathbf{x} = A\mathbf{y}$ and $\mathbf{x} = A\alpha$. The set of vectors $A\alpha$ is called the *saturation* of $\tilde{\mathbb{N}}A$. See [5, 9] for more details on normality.

For an integer matrix $A \in \mathbb{Z}^{d\times m}$, we consider the Smith normal form $D$ of $A$, which is a diagonal matrix $D$ for which there exist unimodular matrices $U \in \mathbb{Z}^{d\times d}$ and $V \in \mathbb{Z}^{m\times m}$, such that $UAV = D$. The Smith normal form encodes the $\mathbb{Z}$-module structure of the abelian group $\mathbb{Z}A := \{n_1A_1 + \cdots + n_mA_m \mid n_i \in \mathbb{Z}\}$. Some additional material about the Smith normal form for matrices with entries over a PID can be found in the book of C. Yap [11]. The Smith normal form is important for studying the normality of the toric ideal associated to the model.

**Proposition 2.3** (Proposition 3.2 in [4]). *Let $A^{S,T}$ be the design matrix for the THMC without initial parameters and no self-loops on $S > 1$ states with time $T > 0$. For $S \geq 3$ and $T \geq 4$, the Smith normal form of the design matrix $A^{S,T}$ is $D = \mathrm{diag}(1, \ldots, 1, T-1)$.*

# 3 Facets of the design polytope

## 3.1 Polytopes

We recall some necessary definitions from polyhedral geometry and we refer the reader to the book of Schrijver [8] for more details. The *convex hull* of $\{\mathbf{a}_1, \ldots, \mathbf{a}_m\} \subset \mathbb{R}^n$ is defined as

$$\operatorname{conv}(\mathbf{a}_1, \ldots, \mathbf{a}_m) := \left\{ \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} = \sum_{i=1}^{m} \lambda_i \mathbf{a}_i,\ \sum_{i=1}^{m} \lambda_i = 1,\ \lambda_i \geq 0 \right\}.$$

A *polytope* $P$ is the convex hull of finitely many points. We say $F \subseteq P$ is a *face* of the polytope $P$ if there exists a vector $\mathbf{c}$ such that $F = \arg\max_{\mathbf{x} \in P} \mathbf{c} \cdot \mathbf{x}$. Every face $F$ of $P$ is also a polytope. If the dimension of $P$ is $d$, a face $F$ is a *facet* if it is of dimension $d-1$. For $k \in \mathbb{N}$, we define the $k$-th dilation of $P$ as $kP := \{\, k\mathbf{x} \mid \mathbf{x} \in P, \}$. A point $\mathbf{x} \in P$ is a *vertex* if and only if it can not be written as a convex combination of points from $P \backslash \{\mathbf{x}\}$.

The *cone* of $\{\mathbf{a}_1, \ldots, \mathbf{a}_m\} \subset \mathbb{R}^n$ is defined as

$$\operatorname{cone}(\mathbf{a}_1, \ldots, \mathbf{a}_m) := \left\{ \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} = \sum_{i=1}^{m} \lambda_i \mathbf{a}_i,\ \lambda_i \geq 0 \right\}.$$

Thus, $\operatorname{cone}(A)$ denote the cone over the columns of the matrix $A$. We are interested in the polytope given by the convex hull of the columns of the design matrices of our model. We define the *design polytope* $P^T$ as the convex hull $\operatorname{conv}(A^T)$ of the columns of the design matrix $A^T$. Notice that in this case, $P^T$ has dimension 6.

If $\mathbf{x} \in \mathbb{R}^6$, we index $\mathbf{x}$ by $\{\, ij \mid 1 \leq i, j \leq 3,\ i \neq j \,\}$. We define $\mathbf{e}_{ij} \in \mathbb{R}^6$ to be the vector of all zeros, except 1 at index $ij$. We also adopt the notation $x_{i+} := \sum_j x_{ij}$ and $x_{+i} := \sum_j x_{ji}$. For any $\mathbf{x} \in \mathbb{N}^6$ we can define a directed multigraph $G(\mathbf{x})$ on three vertices, where there are $x_{ij}$ directed edges from vertex $i$ to vertex $j$. One would like to identify the vectors $\mathbf{x} \in \mathbb{N}^6$ for which the graph $G(\mathbf{x})$ is a state graph. Nevertheless, observe that $x_{i+}$ is the out-degree of vertex $i$ and $x_{+i}$ is the in-degree of vertex $i$ with respect to $G(\mathbf{x})$.

We now give some properties which will be used later for describing the facets of the design polytope $P^T$ given by the design matrix for our model, and to prove normality of the semigroup associated with the design matrix.

**Proposition 3.1** (Proposition 5.1 in [4])**.** *Let $A^T$ be the design matrix for the THMC without loops and initial parameters. If $\mathbf{x} \in \mathbb{Z}A^T \cap \operatorname{cone}(A^T)$ then $\sum_{i \neq j} x_{ij} = k(T-1)$ for some $k \in \mathbb{N}$ and $|x_{i+} - x_{+i}| \leq k$ for all $i \in \{1, 2, 3\}$.*

Proposition 3.1 states that for $\mathbf{x} \in \mathbb{Z}A^T \cap \operatorname{cone}(A^T)$ the multigraph $G(\mathbf{x})$ will have in-degree and out-degree bounded by $\|\mathbf{x}\|_1/(T-1)$ at every vertex. This implies nice properties when $\|\mathbf{x}\|_1 = (T-1)$. Recall that a path in a directed multigraph is *Eulerian* if it visits every edge only once.

**Proposition 3.2** (Proposition 5.2 in [4]). *If $G$ is a directed multigraph on three vertices, with no self-loops, $T-1$ edges, and satisfying*

$$|G_{i+} - G_{+i}| \leq 1 \qquad i = 1, 2, 3;$$

*then, there exists an Eulerian path in $G$.*

Note that every word $\mathbf{w} \in \mathcal{P}^*(\langle 3 \rangle^T)$ gives an Eulerian path in $G(\{\mathbf{w}\})$ containing all edges. Conversely, for every multigraph $G$ with an Eulerian path containing all edges, there exists $\mathbf{w} \in \mathcal{P}^*(\langle 3 \rangle^T)$ such that $G(\{\mathbf{w}\})= G$. More specifically, $\mathbf{w}$ is the Eulerian path in $G(\{\mathbf{w}\})$. Throughout this paper we use the terms *path* and *word* interchangeably.

**Lemma 3.3** (Lemma 5.2 in [4]). *Let $A^T$ be the design matrix for the THMC. If $T \geq 4$, then* $\operatorname{conv}(A^T) \cap \mathbb{Z}^6 = A^T$*, where the right hand side is taken as the set of columns of the matrix $A^T$.*

We define

$$H_{k(T-1)} := \left\{ \mathbf{x} \in \mathbb{R}^6 \mid \sum_{i \neq j} x_{ij} = k(T-1) \right\}.$$

**Proposition 3.4** (Proposition 5.3 in [4]). *Let $A^T$ be the design matrix for the THMC without initial parameters and no loops.*

1. *For $T \geq 4$ and $k \in \mathbb{N}$,*

$$k \operatorname{conv}(A^T) = \operatorname{cone}(A^T) \cap H_{k(T-1)}.$$

2. *For $T \geq 4$,*

$$\operatorname{cone}(A^T) \cap \mathbb{Z}A^T = \bigoplus_{k=0}^{\infty} \left( k \operatorname{conv}(A^T) \cap \mathbb{Z}^6 \right).$$

## 3.2 Facets

Here we summarize all the inequalities in their original form and in their inhomogeneous form. Below we only present one of six of the inequalities with the understanding that for each case that any permutation of the labels $\{1, 2, 3\}$ gives another facet. The inhomogeneous form is derived by substituting the equality $n(T-1) = x_{12} + x_{13} + x_{21} + x_{23} + x_{31} + x_{32}$ into the original form. Inhomogeneous form is essential for proving the normality of semigroup associated with the design matrix $A^T$.

For any $T \geq 5$ homogeneous

$$\mathbf{c} = [1, 0, 0, 0, 0, 0] \cdot \mathbf{x} \geq 0$$

For any $T \geq 5$, homogeneous

$$\mathbf{c} = [T, T, -(T-2), 1, -(T-2), 1)] \cdot \mathbf{x} \geq 0$$

inhomogeneous

$$\mathbf{c} = [1, 1, -1, 0, -1, 0] \cdot \mathbf{x} \geq -n.$$

For any $T$ odd, $T \geq 5$, homogeneous

$$\mathbf{c} = [1, 1, -1, -1, 1, 1] \cdot \mathbf{x} \geq 0.$$

For any $T \geq 4$ of the form $T = 3k + 1$, $k \geq 1$, homogeneous

$$\mathbf{c} = [2, -1, -1, -1, 2, 2] \cdot \mathbf{x} \geq 0.$$

For any $T \geq 5$ of the form $T = 3k + 2$, $k \geq 1$, homogeneous

$$\mathbf{c} = [2k + 1, -k, -k, -k, 2k + 1, 2k + 1] \cdot \mathbf{x} \geq 0$$

inhomogeneous

$$x_{12} + x_{31} + x_{32} \geq kn = \frac{T-2}{3}n.$$

For any $T \geq 6$, $T$:even, homogeneous

$$\mathbf{c} = [\frac{3}{2}T - 1, \frac{T}{2}, -\frac{T}{2} + 1, -\frac{T}{2} + 1, -\frac{T}{2} + 1, \frac{T}{2}] \cdot \mathbf{x} \geq 0$$

inhomogeneous

$$3x_{12} + x_{13} - x_{21} - x_{23} - x_{31} + x_{32} \geq -n.$$

For $T = 6k + 3$, homogeneous

$$\mathbf{c} = [5k + 2, 2k + 1, -4k - 1, -k, -k, 2k + 1] \cdot \mathbf{x} \geq 0$$

inhomogeneous

$$2x_{12} + x_{13} - x_{21} + x_{32} \geq \frac{T-3}{3}n.$$

For $T = 6k$, homogeneous

$$\mathbf{c} = [10k - 1, 4k, -8k + 2, -2k + 1, -2k + 1, 4k] \cdot \mathbf{x} \geq 0$$

inhomogeneous

$$2x_{12} + x_{13} - x_{21} + x_{32} \geq \frac{T-3}{3}n.$$

## 3.3 There are only 24 facets

In the previous section, we give 24 facets of the polytope $P^T$ for every $T \geq 3$, where death of the 24 facets depend on $T \mod 6$. Here, we discuss how these 24 facets are enough to describe the polytope $P^T$ (the convex hull of the columns of $A^T$), depending on $T$. Let $C^T := \text{cone}(A^T)$.

Recall that the columns of $A^T$ are on the following hyperplane

$$H_T = \{(x_{12}, \ldots, x_{32}) \mid T - 1 = x_{12} + \cdots + x_{32}\}.$$

Then it is clear by Proposition 3.4 that

$$P^T = C^T \cap H_T.$$

Let $\mathcal{F}_T$ denote the set of facets of the pointed cone $C^T$. Then the facets $F$ of $P^T$ (within $H_T$) are of the form $F \subseteq H_T, F \subseteq \mathcal{F}_T$.

For every $T$, let $\tilde{\mathcal{F}}_T$ denote the 24 facets prescribed in the previous section, and let $\mathcal{F}_T$ denote the set of all facets of $P^{3,T}$. Therefore we have a certain subset $\tilde{\mathcal{F}}_T \subset \mathcal{F}_T$ and we need to show that $\tilde{\mathcal{F}}_T = \mathcal{F}_T$. Let $\tilde{\mathcal{C}}_T$ denote the polyhedral cone defined by $\tilde{\mathcal{F}}_T$. It follows that $\tilde{\mathcal{C}}_T \supset \mathcal{C}_T$. Note that $\tilde{\mathcal{F}}_T = \mathcal{F}_T$ if and only if $\tilde{\mathcal{C}}_T = \mathcal{C}_T$. Also let

$$\tilde{P}_T = \tilde{\mathcal{C}}_T \cap H_T.$$

Then $\tilde{P}_T \supset P^T$ and $\tilde{P}_T = P^T$ if and only if $\tilde{\mathcal{C}}_T = \mathcal{C}_T$.

The above argument shows that to prove $\tilde{\mathcal{F}}_T = \mathcal{F}_T$ it suffices to show that

$$\tilde{P}_T \subset P^T. \tag{3.1}$$

Let $\tilde{V}_T$ be the set of vertices of $\tilde{P}_T$. Then in order to show (3.1), it suffices to show that

$$\tilde{V}_T \subset P^T.$$

Hence, if we can obtain explicit expressions of the vertices of $\tilde{V}_T$ and can show that each vertex belongs to $P^T$, we are done.

In the previous section, we used only the condition $T-1 = x_{12} + \cdots + x_{32}$ to settle the equivalence between the homogeneous and inhomogeneous inequalities defining the 24 the facets of $P^T$. Hence the homogeneous and the inhomogeneous inequalities are equivalent on $H_T$. Therefore, for each $r = 0, \ldots, 5$, there exists a polyhedral region defined by 24 fixed affine half-spaces, say $Q^r$, such that

$$\tilde{\mathcal{P}}_T = Q^r \cap H_T, \quad T = 6k + r, k = 1, 2, \ldots$$

Since $Q^r$ is a polyhedral region it can be written as a Minkowski sum of a polytope $P^r$ and a cone $C^r$:

$$Q^r = P^r + C^r.$$

Please note that $r$ is modulo 6, but $T$ is not. Recall the Minkowski sum of two sets $A, B \subseteq \mathbb{R}^n$ is simply $\{\, a+b \mid a \in A,\, b \in B \,\}$. For each vertex $\mathbf{v}$ of $P^r$ and each extreme ray $\mathbf{e}$ of $C^r$ let $l_{\mathbf{v},\mathbf{e}}$ denote the half-line emanating from $\mathbf{v}$ in the direction $\mathbf{e}$:

$$l_{\mathbf{v},\mathbf{e}} = \{\mathbf{v} + t\mathbf{e} \mid t \geq 0\}$$

Given the explicit expressions of $v$ and $e$ we can solve

$$[1,1,1,1,1,1] \cdot (\mathbf{v} + t\mathbf{e}) = T - 1$$

for $t$ and get

$$t := t(T, \mathbf{v}, \mathbf{e}) = \frac{T - 1 - [1, \ldots, 1]\mathbf{v}}{(1, \ldots, 1)\mathbf{e}}.$$

Then $v + t(T, \mathbf{v}, \mathbf{e})\mathbf{e} \in H_T$. Note that

$$\tilde{V}_T \subset \{\mathbf{v} + t(T, \mathbf{v}, \mathbf{e})\mathbf{e} \mid \mathbf{v} : \text{vertex of } P^r,\ \mathbf{e} : \text{extreme ray of } C^r\}.$$

Also clearly

$$\{v + t(T, \mathbf{v}, \mathbf{e})\mathbf{e} \mid \mathbf{v} : \text{vertex of } P^r,\ \mathbf{e} : \text{extreme ray of } C^r\} \in \tilde{P}_T = \text{conv}(\tilde{V}_T).$$

The above argument shows that for proving $\tilde{\mathcal{F}}_T = \mathcal{F}_T$ it suffices to show that

$$\{\mathbf{v} + t(T, \mathbf{v}, \mathbf{e})\mathbf{e} \mid \mathbf{v} : \text{vertex of } P^r, \mathbf{e} : \text{extreme ray of } C^r\} \in P^T. \qquad (3.2)$$

For proving (3.2) the following lemma is useful.

**Lemma 3.5.** *Let $\mathbf{v} \in P^r$ and $\mathbf{e} \in C$. If $\mathbf{v} + t(T, \mathbf{v}, \mathbf{e})\mathbf{e} \in P^T \cap \mathbb{Z}^6$ for some $T$, then $\mathbf{v} + t(T + 6k, \mathbf{v}, \mathbf{e})\mathbf{e} \in P^{T+6k}$ for all $k \geq 0$.*

*Proof.* If $\mathbf{x} := \mathbf{v} + t(T, \mathbf{v}, \mathbf{e})\mathbf{e} \in P^T \cap \mathbb{Z}^6$ for some $T$ then $\mathbf{x}$ corresponds to a path of length $T$ on three states with no loops (word in $\langle 3 \rangle^T$). Suppose $\mathbf{e}$ is a two-loop (three-loop) e.g. 121 (1231). Then $\mathbf{x} + (3k)\mathbf{e} \in P^{T+6k}$ ($\mathbf{x} + (2k)\mathbf{e} \in P^{T+6k}$). That is, since $\mathbf{x}$ is an integer point (a path) contained in $P^T$, we can simply add three (or two depending on the loop) copies of the loop $\mathbf{e}$ and we will be guaranteed to have a path of the correct length meaning it will be contained in $P^{T+6}$. □

By this lemma we need to compute $C^r$ only for some special small $T$'s. We computed all vertices and all rays for the cases $T = 12, 7, 20, 9, 16, 11$. The software to generate the design matrices can be found at `https://github.com/dchaws/GenWordsTrans` and the design matrices and some other material can be found at `http://www.davidhaws.net/THMC.html`. By our computational result and Lemma 3.5 we verified the following proposition.

**Proposition 3.6.** *The rays of the cones $C^r$ for $r = 0, \ldots, 5$ are ((1,0,1,0,0,0), (1,0,0,1,1,0), (0,1,1,0,0,1), (0,1,0,0,1,0), (0,0,0,1,0,1)). In terms of the state graph, the rays correspond to the five loops 121, 131, 232, 1231, and 1321.*

Note that $C^r$, $r = 0, \ldots, 5$ are common and we denote them as $C$ hereafter. Also note that the rays of the cone $C^r$ are very simple. Proposition 3.6 implies the following theorem.

**Theorem 3.7.** *The 24 facets given in above (depending on $T \mod 6$) are all the facets of $P^T = \text{conv}(A^T)$.*

# 4 Normality of the semigroup

From the definition of normality of a semigroup defined in Section 2.3, the semigroup $\mathbb{N}A^T$ defined by the design matrix is normal if it coincides with the elements in both, the integer lattice $\mathbb{Z}A^T$ and the cone $\text{cone}(A^T)$.

In this section, we provide an inductive prove on the normality of the semigroup $\mathbb{N}A^T$ for arbitrary $T$.

**Theorem 4.1.** *The semigroup $\mathbb{N}A^T$ is normal for any $T \in \mathbb{N}$.*

See [3] for the proof.

# 5 Discussion

In this paper, we considered only the situation of the toric homogeneous Markov chain (THMC) model (1.1) for $S = 3$, with the extra assumption of having non-zero transition probabilities only when the transition is between two different states. In this setting, we described the hyperplane representations of the design polytope for any $T \geq 4$, and from this representation we showed that the semigroup generated by the columns of the design matrix $A^T$ is normal.

We recall from Lemma 4.14 in [9], that a given set of integer vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ is a graded set, if there exists $\mathbf{w} \in \mathbb{Q}^{S^2}$ such that $\mathbf{a}_i \cdot \mathbf{w} = 1$. In our setting, the set of columns of the design matrix $A^T$ is a graded set, as each of its columns add up to $T - 1$, so we let $\mathbf{w} = (\frac{1}{T-1}, \ldots, \frac{1}{T-1})$.

In his same book, Sturmfels provided a way to bound the generators of the toric ideal associated to an integer matrix $A$, the precise statement is the following.

**Theorem 5.1** (Theorem 13.14 in[9])**.** *Let $A \subset \mathbb{Z}^d$ be a graded set such that the semigroup generated by the elements in $A$ is normal. Then the toric ideal $I_A$ associate with the set $A$ is generated by homogeneous binomials of degree at most $d$.*

In our setting, Theorem 4.1 demonstrates the normality of the semigroup generated by the columns of the design matrix $A^T$, so as a consequence we obtain the following theorem:

**Theorem 5.2.** *For $S = 3$ and for any $T \geq 4$, a Markov basis for the toric ideal $I_{A^T}$ associated to the THMC model (without loops and initial parameters) consists of binomials of degree at most 6.*

The bound provided by Theorem 5.2 seems not to be sharp, in the sense that there exists Markov basis whose elements have degree strictly less than 6. In our computational experiments, we found evidence that more should be true. Our observations hold in a more general setting. For any $S$ and $T$, let $A^{S,T}$ denote the design matrix for the THMC model in $S$ states.

**Conjecture 5.3.** *Fix $S \geq 3$; then, for every $T \geq 4$, there is a Markov basis for the toric ideal $I_{A^{S,T}}$ consisting of binomials of degree at most $S-1$, and there is a Gröbner basis with respect to some term ordering consisting of binomials of degree at most $S$.*

# References

[1] Persi Diaconis and Bernd Sturmfels. Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26(1):363–397, 1998.

[2] Hisayuki Hara and Akimichi Takemura. A markov basis for two-state toric homogeneous markov chain model without initial paramaters. *Journal of Japan Statistical Society*, 41:33–49, 2011.

[3] D. Haws, A. Martín del Campo, A. Takemura, and R. Yoshida, 2012.

[4] David Haws, Abraham Martin del Campo, and Ruriko Yoshida. Degree bounds for a minimal markov basis for the three-state toric homogeneous markov chain model. *Proceedings of the Second CREST–SBM International Conference, "Harmony of Grobner Bases and the Modern Industrial Society"*, pages 99 – 116, 2012.

[5] Ezra Miller and Bernd Sturmfels. *Combinatorial commutative algebra*. Graduate texts in mathematics. Springer, 2005.

[6] Lior Pachter and Bernd Sturmfels. *Algebraic Statistics for Computational Biology.* Cambridge University Press, Cambridge, UK, 2005.

[7] Etienne Pardoux. *Markov Processes and Applications: Algorithms, Networks, Genome and Finance*. Wiley Series in Probability and Statistics Series. Wiley, John & Sons, Incorporated, 2009.

[8] Alexandeer Schrijver. *Theory of Linear and Integer Programming.* John Wiley & Sons, Inc., New York, NY, USA, 1986.

[9] Bernd Sturmfels. *Gröbner Bases and Convex Polytopes*, volume 8 of *University Lecture Series*. American Mathematical Society, Providence, RI, 1996.

[10] Akimichi Takemura and Hisayuki Hara. Markov chain monte carlo test of toric homogeneous markov chains. *Statistical Methodology. doi:10.1016/j.stamet.2011.10.004.*, 9:392–406, 2012.

[11] Chee-Keng Yap. *Fundamental problems of algorithmic algebra*. Oxford University Press, 2000.

# Function Values of Graded Ill-known Sets

**Masahiro Inuiguchi**
Department of Systems Innovation
Graduate School of Engineering Science
Osaka University
inuiguti@sys.es.osaka-u.ac.jp

**Abstract**

In this paper, we investigate the extension principle for graded ill-known sets. Because a graded ill-known set is characterized by a possibility distribution in the power set, calculations of function values of graded ill-known sets are generally complex. To reduce the complexity, lower and upper approximations of a given graded ill-known set are used at the expense of precision. We give a necessary and sufficient condition that lower and upper approximations of function values of graded ill-known sets are obtained as function values of lower and upper approximations of graded ill-known sets.

## 1 Introduction

As a model to represent a set whose members known partially, the graded ill-known set [1, 2] is proposed. A graded ill-known set is characterized by a possibility distribution in the power set of the universe. It can be seen as a possibilistic counterpart of random set [3] and as an extension of possibility distribution [4]. Then there would be a lot of research topics about graded ill-known sets in analogy with random sets and possibility theory.

In this paper, we investigate the function calculations of graded ill-known sets. The function calculations of graded ill-known sets can be done though an extension principle. The extension principle, generalizing a function of real numbers to the function of graded ill-known sets, is indispensable for the applications of graded ill-known sets in decision making and data analysis under uncertainty. Because a graded ill-known set is defined in the power set, the calculations of function values of graded ill-known sets are generally complex. To reduce the complexity, lower and upper approximations of a given graded ill-known set are often used at the expense of precision. Using those approximations, we obtain the approximated values, or exact values in some special cases, are calculated on the universe.

In this paper, we give a necessary and sufficient condition that lower and upper approximations of function values of graded ill-known sets are obtained by function values of lower and upper approximations of graded ill-known sets. We demonstrate some cases where the condition is satisfied.

Graded ill-known sets are briefly introduced in Section 2. Extension principle for graded ill-known sets and previous results are described in Section 3. In Section 4, the main result is shown and applied to a few cases. Concluding remarks are given in Section 5.

## 2 Graded Ill-known Sets

Let $X$ be a universe. Let $\boldsymbol{A}$ be a crisp set whose members are not known exactly. To represent such an ill-known set, collecting possible realizations of $\boldsymbol{A}$, we obtain the following family:

$$\mathcal{A} = \{A_1, A_2, \ldots\}, \tag{1}$$

where $A_i$ is a crisp set such that $\boldsymbol{A} = A_i$ is possible.

Given $\mathcal{A}$, we obtain a set of elements which certainly belong to $A$, say $A^-$ and a set of elements which possibly belong to $A$, say $A^+$ are defined as

$$A^- = \bigcap \mathcal{A} = \bigcap_{i=1,2,\ldots} A_i, \quad A^+ = \bigcup \mathcal{A} = \bigcup_{i=1,2,\ldots} A_i. \tag{2}$$

We call $A^-$ and $A^+$ "the lower approximation" of $\boldsymbol{A}$ and "the upper approximation" of $\boldsymbol{A}$, respectively.

In the real world, we may know sure members and sure non-members only. In other words, we know the lower approximation $A^-$ as a set of sure members and the upper approximation $A^+$ as a complementary set of sure non-members. Given $A^-$ and $A^+$ (or equivalently, the complement of $A^+$), we obtain a family $\hat{\mathcal{A}}$ of possible realizations of $\boldsymbol{A}$ as

$$\hat{\mathcal{A}} = \{A_i \mid A^- \subseteq A_i \subseteq A^+\}. \tag{3}$$

We note that $A^-$ and $A^+$ are recovered by applying (2) to the family $\hat{\mathcal{A}}$ induced from $A^-$ and $A^+$ by (3). On the other hand, $\mathcal{A}$ cannot be always recovered by applying (3) to $A^-$ and $A^+$ defined by (2).

If all $A_i$'s of (1) are not regarded as equally possible, we may assign a possibility degree $\pi_{\mathcal{A}}(A)$ to each $A \subseteq X$ so that

$$\exists A \subseteq X, \quad \pi(A) = 1. \tag{4}$$

A possibility distribution $\pi_{\mathcal{A}} : 2^X \to [0,1]$ can be seen as a membership function of a fuzzy set $\mathcal{A}$ in $2^X$. Thus, we may identify $\boldsymbol{A}$ with $\mathcal{A}$. The ill-known set having such a possibility distribution is called "a graded ill-known set".

In this case, the lower approximation $A^-$ and the upper approximation $A^+$ are defined as fuzzy sets with the following membership functions:

$$\mu_{A^-}(x) = \inf_{\substack{A \subseteq X \\ x \notin A}} n(\pi_{\mathcal{A}}(A)), \quad \mu_{A^+}(x) = \sup_{\substack{A \subseteq X \\ x \in A}} \varphi(\pi_{\mathcal{A}}(A)), \tag{5}$$

where $n : [0,1] \to [0,1]$ and $\varphi : [0,1] \to [0,1]$ are non-increasing and non-decreasing functions such that $n(0) = \varphi(1) = 1$ and $n(1) = \varphi(0) = 0$. Dubois and Prade [1] defined $\mu_{A^-}$ and $\mu_{A^+}$ with $n(s) = 1 - s$ and $\varphi(s) = s$, $\forall s \in [0, 1]$. Inuiguchi [2] defined $\mu_{A^-}$ and $\mu_{A^+}$ with $n(s) = I(s, 0)$ and $\varphi(s) = T(s, 1)$, where $I$ and $T$ are implication and conjunction functions [2], respectively.

We have the following property:

$$\forall x \in X,\ \mu_{A^-}(x) > 0 \text{ implies } \mu_{A^+}(x) = 1. \tag{6}$$

Because the specification of possibility distribution $\pi_{\mathcal{A}}$ may need a lot of information, as is in the usual ill-known sets, we may know only the lower approximation $A^-$ and the upper approximation $A^+$ as fuzzy sets satisfying (6). To have a consistent possibility distribution $\pi_{\mathcal{A}}$ for any $A^-$ and $A^+$, we need to assume that $n$ and $\varphi$ are surjective, i.e.,

$$\{n(s) \mid s \in [0, 1]\} = [0, 1], \quad \{\varphi(s) \mid s \in [0, 1]\} = [0, 1]. \tag{7}$$

Although (7) is assumed, there are many possibility distributions $\pi_{\mathcal{A}}$ having given $A^-$ and $A^+$ as their lower and upper approximations. However, there is the following maximal possibility distribution $\pi^*_{\mathcal{A}}(A)$:

$$\pi^*_{\mathcal{A}}(A) = \min\left(\inf_{x \notin A} n^*(\mu_{A^-}(x)), \inf_{x \in A} \varphi^*(\mu_{A^+}(x))\right), \tag{8}$$

where we define $\inf \emptyset = 1$ and

$$n^*(a) = \sup\{s \in [0, 1] \mid n(s) \geq a\}, \quad \varphi^*(a) = \sup\{s \in [0, 1] \mid \varphi(s) \leq a\}. \tag{9}$$

Then we identify the maximal possibility distribution $\pi^*_{\mathcal{A}}(A)$ with the given fuzzy sets $A^-$ and $A^+$ unless the other information is available. The graded ill-known set corresponding to $\pi^*_{\mathcal{A}}(A)$ is denoted by $\langle A^-, A^+ \rangle$ and $\pi^*_{\mathcal{A}}(A)$ is written also as $\pi_{\langle A^-, A^+ \rangle}(A)$.

In what follows, we assume that $n$ and $\varphi$ are bijective so that we have $n^* = n^{-1}$ and $\varphi^* = \varphi^{-1}$, where $n^{-1}$ and $\varphi^{-1}$ are inverse functions of $n$ and $\varphi$.

# 3 Extension Principle for Graded Ill-known Sets

In this paper, we consider graded ill-known sets in real line $\mathbf{R}$ and investigate the calculations of graded ill-known sets in $\mathbf{R}$. Graded ill-known sets in real line $\mathbf{R}$ are called "graded ill-known sets of quantities". The set of graded ill-known sets of quantities is denoted by $\mathcal{IQ}$.

Because graded ill-known sets are characterized by possibility distributions on the power set which can be seen as a membership function of a fuzzy set in the power set, the function values of ill-known sets of quantities can be defined by the extension principle [5] in fuzzy set theory.

When a function $\psi : (2^{\mathbf{R}})^m \to 2^{\mathbf{R}}$ is given, we extend this function to a function from $\mathcal{IQ}^m$ to $\mathcal{IQ}$ in the following definition.

**Definition 1.** Let $\mathcal{A}_i$, $i = 1, 2, \ldots, m$ be graded ill-known sets of quantities. Given a function $\psi : (2^{\mathbf{R}})^m \to 2^{\mathbf{R}}$, the image $\psi(\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_m)$ is defined by a graded ill-known set of quantities associated with the following possibility distribution:

$$\pi_{\psi(\mathcal{A}_1,\mathcal{A}_2,\ldots,\mathcal{A}_m)}(Y) = \begin{cases} \displaystyle\sup_{\substack{Q_1,Q_2,\ldots,Q_m\subseteq\mathbf{R}\\ Y=\psi(Q_1,\ldots,Q_m)}} \min\left(\pi_{\mathcal{A}_1}(Q_1),\pi_{\mathcal{A}_2}(Q_2),\cdots,\pi_{\mathcal{A}_m}(Q_m)\right), & \text{if } \psi^{-1}(Y)\neq\emptyset,\\ 0, & \text{if } \psi^{-1}(Y)=\emptyset, \end{cases} \tag{10}$$

where $\pi_{\mathcal{A}_i}$ is a possibility distribution associated with graded ill-known set of quantities $\mathcal{A}_i$ and $\psi^{-1}$ is the inverse image of $\psi$.

Note that, function $f:\mathbf{R}^m\to\mathbf{R}$ can be extended to a function $f:(2^{\mathbf{R}})^m\to 2^{\mathbf{R}}$ by $f(A_1,A_2,\ldots,A_m)=\{f(x_1,x_2,\ldots,x_m)\mid x_i\in A_i,\ i=1,2,\ldots,m\}$. The extended function $f:(2^{\mathbf{R}})^m\to 2^{\mathbf{R}}$ can be further extended to a function $f:\mathcal{IQ}^m\to\mathcal{IQ}$ by Definition 1.

The calculation of $\psi(\mathcal{A}_1,\mathcal{A}_2,\ldots,\mathcal{A}_m)$ is very complex because we should consider all elementary sets of power set $2^{\mathbf{R}}$. This implies that at least an exponential order of calculations are requested. In this paper, we investigate the necessary and sufficient condition for the lower and upper approximations of $\psi(\mathcal{A}_1,\mathcal{A}_2,\ldots,\mathcal{A}_m)$ to be calculated in smaller order of complexity when $\psi$ is the extension of $f:\mathbf{R}^m\to\mathbf{R}$.

The following theorem about the upper approximation is given by Inuiguchi [6].

**Theorem 1.** The upper approximation $f^+(\mathcal{A}_1,\mathcal{A}_2,\ldots,\mathcal{A}_m)$ of $f(\mathcal{A}_1,\mathcal{A}_2,\ldots,\ \mathcal{A}_m)$ can be calculated by upper approximations of $\mathcal{A}_i$, $i=1,2,\ldots,m$. More concretely, we obtain

$$\begin{aligned}\mu_{f^+(\mathcal{A}_1,\mathcal{A}_2,\ldots,\mathcal{A}_m)}(y) &= \sup_{y\in Y}\varphi(\pi_{f(\mathcal{A}_1,\mathcal{A}_2,\ldots,\mathcal{A}_m)}(Y))\\ &= \sup_{\substack{x_1,x_2,\ldots,x_m\in\mathbf{R}\\ y=f(x_1,x_2,\ldots,x_m)}} \min(\mu_{A_1^+}(x_1),\mu_{A_2^+}(x_2),\ldots,\mu_{A_m^+}(x_m))) = \mu_{f(A_1^+,A_2^+,\ldots,A_m^+)}(y),\end{aligned} \tag{11}$$

where $\mu_{f^+(\mathcal{A}_1,\mathcal{A}_2,\ldots,\mathcal{A}_m)}$ is the membership function of $f^+(\mathcal{A}_1,\mathcal{A}_2,\ldots,\mathcal{A}_m)$ and $\mu_{A_i^+}$ is the membership function of the upper approximation $A_i^+$ of $\mathcal{A}_i$. Similarly, $\mu_{f(A_1^+,A_2^+,\ldots,A_m^+)}$ is the membership function of the image $f(A_1^+,A_2^+,\ldots,A_m^+)$.

For the lower approximation, we only have an inequality as shown in the following theorem (see Inuiguchi [6]).

**Theorem 2.** The membership function of lower approximation $f^-(\mathcal{A}_1,\mathcal{A}_2,\ldots,\mathcal{A}_m)$ of $f(\mathcal{A}_1,\mathcal{A}_2,\ldots,\mathcal{A}_m)$ is not smaller than that of $f(A_1^-,A_2^-,\ldots,A_m^-)$, i.e.,

$$\begin{aligned}\mu_{f^-(\mathcal{A}_1,\mathcal{A}_2,\ldots,\mathcal{A}_m)}(y) &= \inf_{y\notin Y} n(\pi_{f(\mathcal{A}_1,\mathcal{A}_2,\ldots,\mathcal{A}_m)}(Y))\\ &\geq \sup_{\substack{x_1,x_2,\ldots,x_m\in\mathbf{R}\\ y=f(x_1,x_2,\ldots,x_m)}} \min(\mu_{A_1^-}(x_1),\mu_{A_2^-}(x_2),\ldots,\mu_{A_m^-}(x_m)) = \mu_{f(A_1^-,A_2^-,\ldots,A_m^-)}(y),\end{aligned} \tag{12}$$

where $\mu_{A_i^-}$ is the membership function of lower approximation $A_i^-$ of $\mathcal{A}_i$. $\mu_{f(A_1^-,A_2^-,\ldots,A_m^-)}$ is the membership function of the image $f(A_1^-, A_2^-, \ldots, A_m^-)$ of fuzzy sets $A_1^-, A_2^-, \ldots, A_m^-$.

To have the equality of (12), Inuiguchi [6] considered a special class $\mathcal{IQ}_{\rm ci} \subseteq \mathcal{IQ}$ of graded ill-known sets of quantities $\mathcal{A}$ satisfying the following properties:

$$\begin{aligned}\forall \alpha \in (0,1],\ \hat{A}(\alpha) = \bigcap \{Q \subseteq \mathbf{R} \mid \pi_{\mathcal{A}}(Q) \geq \alpha\} \text{ is nonempty, closed and convex,}\\ \text{and there exist convex sets } Q_j,\ j = 1,2,\ldots,k \text{ such that}\\ \pi_{\mathcal{A}}(Q_j) \geq \alpha, j = 1,2,\ldots,k \text{ and } \hat{A}(\alpha) = \bigcap_{j=1,2,\ldots,k} Q_j. \end{aligned} \tag{13}$$

A graded ill-known set of quantities $\mathcal{A}$ satisfying (13) can be seen as an extension of a closed interval in $\mathbf{R}$. Then $\mathcal{IQ}_{\rm ci}$ is considered the set of ill-known closed intervals.

Then Inuiguchi [6] has proved the following theorem.

**Theorem 3.** Let $f : \mathbf{R}^m \to \mathbf{R}$ be continuous and monotone (monotonically increasing or monotonically decreasing with respect to each argument). Let $\mathcal{A}_i \in \mathcal{IQ}_{\rm ci}$, $i = 1,2,\ldots,m$. If $\{\pi_{\mathcal{A}_i}(A) \mid A \subseteq \mathbf{R}\}$ is finite for $i = 1,2,\ldots,m$ then we have

$$\begin{aligned}&\mu_{f^-(\mathcal{A}_1,\mathcal{A}_2,\ldots,\mathcal{A}_m)}(y) = \inf_{y \notin Y} n(\pi_{f(\mathcal{A}_1,\mathcal{A}_2,\ldots,\mathcal{A}_m)}(Y))\\ &= \sup_{\substack{x_1,x_2,\ldots,x_m \in \mathbf{R}\\ y=f(x_1,x_2,\ldots,x_m)}} \min(\mu_{A_1^-}(x_1), \mu_{A_2^-}(x_2), \ldots, \mu_{A_m^-}(x_m)) = \mu_{f(A_1^-,A_2^-,\ldots,A_m^-)}(y).\end{aligned} \tag{14}$$

We note that $f(\mathcal{A}_1, \ldots, \mathcal{A}_m)$ requires the calculations on the power set while $f(A_1^-, \ldots, A_m^-)$ and $f(A_1^+, \ldots, A_m^+)$ require the calculations on the universe. Therefore, the most right-hand sides values of (11) and (14) are calculated much more efficiently than the most left-hand sides values.

In this paper, to generalize Theorem 3 as well as to capture the essence, we give the necessary and sufficient condition of (14).

# 4 The Main Result and Its Applications

We have the following theorem.

**Theorem 4.** We have $f^-(\mathcal{A}_1, \ldots, \mathcal{A}_m) = f(A_1^-, \ldots, A_m^-)$ if and only if

$$\begin{aligned}\forall \alpha \in [0,1),\ &\bigcap \{f(Q_1, \ldots, Q_m) \mid Q_1 \in (\mathcal{A}_1)_\alpha, \ldots, Q_m \in (\mathcal{A}_m)_\alpha\}\\ &= f\left(\bigcap (\mathcal{A}_1)_\alpha, \ldots, \bigcap (\mathcal{A}_m)_\alpha\right),\end{aligned} \tag{15}$$

where $(\mathcal{A}_i)_\alpha = \{Q \mid \pi_{\mathcal{A}_i}(Q) > \alpha\}$.

(Proof) For the sake of simplicity, we prove when $m = 2$. In cases where $m \neq 2$, it can be proved in the same way. From Theorem 2, we consider the necessary and sufficient condition of

$$\mu_{f^-(\mathcal{A}_1,\mathcal{A}_2)}(y) \le \mu_{f(A_1^-,A_2^-)}(y).$$

This is equivalent to

$$\forall\alpha \in (0,1],\ \mu_{f^-(\mathcal{A}_1,\mathcal{A}_2)}(y) \ge \alpha \text{ implies } \mu_{f(A_1^-,A_2^-)}(y) \ge \alpha. \qquad (*)$$

Then we consider the equivalent condition of (a) $\mu_{f^-(\mathcal{A}_1,\mathcal{A}_2)}(y) \ge \alpha$ and that of (b) $\mu_{f(A_1^-,A_2^-)}(y) \ge \alpha$.

First let us investigate the equivalent condition of (a). By definition, we have

$$\begin{aligned}
&\mu_{f^-(\mathcal{A}_1,\mathcal{A}_2)}(y) \ge \alpha \Leftrightarrow \inf_{Y \not\ni y} n(\pi_{f(\mathcal{A}_1,\mathcal{A}_2)}(Y)) \ge \alpha\\
&\quad\Leftrightarrow y \notin Y \text{ implies } \pi_{f(\mathcal{A}_1,\mathcal{A}_2)}(Y) \ge n^{-1}(\alpha)\\
&\quad\Leftrightarrow \pi_{f(\mathcal{A}_1,\mathcal{A}_2)}(Y) > n^{-1}(\alpha) \text{ implies } y \in Y\\
&\quad\Leftrightarrow \sup_{Q_1,Q_2:Y=f(Q_1,Q_2)} \min(\pi_{\mathcal{A}_1}(Q_1), \pi_{\mathcal{A}_2}(Q_2)) > n^{-1}(\alpha) \text{ implies } y \in Y\\
&\quad\Leftrightarrow y \in \bigcap \left\{ f(Q_1,Q_2) \mid Q_1 \in (\mathcal{A}_1)_{n^{-1}(\alpha)},\ Q_2 \in (\mathcal{A}_2)_{n^{-1}(\alpha)} \right\}.
\end{aligned}$$

Now let us investigate the equivalent condition of (b). By definition and the continuity of $n^{-1}$, we obtain

$$\begin{aligned}
&\mu_{f(A_1^-,A_2^-)}(y) \ge \alpha \Leftrightarrow \sup_{x_1,x_2:y=f(x_1,x_2)} \min\left(\mu_{A_1^-}(x_1), \mu_{A_2^-}(x_2)\right) \ge \alpha\\
&\quad\Leftrightarrow \forall\varepsilon > 0,\ \exists x_1,\ x_2,\ y = f(x_1,x_2),\ \mu_{A_1^-}(x_1) > \alpha - \varepsilon,\ \mu_{A_2^-}(x_2) > \alpha - \varepsilon\\
&\quad\Leftrightarrow \forall\varepsilon > 0,\ \exists x_1,\ x_2,\ y = f(x_1,x_2),\ \inf_{Q_i \not\ni x_i} n(\pi_{\mathcal{A}_i}(Q_i)) > \alpha - \varepsilon,\ i = 1,2\\
&\quad\Leftrightarrow \forall\varepsilon > 0,\ \exists x_1,\ x_2,\ y = f(x_1,x_2),\\
&\qquad\qquad\qquad x_i \in \bigcap \left\{ Q_i \mid \pi_{\mathcal{A}_i}(Q_i) \ge n^{-1}(\alpha - \varepsilon) \right\},\ i = 1,2\\
&\quad\Leftrightarrow \exists x_1,\ x_2,\ y = f(x_1,x_2),\ x_i \in \bigcap \left\{ Q_i \mid \pi_{\mathcal{A}_i}(Q_i) > n^{-1}(\alpha) \right\}, \imath = 1,2\\
&\quad\Leftrightarrow y \in f\left( \bigcap (\mathcal{A}_1)_{n^{-1}(\alpha)}, \bigcap (\mathcal{A}_2)_{n^{-1}(\alpha)} \right).
\end{aligned}$$

From those equivalent conditions of (a) and (b) and the fact that $\{n^{-1}(\alpha) \mid \alpha \in (0,1]\} = [0,1)$, the necessary and sufficient condition of $(*)$ is obtained as

$$\begin{aligned}
\forall\alpha \in [0,1),\ &\bigcap \{ f(Q_1,Q_2) \mid Q_1 \in (\mathcal{A}_1)_\alpha,\ Q_2 \in (\mathcal{A}_2)_\alpha \}\\
&= f\left( \bigcap (\mathcal{A}_1)_\alpha, \bigcap (\mathcal{A}_2)_\alpha \right).
\end{aligned}$$

(Q.E.D.)

From Theorem 4, we prove a more general sufficient condition for $f^-(\mathcal{A}_1, \ldots, \mathcal{A}_m) = f(A_1^-, \ldots, A_m^-)$ than Theorem 3. Before describing it, we define a class $\mathcal{IQ}_{\text{int}} \subseteq$

$\mathcal{IQ}$ of graded ill-known sets of quantities $\mathcal{A}$ satisfying the following properties:

$$\begin{aligned}\forall\alpha\in[0,1),\ A(\alpha)=\bigcap(\mathcal{A})_\alpha \text{ is nonempty and convex, and}\\ \text{there exists a family of convex sets } \{Q_j\}_{j\in J}\\ \text{such that } Q_j\in(\mathcal{A})_\alpha, j\in J \text{ and } A(\alpha)=\bigcap_{j\in J}Q_j.\end{aligned}\tag{16}$$

A graded ill-known set of quantities $\mathcal{A}$ satisfying (16) can be seen as an extension of an interval in $\mathbf{R}$. Then $\mathcal{IQ}_{\text{int}}$ is considered the set of ill-known intervals.

Then we obtain the following theorem.

**Theorem 5.** Let $f:\mathbf{R}^m\to\mathbf{R}$ be continuous and monotone (monotonically increasing or monotonically decreasing with respect to each argument). Let $\mathcal{A}_i\in\mathcal{IQ}_{\text{int}}$, $i=1,2,\ldots,m$. Then we have (14).

(Proof) By the same reason as Theorem 4, we prove when $m=2$. Without loss of generality, we assume $f$ is monotonically increasing with respect to all arguments.

From $A_i(\alpha)=\bigcap(\mathcal{A}_i)_\alpha\subseteq Q_i$ for $Q_i\in(\mathcal{A}_i)_\alpha$, $f(A_1(\alpha),A_2(\alpha))\subseteq\bigcap\{f(Q_1,Q_2)\mid Q_1\in(\mathcal{A}_1)_\alpha,\ Q_2\in(\mathcal{A}_2)_\alpha\}$. Then we prove

$$y\notin f(A_1(\alpha),A_2(\alpha)) \text{ implies } y\notin\bigcap\{f(Q_1,Q_2)\mid Q_1\in(\mathcal{A}_1)_\alpha,\ Q_2\in(\mathcal{A}_2)_\alpha\}.\tag{$*$}$$

Because $f$ is continuous and $A_i(\alpha)$, $i=1,2$ are nonempty and convex, $f(A_1(\alpha), A_2(\alpha))$ becomes an interval (a convex set in the real line). Then we prove $(*)$ dividing into two cases: (a) $y\le\inf f(A_1(\alpha),A_2(\alpha))$ and $y\notin f(A_1(\alpha),A_2(\alpha))$ and (b) $y\ge\sup f(A_1(\alpha),A_2(\alpha))$ and $y\notin f(A_1(\alpha),A_2(\alpha))$.

Because $\mathcal{A}_i\in\mathcal{IQ}_{\text{int}}$, there exists a family $\mathcal{Q}_i$ of convex sets $\{Q_{ij}\}_{j\in J_i}$ such that $Q_{ij}\in(\mathcal{A}_i)_\alpha$ and $A_i(\alpha)=\bigcap_{j\in J_i}Q_{ij}$ for $i=1,2$. From the convexity of $Q_{ij}$, $j\in J_i$, $i=1,2$, there exist subfamilies $\underline{\mathcal{Q}}_i=\{\underline{Q}_{ij}\}_{j\in\underline{J}_i}\subseteq\mathcal{Q}_i$ and $\overline{\mathcal{Q}}_i=\{\overline{Q}_{ij}\}_{j\in\overline{J}_i}\subseteq\mathcal{Q}_i$ such that $\sup_{j\in\underline{J}_i}\inf\underline{Q}_{ij}=\inf A_i(\alpha)$ and $\inf_{j\in\overline{J}_i}\sup\overline{Q}_{ij}=\sup A_i(\alpha)$.

From the monotonicity, we obtain

$$\begin{aligned}&\forall r_1\in A_1(\alpha),\ \forall r_2\in A_2(\alpha),\ y<f(r_1,r_2) \text{ implies}\\ &\qquad\exists k_1\in\underline{J}_1,\ \exists k_2\in\underline{J}_2,\ \forall q_1\in\underline{Q}_{1k_1},\ \forall q_2\in\underline{Q}_{2k_2},\ y<f(q_1,q_2),\\ &\forall r_1\in A_1(\alpha),\ \forall r_2\in A_2(\alpha),\ y>f(r_1,r_2) \text{ implies}\\ &\qquad\exists l_1\in\overline{J}_1,\ \exists l_2\in\overline{J}_2,\ \forall q_1\in\overline{Q}_{1l_1},\ \forall q_2\in\overline{Q}_{2l_2},\ y>f(q_1,q_2).\end{aligned}$$

Therefore, in case (a) $y\le\inf f(A_1(\alpha),A_2(\alpha))$ and $y\notin f(A_1(\alpha),A_2(\alpha))$, we have $y\notin f(\underline{Q}_1,\underline{Q}_2)$. This implies that $y\notin\bigcap\{f(Q_1,Q_2)\mid Q_1\in(\mathcal{A}_1)_\alpha,\ Q_2\in(\mathcal{A}_2)_\alpha\}$. Similarly, in case (b) $y\ge\sup f(A_1(\alpha),A_2(\alpha))$ and $y\notin f(A_1(\alpha),A_2(\alpha))$, we have $y\notin f(\overline{Q}_1,\overline{Q}_2)$. This implies that $y\notin\bigcap\{f(Q_1,Q_2)\mid Q_1\in(\mathcal{A}_1)_\alpha,\ Q_2\in(\mathcal{A}_2)_\alpha\}$. Hence, $(*)$ is proved. (Q.E.D.)

Theorem 5 generalizes Theorem 3. In Theorem 5, $\{\pi_{\mathcal{A}_i}\mid A\subseteq\mathbf{R}\}$ can be infinite and the restriction $\mathcal{A}_i\in\mathcal{IQ}_{\text{ci}}$ is relaxed to $\mathcal{A}_i\in\mathcal{IQ}_{\text{int}}$.

If $\mathcal{A}_i(\alpha) \in (\mathcal{A}_i)_\alpha$, $i = 1, 2, \ldots, m$ for any $\alpha \in [0, 1)$, we have (15). From Theorem 4, we have the following corollary.

**Corollary 1.** If $\mathcal{A}_i(\alpha) \in (\mathcal{A}_i)_\alpha$, $i = 1, 2, \ldots, m$ for any $\alpha \in [0, 1)$, then we have $f^-(\mathcal{A}_1, \ldots, \mathcal{A}_m) = f(A_1^-, \ldots, A_m^-)$.

When $\mathcal{A}_i(\alpha) \in (\mathcal{A}_i)_\alpha$, $i = 1, 2, \ldots, m$ for any $\alpha \in [0, 1)$, we have (14) without any condition on $f$. The strong condition $\mathcal{A}_i(\alpha) \in (\mathcal{A}_i)_\alpha$, $i = 1, 2, \ldots, m$ for any $\alpha \in [0, 1)$ is satisfied by a graded ill-known set of quantities defined by lower and upper approximations. This can be understood directly from the following proposition.

**Proposition 1.** Let $\mathcal{A}$ be a graded ill-known set defined by lower and upper approximations $A^-$ and $A^+$. Then we have

$$(\mathcal{A})_\alpha = \left\{ A \;\middle|\; [A^-]_{n(\alpha)} \subseteq A \subseteq (A^+)_{\varphi(\alpha)} \right\}, \tag{17}$$

where $[A^-]_\beta = \{x \mid \mu_{A^-}(x) \geq \beta\}$, $\beta \in (0, 1]$ and $(A^+)_\gamma = \{x \mid \mu_{A^+}(x) > \gamma\}$, $\beta \in [0, 1)$.
(Proof) From (8), we obtain the following equivalences:

$$\begin{aligned}
& A \in (\mathcal{A})_\alpha \\
& \Leftrightarrow \inf_{x \notin A} n^{-1}(\mu_{A^-}(x)) > \alpha \text{ and } \inf_{x \in A} \varphi^{-1}(\mu_{A^+}(x)) > \alpha \\
& \Leftrightarrow (x \in A \text{ implies } \mu_{A^+}(x) > \varphi(\alpha)) \text{ and } (\mu_{A^-}(x) \geq n(\alpha) \text{ implies } x \in A) \\
& \Leftrightarrow [A^-]_{n(\alpha)} \subseteq A \subseteq (A^+)_{\varphi(\alpha)}.
\end{aligned}$$

(Q.E.D.)

From Proposition 1, we know that $\mathcal{A}(\alpha) = [A^-]_{n(\alpha)}$ if $\mathcal{A}$ is defined by lower and upper approximations $A^-$ and $A^+$. Because $A^+$ is not related $\mathcal{A}(\alpha)$, we may have a weaker sufficient condition for $\mathcal{A}_i(\alpha) \in (\mathcal{A}_i)_\alpha$. Namely, we know that $\mathcal{A}_i(\alpha) \in (\mathcal{A}_i)_\alpha$ is satisfied if the possibility distribution $\pi_{\mathcal{A}_i}$ of a graded ill-known set of quantities $\mathcal{A}_i$ satisfies

$$\begin{aligned}
& \pi_{\mathcal{A}_i}(A) = \inf_{x \notin A} n^{-1}(\mu_{A_i^-}(x)), \\
& \qquad\qquad \forall A \text{ such that } \inf_{x \notin A} n^{-1}(\mu_{A_i^-}(x)) \leq \inf_{x \in A} \varphi^{-1}(\mu_{A_i^+}(x)),
\end{aligned} \tag{18}$$

where $A_i^-$ and $A_i^+$ are lower and upper approximations of $\mathcal{A}_i$, respectively, and $\mu_{A_i^-}$ and $\mu_{A_i^+}$ are their membership functions.

Finally, we investigate whether $f(\mathcal{A}_1, \ldots, \mathcal{A}_m)$ is obtained from $f(A_-, \ldots, A_m^-)$ and $f(A_1^+, \ldots, A_m^+)$ through (8) when $\mathcal{A}_i$, $i = 1, 2, \ldots .m$ are graded ill-known sets of quantities defined by the lower and upper approximations $A_i^-$ and $A_i^+$. Namely we consider whether $f(\langle A_1^-, A_1^+ \rangle, \ldots, \langle A_m^-, A_m^+ \rangle)$ defined by Definition 1 equals to $\langle f(A_1^-, \ldots, A_m^-), f(A_1^+, \ldots, A_m^+) \rangle$.

Contrary to our expectation, the answer is negative. A counter example is given as follows. Let $f_1 : \mathbf{R}^2 \to \mathbf{R}$ be a function defined by

$$f_1(x_1, x_2) = \begin{cases} x_1 + x_2, & \text{if } x_1 + x_2 \le 6, \\ 0, & \text{if } x_1 + x_2 \in (6, 10], \\ x_1 + x_2 - 4, & \text{if } x_1 + x_2 > 10. \end{cases}$$

Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be ill-known sets defined by lower approximations $A_1^- = [2, 3]$ and $A_2^- = [2, 3]$ and upper approximations $A_1^+ = [1, 7]$ and $A_2^+ = [1, 8]$, respectively. Then we have $[4, 8] \notin f_1(\mathcal{A}_1, \mathcal{A}_2) = f_1(\langle A_1^-, A_1^+\rangle, \langle A_2^-, A_2^+\rangle)$ but $\{0\} \cup [4, 8] \in f_1(\mathcal{A}_1, \mathcal{A}_2) = f_1(\langle A_1^-, A_1^+\rangle, \langle A_2^-, A_2^+\rangle)$. On the other hand, we obtain $f_1(A_1^-, A_2^-) = [4, 6]$ and $f_1(A_1^+, A_2^+) = \{0\} \cup [2, 11]$. Then we have $[4, 8] \in \langle f_1(A_1^-, A_2^-), f_1(A_1^+, A_2^+)\rangle$. Therefore, $f_1(\langle A_1^-, A_1^+\rangle, \langle A_2^-, A_2^+\rangle) = \langle f_1(A_1^-, A_2^-),\ f_1(A_1^+, A_2^+)\rangle$ does not always hold.

Even when function is continuous and monotone, we have a similar result. Namely, let $f_2 : \mathbf{R}^2 \to \mathbf{R}$ be a function defined by $f_2(x_1, x_2) = x_1 + x_2$. Let $A_i^-$ and $A_i^+$ $(i = 1, 2)$ be the same as above, i.e., $A_1^- = [2, 3]$, $A_2^- = [2, 3]$, $A_1^+ = [1, 7]$ and $A_2^+ = [1, 8]$. We have $f_2(A_1^-, A_2^-) = [4, 6]$ and $f_2(A_1^+, A_2^+) = [2, 15]$. Then $[4, 6] \cup [11, 12] \in \langle f_2(A_1^-, A_2^-), f_2(A_1^+, A_2^+)\rangle$. On the contrary, $[4, 6] \cup [11, 12] \notin f_2(\langle A_1^-, A_1^+\rangle, \langle A_2^-, A_2^+\rangle)$. This is because there is no $Q_1 \subseteq \mathbf{R}$ and $Q_2 \subseteq \mathbf{R}$ such that $f_2(Q_1, Q_2) = [4, 6] \cup [11, 12]$.

From the examples above, we know that we may have

$$\pi_{f(\langle A_1^-, A_1^+\rangle, \ldots, \langle A_m^-, A_m^+\rangle)}(Y) = \pi_{\langle f(A_1^-, \ldots, A_m^-), f(A_1^+, \ldots, A_m^+)\rangle}(Y), \tag{19}$$

only for $Y \in f(2^{\mathbf{R}}, \ldots, 2^{\mathbf{R}})$.

The following theorem shows that (19) holds for a convex set $Y \subseteq \mathbf{R}$ and a monotone continuous function.

**Theorem 6.** Let $f : \mathbf{R}^m \to \mathbf{R}$ be continuous and monotone. Let $A_i^-$ and $A_i^+$ be fuzzy sets showing lower and upper approximations of a graded ill-known set $\mathcal{A}_i$, $i = 1, 2, \ldots, m$. Then (19) holds for a convex set $Y \subseteq \mathbf{R}$.

(Proof) We prove (19) when $m = 2$. (19) can be proved in the same way even when $m > 2$. Let $f^{-1}(Y) = \{(Q_1, Q_2) \mid f(Q_1, Q_2) = Y\}$ for $Y \subseteq \mathbf{R}$ and $f^{-1}(y) = \{(x_1, x_2) \mid f(x_1, x_2) = y\}$ for $y \in \mathbf{R}$. For the sake of simplicity, we define graded ill-known sets $\mathcal{A}_i = \langle A_i^-, A_i^+\rangle$, $i = 1, 2$ and $\mathcal{F} = \langle f(A_1^-, A_2^-), f(A_1^+, A_2^+)\rangle$.

When $f^{-1}(Y) = \emptyset$, we may have two cases: (a) $\exists y \in Y$, $f^{-1}(y) = \emptyset$, and (b) $\forall y \in Y$, $f^{-1}(y) \neq \emptyset$ and $\forall \hat{Q}_1 \times \hat{Q}_2 \subseteq \mathbf{R}^2$ such that $\forall y \in Y$, $(\hat{Q}_1 \times \hat{Q}_2) \cap f^{-1}(y) \neq \emptyset$ and $f(\hat{Q}_1 \times \hat{Q}_2) \supset Y$.

Because $f$ is continuous and monotone, if $\forall y \in Y$, $f^{-1}(y) \neq \emptyset$, there exists $\hat{Q}_1 \times \hat{Q}_2 \subseteq \mathbf{R}^2$ such that $\forall y \in Y$, $(\hat{Q}_1 \times \hat{Q}_2) \cap f^{-1}(y) \neq \emptyset$ and $f(\hat{Q}_1 \times \hat{Q}_2) \supset Y$. Then (b) is never satisfied if $f^{-1}(Y) = \emptyset$. Then $f^{-1}(Y) = \emptyset$ if and only if $\exists y \in Y$, $f^{-1}(y) = \emptyset$.

When $f^{-1}(Y) = \emptyset$, we have $\pi_{f(\mathcal{A}_1, \mathcal{A}_2)}(Y) = 0$ by Definition 1 and $\inf_{y \in Y} \varphi^{-1}\left(\mu_{f(A_1^+, A_2^+)}(y)\right) = 0$ because $\mu_{f(A_1^+, A_2^+)}(y) = 0$ for $f^{-1}(y) = \emptyset$ from the extension principle in fuzzy sets. The latter implies $\pi_{\mathcal{F}}(Y) = 0$. Hence, we have (19) when $f^{-1}(Y) = \emptyset$.

Then we consider a case where $f^{-1}(Y) \neq \emptyset$. Because $\pi_{\mathcal{F}}$ is the maximal possibility distribution of graded ill-known sets having lower and upper approximations

$f(A_1^-, A_2^-)$ and $f(A_1^+, A_2^+)$. On the other hand, we have $f^-(\mathcal{A}_1, \mathcal{A}_2) = f(A_1^-, A_2^-)$ and $f^+(\mathcal{A}_1, \mathcal{A}_2) = f(A_1^+, A_2^+)$. Then we have $\pi_{f(\mathcal{A}_1,\mathcal{A}_2)}(Y) \le \pi_{\mathcal{F}}(Y)$. Therefore, we prove

$$\pi_{f(\mathcal{A}_1,\mathcal{A}_2)}(Y) \ge \pi_{\mathcal{F}}(Y). \tag{$*$}$$

We have

$$\begin{aligned}
\pi_{f(\mathcal{A}_1,\mathcal{A}_2)}(Y) &= \sup_{\substack{Q_1,Q_2\subseteq\mathbf{R}\\ Y=f(Q_1,Q_2)}} \min\left(\pi_{\mathcal{A}_1}(Q_1), \pi_{\mathcal{A}_2}(Q_2)\right)\\
&= \sup_{\substack{Q_1,Q_2\subseteq\mathbf{R}\\ Y=f(Q_1,Q_2)}} \min\left(\min\left(\inf_{x\notin Q_1} n^{-1}(\mu_{A_1^-}(x)), \inf_{x\in Q_1} \varphi^{-1}(\mu_{A_1^+}(x))\right),\right.\\
&\qquad\qquad \left.\min\left(\inf_{x\notin Q_2} n^{-1}(\mu_{A_2^-}(x)), \inf_{x\in Q_2} \varphi^{-1}(\mu_{A_2^+}(x))\right)\right).
\end{aligned}$$

Applying Proposition 1, we obtain

$$\begin{aligned}
&\pi_{f(\mathcal{A}_1,\mathcal{A}_2)}(Y) > \alpha\\
&\Leftrightarrow\ \exists(Q_1, Q_2) \text{ such that } Y = f(Q_1, Q_2),\\
&\qquad [A_1^-]_{n(\alpha)} \subseteq Q_1 \subseteq (A_1^+)_{\varphi(\alpha)} \text{ and } [A_2^-]_{n(\alpha)} \subseteq Q_2 \subseteq (A_2^+)_{\varphi(\alpha)}.
\end{aligned} \tag{#}$$

On the other hand, from the extension principle in fuzzy sets, we obtain

$$\begin{aligned}
\pi_{\mathcal{F}}(Y) = \min&\left(\inf_{y\notin Y} n^{-1}\left(\sup_{\substack{x_1,x_2\in\mathbf{R}\\ y=f(x_1,x_2)}} \min\left(\mu_{A_1^-}(x_1), \mu_{A_2^-}(x_1)\right)\right),\right.\\
&\left.\inf_{y\in Y} \varphi^{-1}\left(\sup_{\substack{x_1,x_2\in\mathbf{R}\\ y=f(x_1,x_2)}} \min\left(\mu_{A_1^+}(x_1), \mu_{A_2^+}(x_1)\right)\right)\right).
\end{aligned}$$

Then we obtain

$$\pi_{\mathcal{F}}(Y) > \alpha \Leftrightarrow \bigcap_{\varepsilon>0} f([A_1^-]_{n(\alpha)-\varepsilon}, [A_1^-]_{n(\alpha)-\varepsilon}) \subseteq Y \subseteq f((A_1^+)_{\varphi(\alpha)}, (A_2^+)_{\varphi(\alpha)}). \tag{$**$}$$

Now we prove $(*)$ by showing

$$\pi_{\mathcal{F}}(Y) > \alpha \text{ implies } \pi_{f(\mathcal{A}_1,\mathcal{A}_2)}(Y) > \alpha.$$

For lower and upper approximations $A_i^-$ and $A_i^+$, we have $\mu_{A_i^-}(x) > 0$ implies $\mu_{A_i^+}(x) = 1$. Then we obtain $[A_i^-]_{n(\alpha)} \subseteq [A_i^-]_{n(\alpha)-\varepsilon} \subseteq (A_i^+)_{\varphi(\alpha)}$ for any $\varepsilon > 0$ and for $i = 1, 2$.

Assume $\pi_{\mathcal{F}}(Y) > \alpha$, from $(**)$, we obtain $f([A_1^-]_{n(\alpha)}, [A_1^-]_{n(\alpha)}) \subseteq Y$. From the continuity and monotonicity of $f$, the convexity of $Y$ and $(**)$, $Q_i$, $i = 1, 2$ can be enlarged continuously from $Q_i = [A_i^-]_{n(\alpha)}$ to $\bar{Q}_i$, $i = 1, 2$ such that $Y = f(\bar{Q}_1, \bar{Q}_2)$ in $(A_i^+)_{\varphi(\alpha)}$. Therefore, we have $\bar{Q}_i$ $(i = 1, 2)$ such that $[A_i^-]_{n(\alpha)} \subseteq \bar{Q}_i \subseteq (A_i^+)_{\varphi(\alpha)}$ and $Y = f(\bar{Q}_1, \bar{Q}_2)$ (see Figure 1). Hence, from (#), we obtain $\pi_{f(\mathcal{A}_1,\mathcal{A}_2)}(Y) > \alpha$. (Q.E.D.)

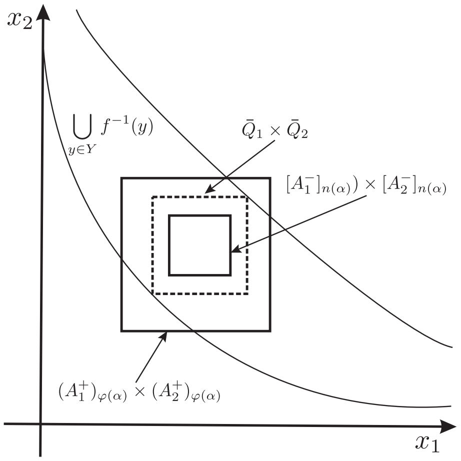

$x_2$
$\bigcup_{y\in Y} f^{-1}(y)$
$\bar{Q}_1 \times \bar{Q}_2$
$[A_1^-]_{n(\alpha)}) \times [A_2^-]_{n(\alpha)}$
$(A_1^+)_{\varphi(\alpha)} \times (A_2^+)_{\varphi(\alpha)}$
$x_1$

Figure 1: $[A_i^-]_{n(\alpha)} \subseteq \bar{Q}_i \subseteq (A_i^+)_{\varphi(\alpha)}$

## 5 Concluding Remarks

We have shown the necessary and sufficient condition that lower and upper approximations of function values of graded ill-known sets are obtained by function values of lower and upper approximations of graded ill-known sets. Using this condition, we have weakened the previously obtained sufficient condition. We have shown also that lower and upper approximations of function values of graded ill-known sets defined by lower and upper approximations are always obtained by function values of the given lower and upper approximations. Moreover, we have demonstrated that function values of graded ill-known sets defined by lower and upper approximations are not always obtained from function values of the given lower and upper approximations while their lower and upper approximations are. Degrees of their possibility distributions can take same values only for function values of sets. We have given a sufficient condition that those possibility degrees are equal. The necessary and sufficient condition and other sufficient conditions would be future topics in the calculations of graded ill-known sets. The results obtained in this paper are valuable for applications of graded ill-known sets to systems optimization, decision making, data analysis and so on.

