# COLLECTIVE REASONING UNDER UNCERTAINTY AND INCONSISTENCY

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2014

**Martin Adamčík**
School of Mathematics

# Contents

Word count 47941

# List of Tables

# List of Figures

# The University of Manchester

**Martin Adamčík**
**Doctor of Philosophy**
**Collective Reasoning under Uncertainty and Inconsistency**
**February 19, 2014**

In this thesis we investigate some global desiderata for probabilistic knowledge merging given several possibly jointly inconsistent, but individually consistent knowledge bases. We show that the most naive methods of merging, which combine applications of a single expert inference process with the application of a pooling operator, fail to satisfy certain basic consistency principles.

We therefore adopt a different approach. Following recent developments in machine learning where Bregman divergences appear to be powerful, we define several probabilistic merging operators which minimise the joint divergence between merged knowledge and given knowledge bases. In particular we prove that in many cases the result of applying such operators coincides with the sets of fixed points of averaging projective procedures — procedures which combine knowledge updating with pooling operators of decision theory.

We develop relevant results concerning the geometry of Bregman divergences and prove new theorems in this field. We show that this geometry connects nicely with some desirable principles which have arisen in the epistemology of merging. In particular, we prove that the merging operators which we define by means of convex Bregman divergences satisfy analogues of the principles of merging due to Konieczny and Pino-Pérez. Additionally, we investigate how such merging operators behave with respect to principles concerning irrelevant information, independence and relativisation which have previously been intensively studied in case of single-expert probabilistic inference.

Finally, we argue that two particular probabilistic merging operators which are based on Kullback-Leibler divergence, a special type of Bregman divergence, have overall the most appealing properties amongst merging operators hitherto considered. By investigating some iterative procedures we propose algorithms to practically compute them.

# Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

# Copyright Statement

# Acknowledgements

The author is in the first place indebted to his supervisor — Dr George Wilmers — who, as a father of multi-expert probabilistic reasoning, was the best source of expertise and support. Moreover, he stepped over the usual role of a supervisor and continued to supervise me even after the University of Manchester refused to provide him with the facilities to do so.

The second big thanks goes to Dr Alena Vencovská who carefully checked the proofs, found many mistakes and explained to me the complicated behaviour of entropy. My special thanks is devoted to Prof Jon Williamson who accommodated me in the University of Kent for two months as a visiting researcher. The author is also grateful to many people who work in the field of uncertain reasoning and who contributed with ideas, references and connections to other fields; namely to Prof Jeff Paris, Dr Jürgen Landes, Dr Soroush Rad, Dr Jiří Vomlel and Dr František Matúš.

The research leading to these results has received funding from the [European Community's] Seventh Framework Programme [FP7/2007-2013] under grant agreement no 238381.

# Chapter 1

# Introduction

## 1.1 Motivation

A typical problem of real-life is uncertainty. While classical propositional logic handles well the situation where each sentence is either true or false, several frameworks have been introduced to deal with reasoning in the context of uncertain knowledge. In the present work we consider the framework of *collective probabilistic knowledge*. Such knowledge is thought of as arising from several experts or sources each of which provides consistent probabilistic knowledge, while collectively their knowledge is typically inconsistent. Throughout this thesis we will assume that this representation of uncertainty is in particular suitable for developing a probabilistic expert system.

Several designs of expert systems have been developed for use in industry. A well known recent example of an expert system is IBM's *Watson* which has been to outperform contestants on a quiz show. However, the largest consumer of expert systems is the banking sector, see e.g. [47]. In relation to the recent banking crisis, there is even more need to deal with inconsistent opinion of several experts — a common issue with economic data.

In our opinion a new area where expert systems can make a significant difference is the public sector. This is because the public sector is challenged to adopt decisions which make people less unhappy, and to do it transparently and promptly. If the expert system satisfies certain natural principles in order better to reflect the joint opinion of a group, then this can make the process more transparent as it will be clear for the user of such a system which principles are engaged in the process.

In this thesis, however, we are not going as far as to investigate how a *decision procedure* of an expert system should be designed. We aim only to develop a mathematical framework for evaluating the probabilities on which a potential expert system may be based. In order to develop such a framework the initial assumption we make is that the probabilistic knowledge of each particular expert is consistent with the *laws of probability*. Moreover, the objective of reasoning, given collective probabilistic knowledge is to use rational criteria to produce consistent knowledge, which optimally represents the joint knowledge *declared* by the experts, but *no more*. It will be assumed that experts have equal status but we can in fact give different weights to experts by cloning them. We also assume that reasoning does not depend on the order in which the knowledge bases are considered, the way in which the knowledge was obtained is considered irrelevant, and each expert has incorporated all her relevant knowledge into what she is declaring.

We illustrate the motivation behind this idea by a toy two-expert example which was introduced in [1].

*Imagine that two safety experts are evaluating safety in a chemical factory producing nitrogen fertilizers. For simplicity we consider only the ammonia supply which is stored in a tank connected to the rest of the factory by a valve which is controlled by an electronic switch.*

*The first expert believes that there is a* 4% *chance that a mechanical problem will cause the valve to fail. The second expert comes up with a different opinion that there is an* 8% *chance that a mechanical problem will cause the valve to fail. Moreover, the first safety expert thinks that there is a* 7% *chance that the electronic switch will fail. We suppose that both experts have no other knowledge related to this problem.*

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

*How should a rational adjudicator, whose only knowledge consists of what is related to him by the two experts above, evaluate, for instance, the probability that both the valve and the electronic switch will be faulty, based only on the experts' subjective knowledge specified above and without any other assumptions?*

Under our previous assumption that each expert's 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: there will be (1) a fault on the valve and a fault on the electronic switch, (2) a fault on the valve and no fault on the electronic switch, (3) no fault on the valve and a fault on the electronic switch and (4) no faults on the valve or on the electronic switch. 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 $w_1 + w_2 = 0.04$ and $w_1 + w_3 = 0.07$. Any probability function $(x, 0.04 - x, 0.07 - x, x + 0.89)$, where $x \in [0, 0.04]$, is consistent with the knowledge of the first expert. Similarly, the second expert admits any probability function $(x, 0.08 - x, y, 0.92 - y)$ where $x \in [0, 0.08]$ and $y \in [0, 0.92]$. 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 probabilistic knowledge of corresponding experts as in the above example, which set of probability functions best represents the combined probabilistic knowledge of the experts?*

Naturally, we would like to find a general procedure answering the above question for any knowledge bases, and such that it satisfies some rational principles. We do not want to restrict the set of probability functions that should represent the knowledge of several experts otherwise than by rationality criteria. Therefore, we do not *a priori* suppose that this set is singleton; yet this will be sometimes a consequence of those criteria.

To conclude this section we make some further remarks on our initial assumptions. Firstly, while in practice humans do not always espouse belief values consistent with the laws of probability (see [28]), there are very strong arguments in the literature as to why *rational beliefs* should obey the laws of probability. Among them the *Dutch Book* argument by Ramsey and de Finetti [15] is perhaps the most compelling reason why the subjective belief of an expert should be represented by a probability distribution. The idea of this argument is that it identifies belief with willingness to bet. If we assume that there is no possibility that an expert will rationally agree to make a set of bets which will subject her to a sure loss, then her willingness to bet must be represented by a probability distribution. While nonprobabilistic frameworks of reasoning such as fuzzy logic and combinations of probabilistic framework and rule-based systems such as Dempster-Shafer theory also have important, but distinct traditions, we seek to extend the well-established theory of the classical probabilistic framework as developed by Jaynes, Cox, Johnson, Shore, Paris, Vencovská and others. For more details as to how the Dutch Book argument relates to our framework see [39]. For further justification of belief as probability see, for example, [50].

Secondly, the assumption that the representation of the knowledge declared by each of the experts by a set of probability distributions contains all the relevant information for answering the above question is sometimes referred to as the *Principle of Total Evidence* [8] or the *Watts assumption* [39]. In practice this assumption is trivially never satisfied as indeed our toy example illustrates. Overall knowledge of any human expert can never be fully formalised as a formalisation is always an abstraction from reality. However, the Principle of Total Evidence needs to be imposed in order to avoid confusion in any discussion related to methods of representing the collective knowledge of experts. Otherwise there would be an inexhaustible supply of invalid arguments as Wilmers in [1] and [2] pointed out:

> "... it is often extremely hard to give illustrative real world examples of abstract principles of probabilistic inference without a philosophical opponent being tempted to challenge one's reasoning using implicit background information concerning the example which is not included in its formal representation as a knowledge base."

In this thesis we argue that it is possible to answer the question which we have formulated here under the assumptions we have imposed. In order to do that, we first need to find a formal representation of knowledge.

## 1.2 Knowledge representation

In this section we introduce three basic notions used in this thesis — a probability function, a probabilistic knowledge base and a probabilistic merging operator. We will mostly follow the papers [1] and [2] by Wilmers and the author.

Let $L = \{a_1, \dots, a_h\}$ be a finite *propositional language* where $a_1, \dots, a_h$ are propositional variables. In our previous example $n = 2$, $a_1$ stands for sentence "a fault on the valve" and $a_2$ stands for sentence "a fault on the electronic switch". We denote the set of all *propositional sentences* which can be defined over $L$ by $SL$. By the disjunctive normal form theorem any sentence in $SL$ is logically equivalent to a disjunction of atomic sentences (atoms) where each atom is of the form $\bigwedge_{i=1}^{h} \pm a_i$ and $\pm a_i$ denotes either $a_i$ or $\neg a_i$. We denote an enumeration of these atoms in some fixed order by $\alpha_1, \dots, \alpha_J$, where $J = 2^h$. The set $\{\alpha_1, \dots, \alpha_J\}$ of all atoms of $L$ will be denoted by $\mathrm{At}(L)$. Notice that atoms of $\mathrm{At}(L)$ are 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 sentence $\varphi \in SL$ may then be defined by setting

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

Note that formula $\varphi$ which is not satisfiable, e.g. $a_1 \wedge \neg a_1$, is defined as the disjunction of an empty set of atoms and we set $\mathbf{w}(\varphi) = 0$ in this case. 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 that 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.

In several places we will need to consider language $L$ to be not fixed. Therefore, at this point, we introduce useful notation which we will use to examine such a situation. 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 manner as for an $L_1$-probability function $\mathbf{v}$ we write $v_j$ instead of $\mathbf{v}(\alpha_j)$.

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 $U$ is a set of $L_1 \cup L_2$-probability functions, we denote the set $\{\mathbf{v}|_{L_1} : \mathbf{v} \in U\}$ by $U|_{L_1}$. 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 $\mathbf{v} \cdot \mathbf{w}(\alpha_j \wedge \beta_i) = v_j w_i$ is an $L_1 \cup L_2$-probability function such that $(\mathbf{v} \cdot \mathbf{w})|_{L_1} = \mathbf{v}$.

A particular probability function $\mathbf{v} \in \mathbb{D}^L$ gives us the full information about the distribution of the atomic events $\mathrm{At}(L)$. Where several probability functions are given it is not known which probability function applies. From this point of view the set $\mathbb{D}^L$ carries no information — any probability function is possible. In our toy example the first expert stated two constraints on the possible probability functions. In effect these constraints generate a subset of $\mathbb{D}^L$. A particular kind of such subsets will now play a prominent role in our *knowledge representation*.

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 $V_{\mathbf{K}}^L$ of $\mathbb{D}^L$. For brevity we shall use the terminology *knowledge base* instead of probabilistic knowledge base. $V_{\mathbf{K}}^L$ will be semantically thought of as the set of possible probability functions of a particular expert which are consistent with her subjective probabilistic knowledge base $\mathbf{K}$ and we will be never concerned with the syntactical definition of $V_{\mathbf{K}}^L$. In particular if $\mathbf{K}_1$ and $\mathbf{K}_2$ are such that $V_{\mathbf{K}_1}^L = V_{\mathbf{K}_2}^L$ we shall say that $\mathbf{K}_1$ and $\mathbf{K}_2$ are *equivalent*. In practice we shall only be interested in knowledge bases up to equivalence, and consequently we may sometimes informally identify a knowledge base $\mathbf{K}$ with its extension $V_{\mathbf{K}}^L$, and with

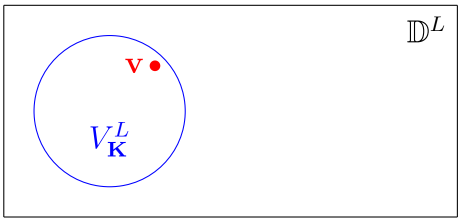


Figure 1.1: The illustration of a knowledge base.

slight abuse of language we may also refer to a nonempty closed convex subset of $\mathbb{D}^L$ as a knowledge base. Note that the non-emptiness of $V^L_{\mathbf{K}}$ 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 $V^L_{\mathbf{K}\cup\mathbf{F}} = V^L_{\mathbf{K}} \cap V^L_{\mathbf{F}}$. The set of all knowledge bases $V^L_{\mathbf{K}}$ over fixed $L$ is denoted by $CL$. Figure 1.1 illustrates $\mathbf{v} \in V^L_{\mathbf{K}} \subseteq \mathbb{D}^L$.

Note that in our definition a knowledge base is associated with the underlying language. We, however, take some liberty to loosen this connection. This is because if $L_1 \subset L_2$ and $\mathbf{K} \in CL_1$ then $\mathbf{K}$ is also in $CL_2$ and $V^{L_2}_{\mathbf{K}} = \{\mathbf{w} \in \mathbb{D}^{L_2}\colon \mathbf{w}|_{L_1} \in V^{L_1}_{\mathbf{K}}\}$. Given $\mathbf{K}$, the underlying language $L_1$ is usually only implicitly understood.

The reason why we have restricted our knowledge bases only to a nonempty closed convex subset of $\mathbb{D}^L$ might be unclear at the moment but later on we will see that this restriction is useful in order to define a relationship between a single probability function and a knowledge base. On the other hand, neither of these requirements is severe in a real world situation. Any set of linear constraints such as in the toy example generates always a nonempty closed convex set of probability functions. In the following we will give such a knowledge base a special name.

Given a language $L$ consider the following system $\mathbf{K}$ of linear constraints for a probability function $Bel$:

$$\Big\{ \sum_{s=1}^{m} q_{si} Bel(\theta_s) = p_i\colon\ i \in I \Big\},$$
$$\sum_{j=1}^{J} Bel(\alpha_j) = 1,$$
$$\Big\{ 0 \leq Bel(\alpha_j),\ 0 \leq j \leq J \Big\},$$

where $\alpha_1, \ldots, \alpha_J$ are all the atoms in $\mathrm{At}(L)$, $I$ is an index set, $q_{si}, p_i \in \mathbb{R}$, and $\theta_s$,

$1 \leq s \leq m$, are satisfiable $L$-sentences.[1] If there exists a probability function $Bel \in \mathbb{D}^L$ which satisfies the system $\mathbf{K}$ above then we call it a *linearly constrained set*. Note that not all knowledge bases are linearly constrained sets. To avoid confusion we will use the notation $Bel$ solely to define constraints in a linearly constrained set. We say that $\mathbf{w} \in \mathbb{D}^L$ satisfies $\mathbf{K}$ if every constraint in $\mathbf{K}$ holds with $Bel$ replaced by $\mathbf{w}$.

In the toy example, the knowledge of the first expert can be represented by the knowledge base $\mathbf{K}$ which consists of a set of linear constraints on a probability function $Bel$ defined over the atomic sentences $a_1 \wedge a_2$, $a_1 \wedge \neg a_2$, $\neg a_1 \wedge a_2$ and $\neg a_1 \wedge \neg a_2$. That is $\mathbf{K} = \{Bel(a_1 \wedge a_2) + Bel(a_1 \wedge \neg a_2) = 0.04,\ Bel(a_1 \wedge a_2) + Bel(\neg a_1 \wedge a_2) = 0.07\}$ and $V^L_{\mathbf{K}} = \{(x, 0.04 - x, 0.07 - x, x + 0.89) : x \in [0, 0.04]\}$, where $(x, 0.04 - x, 0.07 - x, x + 0.89)$ denotes values of a probability function $\mathbf{w} \in V^L_{\mathbf{K}}$ over atoms listed in the above order. With slight abuse of notation we may write $\mathbf{w} = (x, 0.04 - x, 0.07 - x, x + 0.89)$ when the order of atoms is implicitly understood.

Another restricted notion of knowledge base is a knowledge base *which bounds probability functions away from zero*. This is a knowledge base $\mathbf{K} \in CL$ such that $\mathbf{w} \in V^L_{\mathbf{K}}$ satisfies a set of constraints of the form

$$\{a_j \leq w_j \colon 1 \leq j \leq J\},$$

where $0 < a_j < 1$ for all $j = 1, \ldots, J$. We call such a knowledge base *bounded*, and we will denote the set of all bounded knowledge bases for a given language $L$ by $BCL$. A slightly more general notion is that of a knowledge base $\mathbf{K} \in CL$ which does not "force" any atom to take the value zero. More precisely, we call $\mathbf{K}$ *weakly bounded* if for every $1 \leq j \leq J$ there is $\mathbf{w} \in V^L_{\mathbf{K}}$ such that $w_j \neq 0$. The set of weakly bounded knowledge bases for $L$ will be denoted by $WBCL$. Note that $BCL \subset WBCL \subset CL$ and that by convexity if $\mathbf{K} \in WBCL$ then there exists some $\mathbf{w} \in V^L_{\mathbf{K}}$ such that $w_j \neq 0$ for all $j = 1, \ldots, J$. For example the linearly constrained set $\mathbf{K}$ of the first expert in our toy example is a weakly bounded knowledge base but $\mathbf{K} \notin BCL$.

There are several possible motivations for studying knowledge bases with a boundedness condition imposed. Broadly speaking, the imposition of such a condition may avoid some of the potentially intractable technical and philosophical difficulties which

[1] A constraint such as $Bel(\psi \mid \theta) = c$, where $\psi$ and $\theta$ are satisfiable L-sentences, is interpreted as $Bel(\psi \wedge \theta) = c \cdot Bel(\theta)$ which makes sense as a linear constraint even though $Bel(\theta)$ may take the value zero (see [39] for details).

arise from treating zero probabilities in certain contexts. In this thesis we will confine ourselves to stating and proving some theorems concerning certain classes of knowledge base, but will not consider further the epistemological status of the various notions of knowledge base.

We now define the central notion of this thesis. Let $\Delta$ denote an operator defined for all $n \geq 1$ and all $L$ as a mapping

$$\Delta_L : \underbrace{CL \times \ldots \times CL}_{n} \to \mathcal{P}(\mathbb{D}^L),$$

where $\mathcal{P}(\mathbb{D}^L)$ denotes the power set of $\mathbb{D}^L$. The number $n$ here intuitively represents the number of distinct experts or distinct sources of information. We will call such a $\Delta$ a *probabilistic merging operator*, abbreviated to *p-merging operator*, if it satisfies the following

**(K1) Defining Principle.** If $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ then the set $\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is closed, convex and nonempty.

**(K1)** ensures that a p-merging operator applied to a product of knowledge bases yields a knowledge base. A p-merging operator which always results in a single probability function (i.e. $\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is a singleton for all $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$) is called a *social inference process.*

The idea of probabilistic merging has appeared in many places in the literature, including [29], [38], [42], [50], [51] and [52] which will be discussed in more detail later. However, only a few authors have tried to investigate global desiderata of knowledge merging. We challenge that in this thesis. In particular the definition above together with the notion of knowledge base allows us formally state the problem which has been outlined in section 1.1 by the following list of aims to be investigated in this thesis.

- What p-merging operator $\Delta$ should be used to represent knowledge of any college (group) of experts?

- What general principles should $\Delta$ satisfy?

- How can the choice of p-merging operator be justified subject to rationality and computability?

In other words we would like to find a general procedure which represents several given knowledge bases as only a single knowledge base. At this point one needs to realize that although there are several possible options for defining such a procedure $\Delta$, we need to find a way to investigate whether any such definition is rational; whether it satisfies our intuition about well behaved probabilistic expert systems. For instance the operator $\Delta$ which returns always the set $\mathbb{D}^L$ is obviously a p-merging operator however it is hardly a *rational one*.

A general idea is to introduce an additional structure on the set of probability functions. In chapter 2 we will define one which allows us to talk about a measure between probability functions. This resembles a geometry imposed on the set of probability functions and it will prove to be fairly useful. In the next section we show how a different kind of structure appears to have solved a restricted version of the central problem defined above.

## 1.3 Previous research on inference processes

If we restrict the central question which we have formulated in the previous section to the case when only one expert is considered and the operator $\Delta$ is required to produce a single probability function then we obtain the following problem which has been extensively studied in the literature: Given $\mathbf{K} \in CL$, by which procedure we should choose a specific probability function from $V^L_{\mathbf{K}}$? In general, such a procedure is referred to as an *inference process*. Formally, an inference process $\mathcal{S}$ for any $L$ gives a mapping $\mathcal{S}_L : CL \to \mathbb{D}^L$.

In this section we list the results which support the claim that the most rational answer is given by the associated structure with probability functions whose origins go back to nineteenth century statistical mechanics as in [25] or [40] — *the Shannon entropy* $H$. For any $\mathbf{w} \in \mathbb{D}^L$ entropy is defined by

$$H(\mathbf{w}) = -\sum_{j=1}^{J} w_j \log w_j,$$

where log denotes the natural logarithm. For $w \geq 0$ we define $-w \log 0 = 0$ if $w = 0$ and $-w \log 0 = +\infty$ otherwise.

Entropy can be interpreted both as a measure of disorder and as a measure of information content. The higher entropy of $\mathbf{w}$ is, the less information is carried by $\mathbf{w}$. One can then argue that given several seemingly equally probable choices of a probability function one should choose the one which carries the least additional information. Jaynes [26] supported this argument by the following claim.

> "If the information incorporated into the maximum-entropy analysis includes all the constraints actually operating in the random experiment, then the distribution predicted by maximum entropy is overwhelmingly the most likely to be observed experimentally."

Since $H$ is a strictly concave function, and therefore attains a unique maximum over any nonempty closed convex region $V^L_{\mathbf{K}}$, there is always a unique maximum entropy point in $V^L_{\mathbf{K}}$, which we denote by $\mathbf{ME}_L(\mathbf{K})$ and refer to informally as *the most entropic point* of $V^L_{\mathbf{K}}$. The family of mappings $\mathbf{ME}_L : CL \to \mathbb{D}^L$ is called the *maximum entropy inference process*. By only a little abuse of notation given a nonempty closed convex set $V \subseteq \mathbb{D}^L$ we will denote the most entropic probability function in $V$ by $\mathbf{ME}_L(V)$.

The result of applying $\mathbf{ME}$ can be computed for a linearly constrained set. Let $|\operatorname{At}(L)| = J$. Given a linearly constrained set $\mathbf{K} = \{\sum_{r=1}^J a_{rk} Bel(\alpha_r) = b_k,\, k = 1,\dots,m\}$ over $L$ one can compute $\mathbf{ME}_L(\mathbf{K})$ by using Lagrange multipliers, that is by solving the system

$$\Big\{\frac{\partial}{\partial w_j}\Big[H(\mathbf{w}) + \sum_{k=1}^m \lambda_k(\sum_{r=1}^J a_{rk}w_r - b_k)\Big] = 0,\, j = 1,\dots,J\Big\}$$

for variables $w_1,\dots,w_J$ and $\lambda_1,\dots,\lambda_m$ subject to $\mathbf{K}$. Since $H$ is a strictly concave function the above system has a unique solution. Note that one of the $m$ linear constraints is the equation $\sum_{j=1}^J w_j = 1$. However *practically* the problem of actually computing weights to any reasonable approximation is NP-hard, see p. 148 of [39]. We will discuss related computational issues later in this thesis. For more details on computation of $\mathbf{ME}$ see [22].

There are several other inference processes which have been extensively studied by Paris, Vencovská and others ([24] or [39]) but it is perhaps the maximum entropy inference process which is best justified. We will now list some of these justifications.

**Model theoretic justification.**

By the traditional possible worlds modelling or information theoretic arguments, the maximum entropy inference process **ME** has been justified as being optimal by Paris and Vencovská in [40] extending earlier work of Jaynes, Aczel and others. In that paper distributions of a large universe of examples were studied, examples have properties described by some propositional language $L$ and the proportion of examples possessing certain properties were subject to linear constraints. They proved that nearly all possible distributions of examples from given universe subject to the given linear constraints are distributed close to the distribution which has the largest entropy and actually satisfies those linear constraints, thus providing a justification for the maximum entropy inference process. This is an analogous argument to the one used in the statistical mechanic of gases which states that the most likely distribution of molecules of gas has the most entropy — the most level of disorder. □

**Axiomatic justification.**

A different justification for **ME** to the traditional ones was described in [27] by Johnson and Shore. Their work was developed by Paris and Vencovská in [41] where they showed that a list of principles based on symmetry and consistency uniquely characterises **ME** once we have confined inference processes only to linearly constrained sets.

Let $\mathcal{S}$ be an inference process. Paris and Vencovská argue that the following list of principles is rational for any such inference process.

**Equivalence.** If $\mathbf{K}_1, \mathbf{K}_2 \in CL$ are such that $V^L_{\mathbf{K}_1} = V^L_{\mathbf{K}_2}$ then $\mathcal{S}_L(\mathbf{K}_1) = \mathcal{S}_L(\mathbf{K}_2)$.

**Atomic Renaming.** If $\mathbf{K}_1, \mathbf{K}_2 \in CL$ are

$$\mathbf{K}_1 = \Big\{ \sum_s p_{si} Bel(\gamma_s) = q_i,\ i = 1, \ldots, n \Big\},$$

$$\mathbf{K}_2 = \Big\{ \sum_s p_{si} Bel(\delta_s) = q_i,\ i = 1, \ldots, n \Big\},$$

where $\gamma_1, \ldots, \gamma_J$, and $\delta_1, \ldots, \delta_J$ are permutations of $L$-atoms $\alpha_1, \ldots, \alpha_J$. Then for all $1 \leq j \leq J$

$$\mathcal{S}_L(\mathbf{K}_1)(\gamma_j) = \mathcal{S}_L(\mathbf{K}_2)(\delta_j).$$

**Irrelevant Information.** Let $L_1 \cap L_2 = \emptyset$, $\mathbf{K}_1 \in CL_1$, $\mathbf{K}_2 \in CL_2$ and $\varphi$ is $L_1$-sentence. Then $\mathcal{S}_{L_1 \cup L_2}(\mathbf{K}_1 \cup \mathbf{K}_2)(\varphi) = \mathcal{S}_{L_1 \cup L_2}(\mathbf{K}_1)(\varphi)$.

**Obstinacy.** Let $\mathbf{K}_1, \mathbf{K}_2 \in CL$ and $\mathcal{S}_L(\mathbf{K}_1) \in V^L_{\mathbf{K}_2}$. Then $\mathcal{S}_L(\mathbf{K}_1 \cup \mathbf{K}_2) = \mathcal{S}_L(\mathbf{K}_1)$.

**Independence.** Let $L = \{p_1, p_2, p_3\}$ and $\mathbf{K} = \{Bel(p_3) = c, Bel(p_1 \wedge p_3) = a, Bel(p_2 \wedge p_3) = b\}$, $c > 0$. Then

$$\mathcal{S}_L(\mathbf{K})(p_1 \wedge p_2 \wedge p_3) = \frac{ab}{c}.$$

**Open-Mindedness.** If $\mathbf{K} \in CL$, $\varphi \in SL$ and $\mathbf{K} \cup \{Bel(\varphi) = c\}$ for $c > 0$ is consistent then $\mathcal{S}_L(\mathbf{K})(\varphi) \neq 0$.

**Relativisation.** Suppose $\mathbf{K}_1, \mathbf{K}_2 \in CL$, $0 < c < 1$ and

$$\mathbf{K}_1 = \{Bel(\theta) = c\} \cup \Big\{ \sum_j a_{ji} Bel(\varphi_j|\theta) = d_i,\ i = 1, \ldots, h \Big\},$$

$$\mathbf{K}_2 = \mathbf{K}_1 \cup \Big\{ \sum_j b_{ji} Bel(\psi_j|\neg\theta) = e_i,\ i = 1, \ldots, h' \Big\},$$

where $\theta$ is given and $\varphi_j, \psi_j \in SL$. Then for any $\varphi \in SL$

$$\mathcal{S}_L(\mathbf{K}_1)(\varphi|\theta) = \mathcal{S}_L(\mathbf{K}_2)(\varphi|\theta).$$

**Continuity.** If $\theta \in SL$, $\mathbf{K}, \mathbf{K}_1, \mathbf{K}_2, \ldots \in CL$ and $\lim_{i\to\infty} ||V^L_{\mathbf{K}}, V^L_{\mathbf{K}_i}||_B = 0$, where $||V, W||_B$ is *Blaschke metric* defined for convex subsets $V, W \subseteq \mathbb{D}^L$ by

$$||V, W||_B = \inf\{\delta \in \mathbb{R} \colon \forall \mathbf{v} \in V\ \exists \mathbf{w} \in W\ ||\mathbf{w}, \mathbf{v}|| \leq \delta \text{ and}$$

$$\forall \mathbf{w} \in W\ \exists \mathbf{v} \in V\ ||\mathbf{w}, \mathbf{v}|| \leq \delta\},$$

where $||\mathbf{w}, \mathbf{v}||$ is the usual Euclidean distance. Then

$$\lim_{i\to\infty} \mathcal{S}_L(\mathbf{K}_i)(\theta) = \mathcal{S}_L(\mathbf{K})(\theta).$$

The following theorem combines significant results of Paris and Vencovská, see [39] and [41].

**Theorem 1.3.1.** *1.* **ME** *satisfies all the principles above.*

*2. On the other hand, if some other inference process $\mathcal{S}$ satisfies all these principles once we have restricted knowledge bases to linearly constrained sets then $\mathcal{S}$ is equivalent to* **ME** *over linearly constrained sets.*

It seems fruitful to look at the axiomatic approach also for p-merging operators. We might hope that ultimately a set of rational principles may determine uniquely a particular p-merging operator. This idea in multi-expert context was originally put forward by Wilmers in [52] for a special case of social inference processes — i.e. for p-merging operators which always produce single probability functions. We investigate principles for p-merging operators in section 1.4 and chapter 4. □

**Game theoretic justification.**

Here we consider a justification for **ME** which is connected to game theory. First, we define a new binary function — a *loss function* — a structure associated with probability functions. A loss function Ls is any mapping

$$\mathrm{Ls}\colon \mathbb{D}^L \times \mathrm{At}(L) \to \mathbb{R}^*,$$

where $\mathrm{At}(L)$ is the set of all atomic events or atomic sentences and $\mathbb{R}^*$ is the *extended real line* $\mathbb{R} \cup \{+\infty\}$.

*A forecaster* is a decision maker whose decision depends on a random variable $X$ which takes the values from the set $\mathrm{At}(L)$ of some atomic events generated by a language $L$. The forecaster is willing to accept that the distribution of $X$ lies in some (closed and convex) set $V \subseteq \mathbb{D}^L$; however, forecaster does not know which of these probability functions applies.

Once the forecaster has chosen one probability function $\mathbf{v}$ in $V$, the loss function materializes the loss (or the reward when the loss is negative) of the forecaster when the atomic event $\alpha$ happens to be true. From a game-theoretic point of view this can be considered, according to [21], as a game between Nature and the forecaster. Nature takes no account of the forecaster's prediction and plays her event $\alpha$. After this the forecaster needs to pay Nature the value $\mathrm{Ls}(\mathbf{v}, \alpha)$.

A *scoring rule* is defined as the extended real-valued function[2] $S : \mathbb{D}^L \times \mathbb{D}^L \to \mathbb{R}^*$

$$S(\mathbf{w}, \mathbf{v}) = \sum_{\alpha \in \mathrm{At}(L)} \mathbf{w}(\alpha)\, \mathrm{Ls}(\mathbf{v}, \alpha).$$

If the forecaster knows with what distribution $\mathbf{w}$ Nature gives her outcomes then the scoring rule expresses the expected loss for the forecaster.

[2]Note that this definition corresponds to the categorical scoring rules in the literature, see e.g. [19]. Scoring rules are usually defined over a more general space, not just over a finite set of atomic events.

An example of a loss function is the *logarithmic loss* $\mathrm{Ls}(\mathbf{v},\alpha) = -\log \mathbf{v}(\alpha)$ first used by Good in 1952, [20]. This loss function is real valued with exception that the loss $+\infty$ is given only when the event claimed to be impossible is realized. The corresponding *logarithmic scoring rule* is $S(\mathbf{w},\mathbf{v}) = -\sum_j w_j \log v_j$. Now consider the following lemma.

**Lemma 1.3.2.** *Let $\mathbf{v},\mathbf{w} \in \mathbb{D}^L$. Then for fixed $\mathbf{w}$ the function $f : \mathbb{D}^L \to \mathbb{R}$ defined by*

$$f(\mathbf{v}) = -\sum_{j=1}^{J} w_j \log v_j$$

*is strictly minimal for $\mathbf{w} = \mathbf{v}$.*[3]

*Proof.* Whenever $w_j \neq 0$ we can always get a lower loss than $+\infty$ with $v_j = 0$. On the other hand, if $w_j = 0$ then we can take the corresponding index $j$ out of the summation forming an index set $I$ and we may suppose that both $v_j \neq 0$ and $w_j \neq 0$ for all $j \in I$. The Hessian matrix for the function $f$ is positive definite

$$\begin{pmatrix} \frac{w_{j_1}}{(v_{j_1})^2} & 0 & \dots & 0 \\ 0 & \frac{w_{j_2}}{(v_{j_2})^2} & \dots & 0 \\ \vdots & & \dots & \vdots \\ 0 & 0 & \dots & \frac{w_{j_n}}{(v_{j_n})^2} \end{pmatrix},$$

where $I = \{j_1, \dots, j_n\}$, and therefore the function has a global minimum at the only critical point determined by $w_j = v_j$ for all $j \in I$. It follows that $v_j = 0$ for all $j \notin I$ since $\sum_{j=1}^{J} v_j = \sum_{j=1}^{J} w_j = 1$. ☐

It follows that for the logarithmic scoring rule $S$ and for a probability function $\mathbf{w}$ played by Nature

$$\arg\min_{\mathbf{v}\in\mathbb{D}^L} S(\mathbf{w},\mathbf{v}) = \mathbf{w},$$

where $\arg\min_{\mathbf{v}\in\mathbb{D}^L} S(\mathbf{w},\mathbf{v})$ denotes that unique point in $\mathbb{D}^L$ at which the function $S(\mathbf{w},\mathbf{v})$ is minimal. In other words if the forecaster wants to minimise his loss (and it is always a loss since $-\log v_j \geq 0$ for all $j$) then he is obliged to choose the probability function which was played by Nature. Hence, there is no reason why the forecaster in his prediction should not act honestly according to his true opinion about Nature, otherwise he will be subject to sure loss.