## Acknowledgement

This study is supported by Grant-in-Aid for Scientific Research (C) No.23510169.

## References

[1] Dubois, D., Prade, H. (1988), Incomplete Conjunctive Information, *Comput. Math. Applic.*, Vol.15, 797–810

[2] Inuiguchi, M. (2012), Rough Representations of Ill-Known Sets and Their Manipulations in Low Dimensional Space, In: Skowron, A., Suraj, Z. (eds.) *Rough Sets and Intelligent Systems: To the Memory of Professor Zdzisław Pawlak*, Vol.1, Springer, Heidelberg, 305–327

[3] Ngyuen, H. T. (2006), *Introduction to Random Set*, Chapman and Hall/CRC, Boca Raton, FL

[4] Zadeh, L. A. (1978), Fuzzy Sets as the Basis for a Theory of Possibility, *Fuzzy Sets and Systems*, Vol.1, 3–28

[5] Zadeh, L. A. (1975), The Concept of Linguistic Variable and Its Application to Approximate Reasoning, *Inform. Sci.*, Vol.8, 199–246

[6] Inuiguchi, M. (2012), Ill-known Set Approach to Disjunctive Variables: Calculations of Graded Ill-known Intervals *Proceedings of IPMU 2012*, (2012)

# Affiliated Ratio-implicational and Equivalency Data-mining Quantifiers and their Truth Configurations

**Jiří Ivánek**
Department of Information and Knowledge Engineering
University of Economics, Prague
ivanek@vse.cz

**Abstract**

Relations between two Boolean attributes derived from data can be quantified by functions defined on four-fold tables corresponding to pairs of the attributes. In the paper, a class of ratio-implicational quantifiers is investigated. The method of construction of the affiliated (logically nearest) double implicational and equivalence quantifiers to a given implicational quantifier is recalled and applied to ratio-implicational quantifiers. Possible truth-configurations of the obtained set of seven formulae (given data and some treshold) are discussed in details.

## 1 Introduction

Assume having a data file and consider two Boolean (binary, dichotomic) attributes $\varphi$ and $\psi$. A four-fold table $< a, b, c, d >$ corresponding to these attributes is composed from numbers of objects in data satisfying four different Boolean combinations of attributes:

| | $\psi$ | $\neg\psi$ |
|---|---|---|
| $\varphi$ | $a$ | $b$ |
| $\neg\varphi$ | $c$ | $d$ |

$a$ - number of objects satisfying both $\varphi$ and $\psi$,
$b$ - number of objects satisfying $\varphi$ and not satisfying $\psi$,
$c$ - number of objects not satisfying $\varphi$ and satisfying $\psi$,
$d$ - number of objects not satisfying $\varphi$ and not satisfying $\psi$.

*Four-fold table quantifier* $\sim$ is a function with values from the interval $[0, 1]$ defined on the set of all four-fold tables $< a, b, c, d >$. Several classes of quantifiers (implicational, equivalency) have been studied in the theory of the GUHA method ([2],[3]).

A quantifier $\sim (a, b)$ is *implicational* if $\sim (a', b') \geq \sim (a, b)$ when $a' \geq a, b' \leq b$. The most common example of implicational quantifier is the quantifier of basic implication (corresponds to the notion of a confidence or accuracy of an association rule used in data mining, see [1],[8]):

$$\Rightarrow_\phi (a, b) = \frac{a}{a + b}.$$

In the paper, a subclass of ratio-implicational quantifiers is investigated. The method of construction of the affiliated (logically nearest) double implicational and equivalence quantifiers to a given implicational quantifier is recalled and applied to the subclass of ratio-implicational quantifiers. Possible truth-configurations of the obtained set of seven formulae (given data and some treshold) are discussed in details.

## 2 Ratio-implicational quantifiers

This is one of the main properties of the basic implicational quantifier: the greater the ratio $a/b$, the greater the value of the quantifier. This property is stronger than that used in the definition of implicational quantifiers. Therefore we introduced a subclass of implicational quantifiers with this property [6]:

A quantifier $\sim (a,b)$ is *ratio-implicational*, if $\sim (a',b') \geq \sim (a,b)$ when $a'b \geq ab'$. For any $\theta > 0$ the following quantifier is ratio-implicational:

$$\Rightarrow_\theta (a,b) = \frac{a}{a+\theta b}.$$

It is clear that each ratio-implicational quantifier is also implicational. There are some other properties of ratio-implicational quantifiers proved in [6]:

(i) if $a'b = ab'$ then $\Rightarrow^* (a',b') = \Rightarrow^* (a,b)$.

(ii) there are numbers $m^*, M^*$ from $[0,1]$ such that

$$\begin{aligned} m^* &= \Rightarrow^* (0,b) \text{ for all } b > 0, \\ M^* &= \Rightarrow^* (a,0) \text{ for all } a > 0, \\ m^* \leq \Rightarrow^* &(a,b) \leq M^* \text{ for all } a,b > 0. \end{aligned}$$

(iii) there is a non-decreasing function $g^*$ defined on non-negative rationals and $\infty$ such that

$$\Rightarrow^* (a,b) = g^* \left(\frac{a}{b}\right).$$

The function g* is defined as follows:

$$g^*(0) = m^*, \qquad g^*(\infty) = M^*, \qquad g^* \left(\frac{i}{j}\right) = \Rightarrow^* (i,j) \qquad \text{for all integers } i,j > 0.$$

Correctness of this definition follows from (i), (ii); monotonicity follows from the definition of ratio-implicational quantifiers.

The class of ratio-implicational quantifiers is a proper subclass of the class of implicational quantifiers. For a counterexample, let us observe that statistically motivated quantifier $\Rightarrow^?_p$ (where $p$ is a parameter, $0 < p < 1$)

$$\Rightarrow^?_p (a,b) = \sum_{i=0}^{a} \frac{(a+b)!}{i!\,(a+b-i)!}\, p^i\,(1-p)^{a+b-i}$$

is implicational (see [3]), but is not ratio-implicational because for instance

$$\Rightarrow_p^? (0,b) = (1-p)^b \neq \Rightarrow_p^? (0,b+1) = (1-p)^{b+1}.$$

Another important property of ratio-implicational quantifiers is their special relation to fuzzy implications. Let us recall that in fuzzy logic (see e.g. [4]), a binary operator $I$ on the unit interval is called *fuzzy implication* if

$$I(0,0) = I(1,1) = 1, \qquad I(1,0) = 0,$$

and for all $x, x', y, y'$ the following property holds:

$$\text{if } x' \leq x \text{ and } y' \geq y \text{ then } I(x',y') \geq I(x,y).$$

The next theorem (proved in [6]) shows the correspondence between ratio-implicational quantifiers and fuzzy implications.

**Theorem**

(i) Let $I$ be a fuzzy implication. Then

$$\Rightarrow_I^0 (a,b) = I\left(\frac{b}{a+b}, \frac{a}{a+b}\right)$$

is ratio-implicational quantifier with $m^* = 0, M^* = 1$.

(ii) Let $\Rightarrow^*$ be a ratio-implicational quantifier with $m^* = 0, M^* = 1$. Extend the associated function $g^*\left(\frac{a}{b}\right) = \Rightarrow^* (a,b)$ to reals by

$$g^*(r) = \sup\left\{g^*\left(\frac{a}{b}\right) : \frac{a}{b} \leq r\right\},$$

and define

$$I^*(x,y) = g^*\left(\sqrt{\tfrac{(1-x)y}{x(1-y)}}\right) \text{ for } x,y \in [0,1], x \neq 0, y \neq 1,$$
$$I^*(0,y) = I^*(x,1) = 1 \text{ for } x,y \in [0,1].$$

Then $I^*$ is a fuzzy implication such that

$$\Rightarrow^* (a,b) = I^*\left(\frac{b}{a+b}, \frac{a}{a+b}\right).$$

# 3 Affiliated double-implication and equivalency quantifiers

In the paper [5], the method of construction of triads of quantifiers is described.

Starting from an implicational quantifier $\Rightarrow^*$, *affiliated double-implicational quantifier* $\Leftrightarrow^*$ is given by the formula

$$\Leftrightarrow^* (a,b,c) = \Rightarrow^* (a,b+c),$$

and *affiliated equivalency quantifier* $\equiv^*$ is given by the formula

$$\equiv^* (a,b,c,d) = \Rightarrow^* (a+d, b+c).$$

Double-implicational quantifier $\Leftrightarrow^*$ measures the validity of bi-implication $(\varphi \Rightarrow \psi) \wedge (\psi \Rightarrow \varphi)$ in data taking into account only cases where $\varphi$ or $\psi$ is satisfied. Equivalency quantifier $\equiv^*$ measures the validity of equivalency $\varphi \equiv \psi$ in the whole data. Both affiliated quantifiers $\Leftrightarrow^*, \equiv^*$ naturally extend quantification of implication (given by a definition of particual implicational quantifier $\Rightarrow^*$) for covering also two types of symmetric relations between $\varphi$ and $\psi$ in data.

It is proved that the above constructed double-implicational quantifier $\Leftrightarrow^*$ is in some sense the least strict one (out of the class of so-called $\Sigma$-double implication quantifiers, see [7], [5]) satisfying required inequality:

$$\Leftrightarrow^* (a,b,c) \leq \min(\Rightarrow^* (a,b),\ \Rightarrow^* (a,c)).$$

Analogically, the above constructed equivalency quantifier $\equiv^*$ is in some sense the most strict one (out of the class of so-called $\Sigma$-equivalency quantifiers, see [7], [5]) satisfying inequality:

$$\equiv^* (a,b,c,d) \geq \max(\Leftrightarrow^* (a,b,c), \Leftrightarrow^* (d,b,c)).$$

In the case when the starting quantifier $\Rightarrow^*$ is ratio-implicational, the next theorem shows further useful connections between it and affiliated equivalency quantifier $\equiv^*$.

### Theorem

Let $\Rightarrow^*$ be a ratio-implicational quantifier and $\equiv^*$ be its affiliated equivalency quantifier. Then for all $a, b, c, d$ the value $\equiv^* (a,b,c,d)$ lies both

(i) between the values $\Rightarrow^* (a,b)$, and $\Rightarrow^* (d,c)$;

(ii) between the values $\Rightarrow^* (a,c)$, and $\Rightarrow^* (d,b)$.

### Proof.

Recall that the affiliated equivalency quantifier $\equiv^*$ is defined by

$$\equiv^* (a,b,c,d) = \Rightarrow^* (a+d, b+c).$$

(i) Let us discuss possible relations between multiplications of frequencies $ac, bd$:

Let $bd \geq ac$. Then $bd + ab \geq ac + ab$, so $b(a+d) \geq a(b+c)$.

Applying the defining property of ratio-implicational quantifiers for $a, b, a' = a+d, b' = b+c$, we obtain from the last inequality

$$\Rightarrow^* (a+d, b+c) \geq \Rightarrow^* (a,b).$$

Also $bd + cd \geq ac + cd$, so $d(b+c) \geq (a+d)c$, hence

$$\Rightarrow^* (d,c) \geq \Rightarrow^* (a+d, b+c).$$

Let $bd \leq ac$. Using the same steps but with opposite inequalities, we obtain

$$\Rightarrow^* (a, b) \geq \Rightarrow^* (a + d, b + c) \geq \Rightarrow^* (d, c).$$

(ii) The proof is analogical as for point (i) - is based on possible relations between $ab$, and $cd$.

# 4 Discussion of possible truth configurations

Let $\Rightarrow^*$ be a ratio-implicational quantifier, $\Leftrightarrow^*$ and $\equiv^*$ be its affiliated double-implicational and equivalency quantifiers. Assume some truth treshold $t$ from $[0, 1]$ is given. A formulae $\varphi \sim \psi$ is treated as true in data if the value of the quantifier $\sim$ in the four-fold table $< a, b, c, d >$ corresponding to the attributes $\varphi, \psi$ is greater or equal to $t$. Using inequalities presented in the previous paragraph, we shall discuss possible truth configurations of the set of formulae

$$\begin{array}{rllll} \text{Implications:} & \varphi \Rightarrow^* \psi, & \psi \Rightarrow^* \varphi, & \neg\varphi \Rightarrow^* \neg\psi, & \neg\psi \Rightarrow^* \neg\varphi; \\ \text{Double-implications:} & \varphi \Leftrightarrow^* \psi, & \neg\varphi \Leftrightarrow^* \neg\psi; & & \\ \text{Equivalency:} & \varphi \equiv^* \psi. & & & \end{array}$$

There are formally $2^7 = 128$ configurations, but we shall show that most of them are not possible in any data.

## 4.1 Discussion according to truthfullness of equivalency

e0) $\varphi \equiv^* \psi$ is not true. Then

- at most two implications could be true (but excluding the pair $\varphi \Rightarrow^* \psi, \neg\varphi \Rightarrow^* \neg\psi$, and the pair $\psi \Rightarrow^* \varphi, \neg\psi \Rightarrow^* \neg\varphi$);
- no double implications is true.

e1) $\varphi \equiv^* \psi$ is true. Then

- at least two implications are true;
- 0,1 or 2 double-implications could be true.

## 4.2 Discussion according to truthfullness of double-implications

d0) Both $\varphi \Leftrightarrow^* \psi, \neg\varphi \Leftrightarrow^* \neg\psi$ are not true. Then

- the equivalency $\varphi \equiv^* \psi$ could be true or not;
- 0,1,2,3 or 4 implications could be true.

d1) One of double-implications is true. Then

- the equivalency is true;
- at least two implications are true.

d2) Both double-implications are true. Then

- the equivalency is true;
- all implications are true.

## 4.3 Discussion according to truthfullness of implications

This discussion will be provided in details with distinction of different situations. Types of configurations will be described by numbers $i, d, e$ of true implications, double-implications and equivalency, respectively. Each type of configuration will be provided with an apropriate example of four-fold table $< a, b, c, d >$ quantified by the triad of the basic quantifiers

$$\Rightarrow^* (a,b) = \frac{a}{a+b}, \qquad \Leftrightarrow^* (a,b,c) = \frac{a}{a+b+c}, \qquad \equiv^* (a,b,c,d) = \frac{a+d}{a+b+c+d}$$

with the list of true formulae given the treshold $t = 0.7$.

i0) No implication is true. Then no formulae of double-implication or equivalency could be true.
Type: 0/0/0 Example: 10,6,5,9.

i1) Exactly one implication is true. Then no formulae of double-implication or equivalency could be true.
Type: 1/0/0 Example: 10,4,7,9 $\varphi \Rightarrow^* \psi$.

i2) Exactly two implications are true. Two pairs out from six pairs of implications are excluded (namely the pair $\varphi \Rightarrow^* \psi, \neg\varphi \Rightarrow^* \neg\psi$ and the pair $\psi \Rightarrow^* \varphi, \neg\psi \Rightarrow^* \neg\varphi$, because in these cases also the equivalency would be true, hence some third implication also would be true). Remaining four possible pairs of true implications lead to two different cases:

i2a) Either $\varphi \Rightarrow^* \psi, \psi \Rightarrow^* \varphi$ are true or $\neg\varphi \Rightarrow^* \neg\psi, \neg\psi \Rightarrow^* \neg\varphi$ are true. Then corresponding double-implication (either $\varphi \Leftrightarrow^* \psi$ or $\neg\varphi \Leftrightarrow^* \neg\psi$) could be true and also the equivalency could be true. Possible types of configurations:
Type: 2/0/0 Example: 10,1,4,1 $\varphi \Rightarrow^* \psi$, $\psi \Rightarrow^* \varphi$.
Type: 2/0/1 Example: 10,2,3,2 $\varphi \Rightarrow^* \psi$, $\psi \Rightarrow^* \varphi$, $\varphi \equiv^* \psi$.
Type: 2/1/1 Example: 10,1,1,2 $\varphi \Rightarrow^* \psi$, $\psi \Rightarrow^* \varphi$, $\varphi \Leftrightarrow^* \psi$,
$\varphi \equiv^* \psi$.

i2b) Either $\varphi \Rightarrow^* \psi, \neg\psi \Rightarrow^* \neg\varphi$ are true or $\psi \Rightarrow^* \varphi, \neg\varphi \Rightarrow^* \neg\psi$ are true. Then no double-implication could be true. Possible types of configurations:
Type: 2/0/0 Example: 10,3,8,9 $\varphi \Rightarrow^* \psi$, $\neg\psi \Rightarrow^* \neg\varphi$.
Type: 2/0/1 Example: 10,1,8,11 $\varphi \Rightarrow^* \psi$, $\neg\psi \Rightarrow^* \neg\varphi, \varphi \equiv^* \psi$.

i3) Exactly three implications are true. Then the equivalency is true and at most one double-implication is true. Possible types of configurations:
Type: 3/0/1 Example: 10,1,4,3 $\varphi \Rightarrow^* \psi$, $\psi \Rightarrow^* \varphi$, $\neg\psi \Rightarrow^* \neg\varphi$,
$\varphi \equiv^* \psi$.
Type: 3/1/1 Example: 10,1,3,3 $\varphi \Rightarrow^* \psi$, $\psi \Rightarrow^* \varphi$, $\neg\psi \Rightarrow^* \neg\varphi$,
$\varphi \Leftrightarrow^* \psi$, $\varphi \equiv^* \psi$.

i4) All four implications are true. Then the equivalency is true and 0,1, or 2 double-implications could be true. Possible types of configurations:

| | | | | |
|---|---|---|---|---|
| Type: | 4/0/1 | Example: | 10,3,4,13 | $\varphi \Rightarrow^* \psi$, $\psi \Rightarrow^* \varphi$, $\neg\psi \Rightarrow^* \neg\varphi$, $\neg\varphi \Rightarrow^* \neg\psi$, $\varphi \equiv^* \psi$. |
| Type: | 4/1/1 | Example: | 10,2,3,15 | $\varphi \Rightarrow^* \psi$, $\psi \Rightarrow^* \varphi$, $\neg\psi \Rightarrow^* \neg\varphi$, $\neg\varphi \Rightarrow^* \neg\psi$, $\neg\varphi \Leftrightarrow^* \neg\psi$, $\varphi \equiv^* \psi$. |
| Type: | 4/2/1 | Example: | 10,2,1,17 | $\varphi \Rightarrow^* \psi$, $\psi \Rightarrow^* \varphi$, $\neg\psi \Rightarrow^* \neg\varphi$, $\neg\varphi \Rightarrow^* \neg\psi$, $\varphi \Leftrightarrow^* \psi$, $\neg\varphi \Leftrightarrow^* \neg\psi$, $\varphi \equiv^* \psi$. |

Summary of discussion: only 10 types of configurations are possible out from $5{\cdot}3{\cdot}2 = 30$ formally existing types. More detailed analysis would show that there are exactly 27 possible configurations out of 128 formally existing ones.

# 5 Conclusions

In the paper, the class of ratio-implicational quantifiers was introduced and following properties were presented:

- each ratio-implicational quantifier can be represented by a non-decreasing function on rationals;
- there is a correspondence between ratio-implicational quantifiers and fuzzy implications;
- there are double-implication and equivalency quantifiers affiliated to a given ratio-implicational quantifier which leads to the set of seven formulae

  $$\varphi \Rightarrow^* \psi, \psi \Rightarrow^* \varphi, \neg\psi \Rightarrow^* \neg\varphi, \neg\varphi \Rightarrow^* \neg\psi, \varphi \Leftrightarrow^* \psi, \neg\varphi \Leftrightarrow^* \neg\psi, \varphi \equiv^* \psi$$

  connected by the set of inequalities among their values in data;
- number of possible truth-configurations of these formulae in data given some treshold is significantly reduced.

**Acknowledgements**

This work has been supported by the Ministry of Education, Youths and Sports of the Czech Republic (MSM 6138439910).

# References

[1] Aggraval, R. et al. (1996): Fast Discovery of Association Rules. In Fayyad, V.M. et al.: *Advances in Knowledge Discovery and Data Mining.* AAAI Press / MIT Press, pp. 307-328.

[2] Hájek, P., Havránek, T. (1978): *Mechanising Hypothesis Formation - Mathematical Foundations for a General Theory.* Springer-Verlag, Berlin, 396 p.

[3] Hájek, P., Havránek, T., Chytil, M. (1983): *Metoda GUHA*. Academia, Praha, 314 p. (in Czech)

[4] Hájek, P. (1998): *Metamathematics of Fuzzy Logic.* Kluwer Academic Publishers, Dordrecht. 297 p.

[5] Ivánek, J. (1999): On the Correspondence between Classes of Implicational and Equivalence Quantifiers. In: *Principles of Data Mining and Knowledge Discovery. Proc. PKDD'99 Prague* (Zytkow, J., Rauch, J., eds.), Springer-Verlag, Berlin, pp. 116-124.

[6] Ivánek, J. (2005): Construction of Implicational Quantifiers from Fuzzy Implications. In: *Fuzzy Sets and Systems 151*, pp. 381-391.

[7] Rauch, J. (1998): Classes of Four-Fold Table Quantifiers. In: *Principles of Data Mining and Knowledge Discovery*, (Quafafou, M. and Zytkow, J., eds.), Springer Verlag, Berlin, pp. 203-211.

[8] Zembowicz, R., Zytkow, J. (1996): From Contingency Tables to Various Forms of Knowledge in Databases. In: Fayyad, U.M. et al.: *Advances in Knowledge Discovery and Data Mining.* AAAI Press / The MIT Press, pp. 329-349.

# Some Results on Set-Valued Possibilistic Distributions

**Ivan Kramosil**[*]
Institute of Computer Science
Academy of Sciences of the Czech Republic
kramosil@cs.cas.cz

**Abstract**

When proposing and processing uncertainty decision making algorithms of various kinds and purposes we meet more and more often probability distributions ascribing to random events non-numerical uncertainty degrees. The reason is that we have to process systems of uncertainties for which the classical conditions like $\sigma$-additivity or linear ordering of values are too restrictive to define sufficiently closely the nature of uncertainty we would like to specify and process. For the case of non-numerical uncertainty degrees at least the two criteria may be considered. First systems with rather complicated, but sophisticated and non-trivially formally analyzable uncertainty degrees. E.g., uncertainties supported by some algebras or partially ordered structures. Contrary, we may consider more easy non-numerical, but on the intuitive level interpretable relations. Well-known examples of such structures are set-valued possibilistic measures. Some perhaps interesting particular results in this direction will be introduced and analyzed in the contribution.

## 1 Introduction

In the measure theory and, consequently, in probability theory, the sizes of sets and uncertainty (in the sense of randomness as well as of fuzziness and possibility degrees) were quantified by numbers, going from finite natural numbers to rational and then real (or, perhaps, complex-valued numbers). The development of real-valued probability theory took their tops by Kolmogorov axiomatic theory of probability as systematically explained and applied in [4], [6] or elsewhere.

On the other side, the correctness and legality of application of the classical probability theory and its consequences (mathematical statistics, Shannon entropy and information theory, ...) to problems from real life is based on the assumption that certain non-trivial assumptions are satisfied and verified (the precise knowledge of apriori probabilities, statistical independence of some random variables and/or precisely

[*] This research is supported by the grant P202/10/1826 of the Grant Agency of the Czech Republic. Partial institutional support RVO: 67985807 is also acknowledged.

known type and degrees of their dependencies together with the detailed conditional probabilities, ...).

Qualitatively different models of uncertainty quantification and processing, even if still with numerical degrees, are real-valued fuzzy sets, defined by mappings taking the basic space $\Omega$ into the unit interval $[0,1]$, hence, extending the binary-valued characteristic functions of standard set, to functions with values in the closed interval $[0,1]$.

The pioneering Zadeh's idea of fuzzy sets emerged in 1965 in [13, 5] and, as soon as in 1967. J. A. Goguen entered on scene with the further step – fuzzy sets with non-numerical membership degrees. In particular, J. A. Goguen considered uncertainty, in the sense of fuzziness degrees, as elements of complete lattice, let us recall that complete lattice is defined as the p.o.set (partially ordered set) in which for each nonempty subset supremum and infimum are defined.

When quantifying sizes by numbers we have to keep in mind that this introduces into the model the complete ordering of numbers which need not correspond to sizes of pieces of uncertainty in question. Among the structures working with uncertainties and keeping in mind the idea to classify as incomparable also set-quantified degrees of uncertainty with the same values of real-valued measures, set-valued possibilistic measures seem to be sufficiently elastic and resilient to be taken as intuitively acceptable non-numerical size-quantifying mathematical model.

Let us survey, very briefly, the contents of particular sections. Our aim will be to minimize the quantity and complexity of preliminaries necessary for a non-fully oriented reader in order to understand the text. In Section 2 we introduce the structures for quantifying uncertainty (or uncertainties) by set values. It is perhaps worth being so-called just now that probability measure and probability theory is based on standard combination of set-valued uncertainty quantification (random events are sets) with also the standard real-valued quantification of the set-valued random events.

In Section 3 we introduce three alternative ways how to define mappings keeping at least some properties of conditional probabilities. This problem seems to be promising for some new and interesting results In Section 4 we define and analyze set-valued entropy function over set-valued possibilistic function with the aim to solve the problem arising when the possibilistic distribution takes the maximum value $\mathbf{1}_{\mathcal{T}}(=X)$ for at last two different arguments. Analogously to the case of real-valued probability measure the Shannon entropy function [11] takes the maximum value $\mathbf{1}_{\mathcal{T}}(=X)$ so that the qualities of this entropy function cannot be used as a tool for neither a partial ordering of different alternatives of possibilistic distrubution when choosing the best one for the application in question. Very roughly speaking, the idea is to modify the space of values in which set-valued entropy function takes its values, in such a way that the supremum value of the set-valued entropy function is taken for just one value $\omega_0$ from the basic space $\Omega$ of the possibilistic distribution in question.

## 2 Set-valued possibilistic distributions

Let $\Omega$ and $X$ be nonempty sets, let $\mathcal{P}(X)$ be the set of all subsets of $X$ (the power-set over $X$), let $\pi : \Omega \to \mathcal{P}(X)$ be a mapping ascribing to each $\omega \in \Omega$ a subset $\pi(\omega) \subset X$

(i.e., $\pi(\omega) \in \mathcal{P}(X)$). The mapping $\pi$ is called *set-valued possibilistic distribution* on $\Omega$, if $\bigcup_{\omega \in \Omega} \pi(\omega) = X$.

For each $A \subset \Omega$, set $\Pi(A) = \bigcup_{\omega \in A} \pi(\omega)$. The mapping $\Pi : \mathcal{P}(\omega) \to \mathcal{P}(X)$ is called the $\mathcal{P}(X)$-*valued possibilistic measure* induced on $\mathcal{P}(\Omega)$ by the set-valued possibilistic distribution $\pi$ on $\Omega$. The important characteristic of the $\mathcal{P}(X)$-valued possibilistic distribution $\pi$ (and of the related $\mathcal{P}(X)$-valued possibilistic measure $\Pi$ induced by $\pi$) is the so called *possibilistic* (or *Sugeno*) *entropy* defined by the *Sugeno integral* $I(\pi)$. For the particular case of the set-valued possibilistic distribution $\pi$ on $\Omega$ defined as above the definition reads as follows:

$$I(\pi) = \bigcup_{\omega \in \Omega} [\Pi(\Omega - \{\omega\}) \cap \pi(\omega)] \subset X. \tag{2.1}$$

E.g., in the most simple case when $\Omega = X$ and $\pi(\omega) = \{\omega\}$, we obtain that $\Pi(A) = \bigcup_{\omega \in A} \pi(\omega) = \bigcup_{\omega \in A} \{\omega\} = A$. For the entropy $I(\pi)$ we obtain that

$$I(\pi) = \bigcup_{\omega \in \Omega} [\Pi(\Omega - \{\omega\}) \cap \pi(\omega)] = \bigcup_{\omega \in \Omega} ((\Omega - \{\omega\}) \cap \{\omega\}) = \emptyset \tag{2.2}$$

let us recall that the empty subset of $X$ denotes the zero element of the complete lattice (complete Boolean algebra, as a matter of fact) $\langle \mathcal{P}(X), \subseteq \rangle$.

**Fact 2.1** *Let $\Omega$ and $X$ be nonempty sets, let $\pi : \Omega \to \mathcal{P}(X)$ be a $\mathcal{P}(X)$-valued possibilistic distribution on $\Omega$ such that, for each $\omega_1, \omega_2, \omega_1 \neq \omega_2$, $\pi(\omega_1) \cap \pi(\omega_2) = \emptyset$ holds. Then for each $A, B \subset X, A \cap B = \emptyset$, we obtain that $\Pi(A) \cap \Pi(B) = \emptyset$ holds.*

*Proof:* An easy calculation yields that

$$\Pi(A) \cap \Pi(B) = \left( \bigcup_{\omega \in A} \pi(\omega) \right) \cap \left( \bigcup_{\omega \in B} \pi(\omega) \right) =$$

$$= \bigcup_{\omega_1 \in B} \left[ \left( \bigcup_{\omega \in A} \pi(\omega) \right) \cap \pi(\omega_1) \right] = \bigcup_{\omega_1 \in B} \bigcup_{\omega \in A} (\pi(\omega_1) \cap \pi(\omega)) = \emptyset, \tag{2.3}$$

as the sets $A$ and $B$ are disjoint. The assertion is proved. □

**Lemma 2.1** *Let $\Omega$ and $X$ be nonempty sets, let $\pi : \Omega \to \mathcal{P}(X)$ be a $\mathcal{P}(X)$-valued possibilistic distribution on $\Omega$. Then $I(\pi) = \emptyset$ iff $\pi(\omega_1) \cap \pi(\omega_2) = \emptyset$ for each $\omega_1, \omega_2 \in \Omega, \omega_1 \neq \omega_2$.*

*Proof:* If $\pi(\omega_1) \cap \pi(\omega_2) = \emptyset$ for each $\omega_1, \omega_2 \in \Omega, \omega_1 \neq \omega_2$, then

$$I(\pi) = \bigcup_{\omega \in \Omega} [\pi(\Omega - \{\omega\}) \cap \pi(\omega)] = \bigcup_{\omega \in \Omega} [\pi(\Omega - \{\omega\}) \cap \Pi(\{\omega\})] = \emptyset \tag{2.4}$$

holds, due to Fact 2.1.

On the other side, let $\omega_1, \omega_2 \in \Omega, \omega_1 \neq \omega_2$, be such that $\pi(\omega_1) \cap \pi(\omega_2) \neq \emptyset$ Then $\omega_2 \in \Omega - \{\omega_1\}$ holds, so that

$$\Pi(\Omega - \{\omega_1\}) \cap \pi(\omega_1) \supset \pi(\omega_2) \cap \pi(\omega_1) \neq \emptyset, \tag{2.5}$$

consequently,

$$I(\pi) \supset \Pi(\Omega - \{\omega_1\}) \cap \pi(\omega_1) \supset \pi(\omega_2) \cap \pi(\omega_1) \neq \emptyset, \tag{2.6}$$

follows. The assertion is proved. □

**Theorem 2.1** *Let $\Omega$ and $X$ be nonempty sets, let $\pi_1$, $\pi_2$ be $\mathcal{P}(X)$-valued possibilistic distributions such that, for each $\omega \in \Omega, \pi_1(\omega) \subset \pi_2(\omega)$ holds. Then $I(\pi_1) \subset I(\pi_2)$ holds.*

*Proof:* By definition,

$$I(\pi_1) = \bigcup_{\omega \in \Omega} [\pi_1(\Omega - \{\omega\}) \cap \pi_2(\omega)]. \tag{2.7}$$

For each $\omega \in \Omega$, the inclusion

$$\Pi_1(\Omega - \{\omega\}) = \bigcup_{\omega^* \in \Omega - \{\omega\}} \pi_1(\omega^*) \subset \bigcup_{\omega^* \in \Omega - \{\omega\}} \pi_2(\omega^*) = \Pi_2(\Omega - \{\omega\}) \tag{2.8}$$

is valid, as $\pi_1(\omega^*) \subseteq \pi_2(\omega^*)$ holds for each $\omega^* \in \Omega$. Consequently, the inclusion

$$\Pi_1(\Omega - \{\omega\}) \cap \pi_1(\omega) \subset \Pi_2(\Omega - \{\omega\}) \cap \pi_2(\omega) \tag{2.9}$$

holds for each $\omega \in \Omega$, so that the inclusion $I(\pi_1)$ immediately follows. The assertion is proved. □

**Lemma 2.2** *Let $\Omega, X$ be nonempty sets, let $\pi : \Omega \to \mathcal{P}(X)$ be a $\mathcal{P}(X)$-valued possibilistic distribution. If there are $\omega_1, \omega_2 \in \Omega, \omega_1 \neq \omega_2$, such that $\pi(\omega_1) = \pi(\omega_2) = X$, then $I(\pi) = X = \mathbf{1}_{\mathcal{P}(X)}$.*

*Proof:* Let $\omega_1, \omega_2 \in \Omega, \omega_1 \neq \omega_2$, be such that $\pi(\omega_1) = \pi(\omega_2) = X$, consider the set $\Pi(\Omega - \{\omega_1\}) \cap \pi(\omega_1)$. Then $\omega_2 \in \Omega - \{\omega_1\}$ holds, hence,

$$\Pi(\Omega - \{\omega_1\}) = \bigvee^{\omega^* \in \Omega - \{\omega_1\}} \pi(\omega^*) \supset \pi(\omega_2) = X \tag{2.10}$$

holds and $\Pi(\Omega - \{\omega_1\}) = X$ follows. Replacing mutually $\omega_1$ and $\omega_2$ we obtain that $\Pi(\Omega - \{\omega_2\}) = X$ holds as well, hence,

$$X = \Pi(\Omega - \{\omega_j\}) \cap \pi(\omega_j) = \bigcup_{\omega \in \Omega} [\Pi(\Omega - \{\omega\}) \cap \pi(\omega)] = I(\pi) \tag{2.11}$$

holds for any $j$ and the assertion is proved. □

Denote by $Q$ the set of all $\mathcal{P}(X)$-valued possibilistic distributions on $\Omega$. If $\pi_1, \pi_2$ are $\mathcal{P}(X)$-possibilistic distributions on $\Omega$ such that $\pi_1(\omega) \subset \pi_2(\omega)$ holds for each $\omega \in \Omega$, we write $\pi_1 \leq \pi_2$ and say that $\pi_1$ is *majorized* by $\pi_2$ or that $\pi_2$ is an upper bound for $\pi_1$. As proved in Theorem 2.1 if $\pi_1 \leq \pi_2$ holds, then $I(\pi_1) \subseteq I(\pi_2)$ holds as well.

The universe implication does not hold in general, i.e., if $I(\pi_1) \subseteq I(\pi_2)$ is valid, then $\pi_1 \leq \pi_2$ need not hold. For entropy $I(\pi_1)$ we obtain that

$$\begin{aligned} I(\pi_1) &= \bigcup_{\omega \in \Omega} [\Pi_1(\Omega - \{\omega\}) \cap \pi_2(\omega)] = \\ &= [\Pi_1(\Omega - \{\omega_1\}) \cap \pi_1(\omega_1)] \cup [\Pi_1(\Omega - \{\omega_2\}) \cap \pi_1(\omega_2)] = \\ &= (\pi_1(\omega_2) \cap \pi_1(\omega_1)) \cup (\pi_1(\omega_1) \cap \pi_1(\omega_2)) = \\ &= (\emptyset \cap X) \cup (X \cap \emptyset) = \emptyset. \end{aligned} \tag{2.12}$$

For $I(\pi_2)$ the calculations and the results are the same, so that $I(\pi_1) = I(\pi_2)$, but neither $\pi_1 \leq \pi_2$ nor $\pi_2 \leq \pi_1$ holds.

**Lemma 2.3** *Let $\pi$ be a $\mathcal{P}(X)$-valued distribution on $\Omega$. Then for each $\mathcal{S} \subset \mathcal{P}(\Omega)$ the relation*

$$\begin{aligned} \Pi\left(\bigcup \mathcal{S}\right) &= \Pi\left(\bigcup\{A : A \in \mathcal{S}\}\right) = \bigvee^{\mathcal{T}} \left\{\pi(\omega) : \omega \in \bigcup \mathcal{S}\right\} = \\ &= \bigcup\{\{\pi(\omega) : \omega \in A\} : A \in \mathcal{S}\} : A \in \mathcal{S}\} = \\ &= \bigvee^{\mathcal{T}}\{\Pi(A) : A \in \mathcal{S}\} = \bigcup\{\Pi(A) : A \in \mathcal{S}\} \end{aligned} \tag{2.13}$$

*holds.*

*Proof:* Obvious. □

Let us denote by $Q(\Omega, X)$ the space of all $\mathcal{P}(A)$-valued possibilistic distributions over the space $\Omega$, in symbols,

$$Q(\Omega, X) = \{\pi : \pi : \Omega \to \mathcal{P}(X), \bigcup\{\pi(\omega) : \omega \in \Omega\} = \mathbf{1}_{\mathcal{T}} = X\}. \tag{2.14}$$

Let $\leq^*$ be the binary relation, on $Q(\Omega, X)$, i.e., the subset of the Cartesian product $Q(\Omega, X) \times Q(\Omega, X)$ defined in this way: for each $\pi_1, \pi_2 \in Q(\Omega, X), \pi_1 <^* \pi_2$ holds iff $\pi_1(\omega) \subseteq \pi_2(\omega)$ holds for each $\omega \in \Omega$. It is possible that $\pi_1 <^* \pi_2$ holds for two $\mathcal{P}(X)$-distribution $\pi_1, \pi_2$ such that $\pi_1(\omega) \subset \pi_2(\omega)$ is the case for some $\omega \in \Omega$ and, of course, $\pi_1(\omega^*) \subseteq \pi_2(\omega^*)$ holds for two $\mathcal{P}(X)$-distributions $\pi_1, \pi_2$ such that $\pi_1(\omega) \subset \pi_2(\omega)$ is the case for some $\omega \in \Omega$ and, of course, $\pi_1(\omega^*) \subseteq \pi_2(\omega^*)$ holds for each $\omega^* \in \Omega$.

**Lemma 2.4** *The ordered pair $\mathcal{D} = \langle Q(\Omega, X), \leq^* \rangle$ is a p.o.set which defines a complete upper semilattice, so that for each nonempty subset $E \subset D$ the supremum $\pi(E) = \bigvee^{\mathcal{D}}\{\pi : \pi \in E\}$ is defined. Given explicitly, $\pi^E$ is the mapping which takes $\Omega$ into $\mathcal{P}(X)$ in such a way that for each $\omega \in \Omega$*

$$\pi^E(\omega) = \bigcup\{\pi \in E : \pi(\omega)\} \tag{2.15}$$

*This mapping obviously defines a $\pi$-valued possibilistic distribution on $\Omega$.*

*Proof:* Obvious. □

However, the situation with the infimum of a set $E$ of $\mathcal{P}(X)$-distributions is not dual to $\bigvee^D E$. We may define the mapping $M(E) : \Omega \to \mathcal{P}(X)$ in such a way that, for each $\omega \in \Omega$, $M(E)(\omega) = \bigcap\{\pi(\omega) : \pi \in E\}$, but this mapping does not meet the condition $\bigvee^{\mathcal{D}}\{M(E)(\omega) : \omega \in \Omega\} = \mathbf{1}_{\mathcal{P}(X)} = X$.

**Lemma 2.5** *Let $E \subset Q$ be a nonempty set of $\mathcal{P}(X)$-distributions, for each $\pi \in Q$ let $\Pi_\pi : \mathcal{P}(\Omega) \to \mathcal{P}(X)$ denote the corresponding induced $\mathcal{P}(X)$-possibilistic measure on $\mathcal{P}(\Omega)$. Then, for each $A \subset \Omega$, the relation $\Pi_\pi E(A) = \bigvee^{\mathcal{T}}\{\Pi_\pi(A) : \pi \in E\}$ holds.*

*Proof:* For each $A \subset \Omega$ we obtain that

$$\begin{aligned}
&\overset{\mathcal{T}}{\bigvee}\{\pi_\pi(A) : \pi \in E\} = \overset{\mathcal{T}}{\bigvee}\{\{\overset{\mathcal{T}}{\bigvee}\pi(\omega) : \omega \in A\} : \pi \in E\} = \\
= \;& \overset{\mathcal{T}}{\bigvee}\{\pi(\omega) : \omega \in \Omega, \pi \in E\} = \overset{\mathcal{T}}{\bigvee}\{\{\overset{\mathcal{T}}{\bigvee}\pi(\omega) : \pi \in E\} : \omega \in A\} \\
= \;& \overset{\mathcal{T}}{\bigvee}\{\pi^E(\omega) : \omega \in A\} = \Pi_{\pi^E}(A).
\end{aligned} \tag{2.16}$$

The assertion is proved. □

According to the way in which $\mathcal{P}(X)$-valued possibilistic measure $\Pi$ on $\mathcal{P}(\Omega)$ induced by a $\mathcal{P}(X)$-valued possibilistic distribution $\pi$ on $\Omega$ is defined, the set function $\Pi$ is extensional with respect to the supremum operation $\bigvee^{\mathcal{T}}$ on $\mathcal{T} = \mathcal{P}(X)$ in the sense that for each nonempty system $\mathcal{A}$ of subsets of $\Omega$ the identity

$$\Pi\left(\bigcup\mathcal{A}\right) = \overset{\mathcal{T}}{\bigvee}\{\Pi(A) : A \in \mathcal{A}\} \tag{2.17}$$

holds. In particular, for $\mathcal{A} = \{A_1, A_2\}$, $\Pi(A_1) \cup \Pi(A_2) = \Pi(A_1 \cup A_2)$. For the operation of infimum the relation dual to (2.17) is not the case, in general, only the inclusion $\Pi(A \cap B) \subseteq \Pi(A) \cap \Pi(B)$ is valid, as $\Pi(A \cap B) \subset \Pi(A)$ and $\Pi(A \cap B) \subset \Pi(B)$ holds trivially.

As the most simple $\mathcal{P}(X)$-valued possibilstic distribution $\pi$ for which the induced $\mathcal{P}(X)$-measure $\Pi$ on $\mathcal{P}(\Omega)$ is extensional also w.r.to the operation of infimum $\bigwedge$ let us consider the identity mapping on $\mathcal{P}(\Omega)$. Take $\Omega = X$, take $\pi(\omega) = \{\omega\}$ for every $\omega \in \Omega$, so that, for each $A \subset \Omega$, $\Pi(A) = \bigcap_{A \in \mathcal{A}} \Pi(A)$ follows, in particular, $\Pi(A \cap B) = \Pi(A) \cap \Pi(B)$ holds.

**Definition 2.1** *$\mathcal{P}(X)$-valued possibilistic distribution $\pi$ taking a nonempty set $\Omega$ into the power-set $\mathcal{P}(X)$ over a nonempty set $X$ is called completely extensional, if for each nonempty system $\mathcal{A}$ of subsets of $\Omega$ the relation*

$$\Pi\left(\bigcap \mathcal{A}\right) = \Pi\left(\bigcap_{A \in \mathcal{A}} A\right) = \bigcap_{A \in \mathcal{A}} \Pi(A) \tag{2.18}$$

*holds. The $\mathcal{P}(X)$-distribution $\pi$ is called extensional, if*

$$\Pi(A \cap B) = \Pi(A) \cap \Pi(B) \tag{2.19}$$

*holds for each $A, B \subset \Omega$.*

**Lemma 2.6** *Let $\pi$ be a $\mathcal{P}(X)$-valued possibilistic distribution defined on a nonempty space $\Omega$, taking its values in the power-set $\mathcal{P}(X)$ over a nonempty space $X$ and such that $\pi(\omega_1) \cap \pi(\omega_2) = \emptyset$ holds for each $\omega_1, \omega_2 \in \Omega, \omega_1 \neq \omega_2$. Then the induced $\mathcal{P}(X)$-possibilistic measure on $\mathcal{P}(\Omega)$ is extensional in the sense that $\Pi(A) \cap \Pi(B) = \Pi(A \cap B)$ is valid for each $A, B \subset \Omega$.*

*Proof:* First of all, let us consider the case when the sets $A, B$ are disjoint. Then

$$\begin{aligned}
\Pi(A) \cap \Pi(B) &= \left(\bigvee_{\omega_1 \in A} \pi(\omega_1)\right) \cap \left(\bigvee_{\omega_2 \in B} \pi(\omega_2)\right) = \\
&= \bigcup_{\langle \omega_1, \omega_2 \rangle, \omega_1 \in A, \omega_2 \in B} (\pi(\omega_1) \cap \pi(\omega_2)) = \emptyset = \\
&= \Pi(\emptyset) = A \cap B = \Pi(A \cap B),
\end{aligned} \tag{2.20}$$

as for each $\omega_1 \in A, \omega_2 \in B, \omega_1 \neq \omega_2$ and $\pi(\omega_1) \cap \pi(\omega_2) = \emptyset$ holds.

For each $A, B \subset \Omega, A = (A - B) \cup (A \cap B), B = (B - A) \cup (A \cap B)$ holds, so that

$$\begin{aligned}
&\Pi(A) \cap \Pi(B) = [\Pi((A - B) \cup (A \cap B))] \cap [\Pi((B - A) \cup (A \cap B))] = \\
= \;& [\Pi(A - B) \cup \Pi(A \cap B)] \cap [\Pi(B - A) \cup \Pi(A \cap B)] = \\
= \;& [\Pi(A - B) \cap \Pi(B - A)] \cup [\Pi(A \cap B) \cap \Pi(B - A)] \cup \\
\cup \;& [\Pi(A \cap B) \cap \Pi(A - B)] \cup \Pi(A \cap B) = \Pi(A \cap B),
\end{aligned} \tag{2.21}$$

as

$$(A - B) \cap (B - A) = (A \cap B) \cap (B - A) = (A \cap B) \cap (A - B) = \emptyset, \tag{2.22}$$

so that, due to (2.20)

$$\Pi(A - B) \cap \Pi(B - A) = (A \cap B) \cap (B - A) = (A \cap B) \cap (A - B) = \emptyset, \tag{2.23}$$

The assertion is proved. □

# 3 Conditioned set-valued possibilistic distributions and measures

Conditioned (or conditional) probability distributions are very important tools in probability theory. Within the framework of the standard Kolmogorov axiomatic probability theory the mathematical formalization of this transformation is very simple and well-known. Let $\langle \Omega, \mathcal{A}, P \rangle$ be a probability space. Subsets of $\Omega$ belonging to $\mathcal{A}$ are called *random events*, hence, for each $A \in \mathcal{A}$ the real number $P(A) \in [0,1]$ is ascribed and called the *probability of (the random event) A*. Given another random event $B \in \mathcal{A}$ such that $P(B) > 0$ holds, the conditioned probability of (the random event) $A$ under the condition that (the random event) $B$ holds is denoted by $P(A/B)$ and defined by the well-known formula

$$P(A/B) = P(A \cap B)/P(B). \tag{3.1}$$

This definition cannot the immediately translated into the model and language of $\mathcal{T}$-valued possibilistic distributions because of the fact that operation of division between the values $P(A \cap B)$ and $P(B)$ cannot be defined in $\mathcal{T}$. Let us proceed in this way: we introduce three alternative approaches and for each of them we will examine its role when taken as conditioned probability and measure.

So, let $\mathcal{T} = \langle X, \subseteq \rangle, \Omega, \pi : \Omega \to \mathcal{P}(X)$ such that $\bigcup_{\omega \in \Omega} \pi(\omega) = X = \mathbf{1}_{\mathcal{T}}$ and $\Pi : \mathcal{P}(\Omega) \to \mathcal{P}(X)$ defined by $\Pi(A) = \bigcup_{\omega \in A} \pi(\omega)$ for each $A \subset X$ be as above. Given $B \subset \Omega$, let us define three mappings $\pi^i(\omega/B) : \Omega \to \mathcal{P}(X)$ in this way.

$$\begin{aligned}
&(i) && \pi^1(\omega/B) = \pi(\omega) \cap \Pi(B), && (3.2a)\\
&(ii) && \pi^2(\omega/B) = \pi(\omega), \text{ if } \omega \in B, \pi^2(\omega/B) = \emptyset (= \emptyset_{\mathcal{T}}), &&\\
& && \text{if } \omega \in \Omega - B, && (3.2b)\\
&(iii) && \pi^3(\omega/B) = \Pi(\Omega - B) \cup \pi(\omega) = \Pi((\Omega - B) \cup \{\omega\}). && (3.2c)
\end{aligned}$$

Let us investigate the most elementary properties of these three mappings. Define, for each $i = 1,2,3$ and each $B \subset \Omega$ the mapping $\Pi^i(\cdot/B) : \mathcal{P}(\Omega) \to \mathcal{P}(X)$ in this way: for each $A \subset \Omega$,

$$\Pi^i(A/B) = \bigvee_{\omega \in A}^{\mathcal{T}} \pi^i(\omega/B) = \bigcup_{\omega \in A} \pi^i(\omega/B). \tag{3.3}$$

Hence, for each $i = 1,2,3$ we obtain explicitly that

$$\begin{aligned}
\Pi^1(A/B) &= \bigcup_{\omega \in A} \pi^1(\omega/B) = \bigcup_{\omega \in A} (\pi(\omega) \cap \Pi(B)) = \left( \bigcup_{\omega \in A} \pi(\omega) \right) \cap \Pi(B) = \\
&= \Pi(A) \cap \Pi(B), && (3.4)
\end{aligned}$$

$$\Pi^2(A/B) = \bigcup_{\omega \in A} \pi^2(\omega/B) = \bigcup_{\omega \in A \cap B} \pi(\omega) = \Pi(A \cap B), \tag{3.5}$$

$$\begin{aligned}\Pi^3(A/B) &= \bigcup_{\omega\in A}\pi^3(\omega/B) = \bigcup_{\omega\in A}(\Pi(\Omega-B)\cup\pi(\omega)) = \\ &= \Pi(\Omega-B)\cup\bigcup_{\omega\in A}\pi(\omega) = \\ &= \Pi(\Omega-B)\cup\Pi(A) = \Pi((\Omega-B)\cup A) \qquad (3.6)\end{aligned}$$

For the extremum values $A=\Omega$ or $B=\Omega$ we obtain that

$$\begin{aligned}\Pi^1(\Omega/B) &= \Pi(\Omega)\cap\Pi(B) = \Pi(B), \\ \Pi^2(\Omega/B) &= \Pi(\Omega\cap B) = \Pi(B), \\ \Pi^3(\Omega/B) &= \Pi((\Omega-B)\cup\Omega) = \Pi(\Omega) = \mathbf{1}_{\mathcal{T}}, \\ \Pi^1(A/\Omega) &= \Pi(A)\cap\Pi(\Omega) = \Pi(A), \\ \Pi^2(A/\Omega) &= \Pi(A\cap\Omega) = \Pi(A), \\ \Pi^3(A/\Omega) &= \Pi((\Omega-\Omega)\cup A) = \Pi(A), \qquad (3.7)\end{aligned}$$

So, $\pi^1(\cdot/B)$ and $\pi^2(\cdot/B)$ define $\mathcal{T}$-possibilistic distribution on $B$ (supposing that $B\neq\emptyset$), $\pi^3(\cdot/B)$ defines a $\mathcal{T}$-possibilistic distribution on $\Omega$. Moreover, if $B=\Omega$, then $\Pi^i(\cdot/B)$ is identical with the apriori possibilistic distribution $\pi$ on $\Omega$ for each $i=1,2,3$. Let us recall that in standard probability theory, if $B\subset\Omega$ is such that $P(B)=1$, then for each $A\subset\Omega$ the identity $P(A/B)=P(A\cap B)/P(B)=P(A)$ holds. The intuition behind is quite simple – the occurence of certain (i.e., which the probability 1 valid) random event does not bring any new information, so that no modification of the apriori probability measure results. All the three set functions $\Pi^i(\cdot/B), i=1,2,3$, also possess this important property.

More generally, not only for $A=\Omega$, but for each $A\supseteq B$ the result $\Pi^i(A/B)=\Pi(B)$ (for $i=1,2$) or $\Pi^3(A/B)=\mathbf{1}_{\mathcal{T}}$ holds, as may be easily checked by inspection of the formulas (3.4), (3.5), and (3.6).

When approaching to a more detailed analysis of the three $\mathcal{P}(X)$-valued mappings $\pi^i(\omega/B), i=1,2,3$, let us begin with the mapping $\pi^3(\omega/B)$ defined by (3.2c), so that

$$\pi^3(\omega/B) = \Pi(\Omega-B)\cup\pi(\omega) = \Pi((\Omega-B)\cup\{\omega\}). \qquad (3.8)$$

Hence, for each $A,B\subset\Omega$,

$$\begin{aligned}\pi^3(A/B) &= \bigcup_{\omega\in A}\pi^3(\omega/B) = \bigcup_{\omega\in A}(\Pi(\Omega-B)\cup(\omega)) = \\ &= \Pi(\Omega-B)\cup\bigcup_{\omega\in A}\pi(\omega) = \Pi(\Omega-B)\cup\Pi(A) = \\ &= \Pi((\Omega-B)\cup A). \qquad (3.9)\end{aligned}$$

The reason for this preference given to $\pi^3(\cdot/B)$ consists in the fact that $\pi^3(\omega/B)$ is, for each $B$, the only of the three mappings in question which meets the condition of normalization, i.e., for which

$$\bigcup_{\omega\in\Omega}\pi^3(\omega/B) = \bigcup_{\omega\in\Omega}(\pi(\Omega - B)\cup\pi(\omega)) =$$

$$= \Pi(\Omega - B)\cup\bigcup_{\omega\in\Omega}\pi(\omega) = \Pi(\Omega - B)\cup X = X = \mathbf{1}_{\mathcal{T}}. \tag{3.10}$$

So, the $\mathcal{P}(X)$-valued entropy $I(\pi^3(\cdot/B))$ is defined and, writing $\hat{\pi}(\omega)$ for $\pi^3(\omega/B)$ in order to simplify the rotation, may be written by

$$I(\pi^3(\cdot/B)) = I(\hat{\pi}) = \bigcup_{\omega\in\Omega}(\Pi^3(\Omega - \{\omega\})\cap\hat{\pi}(\omega)). \tag{3.11}$$

Let $\omega_0\in\Omega$ be such that $\hat{\pi}(\omega_0) = X$. Then

$$\begin{aligned}
I(\pi^3(\cdot/B)) &= I(\hat{\pi}) = \bigcup_{\omega\in\Omega,\omega\neq\omega_0}\hat{\Pi}((\Omega - \{\omega\})\cap\hat{\pi}(\omega))\cup\hat{\Pi}(\Omega - \{\omega_0\})\cap\hat{\pi}(\omega_0) = \\
&= \bigcup_{\omega\in\Omega,\omega\neq\omega_0}(X\cap\hat{\pi}(\omega))\cup(\hat{\Pi}(\Omega - \{\omega_0\})\cap X = \\
&= \bigcup_{\omega\in\Omega,\omega\neq\omega_0}\hat{\pi}(\omega)\cup\hat{\Pi}(\Omega - \{\omega_0\}) = \hat{\Pi}(\Omega - \{\omega_0\}) = \\
&= \Pi^3((\Omega - \{\omega_0\})/B).
\end{aligned} \tag{3.12}$$

# 4 Refined set-valued entropy functions

Let us re-consider and analyse, in more detail, Lemma 2.2. According to this result, if there are $\omega_1,\omega_1\in\Omega,\omega_1\neq\omega_2$, such that $\pi(\omega_1) = \pi(\omega_2) = X$, then $I(\pi) = X = \mathbf{1}_{\mathcal{P}(X)}$. Hence, each decision rule picking up just one $\omega_0\in\Omega$ must be based on more input parameters than those expressible by the values of the entropy function $I(\pi)$. However, the same is the situation in the most simple probability space $\langle\Omega,\mathcal{A},P\rangle$, where $\Omega = \{\omega_1,\omega_2\}$ and $P(\{\omega_1\}) = P(\{\omega_2\}) = \frac{1}{2}$. The following lemma may be taken as a complementary formulation of the conditions when $I(\pi)\neq 1_{\mathcal{P}(X)} = X$ is the case.

**Lemma 4.1** *Let $\Omega, X$ be nonempty sets, let $\pi:\Omega\to\mathcal{P}(X)$ be a $\mathcal{P}(X)$-possibilistic distribution on $\Omega$, let $\omega_0\in\Omega$ be such that $\pi(\omega_0) = X$. Then*

$$I(\pi) = \pi(\Omega - \{\omega_0\}) \tag{4.1}$$

*holds. Consequently, if $\Pi(\Omega - \{\omega_0\})\subsetneq X$ holds, then $I(\pi)\subsetneq X$ follows.*

*Proof:* For $I(\pi)$ we have

$$\begin{aligned}
I(\pi) &= \bigcup_{\omega\in\Omega}(\Pi(\Omega - \{\omega\})\cap\pi(\omega)] = \\
&= \bigcup_{\omega\in\Omega,\omega\neq\omega_0}[\Pi(\Omega - \{\omega\})\cap\pi(\omega)]\cup\Pi(\Omega - \{\omega_0\})\cap\pi(\omega_0).
\end{aligned} \tag{4.2}$$

If $\omega \neq \omega_0$, then $\omega_0 \in (\Omega - \{\omega\})$ and $\Pi(\Omega - \{\omega\}) = X = \pi(\omega_0)$ holds, so that

$$\begin{aligned} I(\pi) &= \left( \bigcup_{\omega \in \Omega, \omega \neq \omega_0} \pi(\omega) \right) \cup \Pi(\Omega - \{\omega_0\}) = \\ &= \Pi(\Omega - \{\omega_0\}). \end{aligned} \tag{4.3}$$

is valid and the assertion is proved. □

An easy corollary of Lemma 4.1 reads as follows. Let $\Omega, X$ and $\pi$ be as in Lemma 4.1, let there exist $x_0 \in X$ such that there is only one $\omega_0 \in \Omega$ with the property $x_0 \in \pi(\omega_0)$ and $\Pi(\Omega - \{\omega_0\}) \subsetneq X$. Then $I(\pi) = \Pi(\Omega - \{\omega_0\}) \subsetneq X$ follows.

For several other results on refined set-valued entropy functions and on set-valued possibilistic distributions see [10].

## 5 Conclusions

According to what we told in the introductory section, our aim was to introduce and analyze some possibilistic distributions and related possibilistic measures with non-numerical, but intuitive enough uncertainty (in the sense of fuzziness and vagueness) degrees – as the most simple structure for these purposes we have taken the classical Boolean algebra over the power-set of all subsets of a basic set $\Omega$ with sizes of elements of $\Omega$ and their collections quantified by subsets of another space $X$. The contents of particular sections as scheduled in the introductory one have been more or less tightly kept and it is why we do not take as necessary to repeat them now, rather focusing our attention to some inspirations for further developments.

## References

[1] Birkhoff G. (1967), Lattice Theory, 3rd edition. Providence, Rhode Island.

[2] DeCooman G. (1997), Possibility theory I, II, III. International Journal of General Systems 25, pp. 291-323, 325-351, 353-371.

[3] Faure R., Heurgon E. (1971), Structures Ordonnées et Algèbres de Boole. Gauthier-Villars, Paris.

[4] Fine T. L. (1973), Theories of Probability. An Examination of Foundations. Academic Press, New York–London.

[5] Goguen J. A. (1967), $\mathcal{L}$-fuzzy sets. Journal of Mathematical Analysis and Applications 18, pp. 145-174.

[6] Halmos P. R. (1950), Measure Theory. D. van Nostrand, New York.

[7] Kramosil I. (2006), Extensions of partial lattice-valued possibilistic measures from nested domains. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 14, pp. 175-197.

[8] Kramosil I., Daniel M. (2011), Statistical Estimations of Lattice-Valued Possibilistic Distributions, Proceedings of the 2011 ECSQARU Conference, Belfast, pp. 688-699.

[9] Kramosil I., Possibilistic distributions processed by probabilistic algorithms. Submitted for publication.

[10] Kramosil I. (2012), Some results on set-values possibilistic distributions, Technical report V-1162, Institute of Computer Science of the Czech Republic.

[11] Shannon C. E. (1948), The mathematical theory of communication. The Bell Systems Technical Journal 27, pp. 379-423, 623-656.

[12] Sikorski R. (1964), Boolean Algebras, 2nd edition. Springer, Berlin.

[13] Zadeh L. A. (1965), Fuzzy sets. Information and Control 8, pp. 338-353.

[14] Zadeh L. A. (1968), Probability measures of fuzzy events. Journal of Mathematical Analysis and Applications 23, pp. 421-427.

[15] Zadeh L. A. (1978), Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems 1, pp. 3-28.

# An Attempt to Implement Compositional Models in Dempster-Shafer Theory of Evidence

**Václav Kratochvíl**
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
&
Faculty of Management of University of Economics
Czech Republic
velorex@utia.cas.cz

**Abstract**

It has been published recently that some of the ideas for representation of multidimensional distributions in probability theory can be transferred into Dempster-Shafer theory of Evidence [7], [8]. Namely, they showed that multidimensional basic assignments can be rather efficiently represented in a form of so-called compositional models. These models are based on the iterative application of the operator of composition, whose definition for basic assignments has been introduced in [5]. It appears that a software tool supporting computations within compositional model is necessary for additional theoretical research in this framework. In this paper we will familiarize the reader with our first attempts and basic problems of the implementation itself.

## 1 Introduction

Plenty of applications of Artificial intelligence in the field of quantitative reasoning and decision under uncertainty is dominated by probabilistic models like Bayesian networks and their variants. It these models a multidimensional probability distribution is used to represent the real world problem and capture and represent uncertainty. We can distinguish two types of uncertainty. The first is variability that arises from environmental stochasticity, inhomogenity of materials, fluctuations in time, variation in space, or heterogenity or other differences among components or individuals. This variability is sometimes called *aleatory* uncertainty to emphasize its relation to the randomness in gambling and games of chance. The second kind of uncertainty is the incertitude that comes from scientific ignorance, measurement uncertainty, inobservability, censoring, or other lack of knowledge. This is sometimes called *epistemic* uncertainty.

For situations in which the uncertainty about quantities is purely aleatory, probability theory is usually preferred and it is sufficient for this purpose. When the gasp in our knowledge involve both aleatory and epistemic uncertainty, several competing approaches have been suggested: The common practice is to use probability theory as well. As another example we would like to mention *probability boxes* [21] and especially *Dempster-Shafer theory of Evidence* (D-S) [1] [15] which we will deal within this paper.

There is one problem when using probability framework to handle uncertainty. Assume that we have no information concerning behavior of a variable. Using probability theory, one might assume equal priors and distribute the weight of evidence equally among all possible states of the variable. But, as Shafer pointed out, here one will fail to distinguish between uncertainty (or lack of knowledge), and equal certainty. And it is this kind of uncertainty that can be easily captured in the framework of D-S theory.

In this paper we will deal with D-S theory, especially we will work with the notion of *Compositional models*. Compositional models were originally introduced in the probability framework. The intention was to create an algebraic alternative to the well-known Markov graphical models like Bayesian networks. The important advantage of compositional models is that they can be generalized in the framework of possibility theory as well as D-S theory by introducing a special operator of composition [8]. The recent research [7] [8] revealed the necessity of an software tool supporting compositional models in D-S theory.

The intention of this paper is nothing more than to summarize our initial problems when attempting to implement such a software tool. Here we describe our first steps, ideas and preliminary solutions.

# 2 Notation

For an index set $N = \{1, 2, \ldots, N\}$ let $\{X_i\}_{i \in N}$ be a finite set of finite valued variables, each $X_i$ having its values in $\mathbf{X}_i$. In this paper we deal with multidimensional frame of discernment $\mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_n$, and its subframes (for $K \subseteq N$) $\mathbf{X}_K = \times_{i \in K} \mathbf{X}_i$. The symbol $X_K$ will denote a group of variables $\{X_i\}_{i \in K}$. A projection of $x = (x_1, x_2, \ldots, x_n) \in \mathbf{X}_N$ into $\mathbf{X}_K$ will be denoted $x^{\downarrow K}$, i.e. for $K = \{i_1, i_2, \ldots, i_k\}$

$$x^{\downarrow K} = (x_{i_1}, x_{i_2}, \ldots, x_{i_k}) \in \mathbf{X}_K.$$

Analogously, for $M \subset K \subseteq N$ and $A \subset \mathbf{X}_K$, $A^{\downarrow M}$ will denote a *projection* of $A$ into $\mathbf{X}_M$:

$$A^{\downarrow M} = \{y \in \mathbf{X}_M | \exists x \in A : y = x^{\downarrow M}\}.$$

In addition to the projection, in this text we will need also an opposite operation, which will be called a *join*[1]. By a join of two sets $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_L$ $(K, L \subseteq N)$ we will understand a set

$$A \bowtie B = \{x \in \mathbf{X}_{K \cup L} : x^{\downarrow K} \in A \;\; \& \;\; x^{\downarrow L} \in B\}.$$

[1] This term and notation are taken form the theory of relational databases

Let use note that if $K$ and $L$ are disjoint, then $A \bowtie B = A \times B$, and if $K = L$ $A \bowtie B = A \cap B$.

The symbol $\mathcal{P}(\mathbf{X}_K)$ will denote the powerset of $\mathbf{X}_K$, i.e. the set of all subsets of $\mathbf{X}_K$.

## 2.1 Basic assignments

The role played by a probability distribution in probability theory is replaced by that of a set function in D-S theory: belief function, plausibility function, commonality function, or basic (*probability* or *belief*) assignment. Knowing one of them, one can derive the remaining three. In this paper we will use almost exclusively basic assignments.

If $m(A) > 0$, then $A$ is said to be a *focal element* of $m$. The set of focal elements will be denoted by $S$. A *basic assignment* (bpa) $m$ in $\mathbf{X}_K$ $(K \subseteq N)$ is a function

$$m : \mathcal{P}(\mathbf{X}_K) \to [0, 1],$$

for which

$$\sum_{\emptyset \neq A \subseteq \mathbf{X}_K} m(A) = 1.$$

The quantity $m(A)$ is a measure of that portion of the total belief committed exactly to $A$, where $A$ is an element of $\mathcal{P}(\mathbf{X}_K)$ and the total belief is 1. The portion of belief cannot be further subdivided among the subsets of $A$ and does not include portions of belief committed ti subsets of $A$. Since belief in a subset certainly entails belief in subsets, containing that subset, it would be useful to define a function that computes a total amount of belief in $A$. Such a function is called *belief function.*

On the contrary, *plausible function* characterizes the degree in which a proposal $A$ is plausible based on available evidence $B$ expressed by each basic assignment that contributes to realization of $A$. *Commonality function* doesnt have a simple interpretation but it allows a simple statement of Dempsters combination rule [1].

## 2.2 Operator of composition

Compositional models theory has been introduced in the framework of probability theory [6] as an algebraic alternative to well known and widely used Bayesian networks for efficient representations of multidimensional measures more than twelve years ago. Compositional models are based on recurrent application of an operator of composition. Later, the operator of composition was introduced also within the framework of D-S theory in [5]:

**Definition 2.1.** *For two arbitrary bpa $m_1$ on $\boldsymbol{X}_K$ and $m_2$ on $\boldsymbol{X}_L$ $(K, L \neq \emptyset)$, a composition $m_1 \triangleright m_2$ is defined for each $C \subseteq \boldsymbol{X}_{K \cup L}$ by one of the following expressions:*

*a) if $m_2^{\downarrow K \cap L}(C^{\downarrow K \cap L} > 0$ and $C = C^{\downarrow K} \bowtie C^{\downarrow L}$ then*

$$(m_1 \triangleright m_2)(C) = \frac{m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L})}{m_2^{\downarrow K \cap L}(C^{\downarrow K \cap L})}$$

*b) if* $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) = 0$ *and* $C = C^{\downarrow K} \times \boldsymbol{X}_{L\setminus K}$ *then*

$$(m_1 \triangleright m_2)(C) = m_1(C^{\downarrow K});$$

*c) in all other cases* $(m_1 \triangleright m_2)(C) = 0$.

In the D-S theory, there exists several way how to combine different sources of evidence and the above defined operator of composition seems to be one of them. But this is not the case. The classical way is represented by Dempster's combination rule [1]. A criticism of this rule appeared later caused by its behavior when combining two conflicting evidences and several additional combination rules were designed. Recall for example *Yager's rule* [23], *Inagakis rule* [4], *Zhangs rule* [25], or *Dubois and Prades Disjunctive Consensus* [2]. However, the intention of the operator of composition is not to be another combination rule and combine different sources of evidence. Its intention is completely different.

Despite the success of D-S theory of evidence as a well founded and general model of human reasoning under uncertainty, belief functions are rarely used in concrete applications. One of the most significant arguments raised against using belief functions in practice is their relatively high computational complexity, especially in comparison with methods based on classical probability theory. E.g. combining evidence using relatively simple Dempster's rule of combination is known to be #P-complete in the number of evidential sources. Recall that bpa (as well as belief function, plausibility function, and commonality function) is a set function. We work with the powerset of possible events and the number of sets that can be focal elements of a bpa can be superexponential within the number of involved variables.

To overcome these computational limitations, different approximation methods have been proposed. Previous work can be divided into two categories [3]. The first category consists of *Monte-Carlo* techniques [22]. The idea is to estimate exact values of belief and plausibility by ratios of different outcomes relative to randomly generated samples. The second category consist of simplification procedures. They are motivated by the fact that the most algorithms involving belief functions have a complexity polynomial in the number of focal elements. The underlying idea is therefore to restrict in different ways the number of focal elements. A simple method is called *Bayesian approximation* [20], where only singletons are allowed - which corresponds to the restriction on probability distributions only. Other methods like *k-l-x approximation* [19], *summarization* [10], and others try to reduce the number of focal elements by taking the first $k$-most important assignments. The sum of the omitted assignments is then redistributed in different ways depending on the respective method.

The idea of operator of composition goes in a different way: Practically all methods for efficient computations with multidimensional models take advantage of the fact that the model in question in a way factorizes. It means that it is possible to decompose the model into its low-dimensional parts, each of which can be defined with a reasonable number of parameters. This is the basic idea for computation with probabilistic Graphical Markov Models. Such a factorization not only decreases the storage requirements for representation of a multidimensional distribution but it usually also induces possibility to employ efficient computational procedures.

Since we need efficient methods for representation of probabilistic distributions, which require exponential number of parameters, the more we need of efficient methods for representation of an evidence, which cannot be represented by a point function. For such a representation we need a set function, and thus its space requirements are superexponential.

### 2.2.1 Compositional models

The factorizable evidence will be then represented in a form of the so-called *compositional model*. Assume a system of low-dimensional basic assignments $m_1, m_2, \ldots, m_n$ defined on $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}, \ldots, \mathbf{X}_{K_n}$, respectively. Composing them together by multiple application of operator of composition, one get multidimensional basic assignment on $\mathbf{X}_{K_1 \cup K_2 \cup \ldots \cup K_n}$. Note that the operator of composition is neither commutative, nor associative. By "composing them together" we understand that the operator of composition is performed successfully from left to right and $m_1 \triangleright_2 \triangleright \ldots \triangleright m_n = (\ldots((m_1 \triangleright m_2) \triangleright m_3) \triangleright \ldots) \triangleright m_n$.

### 2.2.2 New Concept of Conditional Independence

For belief functions, two type of factorization were designed in the literature. One is based on various combination rules mentioned above, the other use an operator of composition [5]. It has been shown in [7] that approach concerning Dempster's rule and the operator of compositions are equivalent each other in case of unconditional factorization.

The idea of factorization is closely related to the notion of (un)conditional independence in probabilistic modeling. However, as pointed out by Studený, the original definition of conditional independence (published in [24]) was not consistent with marginalization. That is why a new definition of conditional independence was introduced in D-S theory in [7]:

**Definition 2.2.** *Let $m$ be a basic assignment on $\boldsymbol{X}_N$ and $K, L, M \subset N$ be disjoint, $K, L \neq \emptyset$. We say that groups of variables $X_K$ and $X_L$ are conditionally independent given $X_M$ with respect to $m$ (and denote it by $K \perp\!\!\!\perp L | M[m]$), if the equality*

$$m^{\downarrow K \cup L \cup M}(A) \cdot m^{\downarrow M}(A^{\downarrow M}) = m^{\downarrow K \cup M}(A^{\downarrow K \cup M}) \cdot m^{\downarrow K \cup L}(A^{\downarrow K \cup L})$$

*holds for any $A \subseteq \boldsymbol{X}_{K \cup L \cup M}$ such that $A = A^{\downarrow K \cup M} \bowtie A^{\downarrow L \cup M}$, and $m(A) = 0$ otherwise. If $M = \emptyset$ then we say that groups of variables $X_K$ and $X_L$ are independent with respect to $m$ (in symbol $K \perp\!\!\!\perp L[m]$).*

Above that, it has been shown in [8] that the above defined conditional independence satisfies semigraphoid properties and that there is a link between operator of composition and conditional independence:

**Theorem 2.3.** *Let $m$ be a joint basic assignment on $\boldsymbol{X}_M$, $K, L \subseteq M$. Then $(K \setminus L) \perp\!\!\!\perp (L \setminus K) | (K \cap L)[m]$ iff $m^{\downarrow K \cup L}(A) = (m^{\downarrow K} \triangleright m^{\downarrow L})(A)$ for any $A \subseteq \boldsymbol{X}_{K \cup L}$.*

This theorem justifies the usage of the operator of composition when factorizing an evidence.

# 3 Implementation

To evaluate various hypotheses and support accelerate further theoretical research, it is necessary to create an experimental tool for calculations with compositional models in the framework of D-S theory. In this section we would like to describe several problems when attempting to implement such a tool. The tool itself is developed as an extension package for R-Project[2] and it is available at http://dar1.utia.cas.cz/mudim altogether with another tool supporting compositional models in the probability framework.

During our survey of existing implementation of D-S theory we found out that there is no successful universal tool supporting theoretical research. The majority of existing implementations is usually single purpose and base on restricted assumptions. One can find not very up-to-date, but exhausted overview of applications of D-S theory in [14].

The key problem of the implementation is the representation of belief structures. Restrict ourselves to finite sets. For a finite set $P$ of possible outcomes ($P \subseteq \mathbf{X}_N$) with cardinality $|P|$, there are at most $2^{|P|}$ unique basic probability assignments. We assume that rarely is a full set of $2^{|P|}$ unique bpa used in practice. It corresponds to the limited sources of information. Within the research literature there exists four common subclasses of bpa for finite sets [9]:

1. The trivial case of total ignorance where $m(P) = 1$ and $m(A) = 0$ iff $A \neq P$. This is highlighted as a more accurate representation of total ignorance when compare to traditional probability theory, which must apply Laplace's principle of indifference in these circumstances.

2. Every assignment is made to a singleton of the set $P$. This corresponds to a traditional probability measure on the set $P$.

3. Every assignment is made to a nested set. In other words, for every two sets $A$ and $B$ such that $m(A) > 0$ and $m(B) > 0$, then $A \subset B$ or $B \subset A$. This arrangement of is known as possibility theory.

4. Every assignment can be made to an arbitrary set.

In our case we focus on the most general case - the last one. However, we step aside the fact that involved variables can be either contiguous (discrete or continuous), or categorical. Each of these data types requires a special representation in a computer memory. In a survey of real-world application of D-S theory to infinite sets [16] it has been published that the contiguous frame of elements assigns basic probability statements to the closed intervals $[x_{i_a}, x_{i_b}]$ as a rule. Thus, in case of contiguous variables, we will store the interval boundaries. See the overview of various data types in Table 1.

[2] R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please visit http://www.r-project.org/

| *data type* | *example* | *implementation* |
|---|---|---|
| continuous finite data | age - integer | interval boundaries |
| cont. infinite data | sensor output - real number | interval boundaries |
| categorical data | sex {male, female} | set of elements |

Table 1: Implementation of various variable types

**Problems**

Let $A \subseteq \mathbf{X}_K$, $B \subseteq \mathbf{X}_L$, bpa $m$ on $\mathbf{X}_K$, and $S$ set of focal elements of $m$. The key problem of the implementation itself is the fact that we have to store every focal element $A \in S$ of $m$ and pair it with the value $m(A)$. This does not sound very difficult unless we realize that $S$ is a set of sets of vectors and that every set of vectors $A \in S$ is of various cardinality. The implementation of data structure will will have an enormous impact on overall system performance.

The most basic operation which will be instantly used is the checking whether a set $A$ is a focal element, i.e. whether $A \in S$. It is logical to assume that if the data structure will be optimized with respect to this operation, then the system performance allows to add additional functionalities like operator of composition etc.

There are multiple ways of implementing set (and map) functionality, that is:

- ordered (e.g. tree-based) approaches, and
- unordered (e.g. hash-based) approaches

Here we propose the unordered (hash-based) approach, which naturally builds on top of the value-indexed array technique. The problem here is that we have set of sets of vectors, which significantly complicates the implementation of respective hash function.

However, in this attempt, we simply store the evidence in multidimensional arrays (tables) of vectors and we implement the search of a set in a set of sets simply as a full table scan - i.e. the algorithm gradually passes through all elements of $S$ and compares them with set $A$. Such a comparison is described in Algorithm 1. $A, B$ are two sets of possible outputs. Note that we employ the definition $A = B \Leftrightarrow A \subseteq B \& B \subseteq A$. Then, in case of checking e.g. $A \subseteq B$ (Algorithm 2) we simply check whether $\exists b \in B$ such that $a = b$ for all $a \in A$. In the worst case scenario, the complexity is $2 \cdot |A| \cdot |B|$ of vector comparisons. The improvement of this will have an enormous impact on the efficiency of the tool. Our idea is either implement a specific hash function, or to use embedded relational database [17] with optimized index-based search algorithms.

**Algorithm 1** MySetEqual($A, B \subseteq \mathbf{X}_K$): boolean

```
if MySubset(A,B) and MySubset(B,A) then
    return TRUE;
else
    return FALSE;
end if
```

**Algorithm 2** MySubset($A, SubA \subseteq \mathbf{X}_K$): boolean

```
found: flag if the corresponding element is found in the other set
for $i = 1$ to $|SubA|$ do
    found=FALSE;
    for $j = 1$ to $|A|$ do
        if $SubA[i] == A[j]$ then
            found=TRUE; $\{SubA[i] \in A\}$
            break; {additional search is useless}
        end if
    end for
    if not found then
        return FALSE; $\{SubA[i] \notin A \Rightarrow SubA \nsubseteq A\}$
    end if
end for
return TRUE;
```

The other operations that have to be considered when designing a data structure are:

- marginalization $A^{\downarrow M} = \{y \in \mathbf{X}_M | \exists x \in A : y^{\downarrow M} = x\}$
- join operation $A \bowtie B = \{x \in \mathbf{X}_{K \cup L} : x^{\downarrow K} \in A \;\;\&\;\; x^{\downarrow L} \in B\}$

Using above defined functions one can implement the operation of composition specified in Definition 2.1. See Algorithm 3 for the pseudo-code of the implementation. Here two auxiliary boolean flags are employed - *found* and *marginalComputed*. The first one decides between cases $a)$ and $b)$ of Definition 2.1. The second one highlights whether the respective marginal $m_2^{\downarrow K \cap L}(C^{\downarrow K \cap L})$ from Definition 2.1 has been already computed or not.

The careful reader notices that the loop on lines 18-22 of the previous algorithm may be performed for the same set $C^{\downarrow K \cap L}$ several times. This could be easily improved. Let us define an auxiliary vector $m_2 marginal$ to store computed marginal of $m_2$ and index it in the same way as $m_2$. Then it is enough to use the value $m_2 marginal[k]$ on line 20 if $k < l$ and break respective cycle (lines 18-22).

# Conclusion

Recently, it has been published that some of the ideas for representation of multidimensional distributions in probability theory can be transferred into Dempster-Shafer theory of Evidence [7], [8]. Namely, they showed that multidimensional basic assignments can be rather efficiently represented in a form of so-called compositional models. However, only an application of the theory can show which parts still need to be improved. Our goal is to develop not only an interesting theory but also an efficient tool based on these theoretical results. In other words, we intend to create a software tool which could be used for experiments and additional theoretical research.

**Algorithm 3** $\rhd$ operation of composition: $m_3 = m_1 \rhd m_2$

```
input $S_1$: set of focal elements of $m_1, S_1 \subseteq \mathcal{P}(\mathbf{X}_K)$
input $m_1$: set of basic probability assignments $m_1[i] = m_1(S_1[i])$
input $S_2$: set of focal elements of $m_2, S_2 \subseteq \mathcal{P}(\mathbf{X}_L)$
input $m_2$: set of basic probability assignments $m_2[i] = m_2(S_2[i])$
output $S_3$: set of focal elements of $m_3 = m_1 \rhd m_2, S_3 \subseteq \mathcal{P}(\mathbf{X}_{K \cup L})$
output $m_3$: set of basic probability assignments $m_3 = m_1 \rhd m_2$
$l = 1$;
$S_3 = \emptyset$;
$m_3 = \emptyset$;
for $i = 1$ to $|S_1|$ do
  marginalComputed = FALSE;
  found = FALSE;
  marginalValue = 0;
  for $j = 1$ to $|S_2|$ do
    if MySetEqual($(S_1[i])^{\downarrow K \cap L}, (S_2[i])^{\downarrow K \cap L}$) then
      found = TRUE; {i.e. $m_2((S_2[i])^{\downarrow K \cap L}) > 0$}
      if not marginalComputed then
        for $k = 1$ to $|S_2|$ do
          if MySetEqual($S_2[k], S_2[j]$) then
            marginalValue = marginalValue + $m_2[k]$;
          end if
        end for
        marginalComputed = TRUE;
      end if
      $S_3[l] = S_1[i] \bowtie S_2[j]$; {case $a$ of the Definition 2.1}
      $m_3[l] = (m_1[i] \cdot m_2[j])$/marginalValue;
      $l = l + 1$;
    end if
  end for
  if not found then
    $S_3[l] = S_1[i] \times \mathbf{X}_{L \setminus K}$; {case $b$ of the Definition 2.1}
    $m_3[l] = m_1[i]$;
    $l = l + 1$;
  end if
end for
return $m_3, S_3$;
```

In this paper we have described our problems when implementing compositional models in the framework of Dempster-Shafer theory of Evidence. The tool is implemented as an extension package for R-Project and one can find it, altogether with another tool supporting compositional models in probability framework, at the website `http://dar1.utia.cas.cz/mudim`. In this paper we described our first steps and basic problems which we faced during implementation.

The paper contains just a preliminary ideas and gives answers only to very simple questions. So there are many more that remain to be answered. For example:

- Does it exist an efficient representation of sets of vectors?
- How does an effective hash function for a set of sets of vectors look like?
- Can be an embedded SQL database used for representation of focal elements?
- Let $\mathbf{X}_2 = \{a_2, \bar{a}_2\}$. Is it reasonable to combine two elements $(a_1, a_2)$, $(a_1, \bar{a}_2)$ into $(a_1, \mathbf{X}_2)$ and store the information in this way?

# Acknowledgement

This work was supported by institutional research support of Faculty of Management, University of Economics, Prague.

# References

[1] A.P. Dempster: *Upper and lower probabilities induced by a multi-valued mapping.* Annals of Mathematical Statistics 38, 325-339 (1967)

[2] D. Dubois and H. Prade: *On the combination of evidence in various mathematical frameworks.* Reliability Data Collection and Analysis. J. Flamm and T. Luisi. Brussels, ECSC, EEC, EAFC: 213-241 (1992)

[3] R. Haenni and N. Lehmann: *Resource bounded and anytime approximation of belief function computations*, International Journal of Approximate Reasoning (2002)

[4] T. Inagaki: *Interdependence between Safety-Control Policy and Multiple-Sensor Schemes Via Dempster-Shafer Theory.* IEEE Transactions on Reliability 40(2): 182-188 (1991)

[5] R. Jiroušek, J. Vejnarová, M. Daniel: *Compositional models for belief functions* in Proc. of 5th Int. Symposium on Imprecise Probability: Theories and Applications, Eds: De Cooman G., Vejnarová J., Zaffalon M., Praha(2007)

[6] R. Jiroušek: *Foundations of compositional model theory*, International Journal of General Systems - Volume 40, Issue 6, (2011), pp. 623-678.

[7] R. Jiroušek and P.P. Shenoy: *A Note on Factorization of Belief Functions*, Proceedings of 14th Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty CJS 2011 , Eds: Barták Roman (2011)

[8] R. Jiroušek, J. Vejnarová: *Compositional models and conditional independence in evidence theory*, International Journal of Approximate Reasoning vol.52, 3 (2011), p. 316-334 (2011)

[9] I. Kramosil: *Probabilistic Analysis of Belief Functions.* Kluwer Academic/Plenum Publishers, New York, (2001)

[10] H. Lowrance, T. Garvey, T. Strat: *A framework for evidential-reasoning systems* T. Kehler, S. Rosenschein (Eds.), Proceedings of the 5th National Conference on Artificial Intelligence, vol. 2Morgan Kaufmann, CA (1986), pp. 896903

[11] J. Pearl: *Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference*, Margan Kaufmann, San Mateo, CA (1988)

[12] J. Pearl: *Reasoning with Belief Functions: An Analysis of Compatibility.* The International Journal of Approximate Reasoning 4 (5/6): 363389 (1990)

[13] R Development Core Team: *R - A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org (2008)

[14] K. Sentz and S. Ferson: *Combination of Evidence in DempsterShafer Theory*, Sandia National Laboratories SAND 2002-0835 (2002)

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

[16] M. Spiegel: *A proposal for computing with imprecise probabilities: A framework for multiple representations of uncertainty in simulation software* A Disertation Proposal, University of Virginia (2007)

[17] *SQL*. In Wikipedia. Retrieved July 12, 2012, from
http://en.wikipedia.org/wiki/SQL

[18] M. Studený: *Probabilistic Conditional Independence Structures*, Springer, London, (series Information Science and Statistics), (2005).

[19] B. Tessem: *Approximations for efficient computation in the theory of evidence.* Artificial Intelligence, 61 (2) (1993), pp. 315329

[20] F. Voorbraak: A *computationally efficient approximation of DempsterShafer theory.* International Journal of Man-Machine Studies, 30 (5) (1989), pp. 525536

[21] R.C. Williamson and T. Downs. *Probabilistic arithmetic I: Numerical methods for calculating convolutions and dependency bounds.* International Journal of Approximate Reasoning 4: 89158 (1990)

[22] N. Wilson, S. Moral: *Fast Markov chain algorithms for calculating Dempster-Shafer belief*, in: ECAI96; 12th European Conference on Artificial Intelligence, Wiley, pp. 672678 (1996)

[23] R. Yager: *On the Dempster-Shafer Framework and New Combination Rules.* Information Sciences 41: 93-137 (1987)

[24] B.B. Yaghlane, Ph. Smets, K. Mellouli: *Belief functions independence: II. The conditional case*, International Journal of Approximate Reasoning 31 (2002) 3175.

[25] L. Zhang: *Representation, independence, and combination of evidence in the Dempster-Shafer theory.* Advances in the Dempster-Shafer Theory of Evidence. R. R. Yager, J. Kacprzyk and M. Fedrizzi. New York, John Wiley & Sons, Inc.: 51-69 (1994)

# Mixing Marginals for Decision-Making Based on Marginal Problem

**Otakar Kříž**
Prague, Czech Republic
o.kriz@upcmail.cz

**Abstract**

Four different methods ( $SM, MK_1, MK_2, SK$) are used to order a set $AM$ of admissible marginals ascendingly with respect to wrong classifications. Second, an algorithm called "Enhancement" finds, by reduction, the best *knowledge base* $\mathcal{K}$ (i.e. set of marginals) that used as input for a decision-making algorithm $A$ results in least number of wrong classifications.Third, certain strategies to cope with possible inconsistence of marginals and data for which the testing takes place are suggested.

## 1 Introduction

The layout of the paper is the following one: After introducing **Basic notions** four **Ascending sequences** are defined and in **Mixing marginals**, it is shown how to combine their beginnings to suggest the best knowledge base $\mathcal{K}_0$. Then, an multipurpose **Algorithm Enhancement** tries to improve $\mathcal{K}_0$, by reduction, to get an optimized $\mathcal{K}_{opt}$. **Experimental results** describe some details about testing an algorithm $A$ on real data. **Empty symptom pattern** reminds of an interesting situation that has to be taken care of in decision making algorithm $A$. It may appear due to splitting the available data in learning file $\boldsymbol{L}$ and testing file $\boldsymbol{T}$.

## 2 Basic notions

Let $(\Omega, \mathcal{X}, P)$ be a probabilistic space
$\boldsymbol{\eta} = \boldsymbol{\xi_0}, \boldsymbol{\xi_1}, \boldsymbol{\xi_2}, \ldots \boldsymbol{\xi_n}$ be finite sets and
$\xi_r : (\Omega, \mathcal{X}, P) \longrightarrow (\boldsymbol{\xi}_r, 2^{\boldsymbol{\xi}_r})$ for $r = 0, 1, 2, \cdots n$
be measurable functions
The mutual behaviour of all random variables $\eta, , \xi_1, \xi_2 \cdots \xi_n$ is described by joint probability distribution $P_{\eta\,\xi_1\xi_2\ldots\xi_n}$.
Decision making can be interpreted as the diagnostic problem with the following formulation:
**Diagnostic problem** Find the diagnosis $d(s_1, s_2 \cdots s_n) \in \boldsymbol{\eta}$ that is the most probable

(according to the $P_{\eta\xi_1,\xi_2\cdots\xi_n}$ ) on the set
$\{\omega \in \Omega \mid \xi_1(\omega) = s_1 \;\& \,\xi_2(\omega) = s_2 \;\& \;\cdots \xi_n(\omega) = s_n\}$ for a given (i.e. observed) arbitrary combination $(s_1, s_2 \cdots s_n)$ of values of *symptom variables* from the cartesian product $\boldsymbol{\xi_1} \times \boldsymbol{\xi_2} \times \ldots \boldsymbol{\xi_n}$
If we wish to predict the values of diagnostic variable $\eta$, the conditional probability $P_{\eta|\xi_1\xi_2\ldots\xi_n}$ (derivable from $P_{\eta\,\xi_1\xi_2\ldots\xi_n}$) should be used instead.
The optimal decision (i.e. the value of diagnosis $d$ from $\boldsymbol{\eta}$ that should be selected if the values of symptom variables are $(s_1, s_2 \cdots s_n)$ to keep the wrong classification of $d$ as low as possible), called Bayes solution, is given by the formula

$$d_{opt}(s_1, s_2 \cdots s_n) = \underset{d \in \boldsymbol{\eta}}{\text{argmax}}\; P_{\eta\,|\xi_1\xi_2\ldots\xi_n}(d|s_1, s_2 \cdots s_n) \qquad (1)$$

for each $(s_1, s_2 \cdots s_n) \in \boldsymbol{\xi_1} \times \boldsymbol{\xi_2} \times \ldots \boldsymbol{\xi_n}$

So far the theory. Unfortunately, in the "real world", we are never given the theoretical distribution $P_{\eta\xi_1\xi_2\cdots\xi_n}$ in full and directly. To compensate for this, we expect to have some indirect information about $P_{\eta\xi_1\xi_2\cdots\xi_n}$ that will be called *knowledge base* and denoted by $\mathcal{K}$. It is done by postulating a set of conditions that we believe the theoretical $P_{\eta\xi_1\xi_2\cdots\xi_n}$ fulfills. Using the concept of *marginal problem*, see [1], *knowledge base* $\mathcal{K}$ is given as a set of "low-dimensional" distributions ( e.g. number of variables in the distribution does not exceed e.g. 10. ), postulated as theoretical *marginal distributions* of the $P_{\eta\xi_1,\xi_2\cdots\xi_n}$. Instead of the unknown $P_{\eta\xi_1\xi_2\cdots\xi_n}$, we try to construct its suitable approximation $\hat{P}_{\eta\xi_1\xi_2\cdots\xi_n}$ that could play its role in the diagnostic problem.
If we want to speak about decision errors, three notions should be defined: *statistical file* $\boldsymbol{F}$, *decision-making algorithm* $A$ and *objective functional* $M$, measuring the classification errors. At the same time, it should be explained how the marginals are obtained.

Let $(\omega_1, \omega_2, \cdots \omega_s)$ be a sequence, where individual $\omega_i \in \Omega$ denote realizations of a random selection from $\Omega$,
then the sequence $(\eta(\omega_l), \xi_1(\omega_l), \xi_2(\omega_l) \cdots \xi_n(\omega_l))_{l=1}^{s}$ of points in cartesian product $\boldsymbol{\eta} \times \boldsymbol{\xi_1} \times \boldsymbol{\xi_2} \times \ldots \boldsymbol{\xi_n}$ is a *statistical file* $\boldsymbol{F}$ of size $s$ (i.e. $s = |\boldsymbol{F}|$) and $(\boldsymbol{F})_r$ is the $r$-th member of $\boldsymbol{F}$.
The file $\boldsymbol{F}$ can be used for calculating the *empirical joint distribution* $P^{\boldsymbol{F}}_{\eta\,\xi_1\xi_2\ldots\xi_n}$ by the following formula

$$P^{\boldsymbol{F}}_{\eta\,\xi_1\xi_2\ldots\xi_n}(d, s_1, s_2 \cdots s_n) = |\{r \in N : (\boldsymbol{F})_r = (d, s_1, s_2 \cdots s_n)\}| \,/\, |\boldsymbol{F}|$$

If we denote by $\Xi$ the set of all symptom variables and by $2^{\Xi}$ its potential set, then the set $\mathcal{M}$ of all marginals carrying information about $\eta$ can be expressed as $\mathcal{M} = \{P_{\eta j} \mid j \in 2^{\Xi}\}$.
Then, in its turn, the set$\{\,\mathcal{K}_w\,\}_w$ of all knowledge bases $\mathcal{K}$ is potential set $2^{\mathcal{M}}$ of $\mathcal{M}$ i.e. each $\mathcal{K}$ can be expressed as $\mathcal{K} = \{m_1, m_2, \cdots m_r\} \subset 2^{\mathcal{M}}$. Further, let $\boldsymbol{\Xi}$ denote cartesian product of all symptom variables i.e. $\boldsymbol{\Xi} = \boldsymbol{\xi_1} \times \boldsymbol{\xi_2} \times \ldots \boldsymbol{\xi_n}$. Now, each

*decision-making algorithm* $A$ can be formally defined as

$$\begin{array}{rcl} A: 2^{\mathcal{M}} \times \Xi & \longrightarrow & \boldsymbol{\eta} \\ (\mathcal{K}, (s_1, s_2 \cdots s_n)) & \longmapsto & d \in \boldsymbol{\eta} \end{array}$$

The measuring *objective functional* $M$ evaluates the number of wrong classifications for each $A$ and for each statistical file $\boldsymbol{F}$. If we set $((\boldsymbol{F})_j)_\xi = (\xi_1(\omega_j), \xi_2(\omega_j) \cdots \xi_n(\omega_j))$ and $((\boldsymbol{F})_j)_1 = \eta(\omega_j)$, then we may describe the testing scheme as a mapping M

$$\begin{array}{rcl} M: \{A_i\}_i \times 2^{\mathcal{M}} \times \{\boldsymbol{F}_l\}_l & \longrightarrow & \mathcal{N} \\ (A_i, \mathcal{K}, \boldsymbol{F}) & \longmapsto & |\{j \in \mathcal{N} : A_i(\mathcal{K}, ((\boldsymbol{F})_j)_\xi) \neq ((\boldsymbol{F})_j)_1\}| \end{array}$$

The key question where do the marginals come from is considered as external to the *marginal problem* since their existence as input is just postulated. In theory, they could "be given" from an authoritative source (experts ?), but practically, they cannot be obtained otherwise but from a statistical file $\boldsymbol{F}$. There is a hidden supposition that while there is not enough data (i.e. size $|\boldsymbol{F}|$ of statistical file $\boldsymbol{F}$) to get the theoretical distribution $P_{\eta\xi_1\xi_2\cdots\xi_n}$, the same data is sufficient to get (populate) its marginals. Marginal problem formulation postulates that marginals of the approximation $\hat{P}$ are the same as marginals of the theoretical $P$ distribution since this is the way algorithms $A$ are constructed. There is a bit suspicious that there is not a quantification of low-dimensional marginals but if the given marginals are populated from data (i.e. statistical file $\boldsymbol{F}$), they are also marginals of the empirical distribution $P^{\boldsymbol{F}}$ so that for the given marginals the following equations hold

$$\hat{P}_{\eta\xi_{j_1}\xi_{j_2}\cdots\xi_{j_k}} = P_{\eta\xi_{j_1}\xi_{j_2}\cdots\xi_{j_k}} = P^{\boldsymbol{F}}_{\eta\xi_{j_1}\xi_{j_2}\cdots\xi_{j_k}}$$

## 3 Ascending sequences

In [3], there is described a procedure that generates, in automated way, several thousands admissible marginals $AM$ from $\boldsymbol{F}$. It seems useful to order them ascendingly with respect to the number of wrong classifications. In the sequel, there are suggested four ordering algorithms $(SM, MK_1, MK_2, SK)$ in form of recursive formulae. It is obvious that other strategies can be thought of, as well e.g. we could select marginals that are in the best triplets (i.e. minimizing $M(A, \{m^{,}, m^{,,}, m^{,,,}\}, T)$ etc. However, such strategies are not so easy to realize and not object of this study.

## 3.1 $\boldsymbol{SM}$ sequence

$$(v,w) = \underset{(m',m'')\in AM^2}{\operatorname{argmin}} \; M(A,\{m',m''\},T)$$

$$(\boldsymbol{SM})_1 = v$$

$$(\boldsymbol{SM})_2 = w$$

$$U_2 = \{v,w\}$$

$$u_l = \underset{m\,\in\,AM\setminus U_{l-1}}{\operatorname{argmin}} \; \sum_{m_s\,\in\,U_{l-1}} M(A,\{m,m_s\},T)$$

$$(\boldsymbol{SM})_l = u_l$$

$$U_l = U_{l-1} \cup u_l$$

where $l = 3,4,\cdots k$

## 3.2 $\boldsymbol{MK_1}$ sequence

Algorithm $MK_1$ generates the ascending sequence $\boldsymbol{MK_1}$ of prescribed length $s_{min}$. $U_l$ is a set of marginals so far selected at $l$-th step $U_l \subset AM$. $\boldsymbol{u}_l$ is a pair of marginals $(m_i,m_j) \in AM^2$ that was constructed at $l$-th step. It can be expressed alternatively as $\boldsymbol{u}_l = ((\boldsymbol{u}_l)_1,(\boldsymbol{u}_l)_2)$.

```
    procedure $MK_1(\boldsymbol{MK_1}, s_{min})$
       $U_0 = \emptyset$; $s = 0$; $l = 1$
again:
       $\boldsymbol{u}_l = \underset{(v,w)\,\in\,AM^2\colon\, v\,\in\,AM\setminus U_{l-1} \text{ or } w\,\in\,AM\setminus U_{l-1}}{\operatorname{argmin}} M(A,\{v,w\},T)$
       $U_l = U_{l-1} \cup \{(\boldsymbol{u}_l)_1\} \cup \{(\boldsymbol{u}_l)_2\}$
       if $(\boldsymbol{u}_l)_1 \in AM \setminus U_{l-1}$ then
              $s = s+1;\ (\boldsymbol{MK_1})_s = (\boldsymbol{u}_l)_1$
              if $(\boldsymbol{u}_l)_2 \in AM \setminus U_{l-1}$ then
                  $s = s+1;\ (\boldsymbol{MK_1})_s = (\boldsymbol{u}_l)_2$
              endif
       else
              $s = s+1;\ (\boldsymbol{MK_1})_s = (\boldsymbol{u}_l)_2$
       endif
       if $s < s_{min}$ then
```

```
            $l = l + 1;$ goto again
        endif
    end MK1
```

## 3.3 $\boldsymbol{MK_2}$ sequence

Let $U$ be an ordering on set of admissible marginals $AM$ i.e.

$$U \subset AM^2 : \quad (m', m'') \in U \implies (m'', m') \notin U$$
$$(m', m'') \in U ; (m'', m''') \in U \implies (m', m''') \in U$$

Let $U_l$ denote a set of pairs of marginals. $\boldsymbol{u}_l$ denote a pair of marginals where $(\ \boldsymbol{u}_l)_1$ is the first marginal of the pair and $(\ \boldsymbol{u}_l)_2$ is the second marginal of the pair so that $\boldsymbol{u}_l = ((\ \boldsymbol{u}_l)_1, (\ \boldsymbol{u}_l)_2)$. Further, let $V_l$ denote a set of marginals, $\boldsymbol{MK}_2$ be a sequence (vector,array) of marginals from $AM$ and $(\boldsymbol{MK}_2)_k$ denote its $k$-th member.

```
procedure $MK_2$
  $U_1 = \emptyset ; \{u_{-1}\} = \emptyset ; V_{-1} = \emptyset ; \{(\ \boldsymbol{u}_{-1})_1\} = \emptyset ; \{(\ \boldsymbol{u}_{-1})_2\} = \emptyset$
    for $l = 0, 1, 2, \cdots \min(|AM|, 24)$
      $U_l = U_{l-1} \cup \{u_{l-1}\}$
      $V_l = V_{l-1} \cup \{(\ \boldsymbol{u}_{l-1})_1\} \cup \{(\ \boldsymbol{u}_{l-1})_2\}$

          $u_l = \underset{(m', m'') \in U \setminus U_l \,:\, m', m'' \notin V_l}{\operatorname{argmin}} M(A, \{m', m''\}, T)$

      $(\boldsymbol{MK}_2)_{2l+1} = (\ \boldsymbol{u}_l)_1$
      $(\boldsymbol{MK}_2)_{2l+2} = (\ \boldsymbol{u}_l)_2$
    next $l$
end $MK_2$
```

Then, the sequence $\boldsymbol{MK}_2$ contains $2 * \min(25, |AM|)$ members. Complexity of $MK_2$ can be estimated as $O(50 * |AM|^2)$.

## 3.4 $\boldsymbol{SK}$ sequence

The ascending sequence $\boldsymbol{SK}$, of length $k$, is ordered according to wrong classifications based on individual marginals from $AM$.

$$U_0 = \emptyset$$

$$(\boldsymbol{SK})_l = \underset{m \,\in\, AM \setminus U_{l-1}}{\operatorname{argmin}} M(A, \{m\}, T)$$

$$U_l = U_{l-1} \ \cup\ \{(\boldsymbol{SK})_1\}$$

where $l = 1, 2, \cdots k$

# 4 Mixing marginals

The best approximation $\hat{P}_{\eta\xi_1\xi_2\cdots\xi_n}$ of the theoretical $P_{\eta\xi_1\xi_2\cdots\xi_n}$ that would yield the least number of wrong classification in the diagnostic problem is dependent on two factors. First, it is the algorithm $A$ (decision-making engine) integrating marginals to get $\hat{P}_{\eta\xi_1\xi_2\cdots\xi_n}$ and second, it is a set a set of concrete marginals referred to as knowledge base $\mathcal{K}$. The concrete marginals $m \in \mathcal{K}$ are, in its turn, defined by their carrier $\underline{m}$ (i.e. variables whose behaviour is described by the marginal $m$) and second, they depend on a statistical file $\boldsymbol{F}$ that is used for their populating. In classical marginal problem [1], the set of marginals from which the joint distribution $P_{\eta\xi_1\xi_2\ldots\xi_n}$ is to be "integrated", is given and fixed ! Leaving aside what different algorithms can be used to do it (e.g. Bayes nets or see [2]), one may ask whether selection of a different (i.e. better) set of marginals could improve the decision-making. In fact, what is explicitly given are not the marginals $m$ in $\mathcal{K}$ but observed data (i.e. statistical file $\boldsymbol{F}$ ) and what is left to our free choice is the selection of the carriers $\underline{m}$. The best situation would be if this process could be fully automated. An attempt in this direction was done in [3] by suggesting an $\mathcal{SM}$ algorithm where certain "orthogonality" of marginals is respected. It is based on behaviour of pairs of marginals and it generates a sequence of at most 50 marginals in descending order where the "best" marginals should stand at the beginning. Therefore, it seems as a good strategy to select the beginning of the sequence and hope that it represents the optimal set of marginals. (It still remains an open question how far to go in the sequence !) To test if the assumption holds would require to inspect all subsets of the beginning. For 20 marginals, to generate and test all $2^{20}$ subsets is unrealistic. Therefore, an heuristical algorithm "Enhancement" is suggested in section 5. There are some taciturn assumptions about marginals and about knowledge base $\mathcal{K}$. Namely, each marginal $m$ from $AM$ should have diagnostic variable $\eta$ in its carrier $\underline{m}$ and second, none algorithm $A$ can operate on too many marginals. It would result in time increase when evaluating individual decision and it is not certain that the more marginals means the better for decision making. In the meantime, three other algorithms, $MK_1$, $MK_2$ and $SK$, using different optimality criterion, were suggested and "Enhancement" can be applied to their outputs, too. At this moment, it seems natural to prolong the procedure even further. Namely, to apply "Enhancement" not only to outputs of "pure type" algorithms ($\mathcal{SM}$, $MK_1$, $MK_2$ and $SK$), but also to their mixtures, what justifies the title of the paper. E.g. the clause "SM(5)+SK(3)+MK2(10)+ EK(5)" of the control language preparing input for "Enhancement" uses as $\mathcal{K}_0$ first five marginals from the sequence $SM$, first three marginals from $SK$, ten marginals from $MK_2$ and "EK(5)" stands for five marginals whose structure is provided explicitly by experts. In case if some of the first three marginals from $SK$ have appeared already among the first five from $SM$, they are either skipped or it is possible to force out further marginals from sequence $SK$ to have in the mix three contributions from $\mathcal{SK}$ but, this time, not necessarily the first three ones. (In strategy with skipping, the order in the clause does not matter for subsequent processing by Enhancement!).

To sum it up, prior to releasing the final version of the knowledge base $\mathcal{K}$, certain time should be devoted to playing with different mixtures of marginals and their subsequent reduction with Enhancement. That is the semi-automated way to adapt $\mathcal{K}$ to data (i.e. statistical file $\boldsymbol{F}$) for each specific problem area.

# 5 Algorithm Enhancement

The description of the algorithm Enhancement starts with explaining its basic idea. Then, the denotation of used objects and constructions is presented and finally, the algorithm is described in form of a procedure written in a symbolic language.

## 5.1 Basic idea

Let $\mathcal{K}_0$ be a finite set of marginals and $r \in N : r <= |\mathcal{K}_0|$. $V(\mathcal{K}_0, r)$ will denote r-neighbourhood of $\mathcal{K}_0$ so that $V(\mathcal{K}_0, r) = \{\mathcal{K} \subseteq \mathcal{K}_0 \,|\, |\mathcal{K}_0| - |\mathcal{K}| <= r\}$.We can look for the best set $\mathcal{K}_{opt}$ starting from $\mathcal{K}_0$. In other words, we construct recursively a sequence $(\mathcal{K}_l)_{l=1}^{s}$.

$$\mathcal{K}_{l+1} = \underset{\mathcal{K} \in V(\mathcal{K}_l, r)}{\text{argmin}} \quad M(A, \mathcal{K}, T) \qquad l = 0, 1, \cdots s \tag{2}$$

where length $s$ of the sequence $(\mathcal{K}_l)_{l=1}^{s}$ of sets $\mathcal{K}_l$ of marginals is given by stopping condition

$$s \in \{0\} \cup N : M(A, \mathcal{K}_{s+1}, T) \geq M(A, \mathcal{K}_s, T)$$

.
At this moment, we might finish the algorithm by setting $\mathcal{K}_{opt} = \mathcal{K}_s$.
However, in principle, a subset of $\mathcal{K}_0$ better than $\mathcal{K}_s$ may still exist. Therefore, we may reverse the procedure and instead of reducing $\mathcal{K}_0$, we may look for optimum by expanding $\mathcal{K}_s$ up to $\mathcal{K}_0$.
Let $U(\mathcal{K}_s, \mathcal{K}_0, r) = \{\mathcal{K} \subseteq \mathcal{K}_0 \,|\, |\mathcal{K}| - |\mathcal{K}_s| <= r\}$ be a class of sets of marginals that are generated from $\mathcal{K}_s$ by adding at most $r$ marginals from $\mathcal{K}_0 \setminus \mathcal{K}_s$. If no improvement (i.e. minimization) of $M(A, \mathcal{K}_s, T)$ on $U(\mathcal{K}_s, \mathcal{K}_0, r)$ can be found, we have finished setting $\mathcal{K}_{opt} = \mathcal{K}_s$. Otherwise, we proceed the expansion creating a sequence $\mathcal{K}_s \subset \mathcal{K}_{s+1}, \cdots \subset \mathcal{K}_{s+w}$ using the recursive procedure

$$\mathcal{K}_{s+l+1} = \underset{\mathcal{K} \in U(\mathcal{K}_{s+l}, \mathcal{K}_0, r)}{\text{argmin}} \quad M(A, \mathcal{K}, T) \qquad l = 0, 1, \cdots w \tag{3}$$

where length $w$ of the sequence $(\mathcal{K}_l)_{l=1}^{w}$ of sets $\mathcal{K}_l$ of marginals is given by stopping condition

$$w \in \{0\} \cup N : M(A, \mathcal{K}_{s+w+1}, T) \geq M(A, \mathcal{K}_{s+w}, T),$$

and at this moment, we recur to reduction again. But this time, we start not from $\mathcal{K}_0$ but from $\mathcal{K}_{s+w}$ i.e. we set $\mathcal{K}_j = \mathcal{K}_{s+w}$ and enter the reduction cycle.
This changing from reduction to expansion may happen several times and though it should finish in finite number of steps due to finiteness of $\mathcal{K}_0$, we may explicitely limit it to say 10 changes by counting the changes in a variable *count*.

## 5.2 Symbolic description

To shorten the description of the algorithm "Enhancement", following symbols are used: $\overline{a}$, $\underline{b}$, $|K|$, $comb(c,d)$. Let $\mathcal{K} = \{m_1, m_2 \cdots m_s\}$ be a set of $s$ marginals, then $K = \underline{\mathcal{K}} = \{1, 2 \cdots s\}$ is a set of integers denoting indices of marginals from the set $\mathcal{K}$. Inverse mapping (from indices to marginals) is denoted by upper bar over the set of integers. E.g. $\overline{\{5, 8, 15\}} = \{m_5, m_8, m_{15}\}$. The symbol $|\,.\,|$ stands for number of elements of its argument. E.g. $|\,K\,| = |\,\mathcal{K}\,| = s$ and $|\,m_5\,| = |\,\underline{m_5}\,| = 4$.
$|\xi_j|$ denotes number of elements of the set $\boldsymbol{\xi_j}$ the random variable $\xi_j$ takes its value from. Curly parentheses denote a set of elements separated by commas. Bold symbol stands for a sequence. If $\mathbf{e}$ is a sequence, $(\mathbf{e})_j$ is its $l$-th member that can be, in its turn, e.g. a set.
If $\mathbf{I} \subset \mathbf{N} : |\,\mathbf{I}\,| < \infty$ and $i \in \mathbf{N} : i \leq |\,\mathbf{I}\,|$, then function $comb(\mathbf{I}, i)$ returns a sequence of combinations of elements of the set $\mathbf{I}$.
More formally, it is a sequence of length $\binom{|\,\mathbf{I}\,|}{i}$ whose members are sets of $i$-tuples from $\mathbf{I}$.
Let us illustrate some constructions in the following example:
$\mathcal{K}_0 = \{m_1, m_2, m_3, m_4\}$ ; $K_0 = \underline{\mathcal{K}_0} = \{1, 2, 3, 4\}$ ; $|\,K_0\,| = 4$
For $i = 2$, $\boldsymbol{e} = \text{comb}(\underline{\mathcal{K}_0}, 2) = (\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\})$
$\boldsymbol{f} = (\,\mathcal{K}_0 \setminus (\,\boldsymbol{e})_l\,)_{l=1}^{6} = (\{3, 4\}, \{2, 4\}, \{2, 3\}, \{1, 4\}, \{1, 3\}, \{1, 2\})$
For $i = 3$, $(\,\boldsymbol{f})_i = \{2, 3\}$ and $\overline{(\,\boldsymbol{f})_i} = \{m_2, m_3\}$
Finally, the algorithm $A$ operating on the set of marginals $\overline{(\,\mathbf{f})_\mathbf{i}}$ yields $M(A, \{m_2, m_3\}, T)$ wrong classifications for $|\,T\,|$ cases from the testing statistical file $\boldsymbol{T}$ .

```
Algorithm Enhancement ($\mathcal{K}_{opt}, M_{opt}$)
read $\mathcal{K}_0$, read $r$
if  $r > |\mathcal{K}_0|$, then $r = |\mathcal{K}_0|$
$j = 0$;   $M_{min} = |\,F\,|$;   $M_{last} = |\,F\,|$;   $count = 0$
NewMinus:
    for $i = 1,\ \min(r, |\mathcal{K}_j|)$
        $\boldsymbol{e} = \text{comb}(\underline{\mathcal{K}_j}, i)$
        $\boldsymbol{f} = (\,\mathcal{K}_0 \setminus (\,\boldsymbol{e})_l\,)_{l=1}^{|\boldsymbol{e}|}$
            for $l = 1, |\boldsymbol{f}|$
                if $M(A, \overline{(\boldsymbol{f})_l}, T) < M_{min}$ then
                 $M_{min} = M(A, \overline{(\boldsymbol{f})_l}, T)$; $K_{min} = (\boldsymbol{f})_l$
                endif
            next $l$
    next $i$
            if $M_{min} < M_{last}$ then
             $j = j + 1$;  $M_{last} = M_{min}$;  $\mathcal{K}_j = \overline{K_{min}}$;  goto NewMinus
            endif
   $M_{last} = M_{min}$ ; $w = 0$
NewPlus:
   $K_{dif} = \underline{\mathcal{K}_0} \setminus K_{min}$
    for $i = 1,\ \min(r, |K_{dif}|)$
```

```
            $\boldsymbol{g} = \mathrm{comb}(K_{dif}, i)$
            $\boldsymbol{h} = (\ K_{min} \cup (\ \boldsymbol{g})_l\ )_{l=1}^{|\boldsymbol{g}|}$
                   for $l = 1, |\boldsymbol{g}|$
                       if $M(A, \overline{(\boldsymbol{h})_l}, T) < M_{min}$ then
                       $M_{min} = M(A, \overline{(\boldsymbol{h})_l}, T)$; $K_{min} = (\boldsymbol{h})_l$
                       endif
                   next $l$
        next $i$
        if $M_{min} < M_{last}$ then
                       if $w = 0$  goto RegularEnd
                     count = count+1
                   if $count > 10$ goto IrregularEnd
                  $j = j + 1$ ; $\mathcal{K}_j = \overline{K_{min}}$ ; goto NewMinus
        else
                  $w = w + 1$ ; $M_{last} = M_{min}$
                  $j = 0$ ; $\mathcal{K}_j = \overline{K_{min}}$ ; goto NewMinus
        endif
    IrregularEnd:
            Print("After ",count," iterations, Enhancement abnormally ended")
            goto RegularEnd
    RegularEnd:
            Print("$K_{min} =$ ", $K_{min}$, ", $M_{min} =$ ", $M_{min}$);
            $\mathcal{K}_{opt} = \overline{K_{min}}$ ; $M_{opt} = M_{min}$;
    End Enhancement
```

# 6 Experimental results

Experiments with marginals (selecting, mixing) were performed on a statistical file $\boldsymbol{F}$ with 4 diagnoses (i.e. $|\boldsymbol{\eta}| = 4$) and 34 other symptom variables whose ranges have cardinalities from 2 to 9. The file $\boldsymbol{F}$ with 1089 objects was split (i.e. by random selection) in 32 complementary subfiles $\boldsymbol{L}$ and $\boldsymbol{T}$. E.g. a generating control pattern 11212 splits $\boldsymbol{F}$ into 654 cases in $\boldsymbol{L}$ to populate marginals in $\mathcal{K}$ and 435 cases in $\boldsymbol{T}$ used for testing. Improvement achieved by Enhancement was about several percentage points. First few marginals from different ascending sequences seem to coincide and none of sequences ($\boldsymbol{SM}, \boldsymbol{MK}_1, \boldsymbol{MK}_2, \boldsymbol{SK}$ ) seems to dominate the others. The results (i.e. marginals selected to $\mathcal{K}_{opt}$ and the corresponding number of achieved wrong classification $M(A, \mathcal{K}_{opt}, T)$ are dependent on the type of algorithm $A$ and on the splitting into $\boldsymbol{L}$ and $\boldsymbol{T}$

# 7 Empty Symptom Pattern

Let us suppose that for an object $\omega \in \Omega$, following symptom values were observed $\xi_i(\omega) = s_i \in \boldsymbol{\xi_i}$ for $i = 1, 2, \cdots n$. These values are passed to the decision-making algorithm A. The algorithm has, as its knowledge base $\mathcal{K}$, a set of marginals populated from a learning statistical file $\boldsymbol{L}$. Let us suppose there is a marginal $m = P_{\eta\xi_{i_1}\xi_{i_1}\cdots\xi_{i_k}}$ that for its submarginal $m^{\eta}$ holds $m^{\eta}(s_{i_1}, s_{i_2}, \cdots s_{i_k}) = P_{\xi_{i_1}\xi_{i_1}\cdots\xi_{i_k}}(s_{i_1}, s_{i_2}, \cdots s_{i_k}) = 0$. This means we cannot select any $d \in \boldsymbol{\eta}$ since there is no previous information about $\eta$ from the conditioning set $\{\, \omega \in \Omega \mid \xi_{i_1}(\omega) = s_{i_1} \,\&\, \xi_{i_2}(\omega) = s_{i_2} \,\&\, \cdots \xi_{i_k}(\omega) = s_{i_k} \,\}$. However, we may expect that if the conditioning set is increased, by reducing the number of conditions, an evidence about $\eta$ may appear. This increasing will be done, in systematic way, by constructing submarginals of $m$ by excluding single symptom variables, then pairs of them, then triplets of them etc. The process is stopped if an submarginal of $m$ yields a non zero value for respective arguments from $(s_{i_1}, s_{i_2}, \cdots s_{i_k})$. If there are more such submarginals with the same number of excluded variables, the one with minimal entropy (i.e. $H(.)$) with respect to $\eta$ will be selected.

Let $U$ be a finite set, $l \in N : l \le |U|$, then $\boldsymbol{B}(U, l)$ denotes a sequence whose members are $l$-tuples of elements from $U$. Further, let $m = P_{\eta\xi_{i_1}\xi_{i_1}\cdots\xi_{i_k}}$ be a marginal describing behaviour of variables $\eta\xi_{i_1}\xi_{i_1}\cdots\xi_{i_k}$ what can be alternatively expressed as $\underline{m} = \eta\xi_{i_1}\xi_{i_1}\cdots\xi_{i_k}$. Symbol $\underline{m} \setminus \eta$ denotes a set of symptom variables of $m$ and $\boldsymbol{B}(\underline{m} \setminus \eta, l)$ is a sequence of all $l$-tuples created from symptom variables described by $m$. The symbol $m^{(\boldsymbol{B}(\underline{m}\setminus\eta,l))_s}$ is the submarginal of the marginal $m$ created by summing over $s$-th $l$-tuple from symptom variables of $m$. The remaining (active) variables of this submarginal are $\underline{m} \setminus (\boldsymbol{B}(\underline{m} \setminus \eta, l))_s$.

As we need to express formally values of the submarginal, we may use the following construction $(d, s_{i_1}, s_{i_2}, \cdots s_{i_k})^{\underline{m}\setminus(\boldsymbol{B}(\underline{m}\setminus\eta,l))_w}$ to describe arguments that go over from the original vector $(d, s_{i_1}, s_{i_2}, \cdots s_{i_k}) \in \boldsymbol{\eta} \times \boldsymbol{\xi_{i_1}} \times \boldsymbol{\xi_{i_2}} \times \ldots \boldsymbol{\xi_{i_k}}$ as corresponding arguments of the submarginal $m^{(\boldsymbol{B}(\underline{m}\setminus\eta,l))_w}$ so that the number $P_{\eta\xi_{i_1}\xi_{i_2}\cdots\xi_{i_k}}(d, s_{i_1}, s_{i_2}, \cdots s_{i_k})$ is transformed to the number

$$m^{(\boldsymbol{B}(\underline{m}\setminus\eta,l))_w}(d, s_{i_1}, s_{i_2}, \cdots s_{i_k})^{\underline{m}\setminus(\boldsymbol{B}(\underline{m}\setminus\eta,l))_w}$$

.

Similar construction $m^{\eta(\boldsymbol{B}(\underline{m}\setminus\eta,l))_w}(d, s_{i_1}, s_{i_2}, \cdots s_{i_k})^{\underline{m}\setminus(\eta\cup(\boldsymbol{B}(\underline{m}\setminus\eta,l))_w)}$
describes the probability $P(\{\omega \in \Omega \,|\, \&_{j \in T}\ \xi_{i_j}(\omega) = s_{i_j}\})$
where $T = \{j \in N \,|\, \xi_{i_j} \in \{\underline{m} \setminus \{\eta \cup (\boldsymbol{B}(\underline{m} \setminus \eta, l))_w\}\}$

Correction of algorithm $A$ to cope with "Empty Symptom Pattern" consists in the following procedure: Instead of using "common" knowledge base $\mathcal{K}$ selected according to mixing marginals strategy described in sections 3, 4 and 5, its modification $\mathcal{K}'$ is submitted to the $A$. If $\boldsymbol{L} = \boldsymbol{T} = \boldsymbol{F}$, then $\mathcal{K}' = \mathcal{K}$.

However, this modification guarantees a decision even if symptom values combination $(s_1, s_2, \cdots s_n)$ was not found in the file $\boldsymbol{F}$.

```
$\mathcal{K}' = \emptyset$
for $m \in \mathcal{K}$
    if $m(s_1, s_2, \cdots s_n)^{\underline{m}\setminus\eta} \neq 0$ then
        $\mathcal{K}' = \mathcal{K}' \cup \{m\}$
    else
        $q_{min} = \emptyset$
        $x_{min} = |\boldsymbol{\eta}|$
        $k = |\underline{m}| - 1$
        for $l = 1, k$
          for $w = 1, \binom{k}{l}$
            $q = (\boldsymbol{B}(\underline{m} \setminus \eta, l))_w$
            $t = \underline{m} \setminus q$
            $r = t \setminus \eta$
            $h_0 = m^q(s_{i_1}, s_{i_2}, \cdots s_{i_k})^r$
              for $h_0 \ = 0$ then
              else
                for $d \in \boldsymbol{\eta}$
                    $h_\eta(d) = m^q(d, s_{i_1}, s_{i_2}, \cdots s_{i_k})^t / h_0$
                next $d$
              endif
              if $x_{min} > H(h_\eta)$ then
                $x_{min} = H(h_\eta)$
                $q_{min} = q$
              endif
          next $w$
          if $q_{min} \neq \emptyset$ then
            $m' = m^{q_{min}}$; $\mathcal{K}' = \mathcal{K}' \cup \{m'\}$;  goto MarginalEnd
          endif
        next $l$
      endif
MarginalEnd:
    next $m$
    if $\mathcal{K}' = \emptyset$ then
      $\mathcal{K}' = P^{\xi_1 \xi_2, \cdots \xi_n}$
    endif
TotalEnd:
    $A(\mathcal{K}, (s_1, s_2, \cdots s_n)) = A(\mathcal{K}', (s_1, s_2, \cdots s_n))$
```

## 8 Conclusion

1. Algorithm Enhancement improves the decision-making by several percentage points.

2. Changing or setting of different free parameters may be considered as a part of tuning or service each decision-making algorithm should be given before being brought into operation. Bearing in mind that the theoretical joint distribution $P_{\eta\xi_1\xi_2\cdots\xi_n}$ is different for each problem area, the search for optimal input marginals to get the best approximation $\hat{P}_{\eta\xi_1\xi_2\cdots\xi_n}$ is fully justifiable.

3. In general, with other parameters unchanged, it is the size of the marginals that has the greatest influence on the decision.

4. The software necessary for mixing marginals can be seen as a tool kit that helps to find the best approximation of the joint distribution describing the problem area and it is open to further growth if needed.

The author wishes to express his thanks to Jiří Vomlel for comparative runs using the Huggin software for the same input data and for inspiring discussions about the danger of overfitting when testing the influence of marginal selection for real data.

## References

[1] Kellerer H.G.. (1964), Verteilungsfunktionen mit gegebenen Marginalverteilungen (in German), *Z.Wahrsch.Verw.Gebiete, 3, 247-270*

[2] Kříž O. (2007) Comparing algorithms based on marginal problem, *Kybernetika, Vol 43, (2007), No.5, 633–647*

[3] Kříž O. (2009) Selecting marginals for decision making based on marginal problem, *WUPES'09, Proceedings of the 8-th Workshop on Uncertainty Processing, Liblice September 19-23,2009, 144–155*

# Polymatroids and Polyquantoids

**František Matúš**[*]
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
matus@utia.cas.cz

**Abstract**

When studying entropy functions of multivariate probability distributions, polymatroids and matroids emerge. Entropy functions of pure multiparty quantum states give rise to analogous notions, called here polyquantoids and quantoids. Polymatroids and polyquantoids are related via linear mappings and duality. Quantum secret sharing schemes that are ideal are described by selfdual matroids. Expansions of integer polyquantoids to quantoids are studied and linked to that of polymatroids.

## 1 Introduction

A polymatroid $(N, h)$ consists of a finite ground set $N$ and rank function $h$ on the subsets of $N$ that is normalized $h(\emptyset) = 0$, nondecreasing $h(I) \leqslant h(J)$, $I \subseteq J$, and submodular $h(I) + h(J) \geqslant h(I \cup J) + h(I \cap J)$, $I, J \subseteq N$. A polymatroid is entropic if there exists a probability measure $P$ on a finite set $\prod_{i \in N} X_i$ such that $h(I)$ equals Shannon entropy of the marginal of $P$ to $\prod_{i \in I} X_i$, for all $I \subseteq N$. This means that $h$ equals the entropy function of $P$. These functions always induce polymatroids.

In this work, a *polyquantoid* is introduced as a pair $(N, e)$ with a rank function $e$ on the subsets of $N$ that is normalized, complementary $e(I) = e(N \setminus I)$, $I \subseteq N$, and submodular. A polyquantoid is entropic if there exists a quantum state $\rho$ on a complex Hilbert space $\bigotimes_{i \in N} H_i$ of finite dimension such that $e(I)$ equals von Neumann entropy of the reduction of $\rho$ to $\bigotimes_{i \in I} H_i$, for all $I \subseteq N$. This means that $e$ equals the entropy function on $\rho$. These functions always induce polyquantoids, by properties of von Neumann entropy.

A polymatroid/polyquantoid is integer if all values of its rank function are integer numbers. An integer polymatroid whose values on singletons equal zero or one is called matroid. Let *quantoid* be defined as an integer polyquantoid with this property.

This contribution studies interplay between polymatroids, polyquantoids, matroids, quantoids and secret sharing schemes, both classical and quantum. In Section 2, duality of set functions is worked out. Section 3 introduces mutually inverse linear

[*]This work was supported by Grant Agency the Czech Republic under Grant 201/08/0539.

mappings that provide a one-to-one correspondence between tight selfdual polymatroids and polyquantoids, see Theorem 1. This correspondence can serve as a tool for comparing problems on classical and quantum entropy functions.

In Section 4, secret sharing schemes are lifted to the level of polymatroids/polyquantoids. Theorem 2 recalls that the ideal sharing in polymatroids is governed by matroids. This result is translated to polyquantoids in Theorem 3 that describes the ideal quantum sharing via those quantoids that correspond to tight selfdual matroids.

Section 5 departs from the notion of expansions of integer polymatroids to matroids. An analogous construction for integer polyquantoids is introduced to provide expansions of polyquantoids to quantoids, see Theorem 4. Thus, the quantoids play a role of matroids in quantum settings. In Section 6, remarks and discussion of related material and literature are collected.

## 2 Duality

For set functions $h$ with a ground set $N$, the following definition

$$h'(I) \triangleq h(N \setminus I) + h(\emptyset) - h(N) + \sum_{i \in I} \big[ h(i) - h(\emptyset) + h(N) - h(N \setminus i) \big] , \quad I \subseteq N ,$$

gives rise to a duality mapping $h \mapsto h'$. A function $h$ is *selfdual* if $h' = h$. The functions that are complementary, as in polyquantoids, are selfdual.

Let us say that a set function $h$ is *tight* if $h(N \setminus i) = h(N)$ for all $i \in N$. If $h$ is normalized and tight then the definition of duality simplifies to

$$h'(I) = h(N \setminus I) - h(N) + \sum_{i \in I} h(i) , \quad I \subseteq N .$$

**Lemma 1.** *For any function $h$ on the subsets of $N$,*
*(i) $h'(\emptyset) = h(\emptyset)$,*
*(ii) $h'(i) = h(i)$ for $i \in N$,*
*(iii) $h'(N) - h'(N \setminus i) = h(N) - h(N \setminus i)$ for $i \in N$,*
*(iv) $h'' = h$,*
*(v) $h$ is submodular if and only if $h'$ is so,*
*(vi) if $h$ is normalized, submodular and $h(N) \geqslant h(N \setminus i)$, $i \in N$, then $h'$ is nondecreasing.*

*Proof.* First two assertions follow directly from the definition. For $K \subseteq J$ the equality

$$h'(J) - h'(K) = h(N \setminus J) - h(N \setminus K) + \sum_{i \in J \setminus K} \big[ h(i) - h(\emptyset) + h(N) - h(N \setminus i) \big]$$

implies *(iii)*. Choosing $K = N \setminus I$ and $J = N$, it rewrites to

$$h(I) = h'(N \setminus I) + h(\emptyset) - h'(N) + \sum_{i \in I} \big[ h(i) - h(\emptyset) + h(N) - h(N \setminus i) \big] , \qquad I \subseteq N .$$

By *(i)*, *(ii)* and *(iii)*, the right-hand side equals $h''(I)$ which proves *(iv)*. If $h$ is submodular then $I \mapsto h(N \setminus I)$ is so whence $h'$ is submodular. Then, the equivalence *(v)* holds by *(iv)*. If $h$ is normalized and $h(N) \geqslant h(N \setminus i)$, $i \in N$, then for $J \supseteq K$

$$h'(J) - h'(K) \geqslant h(N \setminus J) - h(N \setminus K) + \sum_{i \in J \setminus K} h(i)\,.$$

If $h$ is also submodular then the right-hand side is nonnegative whence *(vi)* holds. □

**Corollary 1.** *The duality mapping restricts to an involution on the (tight) polymatroids.*

## 3 Tight selfdual polymatroids and polyquantoids

Let $h$ and $e$ be set functions with the ground set $N$. The linear mappings $e \mapsto e^\wedge$ and $h \mapsto h^\vee$ introduced here by

$$e^\wedge(I) \triangleq e(I) + \sum_{i \in I} e(i) \quad \text{and} \quad h^\vee(I) \triangleq h(I) - \tfrac{1}{2} \sum_{i \in I} h(i)\,, \quad I \subseteq N\,,$$

are mutually inverse, $(e^\wedge)^\vee = e$ and $(h^\vee)^\wedge = h$. They provide a natural link between the polymatroids and polyquantoids.

**Theorem 1.** *The mappings $e \mapsto e^\wedge$ and $h \mapsto h^\vee$ restrict to mutually inverse bijections between the polyquantoids and the tight selfdual polymatroids.*

*Proof.* Let $(N, e)$ be a polyquantoid. Since $e$ is normalized $e^\wedge(\emptyset) = 0$. The submodularity of $e$ is equivalent to that of $e^\wedge$, and implies $e(N \setminus I) \leqslant e(N \setminus J) + \sum_{i \in J \setminus I} e(i)$ for $I \subseteq J \subseteq N$. By complementarity, $e(I) \leqslant e(J) + \sum_{i \in J \setminus I} e(i)$, and thus $e^\wedge(I) \leqslant e^\wedge(J)$. Therefore, $(N, e^\wedge)$ is a polymatroid. Since $e$ is normalized and complementary

$$e^\wedge(N) = \sum_{j \in N} e(j) = e(N \setminus i) + \sum_{j \in N \setminus i} e(j) = e^\wedge(N \setminus i)\,, \quad i \in N\,,$$

thus $e^\wedge$ is tight. For $I \subseteq N$ it follows that

$$\begin{aligned} (e^\wedge)'(I) &= e^\wedge(N \setminus I) - e^\wedge(N) + \sum_{i \in I} e^\wedge(i) \\ &= \Big[e(N \setminus I) + \sum_{i \in N \setminus I} e(i)\Big] - \sum_{i \in N} e(i) + 2 \sum_{i \in I} e(i) = e^\wedge(I)\,, \end{aligned}$$

thus $e^\wedge$ is selfdual.

Let $(N, h)$ be a tight selfdual polymatroid. Since $h$ is normalized $h^\vee(\emptyset) = 0$. Since $h$ is tight and selfdual $h(I) = h(N \setminus I) - h(N) + \sum_{i \in I} h(i)$, $I \subseteq N$. Then, $h(N)$ is equal to $\frac{1}{2} \sum_{i \in N} h(i)$. It follows that

$$h^\vee(N \setminus I) = \Big[h(I) - h(N) + \sum_{i \in N \setminus I} h(i)\Big] - \tfrac{1}{2} \sum_{i \in N \setminus I} h(i) = h^\vee(I)\,, \quad I \subseteq N\,,$$

thus $h^\vee$ is complementary. The submodularity of $h$ implies that of $h^\vee$. Therefore, $(N, h^\vee)$ is a polyquantoid. □

*Remark* 1. The above proof provides also arguments for the assertion that the mappings $e \mapsto e^{\wedge}$ and $h \mapsto h^{\vee}$ restrict to mutually inverse bijections between the class of normalized complementary functions and the class of normalized tight selfdual functions, dropping submodularity in Theorem 1.

**Corollary 2.** *The mappings* $e \mapsto e^{\wedge}$ *and* $h \mapsto h^{\vee}$ *induce mutually inverse bijections between the integer polyquantoids and the integer tight selfdual polymatroids whose values on all singletons are even.*

**Corollary 3.** *The mappings* $e \mapsto e^{\wedge}$ *and* $h \mapsto h^{\vee}$ *induce mutually inverse bijections between the quantoids and the integer tight selfdual polymatroids whose values on all singletons equal zero or two.*

## 4 Ideal secret sharing

Given a polymatroid $(N,h)$, an element $\mathit{0}$ of $N$ is *perfect* if $h(\mathit{0}\cup I)-h(I)$ equals $h(\mathit{0})$ or zero, for all $I \subseteq N \setminus \mathit{0}$. In the latter case, $I$ is *authorized* for $\mathit{0}$. By submodularity,

$$h(\mathit{0}\cup I)-h(I) \geqslant h(\mathit{0}\cup J)-h(J)\,, \qquad I \subseteq J \subseteq N \setminus \mathit{0}\,.$$

Hence, $h(\mathit{0}\cup I)-h(I)=0$ implies $0 \geqslant h(\mathit{0}\cup J)-h(J)$, and $h(\mathit{0}\cup J)-h(J)=h(\mathit{0})$ implies $h(\mathit{0}\cup I)-h(I) \geqslant h(\mathit{0})$. The two inequalities are tight as $h$ is a polymatroid. Thus, the family of authorized sets for $\mathit{0}$ is closed to supersets and the family of sets $I \subseteq N \setminus \mathit{0}$ with $h(\mathit{0}\cup I)-h(I)$ equal to $h(\mathit{0})$ is closed to subsets. This is referred to as heredity. If $\mathit{0}$ is perfect and $h(\mathit{0})>0$ then the two families are disjoint and cover all subsets of $N \setminus \mathit{0}$, which is referred to as dichotomy.

In a polymatroid $(N,h)$ with a perfect element $\mathit{0} \in N$, an element $i \in N \setminus \mathit{0}$ is *essential* for $\mathit{0}$ if it belongs to some set $I$ that is authorized for $\mathit{0}$ and $h(\mathit{0}\cup I \setminus i)-h(I\setminus i)=h(\mathit{0})$. As a consequence,

$$h(i) \geqslant h(I)-h(I\setminus i) = h(\mathit{0}\cup I)-h(I\setminus i) \geqslant h(\mathit{0}\cup I\setminus i)-h(I\setminus i) = h(\mathit{0})\,,$$

since $h$ is submodular and nondecreasing. A perfect element $\mathit{0}$ in a polymatroid $(N,h)$ is *ideal* if each $i \in N \setminus \mathit{0}$ is essential for $\mathit{0}$ and $h(i)=h(\mathit{0})$.