[3] *Recall that for $v_j = 0$ by definition $-w_j \log v_j = 0$ if $w_j = 0$ and $-w_j \log v_j = +\infty$ otherwise.*

Now we will describe the properties we want scoring rules to satisfy. A scoring rule is *proper* if

$$S(\mathbf{w},\mathbf{w}) \leq S(\mathbf{w},\mathbf{v}), \tag{1.1}$$

for all $\mathbf{w},\mathbf{v} \in \mathbb{D}^L$. It is a *strictly proper* scoring rule if (1.1) holds with equality only if $\mathbf{w} = \mathbf{v}$. A scoring rule is *regular* if $\mathrm{Ls}(\cdot,\alpha)$ is real-valued function for all $\alpha \in \mathrm{At}(L)$ except possibly that $\mathrm{Ls}(\mathbf{v},\alpha) = +\infty$ if $\mathbf{v}(\alpha) = 0$. Obviously the logarithmic scoring rule is strictly proper and regular. Strictly proper scoring rules encourage the forecaster to act honestly and regularity ensures that he is rewarded fairly. Finally, a scoring rule is *local* if the value of the loss $\mathrm{Ls}(\mathbf{w},\alpha)$ depends only on the value of $\mathbf{w}(\alpha)$ and not on the other values — a fairly natural condition. A scoring rule of the form $S(\mathbf{w},\mathbf{v}) = -a \cdot \sum_j w_j \log v_j + b$, where $a, b$ are constants, is the only regular strictly proper local scoring rule, see [35], which indeed justifies a logarithmic scoring rule.

Now assume that the forecaster wants to minimise the worst case loss in the game with Nature. If he knows the set $V \subseteq \mathbb{D}^L$ of distributions that Nature operates with then he is obliged to forecast

$$\arg\min_{\mathbf{v}\in\mathbb{D}^L} \max_{\mathbf{w}\in V} S(\mathbf{w},\mathbf{v}), \tag{1.2}$$

for more details see [21]. In view of the fact that the logarithmic scoring rule is the only strictly proper local scoring rule it seems rational to choose $S$ to be the logarithmic scoring rule.

Now, by [33], theorem 32,[4] we can swap 'min' and 'max' in (1.2):

$$\arg\min_{\mathbf{v}\in\mathbb{D}^L} \max_{\mathbf{w}\in V} S(\mathbf{w},\mathbf{v}) = \arg\max_{\mathbf{w}\in V} \min_{\mathbf{v}\in\mathbb{D}^L} -\sum_{j=1}^{J} w_j \log v_j = \arg\max_{\mathbf{w}\in V} H(\mathbf{w}),$$

where $H$ is the Shannon entropy and $V \subseteq \mathbb{D}^L$ is closed and convex. In particular we can take $V = V^L_{\mathbf{K}}$ for some knowledge base $\mathbf{K}$ and this can be considered as a justification of **ME** in game theory. Given that forecaster is rewarded by a logarithmic scoring rule and he knows the set $V^L_{\mathbf{K}}$ of possible probability functions by which Nature plays, in order to minimise his loss he is obliged to choose $\mathbf{ME}_L(\mathbf{K})$. □

[4]The essential part is due to König's minimax theorem, see [32], theorem 1.3.

There are many other inference processes which have been studied. Among these we mention in detail the $\mathbf{CM}^\infty$-inference process. For any $L$ and any $\mathbf{K} \in CL$

$$\mathbf{CM}^\infty_L(\mathbf{K}) = \arg\max_{\mathbf{w} \in V^L_{\mathbf{K}}} \sum_{j \in \mathrm{Sig}_L(\mathbf{K})} \log w_j,$$

where $\mathrm{Sig}_L(\mathbf{K}) = \{j\colon \exists \mathbf{w} \in V^L_{\mathbf{K}},\ w_j \neq 0\}$. It is not hard to see that by the convexity argument there is always at least one probability function for which the above formula does not give $-\infty$. Also the function $\sum_{j \in \mathrm{Sig}_L(\mathbf{K})} \log w_j$ has a unique maximal point since it is strictly concave on the domain defined by $0 < w_j$, $j \in \mathrm{Sig}_L(\mathbf{K})$.

An alternative characterisation of the $\mathbf{CM}^\infty$-inference process can be given by using the notion of *centre of mass*. Let $\mathbf{K} \in CL_1$ and $L_1 \subseteq L_2$. Define

$$\mathbf{CM}_{L_2}(\mathbf{K})(\alpha) = \frac{\int_{V^{L_2}_{\mathbf{K}}} \mathbf{w}(\alpha) dV}{\int_{V^{L_2}_{\mathbf{K}}} dV},$$

where $\alpha \in \mathrm{At}(L_2)$ and integrals are taken over the relative dimension of $V^{L_2}_{\mathbf{K}}$. Then for all $\varphi \in SL_1$

$$\lim_{\substack{L_1 \subseteq L_2 \\ |L_2| \to \infty}} \mathbf{CM}_{L_2}(\mathbf{K})(\varphi)$$

exists and equals to $\mathbf{CM}^\infty_{L_1}(\mathbf{K})(\varphi)$, see [39].

The $\mathbf{CM}^\infty$-inference process satisfies the principles of Equivalence, Atomic renaming, Obstinacy, Relativisation and Open-mindedness. However, unlike the **ME**-inference process, it does not satisfy the principles of Irrelevant information, Continuity and Independence, c.f. [24] and [39].[5]

Both **ME** and $\mathbf{CM}^\infty$ are a part of so called *Renyi spectrum* of inference processes which we now define. For $1 > r > 0$ we define $\mathbf{REN}^r$ for any $L$ and any $\mathbf{K} \in CL$ by

$$\mathbf{REN}^r_L(\mathbf{K}) = \arg\max_{\mathbf{w} \in V^L_{\mathbf{K}}} \sum_{j=1}^{J} (w_j)^r$$

and for $r > 1$ we define $\mathbf{REN}^r$ for any $L$ and any $\mathbf{K} \in CL$ by

$$\mathbf{REN}^r_L(\mathbf{K}) = \arg\min_{\mathbf{w} \in V^L_{\mathbf{K}}} \sum_{j=1}^{J} (w_j)^r.$$

Since $\sum_{j=1}^{J} (w_j)^r$ is a strictly concave function for $1 > r > 0$ and a strictly convex function for $r > 1$ the above is well defined. **ME** and $\mathbf{CM}^\infty$ are related to Renyi spectrum by the following limit theorems:

[5] Note that reference [39] incorrectly states that $\mathbf{CM}^\infty$ does not satisfy Relativisation.

**Theorem 1.3.3.** *For any $L$ and any $\mathbf{K} \in CL$*

$$\lim_{r \searrow 1} \mathbf{REN}^r_L(\mathbf{K}) = \mathbf{ME}_L(\mathbf{K}) \text{ and } \lim_{r \nearrow 1} \mathbf{REN}^r_L(\mathbf{K}) = \mathbf{ME}_L(\mathbf{K}).$$

**Theorem 1.3.4.** *For any $L$ and any $\mathbf{K} \in CL$*

$$\lim_{r \searrow 0} \mathbf{REN}^r_L(\mathbf{K}) = \mathbf{CM}^{\infty}_L(\mathbf{K}).$$

The first result above is due to Mohamed [37], and the second one is due to Hawes [24].

For a positive $r \neq 1$ the $\mathbf{REN}^r$-inference process satisfies the principles of Equivalence, Atomic Renaming, Continuity and Relativisation. On the other hand it does not satisfy Independence and the Irrelevant information principle and for $r > 1$ not even Open-mindedness. For more details see [24]. We can conclude that none of the other processes in the Renyi spectrum has such appealing properties as the maximum entropy inference process.

## 1.4 Collective reasoning

In the previous section we have described three interesting approaches which, we believe, are also applicable in the multi-expert framework. In this thesis we confine ourselves to investigating the *axiomatic approach*. We will now state several elementary principles applicable to a p-merging operator $\Delta$. It will be implicitly assumed that these principles are required to hold for every $n \geq 1$ and every propositional language $L$.

First, recall that one of our initial assumptions (see section 1.1) is the one that experts are assumed to have equal status. We will now formally state this as a principle which also extends the equivalence principle for inference processes.

**(K2) Equivalence Principle.** If $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ and $\mathbf{F}_1, \ldots, \mathbf{F}_n \in CL$ are such that there exist a permutation $\pi$ of the index set $\{1, \ldots, n\}$ such that $V^L_{\mathbf{K}_i} = V^L_{\mathbf{F}_{\pi(i)}}$ for $1 \leq i \leq n$ and $\Delta$ is a p-merging operator, then $\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \Delta_L(\mathbf{F}_1, \ldots, \mathbf{F}_n)$.

The effect which **(K2)** has is that the order in which the knowledge bases occur when $\Delta$ is applied is immaterial, and therefore we can loosely refer to $\Delta$ as being applied to a multiset of knowledge bases instead of a sequence of such knowledge bases. On the other hand repetitions of knowledge bases *will* in general be significant, so the sequence (or multiset) of knowledge bases cannot be considered as a set; the $\Delta$ we consider will typically share most of the characteristics of what are commonly referred to as majority merging operator, see [30].

**(K3) Atomic Renaming.** Consider a permutation $\sigma$ of the atoms of $L$. By $\sigma(\mathbf{K})$ for $\mathbf{K} \in CL$ we mean a set of constraints $\mathbf{K}$ where the atoms have been renamed according to $\sigma$ and every sentence in $SL$ has been changed accordingly. By $\sigma(\mathbf{w})$ we mean an $L$-probability function such that $\sigma(\mathbf{w})(\alpha) = \mathbf{w}(\sigma(\alpha))$ for all $\alpha \in \mathrm{At}(L)$. Now we say that a p-merging operator $\Delta$ satisfies the atomic renaming principle if for all $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$

$$\Delta_L(\sigma(\mathbf{K}_1), \ldots, \sigma(\mathbf{K}_n)) = \sigma(\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)).$$

This principle is natural counterpart to the principle of atomic renaming for inference processes. It simply says that the result of applying a p-merging operator should not depend on the order in which atoms are listed.

The restricted problem of probabilistic merging when each knowledge base $\mathbf{K}$ determines a single probability function (i.e. $V^L_{\mathbf{K}}$ is a singleton) has been widely studied in decision theory literature as the problem of *Pooling Operators*, see e.g. [16] and [17].

In our presentation, formally, we define a pooling operator **Pool** for each $n \geq 1$ and for each language $L$ as a mapping

$$\mathbf{Pool} : \underbrace{\mathbb{D}^L \times \ldots \times \mathbb{D}^L}_{n} \to \mathbb{D}^L$$

such that the result of applying **Pool** does not depend on the order of probability functions in the argument and for any permutation $\sigma$ of the atoms of $L$

$$\mathbf{Pool}(\sigma(\mathbf{w}^{(1)}), \ldots, \sigma(\mathbf{w}^{(n)})) = \sigma(\mathbf{Pool}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})).$$

Two the most common[6] pooling operators are **LinOp** defined by

$$\mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})(\alpha_j) = \frac{\sum_{i=1}^n w_j^{(i)}}{n},$$

[6]With their weighted variants.

for every atom $\alpha_j \in \mathrm{At}(L)$, corresponding to an arithmetic mean in each coordinate, and **LogOp** defined by

$$\mathbf{LogOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})(\alpha_j) = \frac{\left(\prod_{i=1}^{n} w_j^{(i)}\right)^{\frac{1}{n}}}{\sum_{j=1}^{J}\left(\prod_{i=1}^{n} w_j^{(i)}\right)^{\frac{1}{n}}},$$

for every atom $\alpha_j \in \mathrm{At}(L)$, corresponding to a normalised geometric mean in each coordinate. Notice that $\mathbf{LogOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$ is a probability function if and only if the following holds.

$$\textit{There is some atom } \alpha \textit{ such that for no } i \textit{ is it the case that } \mathbf{w}^{(i)}(\alpha) = 0. \tag{1.3}$$

In relation to expert systems, pooling operators are widely used in economics. For instance see [36] where an application forecasting UK inflation is considered. Extension of pooling operators to our framework of knowledge merging is therefore of a significant interest.

Some axiomatic framework for pooling operators has been also studied (see e.g. [18] for a survey), however the results are perhaps less convincing than the result of Paris and Vencovská for inference processes and we will not rely on them. Instead, we define a proper p-merging operator directly extending the general notions of pooling operators and inference processes.

We define an *Obdurate Merging Operator* **O** by the following two stage process: Given $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$,

1. define $\mathbf{w}^{(i)} = \mathcal{S}_L(\mathbf{K}_i)$, $1 \leq i \leq n$, where $\mathcal{S}$ is an inference process and
2. put $\mathbf{O}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \mathbf{Pool}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$, where **Pool** is a pooling operator which does not depend on the outcome of the first stage[7].

We will shortly see why this operator is called 'obdurate'. Note that the definitions of **Pool** and $\mathcal{S}$ ensure that **O** satisfies principles **(K1)**, **(K2)** and **(K3)** and by the definition obdurate merging operators are single valued.

An obdurate merging operator defined by means of **ME** and **LogOp** is called the *Obdurate Social Entropy Process* (**OSEP**). Note that whenever (1.3) does not hold

[7] This is a non-trivial assumption as shown in section 4.4.

for some $\mathbf{ME}_L(\mathbf{K}_1), \ldots, \mathbf{ME}_L(\mathbf{K}_n)$ we will leave $\mathbf{OSEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ undefined. By only slight abuse of language we still call **OSEP** a p-merging operator. Notice that (1.3) holds whenever the following condition is satisfied.

$$\begin{aligned}&\textit{There is some atom } \alpha \textit{ such that for no } i\\ &\textit{is it the case that for all } \mathbf{w} \in V^L_{\mathbf{K}_i} \;\; \mathbf{w}(\alpha) = 0.\end{aligned} \tag{1.4}$$

This is because **ME** satisfies Open-Mindedness so if $Bel(\alpha) \neq 0$ is consistent with $\mathbf{K} \in CL$ then $\mathbf{ME}_L(\mathbf{K})(\alpha) \neq 0$.

Similarly, the *Obdurate Linear Entropy Process* (**OLEP**) is defined by

$$\mathbf{OLEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)(\alpha_j) = \mathbf{LinOp}(\mathbf{ME}_L(\mathbf{K}_1)(\alpha_j), \ldots, \mathbf{ME}_L(\mathbf{K}_1)(\alpha_j)),$$

for all $\alpha_j \in \mathrm{At}(L)$. **OLEP** has the advantage that it is well defined everywhere.

The notion of an obdurate merging operator seems to be a natural extension of inference processes to our framework indeed, in particular one may argue in favour of the first stage of **OSEP** or **OLEP** on the grounds of **ME**. However, instead, we turn now to some general criticisms of obdurate merging operators. Consider the following

**(K4) Consistency Principle.** For all $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ if $\bigcap_{i=1}^n V^L_{\mathbf{K}_i} \neq \emptyset$ then $\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \subseteq \bigcap_{i=1}^n V^L_{\mathbf{K}_i}$.

**(K4)** can be interpreted as saying that if the knowledge bases of a set of experts are collectively consistent then the merged knowledge base should not consist of anything else than what they agree on.

This principle often fall under the following philosophical criticism. One might imagine a situation when several experts consider a large set of probability functions as admissible while one believe in a single probability function. Although this one is consistent with the beliefs of the rest of the group, one might argue that it is not justified to merge knowledge of the whole group to that single probability function. However such a criticism consists of an additional assumption imposed on rationality of reasoning, namely the lack of trust in a single expert. We argue that once we have confined ourselves only to the narrow problem of merging of knowledge bases this principle is strongly justified.

However, since obdurate merging operators are in fact social inference processes (they always result in a single probability function), the above criticism is hard to accept at all. If we needed to choose only one point then which other one we would choose if not the point of agreement. From this point of view the following theorem is rather striking.

**Theorem 1.4.1.** *Let* $\mathbf{O}$ *be an obdurate merging operator and let* $\mathcal{S}$ *be the inference process used in the first stage of* $\mathbf{O}$*. Assume that for any* $L$ *and any* $\mathbf{K} \in CL$ $\mathcal{S}_L(\mathbf{K})$ *is defined as that unique probability function which maximises some strictly concave function* $f : \mathbb{D}^L \to \mathbb{R}$ *over* $V^L_{\mathbf{K}}$*. (In particular* $\mathbf{ME}$ *and* $\mathbf{CM}^\infty$ *are instances of such an* $\mathcal{S}$*.) Then* $\mathbf{O}$ *does not satisfy the consistency principle* **(K4)***.*

*Proof.* Suppose that $L$ has at least two propositional variables. Let $\mathbf{v} \in \mathbb{D}^L$ be the unique maximiser of $f$ over $\mathbb{D}^L$. Let $\mathbf{w}, \mathbf{u} \in \mathbb{D}^L$ be such that $f(\mathbf{v}) > f(\mathbf{w}) > f(\mathbf{u})$ and $\mathbf{w} = \lambda\mathbf{v} + (1-\lambda)\mathbf{u}$ for some $0 < \lambda < 1$ (in particular $\mathbf{w}$ is a linear combination of $\mathbf{v}$ and $\mathbf{u}$).

Let $\mathbf{a} \in \mathbb{D}^L$ be such that $f(\mathbf{v}) > f(\mathbf{a}) > f(\mathbf{w})$ and $\mathbf{a}$ is not a linear combination of $\mathbf{v}$ and $\mathbf{u}$. Then there is $\mathbf{a}'$ such that $\mathbf{a}' = \lambda\mathbf{a} + (1-\lambda)\mathbf{w}$ for some $0 < \lambda \leq 1$ and $f$ is strictly decreasing along the line from $\mathbf{a}'$ to $\mathbf{w}$. This is because $f$ is strictly concave and $f(\mathbf{a}) > f(\mathbf{w})$. Note that if there was only one propositional variable in $L$ then $\mathbf{a}$ would be always a linear combination of $\mathbf{v}$ and $\mathbf{u}$.

Now we show that $f$ is also strictly decreasing along the line from $\mathbf{a}'$ to $\mathbf{u}$. Assume this is not the case. Then by the same argument as before there is $\mathbf{a}''$ such that $f(\mathbf{a}'') > f(\mathbf{a}')$. Due to the construction the line form $\mathbf{v}$ to $\mathbf{a}''$ intersects the line from $\mathbf{a}'$ to $\mathbf{w}$, let us denote the point of intersection as $\mathbf{r}$. Since $f$ is strictly decreasing along the line from $\mathbf{a}'$ to $\mathbf{w}$ we have that $f(\mathbf{r}) < f(\mathbf{a}') < f(\mathbf{a}'') < f(\mathbf{v})$. This, however, contradicts concavity of $f$. The situation is depicted in figure 1.2.

Now assume that $V^L_{\mathbf{K}_1} = \{\lambda\mathbf{v} + (1-\lambda)\mathbf{w} : \lambda \in [0,1]\}$, $V^L_{\mathbf{K}_2} = \{\lambda\mathbf{a}' + (1-\lambda)\mathbf{w} : \lambda \in [0,1]\}$, $V^L_{\mathbf{F}_1} = \{\lambda\mathbf{v} + (1-\lambda)\mathbf{u} : \lambda \in [0,1]\}$ and $V^L_{\mathbf{F}_2} = \{\lambda\mathbf{a}' + (1-\lambda)\mathbf{u} : \lambda \in [0,1]\}$. Since $\mathbf{v}$ maximises $f$, and along the lines from $\mathbf{a}'$ to $\mathbf{w}$ and from $\mathbf{a}'$ to $\mathbf{u}$ the function $f$ is strictly decreasing we have that

$$\mathbf{O}_L(\mathbf{K}_1, \mathbf{K}_2) = \mathbf{Pool}(\mathbf{v}, \mathbf{a}') = \mathbf{O}_L(\mathbf{F}_1, \mathbf{F}_2), \tag{1.5}$$

where $\mathbf{Pool}$ is a pooling operator used in the second stage of $\mathbf{O}$.

Suppose that $\mathbf{O}$ satisfies **(K4)**. Then $\mathbf{O}_L(\mathbf{K}_1, \mathbf{K}_2) = \mathbf{w}$ and $\mathbf{O}_L(\mathbf{F}_1, \mathbf{F}_2) = \mathbf{u}$ which contradicts (1.5). □

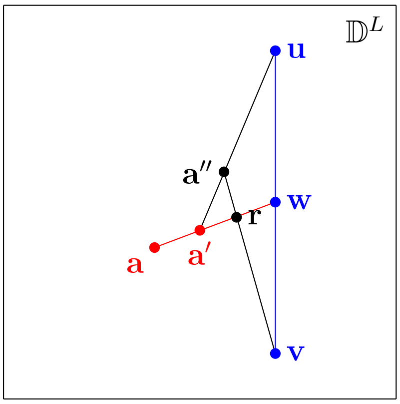


Figure 1.2: The situation in the proof of theorem 1.4.1.

The theorem above reveals a serious flaw within the design of an obdurate merging operator. The first stage does not take into account other experts' opinions and when the second stage is finally performed much of the original information has already been lost. One might say that such a process is obdurate — hence the name.

Note that in [2] the consistency principle is formulated in the following stronger form.

**(K4*) Strong Consistency Principle.** For all $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ if $\bigcap_{i=1}^n V^L_{\mathbf{K}_i} \neq \emptyset$ then $\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \bigcap_{i=1}^n V^L_{\mathbf{K}_i}$

In particular, if there is only one expert with knowledge base $\mathbf{K}$ then this principle just asserts that $\Delta_L(\mathbf{K}) = V^L_{\mathbf{K}}$.

In contrast to the above, the following principle focuses on an extremely weak kind of consistency in the case of an expert possessing zero knowledge.

**Ignorance Principle.** Let $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ and $\mathbf{F} = \emptyset$ (i.e. $V^L_{\mathbf{F}} = \mathbb{D}^L$). Then $\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}) = \Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$.

It says that if we add an expert who knows nothing relevant to the problem then he must not influence the merging procedure.

**Example 1.4.2.** Let $\mathbf{K} \in CL$ be such that $V^L_{\mathbf{K}} = \{\mathbf{v}\}$ is a singleton and $\mathbf{F} = \emptyset$. Assume that $\mathbf{v} \neq \mathbf{ME}_L(\mathbf{F}) = \Big(\underbrace{\frac{1}{J}, \ldots, \frac{1}{J}}_{J}\Big)$. Then obviously **OSEP** and **OLEP** do not satisfy the ignorance principle in this case. □

The ignorance principle is in fact a special case of the following principle due to Wilmers [52].

**(K5) Collegiality Principle.** Let $\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m \in CL$ and $\bigcap_{i=1}^m V_{\mathbf{F}_i}^L \neq \emptyset$. Assume that $\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \subseteq \bigcap_{i=1}^m V_{\mathbf{F}_i}^L$. Then

$$\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m) = \Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

In other words in the case when an expert agrees with the conclusion of a college of experts then including this expert in the decision process should not make any difference to the agreement of the group.

Despite theorem 1.4.1 and the fact that neither **OSEP** nor **OLEP** satisfy the fairly natural ignorance principle they do nevertheless possess some other very appealing properties, as we will see later. In the next chapter, however, we examine how to relate a probability function to a knowledge base in a manner which ultimately allow us to define a class of merging operators which satisfy the consistency principle, the collegiality principle and much more.

# Chapter 2

# Information geometry

## 2.1 Divergences

In this section we will define an associated binary function (a divergence) on the set of probability functions $\mathbb{D}^L$. This function can be interpreted as a measure on ordered pairs of probability functions. As a result a number of geometrical notions arise which we will exploit to define some well behaved probabilistic merging operators.

The concept of a divergence over a manifold is widely used in the differential geometry. For our purposes we define a *divergence* as a function

$$D(\cdot\|\cdot) : \mathbb{D}^L \times \mathbb{D}^L \to \mathbb{R}^*,$$

where for any $\mathbf{w}, \mathbf{v} \in \mathbb{D}^L$

$$D(\mathbf{w}\|\mathbf{v}) \geq 0, \text{ and } D(\mathbf{w}\|\mathbf{v}) = 0 \text{ if and only if } \mathbf{w} = \mathbf{v}. \tag{2.1}$$

Recall that $\mathbb{R}^*$ is the extended real line $\mathbb{R} \cup \{+\infty\}$. A divergence is neither required to be symmetric nor to satisfy the triangular inequality, and therefore, in general, it is not a metric. Due to asymmetry we say that $D(\mathbf{w}\|\mathbf{v})$ is a divergence *from* $\mathbf{v}$ *to* $\mathbf{w}$.

A special type of a divergence which has recently attracted attention in machine learning and plays a major role in optimisation (c.f. [9]) is a *Bregman divergence* ([7]). To define it we first define the useful topological term — a *relative interior*. For any nonempty convex set $C \subseteq \mathbb{R}^J$ the relative interior of $C$ is defined by

$$\mathrm{ri}(C) = \{x \in C\colon\ \forall y \in C\ \exists \lambda > 1(\lambda x + (1-\lambda)y \in C)\}.$$

The relative interior is a version of the concept of interior which is suitable for low dimensional subregions so as to exclude only their outer boundary in a relative dimension. Note that $\mathbf{v} \notin \mathrm{ri}(\mathbb{D}^L)$ only if $v_j = 0$ for some $1 \le j \le J$.

Now consider a mapping $d_f : \mathbb{D}^L \times \mathrm{ri}(\mathbb{D}^L) \to \mathbb{R}$ defined by

$$d_f(\mathbf{w}\|\mathbf{v}) = f(\mathbf{w}) - f(\mathbf{v}) - (\mathbf{w} - \mathbf{v}) \cdot \nabla f(\mathbf{v}),$$

where $f : (\mathbb{R}_0^+)^J \to \mathbb{R}$ is a continuous strictly convex function which is continuously differentiable over $\mathrm{ri}(\mathbb{D}^L)$ where $\mathbb{R}_0^+$ denotes the non-negative real numbers. Note that $\nabla f(\mathbf{v}) = \nabla f(v_1, \ldots, v_J)$ is a gradient of $f$ and $\cdot$ denotes inner (dot) product of two vectors. Therefore

$$(\mathbf{w} - \mathbf{v}) \cdot \nabla f(\mathbf{v}) = \sum_{j=1}^{J} (w_j - v_j) \frac{\partial f(\mathbf{v})}{\partial v_j}. \tag{2.2}$$

To create a divergence from the mapping $d_f$ defined above we need firstly to extend the domain $\mathbb{D}^L \times \mathrm{ri}(\mathbb{D}^L)$ to $\mathbb{D}^L \times \mathbb{D}^L$. This definition will depend on whether $f$ is differentiable over the whole $\mathbb{D}^L$ or not. We will define two notions which depend on this.

For any $\mathbf{v} \in \mathbb{D}^L$ and $\mathbf{w} \in \mathbb{D}^L$ we say that $\mathbf{v}$ *$f$-dominates* $\mathbf{w}$ and write $\mathbf{v} \gg_f \mathbf{w}$ if

1. $f$ is differentiable over $\mathbb{D}^L$ *or*

2. $v_j = 0$ implies $w_j = 0$ for all $1 \le j \le J$.

If the function $f$ is known then we may say for simplicity $\mathbf{v}$ dominates $\mathbf{w}$ ($\mathbf{v} \gg \mathbf{w}$).

Note that the binary relation $\gg$ is reflexive and transitive. To verify transitivity consider $\mathbf{v} \gg_f \mathbf{w} \gg_f \mathbf{u}$. Whenever one of $\mathbf{v} \gg_f \mathbf{w}$ and $\mathbf{w} \gg_f \mathbf{u}$ holds due to condition 1 above, $\mathbf{v} \gg_f \mathbf{u}$ holds for the same reason. If both $\mathbf{v} \gg_f \mathbf{w}$ and $\mathbf{w} \gg_f \mathbf{u}$ hold due to condition 2 above then $v_j = 0$ implies $u_j = 0$ for all $1 \le j \le J$.

A *signature of the gradient* of $f$ at $\mathbf{v}$ is defined as follows.

$$\mathrm{Sig}_f(\mathbf{v}) = \begin{cases} \{1, \ldots, J\}, & f \text{ is differentiable over } \mathbb{D}^L, \\ \{j\colon\ v_j \neq 0\}, & \text{otherwise.} \end{cases}$$

We may omit the index $f$ whenever $f$ is implicitly given.

In addition to the assumption that $f : (\mathbb{R}_0^+)^J \to \mathbb{R}$ is a continuous strictly convex function which is differentiable over $\mathrm{ri}(\mathbb{D}^L)$, suppose that

$$\begin{aligned}&\text{for every } \mathbf{v} \in \mathbb{D}^L \setminus \mathrm{ri}(\mathbb{D}^L) \text{ and every } \mathbf{w} \in \mathbb{D}^L \text{ such that } \mathbf{v} \gg_f \mathbf{w}\\ &\text{there exist a \textit{directional derivative} of } f \text{ at } \mathbf{v} \text{ in the direction to } \mathbf{w},\\ &\text{and is equal to } \sum_{j \in \mathrm{Sig}_f(\mathbf{v})} (w_j - v_j) \frac{\partial f(\mathbf{v})}{\partial v_j}.\end{aligned} \tag{2.3}$$

Note that the existence of directional derivatives above gives that the partial derivative $\frac{\partial f(\mathbf{v})}{\partial v_j}$ exists if $j \in \mathrm{Sig}_f(\mathbf{v})$.

Then we define a *Bregman divergence* as a mapping $D_f : \mathbb{D}^L \times \mathbb{D}^L \to \mathbb{R}^*$ defined by

$$D_f(\mathbf{w}\|\mathbf{v}) = \begin{cases} f(\mathbf{w}) - f(\mathbf{v}) - \sum_{j \in \mathrm{Sig}_f(\mathbf{v})} (w_j - v_j) \frac{\partial f(\mathbf{v})}{\partial v_j}, & \text{if } \mathbf{v} \gg_f \mathbf{w}, \\ +\infty, & \text{otherwise.} \end{cases}$$

Whenever $\mathbf{v} \gg_f \mathbf{w}$ we write $(\mathbf{w} - \mathbf{v}) \cdot \mathbf{v}^*$ as a shorthand of $\sum_{j \in \mathrm{Sig}_f(\mathbf{v})} (w_j - v_j) \frac{\partial f(\mathbf{v})}{\partial v_j}$. To our knowledge the Bregman divergence was so far in the literature only defined over $\mathbb{D}^L \times \mathrm{ri}(\mathbb{D}^L)$ and therefore the above is an extension of the classical definition.

In the forthcoming lemma 2.1.2 we prove that the mapping above is really a divergence. In order to prove this we need the following theorem, see e.g. p. 69 of [6].

**Theorem 2.1.1** (First Convexity Condition)**.** *Suppose that $U$ is convex in $\mathbb{R}^J$ and $f : U \to \mathbb{R}$ is differentiable (i.e., its gradient $\nabla f$ exists at each point in $U$ ). Then $f$ is convex if and only if*

$$f(\mathbf{w}) \geq f(\mathbf{v}) + (\mathbf{w} - \mathbf{v}) \cdot \nabla f(\mathbf{v})$$

*holds for all $\mathbf{w}, \mathbf{v} \in U$.*

**Lemma 2.1.2.** *The function $D_f$ defined above is a divergence.*

*Proof.* Since $f$ is a convex function by the first-order convexity condition (theorem 2.1.1) we have that $f(\mathbf{w}) \geq f(\mathbf{v}) + (\mathbf{w} - \mathbf{v}) \cdot \nabla f(\mathbf{v})$ whenever $\mathbf{v} \in \mathrm{ri}(\mathbb{D}^L)$ and $\mathbf{w} \in \mathbb{D}^L$. Notice that $f(\mathbf{v}) + (\mathbf{w} - \mathbf{v}) \cdot \nabla f(\mathbf{v})$ for the variable $\mathbf{w} \in \mathbb{D}^L$ is the affine tangent space to $f$ at the point $\mathbf{v}$.

For a point $\mathbf{v}$ on the boundary $\mathbb{D}^L \setminus \mathrm{ri}(\mathbb{D}^L)$, assume $\mathbf{v} \gg_f \mathbf{w}$. By convexity of $f$ for any $\lambda \in [0, 1]$ $\lambda f(\mathbf{w}) + (1 - \lambda) f(\mathbf{v}) \geq f(\mathbf{v} + \lambda(\mathbf{w} - \mathbf{v}))$. Then

$$f(\mathbf{w}) \geq f(\mathbf{v}) + \frac{f(\mathbf{v} + \lambda(\mathbf{w} - \mathbf{v})) - f(\mathbf{v})}{\lambda}.$$

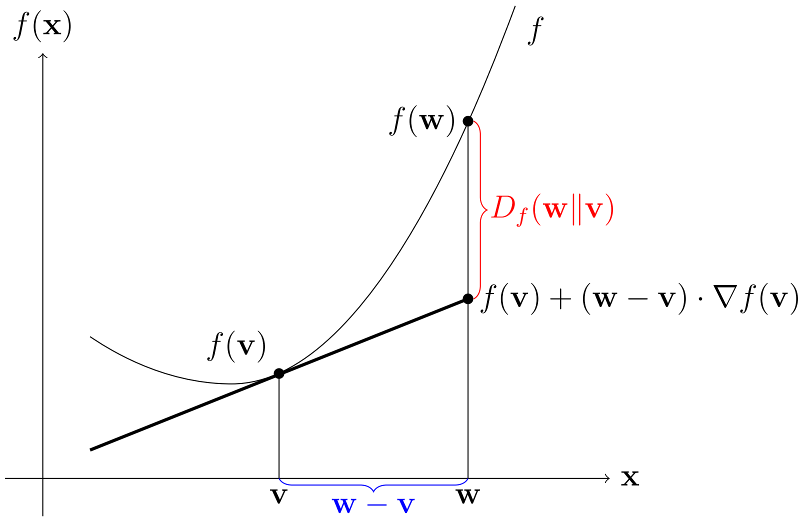


Figure 2.1: A Bregman divergence.