For example, in any matroid $(N,r)$ each element is perfect. Given $\mathit{0} \in N$, a set $I \subseteq N \setminus \mathit{0}$ is authorized for $\mathit{0}$ if and only if a circuit contained in $\mathit{0}\cup I$ contains $\mathit{0}$. If $r(\mathit{0})=0$, thus $\mathit{0}$ is a loop, then all $i \in N\setminus \mathit{0}$ are essential for $\mathit{0}$. Hence, $\mathit{0}$ is ideal if only if $r(N)=0$. Otherwise, when $r(\mathit{0})=1$, $i$ is essential for $\mathit{0}$ if and only if there exists a circuit of the matroid containing $\mathit{0}$ and $i$. Therefore, $\mathit{0}$ is ideal if only if the matroid is connected. Each element of any connected matroid is ideal.

When restricting to the entropic polymatroids, the above notions correspond to the information-theoretical secret sharing schemes.

The following assertion claims that existence of an ideal element implies matroidal structure. It follows from an existing result, see Section 6, but a self-contained proof is presented for convenience.

**Theorem 2.** *If a polymatroid* $(N,h)$ *has an ideal element then there exists a matroid* $(N,r)$ *and* $t>0$ *such that* $h=t\,r$.

*Proof.* Let $\mathit{0} \in N$ be an ideal element of the polymatroid. If $h(\mathit{0}) = 0$ then $h(i) = 0$ for all $i \in N$ whence $(N, h)$ is a matroid and the assertion holds with any $t > 0$. Let $h(\mathit{0}) > 0$.

The idea is to prove that 'if $L \subseteq N$ is nonempty then there exists $\ell \in L$ such that $h(L) - h(L \setminus \ell)$ equals $h(\mathit{0})$ or zero'. This implication and an induction argument on the cardinality of $L$ show that all values of $h$ are multiples of $h(\mathit{0})$. As a consequence, $h$ equals a matroid rank function multiplied by $t = h(\mathit{0}) > 0$.

If $L \subseteq N$ contains $\mathit{0}$ the implication holds with $\ell = \mathit{0}$ because $\mathit{0}$ is perfect.

If $L \subseteq N \setminus \mathit{0}$ is authorized, $h(\mathit{0} \cup L) = h(L)$, then $h(\mathit{0} \cup I) = h(\mathit{0}) + h(I)$ for some $I \subseteq L$, e.g. $I = \emptyset$. Such a set $I$ is chosen to be inclusion maximal. By dichotomy, $I \subsetneq L$. Let $\ell \in L \setminus I$. Since $I$ is maximal and $\mathit{0}$ perfect, $\ell \cup I$ is authorized, $h(\mathit{0} \cup \ell \cup I) = h(\ell \cup I)$. This and submodularity imply

$$h(\mathit{0} \cup L \setminus \ell) + h(\mathit{0} \cup \ell \cup I) \geqslant h(\mathit{0} \cup L) + h(\mathit{0} \cup I) = h(\mathit{0} \cup L) + h(\mathit{0}) + h(I)\,,$$
$$h(\ell) + h(I) \geqslant h(\ell \cup I) = h(\mathit{0} \cup \ell \cup I)\,.$$

As $\mathit{0}$ is ideal, $h(\mathit{0}) = h(\ell)$, and it follows by adding that $h(\mathit{0} \cup L \setminus \ell) \geqslant h(\mathit{0} \cup L)$. Thus, $h(\mathit{0} \cup L \setminus \ell) = h(\mathit{0} \cup L) = h(L)$ because $h$ is nondecreasing and $L$ authorized. This implies that $h(L) - h(L \setminus \ell)$ equals $h(\mathit{0} \cup L \setminus \ell) - h(L \setminus \ell)$ which is zero or $h(\mathit{0})$ by perfectness of $\mathit{0}$. Hence, the implication holds for every $L$ authorized.

By dichotomy, it remains to consider a nonempty subset $L$ of $N \setminus \mathit{0}$ such that $h(\mathit{0} \cup L)$ equals $h(\mathit{0}) + h(L)$. Since $\mathit{0}$ is ideal, any $\ell \in N \setminus \mathit{0}$ is essential for $\mathit{0}$. Taking some $\ell \in L$ there exists an authorized set $K$, $h(\mathit{0} \cup K) = h(K)$, such that $\ell \in K$ and $h(\mathit{0} \cup K \setminus \ell)$ equals $h(\mathit{0}) + h(K \setminus \ell)$. Such a set $K$ is chosen to obtain the cardinality of $K \setminus L$ minimal. By dichotomy, $K$ is not contained in $L$. For every $k \in K \setminus L$ the minimality implies that the set $L \cup K \setminus k$, containing the chosen $\ell$, is not authorized. In turn, since $h$ is submodular, $L \cup K$ authorized and $h$ nondecreasing

$$h(k) + h(L \cup K \setminus k) \geqslant h(L \cup K) = h(\mathit{0} \cup L \cup K) \geqslant h(\mathit{0} \cup L \cup K \setminus k) = h(\mathit{0}) + h(L \cup K \setminus k)\,.$$

The above two inequalities are tight because $h(\mathit{0}) = h(k)$, using that $\mathit{0}$ is ideal. Therefore, $h(L \cup K) = h(k) + h(L \cup K \setminus k)$ for $k \in K \setminus L$. By induction,

$$h(I \cup (K \setminus L)) = h(I) + \sum_{k \in K \setminus L} h(k)\,, \qquad I \subseteq L\,.$$

This implies that $h(L) - h(L \setminus \ell)$ equals $h(L \cup K) - h((L \cup K) \setminus \ell)$. The previous part of the proof is applied to the authorized set $K$ in the role of $L$ and the non-authorized set $K \setminus \ell$ in the role of $I$ to conclude that $h(\mathit{0} \cup K \setminus \ell) = h(\mathit{0} \cup K)$. This implies that $h(\mathit{0} \cup (L \cup K) \setminus \ell)$ equals $h(\mathit{0} \cup L \cup K)$ which coincides with $h(L \cup K)$ because $L \cup K$ is authorized. Hence, $h(L) - h(L \setminus \ell)$ equals $h(\mathit{0} \cup (L \cup K) \setminus \ell) - h((L \cup K) \setminus \ell)$ which is zero or $h(\mathit{0})$ by perfectness of $\mathit{0}$. Thus, the implication holds for all nonempty $L \subseteq N \setminus \mathit{0}$. □

Given a polyquantoid $(N, e)$, an element $\mathit{0}$ of $N$ is *perfect* if $e(\mathit{0} \cup I) - e(I)$ equals $e(\mathit{0})$ or $-e(\mathit{0})$, for all $I \subseteq N \setminus \mathit{0}$. In the latter case, $I$ is *authorized* for $\mathit{0}$. The definition of perfectness does not change when requiring that $e^\wedge(\mathit{0} \cup I) - e^\wedge(I)$ equals $e^\wedge(\mathit{0})$ or zero. Thus, $\mathit{0}$ is perfect in $(N, e)$ if and only if it is perfect in the polymatroid $(N, e^{\wedge})$. Therefore, supersets of authorized sets are authorized and the equality $e(\mathit{0} \cup I) - e(I) = e(\mathit{0})$ with $I \subseteq N \setminus \mathit{0}$ is inherited by the subsets of $I$. The dichotomy takes place whenever $e(\mathit{0}) > 0$.

In a polyquantoid $(N, e)$ with a perfect element $\mathit{0} \in N$, an element $i \in N \setminus \mathit{0}$ is *essential* for $\mathit{0}$ if there exists a set $I$ which authorized for $\mathit{0}$, contains $i$ and $e(\mathit{0} \cup I \setminus i) - e(I \setminus i) = e(\mathit{0})$. This is equivalent to saying that $i \in N \setminus \mathit{0}$ is *essential* for $\mathit{0}$ in the polymatroid $(N, e^{\wedge})$. Hence, $e(i) \geqslant e(\mathit{0})$ once $i$ is essential for $\mathit{0}$ in $(N, e)$. A perfect element $\mathit{0}$ in a polyquantoid $(N, e)$ is *ideal* if each $i \in N \setminus \mathit{0}$ is essential for $\mathit{0}$ and $e(i) = e(\mathit{0})$.

**Theorem 3.** *If a polyquantoid $(N, e)$ has an ideal element then there exists a tight selfdual matroid $(N, r)$ and $t > 0$ such that $e = t\, r^{\vee}$.*

*Proof.* If $\mathit{0} \in N$ is ideal in the polyquantoid then $\mathit{0}$ is ideal in $(N, e^{\wedge})$ which is a tight selfdual polymatroid by Theorem 1. Theorem 2 implies that $e^{\wedge} = t\, r$ for $t > 0$ and a matroid rank function $r$. Hence, $r$ is tight, selfdual, and $e = (e^{\wedge})^{\vee} = (t\, r)^{\vee} = t\, r^{\vee}$. □

As a consequence, if $\mathit{0}$ is an ideal element of a polyquantoid then $I \subseteq N \setminus \mathit{0}$ is authorized for $\mathit{0}$ if and only if $\mathit{0} \in C \subseteq \mathit{0} \cup I$ for some circuit $C$ of the tight selfdual matroid that is assigned to the polyquantoid in Theorem 3.

# 5 Expansions

A set function $h$ with a ground set $N$ *expands* to a set function $h^{\#}$ with a ground set $N^{\#}$ if there exists a mapping $\phi$ on $N$ ranging in the family of subsets of $N^{\#}$ such that $h^{\#}(\bigcup_{i \in I} \phi(i))$ equals $h(I)$ for all $I \subseteq N$.

Each integer polymatroid $(N, h)$ can be expanded to a matroid as follows. Let $\phi$ map $i \in N$ to a set $\phi(i)$ of cardinality $h(i)$ such that these sets are pairwise disjoint. Writing $\phi(I) = \bigcup_{i \in I} \phi(i)$, $I \subseteq N$, the construction

$$h_{\phi} \colon K \mapsto \min_{J \subseteq N} \left[\, h(J) + |K \setminus \phi(J)| \,\right], \qquad K \subseteq \phi(N)\,,$$

defines a matroid $(\phi(N), h_{\phi})$ called a *free expansion* of $(N, h)$. The value $h_{\phi}(K)$ depends on $K$ only through the cardinalities of the sets $\phi(i) \cap K$, $i \in N$. The minimization can be equivalently over the sets that satisfy

$$\{i \in N \colon \phi(i) \cap K \neq \emptyset\} \supseteq J \supseteq \{i \in N \colon \emptyset \neq \phi(i) \subseteq K\}$$

since $h$ is nondecreasing and submodular. Such sets $J$ are termed to be *adapted* to $K$. Hence, $h_{\phi}(\phi(I))$ equals $h(I)$ for all $I \subseteq N$, using that $\{i \in I \colon \phi(i) \neq \emptyset\}$ is the unique adapted set to $\phi(I)$, and thus $h$ expands to $h_{\phi}$.

For any integer polyquantoid $(N, e)$, an analogous construction is introduced as follows. Let $\psi$ map $i \in N$ to a set $\psi(i)$ of cardinality $e(i)$ such that these sets are pairwise disjoint, $\psi(I) = \bigcup_{i \in I} \psi(i)$, $I \subseteq N$, and

$$e_{\psi} \colon K \mapsto \min_{J \subseteq N} \left[\, e(J) + |K \vartriangle \psi(J)| \,\right], \qquad K \subseteq \psi(N)\,.$$

Let $(\psi(N), e_\psi)$ be called a *free expansion* of $(N, e)$. The minimization can be equivalently over the sets adapted to $K$, using that $e$ is normalized, complementary and submodular. Therefore, $e_\psi(\psi(I)) = e(I)$, $I \subseteq N$, thus $e$ expands to $e_\psi$ indeed.

The following assertion shows that from the viewpoint of expansions, quantoids are for polyquantoids what matroids are to polymatroids.

**Theorem 4.** *Any free expansion of an integer polyquantoid is a quantoid.*

*Proof.* Let $(N, e)$ be an integer polyquantoid and $\psi$ a mapping as above. By definition, $e_\psi(K) = |K|$ if $K \subseteq \psi(i)$ for some $i \in N$. In particular, $e_\psi$ is normalized and its values on singletons equal one.

For $J \subseteq N$ adapted to $K \subseteq \psi(N)$, the set $J' = \{i \in N \setminus J \colon \psi(i) \neq \emptyset\}$ is adapted to $\psi(N) \setminus K$ and

$$e(J) + |K \vartriangle \psi(J)| = e(J') + \big|(\psi(N) \setminus K) \vartriangle (\psi(J'))\big|$$

using $e(J) = e(N \setminus J) = e(J')$. Moreover, $J \mapsto J'$ is a bijection between the families of those sets that are adapted to $K$, resp. to $\psi(N) \setminus K$. It follows by minimization that $e_\psi(K)$ equals $e_\psi(\psi(N) \setminus K)$, thus $e_\psi$ is complementary.

To prove that $e_\psi$ is submodular, let $K, L \subseteq \psi(N)$ and

$$e_\psi(K) = e(I) + |K \vartriangle \psi(I)| \quad \text{and} \quad e_\psi(L) = e(J) + |L \vartriangle \psi(J)|$$

where $I$ is adapted to $K$ and $J$ is adapted to $L$. As $e(I) + e(J) \geqslant e(I \cup J) + e(I \cap J)$ and

$$|K \vartriangle \psi(I)| + |L \vartriangle \psi(J)| = |(K \cup L) \vartriangle \psi(I \cup J)| + |(K \cap L) \vartriangle \psi(I \cap J)|$$

the submodularity of $e_\psi$ follows. □

In the remaining part of this section, expansions of polymatroids and polyquantoids are compared by means of the mappings $e \mapsto e^\wedge$ and $h \mapsto h^\vee$.

Let $(N, h)$ be an integer polymatroid with $h(i)$ even for all $i \in N$ and $(\phi(N), h_\phi)$ its free expansion. Then, each set $\phi(i)$ can be partitioned into two-element blocks $m = \{k, \ell\}$ having $k, \ell \in \phi(i)$ different. Let $\phi^*(i)$ denote the set of all blocks in such a partition, $\phi^*(I) = \bigcup_{i \in I} \phi^*(i)$, $I \subseteq N$, and

$$h_{\phi^*}(M) \triangleq h_\phi(\textstyle\bigcup M) = \min_{J \subseteq N} \left[ h(J) + |\textstyle\bigcup M \setminus \phi(J)| \right], \qquad M \subseteq \phi^*(N)\,.$$

This defines a polymatroid $(\phi^*(N), h_{\phi^*})$ called here 2-*factor* of $(\phi(N), h_\phi)$. By definitions, $(N, h)$ expands to $(\phi^*(N), h_{\phi^*})$ which in turn expands to $(\phi(N), h_\phi)$.

The following assertion indicates a correspondence between the free expansions of polymatroids and polyquantoids.

**Lemma 2.** *If $(N, e)$ is an integer polyquantoid, $h = e^\wedge$, $(\phi(N), h_\phi)$ a free expansion of $(N, h)$ and $(\phi^*(N), h_{\phi^*})$ its 2-factor then $(\phi^*(N), (h_{\phi^*})^\vee)$ is a free expansion of $(N, e)$.*

*Proof.* For $M \subseteq \phi^*(N)$

$$(h_{\phi^*})^\vee(M) = h_{\phi^*}(M) - \tfrac{1}{2} \sum_{m \in M} h_{\phi^*}(\{m\}) = h_\phi(\bigcup M) - |M|$$

using that $h_{\phi^*}(\{m\}) = h_\phi(m) = 2$. Since $e(j) = h(j)/2 = |\phi^*(j)|$ for $j \in N$, if $J \subseteq N$ then $h(J) = e(J) + |\phi^*(J)|$. Then, by the definition of polymatroid expansions,

$$(h_{\phi^*})^\vee(M) = \min_{J \subseteq N} \big[ \, e(J) + |\phi^*(J)| + |\bigcup M \setminus \phi(J)| \, \big] - |M|$$

Here, $|\bigcup M \setminus \phi(J)| = 2|M \setminus \phi^*(J)|$. Since $|\phi^*(J)| + |M \setminus \phi^*(J)| - |M|$ equals $|\phi^*(J) \setminus M|$ it follows from definition of polyquantoid expansions that $(h_{\phi^*})^\vee$ coincides with $e_{\phi^*}$. □

In the above lemma, the integer polymatroid $h = e^\wedge$ is tight and selfdual, by Theorem 1. The following two lemmas imply that the expansion $h_\phi$ and its 2-factor $h_{\phi^*}$ have the same properties. Hence, Theorem 4 can be proved alternatively by combining Theorem 1 with Lemmas 2, 3 and 4. This argument is more involved but illustrates the interplay between the two kinds of expansions.

**Lemma 3.** *If an integer polymatroid is tight and selfdual then so are its free expansions.*

*Proof.* Let $(N, h)$ be an integer polymatroid and $\phi$ a mapping with $|\phi(i)| = h(i)$, $i \in N$, as above. For $k \in \phi(N)$ there exists unique $i \in N$ such that $k \in \phi(i)$. Assuming that $h$ is tight $h_\phi(\phi(N \setminus i)) = h(N \setminus i) = h(N) = h_\phi(\phi(N))$. This implies $h_\phi(\phi(N) \setminus k) = h_\phi(\phi(N))$ whence $h_\phi$ is tight.

By definition, $h_\phi(K) = |K|$ if $K \subseteq \phi(i)$ for some $i \in N$. Hence, assuming that $h$ is tight and selfdual, for a set $J \subseteq N$ adapted to $K \subseteq \phi(N)$,

$$\begin{aligned} h(J) + |K \setminus \phi(J)| &= h(N \setminus J) - h(N) + |\phi(J)| + |K \setminus \phi(J)| \\ &= h(N \setminus J) - h_\phi(\phi(N)) + |K| + \big|(\phi(N) \setminus K) \setminus (\phi(N \setminus J))\big| \,. \end{aligned}$$

Minimizing over the adapted sets, it follows that $h_\phi(K) \geqslant h_\phi(\phi(N) \setminus K) - h_\phi(\phi(N)) + |K|$. Since $J$ is adapted to $K$ if and only if $J' = \{i \in N \setminus J \colon \phi(i) \neq \emptyset\}$ is adapted to $\phi(N) \setminus K$ this inequality is tight. Thus, $h_\phi$ is selfdual. □

**Lemma 4.** *If an integer polymatroid is tight, selfdual and takes even values on all singletons then all 2-factors of its free expansions are tight and selfdual.*

*Proof.* Let $(N, h)$ satisfy the assumptions. Keeping the notation of the proof of Lemma 3, for $m \in \phi^*(N)$ there exists unique $i \in N$ such that $m \subseteq \phi(i)$. Since $h$ is tight $h_\phi(\phi(N \setminus i))$ equals $h_\phi(\phi(N))$. Hence,

$$h_{\phi^*}(\phi^*(N) \setminus \{m\}) = h_\phi(\phi(N) \setminus m) \geqslant h_\phi(\phi(N \setminus i)) = h_\phi(\phi(N)) = h_{\phi^*}(\phi^*(N))$$

In turn, $h_{\phi^*}$ is tight.

By Lemma 3, $(\phi(N), h_\phi)$ is selfdual. Hence, for $M \subseteq \phi^*(N)$

$$\begin{aligned}(h_{\phi^*})'(M) &= h_{\phi^*}(\phi^*(N) \setminus M) - h_{\phi^*}(\phi^*(N)) + \sum_{m \in M} h_{\phi^*}(\{m\}) \\ &= h_\phi(\phi(N) \setminus \bigcup M) - h_\phi(\phi(N)) + \sum_{k \in \bigcup M} h_\phi(k) = h_\phi(\bigcup M) = h_{\phi^*}(M)\end{aligned}$$

using that $h_{\phi^*}(\{m\}) = 2 = h_\phi(k) + h_\phi(\ell)$ where $m = \{k, \ell\}$. □

# 6 Discussion

The polymatroids [10, 5, 14] have been studied for decades and history of the matroid theory [16] is even longer. The duality defined in Section 2 is in general different from known ones, as those in [14, 16, 20], since it conserves values on singletons, see Lemma 1*(ii)*. For matroids without loops and coloops, the duality coincides with the usual one [16, 2.1.9]. Functions called above selfdual are in literature also termed identically selfdual. Tightness is a notion suitable for this work but not used elsewhere. A matroid is tight if and only if it has no coloop.

The problem which polymatroid is entropic is of interest for information-theoretical approaches to networks and cryptography, and beyond, for references see e.g. [21, 11, 12]. Its quantum version, asking which polyquantoid is entropic, has also attracted considerable attention [17, 9, 3].

Ideal secret sharing schemes were investigated first in a combinatorial setting [2]. Theorem 2 is a consequence of [1, Theorem 2], building on [2, Theorem 1]. The presented proof is based on the approach of [1]. Quantum secret sharing schemes go back to [4, 7, 6]. Ideal sharing and matroids were discussed recently in [18, 19]. Theorem 3 solves a question related to [18, Fig. 2]. It implies that the access structure of any ideal quantum secret sharing scheme must be generated by circuits of a tight selfdual matroid.

Free expansions were proposed independently by several researchers, see [8, 13, 15]. If an entropic integer polymatroid expands to a matroid then the latter is the limit of entropic polymatroids [12, Theorem 4]. The quantum analogue of this assertion is open.

# Acknowledgement

The author wants to express sincere thanks to Mary Beth Ruskai and Andreas Winter for numerous inspiring discussions and kind hospitality during visits. Discussions with Oriol Farràs Ventura on quantum secret sharing are acknowledged.

# References

[1] G.R. Blakley and G.A. Kabatianski (1997) Generalized ideal secret-sharing schemes and matroids. *Problems of Inf. Transmission* **33** 277–284.

[2] E.F. Brickell and D.M. Davenport (1991) On the classification of ideal secret-sharing schemes. *J. Cryptology* **4** 123–134.

[3] J. Cadney, N. Linden and A. Winter (2012) Infinitely many constrained inequalities for the von Neumann entropy. *IEEE Trans. Inf. Th.* **58** 3657–3663.

[4] R. Cleve, D. Gottesman and H.-K. Lo (1999) How to share a quantum secret. *Ph. Review Letters* **83** 648–651.

[5] S. Fujishige (1991) *Submodular Functions and Optimization.* North-Holland, Amsterdam.

[6] D. Gottesman (2000) Theory of quantum secret sharing. *Phys. Rev. A* **61** 042311.

[7] M. Hillery, V. Bužek and A. Berthiaume (1999) Quantum secret sharing. *Phys. Rev. A* **59** 1829–1834.

[8] T. Helgason (1974) Aspects of the theory of hypermatroids. In: *Hypergraph Seminar* (C. Berge and D.K. Ray-Chaudhuri, eds.), Lecture Notes in Mathematics **411**, Springer-Verlag, Berlin, 191–214.

[9] N. Linden and A. Winter (2005) A new inequality for the von Neumann entropy. *Commun. Math. Phys.* **259** 129–138.

[10] L. Lovász (1982) Submodular functions and convexity. In: *Mathematical Programming – The State of the Art* (A. Bachem, M. Grötchel and B. Korte, eds.), Springer-Verlag, Berlin, 234–257.

[11] F. Matúš (2007) Infinitely many information inequalities. *Proceedings ISIT 2007*, Nice, France, 41–44.

[12] F. Matúš (2007) Two constructions on limits of entropy functions. *IEEE Trans. Inf. Th.* **53** 320–330.

[13] C.J.H. McDiarmid (1975) Rado's theorem for polymatroids. *Math. Proc. Cambridge Phil. Soc.* **78** 263–281.

[14] H. Narayan (1997) *Submodular Functions and Electrical Networks.* Elsevier, Amsterdam.

[15] H.Q. Nguyen (1978) Semimodular functions and combinatorial geometries. *Trans. AMS* **238** 355–383.

[16] J.G. Oxley (1992) *Matroid Theory.* Oxford University Press, Oxford, New York, Tokyo.

[17] N. Pippenger (2003) The inequalities of quantum information theory. *IEEE Trans. Inf. Th.* **49** 773–789.

[18] P. Sarvepalli and R. Raussendorf (2010) Matroids and quantum-secret-sharing schemes. *Physical Review A* **81** 052333.

[19] P. Sarvepalli (2011) Quantum codes and symplectic matroids. (arXiv:1104.1171v1 [quant-ph])

[20] G. Whittle (1992) Duality in polymatroids and set functions. *Combinatorics, Probability and Computing* **1** 275–280.

[21] Z. Zhang and R.W. Yeung (1998) On characterization of entropy function via information inequalities. *IEEE Trans. Inf. Th.* **44** 1440–1452.

# Scaling of Model Approximation Errors and Expected Entropy Distances

**Guido F. Montúfar**
Department of Mathematics
Pennsylvania State University
University Park PA 16802 USA
gfm10@psu.edu

**Johannes Rauh**
Max Planck Institute
for Mathematics in the Sciences
Inselstr. 22 04103 Leipzig Germany
jrauh@mis.mpg.de

**Abstract**

We compute the expected value of the Kullback-Leibler divergence to various fundamental statistical models with respect to canonical priors on the probability simplex. This yields information about the scaling of model approximation errors depending on the cardinality of the sample spaces, and it is a useful reference for more complicated statistical models such as restricted Boltzmann machines.

## 1 Introduction

Let $p, q$ be probability distributions on a finite set $\mathcal{X}$. The *information divergence* or *relative entropy* or *Kullback Leibler divergence*

$$D(p\|q) = \sum_{i\in\mathcal{X}} p_i \log \frac{p_i}{q_i}$$

is a natural measure of dissimilarity between probability distributions that describes how easy it is to distinguish two distributions $p$ and $q$ by means of statistical experiments. In this paper we use the natural logarithm. The divergence is related to the log-likelihood: If $p$ is an empirical distribution, summarizing the outcome of $n$ statistical experiments, then the log-likelihood of a distribution $q$ equals $-n(D(p\|q) + H(p))$. Hence, finding a *maximum likelihood estimator* $q$ within some set of probability distributions $\mathcal{M}$ is the same as finding a minimizer of the divergence $D(p\|q)$ with $q$ restricted to $\mathcal{M}$. The value of $D(p\|q)$ quantifies how well, or bad, the data can be described by $q$ (and by $\mathcal{M}$).

Assume that $\mathcal{M}^{\text{true}}$ is a set of probability distributions for which we do not have a simple mathematical description. We are interested in finding a model $\mathcal{M}$ which does not necessarily include all distributions from $\mathcal{M}^{\text{true}}$, but which approximates them relatively well. What error magnitude should we accept from a good model?

To assess the expressive power of a model $\mathcal{M}$, we study properties of the function $p \mapsto D(p\|\mathcal{M}) = \inf_{q\in\mathcal{M}} D(p\|q)$. For example, the problem of finding the maximizers of this function corresponds to a worst case analysis. The problem of maximizing the divergence from a statistical model was first posed, with different motivation, in [1]. Since then, a lot of progress has been made, notably in the case where $\mathcal{M}$ is an exponential family [5, 4, 8], but also for discrete mixture models and restricted Boltzmann machines [6].

This worst case bound is not the only aspect that decides whether a given model is suited, but also the expected performance and *expected error* are of interest. This leads to the mathematical problem of computing the expectation value

$$\langle D(p\|\mathcal{M})\rangle = \int_\Delta D(p\|\mathcal{M})\,\psi(p)\,\mathrm{d}p,$$

where $p$ is drawn from a probability density $\psi$ on the probability simplex, called the *prior distribution*, or *prior* for short. The correct prior depends on the concrete problem at hand and is often difficult to determine. Given certain conditions on the prior, we also ask, how different is the worst case from the average case, and how much can this behavior be influenced by the choice of the model? We focus on the case that the prior $\psi$ is the uniform distribution or a Dirichlet distribution. It turns out that in most cases the worst-case error is unbounded (as the number of elementary events grows), while the expected error is bounded. Our analysis leads to integrals that have been considered in a Bayesian framework for function estimation in [10], and we can take adventage of the tools developed there.

Our first observation is that, if $\psi$ is the uniform prior, then the expected divergence from the uniform distribution is a monotone function of the system size $N$ (the number of elementary events) and converges to the constant $1-\gamma \approx 0.4228 \approx 0.6099\log(2)$ as $N\to\infty$, where $\gamma$ is the *Euler-Mascheroni* constant. Many natural statistical models contain the uniform distribution, and the expected divergence from such models is then bounded by the same constant. In comparison, for randomly chosen distributions $p$ and $q$, the expected divergence $\langle D(p\|q)\rangle_{p,q}$ equals $1-1/N$. We show, for a class of models including the independence models, partition models, mixtures of product distributions with disjoint supports [6], and decomposable hierarchical models, that the expected divergence actually has the same limit $1-\gamma$, provided that the models remain *small* with respect to $N$ (this is the case in most applications). In contrast, the maximum of the divergence from these models is at least $\log(N/(\dim\mathcal{M}+1))$, see [9]. For reasonable choices of the parameters, the results for Dirichlet priors are similar.

In Section 2 we define the models that we are interested in and collect basic properties of the Dirichlet priors. Section 3 contains analytical results for expectation values of entropies and divergences from these models. The results are interpreted in Section 4. Proofs and calculations are deferred to Appendix A.

# 2 Preliminaries

## 2.1 Models from statistics and machine learning

We consider random variables on a finite set of elementary events $\mathcal{X}$, $|\mathcal{X}| = N$. The set of probability distributions on $\mathcal{X}$ is the $(N-1)$-simplex $\Delta_{N-1}\subset\mathbb{R}^N$. We call any subset $\mathcal{M} \subseteq \Delta_{N-1}$ that can be densely parametrized a model. The support sets of a model $\mathcal{M}$ are the support sets $\operatorname{supp}(p) = \{i \in \mathcal{X} \,|\, p_i > 0\}$ of points $p = (p_i)_{i\in\mathcal{X}}$ in $\mathcal{M}$.

The *k-mixture* of a model $\mathcal{M}$ is the union of all convex combinations of any $k$ of its points, $\mathcal{M}^k := \{\sum_{i=1}^m \lambda_i p^{(i)} \,|\, \lambda_i \geq 0, \sum_i \lambda_i = 1, p^{(i)} \in \mathcal{M}\}$. The *k-mixture with disjoint supports* is the subset of $\mathcal{M}^k$ defined by

$$\mathcal{M}_0^k = \left\{ \sum_{i=1}^k \lambda_i p^{(i)} \in \mathcal{M}^k \,\middle|\, \operatorname{supp}(p^{(i)}) \cap \operatorname{supp}(p^{(j)}) = \emptyset \text{ for all } i \neq j \right\}.$$

Let $\varrho = \{A_1, \dots, A_K\}$ be a partition of $\mathcal{X}$. The *partition model* $\mathcal{M}_\varrho$ consists of all $p \in \Delta_{N-1}$ that satisfy $p_i = p_j$ whenever $i, j$ belong to the same block of $\varrho$. Partition models are closures of convex exponential families with uniform reference measure. The closure of an arbitrary convex exponential family is of the form (see [4])

$$\mathcal{M}_{\varrho,\nu} = \left\{ \sum_{k}^{K} \lambda_k \frac{\mathbb{1}_{A_k}\nu}{\nu(A_k)} \,\middle|\, \lambda_k \geq 0, \sum_{k}^{K} \lambda_k = 1 \right\},$$

where $\nu : \mathcal{X} \to (0, \infty)$ is a positive function on $\mathcal{X}$, called *reference measure*, and $\mathbb{1}_A$ is the indicator function of $A$. Note that all measures $\nu$ with equal conditional distributions $\nu(\cdot|A_k)$ yield the same model. In fact, $\mathcal{M}_{\varrho,\nu}$ equals the $K$-mixture of the set $\{\nu(\cdot|A_k) : k = 1, \dots, K\}$.

For a composite system of $n$ variables, $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, $|\mathcal{X}_i| = N_i$ for all $i$. A *product distribution* is a distribution of the form

$$p(x_1, \dots, x_n) = p_1(x_1) \cdots p_n(x_n),$$

where $p_i \in \Delta_{N_i - 1}$. The *independence model* is the set of all product distributions on a composite system. The support sets of the independence model are the sets of the form $A = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n$ with $\mathcal{Y}_i \subseteq \mathcal{X}_i$ for each $i$.

Let $\mathcal{S}$ be a simplicial complex on $\{0, \dots, n\}$. The *hierarchical model* $\mathcal{M}_\mathcal{S}$ consists of all probability distributions that have a factorization of the form $p(x) = \prod_{S\in\mathcal{S}} \Phi_S(x)$, where $\Phi_S$ is a positive function that depends only on the $S$-components of $x$. The model $\mathcal{M}_\mathcal{S}$ is called *reducible* if there exist simplicial subcomplexes $\mathcal{S}_1, \mathcal{S}_2 \subset \mathcal{S}$ such that $\mathcal{S}_1 \cup \mathcal{S}_2 = \mathcal{S}$ and $\mathcal{S}_1 \cap \mathcal{S}_2$ is a simplex. In this case, the set $(\bigcup_{\mathcal{Y}\in\mathcal{S}_1} \mathcal{Y}) \cap (\bigcup_{\mathcal{Y}\in\mathcal{S}_2} \mathcal{Y})$ is called a *separator*. $\mathcal{M}_\mathcal{S}$ is *decomposable* if it can be iteratively reduced into simplices. The reduction can be described by a *junction tree* (see [2]), which is a tree $(V, E)$ with vertex set the set of facets of $\mathcal{S}$ and such that the following holds: If $(\mathcal{X}, \mathcal{Y})$ is an edge, then $\mathcal{X} \cap \mathcal{Y}$ is a separator, and if this edge is removed from the tree, then the two resulting trees are junction trees of two subcomplexes $\mathcal{S}_1$ and $\mathcal{S}_2$ separated by $\mathcal{X} \cap \mathcal{Y}$. In general the junction tree is not unique, but the multi-set of separators is unique. The independence model is an example of a decomposable model.

For most models it is not possible to find a closed formula for $D(\cdot\|\mathcal{M})$, since there is no closed formula for $\operatorname{arginf}_{q\in\mathcal{M}} D(p\|q)$. However, for some of the above mentioned models a closed formula does exist:

The divergence from the independence model is called *multi-information* and satisfies

$$MI(X_1, \dots, X_n) = D(p\|\mathcal{M}_1) = -H(X_1, \dots, X_n) + \sum_{k=1}^n H(X_k). \tag{1}$$

If $n = 2$ it is also called the *mutual information* of $X_1$ and $X_2$. The divergence from $\mathcal{M}_{\varrho,\nu}$ equals (see [4, eq. (1)])

$$D(p\|\mathcal{M}_{\varrho,\nu}) = D(p\|\sum_{k=1}^{K} p(A_k)\nu(x|A_k)) \ . \tag{2}$$

For a decomposable model $\mathcal{M}_{\mathcal{S}}$ with junction tree $(V, E)$,

$$D(p\|\mathcal{M}_{\mathcal{S}}) = \sum_{S\in V} H_p(X_S) - \sum_{S\in E} H_p(X_S) - H(p). \tag{3}$$

Here, $H_p(X_S)$ denotes the joint entropy of the random variables $\{X_i\}_{i\in S}$ under $p$.

## 2.2 Dirichlet prior

The Dirichlet distribution (or Dirichlet prior) with *concentration parameter* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$, $\alpha_i > 0$ for all $i$, is the probability distribution on $\Delta_{N-1}$ defined by $\mathrm{Dir}_{\boldsymbol{\alpha}}(p) := \frac{1}{\sqrt{N}} \frac{\Gamma(\sum_{i=1}^{N}\alpha_i)}{\prod_{i=1}^{N}\Gamma(\alpha_i)} \prod_{i=1}^{N} p_i^{\alpha_i - 1}$ for $p = (p_1, \ldots, p_N) \in \Delta_{N-1}$, where $\Gamma$ is the gamma function. We write $\alpha = \sum_{i=1}^{N} \alpha_i$.

We will highlight especially the symmetric case $(\alpha_1, \ldots, \alpha_N) = (a, \ldots, a)$, which assigns no preferences to the elementary events. Observe that $\mathrm{Dir}_{(1,\ldots,1)}$ is the uniform probability density on $\Delta_{N-1}$. Furthermore, it is known that $\lim_{a\to 0}\mathrm{Dir}_{(a,\ldots,a)}$ is uniformly concentrated in the point measures (it assigns mass $1/N$ to $p = \delta_x$, $x \in \mathcal{X}$), while $\lim_{a\to\infty}\mathrm{Dir}_{(a,\ldots,a)}$ is concentrated in the uniform distribution $u := (1/N, \ldots, 1/N)$. In general, if $\boldsymbol{\alpha} \in \Delta_{N-1}$, then $\lim_{\kappa\to\infty}\mathrm{Dir}_{\kappa\boldsymbol{\alpha}}$ is the Dirac delta concentrated on $\boldsymbol{\alpha}$.

The Dirichlet distributions satisfy the following *aggregation property*: Consider a partition $\varrho = \{A_1, \ldots, A_K\}$ of $\mathcal{X} = \{1, \ldots, N\}$. If $p = (p_1, \ldots, p_N) \sim \mathrm{Dir}_{(\alpha_1,\ldots,\alpha_N)}$, then $(\sum_{i\in A_1} p_i, \ldots, \sum_{i\in A_K} p_i) \sim \mathrm{Dir}_{(\sum_{i\in A_1}\alpha_i,\ldots,\sum_{i\in A_K}\alpha_i)}$, see, e.g., [3]. We write $\boldsymbol{\alpha}^{\varrho} = (\alpha_1^{\varrho}, \ldots, \alpha_K^{\varrho})$, $\alpha_k^{\varrho} = \sum_{i\in A_k}\alpha_i$ for the concentration parameter induced by the partition $\varrho$. The aggregation property is useful when treating marginals of composite systems. Given a composite system with $\mathcal{X} = \mathcal{X}_1\times\cdots\times\mathcal{X}_n$, $|\mathcal{X}| = N$, $\mathcal{X}_k = \{1, \ldots, N_k\}$ we write $\boldsymbol{\alpha}^k = (\alpha_1^k, \ldots, \alpha_{N_k}^k)$, $\alpha_j^k = \sum_{x\in\mathcal{X}\colon x_k = j}\alpha_x$ for the concentration parameter of the Dirichlet distribution induced on the $\mathcal{X}_k$-marginal

$$\left(\sum_{x\in\mathcal{X}\colon x_k=1} p(x), \ldots, \sum_{x\in\mathcal{X}\colon x_k=N_k} p(x)\right).$$

Note that $\sum_{j=1}^{N_k}\alpha_j^k = \alpha$, and moreover, if $\alpha_x = 1$ for all $x \in \mathcal{X}$, then $\alpha_j^k = N/N_k$ for $j = 1, \ldots, N_k$. For example, if $p$ is drawn uniformly from the simplex of joint distributions $\Delta_{N-1}$, then the sampled marginal probability distribution $p(y_k) = \sum_{x\in\mathcal{X}\colon x_k=y_k} p(x)$, $y_k \in \mathcal{X}_k$ is Dirichlet distributed in $\Delta_{N_k-1}$ with concentration parameter $\boldsymbol{\alpha}^k = (N/N_k, \ldots, N/N_k)$.

# 3 Expected entropies and divergences

For any $k \in \mathbb{N}$ let $h(k) = 1 + \frac{1}{2} + \cdots + \frac{1}{k}$ be the $k$th *harmonic number*. It is known that for large $k$,

$$h(k) = \log(k) + \gamma + O(\frac{1}{k}),$$

where $\gamma \approx 0.57721$ is the *Euler-Mascheroni constant*. Moreover, $h(k) - \log(k)$ is strictly positive and decreases monotonically. We also need the natural analytic extension of $h$ to the non-negative reals given by $h(z) = \partial_z \log(\Gamma(z+1)) + \gamma$, where $\Gamma$ is the gamma function.

The following theorems present formulas for expectation values of divergences from models as well as asymptotic results. The results are based on explicit solutions of the integrals, as done by [10]. The proofs are contained in Appendix A.

**Theorem 1.** *If* $p \sim \mathrm{Dir}_{\boldsymbol{\alpha}}$, *then:*

- $\langle H(p) \rangle = h(\alpha) - \sum_{i=1}^{N} \frac{\alpha_i}{\alpha} h(\alpha_i)$
- $\langle D(p\|u) \rangle = \log(N) - h(\alpha) + \sum_{i=1}^{N} \frac{\alpha_i}{\alpha} h(\alpha_i)$

*In the symmetric case* $(\alpha_1, \ldots, \alpha_N) = (a, \ldots, a)$,

- $\langle H(p) \rangle = h(Na) - h(a)$

$$= \begin{cases} \log(Na) + \gamma - h(a) + O(1/Na) & \textit{for large } N \textit{ and const. } a \\ \log(N) + O(1/a) & \textit{for large } a \textit{ and arb. } N \\ O(aN) & \textit{as } a \to 0 \textit{ with bounded } N \\ h(c) + O(a) & \textit{as } a \to 0 \textit{ with } aN = c \end{cases}$$

- $\langle D(p\|u) \rangle = \log(N) - h(aN) + h(a)$

$$= \begin{cases} h(a) - \log(a) - \gamma + O(1/Na) & \textit{for large } N \textit{ and const. } a \\ O(1/a) & \textit{for large } a \textit{ and arb. } N \\ \log(N) + O(aN) & \textit{as } a \to 0 \textit{ with bounded } N \\ \log(N) - h(c) + O(a) & \textit{as } a \to 0 \textit{ with } aN = c. \end{cases}$$

The maximum of the (Shannon) entropy $H(p) = -\sum_i p_i \log p_i$ on the probability simplex $\Delta_{N-1}$ is attained at the uniform distribution $u$, which satisfies $H(u) = \log(N)$. For large $N$ or $a$, the average entropy is close to the maximum value. It follows that in these cases the expected divergence from the uniform distribution $u$ remains bounded. The fact that the expected entropy is close to the maximal entropy makes it difficult to estimate the entropy. See [7] for a discussion and possible solutions.

**Theorem 2.**

- *For any* $q \in \Delta_{N-1}$, *when* $p \sim \mathrm{Dir}_{\boldsymbol{\alpha}}$, *then*

$$\langle D(p\|q) \rangle = \sum_{i=1}^{N} \frac{\alpha_i}{\alpha} (h(\alpha_i) - \log(q_i)) - h(\alpha) \ .$$

*If $\boldsymbol{\alpha} = (a, \dots, a)$, then this becomes*

$$\langle D(p\|q)\rangle = \log(N) - h(aN) + h(a) + D(q\|u)\ .$$

*When $p \sim \mathrm{Dir}_{\boldsymbol{\alpha}}$ and $q \sim \mathrm{Dir}_{\tilde{\boldsymbol{\alpha}}}$, then*

- $\langle \sum_{i\in\mathcal{X}} p_i \log(q_i)\rangle = \sum_{i=1}^{N} \frac{\alpha_i}{\alpha} h(\tilde{\alpha}_i - 1) - h(\tilde{\alpha} - 1)$,
- $\langle D(p\|q)\rangle = -\sum_{i=1}^{N} \frac{\alpha_i}{\alpha}(h(\tilde{\alpha}_i - 1) - h(\alpha_i)) + h(\tilde{\alpha} - 1) - h(\alpha)$.

*If $\boldsymbol{\alpha} = \tilde{\boldsymbol{\alpha}}$, then $\langle D(p\|q)\rangle = \frac{N-1}{\alpha}$.*

- *For any $q \in \Delta_{N-1}$, when $p$ is drawn uniformly from $\Delta_{N-1}$, then*

$$\langle D(p\|q)\rangle = -\sum_{i=1}^{N} \frac{1}{N}\log(q_i) - h(N) + 1 = D(u\|q) + 1 - \gamma + O(1/N)\ .$$

The divergence is unbounded in $\Delta_{N-1} \times \Delta_{N-1}$, since $D(p\|q) = +\infty$ if $p$ is not absolutely continuous with respect to $q$. Nevertheless, if $p, q \sim \mathrm{Dir}_{\boldsymbol{\alpha}}$, then in the limit $N \to \infty$ the expected divergence $\langle D(p\|q)\rangle$ remains bounded, provided $\frac{1}{N}\sum_{i=1}^{N} \alpha_i = \alpha/N$ is bounded from below by a positive constant.

Consider a sequence of distributions $q_N \in \Delta_{N-1}$, $N \in \mathbb{N}$. As $N \to \infty$ the expected divergence $\langle D(\cdot\|q_N)\rangle$ with respect to the uniform prior is bounded from above by $1 - \gamma + \varepsilon$, $\varepsilon > 0$ if and only if $\limsup_{N\to\infty} D(u\|q_N) \le \varepsilon$. If $q_x \ge \frac{1}{N}e^{-\varepsilon}$ for all $x \in \mathcal{X}$, then $D(u\|q) \le \varepsilon$. Therefore, the expected divergence $\langle D(\cdot\|q_N)\rangle$ is unbounded only if the sequence $q_N$ accumulates at the boundary of the probability simplex, and $\lim_{N\to\infty}\langle D(p\|q_N)\rangle \le 1 - \gamma + \varepsilon$ whenever $q_N$ is in the subsimplex $\mathrm{conv}\{(1 - e^{-\varepsilon})\delta_x + e^{-\varepsilon}u\}_{x\in\mathcal{X}}$. The relative Lebesgue volume of this subsimplex in $\Delta_{N-1}$ is $(1 - e^{-\varepsilon})^{N-1}$.

**Theorem 3.** *Consider a composite system of $n$ random variables $X_1, \dots, X_n$ with joint probability distribution $p$. If $p \sim \mathrm{Dir}_{\boldsymbol{\alpha}}$, then*

- $\langle H(X_k)\rangle = h(\alpha) - \sum_{j=1}^{N_k} \frac{\alpha_j^k}{\alpha} h(\alpha_j^k)$,
- $\langle MI(X_1, \dots, X_n)\rangle = (n-1)h(\alpha) + \sum\limits_{i=1}^{N} \frac{\alpha_i}{\alpha} h(\alpha_i) - \sum\limits_{k=1}^{n}\sum\limits_{j=1}^{N_k} \frac{\alpha_j^k}{\alpha} h(\alpha_j^k)$.

*If $(\alpha_1, \dots, \alpha_N) = (a, \dots, a)$ (symmetric Dirichlet),*

- $\langle H(X_k)\rangle = h(Na) - h(\frac{N}{N_k}a)$,
- $\langle MI(X_1, \dots, X_n)\rangle = (n-1)h(Na) + h(a) - \sum_{k=1}^{n} h(\frac{N}{N_k}a)$.

*If, moreover, $Na/N_k$ is large for all $k$ (this happens, for example, when $a$ remains bounded from below by some $\varepsilon > 0$ and* (i) *all $N_k$ become large, or* (ii) *all $N_k$ are bounded and $n$ becomes large), then:*

- $\langle H(X_k)\rangle = \log(N_k) + O(N_k/Na)$,

- $\langle MI(X_1,\dots,X_n)\rangle = h(a) - \log(a) - \gamma + O(n \max_k N_k/Na)$.

If $Na/N_k$ is large for all $k$, then the expected entropy of a subsystem is also close to its maximum, and hence the expected multi-information is bounded. This follows also from the fact that the independence model contains the uniform distribution, and hence $D(p\|\mathcal{M}_1) \leq D(p\|u)$.

**Theorem 4.** *Let $\varrho = \{A_1,\dots,A_K\}$ be a partition of $\mathcal{X}$ into sets of cardinalities $|A_k| = L_k$, and let $\nu$ be a reference measure on $\mathcal{X}$. If $p \sim \mathrm{Dir}_{\boldsymbol{\alpha}}$, then*

$$\langle D(p\|\mathcal{M}_{\varrho,\nu})\rangle = \sum_{i=1}^{N} \frac{\alpha_i}{\alpha}(h(\alpha_i) - \log(\nu_i)) - \sum_{k=1}^{K} \frac{\alpha_k^{\varrho}}{\alpha}(h(\alpha_k^{\varrho}) - \log(\nu(A_k))),$$

*where $\alpha_k^{\varrho} = \sum_{i\in A_k} \alpha_i$. If $\boldsymbol{\alpha} = (a,\dots,a)$, and (wlog) $\nu(A_k) = L_k/N$,*

$$\langle D(p\|\mathcal{M}_{\varrho,\nu})\rangle = h(a) - \sum_{k=1}^{K} \frac{L_k}{N}(h(L_k a) - \log(L_k)) + D(u\|\nu),$$

*If furthermore $N \gg K$, then*

$$\langle D(p\|\mathcal{M}_{\varrho,\nu})\rangle = h(a) - \log(a) - \gamma + D(u\|\nu) + O(1/N).$$

Partition models (with $\nu = u$) also contain the uniform distribution, and therefore the expected divergence is again bounded. In contrast, the maximal divergence is $\max_{p\in\Delta_{N-1}} D(p\|\mathcal{M}_{\varrho}) = \max_k \log(N_k)$. The result for mixtures of product distributions of disjoint supports is similar:

**Theorem 5.** *Let $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ be the joint state space of $n$ variables, $|\mathcal{X}| = N$, $|\mathcal{X}_k| = N_k$. Let $\varrho = \{A_1,\dots,A_K\}$ be a partition of $\mathcal{X}$ into support sets of the independence model of cardinalities $|A_k| = L_k$, and let $\mathcal{M}_{1,\varrho}^{K}$ be the model containing all mixtures of $K$ product distributions $p^{(1)},\dots,p^{(K)}$ with $\mathrm{supp}(p^{(k)}) \subseteq A_k$.*

- *If $p \sim \mathrm{Dir}_{(\alpha_1,\dots,\alpha_N)}$, then the expected divergence to $\mathcal{M}_{1,\varrho}^{K}$ is*

$$\langle D(p\|\mathcal{M}_{1,\varrho}^{K})\rangle = \sum_{i=1}^{N} \frac{\alpha_i}{\alpha}(h(\alpha_i) - h(\alpha)) + \sum_{k=1}^{K}(|G_k| - 1)\frac{\alpha_k^{\varrho}}{\alpha}(h(\alpha_k^{\varrho}) - h(\alpha))$$
$$- \sum_{k=1}^{K} \sum_{j\in G_k} \sum_{x_j\in\mathcal{X}_{j,k}} \frac{\alpha^{k,x_j}}{\alpha}(h(\alpha^{k,x_j}) - h(\alpha)),$$

  *where $\alpha_k^{\varrho} = \sum_{x\in A_k} \alpha_x$, $\alpha^{k,x_j} = \sum_{y\in A_k:\, y_j = x_j} \alpha_y$, and $G_k \subset [n]$ is the set of variables that take more than one value in the block $A_k$.*

- *Assume that the system is homogeneous $|\mathcal{X}_i| = N_1$ for all $i$ and that, for each $k$, $A_k$ is a cylinder set of cardinality $|A_k| = N_1^{m_k}$, where $m_k = |G_k|$. If $(\alpha_1,\dots,\alpha_N) = (a,\dots,a)$, then*

$$\langle D(p\|\mathcal{M}_{1,\varrho}^{K})\rangle = h(a) + \sum_{k=1}^{K} N_1^{m_k - n}((m_k - 1)h(N_1^{m_k} a) - m_k h(N_1^{m_k - 1} a)).$$

- *If $\frac{N_1^{m_k-1}a}{m_k}$ is large for all $k$, then*

$$\langle D(p\|\mathcal{M}^K_{1,\varrho})\rangle = h(a) - \log(a) - \gamma + O\big(\max_k \frac{m_k}{N_1^{m_k-1}a}\big).$$

The $k$-mixture of binary product distributions with disjoint supports is contained in the restricted Boltzmann machine model with $k-1$ hidden nodes, see [6]. Hence Theorem 5 gives bounds for the expected divergence to these models.

**Theorem 6.** *For a decomposable model $\mathcal{M}_{\mathcal{S}}$ with junction tree $(V,E)$, if $p \sim \mathrm{Dir}_{(\alpha_1,\dots,\alpha_N)}$, then*

$$\langle D(p\|\mathcal{M}_{\mathcal{S}})\rangle = -\sum_{S\in V}\sum_{j\in\mathcal{X}_S}\frac{\alpha_j^S}{\alpha}h(\alpha_j^S) + \sum_{S\in E}\sum_{j\in\mathcal{X}_S}\frac{\alpha_j^S}{\alpha}h(\alpha_j^S) \\ + (|V|-|E|-1)h(\alpha) + \sum_{i=1}^N \frac{\alpha_i}{\alpha}h(\alpha_i),$$

*where $\alpha_j^S = \sum_{x\colon x_S=j}\alpha_x$ for $j\in\mathcal{X}_S$. If $p$ is drawn uniformly at random, then*

$$\langle D(p\|\mathcal{M}_{\mathcal{S}})\rangle = \sum_{S\in V}(h(N)-h(N/N_S)) - \sum_{S\in E}(h(N)-h(N/N_S)) - h(N) + 1.$$

*If $N/N_S$ is large for all $S\in V\cup E$, then*

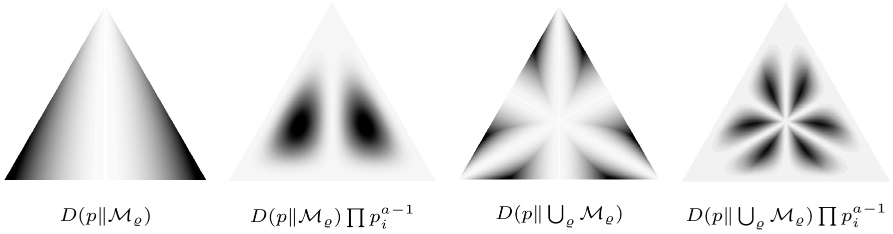
$$\langle D(p\|\mathcal{M}_{\mathcal{S}})\rangle = 1 - \gamma + O\big(\max_k \frac{m_k}{N_1^{m_k-1}a}\big).$$

# 4 Discussion

In the previous section we have shown that the values of $\langle D(p\|\mathcal{M})\rangle$ are very similar for different models $\mathcal{M}$ in the limit of large $N$, provided the Dirichlet parameters $\alpha_i$ remain bounded and the model remains "small." In particular, if $\alpha_i = 1$ for all $i$, then $\langle D(p\|\mathcal{M})\rangle \approx 1-\gamma$ holds for large $N$ and $\mathcal{M} = \{u\}$, for the independence model, for decomposable models, for partition models and for mixtures of product distributions on disjoint supports (for reasonable values of the model parameters $N_k$ and $L_k$). Some of these models are contained in each other, but nevertheless, the expected divergences do not differ too much. The general phenomenon seems to be the following:

- For a low-dimensional model $\mathcal{M}\subset\Delta_{N-1}$ and large $N$, the expected divergence is $\langle D(p\|\mathcal{M})\rangle \approx 1-\gamma$, when $p$ is uniformly distributed on $\Delta_{N-1}$.

Of course, this is not a mathematical statement, because it is very easy to construct counter-examples: Using space-filling curves, it is possible to construct one-dimensional models $\mathcal{M}$ with an arbitrary low value of $\langle D(p\|\mathcal{M})\rangle$ (for arbitrary $N$). However, we expect that the statement is true for most models that appear in practice. In particular, we conjecture that the statement is true for restricted Boltzmann machines.



$D(p\|\mathcal{M}_\varrho)$ $D(p\|\mathcal{M}_\varrho)\prod p_i^{a-1}$ $D(p\|\bigcup_\varrho \mathcal{M}_\varrho)$ $D(p\|\bigcup_\varrho \mathcal{M}_\varrho)\prod p_i^{a-1}$

Figure 1: From left to right: Divergence to a partition model with two blocks on $\mathcal{X} = \{1, 2, 3\}$. Same, multiplied by a symmetric Dirichlet density with parameter $a = 5$. Divergence to the union of the three partition models with two blocks on $\mathcal{X} = \{1, 2, 3\}$. Same, multiplied by the symmetric Dirichlet density with $a = 5$. The shading is scaled on each image individually.

In Theorem 4, if $\alpha = (a, \ldots, a)$, then the expected divergence from $\mathcal{M}_{\varrho,\nu}$ is minimal, if and only if $\nu = u$. In this case $\mathcal{M}_{\varrho,\nu}$ is a partition model. We conjecture that partition models are optimal among all (closures of) exponential families in the following sense:

- For any exponential family $\mathcal{E}$ there is a partition model $\mathcal{M}$ of the same dimension such that $\langle D(p\|\mathcal{E})\rangle \geq \langle D(p\|\mathcal{M})\rangle$.

The statement is, of course, true for zero-dimensional exponential families, i.e., models that consist of a single distribution. The conjecture is related to the following conjecture from [9]:

- For any exponential family $\mathcal{E}$ there is a partition model $\mathcal{M}$ of the same dimension such that $\max_{p\in\Delta_{N-1}} D(p\|\mathcal{E}) \geq \max_{p\in\Delta_{N-1}} D(p\|\mathcal{M})$.

Our findings may be biased by the fact that all the models treated in Section 3 are examples of exponential families. As a slight generalization we did computer experiments with a family of models which are not exponential families, but unions of exponential families.

Let $\Upsilon$ be a family of partitions, and let $\mathcal{M}_\Upsilon = \bigcup_{\varrho\in\Upsilon} \mathcal{M}_\varrho$ be the union of the corresponding partition models. Our interest in these models comes from the fact that such models are contained in more difficult models with hidden variables, like restricted Boltzmann machines and deep belief networks. Figure 1 compares a single partition model on three states with the union of all partition models for bipartitions.

For a given $N$ and $0 \leq k \leq N/2$ let $\Upsilon_k$ be the set of all partitions of $\{1, \ldots, N\}$ into two blocks of cardinalities $k$ and $N - k$. For different values of $a$ and $N$ we computed $D(p\|\mathcal{M}_{\Upsilon_1})$ for 10 000 distributions sampled from $\mathrm{Dir}_{(a,\ldots,a)}$, $D(p\|\mathcal{M}_{\Upsilon_2})$ for 20 000 distributions sampled from $\mathrm{Dir}_{(a,\ldots,a)}$, and $D(p\|\mathcal{M}_{\Upsilon_{N/2}})$ for 20 000 distributions sampled from the uniform prior. The results are shown in Figure 2.

In the first two cases the expected divergence seems to tend to the asymptotic value of $\langle D(p\|u)\rangle$. Observe that $\langle D(p\|\mathcal{M}_{\Upsilon_1})\rangle \geq \langle D(p\|\mathcal{M}_{\Upsilon_2})\rangle$, unless $N = 4$. Intuitively

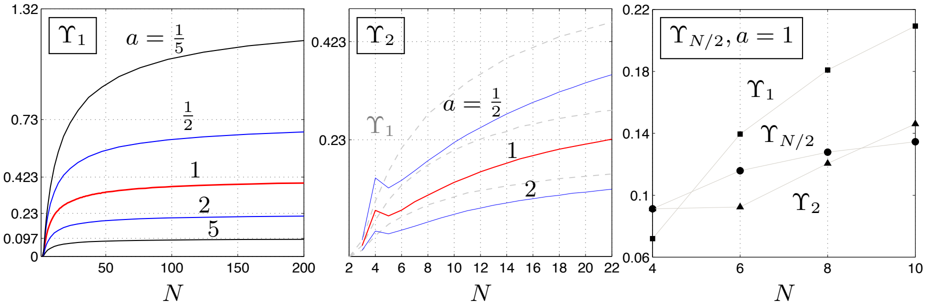


Figure 2: Expected divergence (numerically) from various unions of bipartition models with respect to $\mathrm{Dir}_{(a,\dots,a)}$, for different system sizes $N$ and values of the concentration parameter $a$. Left: Union of all bipartition models with blocks of cardinalities 1 and $(N-1)$. The y-ticks are located at $h(a) - \log(a) - \gamma$, which are the limits of the expected divergence from single bipartition models, see Theorem 4. Middle: Union of all bipartition models with blocks of cardinalities 2 and $(N-2)$. The peak at $N = 4$ is caused by the fact that there are only 3 different partitions when $N = 4$, instead of $\binom{N}{2}$. The dashed plot indicates corresponding results from the left figure. Right: Comparison of the expected divergence from the two previous models and the union of all $\binom{N}{N/2}/2$ bipartition models with two blocks of cardinalities $N/2$, for $a = 1$ and even $N$.

this makes sense for two reasons: First, for $\varrho_1 \in \Upsilon_1$ and $\varrho_2 \in \Upsilon_2$, using Theorem 4 one can show that $\langle D(p\|\mathcal{M}_{\varrho_1})\rangle \geq \langle D(p\|\mathcal{M}_{\varrho_2})\rangle$; and second, the cardinality of $\Upsilon_2$ is much larger than the cardinality of $\Upsilon_1$ if $N \geq 4$. For small values of $N$ this intuition may not always be correct. For example, for $N = 8$, the expected divergence from $\mathcal{M}_{\Upsilon_{N/2}}$ is larger than the one from $\mathcal{M}_{\Upsilon_2}$, although in this case $|\Upsilon_{N/2}| = 35$ and $|\Upsilon_2| = 28$, see Figure 2 right.

For $N = 22$ we computed $D(p\|\mathcal{M}_{\Upsilon_{N/2}})$ for 500 uniformly sampled distributions (in this case $|\Upsilon_{N/2}| = 352\,716$), and found $\langle D(p\|\mathcal{M}_{\Upsilon_{N/2}})\rangle \approx 0.1442$ (with variance 0.0032), which is well below the corresponding expectation values for $\mathcal{M}_{\Upsilon_1}$ and $\mathcal{M}_{\Upsilon_2}$. We expect that, for large $N$, it is possible to make $\langle D(p\|\mathcal{M}_{\Upsilon_k})\rangle$ much smaller than $\langle D(p\|u)\rangle$ by choosing $k \approx N/2$. In this case, the model $\mathcal{M}_{\Upsilon_k}$ has (Hausdorff) dimension only one, but it is a union of exponentially many one-dimensional exponential families.

# A Computations and proofs

The analytic formulas in Theorem 1 are [10, Theorem 7]. The asymptotic expansions are direct.

The proof of Theorem 2 makes use of the following Lemma, see [10, Theorem 3]:

**Lemma 7.** *Let $\{A_1,\dots,A_K\}$ be a partition of $\mathcal{X} = \{1,\dots,N\}$, let $\alpha_1,\dots,\alpha_N$ be* *positive reals, and let* $\alpha^k = \sum_{i\in A_k}\alpha_i$ *for* $k = 1,\dots,K$. *Then*

$$\int_{\Delta_{N-1}} \big(\sum_{i\in A_k} p_i\big)\log\big(\sum_{i\in A_k} p_i\big)\prod_{i=1}^N p_i^{\alpha_i-1}\,\mathrm{d}p = \int_{\Delta_{K-1}} p_k^*\log(p_k^*)\prod_{k'=1}^K (p_{k'}^*)^{\alpha^{k'}-1}\,\mathrm{d}p^*$$

$$= \frac{\alpha^k\prod_{k'=1}^K\Gamma(\alpha^{k'})}{\Gamma(\alpha+1)}(h(\alpha^k)-h(\alpha))\ .$$

*Proof of Theorem 2.* The first statement follows from

$$\int_{\Delta_{N-1}}\log(q_i)p_i\prod_i p_i^{n_i}\,\mathrm{d}p\Big/\int_{\Delta_{N-1}}\prod_i p_i^{n_i}\,\mathrm{d}p = \log(q_i)\frac{(n_i+1)}{(N+n)}$$

and $D(p\|q) = -H(p) - \sum_i p_i\log(q_i)$. By Lemma 7,

$$\int_{\Delta_{N-1}}\log(q_i)\prod_i q_i^{n_i}\,\mathrm{d}q\Big/\int_{\Delta_{N-1}}\prod_i q_i^{n_i}\,\mathrm{d}q = h(n_i) - h(N+n-1)\ ,$$

and the remaining statements follow. □

Theorem 3 is a corollary to Theorem 1, the aggregation property of the Dirichlet priors and the formula (1) for the multi-information. Theorem 4 follows from (2), and Theorem 6 follows from (3). Similarly, Theorem 5 follows from the equality

$$D(p\|\mathcal{M}_0) = \sum_{i=1}^K\sum_{x\in A_i} p(x)\log\frac{p(x)p(A_i)^{n-1}}{\prod_{j=1}^n(\sum_{y\in A_i:y_j=x_j}p(y))}\ ,$$

which can be derived as follows: The unique solution $q\in\operatorname{arginf}_{q'\in M_{1,\varrho}^K} D(p\|q')$ satisfies $p(A_i) = q(A_i)$, and $q(\cdot|A_i)\in\operatorname{arginf}_{q'\in\mathcal{M}_1} D(p(\cdot\|A_i)\|q')$.

**Acknowledgment:** G. F. Montúfar is supported in part by DARPA grant FA8650-11-1-7145. J. Rauh is supported by the VW Foundation.

## References

[1] N. Ay, "An information-geometric approach to a theory of pragmatic structuring," *Annals of Probability*, vol. 30, pp. 416–436, 2002.

[2] M. Drton, B. Sturmfels, and S. Sullivant, *Lectures on Algebraic Statistics*, 1st ed., ser. Oberwolfach Seminars. Birkhuser, Basel, 2009, vol. 39.

[3] B. A. Frigyik, A. Kapila, and M. R. Gupta, "Introduction to the Dirichlet distribution and related processes," Department of Electrical Engineering University of Washington, Tech. Rep., 2010.

[4] F. Matúš and N. Ay, "On maximization of the information divergence from an exponential family," in *Proceedings of the WUPES'03*. University of Economics, Prague, 2003, pp. 199–204.

[5] F. Matúš and J. Rauh, "Maximization of the information divergence from an exponential family and criticality," in *2011 IEEE International Symposium on Information Theory Proceedings (ISIT2011)*, 2011.

[6] G. F. Montúfar, J. Rauh, and N. Ay, "Expressive power and approximation errors of restricted Boltzmann machines," in *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, 2011, pp. 415–423, available at `http://books.nips.cc/papers/files/nips24/NIPS2011_0307.pdf`.

[7] I. Nemenman, F. Shafee, and W. Bialek, "Entropy and inference, revisited," in *NIPS*, 2001, pp. 471–478.

[8] J. Rauh, "Finding the maximizers of the information divergence from an exponential family," Ph.D. dissertation, Universität Leipzig, 2011.

[9] ——, "Optimally approximating exponential families," *submitted*, 2012, available at `http://arxiv.org/abs/1111.0483`.

[10] D. Wolpert and D. Wolf, "Estimating functions of probability distributions from a finite set of samples," *Physical Review E*, vol. 52, no. 6, pp. 6841–6854, 1995.

# A New Common Cause Principle for Bayesian Networks

**Philipp Moritz and Jörg Reichardt**
Institute for Theoretical Physics
University of Würzburg, Germany
pcmoritz@googlemail.com, reichardt@physik.uni-wuerzburg.de

**Nihat Ay**
MPI for Mathematics in the Sciences
Inselstraße 22, Leipzig, Germany
nay@mis.mpg.de

**Abstract**

By postulating the local Markov condition, Bayesian networks can be used to describe conditional indepencence among different parts of a system. If we observe only a subset of variables $Y_1, Y_2, \dots, Y_k$ of a system with an unknown underlying Bayesian network, the existence of a common ancestor for more than $c$ nodes out of $Y_1, Y_2 \dots, Y_k$ may be inferred if a particular information theoretic quantity $I_c(Y_1, \dots, Y_k)$ is positive, as shown by Steudel and Ay (2010). We extend this common cause principle so that it allows a more fine grained discrimination between different causal hypotheses. Our main result is an upper bound on $I_c$ for a given causal structure and the proof that this bound is tight and achieved.

## 1 Introduction

Finding and describing causal relations between parts of a system is a fundamental problem in many scientific disciplines. Examples for this task are

(i) in biology the study of genetic data from pedigrees, where the whole system is the considered family, the parts are the individuals and the causal relations are given by the inheritance structure [4] and

(ii) in the social sciences the analysis of a group of people as the system, with parts of the system being the members of the group and the "causal relations" describing the interactions amongst them.

To deal with such relationships, different types of probabilistic graphical models have been introduced [3], [9]. They describe a system as a graph, whose vertices are the parts of the system. The edges correspond to causal relations between these parts. Such models allow an efficient and natural representation of knowledge and are easy to visualize and manipulate. Bayesian networks use directed graphs and have proven especially valuable, their directed nature makes them particularly suited for the description of causes and effects [6].

Ideally, a scientist will construct a causal model of a system by systematic intervention and subsequent observation of the parts. Often, this is not possible and causality has to be determined by means of observation alone, using methods of statistical inference. Statistical dependencies can be interpreted as causal relations if one employs additional postulates such as the *causal Markov condition* [6]. Once this link between statistical dependence and causal relation is established, tools from information theory prove most valuable for the quantification of correlation and thus causal relations between variables [1].