Since, by assumption (2.3), the directional derivative $\lim_{\lambda\to 0}\frac{f(\mathbf{v}+\lambda(\mathbf{w}-\mathbf{v}))-f(\mathbf{v})}{\lambda}$ exists, we have that $f(\mathbf{w}) \geq f(\mathbf{v}) + \sum_{j\in\mathrm{Sig}_f(\mathbf{v})}(w_j - v_j)\frac{\partial f(\mathbf{v})}{\partial v_j}$. Note that the existence of only partial derivatives $\frac{\partial f(\mathbf{v})}{\partial v_j}$, $j \in \mathrm{Sig}_f(\mathbf{v})$, is not sufficient to establish the claim.

On the other hand if $\mathbf{v} \not\gg_f \mathbf{w}$ then $D_f(\mathbf{w}\|\mathbf{v}) = +\infty > 0$. In any case $D_f(\mathbf{w}\|\mathbf{v}) \geq 0$.

Since $f$ is a strictly convex function, if $\mathbf{v} \gg_f \mathbf{w}$ then

$$f(\mathbf{w}) = f(\mathbf{v}) + \sum_{j\in\mathrm{Sig}_f(\mathbf{v})}(w_j - v_j)\frac{\partial f(\mathbf{v})}{\partial v_j}$$

is possible only if $w_j = v_j$ for all $j \in \mathrm{Sig}_f(\mathbf{v})$. However, since $\mathbf{v} \gg_f \mathbf{w}$ we have that either $\mathrm{Sig}_f(\mathbf{v}) = \{1,\ldots,J\}$, or $v_j = 0$ implies $w_j = 0$. In any case $\mathbf{w} = \mathbf{v}$, where in the latter case we have used the fact that $\sum_{j\in\mathrm{Sig}_f(\mathbf{v})} v_j = 1$.

Consequently $D_f$ satisfies (2.1). □

Notice that for fixed $f$ a Bregman divergence $D_f$ can be defined for any language $L$. Therefore for a particular function $f$ by $D_f$ we will always mean the class of the divergences, one divergence for each language $L$. Figure 2.1 depicts a geometrical interpretation of a Bregman divergence when $\mathbf{v} \in \mathrm{ri}(\mathbb{D}^L)$.

**Lemma 2.1.3.** *Let $g : \mathbb{R}_0^+ \to \mathbb{R}$ be a continuous strictly convex function which is differentiable over $(0,1]$. Let $f : \mathbb{D}^L \to \mathbb{R}$ be defined by $f(\mathbf{v}) = \sum_{j=1}^J g(v_j)$. Then $f$ satisfies condition (2.3).*

*Proof.* First notice that $\mathbf{v} \in \mathbb{D}^L \setminus \mathrm{ri}(\mathbb{D}^L)$ only if $v_j = 0$ for some $1 \leq j \leq J$. Now let $\mathbf{w} \in \mathbb{D}^L$ be such that $\mathbf{v} \gg_f \mathbf{w}$. Then the directional derivative of $f$ at $\mathbf{v}$ in direction to $\mathbf{w}$ exists since it can be expressed only by derivations of $g$ at non-zero points as

$$\sum_{j \in \mathrm{Sig}_f(\mathbf{v})} (w_j - v_j) \frac{\partial g(v_j)}{\partial v_j}.$$

The above is well-defined since, by assumption, $\frac{\partial g(v_j)}{\partial v_j}$ exists for $v_j \neq 0$. □

The following are examples of a Bregman divergence.

**Example 2.1.4** (Squared Euclidean Distance)**.** For any $J \geq 2$ let $f(\mathbf{x}) = \sum_{j=1}^{J} (x_j)^2$. This function is differentiable over $(\mathbb{R}_0^+)^J$ and therefore any $\mathbf{v} \in \mathbb{D}^L$ $f$-dominates any $\mathbf{w} \in \mathbb{D}^L$. Also for all $\mathbf{v} \in \mathbb{D}^L$ $\mathrm{Sig}_f(\mathbf{v}) = \{1, \dots, J\}$. Therefore we can define the divergence $D_f$ at once by

$$D_f(\mathbf{w} \| \mathbf{v}) = \sum_{j=1}^{J} (w_j - v_j)^2.$$

We will denote this divergence by E2 and, exceptionally, this divergence is symmetric. □

**Example 2.1.5** (Kullback-Leibler Divergence)**.** For any $J \geq 2$ let $f(\mathbf{x}) = \sum_{j=1}^{J} x_j \log x_j$ where by definition $x_j \log x_j = 0$ for $x_j = 0$. This function is differentiable only over $(\mathbb{R}^+)^J$ but, by lemma 2.1.3, it satisfies condition (2.3). Therefore $\mathbf{v} \in \mathbb{D}^L$ $f$-dominates $\mathbf{w} \in \mathbb{D}^L$ only if $v_j = 0$ implies $w_j = 0$. $\mathrm{Sig}(\mathbf{v}) = \{j \colon v_j \neq 0\}$. The divergence $D_f$ is then defined by

$$D_f(\mathbf{w} \| \mathbf{v}) = \begin{cases} \sum_{j \in \mathrm{Sig}(\mathbf{v})} w_j \log \frac{w_j}{v_j}, & \text{if } \mathbf{v} \gg \mathbf{w}, \\ +\infty, & \text{otherwise.} \end{cases}$$

This well-known divergence will be denoted by KL.

Note that $\lim_{\epsilon \to 0} v_j \log \frac{v_j}{\epsilon} = +\infty$ for $v_j \neq 0$ and $\lim_{\epsilon \to 0} \epsilon \log \frac{\epsilon}{\delta} = 0$ for any $\delta > 0$. However $\mathrm{KL}(\cdot \| \cdot)$ is not a continuous function. To see that consider the sequence $\epsilon \log \frac{\epsilon}{\delta}$ as $\epsilon \to 0$ and $\delta \to 0$.

The fact that (2.1) holds for the Kullback-Leibler divergence is known as the *Gibbs inequality.* □

**Example 2.1.6** (Renyi-B Divergence for $2 \geq r > 1$)**.** For any $J \geq 2$ and $2 \geq r > 1$ let $f(\mathbf{x}) = \sum_{j=1}^{J}(x_j)^r$. This function is differentiable over $(\mathbb{R}_0^+)^J$ and therefore any $\mathbf{v} \in \mathbb{D}^L$ $f$-dominates any $\mathbf{w} \in \mathbb{D}^L$. Also for all $\mathbf{v} \in \mathbb{D}^L$ $\mathrm{Sig}_f(\mathbf{v}) = \{1, \ldots, J\}$. Therefore we can define the divergence $D_f$ at once by

$$D_f(\mathbf{w}\|\mathbf{v}) = \sum_{j=1}^{J}[(w_j)^r - (v_j)^r - r(w_j - v_j)(v_j)^{r-1}].$$

We will denote this divergence by $D_r$. In particular $D_2 = \mathrm{E2}$. We note that in the literature the Renyi divergence is defined differently, see [43], and the geometry of the Renyi divergence has been studied for example in [4]. That is the reason why we call the above divergence Renyi-B (where 'B' stands for 'Bregman'). $\square$

Bregman divergences are closely related to scoring rules which we have explained in section 1.3. If a scoring rule $S$ is regular and strictly proper then $D(\mathbf{w}\|\mathbf{v}) = S(\mathbf{w}, \mathbf{v}) - S(\mathbf{w}, \mathbf{w})$ is an *associated divergence* from $\mathbf{v} \in \mathbb{D}^L$ to $\mathbf{w} \in \mathbb{D}^L$. If $S$ is regular and proper it follows that $D(\mathbf{w}\|\mathbf{v}) \geq 0$. If moreover $S$ is strictly proper then $D(\mathbf{w}\|\mathbf{v})$ is strictly positive, unless $\mathbf{w} = \mathbf{v}$.

The following theorem follows from a more general theorem by Gneiting and Raftery in [19], theorem 2, (which is due in essence to McCarthy and Savage [45]).

**Theorem 2.1.7.** *For every Bregman divergence $D_f$ there is a regular strictly proper scoring rule $S$ such that $D_f(\mathbf{w}\|\mathbf{v}) = S(\mathbf{w}, \mathbf{v}) - S(\mathbf{w}, \mathbf{w})$.*

*Proof.* Let $|\mathrm{At}(L)| = J$. Assume that $D_f$ is given. Recall that $f : (\mathbb{R}_0^+)^J \to \mathbb{R}$ from the definition of $D_f$ is a continuous strictly convex function which is differentiable over $\mathrm{ri}(\mathbb{D}^L)$ and it satisfies condition (2.3). Define the loss function by

$$\mathrm{Ls}(\mathbf{v}, \alpha) = \begin{cases} \sum_{j \in \mathrm{Sig}_f(\mathbf{v})}(v_j \cdot \frac{\partial f(\mathbf{v})}{\partial v_j}) - \frac{\partial f(\mathbf{v})}{\partial v_k} - f(\mathbf{v}), & \text{if } k \in \mathrm{Sig}_f(\mathbf{v}), \\ +\infty, & \text{otherwise}, \end{cases} \tag{2.4}$$

where $v_k = \mathbf{v}(\alpha)$. The corresponding scoring rule is

$$S(\mathbf{w}, \mathbf{v}) = \sum_{\alpha \in \mathrm{At}(L)} \mathbf{w}(\alpha) \, \mathrm{Ls}(\mathbf{v}, \alpha),$$

where we put by definition $\mathbf{w}(\alpha) \, \mathrm{Ls}(\mathbf{v}, \alpha) = 0$ whenever $\mathbf{w}(\alpha) = 0$.

Now we show that $D_f(\mathbf{w}\|\mathbf{v}) = S(\mathbf{w},\mathbf{v}) - S(\mathbf{w},\mathbf{w})$. If $\mathbf{v} \not\gg_f \mathbf{w}$ then $S(\mathbf{w},\mathbf{v}) = +\infty$ and $D_f(\mathbf{w}\|\mathbf{v}) = +\infty$, hence they are equal. Now assume that $\mathbf{v} \gg_f \mathbf{w}$. Note that

$$S(\mathbf{w},\mathbf{v}) = \sum_{j\in \mathrm{Sig}_f(\mathbf{v})} \Big[ - w_j \frac{\partial f(\mathbf{v})}{\partial v_j} - w_j f(\mathbf{v}) + w_j \sum_{j\in \mathrm{Sig}_f(\mathbf{v})} v_j \frac{\partial f(\mathbf{v})}{\partial v_j} \Big] =$$

$$= - \sum_{j\in \mathrm{Sig}_f(\mathbf{v})} w_j \frac{\partial f(\mathbf{v})}{\partial v_j} - f(\mathbf{v}) + \sum_{j\in \mathrm{Sig}_f(\mathbf{v})} v_j \frac{\partial f(\mathbf{v})}{\partial v_j}.$$

Therefore

$$D_f(\mathbf{w}\|\mathbf{v}) = S(\mathbf{w},\mathbf{v}) - S(\mathbf{w},\mathbf{w}) =$$

$$= - \sum_{j\in \mathrm{Sig}_f(\mathbf{v})} w_j \frac{\partial f(\mathbf{v})}{\partial v_j} - f(\mathbf{v}) + \sum_{j\in \mathrm{Sig}_f(\mathbf{v})} v_j \frac{\partial f(\mathbf{v})}{\partial v_j} +$$

$$+ \sum_{j\in \mathrm{Sig}_f(\mathbf{w})} w_j \frac{\partial f(\mathbf{w})}{\partial w_j} + f(\mathbf{w}) - \sum_{j\in \mathrm{Sig}_f(\mathbf{w})} w_j \frac{\partial f(\mathbf{w})}{\partial w_j} =$$

$$= f(\mathbf{w}) - f(\mathbf{v}) - \sum_{j\in \mathrm{Sig}_f(\mathbf{v})} (w_j - v_j) \frac{\partial f(\mathbf{v})}{\partial v_j}.$$

Moreover from the the above it is obvious that the scoring rule $S(\mathbf{w},\mathbf{v})$ is regular and strictly proper which concludes the proof. □

It seems therefore possible to argue in favour of investigating the Bregman divergence on the grounds of game theory, which, as we have seen before, also offers a justification for the **ME**-inference process.

**Lemma 2.1.8.** *For given $\mathbf{v} \in \mathbb{D}^L$ a Bregman divergence $D_f(\mathbf{w}\|\mathbf{v})$ is a strictly convex function in the first argument over the domain specified by $\mathbf{v} \gg_f \mathbf{w}$.*

*Proof.* Consider $D_f(\mathbf{w}\|\mathbf{v}) = f(\mathbf{w}) - f(\mathbf{v}) - \sum_{j\in \mathrm{Sig}(\mathbf{v})}(w_j - v_j)\frac{\partial f(\mathbf{v})}{\partial v_j}$ where $\mathbf{v}$ is a constant. $f(\mathbf{v})$ is therefore constant as well and $-\sum_{j\in \mathrm{Sig}(\mathbf{v})}(w_j - v_j)\frac{\partial f(\mathbf{v})}{\partial v_j}$ is just a linear term. Since (strict) convexity is not affected by adding a linear term the lemma follows by the strict convexity of the function $f$. □

Note that $D_f(\cdot\|\cdot)$ is not necessarily convex in its second argument as the following example demonstrates.

**Example 2.1.9.** Let $f(\mathbf{x}) = \sum_{j=1}^4 (x_j)^3$ be defined on $(\mathbb{R}_0^+)^4$ so that

$$D_f(\mathbf{w}\|\mathbf{v}) = \sum_{j=1}^4 ((w_j)^3 - (v_j)^3 - 3(w_j - v_j)(v_j)^2) = \sum_{j=1}^4 (2(v_j)^3 - 3w_j(v_j)^2 + (w_j)^3).$$

For instance if $\mathbf{w} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ then clearly the function $D_f(\mathbf{w}\|\mathbf{v})$ is not convex since the Hessian matrix

$$\begin{pmatrix} 12v_1 - 6w_1 & 0 & 0 & 0 \\ 0 & 12v_2 - 6w_2 & 0 & 0 \\ 0 & 0 & 12v_3 - 6w_3 & 0 \\ 0 & 0 & 0 & 12v_4 - 6w_4 \end{pmatrix}$$

is not positive-semidefinite. $\square$

Owing to lemma 2.1.8 if $\mathbf{v} \in \mathbb{D}^L$ is given and $V \subseteq \mathbb{D}^L$ is a *closed* convex set such that there is at least one probability function which $\mathbf{v}$ $f$-dominates then we can define the $D_f$*-projection* of $\mathbf{v}$ to $V$. It is that unique point $\mathbf{w} \in V$ which minimises $D_f(\mathbf{w}\|\mathbf{v})$ subject only to $\mathbf{w} \in V$.

The existence of a $D_f$-projection is heavily used in expert systems which map probability functions to some linearly constrained sets or spaces generated by marginal probability functions, see [9] and [48] respectively. Similar idea of projecting information that we cannot recognize to a known (training) set is used in artificial neural networks which in their most trivial version are just linear projections. We study such use of projections further in section 3.2.

**Lemma 2.1.10.** *For any $L$-probability functions $\mathbf{v} \gg_f \mathbf{w} \gg_f \mathbf{a}$*

$$D_f(\mathbf{a}\|\mathbf{v}) - D_f(\mathbf{w}\|\mathbf{v}) - D_f(\mathbf{a}\|\mathbf{w}) = (\mathbf{a} - \mathbf{w}) \cdot (\mathbf{w}^* - \mathbf{v}^*),$$

*where, with only a slight abuse of notation, $(\mathbf{a} - \mathbf{w}) \cdot (\mathbf{w}^* - \mathbf{v}^*)$ denotes $\sum_{j \in \mathrm{Sig}_f(\mathbf{w})}(a_j - w_j)\frac{\partial f(\mathbf{w})}{\partial w_j} - \sum_{j \in \mathrm{Sig}_f(\mathbf{v})}(a_j - w_j)\frac{\partial f(\mathbf{v})}{\partial v_j}$.*

*Proof.* Immediately by

$$f(\mathbf{a}) - f(\mathbf{v}) - (\mathbf{a} - \mathbf{v}) \cdot \mathbf{v}^* - [f(\mathbf{w}) - f(\mathbf{v}) - (\mathbf{w} - \mathbf{v}) \cdot \mathbf{v}^*] -$$

$$-[f(\mathbf{a}) - f(\mathbf{w}) - (\mathbf{a} - \mathbf{w}) \cdot \mathbf{w}^*] = (\mathbf{a} - \mathbf{w}) \cdot \mathbf{w}^* - (\mathbf{a} - \mathbf{w}) \cdot \mathbf{v}^*.$$

$\square$

The following theorem was proved for the Kullback-Leibler divergence first by Csiszár in [11]. The following proof is a modification of the proof given in [3] for a Bregman divergence.