A system can be described by a Bayesian network that encodes the relations between parts of the system. These parts are modeled by random variables $X_1, \dots, X_n$. Assume that a subset $Y_1, \dots, Y_k$ of the $X_1, \dots, X_n$ is observed and their correlation, measured by a certain information theoretic quantity, exceeds a bound that depends on a number $c$. Then, as shown in [8], in *any* Bayesian network containing $Y_1, \dots, Y_k$, there exists a common ancestor for a subset of $c$ variables out of the $Y_1, \dots, Y_k$. We build on this result and present a new common cause principle that takes the structure of the Bayesian network into account and allows to discriminate between different causal hypotheses.

The paper is organized as follows: In section 2 we formulate a mathematical model for the kind of systems that we study. In section 3 we summarize work that has already been done on the inference of common ancestors. Section 4 is the main part of the paper and describes our new common cause principle. In section 5 we then conclude what has been achieved.

## 2 Describing systems with Bayesian networks

In this section, we define in a rigorous way our notion of a system that was described in the introduction.

**Directed acyclic graphs.** A *directed acyclic graph* (DAG) is a tuple $G = (V, E)$ consisting of *nodes* $V$ and *edges* $E \subseteq V \times V$ that have to fullfil the additional constraint of *acylicity*, specified below. An edge $(u, v) \in E$ is interpreted as a directed connection between the nodes $u$ and $v$, we write $u \to v$ in this case. A *directed path* between two nodes $v_1$ and $v_n$ is a sequence $v_1, v_2, \dots, v_n$ of distinct nodes $v_j$ with $v_j \to v_{j+1}$ for $1 \leq j < n$. We write $v_1 \rightsquigarrow v_n$ if there exists a directed path from $v_1$ to $v_n$. We also admit paths of length 0, so $v \rightsquigarrow v$ for all $v \in V$. An *undirected path* between $v_1$ and $v_n$ is a sequence $v_1, v_2, \dots, v_n$ of distinct nodes $v_j$ with $v_j \to v_{j+1}$ or $v_j \leftarrow v_{j+1}$ for $1 \leq j < n$. We call $G$ *acyclic*, if there is no sequence $v \to \cdots \to v$ for any node $v \in V$.

The set of *parents* of a given node $v \in V$, denoted by $\mathrm{pa}(v) = \{u \in V : (u, v) \in E\}$, contains those nodes that point directly towards $v$. The nodes that have no parents are collected in the set of *root nodes* $\mathrm{roots}(G) = \{v \in G : \mathrm{pa}(v) = \varnothing\}$ of $G$. The *descendants* $\mathrm{de}(u)$ of a node $u \in V$ are the nodes $v \in V$ such that $u \rightsquigarrow v$, thus every $v \in V$ is a descendant of itself. The *non-descendants* of $u \in V$ are given by

$\mathrm{nd}(u) = V \setminus \mathrm{de}(u)$ and the *ancestral set* $\mathrm{an}(v)$ consists of all nodes $u \in V$ with $u \rightsquigarrow v$.[1] The condition of acyclicity means that $\mathrm{de}(v) \cap \mathrm{an}(v) = \varnothing$ for every $v \in V$, which is a natural postulate for a causal model. These concepts are illustrated in Fig. 1.

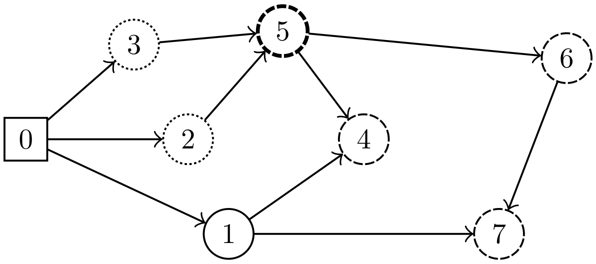


**Fig. 1.** For this example graph $G$ we have $\mathrm{roots}(G) = \{0\}$, the only root is the square node. Consider node 5, its parents are $\mathrm{pa}(5) = \{2, 3\}$, these are dotted. The set of descendants is $\mathrm{de}(5) = \{4, 5, 6, 7\}$, which are dashed.

**Bayesian networks.** Let $G = (V, E)$ be a DAG and $\{X_v : v \in V\}$ a set of random variables, one variable for each node of $G$ (we will sometimes identify nodes with their random variables). How can we encode the conditional independence structure of the joint probability distribution that underlies the random variables $\{X_v : v \in V\}$ in the graph $G$? In the kind of directed models that we study, this can be done as follows.

We call $X = (X_v : v \in V)$ a *Bayesian network* with respect to $G$, if the joint probability function $p$ satisfies

$$p(x) = \prod_{v \in V} p(x_v \mid x_{\mathrm{pa}(v)}). \tag{1}$$

Here, for $A \subseteq V$, the tuple $x_A$ is defined as $x_A = (x_a : a \in A)$. This is the so called *factorization definition* of the Bayesian network and is best interpreted using the equivalent [3] *local Markov condition.* The local Markov condition states that each variable $v \in V$ is conditionally independent of its non-descendants $\mathrm{nd}(v)$ given its parent variables. We write this symbolically as

$$X_v \perp\!\!\!\perp X_{\mathrm{nd}(v)} \mid X_{\mathrm{pa}(v)} \qquad \text{for all } v \in V. \tag{2}$$

This means $p(x_v, x_{\mathrm{nd}(v)} \mid x_{\mathrm{pa}(v)}) = p(x_v \mid x_{\mathrm{pa}(v)}) \cdot p(x_{\mathrm{nd}(v)} \mid x_{\mathrm{pa}(v)})$.

Another but equivalent [3] version of this definition is the *global Markov condition* which tells us for three disjoint sets of nodes $A$, $B$ and $C$ under which condition $X_A$ is independent of $X_B$ given $X_C$, written as $X_A \perp\!\!\!\perp X_B \mid X_C$.

To formulate the global Markov condition, we first of all introduce the concept of $d$-seperation [6]. An undirected path $\gamma$ in $G$ is *d-separated* by a set of nodes $C$ if and only if

[1] Note that $v \in \mathrm{an}(v)$, we do not follow the standard definition of the set of ancestors from [3] here, but rather stick to the definition of an ancestral set given in [1], because the latter definition is more natural in the context of common ancestors that we will study later. The reason for this is that the theorem from [8] allows a common ancestor of nodes $v_1, \ldots, v_n$ to be one of the $v_1, \ldots, v_n$ and intuitively one would expect the common ancestor of $v_1, \ldots, v_n$ to be in the set $\bigcap_{1 \leq j \leq n} \mathrm{an}(v_j)$.

- $\gamma$ contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node $m$ is in $C$, or
- $\gamma$ contains a collider $i \to m \leftarrow j$ such that the middle node $m$ is not in $C$ and such that no descendant of $m$ is in $C$.

The set $C$ does $d$-separate the sets $A$ and $B$ if and only if every undirected path between a node in $A$ and a node in $B$ is $d$-separated by $C$. The global Markov condition then states that $X_A \perp\!\!\!\perp X_B \mid X_C$ whenever $C$ $d$-separates $A$ and $B$.

**Partially observed systems.** A *system* is a fixed Bayesian network $G$ of binary random variables $X_1, \dots, X_n$. A subset $Y_1, \dots, Y_k$ of these variables is *observed*, that means the probability distribution of the $Y_1, \dots, Y_k$ is the probability distribution of the $X_1, \dots, X_n$ with the unobserved variables marginalized out, so

$$p(Y_1, \dots, Y_k) = \sum_A p(X_1, \dots, X_n) \quad \text{with } A = \{X_1, \dots, X_n\} \setminus \{Y_1, \dots, Y_k\}.$$

These definitions are illustrated in Fig. 2. A *common cause* or *common ancestor* of the observed nodes $Y_1, \dots, Y_k$ is then a node $X_j$ with

$$X_j \in \bigcap_{1 \le j \le k} \mathrm{an}(Y_j).$$

In our example of Fig. 2, $X_2$ would be a common ancestor of $X_2$, $X_4$ and $X_6$.

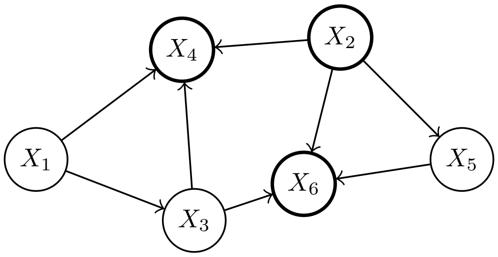


**Fig. 2.** Example of a system with $n = 6$ nodes, of which the $k = 3$ nodes $Y_1 = X_2$, $Y_2 = X_4$ and $Y_3 = X_6$ are observed.

## 3 Inference of common ancestors

In this section we briefly summarize how information theory can be used to get clues about the structure of a Bayesian network underlying a given system. The advantage of these methods is that they do not need intervention into the system but can be applied to observation data alone. First we discuss *Reichenbach's principle of common cause* which is the most elementary common cause principle.

**Reichenbach's principle of common cause.** Reichenbach's principle of common cause is the simplest example for a device that allows us to study the causal structure of a system by observation alone. Reichenbach formulates it in [7] as *"If an improbable coincidence has occured, there must exist a common cause"*. As an example, he considers a fire started by lightning and spread by the wind. This coincidence can be explained by a common cause, the thunderstorm that produced the lightning and the wind.

A more formal version of the principle states that if we observe that two jointly distributed random variables $X$ and $Y$ are dependent, then one of the following must be true: $X$ causes $Y$ or $Y$ causes $X$ or there is a common cause of $X$ and $Y$. In our framework, we can understand Reichenbach's principle in the following way, as noted by [8]. If $X$ and $Y$ are part of a larger system, modeled by a Bayesian network with underlying graph $G$ and they are stochastically dependent, then their ancestral sets must be overlapping, otherwise they would be $d$-separated by the empty set and thus be independent.

**The extended common cause principle.** We will now turn to a quantitative extension of the common cause principle, initially studied in [1] and later extended in [8]. Assume that we have a system with variables $X_1, \dots, X_n$ which form a Bayesian network. Out of these variables, a subset $Y_1, \dots, Y_k$ is observed. On these, we define the *generalized mutual information of degree* $c$ as

$$I_c(Y_1, \dots, Y_k) = \frac{1}{c} \sum_{j=1}^{k} H(Y_j) - H(Y_1, \dots, Y_k). \tag{3}$$

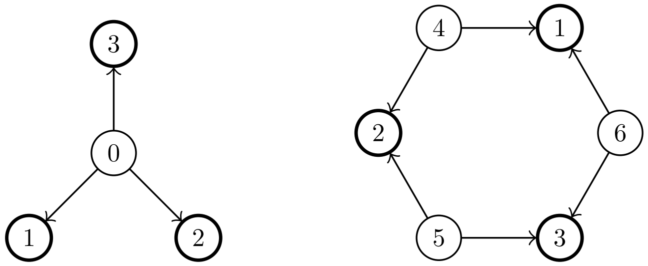
In the case $c = 1$, this is the mutual information from [2]. The quantity $I_c$ is a measure of correlation of the $Y_1, \dots, Y_k$ and allows the following quantitative extension of Reichenbach's principle of common cause, proven in [8].

**Theorem 1** (Extended Common Cause Principle)**.** *Let $X_1, \dots, X_n$ be a system with observed variables $Y_1, \dots, Y_k$. If $I_c(Y_1, \dots, Y_k) > 0$ then in* any *system containing the $Y_1, \dots, Y_k$, there exists a common ancestor of strictly more than $c$ variables out of the $Y_1, \dots, Y_k$.*

The importance of this extended common cause principle stems from the fact that it allows the discrimination between different causal models for a system by observation alone, even where Reichenbach's common cause principle would fail. In Fig. 3 we show two systems from [8] where this is the case. The Reichenbach principle cannot distinguish between (a) and (b), because in both models the observed variables $X_1$, $X_2$ and $X_3$ are not necessarily independent. If we have however $I_2(X_1, X_2, X_3) > 0$, then model (b) can be refused on grounds of the extended common cause principle, because it does not contain a common ancestor of 3 nodes.

# 4 A new common cause principle

In this section, we describe a new common cause principle that is sometimes even stronger than the one from [8]. This principle works by exploiting the maximum



**(a)** one common ancestor of three variables

**(b)** three common ancestors of two varibles each

**Fig. 3.** Two possible Bayesian networks that could underly the observed variables $X_1$, $X_2$ and $X_3$ (observed variables are drawn thick, unobserved ones thin). The Reichenbach principle of common cause cannot discriminate these.

of the generalized mutual information (3) over all probability distributions of the $X_1, \ldots, X_n$ compatible with a given directed graph $G$ is, where the subset $Y_1, \ldots, Y_k$ of the $X_1, \ldots, X_n$ is observed.

Before we state the result, we want to illuminate the problem with two example networks, see Fig 4. We generated a symbolic expression for $I_2$ and used a computer algebra system to maximize $I_2$. There are more than ten variables involved, so the optimization task is not trivial and sometimes we ran into local maxima. For large graphs, this procedure is not tractable, so a better understanding of the problem is crucial. For the example in Fig. 4 (a), the maximum $I_2 = (3/2) \cdot \log 2$ was achieved with

$$P(X_0 = 1) = 1/2, \quad X_1 = X_2 = X_3 = X_4 = X_0,$$

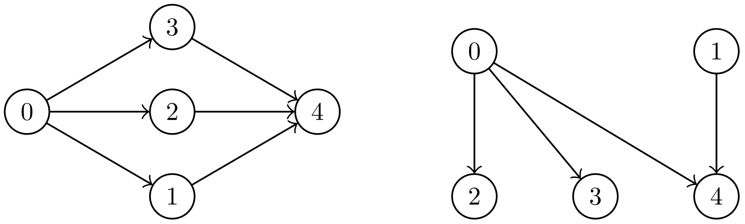
where the last equality means that the conditional probabilities are chosen in such a way that $X_1 = X_2 = X_3 = X_4 = X_0$. This can be achieved by setting $p(x_j = 1 \mid x_0 = 1) = 1$, $p(x_j = 1 \mid x_0 = 0) = 0$ for $1 \leq j \leq 3$ and $p(x_4 = 1 \mid x_1 = x_2 = x_3 = 1) = 1$, $p(x_4 = 1 \mid x_1 = x_2 = x_3 = 0) = 0$. For the example in Fig. 4 (b), we obtain $I_2 = \log 2$ with

$$P(X_0 = 1) = 1/2, \quad P(X_1 = 1) = 0, \quad X_2 = X_3 = X_4 = X_0.$$

Both examples refer to the fully observed case. We will now explain these results in the general case.

In the following theorem and its proof, the set of descendents of a random variable $N$ (which we identify with the corresponding node in the graph) will again play an important role. The important concepts that are needed in the following theorem are summarized in Fig. 5.

We choose the indices in such a way that $X_1, \ldots, X_s$ are the roots. For $1 \leq j \leq s$, the set $A_j$ consists of all the observed nodes from $\mathrm{de}(X_j)$. These can be overlapping. Now the first $s$ indices are rearranged such that the sets $A_j$ are ordered in the following way. The set $A_1$ is the one with most elements in it. The set $A_2$ is chosen such that $|A_2 \setminus A_1|$ is maximal, the set $A_3$ such that $A_3 \setminus (A_1 \cup A_2)$ contains most elements, and so on. For a given $c$, we define the *redundancy* $r$ of the graph as the number of such



**(a)** One root node **(b)** Two root nodes

**Fig. 4.** Example graphs for the maximization of $I_c$

sets with $|A_j \setminus (A_1 \cup \cdots \cup A_{j-1})| > c$, so

$$r = |\{A_j : |A_j \setminus (A_1 \cup \cdots \cup A_{j-1})| > c\}|. \tag{4}$$

In other words, $A_r \setminus (A_1 \cup \cdots \cup A_{r-1})$ is the first set in the above order for which $|A_j \setminus (A_1 \cup \cdots \cup A_{j-1})| \leq c$. We call

$$a = |A_1 \cup A_2 \cup \cdots \cup A_r| \tag{5}$$

the *number of essential nodes*, that is the number of observed nodes in the sets $A_j$ with $|A_j \setminus (A_1 \cup \cdots \cup A_{j-1})| > c$.

The construction of the $A_j$ sounds a bit technical, but it is easily done for a given graph. In Fig. 5, one possible choice would for example be $A_1 = \{X_1, X_6, X_9\}$, $A_2 = \{X_9, X_{10}, X_{12}\}$, $A_3 = \{X_4\}$ and $A_4 = \varnothing$, thus for $c = 2$ we have $r = 1$. In Fig. 6 we could choose $A_1 = \{X_7, X_8, X_9, X_{16}\}$, $A_2 \setminus A_1 = \{X_3, X_{13}\}$ and $A_3 \setminus (A_1 \cup A_2) = \{X_1\}$, thus for $c = 1$ we have $r = 2$.

Before we can determine the maximum of (3), we need the following lemma, which gives a non-trivial bound on $I_c$ for random variables without constraints. It is a small but neccessary (for the following theorem) improvement over the trivial bound $I_c(X_1, \ldots, X_n) \leq (n/c) \cdot \log 2$. The special case $n = 2$ and $c = 1$ is the well known $H(X) + H(Y) - H(X, Y) = H(X) - H(X \mid Y) \leq \log 2$.

**Lemma 1.** *For arbitrary binary random variables $X_1, \ldots, X_n$ we have the bound*

$$I_c(X_1, \ldots, X_n) \leq \left(\frac{n}{c} - 1\right) \cdot \log 2 \quad \textit{if } n \geq c. \tag{6}$$

*Proof.* Define $h(p) = -p \log(p) - (1 - p) \log(1 - p)$ and let $q$ be the largest probability of $p_\alpha = P(X_1 = \alpha_1, \ldots, X_n = \alpha_n)$ over all binary multi-indices $\alpha$. For every marginal $p_j$ with $1 \leq j \leq n$ we have then $p_j \geq q$ or $1 - p_j \geq q$, so $h(p_j) \leq h(q)$ because of symmetry and shape of $h$, thus we get

$$\sum_{j=1}^{n} H(X_j) = \sum_{j=1}^{n} h(p_j) \leq nh(q).$$

For the second part of (6), we have the bound

$$H(X_1, \ldots, X_n) = -\sum_{\alpha} p_\alpha \log p_\alpha = -q \log q - \sum_{\text{rest } \alpha} p_\alpha \log p_\alpha \geq h(q).$$

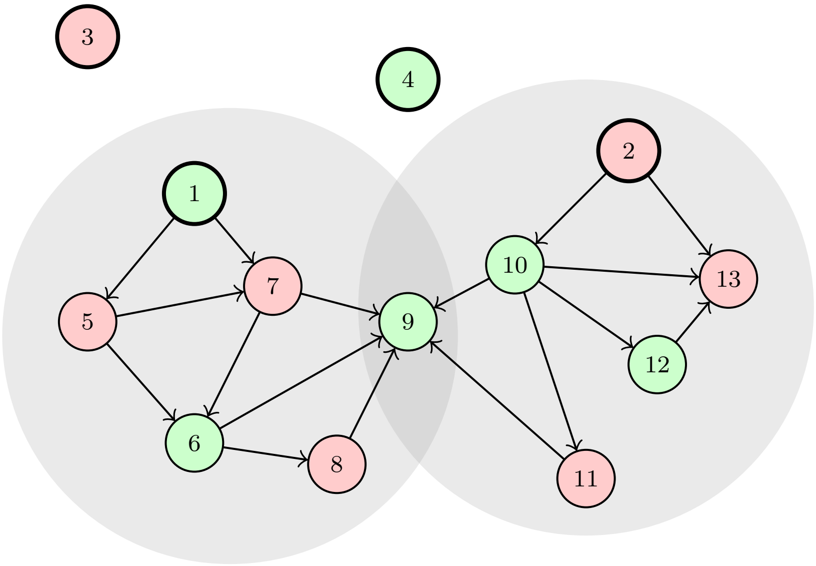


**Fig. 5.** The nodes $X_1$, $X_2$, $X_3$ and $X_4$ are the roots, the descendants of $X_1$ is the shaded circle on the left, the descendants of $X_2$ the right one. The node $X_9$ is descendant of both $X_1$ and $X_2$. The observed nodes $Y_1, \ldots, Y_k$ are green, the unobserved ones red.

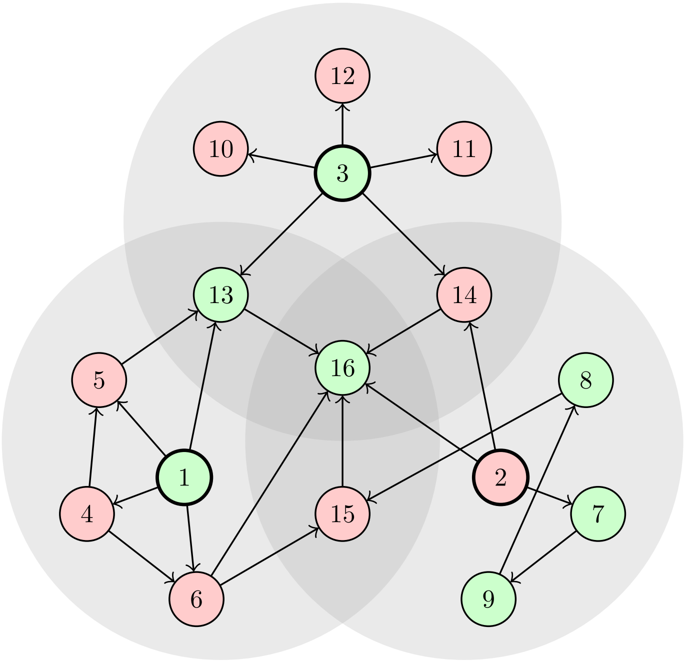


**Fig. 6.** The set $A_1$ contains the largest number of observed variables, that is 4. Then $A_2 \setminus A_1$ contains only 2 variables and $A_3 \setminus (A_1 \cup A_2)$ only one. The ordering is not unique, we could also interchange the names of $A_2$ and $A_3$.

The last step can be justified in the following way. We build a new probability distribution out of the other $p_\alpha$. By the non-negativity of entropy

$$0 \leq -\sum_{\text{rest } \alpha} \frac{p_\alpha}{1-q} \log \frac{p_\alpha}{1-q} = -\frac{1}{1-q} \sum_{\text{rest } \alpha} p_\alpha \log p_\alpha + \log(1-q).$$

Altogether the left hand side of (6) is $\leq (n/c - 1) \cdot h(q)$, and $h$ achieves its maximum for $q = 1/2$. □

We now have prepared all the neccessary tools for our main theorem, which relates the structure of $G$ with the maximum of $I_c$.

**Theorem 2** (New Common Cause Principle)**.** *Let $\mathcal{S}$ be the set of all probability distributions on binary random variables $X_1, \ldots, X_n$ that factorize according to a fixed acyclic graph $G$, so*

$$\mathcal{S} = \{p : \{0,1\}^n \to [0,1] \mid p(x_1, \ldots, x_n) = \textstyle\prod_{1 \leq j \leq n} p(x_j \mid x_{\mathrm{pa}(X_j)})\}.$$

(i) *For any subset $Y_1, \ldots, Y_k$ of observed nodes we have*

$$\sup_{p \in \mathcal{S}} I_c(Y_1, \ldots, Y_k) = \left(\frac{a}{c} - r\right) \cdot \log 2 \qquad (7)$$

*where $a$ and $r$ are defined as in* (4) *and* (5)*.*

(ii) *Certain* deterministic networks*, with $H(X_j \mid \mathrm{pa}(X_j)) = 0$ for all non-root nodes $X_j$ and a specific probability distribution of the root nodes, attain this supremum.*

*Proof.* First of all construct the probability distribution directly to show that (7) can be achieved. Then show that we cannot do better.

For the first part, set $P(X_j = 0) = 1/2$ for $1 \leq j \leq r$, where $r$ is the redundancy, and for all non-roots from the region of influence of these $X_j$ choose the probability distribution such that they copy the value of $X_j$ deterministically (for overlapping regions of influence choose one root from which the value is copied at random). For all the remaining nodes set $P(X_j = 0 \mid \mathrm{pa}(X_j)) = 1$. The joint probability distribution $P(X_1, \ldots, X_n)$ consists of $2^r$ equiprobable events, these are the events for $(X_1, \ldots, X_r) \in \{0,1\}^r$. Because in each $A_j$ for $1 \leq j \leq r$ there is at least one observed node, the marginalized distribution $P(Y_1, \ldots, Y_k)$ also consists of $2^r$ equiprobable events, so we have

$$H(Y_1, \ldots, Y_k) = -\sum_{j=1}^{2^r} \frac{1}{2^r} \log \frac{1}{2^r} = r \log 2.$$

On the other hand, $H(X) = \log 2$ for $X \in \mathrm{de}(X_1) \cup \cdots \cup \mathrm{de}(X_r)$, all other nodes have zero entropy by construction. So we conclude

$$\sum_{j=1}^{k} H(Y_j) = a \cdot \log 2,$$

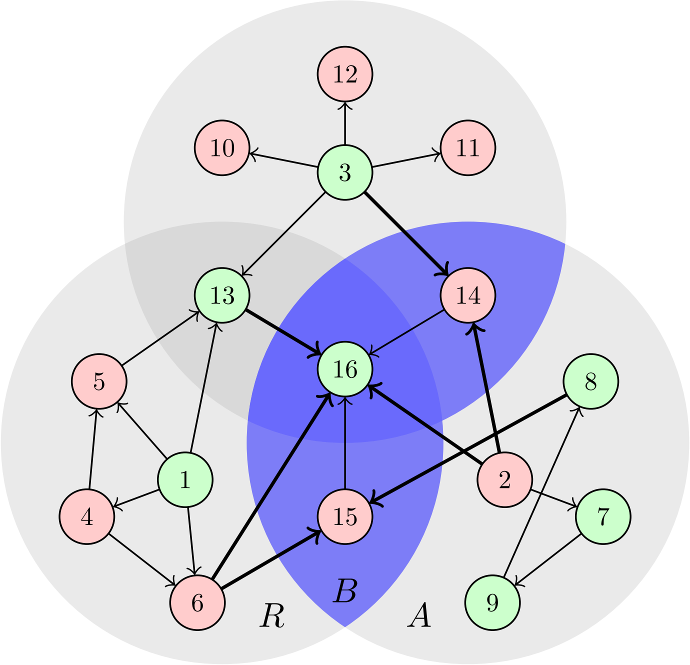
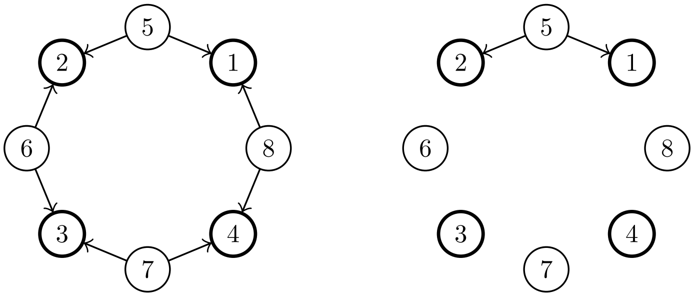


**Fig. 7.** For $R_j = \mathrm{de}(X_j)$, that means $A_j \subseteq R_j$ ($A_j$ contains only the observed nodes from $R_j$), the set $R_1 \cup R_2 \cup \cdots$ is partitioned into $A = R_1 \setminus (R_2 \cup R_3 \cup \cdots)$, $B = R_1 \cap (R_2 \cup R_3 \cup \cdots)$ and the rest $R = (R_2 \cup R_3 \cup \cdots) \setminus R_1$. Note that $X_A$ and $X_R$ are independent because they are $d$-separated by the empty set (the thick arrows are all pointing in).



**(a)** First hypothesis **(b)** Second hypothesis

**Fig. 8.** Two possible Bayesian networks that could underly the observed variables $X_1$, $X_2$, $X_3$ and $X_4$. The extended common cause principle cannot distinguish these.

and $I_c(Y_1, \ldots, Y_k)$ from (7) is achieved.

As for the second part, we use induction on the number of roots. For only one root, the bound follows from lemma 1. The induction step then proceeds as follows. The nodes in $A_1, A_2, \ldots$ are partitioned as shown in Fig. 7, so by the chain rule

$$H(Y_A, Y_B, Y_R) = H(Y_A, Y_R) + H(Y_B \mid Y_A, Y_R) \geq H(Y_A, Y_R),$$

and because $X_A$ and $X_R$ are independent and then also $Y_A$ and $Y_R$, it follows that $H(Y_A, Y_B, Y_R) \geq H(Y_A) + H(Y_R)$. By our induction hypothesis and lemma 1 we then have

$$\begin{aligned} I_c(Y_1, \ldots, Y_k) &\leq \sum_{j \in B} \frac{H(Y_j)}{c} + \sum_{j \in A} \frac{H(Y_j)}{c} - H(Y_A) + \sum_{j \in R} \frac{H(Y_j)}{c} - H(Y_R) \\ &\leq \frac{|B|}{c} \cdot \log 2 + \underbrace{\left( \frac{|A|}{c} - 1 \right)}_{\geq 0} \cdot \log 2 + I_c(Y_R) \leq \left( \frac{a}{c} - r \right) \cdot \log 2, \end{aligned}$$

which is the expected bound. □

## 5 Discussion and Conclusion

Now we describe how this theorem can be used as a new common cause principle. Assume that a scientist measured $X_1$, $X_2$, $X_3$ and $X_4$ from Fig. 8 and now has two hypotheses for the underlying causal model, the ones from Fig. 8 (a) and (b). With the extended common cause principle alone, he would not be able to discriminate between (a) and (b), because in each case there are two of the four observed variables that have a common ancestor. But the new common cause principle sometimes allows to decide between (a) and (b). For $c = 1$ we get $r = 2$ and $a = 4$, so $I_1 \leq 2 \log 2$ for (a) and $r = 1$ and $a = 2$, so $I_1 \leq \log 2$ for (b). Thus for $I_1$ in the range from $\log 2$ to $2 \log 2$ we can reject hypothesis (b), because the correlation is too strong.

Another application of the theorem was studied in [5]. There we showed that if common causes are to be inferred from empirical data, then the results are statistically more significant if $I_c$ is large. Thus, only in the case of large $a/c - r$ we have a chance of making assertions about common causes reliably. The quantity $a/c - r$ is not large if the graph consists of many small unconnected components. Or, in other words, the theorem states that for the prediction of common ancestors to be possible reliably, we must be able to control most of the graph by only few roots, that is, empirical inference of common ancestors works only reliably if the redundancy $r$ is small.

In summary, we presented a new version of the common cause principle in Bayesian networks, originally conceived by [7] and later extended and put in a quantitative form by [8], using concepts from [1]. The new common cause principle uses an inequality that gives an upper bound on the generalized information $I_c$ in a partially observed Bayesian network with fixed causal structure. This common cause principle sometimes allows a more fine grained discrimination between different causal hypotheses underlying an observation of parts of the Bayesian network. It is still an open problem, however, if and in what way lemma 1 and theorem 2 can be generalized to systems with non-binary variables. We hope that this new common cause principle can be of use in algorithms that try to build the causal structure of a Bayesian network from observation of the random variables alone.

**Acknowledgments.** Philipp Moritz thanks his colleagues from the MPI for Mathematics in the Sciences in Leipzig (especially the Information Theory of Cognitive Systems Group, where most of this research was conducted) for the pleasent stay and many stimulating discussions. Furthermore, he thanks the programme "FOKUS Physik" from the University of Würzburg for their support. Jörg Reichardt was supported by a Fellowship Computational Sciences of the Volkswagen Foundation.

## References

[1] Nihat Ay. A refinement of the common cause principle. *Discrete Applied Mathematics*, 157(10):2439–2457, 2009.

[2] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley, second edition edition, 2006.

[3] L. Lauritzen, S. *Graphical Models.* Oxford Science Publications. Clarendon Press, 1996.

[4] Steffen L. Lauritzen and Nuala A. Sheehan. Graphical models for genetic analyses. *Statistical Science*, 18:489–514, 2003.

[5] Philipp Moritz. Information inequalities in bayesian networks. Bachelor thesis, University of Würzburg, 2012.

[6] J. Pearl. *Causality: Models, Reasoning and Inference.* Cambridge University Press, 2000.

[7] H. Reichenbach and M. Reichenbach. *The Direction of Time.* California Library Reprint Series. University of California Press, 1956.

[8] Bastian Steudel and Nihat Ay. Information-theoretic inference of common ancestors. *CoRR*, abs/1010.5720, 2010.

[9] M. Studený. *Probabilistic Conditional Independence Structures.* Information Science and Statistics. Springer, 2005.

# Fuzzy Logic in Broader Sense: A Useful Tool for AI

**Vilém Novák**
Centre of Excellence IT4Innovations,
division of the University of Ostrava
Institute for Research and Applications of Fuzzy Modelling,
30. dubna 22, 701 03 Ostrava 1, Czech Republic
Vilem.Novak@osu.cz

**Abstract**

In this paper, we briefly characterize the main principles and constituents of the, so called, *fuzzy logic in broader sense* (FLb) that is part of mathematical fuzzy logic and extends the fuzzy logic in narrow sense. It is shown that FLb-logic is a reasonable tool using which various problems specific of artificial intelligence can be solved.

## 1 Constituents of mathematical fuzzy logic

Mathematical fuzzy logic includes two branches: *Fuzzy Logic in Narrow Sense* (FLn) and *Fuzzy Logic in Broader Sense* (FLb) which is an extension of the former. FLn generalizes classical mathematical logic (see [7, 10]), i.e., it has clearly distinguished syntax and semantics. The syntax consists of precise definitions of a formula, proof, formal theory, provability, model, etc. Many formal calculi have been developed in FLn. They usually differ from each other on the basis of the assumed structure of truth values, which then determines properties of the respective calculus.

*Fuzzy logic in broader sense* is an extension of FLn, which aims at developing a *formal theory of human reasoning that includes mathematical models of meaning of certain special expressions of natural language and generalized quantifiers with regard to presence of vagueness.* This program was initiated by V. Novák in [9]. It overlaps with two other paradigms proposed in the literature, namely that of commonsense reasoning and precisiated natural language.

*Commonsense reasoning* is an important branch of AI aiming at modeling common human reasoning and its ability to solve complex tasks. Its main tool is logic. Logical sentences are used to represent the knowledge, goals and situations of the agent. In the classical approach, they are represented by sentences of first-order logic, though other logical systems have also been proposed (cf. [3] and the citations therein). The main drawback of these formalizations, in our opinion, is neglecting vagueness present in the semantics of natural language expressions (cf. [5]).

*Precisiated Natural Language* introduced in [19] aims at providing an acceptable and applicable working formalization of the semantics of natural language without pretension to capture it in full detail and fineness. The methodology is based on two main premises:

(a) Much of the world knowledge is perception based (the term "perception" is not considered here as a psychological term but rather as a result of intrinsically imprecise human "measurement").

(b) Perception based information is intrinsically vague (fuzzy, imprecise).

PNL methodology requires presence of *World Knowledge Database* and *Multiagent, Modular Deduction Database.* No exact formalization of it, however, has been developed until recently, and so it should be taken mainly as a reasonable methodology.

The concept of FLb (cf. [13]) is a glue between both paradigms that should consider the best of each. It is a system of formal theories consisting, so far, of the following:

(a) Formal theory of evaluative linguistic expressions.

(b) Formal theory of intermediate and generalized quantifiers and their syllogisms.

(c) Formal theory of the meaning of fuzzy IF-THEN rules and approximate reasoning.

## 2 Fuzzy type theory as the basic tool for FLb

The first attempt at formalization of FLb has been done by V. Novák on the basis of *first-order fuzzy logic with evaluated syntax* (see [17]). More convenient for the goal of FLb is fuzzy type theory, a higher-order fuzzy logic, because the experience indicates that first-order logical systems are not powerful enough for the proper formalization of linguistic semantics.

The role of vagueness in commonsense reasoning has several aspects. First, it enables us to understand the complicated surrounding world, which cannot be known in complete detail. Thus, vagueness enables us to *reduce the necessary amount of information and focus only on its relevant constituents.* Furthermore, vagueness is quite often a feature of the information we have at our disposal; in certain circumstances, either more precise information is not available or obtaining it is too expensive. Thus, we must cope with the lack of detail and still make relevant conclusions. Vagueness is quite often even indispensable: it helps us to increase our awareness of the core of the problem because unnecessary details can be excluded and so we avoid "drowning" in the problem.

*Fuzzy type theory* (FTT) is a higher-order fuzzy logic being generalization of classical type theory initiated by B. Russel, A. Church and L. Henkin (for extensive presentation see [1]). The extension consists especially in replacement of the axiom stating "there are two truth values" by a sequence of axioms characterizing structure of the algebra of truth values.

The truth values should form an IMTL-algebra or EQ-algebra that is the most fundamental structure of truth values for FTT. Recall that the former is a prelinear residuated lattice with double negation while the latter is an algebra $\mathcal{E} = \langle E, \wedge, \otimes, \sim, \mathbf{1}\rangle$ where $\wedge$ is lattice meet, $\otimes$ is monoidal product, and $\sim$ is a fuzzy equality (equivalence). The most distinguished algebra of truth values for FLb-logic is the standard Łukasiewicz$_\Delta$ algebra $\mathcal{L} = \langle [0,1], \vee, \wedge, \otimes, \oplus, \Delta, \rightarrow, 0, 1\rangle$. Important concept in FTT is that of a *fuzzy equality*, which is a reflexive, symmetric and $\otimes$-transitive binary fuzzy relation on a set $M$, i.e. it is a function $\doteq\colon M \times M \longrightarrow L$.

Syntax of FTT is a generalization of the lambda-calculus constructed in a classical way, but differing from classical type theory by definition of additional special connectives, and in logical axioms. It has been proved that FTT (namely, various cases for special algebras of truth values) is complete. The details can be found in [12].

## 3 Constituents of FLb

*Evaluative linguistic expressions* are expressions of natural language, for example, *small, medium, big, about twenty five, roughly one hundred, very short, more or less deep, not very tall, roughly warm or medium hot, quite roughly strong, roughly medium size*, etc. They form a small, syntactically simple, but very important part of natural language which is present in its everyday use any time. The reason is that people regularly need to evaluate phenomena around them and to make important decisions, learn how to control, and many other activities based on their evaluation. In FLb, a special formal theory of FTT has been constructed using which semantics of the evaluative expressions is modeled. It can be demonstrated that the theory of evaluative expressions can well capture the vagueness phenomenon. The details can be found in [11].

In FLb, a special formal theory of FTT has been constructed to capture the meaning of evaluative expressions. We refer to [11] for all technical details. Since it has been proved that this has a model, the completeness theorem enables to prove the following.

**Theorem 1** *The formal theory of evaluative linguistic expressions is consistent.*

We can also demonstrate that the theory of evaluative expressions can capture vagueness. Besides others, it can well model the known *sorites paradox* that can be taken as the typical display of vagueness phenomenon. This result is contained in the following theorem.

**Theorem 2**

*(a)* $\vdash \mathbf{\Delta}\, Sm(0)$,

*(b)* $\vdash (\exists p)(\mathbf{\Delta}\neg\, Sm(p)$,

*(c)* $\vdash \neg(\exists n)(\mathbf{\Delta}\, Sm(n)\,\&\,\mathbf{\Delta}\neg\, Sm(n+1))$,

*(d)* $\vdash (\forall n)(Sm(n) \Rightarrow \mathit{Almost\ true}(Sm(n+1))$.

The statement (a) means that 0 is surely small, (b) means that there is a surely not small number, (c) means that there is no surely small number $n$ such that $n+1$ surely is not small, and (d) means that if $n$ is small then it is *almost true* that $n+1$ is also small. Analogous theorem can be proved also about *big* and *medium*.

*Intermediate quantifiers* are expressions such as *most, a lot of, many, a few, a great deal of, large part of, small part of*. They were informally studied in depth by Peterson in [18]. In FLb, intermediate quantifiers are modeled in FLb by special formulas of fuzzy type theory in a certain extension of the formal theory of evaluative linguistic expressions. The main idea is that intermediate quantifiers are classical general or existential quantifiers for which the universe of quantification is modified and the modification can be imprecise.

Below is formal definition of several specific intermediate quantifiers based on results in [18].

**A:** All $B$ are $A := (\forall x)(Bx \Rightarrow Ax)$,

**E:** No $B$ are $A := (\forall x)(Bx \Rightarrow \neg Ax)$,

**P:** Almost all $B$ are $A :=$

$$(\exists z)((\boldsymbol{\Delta}(z \subseteq B)\,\boldsymbol{\&}(\forall x)(zx \Rightarrow Ax)) \wedge (Bi\ Ex)((\mu B)z)),$$

(*extremely big part of* $B$ *has* $A$)

**B:** Few $B$ are $A$ :=

$$(\exists z)((\boldsymbol{\Delta}(z \subseteq B)\,\boldsymbol{\&}(\forall x)(zx \Rightarrow \neg Ax)) \wedge (Bi\ Ex)((\mu B)z)),$$

(*extremely big part of* $B$ *does not have* $A$)

**T:** Most $B$ are $A :=$

$$(\exists z)((\boldsymbol{\Delta}(z \subseteq B)\,\boldsymbol{\&}(\forall x)(zx \Rightarrow Ax)) \wedge (Bi\ Ve)((\mu B)z)),$$

(*very big part of* $B$ *has* $A$)

**D:** Most $B$ are not $A :=$

$$(\exists z)((\boldsymbol{\Delta}(z \subseteq B)\,\boldsymbol{\&}(\forall x)(zx \Rightarrow \neg Ax)) \wedge (Bi\ Ve)((\mu B)z)),$$

(*very big part of* $B$ *does not have* $A$)

**K:** Many $B$ are $A :=$

$$(\exists z)((\boldsymbol{\Delta}(z \subseteq B)\,\boldsymbol{\&}(\forall x)(zx \Rightarrow Ax)) \wedge \neg(Sm\,\bar{\boldsymbol{\nu}})((\mu B)z)),$$

(*not small part of* $B$ *has* $A$)

**G:** Many $B$ are not $A :=$

$$(\exists z)((\boldsymbol{\Delta}(z \subseteq B)\,\boldsymbol{\&}(\forall x)(zx \Rightarrow \neg Ax)) \wedge \neg(Sm\,\bar{\boldsymbol{\nu}})((\mu B)z)),$$

(*not small part of* $B$ *does not have* $A$)

**I:** Some $B$ are $A := (\exists x)(Bx \wedge Ax)$,

**O:** Some $B$ are not $A := (\exists x)(Bx \wedge \neg Ax)$.

# 4 Human reasoning

For human reasoning it is typical to use natural language so that linguistic semantics is combined with logical inference rules. We argue that our logical theory is rich enough to be able to develop sufficiently well working model of human reasoning. One such possibility was described in [14]. It should be noted that human reasoning is non-monotonic. In the cited paper, the monotonicity was considered and a model with conditional linguistic clauses containing evaluative expressions was presented.

One of many facets of human reasoning is syllogistic reasoning with intermediate quantifiers generalizing the classical Aristotle syllogism. A syllogism in fuzzy logic is a triple of formulas $\langle P_1, P_2, C\rangle$ such that the following is provable:

$$\vdash P_1 \mathbin{\&} P_2 \Rightarrow C.$$

Note a syllogism is valid if $\mathcal{M}(P_1) \otimes \mathcal{M}(P_2) \leq \mathcal{M}(C)$ holds in every model $\mathcal{M}$ of our theory.

Let $Q_1, Q_2, Q_3$ be intermediate quantifiers and $X, Y, M$ be formulas representing properties. Analogously as in classical logic, we will consider four figures of syllogisms:

| **Figure I** | **Figure II** | **Figure III** | **Figure IV** |
|---|---|---|---|
| $Q_1$ $M$ is $Y$ | $Q_1$ $Y$ is $M$ | $Q_1$ $M$ is $Y$ | $Q_1$ $Y$ is $M$ |
| $Q_2$ $X$ is $M$ | $Q_2$ $X$ is $M$ | $Q_2$ $M$ is $X$ | $Q_2$ $M$ is $X$ |
| $Q_3$ $X$ is $Y$ | $Q_3$ $X$ is $Y$ | $Q_3$ $X$ is $Y$ | $Q_3$ $X$ is $Y$ |

In [8] we proved that 105 generalized syllogism analyzed informally in [18] are valid in our theory, too. Let us only demonstrate few of them on an example:

$$\textbf{ATT}\text{-I:}\ \begin{array}{l} \text{All commercials are stupid} \\ \text{Most programs in the in US TV are commercials} \\ \hline \text{Most programs in US TV are stupid} \end{array}$$

$$\textbf{ETO}\text{-II:}\ \begin{array}{l} \text{No lazy people pass exam} \\ \text{Most students pass exam} \\ \hline \text{Some students are not lazy people} \end{array}$$

$$\textbf{PPI}\text{-III:}\ \begin{array}{l} \text{Almost all employed people have a car} \\ \text{Almost all employed people are well situated} \\ \hline \text{Some well situated people have a car} \end{array}$$

$$\textbf{TAI}\text{-IV:}\ \begin{array}{l} \text{Most shares with high value are from computer industry} \\ \text{All shares of computer industry are important} \\ \hline \text{Some important shares have high value} \end{array}$$

The theory of evaluative expressions is the point of departure for the theory of *fuzzy/linguistic IF-THEN rules.* These are conditional clauses of natural language having the form

$$\mathsf{IF}\ X \text{ is } \mathcal{A}\ \mathsf{THEN}\ Y \text{ is } \mathcal{B}, \tag{1}$$

where $X$ is antecedent variable (of course, there can be more of them), $Y$ is consequent variable and $\mathcal{A}, \mathcal{B}$ in (1) are the above mentioned evaluative linguistic expressions. The relation between antecedent and consequent is characterized by fuzzy (many-valued) implication. Sets (finite) of fuzzy IF-THEN rules are called *linguistic descriptions*. They represent a piece of text describing various decision, control, and other kinds of situations. Thus, such a text provides us with more information about the reality. The details about logical analysis of the fuzzy/linguistic IF-THEN rules can be found in [15].

The principal method for derivation of conclusions on the basis of linguistic descriptions is a special reasoning method called *perception-based logical deduction (PbLD)* [16]. Let us remark that this method is also used in the commonsense reasoning model in [14].

## 5 Conclusion

We are convinced that FLb is already sufficiently developed formal theory that can be very useful for further research in AI. For example, since FTT is generalization of classical type theory, it includes also classical predicate logic (namely, if we confine to two truth values only then FTT collapses into classical logic). Since it is not too complicated to translate classical formalism into FTT-formalism, one may immediately include vagueness into its special formal system. This holds, for example, for many systems developed in AI, such as Versatile Event Logic presented in [2] or the formalization from [6]. Our theory of meaning of evaluative expressions can also be extended to include, e.g., sophisticated model of meaning of concepts as presented in [4]. Of course, incorporating features of non-monotonic logic into FLb also causes no significant obstacle though a lot of research still has to be done.

**Acknowledgment.** The research was supported by the European Regional Development Fund in the IT4Innovations Centre of Excellence project
(CZ.1.05/1.1.00/02.0070).

## References

[1] P. Andrews, An Introduction to Mathematical Logic and Type Theory: To Truth Through Proof, Kluwer, Dordrecht, 2002.

[2] B. Bennett, A. Galton, A unifying semantics for time and events, Artificial Intelligence 153 (1-2) (2004) 13–48.

[3] E. Davis, L. Morgenstern, Introduction: Progress in formal commonsense reasoning, Artifical Intelligence 153 (2004) 1–12.

[4] M. Duží B. Jespersen, P. Materna, Procedural Semantics for Hyperintensional Logic, Springer, Dordrecht, 2010.

[5] A. Dvořák, V. Novák, Fuzzy logic as a methodology for the treatment of vagueness, in: L. Běhounek, M. Bílková (eds.), The Logica Yearbook 2004, Filosofia, Prague, 2005, pp. 141–151.

[6] A. Gordon, J. Hobbs, Formalizations of commonsense psychology, AI Magazine 25 (4) (2004) 49–62.

[7] P. Hájek, What is mathematical fuzzy logic, Fuzzy Sets and Systems 157 (2006) 597–603.

[8] P. Murinová, V. Novák, A formal theory of generalized intermediate syllogisms, Fuzzy Sets and Systems 186 (2012) 47–80.

[9] V. Novák, Towards formalized integrated theory of fuzzy logic, in: Z. Bien, K. Min (eds.), Fuzzy Logic and Its Applications to Engineering, Information Sciences, and Intelligent Systems, Kluwer, Dordrecht, 1995, pp. 353–363.

[10] V. Novák, Which logic is the real fuzzy logic?, Fuzzy Sets and Systems 157 (2006) 635–641.

[11] V. Novák, A comprehensive theory of trichotomous evaluative linguistic expressions, Fuzzy Sets and Systems 159 (22) (2008) 2939–2969.

[12] V. Novák, EQ-algebra-based fuzzy type theory and its extensions, Logic Journal of the IGPL 19 (2011) 512–542.

[13] V. Novák, Reasoning about mathematical fuzzy logic and its future, Fuzzy Sets and Systems 192 (2012) 25–44.

[14] V. Novák, A. Dvořák, Formalization of commonsense reasoning in fuzzy logic in broader sense, Journal of Applied and Computational Mathematics 10 (2011) 106–121.

[15] V. Novák, S. Lehmke, Logical structure of fuzzy IF-THEN rules, Fuzzy Sets and Systems 157 (2006) 2003–2029.

[16] V. Novák, I. Perfilieva, On the semantics of perception-based fuzzy logic deduction, International Journal of Intelligent Systems 19 (2004) 1007–1031.

[17] V. Novák, I. Perfilieva, J. Močkoř, Mathematical Principles of Fuzzy Logic, Kluwer, Boston, 1999.

[18] P. Peterson, Intermediate Quantifiers. Logic, linguistics, and Aristotelian semantics, Ashgate, Aldershot, 2000.

[19] L. A. Zadeh, Precisiated natural language, AI Magazine 25 (2004) 74–91.

# Duality in Residuated Structures

**Irina Perfilieva**
University of Ostrava
Centre of Excellence IT4Innovations, division of the University of Ostrava,
Institute for Research and Applications of Fuzzy Modeling,
30. dubna 22, 701 03 Ostrava, Czech Republic
Irina.Perfilieva@osu.cz

**Abstract**

We investigate residuated lattices as structures with dualities. We formulated the *Principle of Duality* for residuated lattice which divides basic properties of this structure into two groups of dual ones.

## 1 Introduction

The notion of commutative residuated $\ell$-monoid has been introduced in [1] with the goal "to outline a common framework for a diversity of monoidal structures which constitute the basis of various papers in fuzzy set theory". This notion is a bit more general than the earlier introduced [2] notion of "residuated lattice" for an integral, residuated, commutative $\ell$-monoid. Both papers became fundamental in the literature related to mathematical fuzzy logic [3, 4, 5], algebraic foundations of fuzzy systems [6], algebraic foundations of triangular norms [7], etc. Moreover, they had significant impact on intensive development of various residuated algebraic structures such as residuated $\ell$-groupoids, residuated $\ell$-semigroups, etc. (see, e.g. [8]). From the point of view of applications, residuated structures play an important role, for example, in mathematical morphology, modeling of linguistic semantics, approximate reasoning, and elsewhere.

In the proposed contribution, we investigate how residuation manifests itself through duality. We will see that existence of a pair of dual semimodules of a commutative semiring implies that its monoidal reduct is a residuated lattice. This result allows to formulate the *Principle of Duality* for residuated lattice which divides basic properties of this structure into two groups of dual ones.

## 2 Residuated, Commutative $\ell$-monoid as a Pair of Dual Semimodules

In this section, we will see how a residuated, commutative $\ell$-monoid can be characterized without any reference to the property of residuation. To characterize residuated operations, we will use two dually ordered semimodules. Moreover, the existence of two dually ordered semimodules is a part of the criterion under which a monoidal reduct of a lattice-ordered commutative semiring is a residuated, commutative $\ell$-monoid. We will formulate and illustrate the *Principle of Duality* for residuated lattice.

## 2.1 Semimodules in residuated, commutative $\ell$-monoid

Let us recall that a *semiring* is an algebraic structure with two associative operations which are connected by distributive laws (cf. [9, 10, 6]). In more details, a semiring $R = (R, +, \cdot, 0, e)$ is an algebraic structure with the following properties:

- $(R, +, 0)$ is a commutative monoid,
- $(R, \cdot, e)$ is a monoid,
- $a \cdot (b + c) = a \cdot b + a \cdot c$ and $(a + b) \cdot c = a \cdot c + b \cdot c$ hold for all $a, b, c \in R$.

A semiring is called *commutative* if $(R, \cdot, e)$ is a commutative monoid. An example of a commutative semiring, which will be used in the sequel, is as follows.

**Example 1** *Let $\mathcal{L} = (L, \leq, *)$ be a residuated, commutative $\ell$-monoid. Then $\mathcal{R}^{\mathcal{L}}_{\vee} = (L, \vee, *, \perp, e)$, where $\perp$ is a bottom element of $(L, \leq)$ and $e$ is the unit with respect to $*$, is a semiring reduct of $\mathcal{L}$. This easily follows from the fact that for every $a \in L$, the map $a * (\cdot) : L \to L$ is residuated and therefore, it preserves the join operation, i.e. for all $x, y \in L$, $a * (x \vee y) = (a * x) \vee (a * y)$ holds.*

Let us give a definition of a (left) *R-semimodule.*

**Definition 1** *Let $R$ be a semiring. A (left) $R$-semimodule is an algebra*

$$M = (M, +, 0, (h_a)_{a \in R}),$$

*where $(M, +, 0)$ is a commutative monoid and each $h_a : M \to M$ is a unary operation so that the following properties are fulfilled:*

$$h_a(x + y) = h_a(x) + h_a(y), \tag{1}$$
$$h_{a+b}(x) = h_a(x) + h_b(x), \tag{2}$$
$$h_{a \cdot b}(x) = h_a(h_b(x)). \tag{3}$$

Let us remark that the above definition is different from those in [9, 6] where there are additional requirements on zero elements of $R$ and $M$.

The unary operation $h_a$ is usually considered as a (left) *scalar multiplication* by $a$. An $R$-semimodule is called *unital* (or *unitary*) if the unit element $e \in R$ determines the identical map on $M$, i.e. for all $x \in M$, $h_e(x) = x$.

**Example 2** *It is easy to see that if $R = (R, +, \cdot, 0, e)$ is a semiring then $M_R = (R, +, 0, (a \cdot (\cdot))_{a \in R})$ is a unital semimodule where $a \cdot (\cdot) : x \mapsto a \cdot x$. We will say that the semimodule $M_R = (R, +, 0, (a \cdot (\cdot))_{a \in R})$ is induced by a semiring $R = (R, +, \cdot, 0, e)$.*

**Example 3** *Let $\mathcal{L} = (L, \leq, *)$ be a residuated, commutative $\ell$-monoid and $\mathcal{R}_\vee^{\mathcal{L}} = (L, \vee, *, \perp, e)$ its semiring reduct. By Example 2, $(L, \vee, \perp, (a * \ (\cdot))_{a \in L})$ is a $\mathcal{R}_\vee^{\mathcal{L}}$-semimodule induced by the respective semiring.*

Two $R$-semimodules $M_1$ and $M_2$ are called *isomorphic* if there is a bijection $\varphi : M_1 \to M_2$ which establishes an isomorphism between respective monoids $(M_1, +, 0)$ and $(M_2, +, 0)$ and which commutes with every unary operation $h_a$, i.e. for all $a \in R$ and for all $x \in M_1$,

$$\varphi(h_a(x)) = h_a(\varphi(x)).$$

Similarly, a homomorphism between $M_1$ and $M_2$ can be defined.

In the below given theorems, we show that any residuated, commutative $\ell$-monoid can be split into two unital left semimodules such that one of them is a homomorphic image of the other one. As a converse statement, we will prove that existence of two (dual) unital left semimodules is a criterion under which a monoidal reduct of a lattice-ordered commutative semiring is a residuated, commutative $\ell$-monoid.

**Theorem 1** *Let $\mathcal{R}_\vee^{\mathcal{L}} = (L, \vee, *, \perp, e)$ be a lattice-ordered commutative semiring where $(L, \vee, \wedge, \perp, \top)$ is a bounded lattice and $e$ is the unit with respect to $*$. Then the monoidal reduct $(L, *, e)$ of $\mathcal{R}_\vee^{\mathcal{L}}$ is a residuated, commutative $\ell$-monoid if and only if there exists a set of unary operations $\{h_a : L \to L \mid a \in L\}$ such that $(L, \wedge, \top, (h_a)_{a \in L}))$ is a $\mathcal{R}_\vee^{\mathcal{L}}$-semimodule and for all $a, b \in L$,*

$$a \leq b \Leftrightarrow e \leq h_a(b). \tag{4}$$

**Corollary 1** *Let $\mathcal{L} = (L, \leq, *)$ be a residuated, commutative $\ell$-monoid and $\mathcal{R}_\vee^{\mathcal{L}} = (L, \vee, *, \perp, e)$ its semiring reduct. Then*

*(i) there are two $\mathcal{R}_\vee^{\mathcal{L}}$-semimodules: $\mathcal{L}_{\vee,*} = (L, \vee, \perp, (a * \ (\cdot))_{a \in L})$ and $\mathcal{L}_{\wedge,\to} = (L, \wedge, \top, (a \to \ (\cdot))_{a \in L})$, that split $\mathcal{L}$,*

*(ii) both $\mathcal{R}_\vee^{\mathcal{L}}$-semimodules $\mathcal{L}_{\vee,*}$ and $\mathcal{L}_{\wedge,\to}$ are unital,*

*(iii) the condition (4) in Theorem 1 is equivalent to the fact that $e = \top$, i.e. $\mathcal{L}$ is integral.*

In the sequel, we will assume that a given residuated, commutative $\ell$-monoid is integral. For its shorter name we will use the name "*residuated lattice*". It will be denoted by $\mathcal{L} = (L, \leq, *)$ with additional operations as follows: $\to$ as a residual of $*$, and $\perp$ and $\top$ as respective bottom and top elements.

**Theorem 2** *Let $\mathcal{L} = (L, \leq, *)$ be a residuated lattice with $\mathcal{R}_\vee^{\mathcal{L}} = (L, \vee, *, \perp, \top)$ as a semiring reduct where $\top$ is the unit with respect to $*$. Then the left $\mathcal{R}_\vee^{\mathcal{L}}$-semimodule $\mathcal{L}_{\wedge,\to} = (L, \wedge, \top, (a \to (\cdot))_{a \in R}))$ is a homomorphic image of the left $\mathcal{R}_\vee^{\mathcal{L}}$-semimodule $\mathcal{L}_{\vee,*} = (L, \vee, \perp, (a * (\cdot))_{a \in L})$ under the homomorphism $\varphi : L \to L$, given by*

$$\varphi(x) = x \to \perp.$$

## 2.2 Principle of Duality in Residuated Lattices

Let $\mathcal{L} = (L, \leq, *)$ be a residuated lattice and $\mathcal{R}_{\vee}^{\mathcal{L}} = (L, \vee, *, \bot, \top)$ its semiring reduct. By Corollary 1, $\mathcal{L}$ splits into two unital $\mathcal{R}_{\vee}^{\mathcal{L}}$-semimodules $\mathcal{L}_{\vee,*} = (L, \vee, \bot, (a * (\cdot))_{a \in L})$ and $\mathcal{L}_{\wedge,\rightarrow} = (L, \wedge, \top, (a \rightarrow (\cdot))_{a \in R}))$. The operations $\vee$ in $\mathcal{L}_{\vee,*}$ and $\wedge$ in $\mathcal{L}_{\wedge,\rightarrow}$ determine dual orderings of the same support $L$ of these semimodules:

$$\begin{aligned} x \leq y &\iff x \vee y = y, \\ x \geq y &\iff x \wedge y = y. \end{aligned}$$

Below, we will refer to semimodules $\mathcal{L}_{\vee,*}$ and $\mathcal{L}_{\wedge,\rightarrow}$ as to dual ones. Because both semimodules are representatives of the same algebraic structure, everything which can be proved for one of them can be proved for the other one. This is used in the below formulated *Principle of Duality* which also uses the fact that both semimodules are parts of a residuated lattice.

**Principle of Duality in residuated, commutative $\ell$-monoids**

*To every theorem that concerns a residuated lattice $\mathcal{L} = (L, \leq, *)$ and is formulated in a language of one of its $\mathcal{R}_{\vee}^{\mathcal{L}}$-semimodules there is a corresponding theorem that concerns its dual semimodule. This is obtained by replacing symbols of operations $\vee$, $\bot$ and $(a * (\cdot))_{a \in L}$ of the semimodule $\mathcal{L}_{\vee,*}$ by the respective dual symbols $\wedge$, $\top$ and $(a \rightarrow\ (\cdot))_{a \in L}$ of the semimodule $\mathcal{L}_{\wedge,\rightarrow}$, and vice versa.*

The following list contains some pairs of dual statements where the elements of the semiring $\mathcal{R}_{\vee}^{\mathcal{L}}$ are denoted by characters $a, b$ to be distinguished from the elements of the respective semimodules denoted by characters $x, y$.

$$a * (x \vee y) = (a * x) \vee (a * y), \quad a \rightarrow (x \wedge y) = (a \rightarrow x) \wedge (a \rightarrow y); \tag{5}$$

$$(a \vee b) * x = (a * x) \vee (b * x), \quad (a \vee b) \rightarrow x = (a \rightarrow x) \wedge (b \rightarrow x); \tag{6}$$

$$(a * b) * x = a * (b * x), \quad (a * b) \rightarrow x = a \rightarrow (b \rightarrow x). \tag{7}$$

**Remark 1** *It is worth stressing that the* Principle of Duality *is applicable only to statements in the language of respective semimodules and only to semimodules operations. It is not applicable to operations of the semiring $\mathcal{R}_{\vee}^{\mathcal{L}}$. The following example illustrates this remark.*

***Example 4*** *In two dual properties (6),*

$$(a \vee b) * x = (a * x) \vee (b * x), \quad (a \vee b) \rightarrow x = (a \rightarrow x) \wedge (b \rightarrow x),$$

*the symbol $\vee$ in the term $(a \vee b)$ (left-hand side of the left equality) does not correspond to the join operation $\vee$ of the semimodule $\mathcal{L}_{\vee,*}$ (it corresponds to the semiring operation). Therefore, it was not changed in the right (dual) equality above. However, the same symbol $\vee$ in the expression $(a * x) \vee (b * x)$ (right-hand side of the left equality) does correspond to the join operation of $\mathcal{L}_{\vee,*}$ and by the* Principle of Duality*, it was replaced by $\wedge$ in the right (dual) equality above. Moreover, all symbols $*$ in the left equality above correspond to (unary) operations of $\mathcal{L}_{\vee,*}$, and therefore, they were replaced by the respective dual symbols $\rightarrow$.*

Let us show how the *Principle of Duality* can be used for discovery dual forms of true equalities.

**Example 5** *Let us consider the following true equality:*

$$(a * b \vee c) * x = a * b * x \vee c * x,$$

*and rewrite it into the language of the semimodule* $\mathcal{L}_{\vee,*}$*:*

$$(a * b \vee c) * x = a * (b * x) \vee c * x. \tag{8}$$

*In the left-hand side of (8), there is one semimodule operation* $*$ *(before* $x$*) which will be replaced by the semimodule operation* $\to$ *of the dual semimodule* $\mathcal{L}_{\wedge,\to}$*. In the right-hand side of (8), all operations are semimodule operations and by the* Principle of Duality*, they will replaced by respective dual operations. Thus, we will come to the following dual equality:*

$$(a * b \vee c) \to x = (a \to (b \to x)) \wedge (c \to x).$$

In the below given example, we will show how the *Principle of Duality* can be used for simplifying formal expressions written in the language of $\mathcal{L}_{\wedge,\to}$.

**Example 6** *Let us consider the following expression in the language of* $\mathcal{L}_{\wedge,\to}$*:*

$$((a * b) \to (x \wedge y)) \wedge (a \to x), \tag{9}$$

*with the purpose to find its simpler form. We propose to rewrite (9) into its dual form, simplify the dual form and then rewrite it back. Let us remark that there is one non-semimodule operation in (9) - the first* $*$ *in the term* $(a * b)$ *(it will not be changed in the dual form of (9)). By the* Principle of Duality*, the dual form of (9) is as follows:*

$$((a * b) * (x \vee y)) \vee (a * x).$$

*The above given expression can be easily rewritten into*

$$(a * b) * x \vee (a * b) * y \vee a * x,$$

*and simplified to*

$$a * (b * y \vee x).$$

*Finally, after rewriting back into the language of* $\mathcal{L}_{\wedge,\to}$ *we will obtain the desired simplified form of (9):*

$$a \to ((b \to y) \wedge x).$$

# 3 Conclusion

In this paper, we investigated residuated lattices as structures with dualities. We proved that the existence of two (dual) unital left semimodules is a criterion under which a monoidal reduct of a lattice-ordered commutative semiring is a residuated, commutative $\ell$-monoid. We formulated the *Principle of Duality* for residuated lattices which divides basic properties of these structures into two groups of dual ones.

**Acknowledgments.** The author acknowledges that this contribution has been prepared in connection with the project IT4Innovations Centre of Excellence, reg. no. CZ.1.05/1.1.00/02.0070.

## References

[1] U. Höhle, Commutative residuated l-monoids, in *Non-Classical Logics and Their Applications to Fuzzy Subsets. A Handbook of the Mathematical Foundations of Fuzzy Set Theory*, eds. U. Höhle and E. P. Klement (Kluwer, Dordrecht, 1995) pp. 53–106.

[2] R. P. Dilworth and M. Ward, *Trans. Amer. Math. Soc.* 45, 335 (1939).

[3] P. Hájek, *Metamathematics of Fuzzy Logic* (Kluwer, Dordrecht, 1998).

[4] V. Novák, I. Perfilieva and J. Močkoř, *Mathematical Principles of Fuzzy Logic* (Kluwer, Boston, 1999).

[5] F. Esteva and L. Godo, *Fuzzy Sets and Systems* 124, 271 (2001).

[6] A. Di Nola, A. Lettieri, I. Perfilieva and V. Novák, *Fuzzy Sets and Systems* 158, 1 (2007).

[7] E. P. Klement, R. Mesiar and E. Pap, *Triangular Norms* (Kluwer, Dordrecht, 2000).

[8] N. Galatos, P. Jipsen, T. Kowalski and H. Ono, *Residuated Lattices: an algebraic glimpse at substructural logics*Studies in Logics and the Foundations of Mathematics, Studies in Logics and the Foundations of Mathematics (Elsevier, Amsterdam, 2007).

[9] J. S. Golan, *Semirings and their Applications* (Kluwer Academic Pulishers, Dordrecht, 1999).

[10] M. Gondran and M. Minoux, *Graphs, Dioids and Semirings* (Springer-Verlag, New York, 2008).

# Conditional Probability: Advantages and Inconveniences of Different Approaches

**Romano Scozzafava**
Dipartimento di Scienze di Base e Applicate per l'Ingegneria, sezione di Matematica
Universitá "La Sapienza"
romscozz@dmmm.uniroma1.it

**Abstract**

A simple example is taken as starting point for a discussion concerning advantages and inconveniences of different approaches to the concept of conditional probability (de Finetti's coherence, classical Kolmogorov's, on MV-algebras, ...).

The example considers only a finite number of conditional events, involving conditioning events (different from the impossible one) of zero probability.

In particular, it is shown that the approach based on coherence allows a proper and convincing treatment and interpretation of conditioning on null events.

## 1 Introduction

Consider a young patient **B** (aged less than 30) with a blurred vision addressing an ophthalmic hospital for a suspect eye disease.

Let $E_1$ be the event "**B** has retinitis", $E_2$ the event "**B** has blepharitis", and $E_3$ the event "**B** has glaucoma".

Consider the assessment

$$P(E_1) = \frac{3}{4} = P(E_1|\Omega) = P(E_1|H_1) \ ,$$

where $\Omega = H_1$ is the *sure* event, and suppose that **B** undertakes a medical test for glaucoma (this test has a history of no positive results for patients aged less than 30). Denoting by $H_2$ the event "the test for glaucoma is negative", consider the assessment

$$P(E_2|H_2) = \frac{1}{4} \ .$$

Denoting by $H_3 = H_2^c$ the event "the test for glaucoma is positive", the probability of the conditional event $E_3^c|H_3$, e.g.

$$P(E_3^c|H_3) = \frac{1}{2}$$

(where $A^c$ denotes the *contrary* of the event $A$), represents the probability of a false positive.

Finally, assume the following logical relations among the given events:

$$E_i \wedge E_j = \emptyset \ \ (i,j = 1,2,3\ , i \neq j)\ \ ,$$

$$H_2 = E_1 \vee E_2\ \ ,\ \ H_3 = E_3 \vee A_4\ ,$$

with $A_4 = E_1^c \wedge E_2^c \wedge E_3^c$ . Then the events $E_i$ can be identified with three *atoms* $A_i$, so that

$$H_1 = \Omega = A_1 \vee A_2 \vee A_3 \vee A_4 = H_2 \vee H_3\ .$$

This simple example will be the starting point for a discussion concerning advantages and inconveniences of different approaches to the concept of conditional probability.

## 2 Checking coherence

Facing the example given in the Introduction, let $P_o$ be a probability on $\Omega$ and put $x_r = P_o(A_r)$, with $r = 1, ..., 4$; consider the system

$$\begin{cases} x_1 = \frac{3}{4} \cdot (x_1 + x_2 + x_3 + x_4) \\ x_2 = \frac{1}{4} \cdot (x_1 + x_2) \\ x_3 = \frac{1}{2} \cdot (x_3 + x_4) \\ x_1 + x_2 + x_3 + x_4 = 1 \\ x_r \geq 0\ , \end{cases}$$

whose only solution is

$$x_1 = \frac{3}{4}\,, \quad x_2 = \frac{1}{4}\,, \quad x_3 = x_4 = 0\,.$$

The above system can be seen as a particular case of the following one

$$(S_o) \qquad \begin{cases} \sum_r\limits_{A_r \subseteq E_i \wedge H_i} x_r = P(E_i|H_i) \sum_r\limits_{A_r \subseteq H_i} x_r \quad (i = 1,2,3), \\ \sum_r\limits_{A_r \subseteq H_o^o} x_r = 1 \\ x_r \geq 0\,, \end{cases}$$

where

$$H_o^o = H_1 \vee H_2 \vee H_3 = \Omega\,.$$

The first three equations correspond to the product rule of conditional probability, that is, taking $P_o(E_i|H_i) = P(E_i|H_i)$,

$$(1) \qquad P_o(E_i \wedge H_i) = P(E_i|H_i)\, P_o(H_i)\,.$$

Notice that if we had required *positivity of the probability of conditioning events*, we should have added to the above system also the conditions

$$P_o(A_1) + P_o(A_2) > 0\,, \quad P_o(A_3) + P_o(A_4) > 0\,,$$

and this enlarged system (as it is easily seen) *has no solutions*.

On the other hand, $x_3 = x_4 = 0$ means $P(H_3) = 0$, an assessment which would not allow to evaluate $P(E_3|H_3)$ by means of the classical Kolmogorov's definition.

**Remark 1** - *The assessment $P(H_3) = 0$ is consistent with the fact that the test for glaucoma had no positive results for patient aged less than 30.*

*The problem of evaluating the probability of such kind of events (not yet occurred in the available data) belongs to the so–called "zero-frequency problems": for a Bayesian approach and related references, see, e.g., [13].*

Going back to $P_o(H_i) = 0$ (which obviously implies $P_o(E_i \wedge H_i) = 0$), notice that the product rule (1) is compatible with **any** value (even negative or greater than 1) of the assessment $P(E_i|H_i)$.

This situation occurs – in our example – for the third equation. So we introduce a "new" probability $P_1$ defined on

$$H_o^1 = H_3 = A_3 \vee A_4$$

and such that the following system, with unknowns $y_r = P_1(A_r)$, is compatible

$$(S_1) \qquad \begin{cases} \sum_{A_r \subseteq E_i \wedge H_i}^{r} y_r = P(E_i|H_i) \sum_{A_r \subseteq H_i}^{r} y_r \quad (i = 3), \\ \sum_{A_r \subseteq H_o^1}^{r} y_r = 1 \\ y_r \geq 0 \quad . \end{cases}$$

Notice that $P_1$ could be seen as the restriction $P(\cdot|H_o^1)$ of the (sought) conditional probability $P(\cdot|\cdot)$ – just as $P_o$ could have been seen as the restriction $P(\cdot|\Omega)$ – and the first equation of system $(S_1)$ corresponds to the product rule, but *with $H_o^1$ playing the role of the sure event $\Omega$*

$$(2) \qquad P_1(E_i \wedge H_i|H_o^1) = P(E_i|H_i)\, P_1(H_i|H_o^1)\,.$$

So in our case the system becomes

$$\begin{cases} y_3 = \frac{1}{2} \cdot (y_3 + y_4) \\ y_3 + y_4 = 1 \\ y_r \geq 0\,, \end{cases}$$

and its only solution is

$$y_3 = y_4 = \frac{1}{2} \quad .$$

We wonder whether from the above procedure (*i.e.*, checking the compatibility of a suitable sequence of linear systems) it follows that the conditional probability assessment given in the Introduction is *coherent*, according to the following (see de Finetti [9])

**Definition** – *The assessment $P(\cdot|\cdot)$ on an arbitrary family $\mathcal{C}=\mathcal{C}_1\times\mathcal{C}_2$ of conditional events is* coherent *if there exists $\mathcal{C}'\supseteq\mathcal{C}$, with $\mathcal{C}'=\mathcal{G}\times\mathcal{B}^o$ ($\mathcal{G}$ a Boolean algebra, $\mathcal{B}$ an additive set – i.e. closed with respect to (finite) logical sums – with $\mathcal{B}^o=\mathcal{B}\setminus\{\emptyset\}$, $\mathcal{B}\subseteq\mathcal{G}$), such that $P(\cdot|\cdot)$ can be* extended *from $\mathcal{C}$ to $\mathcal{C}'$ as a* conditional probability, *i.e. a function $P:\mathcal{C}\to[0,1]$ such that*

*(i)* $P(H|H)=1$, *for every* $H\in\mathcal{B}^o$ ,

*(ii)* $P(\cdot\,|H)$ *is a (finitely additive) probability on* $\mathcal{G}$ *for any given* $H\in\mathcal{B}^o$ ,

*(iii)* $P\big((E\wedge A)|H\big)=P\big(E|(A\wedge H)\big)\,P(A|H)$, *for any* $A,E\in\mathcal{G}\,, H, E\wedge H\in\mathcal{B}^o$.

Notice that the product rule (2) corresponds to axiom (*iii*) with $E=E_i$, $A=H_i$, $H=H_o^1$.

In classical approaches, a conditional probability $P(E|H)$ is not introduced as a *direct* notion, and so there is no meaning given to $E|H$ itself: de Finetti [8] was the first to mention "conditional events" outside the function $P$.

Coherence of the assessments $P(E_i|H_i)$ $(i=1,2,3)$ given in the Introduction follows from the following Theorem (*characterizing* coherent assessments), which is a particular case of a more general one, valid also for *arbitrary* (infinite) families of conditional events.

**Theorem** – *Let $\mathcal{C}$ be a finite family of conditional events*

$$\mathcal{C}=\{E_1|H_1,\ldots,E_n|H_n\}$$

*and denote by $\mathcal{A}_o$ the set of atoms $A_r$ generated by the (unconditional) events $E_1,H_1,\ldots,E_n,H_n$. For a real function $P$ on $\mathcal{C}$ the following two statements are equivalent:*

*(a) $P$ is a* coherent *conditional probability on $\mathcal{C}$;*

*(b) all systems of the following sequence, with unknowns $x_r^\beta=P_\beta(A_r)\geq 0$, $A_r\in\mathcal{A}_\beta$, and $\beta=0,1,2,\ldots,k\leq n$, are compatible:*

$$(S_\beta)\qquad\begin{cases}\underset{A_r\subseteq E_i\wedge H_i}{\sum\nolimits_r} x_r^\beta = P(E_i|H_i)\underset{A_r\subseteq H_i}{\sum\nolimits_r} x_r^\beta\,,\\ \big[\text{for all } E_i|H_i\in\mathcal{C} \text{ such that } \underset{A_r\subseteq H_i}{\sum\nolimits_r}\mathbf{x}_\mathbf{r}^{\beta-\mathbf{1}}=0\,\big]\\ \underset{A_r\subseteq H_o^\beta}{\sum\nolimits_r} x_r^\beta=1\end{cases}$$

*(put, for* **all** *$H_i$'s, $\underset{A_r\subseteq H_i}{\sum\nolimits_r} x_r^{-1}=0$ when $\beta=0$), where $H_o^o=H_o=H_1\vee\ldots\vee H_n$, while $\mathbf{x}_\mathbf{r}^{\beta-\mathbf{1}}$ denotes* a solution *of $(S_{\beta-1})$ and $H_o^\beta$ is, for $\beta\geq 1$, the union of the $H_i$'s such that $\underset{A_r\subseteq H_i}{\sum\nolimits_r} x_r^{\beta-1}=0$.*

Any class $\{P_\beta\}$ singled–out by the condition *(b)* is said *to agree* with the conditional probability $P$.

Slightly different versions of this theorem have been given in recent years (see, e.g., [4], [5], the book [6] and [7]).

# 3 Coherence vs. Kolmogorov approach

Since *we do not presuppose the existence of a probability measure $P$ on the subfamily of (unconditional) events*, a peculiarity of the approach to conditional probability based on coherence is that, due to the direct assignment of $P(E|H)$ as a whole, the knowledge – or the assessment – of the "joint" and "marginal" unconditional probabilities $P(E \wedge H)$ and $P(H)$ is not required. Moreover, the conditioning event $H$ – different from the impossible one – may have zero probability, a situation that is not allowed in the classical Kolmogorov's definition.

And what about the Radon–Nikodym procedure? It has been proved in [2] and [1] that there are situations, when $P(H) = 0$, in which any version of the conditional distribution of $P(E|H)$ obtained by the classic Radon–Nikodym procedure can violate a "natural" property such as $P(E|H) = 1$ if $H \subseteq E$.

But even in a simple (and finite) case such as that discussed in our example, invoking Radon–Nikodym procedure cannot help, since the corresponding conditional distribution – as can be easily checked – is anyway "completely free" on sets of measure zero: in fact (referring to the algebra generated by the atoms $A_1, ..., A_4$), we would get, e.g. (since $H_3 = A_3 \vee A_4$),

$$0 = P(E_3 \wedge H_3) = P(E_3|A_3)P(A_3) + P(E_3|A_4)P(A_4)\,,$$

with $P(A_3) = P(A_4) = 0$, and so there are no constraints for $P(E_3|A_3)$ and $P(E_3|A_4)$, while they should instead be "compulsorily" assessed equal to 1 and 0, respectively.

Not to mention that Radon–Nikodym procedure *requires to refer not just to the given conditioning event*, but rather it needs the knowledge of the whole conditional distribution: this circumstance is clearly misleading from an inferential point of view, since a conditional density $P(E|x)$ turns out to depend not only on $x$, but on the whole $\sigma$-algebra in which $x$ is embedded.

**Remark 2** – *Since $P(E_3|H_3) = \frac{1}{2}$ , another peculiarity of the above example is that* a probability equal to 0 *(recall in fact that $P(E_3) = 0$)* can be updated by conditioning *(in this case, with respect to $H_3$, and we get the "new" value $\frac{1}{2}$ ). The same is true for probabilities equal to* 1 *(just consider $E_3^c$ in place of $E_3$).*

In conclusion, there are many aspects (even if we just consider a very simple example) that render coherent conditional probability a nice generalization of the classical Kolmogorov approach.

# 4 A quick glance on conditional probability on MV-algebras

Another way of generalizing Kolmogorov approach is expressed through the possibility of defining conditional probability on an MV-algebra, which is an important many–valued generalization of a Boolean algebra, apt to capture different kinds of uncertainty.

For the main definitions and results (and also for some relevant literature) we refer to [12] and [10].

We just recall here that (just as in the Kolmogorov approach) usually there is no *direct* definition on suitable families of *conditional* events (with a structure of MV-algebra), but a (so–to–say) "preliminary" introduction of (unconditional) probability on a Boolean algebra and of the concept of *state* $m$ on a relevant MV-algebra. Then, for an MV-algebra $M$ with product $\cdot$ (for details, see [12], [10]), conditioning is introduced through the concept of conditional state $m(a|b)$ defined (for $a, b \in M$) by means of a relation that is similar to the product rule for probability, i.e. a conditional state is any solution $m(a|b)$ of the equation

$$m(b)\, m(a|b) = m(a \cdot b)\,.$$

When the MV-algebra reduces to a Boolean algebra, then the above definition of conditional state coincides with the classical (Kolmogorov) definition of conditional probability.

Clearly, this kind of generalization is in a different sense from that based on coherence. Moreover it cannot be of any help as far as the problem of zero probability for conditioning events is concerned.

An approach with a direct definition of conditional probability on an MV-algebras $M$, taking *conditional events* as elements of $M$, is given in [3]. The relevant operations are defined as follows:

$$A|B \oplus C|D = (A \vee C \vee (B \wedge D))\big|(B \vee D)\ ,$$

$$\neg\,(A|B) = B^c|A^c\ .$$

As it is well–known, by suitably combining these two operations we get another binary operation $\otimes$, which reads, in this case, as

$$A|B \otimes C|D = (A \wedge C)\big|((A \wedge D) \vee (B \wedge C))\ ,$$

and the neutral elements for $\oplus$ and $\otimes$ are, respectively, $\emptyset|\emptyset$ and $\Omega|\Omega$.

Among the various results contained in this paper, we mention a re-formulation of the Theorem characterizing coherence (recalled above, in Sect.2) in terms of "tricotomic" *conditional atoms* (for details, see [3]): these give rise to an MV-partition, in the sense defined in [11].

In particular, going back to our simple example, we could easily extend the family $E_i|H_i$ (i=1,2,3) in such a way to make the enlarged family an MV-algebra as that defined in [3], and then check coherence through the Theorem involving "three–valued" atoms.

Yet for our elementary example it would just correspond to the use, for unfolding a nut, of a tank instead of a nutcracker!

So it is better to bear on the principle (the so called *"Ockham's razor"*) that states (in its original Latin form): "Pluralitas non est ponenda sine necessitate", and could be interpreted, for scientists, "when you have two competing theories which make exactly the same prediction, the one that is simpler is the better".

## References

[1] Blackwell D., Dubins L.E. (1975), On existence and non existence of proper, regular, conditional distributions, *The Annals of Probability*, 3, pp. 741–752.

[2] Blackwell D., Ryll–Nardzewski C. (1963), Non–existence of everywhere proper conditional distributions, *Ann. Math. Stat.*, 34, pp. 223–225.

[3] Capotorti A., Vantaggi B. (1999), A general interpretation of conditioning and its implication on coherence, *Soft Computing*, 3, pp. 148–153.

[4] Coletti G. (1994), Coherent Numerical and Ordinal Probabilistic Assessments, *IEEE Transactions on Systems, Man, and Cybernetics*, 24, pp. 1747–1754.

[5] Coletti G., Scozzafava R. (1996), Characterization of Coherent Conditional Probabilities as a Tool for their Assessment and Extension, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 4, pp. 103–127.

[6] Coletti G., Scozzafava R. (2002), *Probabilistic Logic in a Coherent Setting*, Trends in Logic, Vol.15, Kluwer, Dordrecht.

[7] Coletti G., Scozzafava R. (2005), Conditioning in a coherent setting: theory and applications, *Fuzzy Sets and Systems*, 155, pp. 26–49.

[8] de Finetti B. (1935), La logique de la probabilité, in: *Actes du Congrès International de Philosophie Scientifique*, Hermann, Paris, IV, pp. 1–9.

[9] de Finetti B. (1949), Sull'impostazione assiomatica del calcolo delle probabilità, *Annali Univ. Trieste*, 19, pp. 3–55. (Engl. transl.: Ch.5 in *Probability, Induction, Statistics*, Wiley, London, 1972).

[10] Kroupa T. (2005), Conditional probability on MV–algebras, *Fuzzy Sets and Systems*, 149, pp. 369–381.

[11] Mundici D. (1997), Non Boolean partitions and many–valued logic, In: *Proc. IFSA 97*, Prague, vol. 1, pp. 25–29.

[12] Riečan B., Mundici D. (2002), Probability on MV–algebras, In: *Handbook of Measure Theory* (E. Pap, Ed.), pp. 869–909, North–Holland, Amsterdam.

[13] Scozzafava R., Vantaggi B. (2004), The role of zero probabilities in dealing with zero frequency problems, In: *ÃSoft Methodology and Random Information SystemsÃ* (Eds. M.Lopes-Diaz, A.Gil, P.Grzegorzewski, O.Hryniewicz, J.Lawry), Springer, pp. 265–272.

# A New Method for Conditional Independence Inference

**Kentaro Tanaka**
Department of Industrial Engineering and Management
Graduate School of Decision Science and Technology
Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo 152-8552, Japan
tanaka.k.al@m.titech.ac.jp

**Akimichi Takemura**
Department of Mathematical Informatics
Graduate School of Information Science and Technology
University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo 113-8656, Japan
takemura@stat.t.u-tokyo.ac.jp

**Tomonari Sei**
Department of Mathematics
Faculty of Science and Technology
Keio University
Hiyoshi 3-14-1, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan
sei@math.keio.ac.jp

**Abstract**

We introduce a new algebraic method for the implication problem of conditional independence statements. The method is based on an idea that the implication problem can be transformed into an easier problem by adding extra conditional independence statements to a given set of conditional independence statements.

## 1 Introduction

In this paper, we deal with the implication problem of conditional independence statements, that is, testing whether a conditional independence statement is derived from a set of other conditional independence statements.

It is known that there is no finite axiomatic characterization of the conditional independence implication problem for general discrete probability distributions (Stuený [7]). The situation is different if we restrict the class of the conditional independence implication problem. It is known that there exists the finite axiomatic characterization each for the following restricted conditional independence statements: unconditional independence statements (Geiger et al. [1], Matúš [4]); saturated conditional independence statements (Geiger and Pearl [2], Malvestuto and Studený [3]); conditional independence statements represented by Markov networks (Pearl and Paz [5]), and so forth. See Niepert et al. [6] and Studený [8] for the comprehensive description.

Another way to characterize the conditional independence implication problem is based on algebra. The method of imsets by Studený [8] provides a very powerful algebraic method for testing of conditional independence implications. By using imsets, the conditional independence implication problem is translated into relations among integer-valued vectors. In Bouckaert et al. [9], they develop a method of linear programming for computer testing of conditional independence implications. In this paper, we introduce a new algebraic method for the conditional independence implication problem for positive discrete probability distributions.

## 2 Imsets

In the following we only consider the case that probabilities are positive at every point of the sample space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$. We assume that $\mathcal{X}$ is a finite set. Namely we consider the positive probability distributions for $N$-way contingency tables.

For a subset $A \subseteq N$ and the set of random variables $X$ on $\mathcal{X}$, let $X_A$ denote the set of random variables in $A$. Given disjoint $A, B, C \subseteq N$, we abbreviate $X_A \perp\!\!\!\perp X_B \mid X_C$ as $A \perp\!\!\!\perp B \mid C$. For $A, B \subseteq N$, we abbreviate $A \cup B$ as $AB$.

The identifier $\delta_A$ of a set $A \subseteq N$ is defined as

$$\delta_A(B) = \begin{cases} 1, & B = A, \\ 0, & B \neq A, B \subseteq N. \end{cases}.$$

For any triplet of pairwise disjoint subsets $A, B, C \subseteq N$, the *semi-elementary imset* $u_{\langle A,B \mid C \rangle}$ is defined as

$$u_{\langle A,B \mid C \rangle} = \delta_{ABC} + \delta_C - \delta_{AC} - \delta_{BC}.$$

In the context of Studený [8], we can encode a semi-elementary imset $u_{\langle A,B \mid C \rangle}$ as the corresponding conditional independence statement $A \perp\!\!\!\perp B \mid C$. If $A = a$ and $B = b$ are singletons, the imset $u_{\langle a,b \mid C \rangle}$ is called *elementary*. The set of all elementary imsets is denoted by $\mathcal{E}(N)$.

**Example 1.** *We assume that $N = \{1, 2, 3\}$ and $a, b, c$ are desjoint elements in $N$. Let us consider the following implication problem of conditional independence statements.*

$$a \perp\!\!\!\perp b \,|\, c,\ a \perp\!\!\!\perp c \quad \Rightarrow \quad a \perp\!\!\!\perp bc$$

*The following two imsets correspond to $a \perp\!\!\!\perp b \,|\, c$ and $a \perp\!\!\!\perp c$ respectively: $u_{\langle a,b \mid c \rangle} = \delta_{abc} - \delta_{ac} - \delta_{bc} + \delta_c$ and $u_{\langle a,c \mid \rangle} = \delta_{ac} - \delta_a - \delta_c + \delta_\emptyset$. Then we obtain*

$$u_{\langle a,b \mid c \rangle} + u_{\langle a,c \mid \rangle} = \delta_{abc} - \delta_a - \delta_{bc} + \delta_\emptyset = u_{\langle a,bc \mid \rangle}.$$

*This means that $a \perp\!\!\!\perp b \,|\, c$ and $a \perp\!\!\!\perp c$ imply $a \perp\!\!\!\perp bc$.*

Though the usual arguments based on imsets are very powerful, it is known that they may make a mistake in proving conditional independence implication problem in some cases. As an example, we consider the example in Corollary 2.1 and Example 4.1 of Studený[8]:

$$a \perp\!\!\!\perp b \,|\, cd,\ c \perp\!\!\!\perp d \,|\, a,\ c \perp\!\!\!\perp d \,|\, b,\ c \perp\!\!\!\perp d \,|\, \emptyset. \quad \Rightarrow \quad c \perp\!\!\!\perp d \,|\, ab$$

This relation is true. However, it seems to be hard to prove the relation by usual arguments based on imsets. In fact, in Studený [8], it is shown that it is impossible to prove the above relation by direct use of imsets. In the next section, we give a new method to partially overcome this difficulty. Our method broaden the applicability of techniques based on imsets for the conditional independence implication problem.

# 3 A new method for the conditional independence implication problem

The implication problem of conditional independence statements can be represented as the problem of proving (or disproving)

$$\{A_i \perp\!\!\!\perp B_i \,|\, C_i\}_{i=1}^{I} \ \Rightarrow\ A \perp\!\!\!\perp B \,|\, C. \tag{1}$$

The left-hand side is the set of given conditional independence statements.

Let $p$ be a probability function. The following property holds for conditional independence statement.

$$A \perp\!\!\!\perp C \,|\, B \quad \Leftrightarrow \quad p(ABC) = q(AB)r(BC) \text{ for some } q(AB), r(BC). \tag{2}$$

Here for simplicity we are writing only the sets of variables "$ABC$, $AB$, $BC$" and omitting the sample point $x$.

The right-hand side of (2) is a shorthand notation for

$$\begin{aligned}&\exists q, r \text{ depending on } AB\text{- and } BC\text{-marginal such that}\\ &p(ABC;x) = q(AB;x)r(BC;x) \quad \forall x \in \mathcal{X}.\end{aligned} \tag{3}$$

Note that the left-hand side only depends on the $ABC$-marginal cell $x_{ABC}$ (i.e. components of $x$ in the index set $ABC$) of $x$. Similarly $q(AB;x)$ depends only on $x_{AB}$ and $r(BC;x)$ depends only on $x_{BC}$.

In the following we take the logarithm of $p$:

$$\log p(ABC;x) = \log q(AB;x) + \log r(BC;x) \quad \forall x \in \mathcal{X}. \tag{4}$$

For $A \subseteq N$, let $h_A(x) = h_A(x_A)$ be a real-valued function of $x$ depending only on the marginal $x_A$. We can consider $h_A$ as a vector in $\mathbb{R}^{\mathcal{X}}$. Define $L_A = \mathrm{span}\{h_A(S;x)\}$ denote the linear subspace of $\mathbb{R}^{\mathcal{X}}$ spanned by functions depending only on the marginal cell $x_A$. Define

$$L_{\langle A,C \,|\, B\rangle} = L_{AB} + L_{BC}$$

as the (vector) sum of two spaces $L_{AB}$ and $L_{BC}$. We can equivalently write (4) as

$$\log p(ABC; x) \in L_{\langle A,C \mid B\rangle},$$

where we are again considering $\log p(ABC; x)$ as an element of $\mathbb{R}^{\mathcal{X}}$. For everywhere positive probabilities, by (2) we have

$$A \perp\!\!\!\perp C \mid B \quad \Leftrightarrow \quad \log p(ABC; x) \in L_{\langle A,C \mid B\rangle}.$$

Let

$$\begin{aligned} L_{\{A_i \perp\!\!\!\perp B_i \mid C_i\}_{i=1}^I} &= \mathrm{span}(\{\delta_{A_iB_iC_i}(x) - \delta_{A_iC_i}(x) - \delta_{B_iC_i}(x) + \delta_{C_i}(x)\}_{i=1}^I) \\ &= \{\sum_{i=1}^I \alpha_i(\delta_{A_iB_iC_i}(x) - \delta_{A_iC_i}(x) - \delta_{B_iC_i}(x) + \delta_{C_i}(x)) \mid \alpha_i \in \mathbb{R}, i = 1, \ldots, I\} \end{aligned} \tag{5}$$

be a linear subspace of $\mathbb{R}^{\mathcal{X}}$ spanned by "imsets" $\{u_{\langle A_i,B_i \mid C_i\rangle} = \delta_{A_iB_iC_i}(x) - \delta_{A_iC_i}(x) - \delta_{B_iC_i}(x) + \delta_{C_i}(x)\}_{i=1}^I$, now considered as functions of $x$.

Then we have the following proposition.

**Proposition 1.** *If* $(\log p(ABC; x) + L_{\{A_i \perp\!\!\!\perp B_i \mid C_i\}_{i=1}^I}) \cap L_{\langle A,B \mid C\rangle} \neq \emptyset$, *then* (1) *holds.*

Next we introduce our new method. The method is based on an idea: "adding extra conditional independence statements" to a given set of conditional independence statements. Now we add extra conditional independence statements $\{E_j \perp\!\!\!\perp F_j \mid G_j\}_{j=1}^J$ to the implication problem in (1) and consider

$$\{A_i \perp\!\!\!\perp B_i \mid C_i\}_{i=1}^I \cup \{E_j \perp\!\!\!\perp F_j \mid G_j\}_{j=1}^J \;\Rightarrow\; A \perp\!\!\!\perp B \mid C, \tag{6}$$

which may be easier to prove, if true. The question is what kind of extra conditions we can add without affecting the truth (or non-truth) of (1). To answer this question, we make the following definition.

**Definition 1.** *$E \perp\!\!\!\perp F \mid G$ does not bridge $A \perp\!\!\!\perp B \mid C$ if*

$$(EFG) \cap A = \emptyset \quad \text{or} \quad (EFG) \cap B = \emptyset,$$

*i.e., $EFG$ intersects at most one of $A$ and $B$.*

Then we have the following theorem.

**Theorem 1.** *Suppose that we check the implication of* (1) *only by the criterion of 1. Furthermore, assume that each $E_j \perp\!\!\!\perp F_j \mid G_j$ does not bridge $A \perp\!\!\!\perp B \mid C$. Then* (1) *holds if and only if*

$$\{A_i \perp\!\!\!\perp B_i \mid C_i\}_{i=1}^I \cup \{E_j \perp\!\!\!\perp F_j \mid G_j\}_{j=1}^J \;\Rightarrow\; A \perp\!\!\!\perp B \mid C \tag{7}$$

*holds.*

*Proof.* Suppose that (1) holds. Then by the given conditions $\{A_i \perp\!\!\!\perp B_i \,|\, C_i\}_{i=1}^I$ we already have $\log p(ABC; x) \in L_{\langle A,B \,|\, C\rangle}$. Adding $\{E_j \perp\!\!\!\perp F_j \,|\, G_j\}_{j=1}^J$ only adds more decompositions of $\log p(ABC; x)$, so we might have more conditional independences. However $A \perp\!\!\!\perp B \,|\, C$ still holds.

Suppose that we could prove (7) by the technique of Proposition 1. Then for some coefficients $\alpha_i$, $i = 1, \ldots, I$ and $\beta_j$, $j = 1, \ldots, J$ we can write

$$\begin{aligned}\log p(ABC; x) &+ \sum_{i=1}^{I} \alpha_i(\delta_{A_iB_iC_i}(x) - \delta_{A_iC_i}(x) - \delta_{B_iC_i}(x) + \delta_{C_i}(x)) \\ &= h(x) + \sum_{j=1}^{J} \beta_j(\delta_{E_iF_iG_i}(x) - \delta_{E_iG_i}(x) - \delta_{F_iG_i}(x) + \delta_{G_i}(x)), \end{aligned} \tag{8}$$

where $h(x) \in L_{\langle A,B \,|\, C\rangle}$. Note that (8) holds for every $x \in \mathcal{X}$. Note also that $\log p(ABC; x)$ and $h(x)$ depend only on $x_{ABC}$. Fix components of $x$ other than $x_{ABC}$ to particular values, say $x^0_{N\setminus ABC}$ in (8). Writing $x = (x_{ABC}, x^0_{N\setminus ABC})$, we have

$$\begin{aligned}\log p(ABC; x_{ABC}) &+ \sum_{i=1}^{I} \alpha_i(\delta_{A_iB_iC_i}(x_{ABC}, x^0_{N\setminus ABC}) \\ &\quad - \delta_{A_iC_i}(x_{ABC}, x^0_{N\setminus ABC}) - \delta_{B_iC_i}(x_{ABC}, x^0_{N\setminus ABC}) + \delta_{C_i}(x_{ABC}, x^0_{N\setminus ABC})) \\ = h(x_{ABC}) &+ \sum_{j=1}^{J} \beta_j(\delta_{E_iF_iG_i}(x_{ABC}, x^0_{N\setminus ABC}) \\ &\quad - \delta_{E_iG_i}(x_{ABC}, x^0_{N\setminus ABC}) - \delta_{F_iG_i}(x_{ABC}, x^0_{N\setminus ABC}) + \delta_{G_i}(x_{ABC}, x^0_{N\setminus ABC})), \end{aligned} \tag{9}$$

for every $x_{ABC}$. Now by the non-bridging assumption, the right-hand side of (9) belongs to $L_{\langle A,B \,|\, C\rangle}$. Hence, again by Proposition 1, we see that (1) holds.

□

Theorem 1 shows that given $\{A_i \perp\!\!\!\perp B_i \,|\, C_i\}_{i=1}^I$ and $A \perp\!\!\!\perp B \,|\, C$, we can first add every $E \perp\!\!\!\perp F \,|\, G$, which does not bridge $A \perp\!\!\!\perp B \,|\, C$, to the given conditions. Therefore, if (1) is true, then it suffices to prove an easier problem (7).

**Example 2.** *Let us consider the following implication problem:*

$$a \perp\!\!\!\perp b \,|\, cd,\ c \perp\!\!\!\perp d \,|\, a,\ c \perp\!\!\!\perp d \,|\, b,\ c \perp\!\!\!\perp d \,|\, \emptyset. \quad \Rightarrow \quad c \perp\!\!\!\perp d \,|\, ab. \tag{10}$$

*As discussed in the end of the previous section, though this relation is true, it is impossible to derive it by direct use of imsets. Now we show that it is possible to derive it by algebraic manipulation using Proposition 1 and Theorem 1. From Theorem 1, to prove the relation of (10), we can add extra conditional independence statements, such as $a \perp\!\!\!\perp b \,|\, c$, $a \perp\!\!\!\perp b \,|\, d$ and $a \perp\!\!\!\perp b \,|\, \emptyset$, to a given set of conditional independence statements. Then we obtain*

$$u_{\langle a,b \,|\, cd\rangle} + u_{\langle c,d \,|\, a\rangle} + u_{\langle c,d \,|\, b\rangle} - u_{\langle c,d \,|\, \emptyset\rangle} - u_{\langle a,b \,|\, c\rangle} - u_{\langle a,b \,|\, d\rangle} + u_{\langle a,b \,|\, \emptyset\rangle} = u_{\langle c,d \,|\, ab\rangle}.$$

*From this, we have*

$$(\log p(abcd; x) + L_{\{A_i \perp\!\!\!\perp B_i \,|\, C_i\}_{i=1}^{I}} + L_{\{E_j \perp\!\!\!\perp F_j \,|\, G_j\}_{j=1}^{J}}) \cap L_{\langle c,d \,|\, ab\rangle} \neq \emptyset.$$

*Therefore, from Proposition 1, the relation of* (10) *is true.*

As in Bouckaert et al. [9], we can formulate the implication problem of conditional independence statements as a problem of finding a solution of the system of linear equations.

**Corollary 1.** *Let us consider the conditional independence implication problem of (1). Assume that each $E_j \perp\!\!\!\perp F_j \,|\, G_j$ does not bridge $A \perp\!\!\!\perp B \,|\, C$. Then the relation of (1) is true, if the following system of linear equation*

$$\sum_{i=1}^{I} \mu_i \cdot u_{\langle A_i, B_i \,|\, C_i\rangle} + \sum_{j=1}^{J} \lambda_j \cdot u_{\langle E_j, F_j \,|\, G_j\rangle} = u_{\langle A, B \,|\, C\rangle}$$

*has a solution in $\{\mu_i\}_{i=1}^{I}, \{\lambda_j\}_{j=1}^{J}$.*

## References

[1] Dan Geiger, Azaria Paz, Judea Pearl, Axioms and algorithms for inferences involving probabilistic independence, Information and Computation Volume 91, Issue 1, Pages 128–141, 1991.

[2] Dan Geiger, Judea Pearl, Logical and algorithmic properties of conditional independence and graphical models, *The Annals of Statistics* Volume 21, Issue 4, Pages 2001–2021, 1993.

[3] Francesco M. Malvestuto, A unique formal system for binary decomposition of database relations, probability distributions and graphs, *Information Sciences*, Volume 59, Pages 21–52, 1992 + Francesco M. Malvestuto, M. Studený Comment on "A unique formal ... graphs", *Information Sciences*, Volume 63, Pages 1–2, 1992.

[4] František Matúš Stochastic independence, algebraic independence and abstract connectedness, *Theoretical Computer Science*, Volume 134, Issue 2, Pages 455–471, 1994.

[5] Judea Pearl, Azaria Paz, Graphoids, graph-based logic for reasoning about relevance relations, *Advances in Artificial Intelligence*, Volume II (B. Du Boulay, D. Hogg, L. Steels eds.), North-Holland, Elsevier, Pages 357-363, 1987.

[6] Mathias Niepert, Dirk Van Gucht, and Marc Gyssens, Logical and Algorithmic Properties of Stable Conditional Independence, *International Journal of Approximate Reasoning*, Volume 51, Issue 5, pages 531–543, 2010.

[7] Milan Studený. Conditional independence relations have no finite complete characterization. *Information Theory, Statistical Decision Functions and Random Processes, Transactions of the 11th Prague Conference*, Volume B (S. Kubík, J. Á. Vísěk eds.), Kluwer, Pages 377–396, 1992.

[8] Milan Studený. *Probabilistic Conditional Independence Structures*, Springer-Verlag, London, 2005.

[9] Remco Bouckaert, Raymond Hemmecke, Silvia Lindner, and Milan Studený. Efficient Algorithms for Conditional Independence Inference, *Journal of Machine Learning Research*, Volume 11, Pages 3453–3479, 2010.

# On Weakness of Evidential Networks*

**Jiřina Vejnarová**
Institute Information Theory and Automation
Academy of Sciences of the Czech Republic
vejnar@utia.cas.cz

**Abstract**

In evidence theory several counterparts of Bayesian networks based on different paradigms have been proposed. We will present, through simple examples, problems appearing in two kinds of these models caused either by the conditional independence concept (or its misinterpretation) or by the use of a conditioning rule. The latter kind of problems can be avoided if undirected models are used instead.

## 1 Introduction

When applying models of artificial intelligence to any practical problem one must cope with two basic problems: uncertainty and multidimensionality. The most widely used models managing these issues are, at present, so-called *probabilistic graphical Markov models*.

The problem of multidimensionality is solved in these models with the help of the notion of conditional independence, which enables factorization of a multidimensional probability distribution into small parts, usually marginal or conditional low-dimensional distributions (e.g. in *Bayesian networks*), or generally into low-dimensional factors (e.g. in *decomposable models*). Such a factorization not only decreases the storage requirements for representation of a multidimensional distribution but it usually also induces efficient computational procedures allowing inference from these models.

Probably the most popular representative of these models are *Bayesian networks*, while from the computational point of view so-called *decomposable models* are the most advantageous. Naturally, several attempts to construct an analogy of Bayesian networks have also been made in other frameworks as e.g. in possibility theory [5], evidence theory [4] or in the more general frameworks of valuation-based systems [13] and credal sets [7], while counterparts of decomposable models are, more or less, omitted.

In this contribution we will confine ourselves to evidence theory, where several counterparts of Bayesian networks based on different paradigms have been proposed

*The support of Grant GAČR 201/08/0539 is gratefully acknowledged.

[4, 13, 19]. We will present, through two simple examples, problems appearing in these models caused either by the conditional independence concept (or its misinterpretation) or by the use of different conditioning rules. The latter kind of problems can be avoided if undirected models are used instead.

## 2 Basic Concepts

In this section we will briefly recall basic concepts from evidence theory [12] concerning sets and set functions.

### 2.1 Set Projections and Joins

For an index set $N = \{1, 2, \ldots, n\}$ let $\{X_i\}_{i \in N}$ be a system of variables, each $X_i$ having its values in a finite set $\mathbf{X}_i$. In this paper we will deal with *multidimensional frame of discernment* $\mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_n$, and its *subframes* (for $K \subseteq N$) $\mathbf{X}_K = \times_{i \in K} \mathbf{X}_i$. When dealing with groups of variables on these subframes, $X_K$ will denote a group of variables $\{X_i\}_{i \in K}$ throughout the paper.

For $M \subset K \subseteq N$ and $A \subset \mathbf{X}_K$, $A^{\downarrow M}$ will denote a *projection* of $A$ into $\mathbf{X}_M$:

$$A^{\downarrow M} = \{y \in \mathbf{X}_M \mid \exists x \in A : y = x^{\downarrow M}\},$$

where, for $M = \{i_1, i_2, \ldots, i_m\}$,

$$x^{\downarrow M} = (x_{i_1}, x_{i_2}, \ldots, x_{i_m}) \in \mathbf{X}_M.$$

In addition to the projection, in this text we will also need an opposite operation, which will be called a join. By a *join*[1] of two sets $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_L$ $(K, L \subseteq N)$ we will understand a set

$$A \bowtie B = \{x \in \mathbf{X}_{K \cup L} : x^{\downarrow K} \in A \ \ \& \ \ x^{\downarrow L} \in B\}.$$

Let us note that for any $C \subseteq \mathbf{X}_{K \cup L}$ naturally $C \subseteq C^{\downarrow K} \bowtie C^{\downarrow L}$, but generally $C \neq C^{\downarrow K} \bowtie C^{\downarrow L}$.

### 2.2 Set Functions

In evidence theory [12] two dual measures are used to model the uncertainty: belief and plausibility measures. Both of them can be defined with the help of another set function called a *basic (probability* or *belief) assignment* $m$ on $\mathbf{X}_N$, i.e. ,

$$m : \mathcal{P}(\mathbf{X}_N) \longrightarrow [0, 1],$$

where $\mathcal{P}(\mathbf{X}_N)$ is the power set of $\mathbf{X}_N$, and $\sum_{A \subseteq \mathbf{X}_N} m(A) = 1$. Furthermore, we assume that $m(\emptyset) = 0$.[2]

[1] This term and notation are taken from the theory of relational databases [1].

[2] This assumption is not generally accepted, e.g. in [2] it is omitted. The consequences of this omission will be mentioned several times throughout this paper.

A set $A \in \mathcal{P}(\mathbf{X}_N)$ is a *focal element* if $m(A) > 0$. Let $\mathcal{F}$ denote the set of all focal elements, a focal element $A \in \mathcal{F}$ is called an $m-atom$ if for any $B \subseteq A$ either $B = A$ or $B \notin \mathcal{F}$. In other words, $m-atom$ is a setwise-minimal focal element.

*Belief* and *plausibility measures* are defined for any $A \subseteq \mathbf{X}_N$ by the equalities

$$Bel(A) = \sum_{B \subseteq A} m(B), \qquad Pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \tag{1}$$

respectively. It is well-known (and evident from these formulae) that for any $A \in \mathcal{P}(\mathbf{X}_N)$

$$Bel(A) \leq Pl(A), \qquad Pl(A) = 1 - Bel(A^C), \tag{2}$$

where $A^C$ is the set complement of $A \in \mathcal{P}(\mathbf{X}_N)$. Furthermore, basic assignment can be computed from belief function via Möbius inverse:

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} Bel(B), \tag{3}$$

i.e. any of these three functions is sufficient to define values of the remaining two.

For a basic assignment $m$ on $\mathbf{X}_K$ and $M \subset K$ a *marginal basic assignment* of $m$ is defined (for each $A \subseteq \mathbf{X}_M$):

$$m^{\downarrow M}(A) = \sum_{B \subseteq \mathbf{X}_K : B^{\downarrow M} = A} m(B).$$

# 3 Conditioning

Conditioning belongs to the most important topics of any theory dealing with uncertainty. From the viewpoint of the construction of Bayesian-network-like multidimensional models it seems to be inevitable.

## 3.1 Conditioning of Events

In evidence theory the "classical" conditioning rule is the so-called *Dempster's rule of conditioning* defined for any $\emptyset \neq A \subseteq \mathbf{X}_N$ and $B \subseteq \mathbf{X}_N$ such that $Pl(B) > 0$ by the formulae

$$\begin{aligned} Bel(A|_D B) &= \frac{Bel(A \cup B^C) - Bel(B^C)}{1 - Bel(B^C)}, \\ Pl(A|_D B) &= \frac{Pl(A \cap B)}{Pl(B)}. \end{aligned} \tag{4}$$

Let us note that in [2] a bit different formulae are used: conditional beliefs and plausibilities are not normalized. It corresponds to the omission of the assumption $m(\emptyset) = 0$.

This is not the only possibility how to condition, another — in a way symmetric — conditioning rule is the following one, called *focusing* defined for any $\emptyset \neq A \subseteq \mathbf{X}_N$ and $B \subseteq \mathbf{X}_N$ such that $Bel(B) > 0$ by the formulae

$$\begin{aligned} Bel(A|_F B) &= \frac{Bel(A \cap B)}{Bel(B)}, \\ Pl(A|_F B) &= \frac{Pl(A \cup B^C) - Pl(B^C)}{1 - Pl(B^C)}. \end{aligned} \tag{5}$$

Formulae (4) and (5) are, in a way, evidential counterparts of conditioning in probabilistic framework. Let us note that the seemingly "natural" way of conditioning

$$m(A|_P B) = \frac{m(A \cap B)}{m(B)} \tag{6}$$

is not possible, since $m(A|_P B)$ need not be a basic assignment. It is caused by a simple fact that $m$, in contrary to $Bel$ and $Pl$, is not monotonous with respect to set inclusion.

## 3.2 Conditional Variables

However, from the viewpoint of evidential networks conditioning of variables is of primary interest. In [18] we presented two definitions of conditioning by variables, based on Dempster conditioning rule and focusing, we proved that these definitions are correct, nevertheless, their usefulness for multidimensional models is rather questionable, as thoroughly discussed in the above-mentioned paper.

Therefore, in [19] we proposed a new conditioning rule which is, in a way, a generalization of (6).

**Definition 1** *Let $X_K$ and $X_L$ $(K \cap L = \emptyset)$ be two groups of variables with values in $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively. Then the conditional basic assignment of $X_K$ given $X_L \in B \subseteq \mathbf{X}_L$ (for $B$ such that $m^{\downarrow L}(B) > 0$) is defined as follows:*

$$m_{X_K|_P X_L}(A|_P B) = \frac{\displaystyle\sum_{\substack{C \subseteq \mathbf{X}_{K \cup L}: \\ C^{\downarrow K} = A \& C^{\downarrow L} = B}} m(C)}{m^{\downarrow L}(B)} \tag{7}$$

*for any $A \subseteq \mathbf{X}_K$.*

Although we said above, that it makes little sense for conditioning of events, it is sensible in conditioning of variables, as expressed by Theorem 1 proven in [19]. The above-mentioned problem of non-monotonicity is avoided, because a marginal basic assignment is always greater than (or equal to) the joint one.

**Theorem 1** *The set function $m_{X_K|_P X_L}$ defined for any fixed $B \subseteq \mathbf{X}_L$, such that $m^{\downarrow L}(B) > 0$ by Definition 1 is a basic assignment on $\mathbf{X}_K$.*

# 4 Conditional Independence and Irrelevance

Independence and irrelevance need not be (and usually are not) distinguished in the probabilistic framework, as they are almost equivalent to each other. Similarly, in possibilistic framework adopting De Cooman's measure-theoretical approach [9] (particularly his notion of almost everywhere equality) we proved that the analogous concepts are equivalent (for more details see [15]).

## 4.1 Independence

In evidence theory the most common notion of independence is that of random set independence [6]. It has already been proven [16] that it is also the only sensible one.

**Definition 2** Let $m$ be a basic assignment on $\mathbf{X}_N$ and $K, L \subset N$ be disjoint. We say that groups of variables $X_K$ and $X_L$ are *independent with respect to a basic assignment* $m$ (in notation $K \perp\!\!\!\perp L\ [m]$) if

$$m^{\downarrow K\cup L}(A) = m^{\downarrow K}(A^{\downarrow K}) \cdot m^{\downarrow L}(A^{\downarrow L})$$

for all $A \subseteq \mathbf{X}_{K\cup L}$ for which $A = A^{\downarrow K} \times A^{\downarrow L}$, and $m(A) = 0$ otherwise.

This notion can be generalized in various ways [3, 13, 16]; the concept of conditional non-interactivity from [3], based on conjunctive combination rule, is used for construction of directed evidential networks in [4] (cf. also Section 5.3). In this paper we will use the concept introduced in [10, 16], as we consider it more suitable: in contrary to other conditional independence concepts [3, 13] it is *consistent with marginalization* [14], in other words, the multidimensional model of conditionally independent variables keeps the original marginals (for more details see [16]).

**Definition 3** Let $m$ be a basic assignment on $\mathbf{X}_N$ and $K, L, M \subset N$ be disjoint, $K \neq \emptyset \neq L$. We say that groups of variables $X_K$ and $X_L$ are *conditionally independent given $X_M$ with respect to $m$* (and denote it by $K \perp\!\!\!\perp L|M\ [m]$), if the equality

$$m^{\downarrow K\cup L\cup M}(A) \cdot m^{\downarrow M}(A^{\downarrow M}) \quad = \quad m^{\downarrow K\cup M}(A^{\downarrow K\cup M}) \cdot m^{\downarrow L\cup M}(A^{\downarrow L\cup M})$$

holds for any $A \subseteq \mathbf{X}_{K\cup L\cup M}$ such that $A = A^{\downarrow K\cup M} \bowtie A^{\downarrow L\cup M}$, and $m(A) = 0$ otherwise.

It has been proven in [16] that this conditional independence concept satisfies so-called the semi-graphoid properties taken as reasonable to be valid for any conditional independence concept and it has been shown in which sense this conditional independence concept is superior to previously introduced ones [3, 13].

## 4.2 Irrelevance

Irrelevance is usually considered to be a weaker notion than independence (see e.g. [6]). It expresses the fact that a new piece of evidence concerning one variable cannot influence the evidence concerning the other variable, in other words is irrelevant to it.

More formally: group of variables $X_L$ is *irrelevant* to $X_K$ ($K \cap L = \emptyset$) if for any $B \subseteq \mathbf{X}_L$ such that the left-hand side of the equality is defined

$$m_{X_K|X_L}(A|B) = m(A) \tag{8}$$

for any $A \subseteq \mathbf{X}_K$.[3]

It follows from the definition of irrelevance that it need not be a symmetric relation. Let us note, that in the framework of evidence theory neither irrelevance based on Dempster conditioning rule nor that based on focusing even in cases when the relation is symmetric, imply independence, as can be seen from examples in [18].

Generalization of this notion to conditional irrelevance may be done as follows. Group of variables $X_L$ is *conditionally irrelevant* to $X_K$ given $X_M$ ($K, L, M$ disjoint, $K \neq \emptyset \neq L$) if

$$m_{X_K|X_L X_M}(A|B) = m_{X_K|X_M}(A|B^{\downarrow M}) \tag{9}$$

is satisfied for any $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_{L\cup M}$, such that both sides are defined.

Let us note that the conditioning in equalities (8) and (9) stands for an abstract conditioning rule (any of those mentioned in the previous section or some other [8]). However, the validity of (8) and (9) may depend on the choice of the conditioning rule.

## 4.3 Relationship Between Independence and Irrelevance

As mentioned at the end of preceding section, different conditioning rules lead to different irrelevance concepts. Nevertheless, when studying the relationship between (conditional) independence and irrelevance based on Dempster conditioning rule and focusing we realized that they do not differ too much from each other, as suggested by the following summary.

For both conditioning rules:

- Irrelevance is implied by independence.
- Irrelevance does not imply independence.
- Irrelevance is not symmetric, in general.
- Even in case of symmetry it does not imply independence.
- Conditional independence does not imply conditional irrelevance.

The only difference between these conditioning rules is expressed by the following theorem proven in [18].

**Theorem 2** *Let $X_K$ and $X_L$ be conditionally independent groups of variables given $X_M$ under joint basic assignment $m$ on $\mathbf{X}_{K\cup L\cup M}$ ($K, L, M$ disjoint, $K \neq \emptyset \neq L$). Then*

$$m_{X_K|_F X_L X_M}(A|_F B) = m_{X_K|_F X_M}(A|_F B^{\downarrow M}) \tag{10}$$

*for any $m^{\downarrow L\cup M}$-atom $B \subseteq \mathbf{X}_{L\cup M}$ such that $B^{\downarrow M}$ is $m^{\downarrow M}$-atom and $A \subseteq \mathbf{X}_K$.*

[3] Let us note that somewhat weaker definition of irrelevance one can found in [2], where equality is substituted by proportionality. This notion has been later generalized using conjunctive combination rule [3].

From this point of view focusing seems to be slightly superior to Dempster conditioning rule, but still it is not satisfactory. However, the new conditioning rule introduced by Definition 1 is more promising, as suggested by the following theorem, proven in [19].

**Theorem 3** *Let $K, L, M$ be disjoint subsets of $N$ such that $K, L \neq \emptyset$. If $X_K$ and $X_L$ are independent given $X_M$ (with respect to a joint basic assignment $m$ defined on $X_{K\cup L\cup M}$), then $X_L$ is irrelevant to $X_K$ given $X_M$ under the conditioning rule given by Definition 1.*

The reverse implication is not valid in general, which expresses the expected property: conditional independence is stronger than conditional irrelevance.

However, in Bayesian networks also the reverse implication plays an important role, as for the inference, the network is usually transformed into a decomposable model. Nevertheless, the following assertion proven in [20] holds true.

**Theorem 4** *Let $K, L, M$ be disjoint subsets of $N$ such that $K, L \neq \emptyset$ and $m_{X_K|_P X_{L\cup M}}$ be a (given) conditional basic assignment of $X_K$ given $X_{L\cup M}$ and $m_{X_{L\cup M}}$ be a basic assignment of $X_{L\cup M}$. If $X_L$ is irrelevant to $X_K$ given $X_M$ under the conditioning rule given by Definition 1, then $X_K$ and $X_L$ are independent given $X_M$ (with respect to a joint basic assignment $m = m_{X_K|_P X_{L\cup M}} \cdot m_{X_{L\cup M}}$ defined on $\mathbf{X}_{K\cup L\cup M}$).*

# 5 (Directed) Evidential Networks and Compositional Models

In this section we will deal with directed evidential networks [4] and evidential networks [20]. These two models differ not only by the conditioning rule, but also, and it seems to be more important, by the interpretation of graph structure of the model.

While in evidential networks conditional basic assignment is assigned to every node given its parents (analogously to Bayesian networks), in directed evidential networks conditional beliefs are assigned to arcs, i.e. to every node as many conditionals are assigned as is the number of its parents. These conditionals are subsequently combined by the conjunctive combination rule.

The difference between directed evidential networks and compositional models will be described in Section 5.3 by a simple example, while the lost of information in evidential networks (in comparison with compositional models) in Section 5.4. Before doing that we need to recall the concept of compositional models.

## 5.1 Compositional models

Compositional models are based on the concept of the operator of composition of basic assignments, introduced in [11] in the following way.

**Definition 4** *For two arbitrary basic assignments $m_1$ on $\mathbf{X}_K$ and $m_2$ on $\mathbf{X}_L$ a* composition *$m_1 \triangleright m_2$ is defined for all $C \subseteq \mathbf{X}_{K\cup L}$ by one of the following expressions:*

**(a)** *if* $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) > 0$ *and* $C = C^{\downarrow K} \bowtie C^{\downarrow L}$ *then*

$$(m_1 \triangleright m_2)(C) = \frac{m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L})}{m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L})};$$

**(b)** *if* $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) = 0$ *and* $C = C^{\downarrow K} \times \mathbf{X}_{L\setminus K}$ *then*

$$(m_1 \triangleright m_2)(C) = m_1(C^{\downarrow K});$$

**(c)** *in all other cases*

$$(m_1 \triangleright m_2)(C) = 0.$$

From the basic properties of this operator (proven in [10, 11]) it follows that operator of composition is not commutative in general, but it preserves first marginal (in case of projective basic assignments both of them). In both these aspects it differs from conjunctive combination rule. Furthermore, operator of composition is not associative and therefore its iterative applications must be made carefully, as we will see later.

A lot of other properties possessed by the operator of composition can be found in [10, 11], nevertheless here we will confine ourselves to the following theorem (proven in [10]) expressing the relationship between conditional independence and operator of composition.

**Theorem 5** *Let* $m$ *be a joint basic assignment on* $\mathbf{X}_M$*,* $K, L \subseteq M$*. Then* $(K \setminus L) \perp\!\!\!\perp (L \setminus K)|(K \cap L)\ [m]$ *if and only if*

$$m^{\downarrow K\cup L}(A) = (m^{\downarrow K} \triangleright m^{\downarrow L})(A)$$

*for any* $A \subseteq \mathbf{X}_{K\cup L}$*.*

Now, let us consider a system of low-dimensional basic assignments $m_1, m_2, \ldots, m_n$ defined on $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}, \ldots, \mathbf{X}_{K_n}$, respectively. Composing them together by multiple application of the operator of composition, one gets multidimensional a basic assignment on $\mathbf{X}_{K_1\cup K_2\cup\ldots\cup K_n}$. However, since we know that the operator of composition is neither commutative nor associative, we have to properly specify what "composing them together" means.

To avoid using too many parentheses let us make the following convention. Whenever we write the expression $m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n$ we will understand that the operator of composition is performed successively from left to right:[4]

$$m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n = (\ldots((m_1 \triangleright m_2) \triangleright m_3) \triangleright \ldots) \triangleright m_n. \tag{11}$$

Therefore, multidimensional model (11) is specified by an ordered sequence of low-dimensional basic assignments — a *generating sequence* $m_1, m_2, \ldots, m_n$.

[4] Naturally, if we want to change the ordering in which the operators are to be performed we will do so using parentheses.

## 5.2 Evidential network generated by a perfect sequence

From the point of view of artificial intelligence models used to represent knowledge in a specific area of interest, a special role is played by the so-called *perfect sequences*, i.e., generating sequences $m_1, m_2, \ldots, m_n$, for which

$$\begin{aligned} m_1 \triangleright m_2 &= m_2 \triangleright m_1, \\ m_1 \triangleright m_2 \triangleright m_3 &= m_3 \triangleright (m_1 \triangleright m_2), \\ &\vdots \\ m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n &= m_n \triangleright (m_1 \triangleright \ldots \triangleright m_{n-1}). \end{aligned}$$

The property explaining why we call these sequences "perfect" is expressed by the following assertion proven in [10].

**Theorem 6** *A generating sequence $m_1, m_2, \ldots, m_n$ is perfect if and only if all assignments $m_1, m_2, \ldots, m_n$ are marginal assignments of the multidimensional assignment $m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n$:*

$$(m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n)^{\downarrow K_j} = m_j,$$

*for all $j = 1, \ldots, n$.*

Now, let us recall a simple algorithm for the construction of an evidential network from a perfect sequence of basic assignments [17].

Having a perfect sequence $m_1, m_2, \ldots, m_n$ ($m_\ell$ being the basic assignment of $X_{K_\ell}$), we first order all the variables for which at least one of the basic assignments $m_\ell$ is defined in such a way that first we order (in an arbitrary way) variables for which $m_1$ is defined, then variables from $m_2$ which are not contained in $m_1$, etc.[5] Finally we have

$$\{X_1, X_2, X_3, \ldots, X_k\} = \{X_i\}_{i \in K_1 \cup \ldots \cup K_n}.$$

Then we get a graph of the constructed evidential network in the following way:

1. the nodes are all the variables $X_1, X_2, X_3, \ldots, X_k$;
2. there is an edge $(X_i \to X_j)$ if there exists a basic assignment $m_\ell$ such that both $i, j \in K_\ell$, $j \notin K_1 \cup \ldots \cup K_{\ell-1}$ and either $i \in K_1 \cup \ldots \cup K_{\ell-1}$ or $i < j$.

Evidently, for each $j$ the requirement $j \in K_\ell$, $j \notin K_1 \cup \ldots \cup K_{\ell-1}$ is met exactly for one $\ell \in \{1, \ldots, n\}$. It means that all the parents of node $X_j$ must be from the respective set $\{X_i\}_{i \in K_\ell}$ and therefore the necessary conditional basic assignments $m_{j|pa(j)}$ can easily be computed from basic assignment $m_\ell$ via (7).

It is also evident, that if both $i$ and $j$ are in the same basic assignment and not in previous ones, then the direction of the arc depends only on the ordering of the variables. This might lead to different independences, nevertheless, the following theorem proven in [17] sets forth that any of them is induced by the perfect sequence.

[5] Let us note that variables $X_1, X_2, \ldots, X_k$ may be ordered arbitrarily, nevertheless, for the above ordering proof of Theorem 7 is simpler than in the general case.

Table 1: Basic assignments $m_i$ and conditional basic assignments $m_{.|i}$.

| $A \subseteq \mathbf{C}_i$ | $m_i(A)$ |
|---|---|
| $\{h_i\}$ | 0.49 |
| $\{t_1\}$ | 0.49 |
| $\{h_1, t_1\}$ | 0.02 |

| $D \subseteq \mathbf{B}$ | $m_{.\|i}(D)$ |
|---|---|
| $\{b\}$ | 0.49 |
| $\{\bar{b}\}$ | 0.49 |
| $\{b, \bar{b}\}$ | 0.02 |

Table 2: Joint basic assignment $m$ of variables $C_1, C_2$ and $B$.

| $m$ | $\{b\}$ | | | $\{\bar{b}\}$ | | | $\{b, \bar{b}\}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\{h_2\}$ | $\{t_2\}$ | $\{h_2, t_2\}$ | $\{h_2\}$ | $\{t_2\}$ | $\{h_2, t_2\}$ | $\{h_2\}$ | $\{t_2\}$ | $\{h_2, t_2\}$ |
| $\{h_1\}$ | 0.24 | 0 | 0 | 0 | 0.24 | 0 | 0 | 0 | 0.01 |
| $\{t_1\}$ | 0 | 0.24 | 0 | 0.24 | 0 | 0 | 0 | 0 | 0.01 |
| $\{h_1, t_1\}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0.01 | 0.01 | $\sim 0$ |

**Theorem 7** *For a belief network defined by the above procedure the following independence statements are satisfied for any $j = 2, \ldots k$:*

$$\{j\} \perp\!\!\!\perp (\{i < j\} \setminus pa(j)) \mid pa(j). \tag{12}$$

### 5.3 Example: two coins toss

Let us consider two fair coins toss expressed by variables $C_1$ and $C_2$ with values in $\mathbf{C}_1$ and $\mathbf{C}_2$, respectively ($\mathbf{C}_i = \{h_i, t_i\}$), and the basic assignments $m_1$ and $m_2$ (contained in the left part of Table 1) expressing the fact that the result of any of the coins may from time to time be unknown. The results of tossing two coins are usually considered to be independent, therefore the joint basic assignment $m_{12}$ is just a product of these $m_1$ and $m_2$ (cf. definition of random set independence at the beginning of Section 4).

Now, let us consider one more variable $B$ expressing the fact the bell is ringing, i.e. $\mathbf{B} = \{b, \bar{b}\}$. It happens only if the result on both coins is the same (two heads or two tails). It is evident, that $B$ depends on both $C_1$ and $C_2$, which corresponds to the graph in Figure 5.3 and (due to deterministic dependence of the values of $B$ on the values of $C_1$ and $C_2$) the joint basic assignment of the three variables is in Table 2. The above-mentioned graph can easily be obtained from perfect sequence of basic assignments $m_1, m_2$ and $m_3 \equiv m$ (contained in Tables 1 and 2) via the algorithm presented in the preceding section.

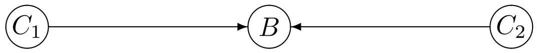


Figure 1: Graph $G$ from Example: two coin toss.

Table 3: Joint basic assignment of variables $C_1, C_2$ and $B$ based conjunctive combination rule; $b^*$ stands for either $b$ or $\bar{b}$.

| $m$ | $\{b^*\}$ | | | $\{b, \bar{b}\}$ | | |
|---|---|---|---|---|---|---|
| | $\{h_2\}$ | $\{t_2\}$ | $\{h_2, t_2\}$ | $\{h_2\}$ | $\{t_2\}$ | $\{h_2, t_2\}$ |
| $\{h_1\}$ | 0.0624 | 0.0624 | 0.0025 | 0.0001 | 0.0001 | $\sim 0$ |
| $\{t_1\}$ | 0.0624 | 0.0624 | 0.0025 | 0.0001 | 0.0001 | $\sim 0$ |
| $\{h_1, t_1\}$ | 0.0025 | 0.0025 | 0.0001 | $\sim 0$ | $\sim 0$ | $\sim 0$ |

The approach suggested by Ben Yaghlane et al. [4] is completely different. The authors start from belief functions of $C_1$ and $C_2$ and conditional belief functions of $B$ given $C_1$ and $C_2$, respectively. To make the difference between these two approaches more apparent we will use basic assignments instead of belief functions (belief functions, nevertheless, can be easily obtained from them by (1)). The conditional basic assignments of $B$ given $C_1$ and $C_2$, respectively, can be found in the right part of Table 1. Let us note that these conditional basic assignments do not depend on the condition, as the results of tossing two coins are independent and therefore also the event that the bell rings does not depend on the result at one coin.

The values of joint basic assignments is computed from Tables 1 using (non-normalized) conjunctive combination rule. Results of these computations can be found in Table 3.

It is evident that the independence (non-interactivity) between coins $C_1$ and $C_2$ is not valid any more — it has been substituted by conditional non-interactivity, which does not make a sense, as $C_1$ is strongly dependent on $C_2$ whenever $B$ is known.

### 5.4 Evidential Network vs Compositional Model

Theorem 3 makes it possible to define evidential networks in a way analogous to Bayesian networks, but simultaneously brings a question: are these networks advantageous in comparison with other multidimensional models in this framework? The following example brings, at least partial, answer to this question.

**Example 1** Let $X_1, X_2$ and $X_3$ be three binary variables with values in $\mathbf{X}_i = \{a_i, \bar{a}_i\}, i = 1, 2, 3$, and $m$ be a basic assignment on $\mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3$ defined as follows

$$\begin{aligned} &m(\mathbf{X}_1 \times \mathbf{X}_2 \times \{\bar{a}_3\}) = .5, \\ &m(\{(a_1, a_2, \bar{a}_3), (\bar{a}_1, \bar{a}_2, a_3)\}) = .5. \end{aligned}$$

Variables $X_1$ and $X_2$ are conditionally independent given $X_3$ with respect to $m$. Therefore also $X_2$ is irrelevant to $X_1$ given $X_3$, i.e.

$$m_{X_1|X_{23}}(A|B) = m_{X_1|X_3}(A|B^{\downarrow\{3\}}), \tag{13}$$

for any focal element $B$ of $m^{\downarrow\{23\}}$. As both $m^{\downarrow\{23\}}$ and $m^{\downarrow\{3\}}$ have only two focal elements, namely $\mathbf{X}_2 \times \{\bar{a}_3\}$ and $\{(a_2, \bar{a}_3), (\bar{a}_2, a_3)\}$ and $\{\bar{a}_3\}$ and $\mathbf{X}_3$, respectively, we have

$$\begin{aligned} m_{X_1|_PX_{23}}(\mathbf{X}_1|\mathbf{X}_2\times\{\bar{a}_3\}) &= m_{X_1|_PX_3}(\mathbf{X}_1|\{\bar{a}_3\}) = 1, && (14)\\ m_{X_1|_PX_{23}}(\mathbf{X}_1|\{(a_2,\bar{a}_3),(\bar{a}_2,a_3)\}) &= m_{X_1|_PX_3}(\mathbf{X}_1|\mathbf{X}_3) = 1. && (15)\end{aligned}$$

Using these conditionals and the marginal basic assignment $m^{\downarrow\{23\}}$ we get a basic assignment $\tilde{m}$ different from the original one, namely

$$\begin{aligned} &\tilde{m}(\mathbf{X}_1\times\mathbf{X}_2\times\{\bar{a}_3\}) = .5,\\ &\tilde{m}(\mathbf{X}_1\times\{(a_2,\bar{a}_3),(\bar{a}_2,a_3)\}) = .5.\end{aligned}$$

Furthermore, if we interchange $X_1$ and $X_2$ we get yet another model, namely

$$\begin{aligned} &\hat{m}(\mathbf{X}_1\times\mathbf{X}_2\times\{\bar{a}_3\}) = .5,\\ &\hat{m}(\mathbf{X}_2\times\{(a_1,\bar{a}_3),(\bar{a}_1,a_3)\}) = .5. \qquad \diamondsuit\end{aligned}$$

The conditional independence of $X_1$ and $X_2$ given $X_3$ and relation (13) correspond to a directed graph in Figure 5.4, which leads to the following system of (conditional) basic assignments:

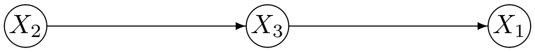


Figure 2: Graph $G$ from Example 1.

$$\begin{aligned} &m^{\downarrow 2}(\mathbf{X}_2) = 1,\\ &m_{X_3|_PX_2}(\{\bar{a}_3\}|\mathbf{X}_2) = m_{X_3|_PX_2}(\mathbf{X}_2|\mathbf{X}_2) = 1,\end{aligned}$$

and $m_{X_1|_PX_3}$ as suggested in right-hand side of (14) and (15).

The final model

$$\begin{aligned} &\check{m}(\mathbf{X}_1\times\mathbf{X}_2\times\{\bar{a}_3\}) = .5,\\ &\check{m}(\mathbf{X}_1\times\mathbf{X}_2\times\mathbf{X}_3) = .5.\end{aligned}$$

is again different, as instead of basic assignment $m^{\downarrow 23}$ (as in Example 1) we used its marginal and conditional.

Therefore it is evident, that evidential networks are less powerful than e.g. compositional models [10], as any of these threedimensional basic assignments can be obtained from two twodimensional ones using the operator of composition.

# 6 Conclusions

This contribution was devoted to two kinds of multidimensional models with directed graph structure, namely directed evidential networks and evidential networks.

In directed evidential networks the graph structure is used in different sense than in Bayesian networks (it resembles rather so-called pseudobayesian networks), which may lead to senseless results, as we presented by a simple example.

Evidential networks, in contrary, keep the sense of the graphical structure known from Bayesian networks, nevertheless their weakness consists in conditioning, which may destroy the structure of the original focal elements.

From this point of view compositional models seem to be more appropriate multidimensional models in the framework of evidence theory than these two kinds of networks.

## References

[1] C. Beeri, R. Fagin, D. Maier, M. Yannakakis, On the desirability of acyclic database schemes, *J. of the Association for Computing Machinery,* **30** (1983), 479–513.

[2] B. Ben Yaghlane, Ph. Smets, K. Mellouli, Belief functions independence: I. the marginal case. *Int. J. Approx. Reasoning*, **29** (2002), 47–70.

[3] B. Ben Yaghlane, Ph. Smets, K. Mellouli, Belief functions independence: II. the conditional case. *Int. J. Approx. Reasoning*, **31** (2002), 31–75.