**Theorem 2.1.11** (Extended Pythagorean Theorem)**.** *Let $D_f$ be a Bregman divergence. Let $\mathbf{w}$ be the $D_f$-projection of $\mathbf{v} \in \mathbb{D}^L$ to a closed convex set $W \subseteq \mathbb{D}^L$. Let $\mathbf{a} \in W$ be such that $\mathbf{v} \gg_f \mathbf{w} \gg_f \mathbf{a}$. Then*

$$D_f(\mathbf{a}\|\mathbf{w}) + D_f(\mathbf{w}\|\mathbf{v}) \leq D_f(\mathbf{a}\|\mathbf{v}).$$

*If, in particular, $W$ is a line segment in $\mathbb{D}^L$ such that $\mathbf{w} \in \mathrm{ri}(W)$ then*

$$D_f(\mathbf{a}\|\mathbf{w}) + D_f(\mathbf{w}\|\mathbf{v}) = D_f(\mathbf{a}\|\mathbf{v}).$$

*Proof.* First of all notice that by lemma 2.1.10 we have that

$$D_f(\mathbf{a}\|\mathbf{v}) - D_f(\mathbf{w}\|\mathbf{v}) - D_f(\mathbf{a}\|\mathbf{w}) = (\mathbf{a} - \mathbf{w}) \cdot (\mathbf{w}^* - \mathbf{v}^*)$$

and the values of all the divergences above are real numbers. Let $\mathbf{w}_\lambda = \lambda\mathbf{a} + (1-\lambda)\mathbf{w}$, for $\lambda \in [0,1]$. Clearly $D_f(\mathbf{w}_\lambda\|\mathbf{v}) \in \mathbb{R}$. Consider

$$\frac{\partial}{\partial\lambda} D_f(\mathbf{w}_\lambda\|\mathbf{v})|_{\lambda=0} = \frac{\partial}{\partial\lambda}(f(\lambda\mathbf{a} + (1-\lambda)\mathbf{w}) - f(\mathbf{v}) - (\lambda\mathbf{a} + (1-\lambda)\mathbf{w} - \mathbf{v}) \cdot \mathbf{v}^*)|_{\lambda=0} =$$

$$= (\mathbf{a} - \mathbf{w}) \cdot \mathbf{w}^* - (\mathbf{a} - \mathbf{w}) \cdot \mathbf{v}^* = (\mathbf{a} - \mathbf{w}) \cdot (\mathbf{w}^* - \mathbf{v}^*).$$

Therefore if $(\mathbf{a} - \mathbf{w}) \cdot (\mathbf{w}^* - \mathbf{v}^*) < 0$ there is $\mathbf{w}_\lambda$ such that $D_f(\mathbf{w}_\lambda\|\mathbf{v}) < D_f(\mathbf{w}\|\mathbf{v})$ for some $\lambda > 0$. This contradicts that $\mathbf{w}$ is the $D_f$-projection of $\mathbf{v}$ to $W$ and the first part of the theorem follows.

If moreover $W$ is a line segment in $\mathbb{D}^L$ such that $\mathbf{w} \in \mathrm{ri}(W)$ then for any $\mathbf{a} \in W$ there is $\lambda < 0$ such that $\mathbf{w}_\lambda \in W$. Now if $(\mathbf{a} - \mathbf{w}) \cdot (\mathbf{w}^* - \mathbf{v}^*) > 0$ then there is $\lambda \leq \epsilon < 0$ such that $\mathbf{w}_\epsilon \in W$ and $D_f(\mathbf{w}_\epsilon\|\mathbf{v}) < D_f(\mathbf{w}\|\mathbf{v})$. This contradicts that $\mathbf{w}$ is the $D_f$-projection of $\mathbf{v}$ to $W$ and the second part of the theorem follows. □

Notice that the squared Euclidean distance has a special role among all other Bregman divergences. It is symmetric and it interprets the extended Pythagorean theorem 'classically' as the relation of the sizes of the squares constructed on the sides of a triangle.

The following theorem in the version for the Kullback-Leibler divergence is folklore in information theory, see [10].

**Theorem 2.1.12** (Parallelogram Theorem)**.** *Let $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}, \mathbf{v} \in \mathbb{D}^L$ be such that $\mathbf{v} \gg_f \mathbf{w}^{(i)}$ for all $1 \leq i \leq n$ and $D_f$ be a Bregman divergence. Then*

$$\sum_{i=1}^{n} D_f(\mathbf{w}^{(i)}\|\mathbf{v}) = \sum_{i=1}^{n} D_f(\mathbf{w}^{(i)}\| \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}))+$$

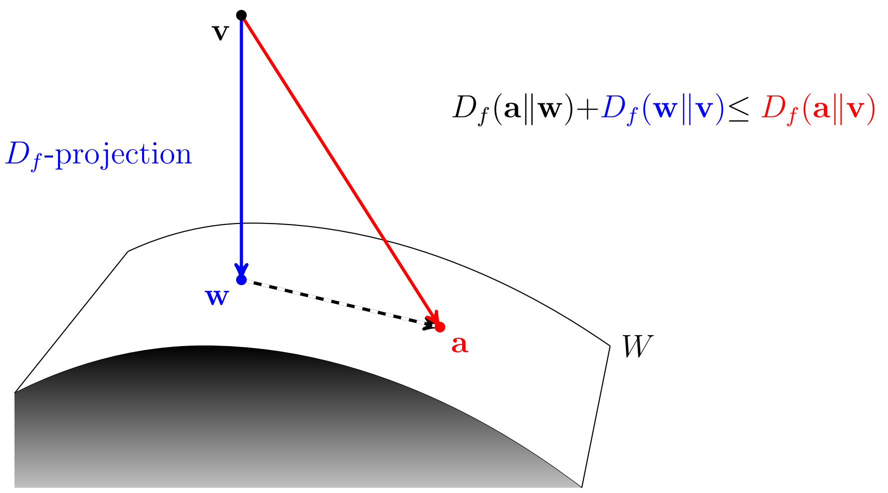


Figure 2.2: The extended Pythagorean theorem.

$$+n\cdot D_f(\mathbf{LinOp}(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)})\|\mathbf{v}).$$

*Proof.* Let $\mathbf{w}=\mathbf{LinOp}(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)})$. The equality is easy to observe by

$$\sum_{i=1}^{n}\left[f(\mathbf{w}^{(i)})-f(\mathbf{v})-\sum_{j\in\mathrm{Sig}_f(\mathbf{v})}(w_j^{(i)}-v_j)\frac{\partial f(\mathbf{v})}{\partial v_j}\right]=$$

$$=\sum_{i=1}^{n}\left[f(\mathbf{w}^{(i)})-f(\mathbf{w})-(\mathbf{w}^{(i)}-\mathbf{w})\cdot\mathbf{w}^*\right]+$$

$$+n\cdot\left[f(\mathbf{w})-f(\mathbf{v})-\sum_{j\in\mathrm{Sig}_f(\mathbf{v})}(w_j-v_j)\frac{\partial f(\mathbf{v})}{\partial v_j}\right]$$

since $\sum_{i=1}^{n}(\mathbf{w}^{(i)}-\mathbf{w})\cdot\mathbf{w}^*=0$. □

Since a Bregman divergence is not necessary convex in its second argument the following result might be a bit surprising. This theorem formulated for random variables was proved in [5].

**Theorem 2.1.13.** *Let $\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)}$ be given L-probability functions and $D_f$ be a Bregman divergence. Then the function*

$$\sum_{i=1}^{n}D_f(\mathbf{w}^{(i)}\|\mathbf{v}),$$

*in the domain given by $\mathbf{v}\in\mathbb{D}^L$, $\mathbf{v}\gg_f\mathbf{w}^{(i)}$ for all $1\leq i\leq n$, is strictly minimal for*

$$\mathbf{v}=\mathbf{LinOp}(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)}).$$

*Proof.* By the parallelogram theorem the minimality of $\sum_{i=1}^{n} D_f(\mathbf{w}^{(i)}\|\mathbf{v})$ for fixed $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ is equivalent to the minimality of $D_f(\mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})\|\mathbf{v})$. Since $D_f(\mathbf{w}\|\mathbf{v}) = 0$ only if $\mathbf{w} = \mathbf{v}$ and otherwise it is positive, the unique minimum of the considered function is at the point $\mathbf{v} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$. □

Let $D$ be a divergence. We say that $D(\cdot\|\cdot)$ is a *convex function* over a domain $V \subseteq \mathbb{D}^L$ if for all $\lambda \in [0, 1]$ and all $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \mathbf{v}^{(1)}, \mathbf{v}^{(2)} \in V$

$$\lambda D(\mathbf{w}^{(1)}\|\mathbf{v}^{(1)}) + (1-\lambda) D(\mathbf{w}^{(2)}\|\mathbf{v}^{(2)}) \geq D(\lambda \mathbf{w}^{(1)} + (1-\lambda)\mathbf{w}^{(2)}\|\lambda \mathbf{v}^{(1)} + (1-\lambda)\mathbf{v}^{(2)}).$$

Note that if $D(\cdot\|\cdot)$ is a convex function then $D(\cdot\|\cdot)$ is a convex function also in each argument separately.

Note that the domain $\{(\mathbf{w}, \mathbf{v}) \in \mathbb{D}^L \times \mathbb{D}^L \colon \mathbf{v} \gg_f \mathbf{w}\}$ is convex and nonempty. In the rest of this section we will study Bregman divergences which are convex over $\{(\mathbf{w}, \mathbf{v}) \in \mathbb{D}^L \times \mathbb{D}^L \colon \mathbf{v} \gg_f \mathbf{w}\}$.

**Example 2.1.14.** The squared Euclidean distance

$$\mathrm{E2}(\mathbf{w}\|\mathbf{v}) = \sum_{j=1}^{J} (w_j - v_j)^2$$

is obviously a convex function over the domain $\mathbb{D}^L \times \mathbb{D}^L$ and hence it is a convex Bregman divergence. □

**Example 2.1.15.** Let $f(\mathbf{x}) = \sum_{j=1}^{J} x_j \log x_j$. The Kullback-Leibler divergence

$$\mathrm{KL}(\mathbf{w}\|\mathbf{v}) = \begin{cases} \sum_{j \in \mathrm{Sig}(\mathbf{v})} w_j \log \frac{w_j}{v_j}, & \text{if } \mathbf{v} \gg \mathbf{w}, \\ +\infty, & \text{otherwise} \end{cases}$$

is a convex function over the domain $\{(\mathbf{w}, \mathbf{v}) \in \mathbb{D}^L \times \mathbb{D}^L \colon \mathbf{v} \gg \mathbf{w}\}$ and hence it is a convex Bregman divergence. For this it is sufficient to prove the following lemma. □

**Lemma 2.1.16.** $\mathrm{KL}(\mathbf{x}\|\mathbf{y})$ *is a convex function over the domain* $\{(\mathbf{x}, \mathbf{y}) \in \mathbb{D}^L \times \mathbb{D}^L \colon y_j = 0$ *implies* $x_j = 0\}$. *Note that over this domain* $\mathrm{KL}(\mathbf{x}\|\mathbf{y})$ *is defined by*

$$\sum_{j \in \mathrm{Sig}(\mathbf{y})} x_j \log \frac{x_j}{y_j}.$$

*Proof.* A differentiation for all $j \in \mathrm{Sig}(\mathbf{y})$ gives

$$\frac{\partial\, \mathrm{KL}(\mathbf{x}\|\mathbf{y})}{\partial x_j} = \log \frac{x_j}{y_j} + 1,$$

$$\frac{\partial\, \mathrm{KL}(\mathbf{x}\|\mathbf{y})}{\partial y_j} = -\frac{x_j}{y_j},$$

$$\frac{\partial^2\, \mathrm{KL}(\mathbf{x}\|\mathbf{y})}{\partial x_j \partial x_j} = \frac{1}{x_j},$$

$$\frac{\partial^2\, \mathrm{KL}(\mathbf{x}\|\mathbf{y})}{\partial x_j \partial y_j} = \frac{\partial^2\, \mathrm{KL}(\mathbf{x}\|\mathbf{y})}{\partial y_j \partial x_j} = -\frac{1}{y_j},$$

$$\frac{\partial^2\, \mathrm{KL}(\mathbf{x}\|\mathbf{y})}{\partial y_j \partial y_j} = \frac{x_j}{(y_j)^2}.$$

Therefore the Hessian matrix is block-diagonal $2 \cdot J \times 2 \cdot J$ matrix

$$\mathbf{H} = \begin{pmatrix} \mathbf{B_1} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{B_2} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{B_3} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where each block is $2 \times 2$ matrix

$$\mathbf{B_j} = \begin{pmatrix} \frac{1}{x_j} & -\frac{1}{y_j} \\ -\frac{1}{y_j} & \frac{x_j}{(y_j)^2} \end{pmatrix},$$

if $j \in \mathrm{Sig}(\mathbf{y})$ and

$$\mathbf{B_j} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

otherwise.

A continuously differentiable function is convex if its Hessian matrix is positive-semidefinite. A matrix $\mathbf{H}$ is positive-semidefinite if $\mathbf{zHz}^T \geq 0$ for all $\mathbf{z} \in \mathbb{R}^{2 \cdot J}$. It is easy to check that it is sufficient to prove that the matrix $\mathbf{B_j}$ is positive-semidefinite for all $j \in \mathrm{Sig}(\mathbf{y})$:

$$\mathbf{zB_jz}^T = \frac{1}{x_j} z_1^2 - \frac{2}{y_j} z_1 z_2 + \frac{x_j}{(y_j)^2} z_2^2 = \left( \frac{1}{\sqrt{x_j}} z_1 - \frac{\sqrt{x_j}}{y_j} z_2 \right)^2 \geq 0.$$

Above expression can attain 0 for a positive vector $\mathbf{z}$ and moreover some of $\mathbf{B_j}$-s might be zero matrices. Therefore the function KL is not *strictly* convex over the specified domain. □

**Example 2.1.17.** For $2 \geq r > 1$ let $f(\mathbf{x}) = \sum_{j=1}^{J} (x_j)^r$. The Renyi-B divergence

$$D_r(\mathbf{w}\|\mathbf{v}) = \sum_{j=1}^{J} [(w_j)^r - (v_j)^r - r(w_j - v_j)(v_j)^{r-1}]$$

is a convex function over the domain $\{(\mathbf{w}, \mathbf{v}) \in \mathbb{D}^L \times \mathbb{D}^L\}$ and hence it is a convex Bregman divergence. For this it is sufficient to prove the following lemma. $\square$

**Lemma 2.1.18.** *For* $2 \geq r > 1$ $\sum_{j=1}^{J} [(x_j)^r - (y_j)^r - r(x_j - y_j)(y_j)^{r-1}]$ *is a convex function over the domain* $\{(\mathbf{x}, \mathbf{y}) \in [0,1]^J \times [0,1]^J\}$.