[4] B. Ben Yaghlane, Ph. Smets, K. Mellouli, Directed evidential networks with conditional belief functions. In: Nielsen TD, Zhang NL (eds.) *Proceedings of ECSQARU 2003,* 291–305.

[5] S. Benferhat, D. Dubois, L. Gracia, H. Prade, Directed possibilistic graphs and possibilistic logic. In: B. Bouchon-Meunier, R.R. Yager, (eds.) *Proc. of IPMU'98*, Editions E.D.K. Paris, 1470–1477.

[6] I. Couso, S. Moral, P. Walley, Examples of independence for imprecise probabilities, *Proceedings of ISIPTA'99,* eds. G. de Cooman, F. G. Cozman, S. Moral, P. Walley, 121–130.

[7] F. G. Cozman, Credal networks, *Artificial Intelligence Journal,* **120** (2000), 199-233.

[8] M. Daniel, Belief conditioning rules for classic belief functions, *Proceedings of WUPES'09,* eds. T. Kroupa, J. Vejnarová, 46–56.

[9] G. De Cooman, Possibility theory I – III. *Int. J. General Systems* **25** (1997), 291–371.

[10] R. Jiroušek, J. Vejnarová, Compositional models and conditional independence in Evidence Theory, *Int. J. Approx. Reasoning,* **52** (2011), 316-334.

[11] R. Jiroušek, J. Vejnarová, M. Daniel, Compositional models for belief functions. In: De Cooman G, Vejnarová J, Zaffalon M (eds.) *Proceedings of ISIPTA'07.* Praha, 243–252.

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

[13] P. P. Shenoy, Conditional independence in valuation-based systems. *Int. J. Approx. reasoning,* **10** (1994), 203–234.

[14] M. Studený, Formal properties of conditional independence in different calculi of artificial intelligence. *Proceedings of ECSQARUÃ93*, eds. K. Clarke, R. Kruse, S. Moral, Springer-Verlag, 1993, 341–348.

[15] J. Vejnarová, Conditional independence relations in possibility theory. *Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems* **8** (2000), 253–269.

[16] J. Vejnarová, On conditional independence in evidence theory, *Proceedings of ISIPTA'09,* eds. T. Augustin, F. P. A. Coolen, S. Moral, M. C. M. Troffaes, Durham, UK, 2009, 431–440.

[17] J. Vejnarová, An alternative approach to evidential network construction, *Combining Soft Computing and Statistical Methods in Data Analysis.* Eds: Borgelt Ch., Gonzales-Rodriguez G., Trutschnig W., Lubiano M.A., Gil M.A., Grzegorzewski P., Hryniewicz O., Oviedo 2010, 619–626.

[18] J. Vejnarová, Conditioning, conditional independence and irrelevance in evidence theory, *Proceedings of ISIPTA'11,* eds. F. Coolen, G. de Cooman, T. Fetz, M. Oberguggenberger, Innsbruck, Austria, 2011, 381–390.

[19] J. Vejnarová, Conditioning in evidence theory from the perspective of multidimensional models, *Proceedings of IPMU'12,* eds. S. Greco et al., Part III, CCIS 299, 2012, 450459.

[20] J. Vejnarová, Evidential Networks from a Different Perspective, *Proeedings of SMPS'12,* to apperar.

# Machine Learning Methods for Mortality Prediction in Patients with ST Elevation Myocardial Infarction*

**Jiří Vomlel**
Institute of Information Theory and Automation of the AS CR
Pod vodárenskou věží 4, 182 08, Prague, Czech Republic
and
Faculty of Management, University of Economics
Jarošovská 1117/II, 377 01, Jindřichův Hradec
vomlel@utia.cas.cz

**Hynek Kružík and Petr Tůma**
Gnomon, s.r.o.
Faltysova 1500/18, 156 00, Prague – Zbraslav, Czech Republic
kruzik@gnomon.cz, tuma@gnomon.cz

**Jan Přeček and Martin Hutyra**
1st Department of Internal Medicine, University Hospital Olomouc
I. P. Pavlova 6, 779 00 Olomouc, Czech Republic
jan.precek@seznam.cz, martinhutyra@seznam.cz

**Abstract**

ST Elevation Myocardial Infarction (STEMI) is the leading cause of death in developed countries. The objective of our research is to design and verify a predictive model of hospital mortality in STEMI based on clinical data about patients that could serve as a benchmark for evaluation of healthcare providers. In this paper we present results of an experimental evaluation of different machine learning methods on a real data about 603 patients from University Hospital in Olomouc.

## 1 Introduction

In developed countries ST Elevation Myocardial Infarction (STEMI) is responsible for more than a half of deaths. Its treatment has a significant socio-economic impact. The

*Our research was supported by the Czech Science Foundation through grant nr. 201/08/0539.

main objective of our research is to design and verify a predictive model of hospital mortality in STEMI based on clinical data about patients available at the beginning of their hospitalization. This model can be used not only as a decision support tool that supports medical decisions about patients' treatments but also as a benchmark for evaluation of healthcare providers, which is our main motivation for the research reported in this paper.

The motivation for this type of benchmarking is that mere mortality does not reflect severity of the illness at the hospital admission. There are hospitals that more often treat complicated cases and mere mortality would not be fair to them. Therefore the mortality should be risk adjusted. For this purpose a good model describing influence of risk factors on the mortality is needed.

In this paper we will present the results of our experimental evaluation of different machine learning methods on a real data from University Hospital in Olomouc.

## 2 Dataset of patients with STEMI

Our dataset contains data of 603 patients admitted to University Hospital in Olomouc for STEMI. The average age was 65 years. There were 431 men (71%) and 172 women (29%) in the dataset. Our goal is to classify patients into two classes according to whether they survive 30 days or not. This criteria is called *30-days mortality* [8]. The value 0 will correspond to survival while the value 1 to non-survival. Since the intended use of a constructed classifier is the evaluation of healthcare quality we use only information about patients' health state at the time of their hospital admission. In data we have 23 attributes of different types and value range. They were selected by cardiologists since they may influence STEMI mortality. The attributes are listed in Table 1. In the first group there are basic demographic characteristics and body measurements. The attributes of the second group describe the location and the mortality risk of STEMI. The third group consists of laboratory tests.

Some attribute values are missing for some patients. In total 3.2% of values are missing. As it can be seen from Table 1 the attributes are of different types by their nature. Some classification methods can handle certain types of attributes only and thus require a transformation of attributes' values.

### 2.1 Ordinal attributes

Ordinal attributes are attributes whose values have an ordering of values that is natural for the quantification of their impact on the class. This is satisfied by all attributes that can take only two values – even if they are nominal, e.g. by Gender[1]. In our data it seems it can be assumed for most real-valued attributes, but note that there might exist laboratory tests whose values deviating from a normal range in both directions (i.e. both lower and higher values) may increase the probability of death[2]. However, there is no natural ordering of the values of the nominal attribute STEMI since its

[1] For this purpose we encode Gender using two numbers: 0 for male and 1 for female.

[2] In order to allow modeling this type of influence we will transform such attributes into two attributes. We will discuss this in the next subsection.

Table 1: Attributes

| Attribute | Code | type | value range in data |
|---|---|---|---|
| Gender | SEX | nominal | {male, female} |
| Age | AGE | real | $[23, 94]$ |
| Height | HT | real | $[145, 205]$ |
| Weight | WT | real | $[35, 150]$ |
| Body Mass Index | BMI | real | $[16.65, 48.98]$ |
| STEMI Location | STEMI | nominal | {inferior, anterior, lateral} |
| Killip classification at admission | KILLIP | integer | $\{1, 2, 3, 4\}$ |
| Kalium | K | real | $[2.25, 7.07]$ |
| Urea | UR | real | $[1.6, 46.5]$ |
| Kreatinin | KREA | real | $[17, 525]$ |
| Uric acid | KM | real | $[109, 935]$ |
| Albumin | ALB | real | $[23, 53.5]$ |
| HDL Cholesterol | HDLC | real | $[0.38, 2.21]$ |
| Cholesterol | CH | real | $[1.8, 9.59]$ |
| Triacylglycerol | TAG | real | $[0.31, 8.13]$ |
| LDL Cholesterol | LDLC | real | $[0.63, 7.79]$ |
| Glucose | GLU | real | $[4.2, 25.7]$ |
| C-reactive protein | CRP | real | $[0.3, 359]$ |
| Cystatin C | CYSC | real | $[0.38, 5.22]$ |
| N-terminal prohormone of brain natriuretic peptide | NTBNP | real | $[22.2, 35000]$ |
| Troponin | TRPT | real | $[0, 25]$ |
| Glomerular filtration rate (based on MDRD) | GFMD | real | $[0.13, 7.31]$ |
| Glomerular filtration rate (based on Cystatin C) | GFCD | real | $[0.09, 7.17]$ |

values are locations. Fortunately, this problem can be simply overcame by creating one binary attribute for each state of STEMI indicating whether STEMI takes this state or not. We denote new binary attributes as STEMI_inferior, STEMI_anterior, and STEMI_lateral. We will refer to data in this form as D.ORD.

## 2.2 Discrete attributes

Some classification methods require a finite number of values of each attribute – i.e., discrete attributes. In order to get statistically reliable estimation the number of values should be as low as possible (and sensible). The transformation of a real-valued attribute into an attribute with finitely many values is called discretization. We performed discretization of all real-valued attributes. We used different number of values depending on the nature of each attribute. Generally, it is difficult to find the optimal number and the values of split points in discretization. Fortunately, there exists the Czech National Code Book that classifies numeric laboratory results, with respect to age and gender, into nine groups $1, 2, \ldots, 9$. Group 5 corresponds to standard values in the standard population. The groups $< 5$ to decreased values and groups $> 5$ to increased values. We discretized all laboratory tests X so that for each test we created two new attributes:

- One attribute for a decreased value of the test – denoted X_low – with state 0 if the value is within the normal range. Values $1, 2, 3, 4$ became values of this attribute.

- Another attribute for the increased value of the test – denoted X_high – again with state 0 if the value is within the normal range. Values $6, 7, 8, 9$ became values of this attribute.

The attributes Age, Height, and Weight were discretized into more than two groups (10, 4, and 4, respectively). We will refer to data in this form as D.DISCR.

## 2.3 Binary attributes

However, as we will see in Section 4 the performance of tested methods using discretization described in Section 2.2 was inferior to discretization to only binary attributes, where all laboratory tests are encoded using two binary attributes. The first attribute takes value 0 for the standard values of the test and value 1 if the values are decreased. The second attribute takes value 0 for the standard values of the test and value 1 if the values are increased. The attribute Killip classification was transformed by replacing value 1 by 0 and by joining the values $2, 3, 4$ into one value 1. The attributes Age, Height, and Weight were removed since they appeared not to be relevant for mortality. From the demographic group of attributes only Gender and the Body Mass Index (BMI) were kept with BMI being encoded using two binary attributes BMI_high and BMI_low. We will refer to data in this form as D.BIN.

## 2.4 Attribute selection

When learning classifiers from datasets we used every dataset in two different ways:

- all attributes were included or

- only attributes selected by the attribute selection method CfsSubsetEval from Weka [6] were included.

CfsSubsetsEval method [5] selects a subsets of attributes that are highly correlated with the class while having low intercorrelation. We searched the space of all subsets by a greedy best first search with backtracking. Data D after the application of this attribute selection method will be suffixed as D.AS.

# 3 Tested classifiers

For tests we used a large subset of classifiers implemented in Weka [6]. Classifiers that performed best in the preliminary tests qualified for the final tests. In the final tests we compared following classifiers:

- Logistic regression (two versions):

  LOG.REG – logistic regression model with a ridge estimator [10].

  LOG.BOOST – LogitBoost with simple regression functions as base learners used for fitting the logistic models [9].

- Decision tree C4.5 – pruned C4.5 decision tree [11].

- Naive Bayes classifier (two versions):

  NB.SIMPL – Naive Bayes classifier which estimates Gaussian distribution when learned from real-valued (numeric) attributes [3].

  NB – Naive Bayes classifier which also uses estimator classes. Numeric estimator precision values are chosen based on analysis of the training data [7].

- NN – Artificial Neural Network Multilayer Perceptron. The nodes in this network model sigmoid functions [2].

- Bayesian network classifier (two versions):

  BN.K2 – Bayesian Network classifier learned by K2 algorithm [1] (with unrestricted number of parents).

  BN.TAN – Tree Augmented Naive Bayes classifier [4].

# 4 Results of experiments

We compared the classifiers using Weka [6]. We used the 10-fold cross-validation methods. The results are summarized in Table 2 using the following two measures of prediction quality:

- Accuracy (ACC), which is the number of true positive and true negative classification divided by total number of classifications. It is reported using percentage scale (i.e. multiplied by 100).

- Area under the ROC curve (AOC). The ROC curve depicts the dependence of True Positive Rate (vertical axis) on False Positive Rate (horizontal axis) both as functions of the threshold.

In Table 2 we can observe several interesting things:

First, if we compare results of a single classifier on different versions of data, we can see that the best results are mostly achieved with D.BIN.AS, i.e. with discretized data, where each attribute is binary. This observation confirms the general recommendation that if the number of data records is not large then the discretization should not be

Table 2: Results of experiments

| Classifier | Criteria | D.ORD | D.ORD.AS | D.DISCR | D.DISCR.AS | D.BIN | D.BIN.AS |
|---|---|---|---|---|---|---|---|
| LOG.BOOST | ACC | 94.03 | 94.20 | 93.86 | 88.23 | 94.03 | 93.86 |
| | AUC | 0.618 | 0.646 | 0.722 | 0.640 | 0.802 | **0.832** |
| LOG.REG | ACC | 92.54 | 93.86 | 90.05 | 87.56 | 92.87 | 93.70 |
| | AUC | 0.792 | 0.821 | 0.646 | 0.607 | 0.743 | 0.798 |
| C4.5 | ACC | 93.86 | 94.69 | 94.20 | 88.72 | 93.53 | 94.53 |
| | AUC | 0.618 | 0.569 | 0.600 | 0.544 | 0.547 | 0.610 |
| NB | ACC | 89.22 | 91.04 | 86.90 | 87.73 | 86.90 | 94.20 |
| | AUC | 0.820 | 0.813 | 0.806 | 0.649 | 0.811 | 0.809 |
| NB.SIMPL | ACC | 89.72 | 90.88 | 86.90 | 87.73 | 86.90 | 94.20 |
| | AUC | 0.828 | 0.769 | 0.806 | 0.649 | 0.811 | 0.809 |
| NN | ACC | 91.38 | 93.86 | 93.20 | 87.40 | 92.04 | 93.53 |
| | AUC | 0.763 | 0.746 | 0.737 | 0.550 | 0.767 | 0.759 |
| BN.K2 | ACC | NA | NA | 92.04 | 94.53 | 94.03 | 94.36 |
| | AUC | NA | NA | 0.769 | 0.783 | 0.769 | 0.821 |
| BN.TAN | ACC | NA | NA | 92.04 | 88.89 | 94.20 | **94.86** |
| | AUC | NA | NA | 0.787 | 0.590 | 0.811 | 0.818 |

```
0.87   + STEMI_lateral * -0.41   + ALB * -0.08
       + HDLC * 0.21   + CYSC * 0.24 + KILLIP * 0.31

-1.64  + ALB_low * 0.76  + CYSC_high * 0.62  + KILLIP * 0.68
```

Figure 1: LOG.BOOST for D.ORD.AS (up) and D.BIN.AS (down).

fine-grained. We were able to improve the classifiers' performance due to a good discretization of original ordinal data based on expert knowledge of the domains of attributes.

Secondly, attribute selection methods also helped to improved performance. Originally, we did't have large number of attributes since we started with only 23 attributes. But the performance of most classifiers improved if only few of the most relevant attributes were included. This also confirms the general recommendation that in order to avoid overfitting of training data the models should be as simple as possible.

Finally, when comparing different classifiers we can see that there is not big difference between their accuracy. Actually, the high accuracy could be achieved by a primitive classifier that would assign all instance to class 0, i.e. all patients would survive 30 days. Its accuracy would be 94.03%, which is the relative number of patients that survive STEMI in our data. However, its AUC would be very low, only 0.465. Therefore we prefer classifiers that maximize both criteria at the same time. From this point of view the classifiers C4.5 and NN seem inferior to LOG, NB, and BN families. There are not huge differences between later three families, but if we should choose two best performing classifiers it would be LOG.BOOST and BN.TAN that have the best AUC and ACC from all classifiers, respectively.

Next we will present our choice of the best performing classifiers in more detail. In Figure 1 we compare LOG.BOOST for two versions of data – original ordinal and binarized data. Both formulas are for logit of probability of Mortality=1. Although there are some similarities between these two classifiers they are not exactly the same. Note that splitting laboratory tests ALB and CYSC into two attributes ALB_low, ALB_high and CYSC_low and CYSC_high helps to make explicit the impact of low values of ALB and high values of CYSC on the mortality. Also note that while in the first formula KILLIP takes values $1, 2, 3, 4$ in the second one it is only 0 (corresponding to the original 1) and 1 corresponding to the original $2, 3, 4$. Albeit the second model is simpler it has substantially higher value of AUC. Actually, according to AUC it is the best performing classifier.

The AUC values of C4.5 classifiers were quite low. However, it is interesting to see that the C4.5 for binarized data despite its extreme simplicity has quite good accuracy ACC and performs actually better than more complex C4.5 build from ordinal data. See Figure 2. In each leaf the first number after colon is the classification. The number in parenthesis is the total number of instances reaching that leaf (since our data has missing attribute values we got decimal numbers).

Finally, we add a comment on two Bayesian network classifiers. In Figure 3 we compare Tree Augmented Naive Naive Bayes classifier (up) and Bayesian Network classifier learned by K2. Despite the BN learned by K2 was allowed to have more

```
CYSC <= 1.64: 0 (553.0)
CYSC > 1.64
|   HDLC <= 0.56: 1 (5.0)
|   HDLC > 0.56
|   |   KILLIP <= 1
|   |   |   ALB <= 25.2: 1 (2.21)
|   |   |   ALB > 25.2: 0 (29.79)
|   |   KILLIP > 1
|   |   |   UR <= 15.8: 1 (6.0)
|   |   |   UR > 15.8: 0 (7.0)
```

```
CYSC_high = 0: 0 (526.0)
CYSC_high = 1
|   ALB_low = 0: 0 (63.29)
|   ALB_low = 1: 1 (13.71)
```

Figure 2: C4.5 for D.ORD.AS (up) and for D.BIN.AS (down).

parents of each attributes than TAN[3] it finally contains less edges (only four edges between attributes) and its performance is comparable with BN.TAN.

# 5 Conclusions

In this paper we compare different machine learning methods using a real medical data from a hospital. The best performance was achieved on discretized data where the discretization was based on the expert knowledge about the attributes (mostly on standard scale of results of laboratory tests) and the attributes had only two values. The best performing classifiers were based on logistic regression and on simple Bayesian networks. In our future research we would like to extend the set of attributes with other clinical data and get datasets with a larger number of patients.

## Acknowledgments

This work would not be so easy without the Weka system. We are grateful to the authors of Weka for making their tool freely available.

# References

[1] G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

[3] In TAN each but one attribute has exactly two parents, the class and one other attribute.

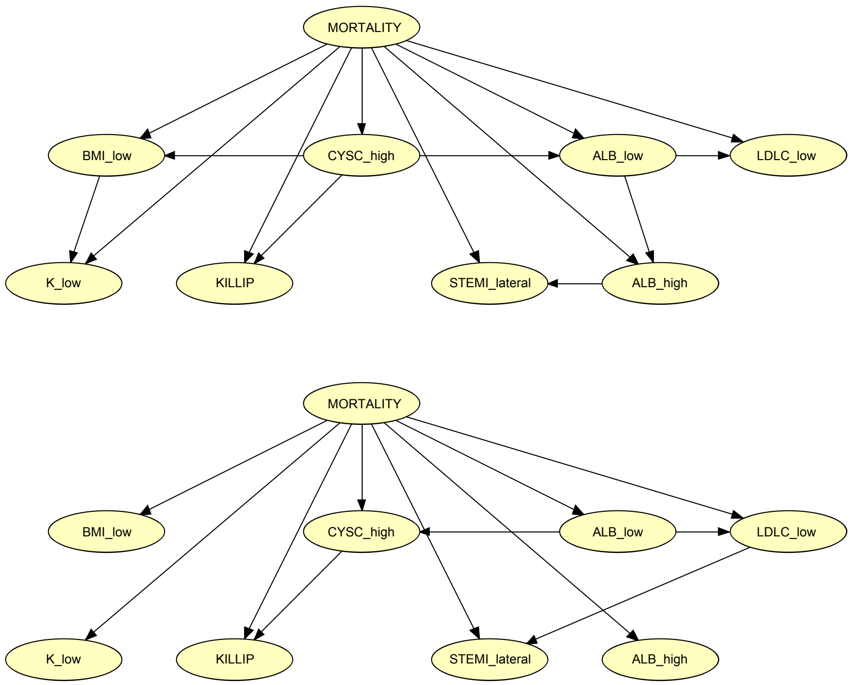


Figure 3: Tree Augmented Naive Bayes classifier (up) and Bayesian Network classifier learned by K2 (down). Both were learned from D.BIN.AS.

[2] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989.

[3] Richard Duda and Peter Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

[4] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2–3):131–163, 1997.

[5] Mark Hall. *Correlation-based Feature Subset Selection for Machine Learning.* PhD thesis, University of Waikato, Hamilton, New Zealand, 1999.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[7] George H. John and Pat Langley. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, UAI'95, pages 338–345, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[8] Harlan M. Krumholz, Sharon-Lise T. Normand, Deron H. Galusha, Jennifer A. Mattera, Amy S. Rich, Yongfei Wang, and Yun Wang. Risk-adjustment models for AMI and HF 30-day mortality, methodology. Technical report, Harvard Medical School, Department of Health Care Policy, 2007.

[9] Niels Landwehr, Mark Hall, and Eibe Frank. Logistic model trees. *Machine Learning*, 59(1–2):161–205, 2005.

[10] S. le Cessie and J.C. van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201, 1992.

[11] Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

# Beware of Too Much Information

**Christian Wallmann and Gernot D. Kleiter**
Department of Psychology
University of Salzburg
christian.wallmann2@sbg.ac.at, gernot.kleiter@sbg.ac.at

**Abstract**

Probability logic studies the properties resulting from the probabilistic interpretation of logical argument forms. Typical examples are the probabilistic Modus Ponens or Modus Tollens. Argument forms with two premises usually lead from precise probabilities of the premises to imprecise or interval probabilities of the conclusion. In the contribution we study generalized inference forms having three or more premises. Recently Gilio has shown that these generalized forms "degrade"—more premises lead to more imprecise conclusions, i.e., to wider intervals. We distinguish different forms of degradation. We analyse Predictive Inference, Modus Ponens, Bayes' Theorem, and Modus Tollens. Special attention is spend to the case where the conditioning events have zero probabilities.

## 1 Introduction

Consider a knowledge base that contains the observations $D_1$, $D_2$, $D_3$. From the knowledge base it follows that $P(H|D_1 \wedge D_2) \in [0.1, 0.12]$. In addition from the knowledge base it follows that $P(H|D_1 \wedge D_2 \wedge D_3) \in [0.6, 0.9]$. On which of the two probability intervals should we base the probability of $H$? Three properties are to be considered for this decision: (i) The width of the intervals; the interval $[0.1, 0.12]$ is tighter than $[0.6, 0.9]$, (ii) the position of the intervals; the positions of $[0.1, 0.12]$ and $[0.6, 0.9]$ are rather different, (iii) the amount of information; $D_1 \wedge D_2$ is less specific than $D_1 \wedge D_2 \wedge D_3$. The principle of total evidence requires to base the updated probability of $H$ on $P(H|D_1 \wedge D_2 \wedge D_3)$. However, this leads to the more imprecise interval. In conditional probability logic we are confronted with the above situation in many cases. Contrary to what one would expect, more specific information leads to more imprecise conclusions. Several investigations [3, 9, 4] have shown that the width of the interval of the conclusion increases as the number of premises increases. This property has been called "degradation in conditional probability logic".

Modus Ponens is the inference from the premises $\{A, A \to B\}$ to the conclusion $B$. Conditional probability logic determines the set of all coherent probability values of the conclusion if a certain coherent probability assessment on the premises is given. This set is according to de Finetti's Fundamental Theorem [2] an interval or a point value. Empirical findings strongly indicate that people interpret the probability of a conditional as the conditional probability of the corresponding conditional event [7]. The probabilistic version of Modus Ponens is consequently the inference from the premises $\{P(A) = \alpha, P(B|A) = \beta\}$ to the conclusion $P(B) \in [\alpha\beta, \alpha\beta + 1 - \alpha]$. Generalized probabilistic Modus Ponens determines the interval $P(H) \in [\delta', \delta'']$ if the premises $\{P(E_1) = \alpha_1, \ldots, P(E_n) = \alpha_n, P(H| \bigwedge\limits_{i=1}^{n} E_i) = \beta\}$ are given.

For a generalized inference form we denote by $I_n$ the interval for the conclusion if $n$ premises are given. Let $|I_i|$ be the width of the interval $I_i$. A generalized inference form *degrades* if and only if for all $i, j \in \mathbb{N}$: If $i < j$, then $|I_i| \leq |I_j|$. To study degradation in more detail, we distinguish two forms of degradation. A generalized inference form *strongly degrades* if and only if for all $i, j \in \mathbb{N}$: If $i < j$, then $I_i \subseteq I_j$. A generalized inference form *weakly degrades* if and only if it degrades and if for some $k, l \in \mathbb{N}$ $I_k \not\subseteq I_l$ and $I_l \not\subseteq I_k$ . Suppose that $l > k$. Because for the prediction of an event the position of it's probability is of main importance, a new position $I_k \not\subseteq I_l$ of the interval $I_l$ compensates for its greater width. Consequently, although both forms of degradation form a problem to the application of conditional probability logic to generalized inference forms, strong degradation is the more serious problem. Since $I_k \subseteq I_l$, no compensation in form of a new position is received for obtaining a wider interval $I_l$.

In this contribution we analyse generalizations of Modus Ponens, Predictive Inference, Conjunction, Bayes' Theorem, and Modus Tollens for the different kinds of degradation. It is common to all the inference forms considered in the present paper—with the exception of Modus Tollens—that a certain form of ultimate degradation occurs. If the number of premises is sufficiently high, then the interval of the conclusion is the unit interval $[0, 1]$. The reason for this is that the lower bound of the conjunction of $n$ events $P(\bigwedge\limits_{i=1}^{n} E_i)$ is 0 if $n$ is large. The fact that the lower bound of the conjunction is often zero has the consequence that the conditioning event of many conditional events has zero probability. We therefore give special attention to this case. In particular, we study the generalization of Bayes' Theorem where the prior of the hypothesis has zero probability or where the data has zero probability. Furthermore, we proof the result for the generalized Modus Tollens stated in [4, 9].

# 2 Degradation of Inferences in Conditional Probability Logic

## 2.1 Terminology

Let $\mathcal{F} = \{E_1|H_1, \ldots, E_n|H_n\}$ be a set of conditional events. If $H_i$ is the sure event, i.e., $H_i = \top$, then we write $E_i$ instead of $E_i|H_i$. A possible outcome or a *constituent* is a conjunction of the form $\pm E_1 \wedge \ldots \wedge \pm E_n \wedge \pm H_1 \wedge \ldots \wedge \pm H_n$, where for all events $A \in \{E_1, \ldots, E_n, H_1, \ldots, H_n\}$ $\pm A$ is either $A$ or $\neg A$ . If the $2n$ events are logically independent, then there are $2^{2n}$ constituents. We denote each constituent by a member of the set $\{C_i\}$ and by $x_i$ the probability of the $i$-th constituent $P(C_i)$. The probability of an event is the sum of the probabilities of the constituents verifying it.

Table 1 shows our notation in the case of three events $H, E_1, E_2$.

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | $C_8$ | Probability |
|---|---|---|---|---|---|---|---|---|---|
| $H$ | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | $P(H) \ = x_1 + x_2 + x_3 + x_4$ |
| $E_1$ | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | $P(E_1) = x_1 + x_2 + x_5 + x_6$ |
| $E_2$ | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | $P(E_2) = x_1 + x_3 + x_5 + x_7$ |

Table 1: Constituents for three events

The interval of the coherent probability values for the conclusion of an inference form can be determined by solving sequences of linear systems. This is a corollary of the following theorem which characterizes coherence [1, p. 81] (original for infinite sets of conditional events).

***Theorem* 1.** (Coletti and Scozzafava, 2002) A probability assessment $P$ on $\mathcal{F} = \{E_1|H_1, \ldots, E_n|H_n\}$ is coherent iff there exists a sequence of compatible systems, with unknowns $x_r^\alpha \geq 0$,

$$\mathcal{S}_\alpha = \begin{cases} \sum\limits_{C_r \subseteq E_i \wedge H_i} x_r^\alpha = P(E_i|H_i) \sum\limits_{C_r \subseteq H_i} x_r^\alpha \\ [\text{if} \sum\limits_{C_r \subseteq H_i} x_r^{\alpha-1} = 0, \ \alpha \geq 1] \ \ (i = 1, \ldots, n) \\ \sum\limits_{C_r \subseteq H_0^\alpha} x_r^\alpha = 1 \end{cases}$$

with $\alpha = 0, 1, \ldots, n$, where $H_0^0 = H_1 \vee \ldots \vee H_n$ and $H_0^\alpha$ denotes, for $\alpha \geq 1$, the union of the $H_i$ such that $\sum\limits_{C_r \subseteq H_i} x_r^{\alpha-1} = 0$.

Consider for example probabilistic Modus Tollens. The premises are $P(\neg E_1) = \alpha$, $P(E_1|H) = \beta$. Employing the notation of Table 1, the lower (resp. upper) bound for the conclusion $P(\neg H)$ can be determined by minimizing (resp. maximizing) the sum $x_5 + x_6 + x_7 + x_8$ in the following linear system

$$\begin{aligned} x_3 + x_4 + x_7 + x_8 &= \alpha \\ \beta(x_1 + x_2 + x_3 + x_4) &= x_1 + x_2 \\ \sum_{i=1}^{8} x_i = 1, \qquad x_i &\geq 0 \ . \end{aligned}$$

In the case of Modus Tollens only one linear system is to be considered. However, if we want to determine the probability of a conditional event with a conditioning event that has zero probability, then more than one linear system are to be considered. We demonstrate this in the proof of Bayes' Theorem where the data has zero probability (Theorem 5, Theorem 6).

## 2.2 Conjunction

For the conjunction of $n$ events it holds the following theorem (see for example [3]).

***Theorem* 2** (Conjunction of $n$ events)**.** If $P(E_i|H) = \alpha_i$, for $i = 1, \ldots, n$, then

$$P(\bigwedge_{i=1}^{n} E_i|H) \in \left[\max\left\{0, \sum_{i=1}^{n} \alpha_i - (n-1)\right\}, \min\{\alpha_i\}\right] .$$

The lower bound of $P(\bigwedge_{i=1}^{n+1} E_i|H)$ is less than or equal to that of $P(\bigwedge_{i=1}^{n} E_i|H)$. Equality holds if and only if $P(E_{n+1}|H) = \alpha_{n+1} = 1$. Moreover, if $n \geq \sum_{i=1}^{n} \alpha_i + 1$, then the lower bound of $P(\bigwedge_{i=1}^{n} E_i|H)$ is 0. We shall soon see that these properties of the conjunction are the reasons for the degradation of many other inferences.

## 2.3 Predictive Inference

Predictive Inference is one of the key inference rules in Bayesian statistics. It determines the predictive probability $P(H|E_1 \wedge \ldots \wedge E_r \wedge \neg E_{r+1} \wedge \ldots \wedge \neg E_n)$ of $H$ after having observed $r$ successes and $n-r$ failures in the set $\{E_i\}_{i=1}^{n}$. It is of main importance if $H$ is regarded exchangeable with the other events. If at least one of the events $\{E_i\}_{i=1}^{n}$ did not occur, i.e., $r < n$, then the interval obtained for the predictive probability is the unit interval [9]. As observed in [9], the case where all previous trials were successes, i.e., $r = n$, is a special case of the System P rule Cautious Monotonicity. Consequently, the following theorem is a corollary of the result for Cautious Monotonicity stated in [3].

***Theorem* 3** (Predictive probability)**.** If $P(H) = \beta$ and $P(E_i) = \alpha_i$, for $i = 1, \ldots, n$, then $P(H|E_1 \wedge \ldots \wedge E_n) \in [\gamma', \gamma'']$, with

$$\gamma' = \begin{cases} \max\left\{0, \frac{\beta + \sum_{i=1}^{n} \alpha_i - n}{\sum_{i=1}^{n} \alpha_i - (n-1)}\right\} & \text{if } \sum_{i=1}^{n} \alpha_i - (n-1) > 0 \\ 0 & \text{if } \sum_{i=1}^{n} \alpha_i - (n-1) \leq 0 \end{cases}$$

$$\gamma'' = \begin{cases} \min\left\{1, \frac{\beta}{\sum_{i=1}^{n} \alpha_i - (n-1)}\right\} & \text{if } \sum_{i=1}^{n} \alpha_i - (n-1) > 0 \\ 1 & \text{if } \sum_{i=1}^{n} \alpha_i - (n-1) \leq 0 \end{cases}$$

We compare this result with the result for the case where the premise $P(E_{n+1}) = \alpha_{n+1}$ is added and the conclusion is $H|E_1 \wedge \ldots \wedge E_n \wedge E_{n+1}$. So that in both cases the same event $H$ is predicted. Theorem 3 shows that the upper bound of the conclusion increases and that its lower bound decreases. Thus, the interval gets wider if a new event is added and the old interval is a subset of the new interval. Consequently, in the case of predictive inference we have a strong degradation. Furthermore, if $n \geq \sum_{i=1}^{n} \alpha_i + 1$, then $P(H|E_1 \wedge \ldots \wedge E_n) \in [0,1]$. The fact that the lower bound of the conjunction decreases is the reason for both, the strong degradation of Predictive Inference and for obtaining the unit interval if $n$ is large.

## 2.4 Modus Ponens

Modus Ponens is a special case of the System P rule Cut. The following theorem is a corollary of Gilio's result for the generalization of the Cut rule [3].

***Theorem* 4** (Modus Ponens)**.** If $P(E_i) = \alpha_i$, for $i = 1, \ldots, n$, and $P(H| \bigwedge_{i=1}^{n} E_i) = \beta$, then

$$P(H) \in [\beta\sigma_n \ , \ \beta\sigma_n + 1 - \sigma_n], \text{with}$$
$$\sigma_n = \max\left\{0, \sum_{i=1}^{n} \alpha_i - (n-1)\right\} \ .$$

We compare this result with the result where $P(E_{n+1}) = \alpha_{n+1}$ is added to the premises and $P(H| \bigwedge_{i=1}^{n} E_i) = \beta$ is replaced by $P(H| \bigwedge_{i=1}^{n+1} E_i) = \gamma$. If $\gamma = \beta$, then Modus Ponens strongly degrades. However, even if $\beta \neq \gamma$, the width of the interval for $P(H)$ increases. This follows from the facts that its width is $1 - \sigma_n$ and that $\sigma_n$ is monotonically decreasing. Consequently, that the lower bound of the conjunction decreases is the reason for the degradation of Modus Ponens. However, since $\beta$ is replaced by $\gamma \neq \beta$, the position of the interval for $P(H)$ may change. Therefore, in the case of Modus Ponens a weak degradation takes place.

**Example 1.** Consider the premise sets $T$ and $T'$ such that
$T = \{P(E_1) = 0.9, P(E_2) = 0.8, P(E_3) = 0.95, P(H|E_1 \wedge E_2 \wedge E_3) = 0.8\}$ and
$T' = \{P(E_1) = 0.9, P(E_2) = 0.8, P(E_3) = 0.95, P(E_4) = 0.8, P(H|E_1 \wedge E_2 \wedge E_3 \wedge E_4) = 0.1\}$.
From $T$ it follows that $P(H) \in [0.52, 0.87]$, whereas from $T'$ it follows that $P(H) \in [0.045, 0.595]$.

The width of the interval $1 - \sigma_n$ depends on the lower bound of the conjunction $\sigma_n$. Since this lower bound is zero if $n \geq \sum_{i=1}^{n} \alpha_i + 1$, the interval for $P(H)$ is the unit interval if the number of premises is sufficiently high.

## 2.5 Bayes' Theorem

Suppose that the prior probability of a certain hypothesis $P(H) = \delta$, the likelihood of the data given both, the hypothesis $H$, $P(D|H) = \beta$, and the alternative hypothesis

$\neg H$, $P(D|\neg H) = \gamma$, are given. The posterior probability of the hypothesis $H$ given the data $D$ is obtained by Bayes' Theorem $P(H|D) = \frac{\beta\delta}{\beta\delta+\gamma(1-\delta)}$. The premises of generalized Bayes' Theorem are $P(H) = \delta$, $P(E_1|H) = \beta_1, \ldots, P(E_n|H) = \beta_n$, $P(E_1|\neg H) = \gamma_1, \ldots, P(E_n|\neg H) = \gamma_n$. In inferential statistics it is often assumed that the $E_i$'s are independent and identically distributed. To be as general as possible, we do neither require conditional independence of the $E_i$'s given $H$ nor do we require that $P(E_i|H) = P(E_j|H)$ for $i \neq j$. The conclusion of generalized Bayes' Theorem is $P(H|E_1 \wedge \ldots \wedge E_n)$. Observe that if $P(E_1 \wedge \ldots \wedge E_n) > 0$, then

$$\begin{aligned} P(H|E_1 \wedge \ldots \wedge E_n) &= \frac{P(H \wedge E_1 \wedge \ldots \wedge E_n)}{P(E_1 \wedge \ldots \wedge E_n)} \\ &= \frac{P(H)P(E_1 \wedge \ldots \wedge E_n|H)}{P(H)P(E_1 \wedge \ldots \wedge E_n|H) + P(\neg H)P(E_1 \wedge \ldots \wedge E_n|\neg H)} \ . \end{aligned} \tag{2.1}$$

To proof the result for the generalization of Bayes' Theorem (Theorem 5 and Theorem 6) we consequently treat two cases for the probability of the data $P(E_1 \wedge \ldots \wedge E_n)$: (i) $P(E_1 \wedge \ldots \wedge E_n) > 0$ and (ii) $P(E_1 \wedge \ldots \wedge E_n) = 0$. In case (ii) it is relevant whether the prior probability $P(H)$ is zero, one, or different from both values. To handle case (ii) properly we make use of Theorem 1. The special case $n = 1$ has been investigated in detail by Coletti and Scozzafava [1, Chapter 16].

***Theorem* 5** (Bayes' Theorem, lower bound)**.** Suppose that $P(H) = \delta$ and that for all $i = 1, \ldots, n$, $P(E_i|H) = \beta_i$ and $P(E_i|\neg H) = \gamma_i$. Then:

- If $\delta(\sum_{i=1}^{n} \beta_i - (n-1)) > 0$, then

$$P(H|E_1 \wedge \ldots \wedge E_n) \geq \frac{\delta(\sum_{i=1}^{n} \beta_i - (n-1))}{\delta(\sum_{i=1}^{n} \beta_i - (n-1)) + (1-\delta)\min\{\gamma_i\}} \ .$$

- If $\delta(\sum_{i=1}^{n} \beta_i - (n-1)) \leq 0$, then $P(H|E_1 \wedge \ldots \wedge E_n) \geq 0$.

***Theorem* 6** (Bayes' Theorem, upper bound)**.** Suppose that $P(H) = \delta$ and that for all $i = 1, \ldots, n$, $P(E_i|H) = \beta_i$ and $P(E_i|\neg H) = \gamma_i$. Then:

- If $(1-\delta)(\sum_{i=1}^{n} \gamma_i - (n-1)) > 0$, then

$$P(H|E_1 \wedge \ldots \wedge E_n) \leq \frac{\delta \min\{\beta_i\}}{\delta \min\{\beta_i\} + (1-\delta)(\sum_{i=1}^{n} \gamma_i - (n-1))} \ .$$

- If $(1-\delta)(\sum_{i=1}^{n} \gamma_i - (n-1)) \leq 0$, then $P(H|E_1 \wedge \ldots \wedge E_n) \leq 1$.

*Proof.* We prove the result for the upper bound. The proof for the lower bound is obtained by analog considerations. We distinguish two cases.

(I) If $(1-\delta)(\sum_{i=1}^{n}\gamma_i-(n-1))>0$, then $P(E_1\wedge\ldots\wedge E_n)>0$. The result is obtained by application of the Conjunction Theorem (Theorem 2) to (2.1).

(II) If $(1-\delta)(\sum_{i=1}^{n}\gamma_i-(n-1))\leq 0$, we distinguish two cases (i) $\delta\min\{\beta_i\}>0$ and (ii) $\delta\min\{\beta_i\}=0$.

In case (i) the upper bound 1 is obtained by setting $P(H\wedge E_1\wedge\ldots\wedge E_n)$ to $\delta\min\{\beta_i\}>0$ and $P(\neg H\wedge E_1\wedge\ldots\wedge E_n)$ to its minimum 0.

In case (ii) we obtain the upper bound by setting the probability of the data $P(E_1\wedge\ldots\wedge E_n)$ to 0. We treat the case $n=2$. The proof generalizes to the case $n>2$ straightforwardly. We build the sequence of linear systems $\mathcal{S}_\alpha$ (Theorem 1). To improve readability we write $x_i$ instead of $x_i^0$, $y_i$ instead of $x_i^1$, and $z_i$ instead of $x_i^2$.

Using the notation of Table 1, the first linear system $\mathcal{S}_0$ is given by

$$\begin{aligned}
x_1+x_5 &=0\\
P(H|E_1\wedge E_2)(x_1+x_5) &=x_1\\
x_1+x_2+x_3+x_4 &=\delta\\
x_1+x_2=\beta_1(x_1+x_2+x_3+x_4),\qquad x_1+x_3 &=\beta_2(x_1+x_2+x_3+x_4)\\
x_5+x_6=\gamma_1(x_5+x_6+x_7+x_8),\qquad x_5+x_7 &=\gamma_2(x_5+x_6+x_7+x_8)\\
x_1+x_2+x_3+x_4+x_5+x_6+x_7+x_8 &=1,\quad x_i\geq 0\ .
\end{aligned}$$

As unique solution of $\mathcal{S}_0$ we obtain $x_1=x_5=0$, $x_2=\beta_1\delta$, $x_3=\beta_2\delta$, $x_4=\delta-(\beta_1+\beta_2)\delta$, $x_6=\gamma_1(1-\delta)$, $x_7=\gamma_2(1-\delta)$, $x_8=(1-\delta)-(\gamma_1+\gamma_2)(1-\delta)$. Since $\min\{\beta_1,\beta_2\}=0$, it is $x_4\geq 0$ and since by assumption $\gamma_1+\gamma_2\leq 1$, it is $x_8\geq 0$, so that the solution is coherent.

If $0<P(H)=\delta<1$, then $H_0^1=E_1\wedge E_2$. The system $\mathcal{S}_1$ is consequently given by

$$\begin{aligned}
P(H|E_1\wedge E_2)(y_1+y_5) &=y_1\\
y_1+y_5=1,\qquad y_i &\geq 0\ .
\end{aligned}$$

So that $P(H|E_1\wedge E_2)=\frac{y_1}{y_1+y_5}$ can attain any value in $[0,1]$.

If $P(H)=\delta=0$ (the case $P(H)=1$ is treated in the same way), then $x_1=x_2=x_3=x_4=x_5=0$ and consequently $H_0^1=H\vee(E_1\wedge E_2)$. In the system $\mathcal{S}_1'$ all constraints that concern conditional events with conditioning event $H$ remain.

$$\begin{aligned}
P(H|E_1\wedge E_2)(y_1+y_5) &=y_1\\
y_1+y_2=\beta_1(y_1+y_2+y_3+y_4),\qquad y_1+y_3 &=\beta_2(y_1+y_2+y_3+y_4)\\
y_1+y_2+y_3+y_4+y_5 &=1,\qquad y_i\geq 0\ .
\end{aligned}$$

We solve $\mathcal{S}_1'$ in such a way that $P(H)>0$ and $P(E_1\wedge E_2)=0$. The unique solution in this case is $y_2=\beta_1$, $y_3=\beta_2$, $y_4=1-(\beta_1+\beta_2)$. Then $H_0^2=E_1\wedge E_2$ and the third system $\mathcal{S}_2'$ is

$$\begin{aligned} P(H|E_1 \wedge E_2)(z_1 + z_5) =& z_1 \\ z_1 + z_5 = 1, \qquad z_i \geq& 0 \ . \end{aligned}$$

So that $P(H|E_1 \wedge E_2) = \frac{z_1}{z_1+z_5}$ can attain any value in $[0, 1]$.
In both cases, $0 < P(H) < 1$ and $P(H) = 0$, we have constructed a sequence of compatible systems $(S_\alpha)$, with unknowns $(x_i^\alpha)$, $i = 1, \dots, 8$, $\alpha = 0, 1, 2$, such that $P(H|E_1 \wedge E_2) \in [0, 1]$. According to Theorem 1 $P(H|E_1 \wedge E_2)$ can coherently attain any value in $[0, 1]$. □

Bayes' Theorem does not degrade. First of all, Bayes' Theorem does not degrade strongly. The lower bound is not monotonically decreasing, because it depends on the minimum of the set $\{\gamma_i\}$. If for a given $n$ the premises $P(E_{n+1}|\neg H) = \gamma_{n+1} < \min\{\gamma_i\}$ and $P(E_{n+1}|H) = \beta_{n+1}$ are added, the lower bound may increase. Similar considerations show that the upper bound is not monotonically increasing. As consequence, intervals with rather different positions may result.

**Example 2.** Suppose that $P(H) = 0.1$, $P(E_1|H) = 0.9$, $P(E_2|H) = 0.8$, $P(E_3|H) = 0.4$, $P(E_1|\neg H) = 0.9999$, $P(E_2|\neg H) = 0.9999$, $P(E_3|\neg H) = 0.001$. Then $P(H|E_1 \wedge E_2) \in [0.072, 0.081]$, but $P(H|E_1 \wedge E_2 \wedge E_3) = [0.917, 0.982]$.

In Bayes' Theorem even no weak degradation takes place. In general, the probability interval of the conclusion does not get wider as the number of premises increases.

**Example 3.** Suppose that $P(H) = 0.9$, $P(E_1|H) = 0.99$, $P(E_2|H) = 0.99$, $P(E_3|H) = 0.98$, $P(E_1|\neg H) = 0.9999$, $P(E_2|\neg H) = 0.9999$, $P(E_3|\neg H) = 0.001$. Then $P(H|E_1 \wedge E_2) \in [0.898176, 0.89911]$ , but $P(H|E_1 \wedge E_2 \wedge E_3) \in [0.999884, 0.999909]$, so that the width of the first interval is 0.000934 and that of the second interval is 0.000025.

This does by no means imply that additional information makes the situation necessarily better. In many cases the interval does get wider if the number of premises increases. If, for instance, identical probabilities $\beta_i = \beta$ and $\gamma_i = \gamma$ for $i = 1, \dots, n$, are assumed, then Bayes' Theorem strongly degrades. This case is of main importance because it is implied by the assumption of conditional exchangeability. Moreover, Theorem 5 and Theorem 6 show that even in the case of Bayes' Theorem one ends up with the unit interval. If $n \geq \max\left\{\sum\limits_{i=1}^{n} \beta_i + 1, \sum\limits_{i=1}^{n} \gamma_i + 1\right\}$, then $\sum\limits_{i=1}^{n} \beta_i - (n-1) \leq 0$ and $\sum\limits_{i=1}^{n} \gamma_i - (n-1) \leq 0$, so that the interval $[0, 1]$ is obtained.

## 2.6 Modus Tollens

The following holds for the probabilistic Modus Tollens of two events. If $P(\neg E) = \alpha$ and $P(E|H) = \beta$, then $P(\neg H) \in [\gamma', 1]$, where

$$\gamma' = \max\left\{1 - \frac{\alpha}{1-\beta}, 1 - \frac{1-\alpha}{\beta}\right\} \ . \tag{2.2}$$

Wagner [8] has shown the result for the lower bound. However, Wagner's upper bound is different from 1. The reason for this is that Wagner defined the conditional probability $P(E|H)$ by the fraction $\frac{P(E\wedge H)}{P(H)}$. If $P(\neg H) = 1$, then $P(H) = 0$ and $P(E|H)$ would consequently be undefined. As already pointed out, in the coherence approach conditionalizing on events with zero probability is possible, so that the correct upper bound $P(\neg H) = 1$ is obtained.

The result for the generalized Modus Tollens has been presented without proof in [4, 9].

***Theorem 7*** (Modus Tollens)**.** If $P(\neg E_i) = \alpha_i$, for $i = 1, 2, \ldots, n$, and if $P(E_1 \wedge E_2 \wedge \ldots \wedge E_n|H) = \beta$, then $P(\neg H) \in [\delta', 1]$, with

$$\delta' = \begin{cases} 1 - \frac{1-\alpha^*}{\beta} & \text{if} \quad \alpha^* + \beta > 1 \\ 1 - \frac{\sum\limits_{i=1}^{n} \alpha_i}{1-\beta} & \text{if} \quad \alpha^* + \beta \leq 1 \text{ and } \sum\limits_{i=1}^{n} \alpha_i + \beta < 1 \\ 0 & \text{if} \quad \alpha^* + \beta \leq 1 \text{ and } \sum\limits_{i=1}^{n} \alpha_i + \beta \geq 1 \ , \end{cases}$$

where $\alpha^* = \max\{\alpha_i\}$.

*Proof.* First, we treat the special case $n = 2$ and then we outline the proof for the general case.
*Two events:* If $n = 2$, then by employing the notation of Table 1 we obtain the linear system

$$\beta(x_1 + x_2 + x_3 + x_4) = x_1 \tag{2.3}$$

$$x_3 + x_4 + x_7 + x_8 = \alpha_1 \tag{2.4}$$

$$x_2 + x_4 + x_6 + x_8 = \alpha_2 \tag{2.5}$$

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 = 1, \qquad x_i \geq 0 \ .$$

To maximize (resp. minimize) $P(\neg H)$ we minimize (resp. maximize)

$$P(H) = x_1 + x_2 + x_3 + x_4 \ .$$

Manipulation of (2.3) shows that

$$P(H) = \frac{x_1}{\beta} \ . \tag{2.6}$$

Therefore, to maximize (minimize) $P(H)$ we maximize (minimize) $x_1 = P(H \wedge E_1 \wedge E_2)$.
(A) The minimum 0 for $x_1$ is obtained by setting $x_8 = \min\{\alpha_1, \alpha_2\}$. If $\alpha_1 \leq \alpha_2$, then we set $x_6 = \alpha_2 - \alpha_1$ and $x_5 = 1 - \alpha_2$. If $\alpha_1 > \alpha_2$, then we set $x_7 = \alpha_1 - \alpha_2$ and $x_5 = 1 - \alpha_1$. Hence, by (2.6) the maximum of $P(\neg H)$ is 1.
(B) For the maximum of $x_1$ observe that according to the Conjunction rule $x_1 \leq 1-\alpha^*$. Furthermore, since $P(H \wedge E_1 \wedge E_2) \leq P(E_1 \wedge E_2|H)$, we have $x_1 \leq \beta$. Therefore, we distinguish two cases: (I) $1 - \alpha^* < \beta$ and (II) $1 - \alpha^* \geq \beta$.

In case (I) we set $x_1$ to its maximum $1 - \alpha^*$. By (2.6) we obtain $P(H) = \frac{1-\alpha^*}{\beta}$, so that the minimum of $P(\neg H)$ is $1 - \frac{1-\alpha^*}{\beta}$. A solution of the linear system that establishes that $x_1 = 1 - \alpha^*$ is coherent, is uniquely determined by setting $x_5 = 0$ and $x_8 = \min\{1 - \frac{1-\alpha^*}{\beta}, \alpha_1, \alpha_2\}$.
In case (II) we distinguish two cases:

1. If $\alpha_1 + \alpha_2 + \beta \geq 1$, then the maximum of $x_1$ is $\beta$ and consequently according to (2.6) $P(H) = 1$. Thus, $P(\neg H) = 0$. The following solution of the linear system establishes that setting $x_1 = \beta$ is coherent: $x_2 = -\alpha_1 + (1-\beta)$, $x_3 = -\alpha_2 + (1-\beta)$, $x_4 = \alpha_1 + \alpha_2 - (1 - \beta)$.

2. If $\alpha_1 + \alpha_2 + \beta < 1$, then $x_1 < \beta$. Suppose on the contrary $x_1 = \beta$, then $x_1 + x_2 + x_3 + x_4 = 1$. But according to (2.4) and (2.5) it is $x_3 + x_4 = \alpha_1$ and $x_2 + x_4 = \alpha_2$. Consequently, $x_1 + x_2 + x_3 + x_4 \leq \alpha_1 + \alpha_2 + \beta < 1$, which is a contradiction. In this case $x_1 = \frac{(x_2+x_3+x_4)\beta}{1-\beta}$ is maximized if $x_2 + x_3 + x_4$ is maximized. This is the case if $x_2 = \alpha_1$, $x_3 = \alpha_2$, and $x_4 = 0$, so that $x_1 = \frac{(\alpha_1+\alpha_2)\beta}{1-\beta}$. Consequently

$$x_5 = P(\neg H) = 1 - (x_1 + x_2 + x_3 + x_4) = 1 - \frac{\frac{(\alpha_1+\alpha_2)\beta}{1-\beta}}{\beta} = 1 - \frac{\alpha_1 + \alpha_2}{1 - \beta} \quad .$$

*n events:* The proof generalizes to $n$ events as follows.
(A) The maximum is obtained from the solution $P(\neg H \wedge \neg E_1 \wedge \ldots \wedge \neg E_n) = \min\{\alpha_i\}$. Let $\alpha^1 = \min\{\alpha_i\}$ and $\alpha^k = \min\{\alpha_i | i = 1, \ldots, n\} \setminus \{\alpha^i | i = 1, \ldots, k-1\}$. Let $E^k = E_i$ iff $\alpha_i = \alpha^k$. Suppose that the events $\{E_i\}$ have been renumbered appropriately and set $P(E^1 \wedge \ldots \wedge E^{k-1} \wedge \neg E_k \wedge \ldots \wedge \neg E_n \wedge \neg H) = \alpha^k - \alpha^{k-1}$. The sum of these constituents is $\alpha^n$. The remaining probability $1 - \alpha^n$ is given to $P(\neg H \wedge E_1 \wedge \ldots \wedge E_n)$.
(B) For the minimum we make the same case distinction as in the case $n = 2$. In case (I) $\alpha^* + \beta > 1$, set $x_1 = P(H \wedge E_1 \wedge \ldots \wedge E_n) = 1 - \alpha^*$, then $P(\neg H) = 1 - \frac{1-\alpha^*}{\beta}$. In case (II) we distinguish between (i) $\sum_{i=1}^{n} \alpha_i + \beta \geq 1$ and (ii) $\sum_{i=1}^{n} \alpha_i + \beta < 1$. In case (i) the lower bound $P(\neg H) = 0$ is obtained by setting $x_1 = P(H \wedge E_1 \wedge \ldots \wedge E_n) = \beta$. In case (ii) we set $P(H \wedge E_1 \wedge \ldots \wedge \neg E_i \wedge \ldots \wedge E_n)$ to $\alpha_i$. Then $x_1 = P(H \wedge E_1 \wedge \ldots \wedge E_n) = \frac{\beta \sum_{i=1}^{n} \alpha_i}{1-\beta}$. The remaining probability $1 - \frac{\beta \sum_{i=1}^{n} \alpha_i}{1-\beta}$ is given to the constituent $\neg H \wedge E_1 \wedge \ldots \wedge E_n$. □

Modus Tollens has very interesting properties with respect to degradation. Suppose that $P(\neg E_{n+1})$ is added to the premises and $P(E_1 \wedge E_2 \wedge \ldots \wedge E_n | H) = \beta$ is replaced by $P(E_1 \wedge E_2 \wedge \ldots \wedge E_{n+1} | H) = \gamma$. While the upper bound 1 for $P(\neg H)$ is already most “degraded”. The lower bound does not decrease as the number of premises $n$ increases. Depending on the values of $\gamma$ and $\alpha^*$ we jump back and forth between the cases (I) $\alpha^* + \gamma > 1$ and (II) $\alpha^* + \gamma \leq 1$ In case (II) since $\sum_{i=1}^{n} \alpha_i$ increases as $n$ increases, the lower bound 0 is obtained rapidly. In case (I) the lower bound strongly depends on the values of $\gamma$ and $\alpha^*$. As a consequence it can attain any value $c \in (0, 1]$.

If $\alpha^* < 1$, then $P(\neg H) \in [c, 1]$ if $\beta = \frac{1-\alpha^*}{1-c}$. Consequently, Modus Tollens does not degrade. Moreover, contrary to the other inferences considered in this paper, the unit interval is not necessarily obtained if the number of premises is large.

## 3 Discussion

We have seen that Predictive Inference strongly degrades, Modus Ponens weakly degrades, and that Bayes' Theorem and Modus Tollens do not degrade. Moreover, in all the inference forms considered—with the exception of Modus Tollens—the unit interval is obtained if the number of premises is sufficiently large. These facts cast a dark shadow on the utility of probability logic in the case of generalized inference forms. Surely, a narrower interval is better than a wider interval and a more complete knowledge base is better than a truncated one [5]. While in general the number of premises and the precision of the conclusion *may* conflict, in probability logic they often *must* conflict.

This conflict between amount of information and precision cannot be ignored. On the one hand, to select the most "recent" interval obtained by the most specific information leads to wide intervals. In many case it even leads to the unit interval. On the other hand, to select the narrowest interval requires to base the interval of the conclusion on the most unspecific information. Since all additional premises are discarded, it would be useless to apply probability logic to generalized inference forms.

Although additional premises often yield more imprecise intervals, they do not necessarily make inference in conditional probability logic worse. Contrary to strong degradation, in the case of weak degradation obtaining intervals with different positions to a certain degree compensates for obtaining wider intervals. Clearly, a solution to the conflict between specificity and precision has to be different for both forms of degradation. If an inference form strongly degrades, it is more reasonable to follow a take-the-best strategy, i.e., take the most precise interval. This is to be done by discarding all information but the most unspecific. It is then not reasonable to search for further information, because it simply cannot be used. If an inference form weakly degrades, a take-the-best strategy is less reasonable. Obtaining conflicting intervals changes the opinion about the position of the interval (as, for instance, in Example 2). Since the new position is based on more information, it is more "recent" than the old position. The knowledge of the position of the interval is of main importance, so that it is not reasonable to discard the new information. In this case to solve the conflict between precision and specificity requires to counterbalance (i) the width of an interval, (ii) the amount of information it is based upon, and (iii) the position of the interval. However, whether a take-the-best strategy is rational or how to counterbalance (i), (ii), and (iii) are questions that cannot be answered by the formal results of probability logic. They are a matter of subjective preferences.

**Acknowledgement** This work is supported by the Austrian Science Foundation (FWF, I 141-G15) within the LogICCC Programme of the European Science Foundation.

## References

[1] Coletti, G. and Scozzafava, R.: Probabilistic Logic in a Coherent Setting. Kluwer, Dordrecht (2002)

[2] De Finetti, B.: Theory of Probability. A critical introductory treatment. Volume 1. Wiley, New York (1974)

[3] Gilio, A.: Generalization of inference rules in coherence-based probabilistic default reasoning. International Journal of Approximate Reasoning 53 (2012), 413–434

[4] Kleiter, G. D.: Ockham's razor in probability logic. To appear in the proceedings of SMPS 2012.

[5] Kyburg, H. E. and Teng, C. M.: Uncertain Inference. Cambridge University Press, Cambridge (2001)

[6] Lad, F.: Operational Subjective Statistical Methods. Wiley, New York (1996)

[7] Pfeifer, N. and Kleiter, G. D.: Inference in conditional probability logic. Kybernetika, 42 (2006) 391–404

[8] Wagner, C. G.: Modus tollens probabilized. British Journal for the Philosophy of Science 55 (4):747–753 (2004)

[9] Wallmann, C. and Kleiter, G. D.: Exchangeability in Probability Logic. In S. Greco et al. (Eds.): IPMU 2012, Part IV, CCIS 300 (2012) 157–167

# The Characteristic Imset Polytope for Diagnosis Models

**Jing Xi and Ruriko Yoshida**
Department of Statistics
University of Kentucky
jing.xi@uky.edu, ruriko.yoshida@gmail.com

**Abstract**

In 2010, M. Studený, R. Hemmecke, and Linder explored a new algebraic description of graphical models, characteristic imsets. Compare with standard imsets, characteristic imsets have several advantages: they are still unique vector representative of conditional independence structures, they are 0-1 vectors, and they are more intuitive in terms of graphs than standard imsets. After defining characteristic imset polytope as the convex hull of all characteristic imsets for a given set of nodes, they also showed that a model selection in graphical models, which essentially is a problem of maximizing a quality criterion, can be converted into an integer programming problem on the characteristic imset polytope. However, this integer programming problem is very hard in general. Therefore, here we focus on diagnosis models which can be described by Bipartite graphs with a set of $m$ nodes and a set of $n$ nodes for any $m, n \in \mathbb{Z}_+$, and their characteristic imset polytope. In this paper, first, we will show that the characteristic imsets for diagnosis models have very nice properties including that the number of non-zero coordinates is at most is $n \cdot (2^m - 1)$, and with these properties we are able to find a combinatorial description of all edges of the characteristic imset polytopes for diagnosis models. Then we prove that these characteristic imset polytopes are direct products of n many $(2^m - 1)$ dimensional simplicies. Finally, we end the paper with further questions in this topic.

## 1 Introduction

Bayesian networks (BNs), also known as belief networks, Bayes networks, Bayes(ian) models or probabilistic directed acyclic graphical models, find their applications to model knowledge in many areas, such as computational biology and bioinformatics (gene regulatory networks, protein structure, gene expression analysis [3] learning epistasis from GWAS data sets [4]) and medicine [13]. BNs are a part of the family of probabilistic graphical models (GMs). These graphical structures represent knowledge about probabilistic structures for a statistical model. More precisely, each node in the graph represents a random variable and an edge between the nodes represents probabilistic dependencies among the random variables corresponding to the nodes adjacent to the edge [6]. BNs correspond to GM structure known as a directed acyclic graph (DAG) defined by the set of nodes (vertices) and the set of directed edges.

In order to infer parameters from the observed data set, we first apply a model selection criterion called *quality criterion*, which provides a way to construct highly predictive BN models from data by choosing the graph which gives the given criteria, such as Bayesian Information Criteria (BIC) [8] or Akaike Information Criteria (AIC) [1], maximum (see [10] for more details on quality criterions). Intuitively a quality criterion is a function, $\mathcal{Q}(G, D)$, which takes a DAG, $G$, and an observed data set, $D$, to evaluate how good the DAG $G$ to explain the observed data $D$. Note that different DAGS, $G_1, G_2$ may have the same conditional independences (CIs). In that case we say $G_1, G_2$ are *Markov equivalent*. When researchers wish to infer the CIs of the BN structure from the observed data set one represents each set of Markov equivalent graphs by one graph called the *essential graph* the corresponding Markov equivalence class of DAGs [2]. In this paper we focus on quality criterions $\mathcal{Q}(G, D)$, such that $\mathcal{Q}(G_1, D) = \mathcal{Q}(G_2, D)$ if and only if $G_1, G_2$ are Markov equivalent.

Since in general there are super exponentially many essential graphs with a fixed set of nodes $N$, maximizing the quality criterion, $\mathcal{Q}(G, D)$, over all possible essential graphs with $N$ is known to be NP-hard. Studený developed an algebraic representation of each essential graph $G$ called a *standard imset*, of $G$, which is an integral vector representation of $G$ in $\mathbb{R}^{2^{|N|}-|N|-1}$. From the view of this setting a criterion function $\mathcal{Q}(G, D)$ is a dot product of vectors in $\mathbb{R}^{2^{|N|}-|N|-1}$. In 2010, M. Studený, J. Vomlel, and R. Hemmecke showed that maximizing the $\mathcal{Q}(G, D)$ over all essential graphs can be formulated as a linear programming problem over the convex hull of standard imsets for all possible essential graphs [12]. This gives us a systematic way to find the best criterion with the optimality certificate rather than finding the best criterion by the brute-force search. Then M. Studený, R. Hemmecke, and Linder explored an alternative vector representative of the BN structure, called *characteristic imsets*. Compare with standard imsets, characteristic imsets have several advantages: they are still unique vector representative of conditional independence structures; they are 0-1 vectors; and they are more intuitive in terms of graphs than standard imsets [11].

In general, however, the dimension of the convex hull of the characteristic imsets with the fixed set of nodes $N$, called *characteristic imset polytope*, is exponentially large and there are double exponentially many vertices as well as facets of the characteristic imset polytope. Thus it is infeasible to optimize by software if $|N| > 6$. In order to solve the LP problem for a larger $|N|$, we need to understand the structure of the characteristic imset polytope, such as combinatorial description of edges and facets of the polytope so that we might be able to apply a simplex method to find an optimal solution. However, in general, it is challenging because there are too many facets and too many edges of the polytope. Therefore here we focus on a particular family of BN models, namely *diagnosis models*.

In medical studies, researchers are often interested in probabilistic models in order for them to correctly diagnose a disease from a patient symptoms. The diagnoses models, also known as the Quick Medical Reference (QMR) diagnostic model, is introduced in [9] to diagnose a disease from a given set of symptoms of a patient. Therefore, here we focus on diagnosis models (e.g., [7]). Under this model, a DAG representing the model is a bipartite graph with two sets of nodes, one representing $m$ diseases and one representing $n$ symptoms, and set of directed edges from nodes representing diseases to nodes representing symptoms (see Figure 1 for an example).

In this paper, first, we will show that the dimension of the characteristic imset polytope for diagnosis models with $m$ diseases and $n$ symptoms is $n \cdot (2^m - 1)$ which is much smaller than $2^{(m+n)} - (n + m) - 1$. Second, we are able to find an explicit combinatorial description of all edges of the characteristic imset polytopes for diagnosis models with fixed $m$ and $n$, that is, if $G_1$, $G_2$ are graphs representing two diagnosis models such that all symptoms have the same parents in $G_1$ and in $G_2$ except one symptom, then the characteristic imsets representing $G_1$, $G_2$ form an edge of the characteristic imset polytope for diagnosis models. Then we prove that these characteristic imset polytopes are direct products of $n$ many $(2^m - 1)$ dimensional simplicies.

This paper is organized as follows. In Section 2 we introduce notation and state some definitions. Section 3 shows propositions and their proofs, and Section 4 shows our main results.

## 2 Notation and definitions

In this section we state some notation and remind readers some definitions.

**Definition 2.1.** A Diagnosis Model can be described by a Bipartite Graph whose nodes $N = \{a_1, \ldots, a_m\} \cup \{b_1, \ldots, b_n\}$ can be divided into disjoint sets $A = \{a_1, \ldots, a_m\}$ and $B = \{b_1, \ldots, b_n\}$. Nodes in $A$ can be interpreted as diseases and nodes in $B$ can be interpreted as symptoms. Every single edge can only be drawn from a disease to a symptom. An example is given by Figure 1.
Define notation: $\mathcal{G}_{m,n} = \{$All possible Bipartite graphs defined in Definition 2.1 for fixed m and n$\}$.

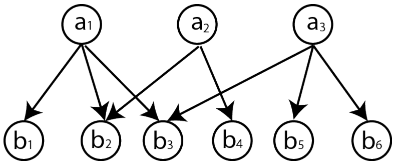


Figure 1: An example of Bipartite Graph, $m = 3$, $n = 6$.

## 3 Propositions

All proofs in this section can be found in [14].
Recall that we have the definition of Characteristic Imset.

**Definition 3.1.** Let $G$ be an acyclic directed graph over $N$. The **characteristic imset** for $G$ can be introduced as a zero-one vector $c_G$ with components $c_G(S)$ where $S \subseteq N$, $|S| \geq 2$ given by

$$c_G(S) = 1 \Longleftrightarrow \exists\ i \in S \text{ such that } j \in pa_G(i) \text{ for } \forall\ j \in S\backslash\{i\}$$

where $j \in pa_G(i)$ means $G$ includes the edge from $j$ to $i$.

**Proposition 3.2.** *Assume $|N| > 2$ and $G$ is a Bipartite graph. $A = \{a_1, \ldots, a_m\}$ and $B = \{b_1, \ldots, b_n\}$ are defined in Definition 2.1. Then $c_G(T)$ is possible to take value 1 if and only if $T$ has the form of $a_{i_1} \ldots a_{i_k} b_j$, where $1 \leq k \leq m$, $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$.*

**Proposition 3.3.** *Notation same as Proposition 3.2. Suppose $T$ has the form of $a_{i_1} \ldots a_{i_k} b_j$, where $1 \leq k \leq m$, $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$, then $c_G(T) = \prod_{s=i_1,\ldots,i_k} c_G(a_s b_j)$.*

**Remark 3.4.** Another way to see Proposition 3.3 is that for a diagnosis model, the whole characteristic imset is determined by only $m \cdot n$ coordinates, $c_G(a_i b_j)$, $i \in \{1, \ldots, m\}$, $j \in \{1, \ldots, n\}$, which can also be interpreted as the existence of edge point from $a_i$ to $b_j$. From this remark, it is straightforward to get Proposition 3.6.

**Remark 3.5.** Proposition 3.3 also implies that $\forall\ G \in \mathcal{G}_{m,n}$, $G$ can be determined by $pa_G(b_j)$, $b_j \in B$, and $pa_G(b_j)$, $b_j \in B$ are completely irrelevant.

**Proposition 3.6.** *Notation same as Definition 2.1. Fix $m$ and $n$. The number of elements in $\mathcal{G}_{m,n}$ is $2^{mn}$.*

**Proposition 3.7.** *Notation same as Definition 2.1. Fix $m$ and $n$. If we consider the characteristic imset for an arbitrary Bipartite graph in $\mathcal{G}_{m,n}$, the number of non-zero coordinates is at most $n \cdot (2^m - 1)$.*

**Remark 3.8.**

- Notation same as Definition 2.1. For a fixed $N$, by Proposition 3.2 and 3.7, we can define $\mathcal{S}_{m,n}$ as the support of $\{c_G : G \in \mathcal{G}_{m,n}\}$. We know that:

  $$\mathcal{S}_{m,n} = \{T : \exists\ G \in \mathcal{G}_{m,n} \text{ such that } c_G(T) = 1\} \subset \mathcal{P}(N)$$

  where $\mathcal{P}(N)$ is the power set of $N$.

- Recall that in elementary geometry,
  - a **closed convex polyhedron** (which will be indicated as **polyhedron** for short) in $\mathbb{R}^q$ can be defined by a system of linear inequalities:

    $$\{\mathbf{x} \in \mathbb{R}^q : A\mathbf{x} \leq \mathbf{b}\}$$

    where $A$ is a $p \times q$ matrix in $\mathbb{R}^{p\times q}$ and $\mathbf{b}$ is a $p \times 1$ vector in $\mathbb{R}^{p\times 1}$;
  - a **closed convex polytope** (which will be indicated as **polytope** for short) is defined as the convex hull of a finite set of points;
  - if a polyhedron is bounded, then it is a polytope.

- A **d-simplex** is a d-dimensional polytope which has exactly $d + 1$ vertices. It is notated as $\Delta_d$.

- For more details on polyhedral geometry see [15].
- Notice that Proposition 3.7 also implies: for fixed $m$ and $n$, the dimension of the polytope of characteristic imsets for all elements in $\mathcal{G}_{m,n}$ (we call it **characteristic imset polytope** in the following and use $\mathbf{P}_{m,n}$ as the notation) is at most $n \cdot (2^m - 1)$. We will prove that it is actually exactly $n \cdot (2^m - 1)$.
- Because characteristic imsets are all 0-1 vectors, it is obvious that for fixed $m$ and $n$, the set of vertices of characteristic imset polytope is exactly $\{c_G : G \in \mathcal{G}_{m,n}\}$.

# 4 Theorems and Proofs

**Theorem 4.1.** *Notation same as Definition 2.1. Fix $m$ and $n$. The dimension of the characteristic imset polytope is exactly $n \cdot (2^m - 1)$.*

*Proof.* Just need to show that all characteristic imsets are linearly independent. For details please see [14]. □

**Remark 4.2.** Notice that for the special case $n = 1$, Theorem 4.1 and Proposition 3.6 claim that $\mathbf{P}_{m,1}$ has $2^m$ vertices and the dimension of $\mathbf{P}_{m,n}$ is $n \cdot (2^m - 1)$. This directly lead to Theorem 4.3.

**Theorem 4.3.** *Fix $m$, $\mathbf{P}_{m,1}$ is a simplex with dimension $2^m - 1$, i.e. $\mathbf{P}_{m,1} = \Delta_{2^m-1}$.*

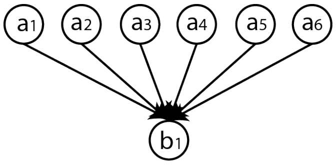
**Definition 4.4.** Graphs $G$, $H \in \mathcal{G}_{m,n}$ are called **neighbors** if $c_G$ and $c_H$ form an edge in the characteristic imset polytope.

**Lemma 4.5.** *Notation same as Definition 2.1. Fix $m$, then for arbitrary two distinct graphs, $G_1$, $G_2 \in \mathcal{G}_{m,1}$, $G_1$ and $G_2$ are neighbors, i.e. $c_{G_1}$ and $c_{G_2}$ form an edge in the characteristic imset polytope.*

*Proof.* The proof is omitted here. For details please see [14]. □

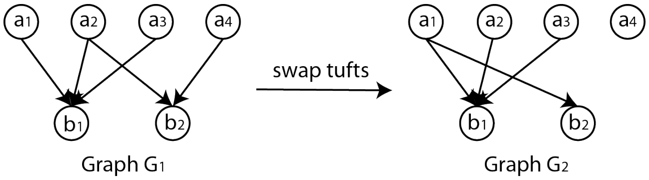
**Remark 4.6.** We can understand Lemma 4.5 better after we introduce a new concept "tuft".

- A structure is called **k-tuft** if it includes $(k + 1)$ nodes such that there exists one node (the **offspring**) that all other nodes are its parents. There are two important special cases: an 1-tuft is just an edge and a 2-tuft is just an immorality. An example of 6-tuft is given in the right hand side graph.



A 6-Tuft Structure

- Fix $m$ and $n$. We say a change from one graph $G_1$ to another one $G_2$ is **swapping tufts** if: $G_2$ can be obtained from $G_1$ by removing a tuft and adding another tuft, where these two tufts share the same offspring but disjoint parents. An example is given as below: $G_2$ can be obtained from $G_1$ by removing a 2-tuft and add an 1-tuft, where the two tufts share the same offspring $b_2$.



- Lemma 4.5 gives us a principal idea of finding neighbors for a graph: add or remove a tuft, or swap tufts once. We are going to prove this in general case in Theorem 4.7.

**Theorem 4.7.** *Notation same as Definition 2.1. Fix $m$ and $n$. Two graphs, $G_1$, $G_2 \in \mathcal{G}_{m,n}$ are neighbors if and only if $\exists\ b_i \in B$ such that $pa_{G_1}(b_i) \neq pa_{G_2}(b_i)$ and $pa_{G_1}(b_j) = pa_{G_2}(b_j)$, for $\forall\ b_j \in B$ and $b_j \neq b_i$, i.e., $G_2$ can be obtained from $G_1$ by removing or adding a tuft or swapping tufts.*

*Proof.* We will prove "if" and "only if" separately.

(1) Prove "if" part.
Suppose we have two graphs, $G_1$, $G_2 \in \mathcal{G}_{m,n}$, and $\exists\ b_i \in B$ such that $pa_{G_1}(b_i) \neq pa_{G_2}(b_i)$ and $pa_{G_1}(b_j) = pa_{G_2}(b_j)$, for $\forall\ b_j \in B$ and $b_j \neq b_i$. We need to prove $G_1$ and $G_2$ are neighbors.
Consider an arbitrary graph $G_3 \in \mathcal{G}_{m,n}$. We want to prove that $\exists$ a cost vector $w$ such that $w \cdot c_{G_1} = w \cdot c_{G_2} \geq w \cdot c_{G_3}$ where "=" holds if and only if $G_3 = G_1$ or $G_2$.
Define the following graphs (see Remark 4.8 for example):

- ⋆ $G_1'$, $G_2'$, $G_3' \in \mathcal{G}_{m,1}$ with symptom $B_{m,1} = \{b_i\}$ such that $pa_{G_1'}(b_i) = pa_{G_1}(b_i)$, $pa_{G_2'}(b_i) = pa_{G_2}(b_i)$ and $pa_{G_3'}(b_i) = pa_{G_3}(b_i)$;
- ⋆ $G_0$, $G_3'' \in \mathcal{G}_{m,(n-1)}$ with symptoms $B_{m,(n-1)} = B\backslash\{b_i\}$ such that $pa_{G_0}(b_j) = pa_{G_1}(b_j) = pa_{G_2}(b_j)$ and $pa_{G_3''}(b_j) = pa_{G_3}(b_j)$, $\forall\ b_j \in B_{m,(n-1)}$.

Notice that the connections of the characteristic imsets of these graphs are very simple. By Remark 3.5, after moving the coordinates properly, we can write the characteristic imsets of $G_1$, $G_2$ and $G_3$ in the form of:

$$\begin{aligned} c_{G_1} &= \ (c_{G_1'} \quad c_{G_0}) \\ c_{G_2} &= \ (c_{G_2'} \quad c_{G_0}) \\ c_{G_3} &= \ (c_{G_3'} \quad c_{G_3''}) \end{aligned}$$

- As proved in Lemma 4.5, $G_1'$ and $G_2'$ are neighbors. This means that $\exists$ related cost vector $w_1$ such that $w_1 \cdot c_{G_1'} = w_1 \cdot c_{G_2'} \geq w_1 \cdot c_{G_3'}$ for $\forall\ G_3' \in \mathcal{G}_{m,1}$, where "=" holds if and only if $G_3' = G_1'$ or $G_2'$.

– Because $c_{G_0}$ is a vertex of the characteristic imset polytope related to $\mathcal{G}_{m,(n-1)}$, we can find a related cost vector $w_2$ such that $w_2 \cdot c_{G_0} \geq w_2 \cdot c_{G''_3}$ for $\forall\ G''_3 \in \mathcal{G}_{m,(n-1)}$, where "=" holds if and only if $G''_3 = G_0$.

Now let $w = (w_1\ w_2)$ with the new permutation of coordinates. We have:

$$\begin{array}{rclcl} w \cdot c_{G_1} & = & w_1 \cdot c_{G'_1} + w_2 \cdot c_{G_0} & & \\ & = & w_1 \cdot c_{G'_2} + w_2 \cdot c_{G_0} & = & w \cdot c_{G_2} \\ & \geq & w_1 \cdot c_{G'_3} + w_2 \cdot c_{G''_3} & = & w \cdot c_{G_3} \end{array}$$

where "=" holds if and only if i) $G'_3 = G'_1$ or $G'_2$, and ii) $G''_3 = G_0$, i.e. $G_3 = G_1$ or $G_2$.

(2) Prove "only if" part.
Suppose we have two graphs, $G_1$, $G_2 \in \mathcal{G}_{m,n}$, which are neighbors. i.e. $\exists$ a cost vector $w$ such that $w \cdot c_{G_1} = w \cdot c_{G_2} > w \cdot c_G$ for $\forall\ G \in \mathcal{G}_{m,n}$, $G \neq G_1$, $G_2$. We are going to prove by contradiction.
Suppose $\exists\ b_i$, $b_j \in B$ distinct, $pa_{G_1}(b_i) \neq pa_{G_2}(b_i)$ and $pa_{G_1}(b_j) \neq pa_{G_2}(b_j)$.
Define the following graphs (see Remark 4.8 for example):

- ⋆ $G'_1$, $G'_2 \in \mathcal{G}_{m,1}$ with symptom $B_{m,1} = \{b_i\}$ such that $pa_{G'_1}(b_i) = pa_{G_1}(b_i)$ and $pa_{G'_2}(b_i) = pa_{G_2}(b_i)$;
- ⋆ $G''_1$, $G''_2 \in \mathcal{G}_{m,1}$ with symptom $B_{m,1} = \{b_j\}$ such that $pa_{G''_1}(b_j) = pa_{G_1}(b_j)$ and $pa_{G''_2}(b_j) = pa_{G_2}(b_j)$;
- ⋆ $G'''_1$, $G'''_2 \in \mathcal{G}_{m,(n-2)}$ with symptoms $B_{m,(n-2)} = B \backslash \{b_i, b_j\}$ such that $pa_{G'''_1}(b_k) = pa_{G_1}(b_k)$ and $pa_{G'''_2}(b_k) = pa_{G_2}(b_k)$, $\forall\ b_k \in B_{m,(n-2)}$;
- ⋆ $G_3 \in \mathcal{G}_{m,n}$ is all same with $G_1$ but $pa_{G_3}(b_i) = pa_{G_2}(b_i)$;
- ⋆ $G_4 \in \mathcal{G}_{m,n}$ is all same with $G_1$ but $pa_{G_4}(b_j) = pa_{G_2}(b_j)$;
- ⋆ $G_5 \in \mathcal{G}_{m,n}$ is all same with $G_2$ but $pa_{G_5}(b_i) = pa_{G_1}(b_i)$ and $pa_{G_5}(b_j) = pa_{G_1}(b_j)$, notice that $G_5$ might be same with $G_1$.

Similar with (1), after moving the coordinates properly, we can write the characteristic imsets of $G_1$, $G_2$, $G_3$, $G_4$ and $G_5$ in the following form:

$$\begin{array}{rl} c_{G_1} = & (c_{G'_1} \quad c_{G''_1} \quad c_{G'''_1}) \\ c_{G_2} = & (c_{G'_2} \quad c_{G''_2} \quad c_{G'''_2}) \\ c_{G_3} = & (c_{G'_2} \quad c_{G''_1} \quad c_{G'''_1}) \\ c_{G_4} = & (c_{G'_1} \quad c_{G''_2} \quad c_{G'''_1}) \\ c_{G_5} = & (c_{G'_1} \quad c_{G''_1} \quad c_{G'''_2}) \end{array}$$

Now use the same permutation of coordinates in $w$ and do the same partition, we can write $w$ in the form of $w = (w_1\ w_2\ w_3)$. By the assumption we indicated at the beginning of this part, we have:

∗ It is obvious that $G_3 \neq G_1$ or $G_2$. So:

$$\begin{array}{rrcl} & w \cdot c_{G_1} & = & w_1 \cdot c_{G'_1} + w_2 \cdot c_{G''_1} + w_3 \cdot c_{G'''_1} \\ > & w \cdot c_{G_3} & = & w_1 \cdot c_{G'_2} + w_2 \cdot c_{G''_1} + w_3 \cdot c_{G'''_1} \\ \Longrightarrow & w_1 \cdot c_{G'_1} & > & w_1 \cdot c_{G'_2} \end{array}$$

- $*$ It is obvious that $G_4 \neq G_1$ or $G_2$. So:

$$\begin{array}{rrcl} & w \cdot c_{G_1} & = & w_1 \cdot c_{G_1'} + w_2 \cdot c_{G_1''} + w_3 \cdot c_{G_1'''} \\ > & w \cdot c_{G_4} & = & w_1 \cdot c_{G_1'} + w_2 \cdot c_{G_2''} + w_3 \cdot c_{G_1'''} \\ \Longrightarrow & w_2 \cdot c_{G_1''} & > & w_2 \cdot c_{G_2''} \end{array}$$

Then we have:

$$\begin{array}{rrcll} & w \cdot c_{G_2} & = & w_1 \cdot c_{G_2'} + w_2 \cdot c_{G_2''} + w_3 \cdot c_{G_2'''} & \\ & & < & w_1 \cdot c_{G_1'} + w_2 \cdot c_{G_1''} + w_3 \cdot c_{G_2'''} & = w \cdot c_{G_5} \\ \Longrightarrow & w \cdot c_{G_2} & < & w \cdot c_{G_5} & \end{array}$$

which is a contradiction with our assumption. Therefore $G_1$ and $G_2$ cannot be neighbors.

□

**Remark 4.8.** In the two parts of proof of Theorem 4.7, it might be more intuitive to see how we define the new graphs using examples.

- Part (1), prove "if" condition. Let $m = 4$ and $n = 3$. In the example in Figure 2, $b_i = b_1$ .

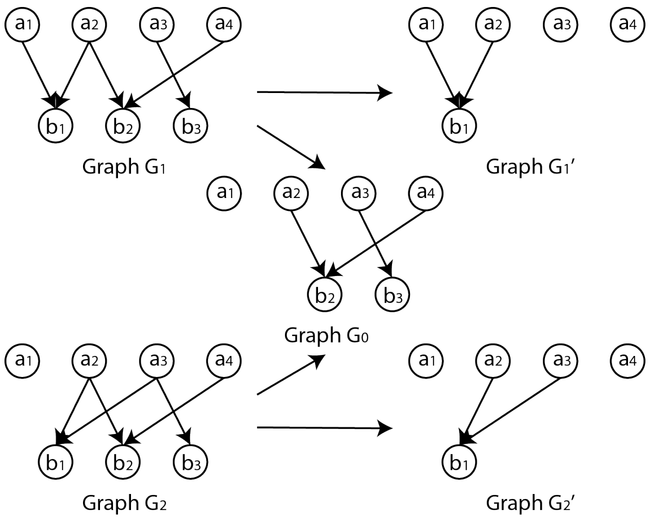


Figure 2: An example of proof for "if" condition. Here $m = 4$, $n = 3$ and $b_i = b_1$

- Part (2), prove "only if" condition. Let $m = 4$ and $n = 3$. In the example in Figure 3, $b_i = b_1$ and $b_j = b_2$.

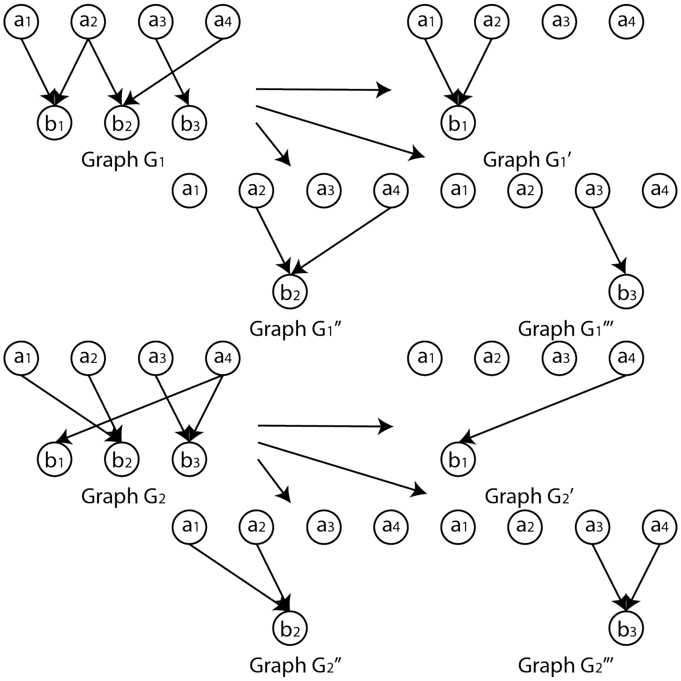


Figure 3: An example of proof for "only if" condition. Here $m = 4$, $n = 3$, $b_i = b_1$ and $b_j = b_2$.

**Theorem 4.9.** *Notation same as Definition 2.1. Fix $m$ and $n$. For $\forall\ G \in \mathcal{G}_{m,n}$, $G$ has $n \cdot (2^m - 1)$ many neighbors.*

*Proof.* Proof is omitted here. Please see [14] for details. □

**Remark 4.10.** Recall that we proved in Theorem 4.1, for fixed $m$ and $n$, the dimension of the characteristic imset polytope is $n \cdot (2^m - 1)$. Thus Theorem 4.9 implies that the number of neighbors for each vertex equals to the dimension, i.e. the characteristic imset polytope for Bipartite graphs is a **simple polytope**. In 2000, V. Kaibel and M. Wolff proved that a zero-one polytope is simple if and only if it equals to a direct product of zero-one simplices [5]. But here, we are going to prove a even stronger result.

**Theorem 4.11.** *Notation same with Definition 2.1. $\mathbf{P}_{m,n}$ is the direct product of $n$ many $\Delta_{2^m-1}$, i.e.*

$$\mathbf{P}_{m,n} = \underbrace{\Delta_{2^m-1} \times \Delta_{2^m-1} \times \cdots \times \Delta_{2^m-1}}_{n \text{ many}}.$$

*And the $i_{th}$ simplex is $\mathbf{P}_{m,1}$ with the same diseases $A$ and one symptom $\{b_i\}$.*

*Proof.* Fix $m$, we are going to prove the equality holds using induction on $n$.

- $n = 1$. Proved in Theorem 4.3;
- Fix $q \in \mathbb{Z}^+$. Suppose the equality holds for $\mathbf{P}_{m,n}$, $\forall\ n < q$, and we need to prove that it also holds for $\mathbf{P}_{m,q}$. Recall that for $\mathcal{G}_{m,q}$, we have notation for all symptoms: $B = \{b_1, b_2, \ldots, b_q\}$.
  First, we want to prove: $\mathbf{P}_{m,n} \subseteq \mathbf{P}_{m,q-1} \times \mathbf{P}_{m,1}$.
  Similar with the proof of Theorem 4.7, for $\forall\ G \in \mathcal{G}_{m,n}$, we can define graphs:
  - ⋆ $G' \in \mathcal{G}_{m,(q-1)}$ with symptoms $B_{m,(q-1)} = B \backslash \{b_q\}$ such that $pa_{G'}(b_i) = pa_G(b_i)$, $\forall\ b_i \in B_{m,(q-1)}$. This implies $c_{G'} \in \mathbf{P}_{m,q-1}$;
  - ⋆ $G'' \in \mathcal{G}_{m,1}$ with symptom $B_{m,1} = \{b_q\}$ such that $pa_{G''}(b_q) = pa_G(b_q)$. This implies $c_{G''} \in \mathbf{P}_{m,1}$.

  Again, with a proper permutation of coordinates, we can write $c_G$ in form of:

  $$c_G = (c_{G'}\ \ c_{G''}).$$

  Now because the set of vertices of $\mathbf{P}_{m,q}$ is $\{c_G : G \in \mathcal{G}_{m,q}\}$, so for $\forall\ x \in \mathbf{P}_{m,q}$, with the same permutation of coordinates, we have:

  $$x = \sum_{G \in \mathcal{G}_{m,q}} \alpha_G c_G = (\sum_{G \in \mathcal{G}_{m,q}} \alpha_G c_{G'}\ ,\ \sum_{G \in \mathcal{G}_{m,q}} \alpha_G c_{G''}),$$

  where $0 \leq \alpha_G \leq 1$, $\forall\ G \in \mathcal{G}_{m,q}$ and $\sum_{G \in \mathcal{G}_{m,q}} \alpha_G = 1$.
  Notice that $\sum_{G \in \mathcal{G}_{m,q}} \alpha_G c_{G'} \in \mathbf{P}_{m,q-1}$ and $\sum_{G \in \mathcal{G}_{m,q}} \alpha_G c_{G''} \in \mathbf{P}_{m,1}$, the above equality implies $x \in \mathbf{P}_{m,q-1} \times \mathbf{P}_{m,1}$. Hence:

$$\mathbf{P}_{m,q} \subseteq \mathbf{P}_{m,q-1} \times \mathbf{P}_{m,1}.$$

Second, we want to prove: $\mathbf{P}_{m,q-1} \times \mathbf{P}_{m,1} \subseteq \mathbf{P}_{m,n}$.
Let $\mathcal{G}_{m,q-1}$ has symptoms $B_{m,(q-1)} = B\backslash\{b_q\}$ and $\mathcal{G}_{m,1}$ has symptom $B_{m,1} = \{b_q\}$. For $\forall\ G' \in \mathcal{G}_{m,(q-1)}$ and $G'' \in \mathcal{G}_{m,1}$, we can define $G \in \mathcal{G}_{m,q}$ such that $pa_G(b_i) = pa_{G'}(b_i)$, $\forall\ b_i \in B_{m,(q-1)}$, and $pa_G(b_q) = pa_{G''}(b_q)$. We can write $c_G$ in form of $c_G = (c_{G'} \ \ c_{G''})$.
Now for $\forall\ x \in \mathbf{P}_{m,q-1} \times \mathbf{P}_{m,1}$, by definition of direct product, it can be written as:

$$\begin{aligned} x &= \Big(\sum_{G' \in \mathcal{G}_{m,q-1}} \beta_{G'} c_{G'}\ ,\ \sum_{G'' \in \mathcal{G}_{m,1}} \gamma_{G''} c_{G''}\Big) = \sum_{G' \in \mathcal{G}_{m,q-1}} \sum_{G'' \in \mathcal{G}_{m,1}} \beta_{G'}\gamma_{G''}(c_{G'}\ ,\ c_{G''}) \\ &= \sum_{G' \in \mathcal{G}_{m,q-1}} \sum_{G'' \in \mathcal{G}_{m,1}} (\beta_{G'}\gamma_{G''})\ c_G\ \ , \end{aligned}$$

where $0 \leq \beta_{G'}, \gamma_{G''} \leq 1$, $\forall\ G' \in \mathcal{G}_{m,q-1}$, $G'' \in \mathcal{G}_{m,1}$, and $\sum_{G' \in \mathcal{G}_{m,q-1}} \beta_{G'} = 1$, $\sum_{G'' \in \mathcal{G}_{m,1}} \gamma_{G''} = 1$.
Notice that

$$\sum_{G' \in \mathcal{G}_{m,q-1}} \sum_{G'' \in \mathcal{G}_{m,1}} (\beta_{G'}\gamma_{G''}) = \sum_{G' \in \mathcal{G}_{m,q-1}} \beta_{G'}\Big(\sum_{G'' \in \mathcal{G}_{m,1}} \gamma_{G''}\Big) = \sum_{G' \in \mathcal{G}_{m,q-1}} \beta_{G'} = 1\ .$$

This leads to $x \in \mathbf{P}_{m,q}$. Hence:

$$\mathbf{P}_{m,q-1} \times \mathbf{P}_{m,1} \subseteq \mathbf{P}_{m,n}.$$

Now using the assumption we made before, we can finish the proof because:

$$\begin{gathered} \mathbf{P}_{m,q} = \mathbf{P}_{m,q-1} \times \mathbf{P}_{m,1} = \underbrace{\Delta_{2^m-1} \times \cdots \times \Delta_{2^m-1}}_{\text{q-1 many}} \times \Delta_{2^m-1} = \\ \underbrace{\Delta_{2^m-1} \times \cdots \times \Delta_{2^m-1}}_{\text{q many}}\ . \end{gathered}$$

□

# References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. *In B. Petrox & F. Caski (Eds.), Proceedings of the second international symposium on information theory*, pages 267–281, 1973.

[2] S. A. Andersson, D. Madigan, and M. D. Perlman. A characterization of markov equivalence classes for acyclic digraphs. *Annals of Statistics*, 25:505–541, 1997.

[3] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620, 2000.

[4] X. Jiang, R.E. Neapolitan, M.M. Barmada, and S. Visweswaran. Learning genetic epistasis using bayesian network scoring criteria. *BMC Bioinformatics*, 12(89), 2011.

[5] V. Kaibel and M. Wolff. Simple 0/1-polytope. *Europ. J. Combinatorics*, 21:139–144, 2000.

[6] S. L. Lauritzen. *Graphical Models.* Clarendon Press, 1996.

[7] P.J.F. Lucas. Bayesian model-based diagnosis. *International Journal of Approximate Reasoning*, 27(2):99–119, 2001.

[8] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[9] M. A. Shwe, D. E. Heckerman, M. Henrion, H. P. Lehmann, and G. F. Cooper. Probabilistic diagnosis using a reformulation of the internist-1/qmr knowledge base: I. the probabilistic model and inference algorithms. *Methods of Information in Medicine*, 30:241–255, 1991.

[10] M. Studený. *Probabilistic Conditional Independence Structures.* Springer Verlag, 2005.

[11] M. Studený, R. Hemmecke, and S. Lindner. Characteristic imset: a simple algebraic representative of a bayesian network structure. *In Proceedings of the 5th European Workshop on Probabilistic Graphical Models*, pages 257–264, 2010.

[12] M. Studený, J. Vomlel, and R. Hemmecke. A geometric view on learning bayesian network structures. *International Journal of Approximate Reasoning*, 51(5):573–586, 2010.

[13] J. Uebersax. Pgenetic counseling and cancer risk modeling: An application of bayes nets. marbella. *Spain: Ravenpack International*, 2004.

[14] J. Xi and R. Yoshida. The characteristic imset polytope for diagnosis models, 2012.

[15] G. Ziegler. *Lectures on Polytopes.* Springer Verlag, New York, New York, 1994.

# ASSURANCE SERVICES FOR SUSTAINABILITY REPORTING UNDER DEMPSTER-SHAFER THEORY OF BELIEF FUNCTIONS

**Rajendra P. Srivastava**[a], **Sunita S. Rao**[a] **and Theodore J. Mock**[b]
[a]School of Business, The University of Kansas, Lawrence, Kansas 66045, USA
rsrivastava@ku.edu, ssrao@ku.edu
[b]Anderson Graduate School of Management, University of California-Riverside, USA
University Maastricht, tmock@ucr.edu

### Abstract

This study develops a framework based on the Dempster-Shafer theory of belief functions for the purpose of planning, performing and evaluating evidence obtained to assess the quality of, that is to provide assurance on, sustainability reports. Such an approach facilitates auditors in planning and performing efficient and effective assurance services.

Sustainability reporting, which is also known as corporate sustainability reporting (SR), and triple bottom line (TBL) reporting, provides data on financial and non-financial factors related to environmental, social and economic performance. To demonstrate the application of this framework we evaluate assertions, sub-assertions and audit evidence relevant to sustainability reporting based on the G3 Reporting framework developed by the Global Reporting Initiative (GRI).

## 1 Introduction

Research on sustainability reporting and quality assurance is important because an increasing number of entities are striving to measure, report and reduce their environmental and social impact [2] [4]. We focus on sustainability reporting and its assurance because a growing number of companies are issuing sustainability reports and a substantial number of companies are also getting these reports assured (*audited*) [1] [4] [16].

Sustainability reporting, which is also known as corporate sustainability reporting (SR), corporate social responsibility reporting and triple bottom line (TBL) reporting, provides economic, environmental and social information on financial and non-financial factors related to any entities' performance. Most companies providing SR Reports follow the reporting guidelines developed by the Global Reporting Initiative (GRI), a decade-old non-profit organization based in Amsterdam. GRI issued its G3 Reporting Framework in October 2006 [3]. The G3 Reporting Framework is a set of sustainability reporting guidelines and consists of reporting principles, reporting guidance, and standard disclosures including performance indicators (see also [4]). Assurance guidelines, developed to provide information quality assurance of the SRs, are being developed, but most assurance firms use IAASB guidelines [5] [6].

The primary objective of this paper is to construct and demonstrate the use of an evidential reasoning framework under the Dempster-Shafer theory of belief functions [7] (hereafter DS theory) for planning, performing and evaluating evidence for assurance of SRs. Under the evidential reasoning framework using DS theory, there are four important steps the assurance provider needs to consider:

1) Prepare an appropriate evidential diagram and identify the corresponding items of evidence;
2) Perform the audit procedures, i.e., collect and evaluate various items of evidence;
3) Combine all items of evidence in the evidential diagram to determine the overall level of assurance achieved; and
4) Assess and report on the level assurance that may be provided.

The evidential network development is elaborated in Section 2.

The rest of the paper is divided into four additional sections. The following section describes the evidential reasoning approach and the use of evidential diagrams. In Section 3 we discuss the theoretical framework used in the paper. In Section 4, a hypothetical case is presented that demonstrates how an assurance provider can use an evidential reasoning approach to provide SR assurance. The final section summarizes the conclusions and discusses directions for future research along with the limitations of the study.

## 2 Evidential Diagram

In the auditing contexts discussed in this paper, evidential reasoning entails determining the important ‘assertions’ within the SR that needs to be assured; determining what sort of evidence is relevant to these assertions; and deciding what level of support for the assertions is obtained. The ‘assertions’ are the statements made by the reporting entity such as its impact on its environment or its economic performance. To model such settings, an evidential diagram may be developed comprising of the variables involved in providing assurance along with their mutual relationships and items of evidence pertaining to those variables. Once the evidential diagram is completed, the auditor can judge the influence of evidence on all of the assertions being audited. In general, the assurance provider must decide whether the statements (assertions) in the SRs are supported by the evidence or not.

In such models, knowledge about one or more variables can be used to make assessments concerning other variables, if we know how these variables are interrelated [11] [15]. Normally, knowledge about the states of these variables is incomplete. That is, there is uncertainty associated with what an assurance provider knows about these variables. These uncertainties translate into the audit risks that must be controlled [13].

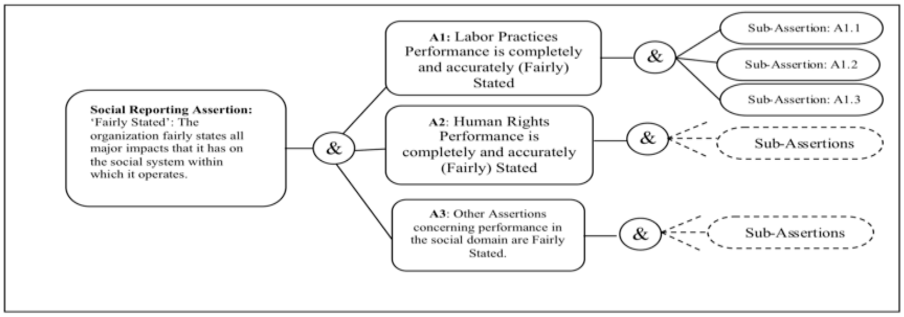
### 2.1 Underlying Framework

G3 [3] guidelines are used to structure the evidential reasoning for conducting a SR audit. Table 1 lists the three major assertion categories that must be considered: 'Social Assertions', 'Environmental Assertions' and 'Economic Assertions'. Table 1, column 1, labeled ‘Main Assertions’, describes the assertion categories. The related sub-assertions are listed in the second column. According to the G3 guidelines, social assertion category reporting requires that the entity disclose all major impacts that it has on the social system within which it operates. This includes labor practices, human rights, social interaction, and product responsibility. These impacts are expressed as assertions in column 2 of Table 1 and in Figure 1. The assurance provider needs to plan and collect adequate and pertinent evidence to audit each of these assertions. The sub-assertions are assumed to be related to the corresponding main assertions through an ‘and’ relationship. This relationship conveys that the main assertions are valid if and only if the corresponding sub-assertions are valid.

## Table 1: Assertions and Sub-assertions for Sustainability Reporting Services (Taken from GRI guidelines, GRI 2006)

| Main Assertions | Sub-Assertions |
|---|---|
| **A1. Social reporting assertion:** The organization fairly presents all major impacts that it has on the social system that it operates in. | **A.1.1: Labor Practices -** Complete and Accurate disclosure of Labor Practices**.** The organization fairly presents its labor practices and whether it meets internationally recognized standards. |
| | **A1.2 Human Rights:** The organization fairly presents the extent to which human rights plays a part in its operations and activities. |
| | **A1.3 Social Interaction:** The organization fairly presents the major risks that arise from interaction with other social institutions. |
| | **A1.4 Product Responsibility:** The organization fairly presents how its products and services directly affect customers. |
| **A2. Environmental Reporting assertion:** The organization fairly presents its performance and all major impacts that it has on the environment that it operates within. | **A2.1 Materials:** The organization fairly presents the extent to which it uses different materials by weight and by volume and the percentage of materials used that are recycled input materials.<br>**A2.2 Energy:** The organization fairly presents the extent to which it consumes energy by energy source.<br>**A2.3 Water:** The organization fairly presents the extent to which it withdraws water by source.<br>**A2.4 Biodiversity:** The organization fairly presents the location, size of land owned, leased, managed in or adjacent to protected areas and areas of high diversity value, description of significant impacts of activities, products and services on biodiversity in protected areas and areas of high biodiversity value.<br>**A2.5 Emissions, Effluents and Waste:** The organization fairly presents total direct and indirect greenhouse gas emissions by weight, emissions of ozone-depleting substances by weight, $NO_x$ and $SO_x$ and other significant air-emissions by type and weight, total water discharge by quality and destination, total weight of waste by type and disposal method, total number and volume of significant spills.<br>**A2.6 Products and Services:** The organization fairly presents initiatives to mitigate environmental impacts of products and services and the extent of impact mitigation, percentage of products sold and their packing materials that are reclaimed by category.<br>**A2.7 Compliance:** The organization fairly presents monetary value of significant fines and total number of non-monetary sanctions for non-compliance with environmental laws and regulations. |
| **A3. Economic Reporting assertion:** The organization fairly presents its economic performance | **A.3.1 Economic Performance:** The organization fairly presents direct economic value generated and distributed, which includes revenues, operating costs, employee compensation, donations, community investments, retained earnings and payments to capital providers and governments, coverage of the company's defined benefit plan obligations and significant assistance received from government.<br>**A3.2 Financial Performance:** The organization fairly presents financial implications, risks and opportunities, of the organizations activities due to climate change.<br>**A.3.3 Market Presence:** The organization fairly presents policy, practices and proportion of spending on locally based suppliers at significant locations of operation, procedures for local hiring and proportion of senior management hired from local community at significant locations of operation.<br>**A.3.4 Indirect Economic Impacts:** The organization fairly presents development and impact of infrastructure investments and services provided primarily for public benefit through commercial, in-kind or pro-bono engagement. |

**Figure 1: Evidential Diagram for an entity reporting on its performance in the Social Category. Assertions A.1.1 - A1.1.3 are described in Table 2**



Social Reporting Assertion: 'Fairly Stated': The organization fairly states all major impacts that it has on the social system within which it operates.
&
A1: Labor Practices Performance is completely and accurately (Fairly) Stated
A2: Human Rights Performance is completely and accurately (Fairly) Stated
A3: Other Assertions concerning performance in the social domain are Fairly Stated.
&
Sub-Assertion: A1.1
Sub-Assertion: A1.2
Sub-Assertion: A1.3
&
Sub-Assertions
&
Sub-Assertions

## 2.2 Construction of an Evidential Diagram

Figures 1 and 2 illustrate the structure of evidential diagrams. First, all the assertions and items of evidence pertaining to these assertions must be identified. Consider Figure 2 where the assertions are depicted as rectangular boxes with rounded corners. The main assertion on the left (A1.1) states a 'completeness accuracy assertion'[1] that the reporting entity has presented a complete and accurate disclosure of labor practices. This assertion is connected through an 'and' relationship, represented by a circle with an '&', to six sub-assertions labeled A1.1.1 through A1.1.6. All sub-assertions and the corresponding main assertion are based on the G3 guidelines. The variables representing assertions and sub-assertions have three associated values. For example for A1.1 and following the syntax of DS theory, an assessment is shown of the believe that the assertion is valid (true) of .846; that the assertion is not valid (false) of .032 and an assessment of the level of unresolved uncertainty given available evidence of .122.

----- Figure 2 about here -----

Of course, relevant items of evidence pertaining to various assertions must be indicated within the evidential network. These items of evidence result from audit procedures performed by the assurance provider. Rectangular boxes are used to represent items of evidence and these are connected to the assertion or assertions that they help inform.

In Figure 2, the six sub-assertions to the right of the main assertion are related to it through an 'and' relationship. This relationship suggests that the main assertion is valid or true if and only if the six sub-assertions are valid. In Figure 2, the evidential diagram drawn is a network diagram, that is, a network where at least one item of evidence pertains to more than one assertion. In order to determine whether the main assertion is true, the assurance provider would plan and perform the procedures described in the rectangular boxes (evidence nodes). Each evidence node represents an audit procedure which provides positive, negative, or mixed evidence concerning the assertion to which it is linked. Based on what is ascertained from each of the procedures, the assurance provider must estimate the level of support or negation from each item of evidence for each corresponding assertion.

[1] Completeness and accuracy are used for illustrative purposes.

The audit procedures illustrated throughout the paper are intended to be comprehensive, but not exhaustive. That is, there could be other items of evidence that could be created using G3 guidelines. Our intention is to show how an assurance provider can use the evidential reasoning framework for planning and performing an efficient and effective SR audit. First, in Section 3, we discuss the theoretical foundation on how audit evidence propagates through a SR evidential network such as that represented in Figure 2. Then in Section 4 we illustrate the process of aggregating evidence through a numerical example pertaining to Figure 2. This will demonstrate the level of assurance obtained on the assertions of interest.

## 3 Propagation of Beliefs in Evidential Diagram

Shenoy and Shafer [8] have developed a theoretical framework for propagating beliefs and probabilities through an evidential network. We use the models derived by Srivastava, Shenoy, and Shafer [14] for propagating beliefs in an 'and' tree evidential network and use Srivastava [9] for combining beliefs from multiple items of evidence on the same variable. The combined beliefs at the main assertion X in an 'and' tree and the combined belief on any one of the objectives (sub-assertions) that is connected to the main assertion through an 'and' relationships can be expressed in terms of the following two propositions [14].

**Proposition 1** (*Propagation of **m**-values from sub-objectives to the main objective* ): The resultant **m**-values propagated from n sub-objectives ($O_i$, i = 1, 2, . . . n) to the main objective X in an AND-tree are given as follows.

$$\mathrm{m}_{\mathrm{X}\leftarrow\text{all O's}}(x) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{m}_{\mathrm{O_i}}(o_\mathrm{i}),\ \mathrm{m}_{\mathrm{X}\leftarrow\text{all O's}}(\sim x) = 1\text{-}\prod_{\mathrm{i}=1}^{\mathrm{n}}[1\text{-}\ \mathrm{m}_{\mathrm{O_i}}(\sim o_\mathrm{i})]$$

and

$$\mathrm{m}_{\text{X-all O's}}(\{x, \sim x\}) = 1 - \mathrm{m}_{\text{X-all O's}}(x) - \mathrm{m}_{\text{X-all O's}}(\sim x)$$

**Proposition 2** (*Propagation of **m**-values to a given sub-objective from the main objective and the other sub-objectives* ): The resultant **m**-values propagated to a given sub-objective $O_i$ from the main objective X and the other n-1 sub-objectives in an AND-tree are given as follows.

$$\mathrm{m}_{\mathrm{O_i}\leftarrow\text{X\&All other O's}}(o_\mathrm{i}) = \mathrm{K}_\mathrm{i}^{-1}\mathrm{m}_\mathrm{X}(x)\prod_{\mathrm{j}\neq\mathrm{i}}[1\text{-}\ \mathrm{m}_{\mathrm{O_j}}(\sim o_\mathrm{j})]$$

$$\mathrm{m}_{\mathrm{O_i}\leftarrow\text{X\&All other O's}}(\sim o_\mathrm{i}) = \mathrm{K}_\mathrm{i}^{-1}\mathrm{m}_\mathrm{X}(\sim x)\prod_{\mathrm{j}\neq\mathrm{i}}\mathrm{m}_{\mathrm{O_j}}(o_\mathrm{j})$$

$$\mathbf{m}_{\mathrm{O_i}\leftarrow\text{X\&All other O's}}(\{o_\mathrm{i}, \sim o_\mathrm{i}\}) = 1 - \mathbf{m}_{\mathrm{O_i}\leftarrow\text{X\&All other O's}}(o_\mathrm{i})$$
$$- \mathbf{m}_{\mathrm{O_i}\leftarrow\text{X\&All other O's}}(\sim o_\mathrm{i}).$$

where $K_i$ is the renormalization constant which is given by $K_i = [1\text{-}\ \mathbf{m}_X(x)C_i]$, where $C_i$ is given by $C_\mathrm{i} = 1\ \text{-} \prod_{\mathrm{j}\neq\mathrm{i}}[1\ \text{-}\ \mathrm{m}_{\mathrm{O_j}}(\sim o_\mathrm{j})]$.

To combine multiple independent items of evidence, say n, on a single binary variable, $\{x, \sim x\}$, we use the formulas developed by Srivastava [9]. The combined belief masses can be expressed as:

$$\mathrm{m}(x) = 1 - \prod_{\mathrm{i}=1}^{\mathrm{n}}\left(1 - \mathrm{m}_{\mathrm{i}}(x)\right)/\mathrm{K}\,,\ \ \mathrm{m}(\sim x) = 1 - \prod_{\mathrm{i}=1}^{\mathrm{n}}\left(1 - \mathrm{m}_{\mathrm{i}}(\sim x)\right)/\mathrm{K}\,,\ \ \mathrm{m}(\Theta) = \prod_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{m}_{\mathrm{i}}(\Theta)/\mathrm{K}$$

where K is given by

$$\mathrm{K} = \prod_{\mathrm{i}=1}^{\mathrm{n}}\left(1 - \mathrm{m}_{\mathrm{i}}(x)\right) + \prod_{\mathrm{i}=1}^{\mathrm{n}}\left(1 - \mathrm{m}_{\mathrm{i}}(\sim x)\right) - \prod_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{m}_{\mathrm{i}}(\Theta)\,.$$

We program the above formulas in MS Excel to perform the analysis.

## 4 Evidential Reasoning Approach applied to Sustainability Reporting

In this section, the hypothetical case presented in Figure 2 is used to illustrate the propagation of strength of evidence (i.e., belief masses[2] or m-values) obtained from various items of SR evidence to a set of assertions of interest. A similar example is then used in Section 5 to illustrate audit planning.

First, we illustrate the propagation of strength of evidence in terms of m-values (belief masses) from sub-assertions to the main assertion which is *Complete and Accurate disclosure of Labor Practices* and is abbreviated as A1.1. Then we illustrate the propagation of m-values to a particular sub-assertion from the main assertion and all other sub-assertions. In particular, we choose Assertion A1.1.1: *Complete & Accurate disclosure of Conditions & Benefits of Employment* as the sub-assertion of interest. We use upper case letters to represent the name of the variables such as 'A1.1.1' for the assertion A.1.1.1 and lower case letters to represent their values. For example, 'a111' represents the situation where 'A1.1.1' is true and '~a111' the state where A1.1.1 is not true. Additionally, we label the evidence items with 'En' to signify the evidence number. Abbreviations and symbols used are listed in Table 2.

----- Table 2 about here -----

### 4.1 Combination of Audit Evidence Relevant to the Main Assertion

Consider the propagation of strength of evidence from sub-assertions (A1.1.1, A1.1.2, … A1.1.N) to the main assertion (A1.1) (Figure 2). To simplify the computations, we transform the evidential diagram from a network structure to a tree structure[3] using the following process. Suppose we have evidence that pertains to two sub-assertions. We split this evidence into two different items of evidence relating individually to the two sub-assertions. For example, in Figure 2, evidence E4 is linked to sub-assertion A1.1.1 and to sub-assertion A1.1.2. The partitioned input m-values are assumed to be as follows:

$$\mathrm{m}_{\mathrm{E4}}(\mathrm{a111}) = 0.7,\ \mathrm{m}_{\mathrm{E4}}(\sim\mathrm{a111}) = 0.1,\ \mathrm{m}_{\mathrm{E4}}(\{\mathrm{a111}, \sim\mathrm{a111}\}) = 0.2,$$
$$\mathrm{m}_{\mathrm{E4}}(\mathrm{a112}) = 0.7,\ \mathrm{m}_{\mathrm{E4}}(\sim\mathrm{a112}) = 0.1,\ \mathrm{m}_{\mathrm{E4}}(\{\mathrm{a112}, \sim\mathrm{a112}\}) = 0.2.$$

That is, we assume equal evidential support for the two sub-assertions. However, in general, one can choose different levels of support for each sub-assertion.

[2] We assume that readers have basic background on the DS theory of belief functions [12].

[3] Srivastava and Lu [10] demonstrate that a tree-structured evidential diagram is a good approximation of a network structure under conditions that are relevant here.

We combine multiple items of evidence at each sub-assertion using the approach described in Section 3 [9] to obtain updated m-values at each sub-assertion. Next, we use Srivastava, Shenoy and Shafer [14] to propagate the evidence impounded in the above m-values from the six sub-assertions to the assertion 'A1.1' through the 'and' relationship. Finally, we combine the above m-values propagated to 'A1.1' from the six sub-assertions with the m-values obtained from the evidence directly bearing on 'A1.1'. The resulting m-values are the updated belief masses at ‘A1.1’ given all of the audit evidence bearing on the six sub-assertions (i.e. E2, E3 … E9), the desired result.

Consider the following scenario for our illustration. Suppose an assurance provider is collecting evidence pertaining to sub-assertion A1.1.1 and plans and obtains three relevant items of evidence for A1.1.1, namely E2, E3 and E4 (See Figure 2). The assurance provider examines evidence E2, that is, reviews and recalculates payroll data and confirms minimum wages and pay scales with a sample of employees. The auditor then decides that these procedures provide positive support for A.1.1.1 to the extent of 0.7, no support for its negation and thus a resulting in an unresolved uncertainty level of 0.3. In other words, these audit tests allow the auditor to be 70% confident that the client has complete and accurate disclosure of employment conditions and benefits. However, as the audit test provides no evidence to the contrary, thus there is still 30% uncertainty.

The assurance provider then reviews benefits provided to full-time employees that are not provided to part-time employees (E3) and decides that these audit procedures provide evidence in support of A1.1.1 of 0.6. Again, the assurance provider does not find any evidence that provides negative support for A1.1.1. Here, the resulting level of uncertainty is 0.4.

The assurance provider proceeds to review labor lawsuits to find out the number and cause of such lawsuits (E4) and decides that the evidence provides support in favor of A1.1.1 of 0.7 and provides negative support for A1.1.1 of 0.1, which leaves the level of uncertainty to 0.2. These input m-values are based on the assurance provider’s assessment of the evidence and judgment. Similarly, the assurance provider determines m-values for all other items of evidence as given in columns 3-5 in Table 3.

As noted, the first step in propagating belief masses from the sub-assertions to the main assertion is to determine the total belief masses at each sub-assertion based on all items of evidence directly bearing on each sub-assertion. Using Dempster’s rule, the combined m-values of the three items of evidence, E2, E3, and E4, bearing directly on the sub-assertion A1.1.1 are m(a111) = 0.961, m(~a111) = 0.013, m({a111, ~ a111}) = 0.026. This means that when evidence E2, E3 and E4 are combined, the combined strength of evidence indicating that A1.1.1 is valid is 0.961, the combined strength of evidence implying that A1.1.1 is not valid is 0.013, and the remaining uncertainty about A1.1.1 is 0.026. Similarly, we determine the total m-values at each sub-assertion. These values are listed in columns 6-8 in Table 3.

Next, we use Proposition 1 of [14] to propagate m-values from sub-assertions to the main assertion A1.1. The combined strength of evidence at A1.1 propagated from the sub-assertions yields the following m-values:

$$m(a11) = 0.521, m(\sim a11) = 0.101\ m(\{a11, \sim a11\}) = 0.378.$$

The assurance provider has one additional item of evidence to consider, specifically E1. Regarding E1, suppose that the assurance provider examines a sample of labor reports filed by the client and decides that they provide evidence in support of A1.1 to the extent of 0.7, as the labor reports are judged to have a good degree of objectivity and reliability. In the assurance provider's opinion, these labor reports provide no negative evidence for A1.1, leaving the level of uncertainty about A1.1 to be 0.3 given this particular audit test. Thus, we derive the following set of belief masses obtained from E1 for A1.1:

$$m(a11) = 0.7,\ m(\sim a11) = 0.0,\ m(\{a11, \sim a11\}) = 0.3.$$

To determine the overall belief masses at the main assertion level, A1.1, we combine the belief masses obtained from E1 with the belief masses propagated from the sub-assertions, A1.1.1. – A1.1.6 (see Figure 2). This yields the following overall m-values: m(a11) = 0.846, m(~a11) = 0.032, m({a11, ~a11}) = 0.122 (see columns 9-11 in Table 3). This means that the combined audit evidence confirming the assertion that the organization completely and accurately discloses its labor practices is 0.846, the combined evidence disconfirming the assertion is 0.032 and the level of uncertainty is 0.122.

----- Table 3 about here -----

The assurance provider can then use the above information to decide whether the 'Labor Practices' assertion is valid or not or whether additional evidence needs to be collected. In the illustration, the evidence confirming the assertion is a moderate level of 0.846, the evidence disconfirming the assertion is only 0.032. However, the plausibility that the assertion is not valid is 0.154. If the Srivastava and Shafer [13] plausibility definition of audit risk is used, the audit risk that the assurance is not true is 0.154 (i.e., 15.4%).

Given that the belief that the assertion is true is 0.846, the SR assurance provider has two main alternatives. First, the auditor could conclude and report that the assertion is fairly stated at what might be considered a 'moderate' level of assurance. Or, the auditor could continue to collect audit evidence to the point where the plausibility of misstatement was much lower (it is conventional to use 5% in a financial audit). An approach to obtaining such evidence at minimum cost is a topic of future study.

A third possibility is to conclude that the evidence suggests that the assertion is not valid, but this would be unlikely given the evidence only supports a very small belief of 0.032 supporting such a conclusion. Given the low belief in misstatement, the assurance provider could opine that the main assertion is fairly stated at an acceptable level of audit risk; describe the nature of any observed deficiencies in labor practices; and identify specific areas the management should focus on to mitigate such deficiencies. SR assurance standards and practices provide much more flexibility than conventional financial statement audit reports in what may be communicated [5] [6].

### 4.2 Combination of Evidence at a Sub-assertion

Evidential networks are somewhat peculiar in that the information obtained at each node flows to all other connected nodes [11] [15]. To consider this aspect, we use sub-assertion A1.1.1: *Complete & Accurate disclosure of Conditions & Benefits of Employment* – to exemplify the propagation of strength of evidence from assertion A1.1 and from the other sub-assertions to the chosen sub-assertion (A1.1.1).

For the Figure 2 case, the m-values from various items of evidence at the sub-assertions and the overall assertion are given in Table 3. The input m-values are assumed to be based on the assurance provider's assessment of the various strength of evidence provided by each audit procedure as indicated in columns 3, 4 and 5.

As in the previous case, we use the same input m-values at A1.1 from evidence E1: $m_{E1}(a11) = 0.7$, $m_{E1}(\sim a11) = 0$, and $m_{E1}(\{a11,\sim a11\}) = 0.3$ (see row 1, and columns 3-5 in Table 3). To determine the overall combined m-values at sub-assertion A1.1.1, three sets of m-values must be combined. One set comes from A1.1 (i.e., from E1), another from the other sub-assertions, and the last set of m-values are defined at A1.1.1 originating from evidence E2, E3, and E4. We again use Dempster's rule and Srivastava [9] to combine the above m-values.

The resulting overall combined belief masses at A1.1.1 are: m(a111) = 0.988, m(~a111) = 0.004, m({a111,~a111}) = 0.008 (see columns 9-11 in Table 3). These values indicate that there is a very high positive support for A1.1.1 (0.988) and almost no support for the negation of the sub-assertion (0.004). Given this situation, the assurance provider should be confident that the sub-assertion A1.1.1 is valid, could provide a high level of assurance with little audit risk on this assertion, and thus would not need to perform any additional audit procedures. However, if the evidence provided less than the assurance provider's target acceptable level of belief, say 0.95, then the assurance provider should either perform additional procedures to obtain a higher level of assurance, qualify the opinion by listing any shortcomings or even provide a negative opinion of some sort suggesting that the assertion may not be 'fairly stated'.

## 5 Summary and Conclusion

We have demonstrated the use of an evidential reasoning framework based on the Dempster-Shafer theory of belief functions for SR assurance services. We use the G3 sustainability reporting guidelines to develop the evidential network illustrations.

Since this paper is a first attempt to apply the evidential reasoning approach to the assurance of sustainability reports, there are limitations as well as opportunities for future research. Our models likely do not identify all of the relevant variables or associated items of audit evidence. As a future project, we plan to incorporate cost of performing audit procedures in order to evaluate cost effectiveness. Since we use DS theory to represent uncertainties in the SR setting, future research should examine the empirical ramifications of using this approach. Future research should also explore other possibilities such as empirically derived cost functions.

## References

[1] Ballou, B, D. L. Heitger and C. E. Landes. (2006), The Rise of Corporate Sustainability Reporting. *Journal of Accountancy*, (December) 202, 6: 65-73.

[2] Burdge, R. J. (2002), Why is Social Impact Assessment the Orphan of the Assessment Process? *Impact Assessment and Project Appraisal* 20 (1):3-9.

[3] Global Reporting Initiative. (2006), G3 Sustainability Reporting Guidelines, www.globalreporting.org.

[4] Labuschagne, C., A. C. Brent, and R. P. G. van Erck. (2005), Assessing the sustainability performances of industries. *Journal of Cleaner Production* 13 (4):373-385.

[5] Mock, T. J., C. Strohm and K. M. Swartz. (2007), An Examination of Assured Sustainability Reporting. *Australian Accounting Review* 17 (1):67-77.

[6] Mock, T.J., S. Rao, and R.P. Srivastava. (2013), The Development of Worldwide Assured Sustainability Reporting. *Australian Accounting Review* (forthcoming).

[7] Shafer, G. (1976), A Mathematical Theory of Evidence, Princeton, N.J.:

*Princeton University Press.*

[8] Shenoy, P. P. and G. Shafer. (1990), Axioms for Probability and Belief-Function Computation, in Shachter, R. D., T. S. Levitt, J. F. Lemmer, and L. N. Kanal, eds., *Uncertainty in Artificial Intelligence*, 4, North-Holland: 169-198.

[9] Srivastava, R. P. (2005), Alternative Form of Dempster's Rule for Binary Variables. *International Journal of Intelligent Systems* 20, 8: 789-797.

[10] Srivastava, R. P. and H. Lu. (2002), Structural Analysis of Audit Evidence using Belief Functions," *Fuzzy Sets and Systems,* Vol. 131, Issues No. 1, October: 107-120.

[11] Srivastava, R. P. and T. J. Mock. (1999-2000), Evidential Reasoning for WebTrust Assurance Services. *Journal of Management Information Systems*. Winter 1999-2000, Vol.16.

[12] Srivastava, R. P., and T. Mock. (2002), In *Belief Functions in Business Decisions,* Physica-Verlag, Heidelberg, Springer-Verlag Company.

[13] Srivastava, R.P., and G. Shafer. (1992) Belief-Function Formulas for Audit Risk. *The Accounting Review*. 67, 2: 249-283.

[14] Srivastava, R.P., P. P. Shenoy and G. Shafer. (1995), Propagating Belief Functions in AND-Trees. *International Journal of Intelligent Systems* 10, 0: 647-664.

[15] Sun, L., R. P. Srivastava, and T. Mock. (2006), An Information Systems Security Risk Assessment Model under Dempster-Shafer Theory of Belief Functions. *Journal of Management Information Systems,* 22, 4: 109-142.

[16] Wallage, P. (2000), Assurance on Sustainability Reporting: An Auditor's View. *Auditing: A Journal of Practice & Theory* 19, Supplement: 53-65.

Table 2: List of Symbols and Their Descriptions

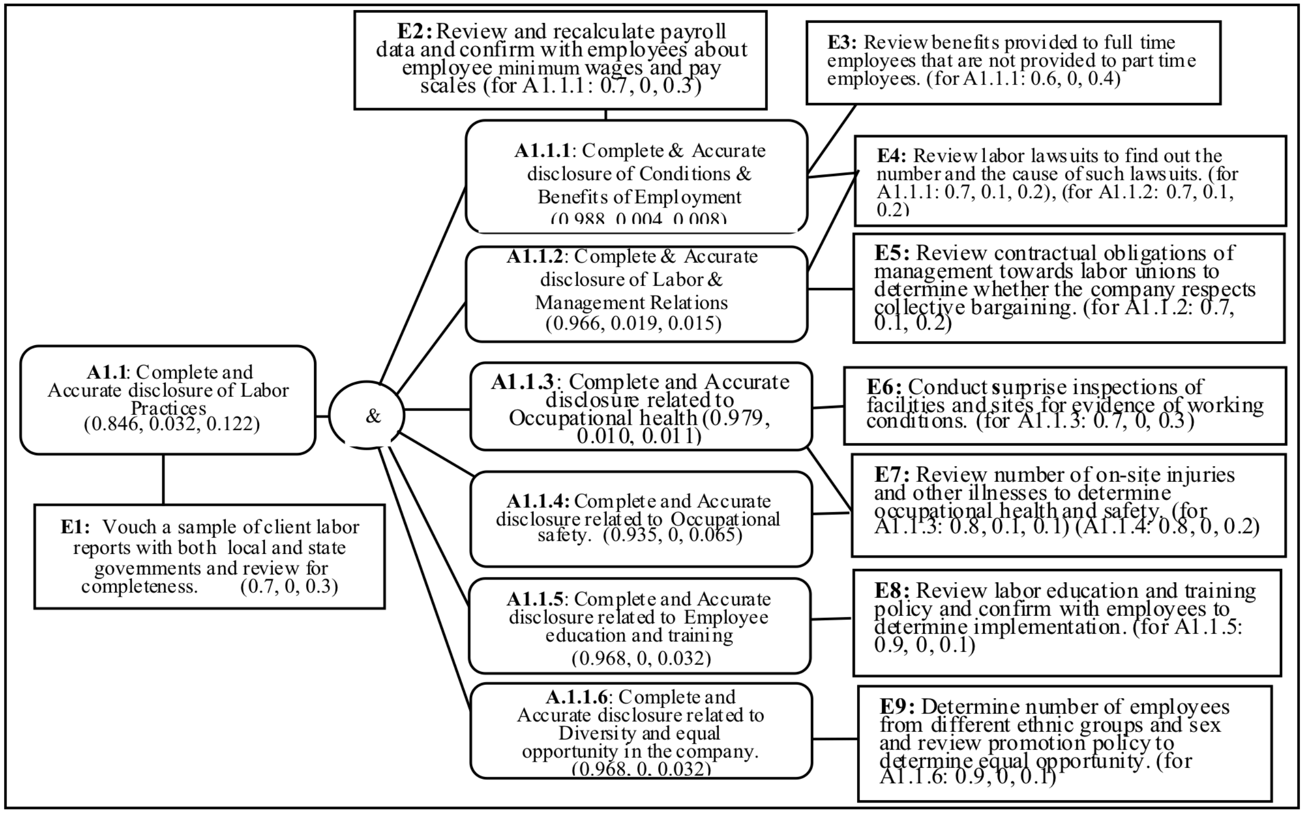
| Assertion and Sub-Assertion | Description of Assertion and Sub Assertion | Evidence/Related Assertion[s] | Audit Procedure |
|---|---|---|---|
| A1.1 | Complete and Accurate disclosure of Labor Practices | E1/ A1.1 | Vouch a sample of client labor reports with both local and state governments and review for completeness. |
| A1.1.1 | Complete & Accurate disclosure of Conditions & Benefits of Employment | E2/A1.1.1 | Review and recalculate payroll data and confirm with employees about employee minimum wages and pay scales |
| A1.1.2 | Complete & accurate disclosure of Labor & Management Relations | E3/A1.1.1 | Review benefits provided to full time employees that are not provided to part time employees. |
| A1.1.3 | Complete and accurate disclosure related to Occupational health | E4/A1.1.1 & A1.1.2 | Review labor lawsuits to find out the number and the cause of such lawsuits |
| A1.1.4 | Complete and accurate disclosure related to Occupational safety | E5/A1.1.2 | Review contractual obligations of management towards labor unions to determine whether the company respects collective bargaining. |
| A1.1.5 | Complete and accurate disclosure related to Employee education and training | E6/A1.1.3 | Conduct surprise inspections of facilities and sites for evidence of working conditions |
| A1.1.6 | Complete and accurate disclosure related to Diversity and equal opportunity in the company. | E7/A1.1.3 & A1.1.4 | Review number of on-site injuries and other illnesses to determine occupational health and safety. |
|  |  | E8/A1.1.5 | Review labor education and training policy and confirm with employees to determine implementation. |
|  |  | E9/A1.1.6 | Determine number of employees from different ethnic groups and sex and review promotion policy to determine equal opportunity. |

**Table 3: List of Input m-values and Overall m-values. The Assertion and Sub-Assertions along with the Corresponding Items of Evidence are defined in Table 2.**

| Assertion and Sub-assertion | Item of Evidence Pertaining to Assertion or Sub-Assertion | Positive | Negative | Θ* | Total m-values as a result of combining all the evidence directly bearing on the assertion and sub-assertions | | | Overall m-values | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Positive | Negative | Θ* | Positive | Negative | Θ* |
| A1.1 | E1 | 0.7 | 0 | 0.3 | 0.7 | 0 | 0.3 | 0.846 | 0.032 | 0.122 |
| A1.1.1 | E2 | 0.7 | 0 | 0.3 | 0.961 | 0.013 | 0.026 | 0.988 | 0.004 | 0.008 |
| | E3 | 0.6 | 0 | 0.4 | | | | | | |
| | E4 | 0.7 | 0.1 | 0.2 | | | | | | |
| A1.1.2 | E4 | 0.7 | 0.1 | 0.2 | 0.895 | 0.058 | 0.047 | 0.966 | 0.019 | 0.015 |
| | E5 | 0.7 | 0.1 | 0.2 | | | | | | |
| A1.1.3 | E6 | 0.7 | 0 | 0.3 | 0.935 | 0.032 | 0.032 | 0.979 | 0.010 | 0.011 |
| | E7 | 0.8 | 0.1 | 0.1 | | | | | | |
| A1.1.4 | E7 | 0.8 | 0 | 0.2 | 0.8 | 0 | 0.2 | 0.935 | 0 | 0.065 |
| A1.1.5 | E8 | 0.9 | 0 | 0.1 | 0.9 | 0 | 0.1 | 0.968 | 0 | 0.032 |
| A1.1.6 | E9 | 0.9 | 0 | 0.1 | 0.9 | 0 | 0.1 | 0.968 | 0 | 0.032 |

* The values in the column with heading Θ represent ignorance about the corresponding assertion or sub-assertion

**Figure 2. Evidential Diagram for Social Assertion Category A1.1: Labor Practices Performance is completely and accurately (Fairly) Stated**


**A1.1**: Complete and Accurate disclosure of Labor Practices (0.846, 0.032, 0.122)
&
**E1:** Vouch a sample of client labor reports with both local and state governments and review for completeness. (0.7, 0, 0.3)
**E2:** Review and recalculate payroll data and confirm with employees about employee minimum wages and pay scales (for A1.1.1: 0.7, 0, 0.3)
**A1.1.1**: Complete & Accurate disclosure of Conditions & Benefits of Employment (0.988, 0.004, 0.008)
**A1.1.2**: Complete & Accurate disclosure of Labor & Management Relations (0.966, 0.019, 0.015)
**A1.1.3**: Complete and Accurate disclosure related to Occupational health (0.979, 0.010, 0.011)
**A1.1.4:** Complete and Accurate disclosure related to Occupational safety. (0.935, 0, 0.065)
**A1.1.5**: Complete and Accurate disclosure related to Employee education and training (0.968, 0, 0.032)
**A.1.1.6**: Complete and Accurate disclosure related to Diversity and equal opportunity in the company. (0.968, 0, 0.032)
**E3:** Review benefits provided to full time employees that are not provided to part time employees. (for A1.1.1: 0.6, 0, 0.4)
**E4:** Review labor lawsuits to find out the number and the cause of such lawsuits. (for A1.1.1: 0.7, 0.1, 0.2), (for A1.1.2: 0.7, 0.1, 0.2)
**E5:** Review contractual obligations of management towards labor unions to determine whether the company respects collective bargaining. (for A1.1.2: 0.7, 0.1, 0.2)
**E6:** Conduct surprise inspections of facilities and sites for evidence of working conditions. (for A1.1.3: 0.7, 0, 0.3)
**E7:** Review number of on-site injuries and other illnesses to determine occupational health and safety. (for A1.1.3: 0.8, 0.1, 0.1) (A1.1.4: 0.8, 0, 0.2)
**E8:** Review labor education and training policy and confirm with employees to determine implementation. (for A1.1.5: 0.9, 0, 0.1)
**E9:** Determine number of employees from different ethnic groups and sex and review promotion policy to determine equal opportunity. (for A1.1.6: 0.9, 0, 0.1)

**Title: Proceedings of WUPES 2012**

**Publisher: University of Economics**

**Editors: Tomáš Kroupa, Jiřina Vejnarová**

**Cover design: CANDY BOX s.r.o.**

**ISBN 978-80-245-1885-5**