*Proof.* A differentiation for all $1 \leq j \leq J$ gives

$$\frac{\partial D_r(\mathbf{x}\|\mathbf{y})}{\partial x_j} = r(x_j)^{r-1} - r(y_j)^{r-1},$$

$$\frac{\partial D_r(\mathbf{x}\|\mathbf{y})}{\partial y_j} = -r(r-1)(x_j - y_j)(y_j)^{r-2},$$

$$\frac{\partial^2 D_r(\mathbf{x}\|\mathbf{y})}{\partial x_j \partial x_j} = r(r-1)(x_j)^{r-2},$$

$$\frac{\partial^2 D_r(\mathbf{x}\|\mathbf{y})}{\partial x_j \partial y_j} = \frac{\partial^2 D_r(\mathbf{x}\|\mathbf{y})}{\partial y_j \partial x_j} = -r(r-1)(y_j)^{r-2},$$

$$\frac{\partial^2 D_r(\mathbf{x}\|\mathbf{y})}{\partial y_j \partial y_j} = r(r-1)^2(y_j)^{r-2} - r(r-1)(r-2)x_j(y_j)^{r-3}.$$

Therefore the Hessian matrix is block-diagonal $2 \cdot J \times 2 \cdot J$ matrix

$$\mathbf{H} = \begin{pmatrix} \mathbf{B_1} & \mathbf{0} & \mathbf{0} & \ldots \\ \mathbf{0} & \mathbf{B_2} & \mathbf{0} & \ldots \\ \mathbf{0} & \mathbf{0} & \mathbf{B_3} & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where each block is $2 \times 2$ matrix

$$\mathbf{B_j} = \begin{pmatrix} r(r-1)(x_j)^{r-2}, & -r(r-1)(y_j)^{r-2} \\ -r(r-1)(y_j)^{r-2}, & r(r-1)^2(y_j)^{r-2} - r(r-1)(r-2)x_j(y_j)^{r-3} \end{pmatrix}.$$

A continuously differentiable function is convex if its Hessian matrix is positive-semidefinite. A symmetric matrix $\mathbf{H}$ is positive-semidefinite if determinants of all its principal sub-matrices are non-negative. It is not hard to see that we need to verify that the matrix $\mathbf{B_j}$ is positive-semidefinite for all $j \in \mathrm{Sig}(\mathbf{y})$:

The determinant of the first principal sub-matrix of $\mathbf{B_j}$ is $r(r-1)(x_j)^{r-2} \geq 0$, since $2 \geq r > 1$ and $x_j \in [0,1]$. After some algebraic manipulation the determinant of $\mathbf{B_j}$ is equal to

$$r^2(r-1)^2(y_j)^{r-3}[-(r-2)(x_j)^{r-1} + (r-1)y_j(x_j)^{r-2} - (y_j)^{r-1}].$$

One can easily see that for $x_j = y_j$ the term above is zero. Fix $y_j$. Then differentiating according to $x_j$ gives

$$r^2(r-1)^3(2-r)(y_j)^{r-3}(x_j)^{r-3}[x_j - y_j].$$

Clearly, for $1 \geq x_j > y_j$ the value of the determinant is increasing, and for $0 \leq x_j < y_j$ this value is decreasing. Therefore at $x_j = y_j$ we get a global minimum (for $y_j$ fixed) and we conclude that the determinant is non-negative. □

**Lemma 2.1.19.** *Assume that a Bregman divergence $D_f(\mathbf{w}\|\mathbf{v})$ is a convex function over the domain $\{(\mathbf{w},\mathbf{v}) \in \mathbb{D}^L \times \mathbb{D}^L\colon \mathbf{v} \gg_f \mathbf{w}\}$. Then*

$$\sum_{i=1}^{n} D_f(\mathbf{w}^{(i)} \| \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}))$$

*is also a convex function over the domain $\{(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in \underbrace{\mathbb{D}^L \times \ldots \times \mathbb{D}^L}_{n}\}$.*

*Proof.* Since $D_f(\cdot\|\cdot)$ is a convex function and the sum of convex functions is convex we have that

$$\lambda \sum_{i=1}^{n} D_f(\mathbf{w}^{(i)}\|\mathbf{v}) + (1-\lambda)\sum_{i=1}^{n} D_f(\mathbf{u}^{(i)}\|\mathbf{s}) \geq \sum_{i=1}^{n} D_f(\lambda\mathbf{w}^{(i)} + (1-\lambda)\mathbf{u}^{(i)}\|\lambda\mathbf{v} + (1-\lambda)\mathbf{s})$$

for any $\lambda \in [0,1]$ and $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{n} \in \mathbb{D}^L$, $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{n} \in \mathbb{D}^L$ and $\mathbf{v}, \mathbf{s} \in \mathbb{D}^L$ such that $\mathbf{v} \gg_f \mathbf{w}^{(i)}$ and $\mathbf{s} \gg_f \mathbf{u}^{(i)}$ for all $1 \leq i \leq n$. By substituting $\mathbf{v} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$ and $\mathbf{s} = \mathbf{LinOp}(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)})$ we have also

$$\lambda \sum_{i=1}^{n} D_f(\mathbf{w}^{(i)}\| \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})) + (1-\lambda)\sum_{i=1}^{n} D_f(\mathbf{u}^{(i)}\| \mathbf{LinOp}(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)})) \geq$$

$$\geq \sum_{i=1}^{n} D_f(\lambda\mathbf{w}^{(i)} + (1-\lambda)\mathbf{u}^{(i)}\|\lambda\, \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) + (1-\lambda)\, \mathbf{LinOp}(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)})) =$$

$$= \sum_{i=1}^{n} D_f(\lambda\mathbf{w}^{(i)} + (1-\lambda)\mathbf{u}^{(i)}\| \mathbf{LinOp}(\lambda\mathbf{w}^{(1)} + (1-\lambda)\mathbf{u}^{(1)}, \ldots, \lambda\mathbf{w}^{(n)} + (1-\lambda)\mathbf{u}^{(n)}))$$

which concludes the proof. □

The following theorem will have later some strong consequences.

**Theorem 2.1.20.** *Let $D_f$ be a convex Bregman divergence. Let $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)} \in \mathbb{D}^L$ and $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(n)} \in \mathbb{D}^L$ be such that*

$$\sum_{i=1}^{n} D_f(\mathbf{u}^{(i)} \| \mathbf{v}) > \sum_{i=1}^{n} D_f(\mathbf{a}^{(i)} \| \mathbf{a}),$$

*where $\mathbf{v} = \mathbf{LinOp}(\mathbf{u}^{(1)} \ldots, \mathbf{u}^{(n)})$ and $\mathbf{a} = \mathbf{LinOp}(\mathbf{a}^{(1)} \ldots, \mathbf{a}^{(n)})$. Assume that $\mathbf{u}^{(i)} \gg_f \mathbf{a}^{(i)}$ for all $1 \leq i \leq n$. Then*

$$\sum_{i=1}^{n} (\mathbf{a}^{(i)} - \mathbf{u}^{(i)}) \cdot ((\mathbf{u}^{(i)})^* - \mathbf{v}^*) < 0.$$

*Proof.* First of all notice that by lemma 2.1.10 and by the assumption $\mathbf{u}^{(i)} \gg_f \mathbf{a}^{(i)}$, for all $i$, we have that

$$D_f(\mathbf{a}^{(i)} \| \mathbf{v}) - D_f(\mathbf{u}^{(i)} \| \mathbf{v}) - D_f(\mathbf{a}^{(i)} \| \mathbf{u}^{(i)}) = (\mathbf{a}^{(i)} - \mathbf{u}^{(i)}) \cdot ((\mathbf{u}^{(i)})^* - \mathbf{v}^*).$$

The above has a sense since $\mathbf{v} \gg_f \mathbf{u}^{(i)}$ for all $1 \leq i \leq n$. By the parallelogram theorem

$$\sum_{i=1}^{n} D_f(\mathbf{a}^{(i)} \| \mathbf{v}) = \sum_{i=1}^{n} D_f(\mathbf{a}^{(i)} \| \mathbf{a}) + n \cdot D_f(\mathbf{a} \| \mathbf{v}).$$

Hence

$$\begin{aligned} &\sum_{i=1}^{n} D_f(\mathbf{a}^{(i)} \| \mathbf{a}) - \sum_{i=1}^{n} D_f(\mathbf{u}^{(i)} \| \mathbf{v}) + n \cdot D_f(\mathbf{a} \| \mathbf{v}) - \\ &- \sum_{i=1}^{n} D_f(\mathbf{a}^{(i)} \| \mathbf{u}^{(i)}) = \sum_{i=1}^{n} (\mathbf{a}^{(i)} - \mathbf{u}^{(i)}) \cdot ((\mathbf{u}^{(i)})^* - \mathbf{v}^*). \end{aligned} \tag{2.5}$$

Since we assume that $D_f(\mathbf{w} \| \mathbf{v})$ is a convex function in both arguments whenever $\mathbf{v} \gg_f \mathbf{w}$ by the Jensen inequality

$$n \cdot D_f(\mathbf{a} \| \mathbf{v}) - \sum_{i=1}^{n} D_f(\mathbf{a}^{(i)} \| \mathbf{u}^{(i)}) \leq 0. \tag{2.6}$$

The inequality (2.6) together with the assumption that

$$\sum_{i=1}^{n} D_f(\mathbf{u}^{(i)} \| \mathbf{v}) > \sum_{i=1}^{n} D_f(\mathbf{a}^{(i)} \| \mathbf{a})$$

gives that left-hand side of the equality (2.5) is negative and so the right-hand side is too, whence

$$\sum_{i=1}^{n} (\mathbf{a}^{(i)} - \mathbf{u}^{(i)}) \cdot ((\mathbf{u}^{(i)})^* - \mathbf{v}^*) < 0$$

as required. □

Figure 2.3 depicts the situation in the proof above for $n = 2$. Arrows indicate corresponding divergences.

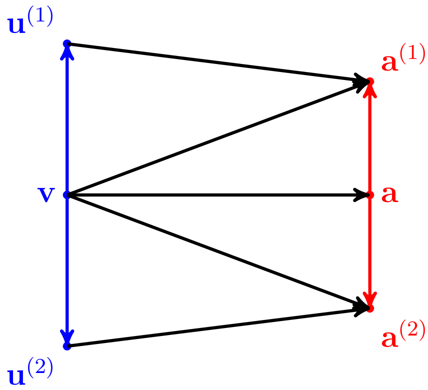


Figure 2.3: The situation in the proof of theorem 2.1.20 for $n = 2$.

## 2.2 Conjugated geometry

In the previous section we have defined a projection exploiting fact that a Bregman divergence is a strictly convex function in the first argument. Since the same is not true for the second argument, not even if we confine ourselves to KL-divergence, we have not defined a projection by means of the second argument. However due to the fact that $\mathrm{KL}(\cdot\|\cdot)$ is a convex function we can define a notion of conjugated KL-projection, introduced by Matúš in [34], as follows:

Let $\mathbf{v} \in \mathbb{D}^L$ and $W$ be a closed convex set of $L$-probability functions such that there is at least one $\mathbf{w} \in W$ which dominates $\mathbf{v}$. Then the set $V \subseteq W$ of all $L$-probability functions $\mathbf{w}$ which minimise $\mathrm{KL}(\mathbf{v}\|\mathbf{w})$ subject only to $\mathbf{w} \in W$ is nonempty, closed and convex (since KL is a convex function). Whenever $v_j > 0$ for all $1 \le j \le J$ we have that $V$ is a singleton — the *conjugated* KL-*projection* of $\mathbf{v}$ to $W$. This is proved in forthcoming lemma 2.2.1. If $V$ is not a singleton then the conjugated KL projection of $\mathbf{v}$ to $W$ is defined to be $\mathbf{ME}_L(V)$.

**Lemma 2.2.1.** *Let $\mathbf{v} \in \mathbb{D}^L$ be such that $v_j > 0$, $1 \le j \le J$. Then $\mathrm{KL}(\mathbf{v}\|\mathbf{w})$ is a strictly convex function over the domain $\{\mathbf{w} \in \mathbb{D}^L : \ w_j > 0,\ 1 \le j \le J\}$.*

*Proof.* $\frac{\partial}{\partial w_j}\mathrm{KL}(\mathbf{v}\|\mathbf{w}) = -\frac{v_j}{w_j}$ and $\frac{\partial^2}{\partial w_j \partial w_j}\mathrm{KL}(\mathbf{v}\|\mathbf{w}) = \frac{v_j}{(w_j)^2}$. Then the Hessian matrix is positive definite and the considered function is strictly convex over the specified domain. □

**Theorem 2.2.2** (Four Points Property)**.** *Let $V$ be a convex closed subset of $\mathbb{D}^L$ and let $\mathbf{v}, \mathbf{u}, \mathbf{w}, \mathbf{a} \in \mathbb{D}^L$ be such that $\mathbf{v} \gg \mathbf{u}$, and $\mathbf{w}$ is the conjugated* KL*-projection of* $\mathbf{v}$

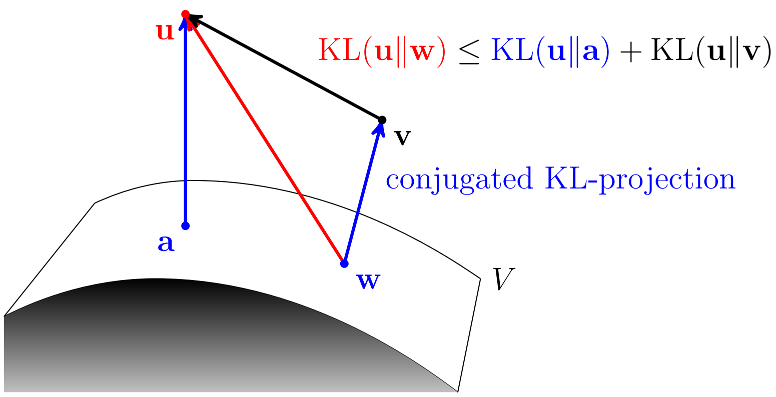


Figure 2.4: The illustration of theorem 2.2.2.

*into* $V$ *and* $\mathbf{a} \in V$ *be such that* $\mathbf{a} \gg \mathbf{u}$. *Then*

$$\mathrm{KL}(\mathbf{u}\|\mathbf{w}) \leq \mathrm{KL}(\mathbf{u}\|\mathbf{a}) + \mathrm{KL}(\mathbf{u}\|\mathbf{v}).$$

The theorem above is a specific instance of a result due to Csiszár and Tusnády, see [12], lemma 3 (the above formulation using the term 'conjugated KL-projection' appeared in [34]). We will not prove it here since a proof can be derived from our forthcoming proof of theorem 2.2.3. The theorem is illustrated in figure 2.4.

It is a remarkable fact that we can abstractly define 'conjugated projection' with respect to a sum of KL-divergences: Let $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)} \in \mathbb{D}^L$. Consider the following minimisation problem. For which $\mathbf{v} \in \mathbb{D}^L$ is the sum

$$\sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)}\|\mathbf{v})$$

minimal subject only to $\mathbf{v} \gg \mathbf{w}^{(i)}$, $1 \leq i \leq n$?
By theorem 2.1.13

$$\mathbf{v} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}).$$

We can abstractly think of $\mathbf{v}$ as the conjugated projection of $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$ into $\mathbb{D}^L$ since the following analogue of the four points property holds.

**Theorem 2.2.3.** *Let* $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$, $\mathbf{w}^{(1)} \in V^L_{\mathbf{K}_1}, \ldots, \mathbf{w}^{(n)} \in V^L_{\mathbf{K}_n}$ *and* $\mathbf{v} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$. *Let* $\mathbf{a}^{(1)} \in V^L_{\mathbf{K}_1}, \ldots, \mathbf{a}^{(n)} \in V^L_{\mathbf{K}_n}$ *be such that* $\mathbf{w}^{(i)} \gg \mathbf{a}^{(i)}$, $1 \leq i \leq n$, *and* $\mathbf{u} \in \mathbb{D}^L$ *be such that* $\mathbf{u} \gg \mathbf{a}^{(i)}$, $1 \leq i \leq n$. *Then*

$$\sum_{i=1}^{n} \mathrm{KL}(\mathbf{a}^{(i)}\|\mathbf{v}) \leq \sum_{i=1}^{n} \mathrm{KL}(\mathbf{a}^{(i)}\|\mathbf{u}) + \sum_{i=1}^{n} \mathrm{KL}(\mathbf{a}^{(i)}\|\mathbf{w}^{(i)}).$$

*Proof.* By lemma 2.1.10

$$\sum_{i=1}^{n} \mathrm{KL}(\mathbf{a}^{(i)} \| \mathbf{w}^{(i)}) = \sum_{i=1}^{n} \mathrm{KL}(\mathbf{a}^{(i)} \| \mathbf{v}) - \sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{v}) - \sum_{i=1}^{n} \sum_{j \in \mathrm{Sig}(\mathbf{v})} (a_j^{(i)} - w_j^{(i)}) \log \frac{w_j^{(i)}}{v_j}.$$

Notice that the assumption $\mathbf{w}^{(i)} \gg \mathbf{a}^{(i)}$, $1 \leq i \leq n$, is necessary here. We can rewrite the above as

$$\sum_{i=1}^{n} \mathrm{KL}(\mathbf{a}^{(i)} \| \mathbf{w}^{(i)}) - \sum_{i=1}^{n} \mathrm{KL}(\mathbf{a}^{(i)} \| \mathbf{v}) + \sum_{i=1}^{n} \mathrm{KL}(\mathbf{a}^{(i)} \| \mathbf{u}) =$$

$$= \sum_{i=1}^{n} \mathrm{KL}(\mathbf{a}^{(i)} \| \mathbf{u}) - \sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{v}) - \sum_{i=1}^{n} \sum_{j \in \mathrm{Sig}(\mathbf{v})} (a_j^{(i)} - w_j^{(i)}) \log \frac{w_j^{(i)}}{v_j}. \tag{2.7}$$

Since $\mathrm{KL}(\cdot \| \cdot)$ is a convex function, by applying the first convexity condition (theorem 2.1.1) twice, we have that

$$\sum_{i=1}^{n} \mathrm{KL}(\mathbf{a}^{(i)} \| \mathbf{u}) \geq \sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{v}) +$$

$$+ \sum_{i=1}^{n} \sum_{j \in \mathrm{Sig}(\mathbf{v})} (a_j^{(i)} - w_j^{(i)}) \frac{\partial}{\partial x_j} \Big[ \mathrm{KL}(\mathbf{x} \| \mathbf{v}) \Big] \Big|_{\mathbf{x} = \mathbf{w}^{(i)}}$$

$$+ \sum_{i=1}^{n} \sum_{j \in \mathrm{Sig}(\mathbf{v})} (u_j - v_j) \frac{\partial}{\partial x_j} \Big[ \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{x}) \Big] \Big|_{\mathbf{x} = \mathbf{v}}. \tag{2.8}$$

(2.7) and (2.8) give that

$$\sum_{i=1}^{n} \mathrm{KL}(\mathbf{a}^{(i)} \| \mathbf{v}) \leq \sum_{i=1}^{n} \mathrm{KL}(\mathbf{a}^{(i)} \| \mathbf{u}) + \sum_{i=1}^{n} \mathrm{KL}(\mathbf{a}^{(i)} \| \mathbf{w}^{(i)}) -$$

$$- \sum_{i=1}^{n} \sum_{j \in \mathrm{Sig}(\mathbf{v})} (u_j - v_j) \frac{\partial}{\partial x_j} \Big[ \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{x}) \Big] \Big|_{\mathbf{x} = \mathbf{v}}.$$

Notice that

$$- \sum_{i=1}^{n} \sum_{j \in \mathrm{Sig}(\mathbf{v})} (u_j - v_j) \frac{\partial}{\partial x_j} \Big[ \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{x}) \Big] \Big|_{\mathbf{x} = \mathbf{v}} =$$

$$= \sum_{j \in \mathrm{Sig}(\mathbf{v})} (u_j - v_j) \cdot n = -n \cdot \Big( 1 - \sum_{j \in \mathrm{Sig}(\mathbf{v})} u_j \Big).$$

Since $\sum_{j \in \mathrm{Sig}(\mathbf{v})} u_j \leq 1$ the above is not positive and therefore the theorem follows. □

Note that a proof for theorem 2.2.2 can be constructed analogously to the proof above.

The following is a counterpart to the parallelogram theorem for the *conjugated geometry*.

**Theorem 2.2.4.** *Let* $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}, \mathbf{v} \in \mathbb{D}^L$ *be such that* $\mathbf{w}^{(i)} \gg \mathbf{v}$ *for all* $1 \le i \le n$. *Then*

$$\sum_{i=1}^{n} \mathrm{KL}(\mathbf{v} \| \mathbf{w}^{(i)}) = \sum_{i=1}^{n} \mathrm{KL}(\mathbf{LogOp}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}) \| \mathbf{w}^{(i)}) + \\ + n \cdot \mathrm{KL}(\mathbf{v} \| \mathbf{LogOp}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})).$$

*Proof.* Let $\mathbf{w} = \mathbf{LogOp}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$. First note that

$$\sum_{i=1}^{n} \sum_{j \in \mathrm{Sig}(\mathbf{w}^{(i)})} v_j \log \frac{v_j}{w_j^{(i)}} = \sum_{j \in \bigcap_{i=1}^{n} \mathrm{Sig}(\mathbf{w}^{(i)})} v_j \log \frac{(v_j)^n}{\prod_{i=1}^{n} w_j^{(i)}}.$$

This is because whenever $w_j^{(i)} = 0$ we have $v_j = 0$ since $\mathbf{w}^{(i)} \gg \mathbf{v}$ for all $1 \le i \le n$. Now

$$\sum_{j \in \bigcap_{i=1}^{n} \mathrm{Sig}(\mathbf{w}^{(i)})} v_j \log \frac{(v_j)^n}{\prod_{i=1}^{n} w_j^{(i)}} = n \sum_{j \in \bigcap_{i=1}^{n} \mathrm{Sig}(\mathbf{w}^{(i)})} v_j \log \frac{v_j}{w_j} - \\ -n \Big( \sum_{j \in \bigcap_{i=1}^{n} \mathrm{Sig}(\mathbf{w}^{(i)})} v_j \Big) \log \Big( \sum_{j \in \bigcap_{i=1}^{n} \mathrm{Sig}(\mathbf{w}^{(i)})} (\prod_{i=1}^{n} w_j^{(i)})^{\frac{1}{n}} \Big) = \\ = n \sum_{j \in \bigcap_{i=1}^{n} \mathrm{Sig}(\mathbf{w}^{(i)})} v_j \log \frac{v_j}{w_j} + \sum_{i=1}^{n} \sum_{j \in \bigcap_{i=1}^{n} \mathrm{Sig}(\mathbf{w}^{(i)})} w_j \log \frac{w_j}{w_j^{(i)}}, \tag{2.9}$$

where $\sum_{j \in \bigcap_{i=1}^{n} \mathrm{Sig}(\mathbf{w}^{(i)})} v_j = 1$. Since $w_j \neq 0$ if and only if $j \in \bigcap_{i=1}^{n} \mathrm{Sig}(\mathbf{w}^{(i)})$, (2.9) is equal to

$$\sum_{i=1}^{n} \sum_{j \in \mathrm{Sig}(\mathbf{w}^{(i)})} w_j \log \frac{w_j}{w_j^{(i)}} + n \sum_{j \in \mathrm{Sig}(\mathbf{w})} v_j \log \frac{v_j}{w_j}.$$

□

Later we will see that the two theorems above have very nice consequences for a p-merging operator based on the conjugated KL-projection.

**Theorem 2.2.5.** *Let* $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}$ *be given L-probability functions satisfying (1.3). Then the function*

$$\sum_{i=1}^{n} \mathrm{KL}(\mathbf{v} \| \mathbf{w}^{(i)})$$

*is strictly minimal for*

$$\mathbf{v} = \mathbf{LogOp}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}).$$

*Proof.* By theorem 2.2.4 the minimality of $\sum_{i=1}^{n} \mathrm{KL}(\mathbf{v} \| \mathbf{w}^{(i)})$ for fixed $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}$ is equivalent to the minimality of $\mathrm{KL}(\mathbf{v} \| \mathbf{LogOp}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}))$. Since $\mathrm{KL}(\mathbf{v} \| \mathbf{w}) = 0$ only if $\mathbf{v} = \mathbf{w}$ and otherwise it is positive the unique minimum of the considered function is at the point $\mathbf{v} = \mathbf{LogOp}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$. □

In this chapter we have offered a complete definition for Bregman divergences which, to our knowledge, has not appeared elsewhere. We have provided a comprehensive list of geometrical properties and we have supplied the proofs for all those properties within our framework. In particular the following new lemmas and theorems have been proved: 2.1.3, 2.1.18, 2.1.19, 2.1.20 and 2.2.4.

# Chapter 3

# Merging based on convex Bregman divergences

## 3.1 Minimum sum of divergences

In this section we will define p-merging operators based on minimization of a sum of convex Bregman divergences, for which we have prepared the ground in the previous chapter. We also aim to show that these p-merging operators have better properties than obdurate ones.

For any $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ the $\hat{\Delta}^D$*-merging operator* is defined as follows:

$$\hat{\Delta}^D_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \Big\{ \arg\min_{\mathbf{v}\in\mathbb{D}^L} \sum_{i=1}^{n} D(\mathbf{w}^{(i)} \| \mathbf{v}) \colon \ \mathbf{w}^{(i)} \in V^L_{\mathbf{K}_i},\ 1 \le i \le n \Big\},$$

where $D$ is a convex Bregman divergence and the right hand-side denotes the set of all possible minimisers. This is the second possible use of the notation 'arg min' in this thesis as a set constructor.

In the special case of the Kullback-Leibler divergence we will call $\hat{\Delta}^{\mathrm{KL}}$ the *Linear Entropy Operator*. Similarly $\hat{\Delta}^{\mathrm{E2}}$ is called the *Linear Euclidean Operator* and this operator was first formulated by Osherson and Vardi in [38]. Finally, in the case of a Renyi-B divergence $D_r$, for $2 \ge r > 1$, we call $\hat{\Delta}^{D_r}$ the *Linear Renyi Operator.*

Dually to the linear entropy operator we define the *Social Entropy Operator* $\Delta^{\mathrm{KL}}$ for any $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ by

$$\Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \Big\{ \arg\min_{\mathbf{v}\in\mathbb{D}^L} \sum_{i=1}^{n} \mathrm{KL}(\mathbf{v} \| \mathbf{w}^{(i)}) \colon \ \mathbf{w}^{(i)} \in V^L_{\mathbf{K}_i},\ 1 \le i \le n \Big\}.$$

This operator was first introduced by Wilmers in [52]. Notice that the Kullback-Leibler divergence in the definition of $\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ takes the divergence between $\mathbf{v}$ and $\mathbf{w}^{(i)}$ in the opposite direction to the one used in the definition of $\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$.

We will show (theorem 3.1.2) that the $\hat{\Delta}^D$-merging operator satisfies the principle **(K1)** for any convex Bregman divergence $D$. But first we will state the following auxiliary lemma.

**Lemma 3.1.1.** *Let $D$ be a convex Bregman divergence. In particular $D$ can be* KL *or* E2*. Then the following are equivalent:*

1. *The $L$-probability functions $\mathbf{v}, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ minimise the quantity*

$$\sum_{i=1}^{n} D(\mathbf{w}^{(i)} \| \mathbf{v})$$

*subject to $\mathbf{w}^{(1)} \in V_{\mathbf{K}_1}^L, \ldots, \mathbf{w}^{(n)} \in V_{\mathbf{K}_n}^L$.*

2. *The $L$-probability functions $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ minimise the quantity*

$$\sum_{i=1}^{n} D(\mathbf{w}^{(i)} \| \, \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$$

*subject to $\mathbf{w}^{(1)} \in V_{\mathbf{K}_1}^L, \ldots, \mathbf{w}^{(n)} \in V_{\mathbf{K}_n}^L$ and*

$$\mathbf{v} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}).$$

*Proof.* This lemma follows directly by theorem 2.1.13. □

By $\hat{\Gamma}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ we denote the set of (ordered) $n$-tuples of probability functions satisfying condition 2 of the previous lemma. Hence clearly $\hat{\Delta}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ can be characterized as the set of all probability functions of the form $\mathbf{v} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$ where $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in \hat{\Gamma}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n)$.

Notice that for any $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)} \in \mathbb{D}^L$ $\mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$ is an $L$-probability function. This, together with the fact that $D$ is convex, implies that $\hat{\Delta}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is nonempty and hence well-defined for any $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$.

It also follows that every $\hat{\Delta}^D$-merging operator extends the **LinOp**-pooling operator, in the sense that it coincides with **LinOp** in the special case when each knowledge base admits only a single probability function. However, unlike the obdurate linear entropy process **OLEP** which possesses the same feature, this operator also satisfies the consistency principle **(K4)**, and in fact the stronger form **(K4*)**, since given $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ the value

$$\hat{\mathrm{M}}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \min\left\{ \sum_{i=1}^n D(\mathbf{w}^{(i)} \| \mathbf{v}) \,:\, \mathbf{v} \in \mathbb{D}^L \,;\, \mathbf{w}^{(1)} \in V_{\mathbf{K}_1}^L, \ldots, \mathbf{w}^{(n)} \in V_{\mathbf{K}_n}^L \right\}$$

lies in the interval $[0, +\infty)$ and is equal to 0 if and only if $\mathbf{v} = \mathbf{w}^{(1)} = \ldots = \mathbf{w}^{(n)}$.

Although it is not hard to see that the $\hat{\Delta}^D$-merging operator satisfies also the collegiality principle **(K5)**, we give an elegant proof of this fact later as a consequence of theorem 4.1.1.

**Theorem 3.1.2.** *Let $D$ be a convex Bregman divergence. Then for all $n$-tuples $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ the set $\hat{\Delta}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is a closed convex region of $\mathbb{D}^L$.*

*Proof.* Let $\mathbf{v}, \mathbf{s} \in \hat{\Delta}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. We need to show that $\lambda\mathbf{v} + (1-\lambda)\mathbf{s} \in \hat{\Delta}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ for any $\lambda \in [0, 1]$.

Assume that $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in \hat{\Gamma}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ are such that

$$\mathbf{v} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$$

and $(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)}) \in \hat{\Gamma}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ are such that

$$\mathbf{s} = \mathbf{LinOp}(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)}).$$

By lemma 2.1.19 the function

$$g(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}) = \sum_{i=1}^n D(\mathbf{x}^{(i)} \| \,\mathbf{LinOp}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}))$$

is convex over the convex region specified by constraints $\mathbf{x}^{(i)} \in V_{\mathbf{K}_i}^L$, $1 \le i \le n$. Moreover, by the definition of the set $\hat{\Delta}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n)$, the function $g$ attains its minimum over this convex region at points $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$ and $(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)})$. We need to show that $g$ also attains its minimum at the point

$$\lambda(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) + (1-\lambda)(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)})$$

for any $\lambda \in [0, 1]$. Since $g$ is convex we have that

$$\lambda g(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) + (1-\lambda) g(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)}) \ge$$

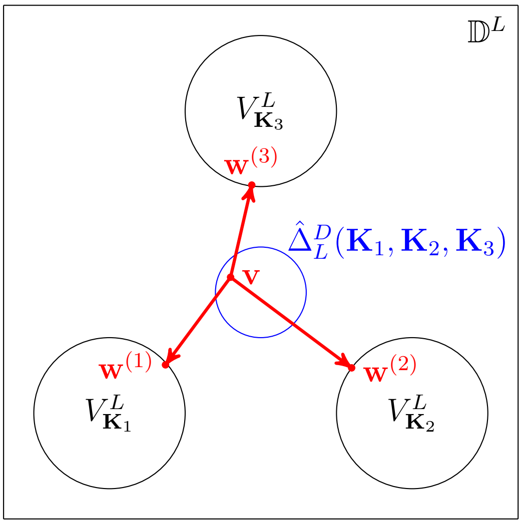


Figure 3.1: $\hat{\Delta}^D$-merging operator.

$$\geq g(\lambda(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) + (1-\lambda)(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)}))$$

by the Jensen inequality. Since $g(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) = g(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)})$ the inequality above can only hold with equality and therefore

$$\lambda\mathbf{v} + (1-\lambda)\mathbf{s} \in \hat{\Delta}^D_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$$

for any $\lambda \in [0, 1]$.

Note that, since $V_{\mathbf{K}_i}$ is a closed set for all $1 \leq i \leq n$, by lemma 3.1.1 the set $\hat{\Gamma}^D_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is closed. It follows that $\hat{\Delta}^D_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is closed as well. □

We can conclude that a $\hat{\Delta}^D$-merging operator is a probabilistic merging operator whenever $D$ is a convex Bregman divergence. It also obviously satisfies principles **(K2)** and **(K3)**. Figure 3.1 illustrates the case $n = 3$.

The social entropy operator $\Delta^{\mathrm{KL}}$ was first introduced by Wilmers in [52] and it was further studied in [1] and [2] by Wilmers and the author. The rest of this section is taken entirely from [2].

Given knowledge bases $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ let

$$\mathrm{M}^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \min\left\{ \sum_{i=1}^n \mathrm{KL}(\mathbf{v}\|\mathbf{w}^{(i)}) \,:\, \mathbf{v} \in \mathbb{D}^L\,;\, \mathbf{w}^{(1)} \in V^L_{\mathbf{K}_1}, \ldots, \mathbf{w}^{(n)} \in V^L_{\mathbf{K}_n} \right\}.$$

It is easy to see that this is well-defined (see [53]). Note that this value lies in the interval $[0, +\infty]$. Also $\mathrm{M}^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = 0$ if and only if $\mathbf{v} = \mathbf{w}^{(1)} = \ldots = \mathbf{w}^{(n)}$

in the definition above in which case the $\mathbf{K}_1, \ldots, \mathbf{K}_n$ are jointly consistent. Also $\mathrm{M}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is finite if and only if condition (1.4) holds. Therefore we redefine $\Delta^{\mathrm{KL}}$ as follows: for any $L$ and any $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ $\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is defined as

$$\{\mathbf{v} \in \mathbb{D}^L \ : \ \exists\, \mathbf{w}^{(1)} \in V_{\mathbf{K}_1}^L, \ldots, \mathbf{w}^{(n)} \in V_{\mathbf{K}_n}^L \text{ s.t. } \sum_{i=1}^{n} \mathrm{KL}(\mathbf{v} \| \mathbf{w}^{(i)}) \ = \ \mathrm{M}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)\}. \tag{3.1}$$

**Example 3.1.3.** Let $L = \{p, q\}$ and $\mathbf{K}_1 = \{Bel(p) = 0.2, Bel(q) = 0\}$, $\mathbf{K}_2 = \{Bel(p) = 0.4, Bel(q) = 1\}$. There is only one probability function $\mathbf{w}^{(1)} \in V_{\mathbf{K}_1}^L$: $(0, 0.2, 0, 0.8)$ (atoms are listed in the obvious order) and the only one $\mathbf{w}^{(2)} \in V_{\mathbf{K}_2}^L$: $(0.4, 0, 0.6, 0)$. Hence by (3.1) $\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \mathbf{K}_2) = \mathbb{D}^L$. □

In [53] it is shown that for any $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ the set $\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is always a nonempty closed convex region of $\mathbb{D}^L$, and hence it follows that $\Delta^{\mathrm{KL}}$ is a p-merging operator (i.e. it satisfies **(K1)**). We note, however, that although $\Delta^{\mathrm{KL}}$ is everywhere defined[1] it is really only interesting as a merging operator for those $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ for which the relatively undemanding condition (1.4) is satisfied, since otherwise applying $\Delta^{\mathrm{KL}}$ simply returns the whole space $\mathbb{D}^L$. The fact that the social entropy operator $\Delta^{\mathrm{KL}}$ satisfies **(K4*)** follows at once from the fact noted above that $\mathrm{M}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n) = 0$ if and only if $\mathbf{v} = \mathbf{w}^{(1)} = \ldots = \mathbf{w}^{(n)}$ in the definition of $\mathrm{M}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. Moreover $\Delta^{\mathrm{KL}}$ satisfies **(K2)** and **(K3)** trivially by definition. Although it is not hard to see that the $\Delta^{\mathrm{KL}}$-merging operator also satisfies the collegiality principle **(K5)**, we give an elegant proof of this fact later as a consequence of theorem 4.1.1.

**Lemma 3.1.4.** *Assume that $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ satisfy condition (1.4). Then the following are equivalent:*

*(i) The L-probability functions $\mathbf{v}$, $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ minimise $\sum_{i=1}^{n} \mathrm{KL}(\mathbf{v} \| \mathbf{w}^{(i)})$ subject only to $\mathbf{w}^{(1)} \in V_{\mathbf{K}_1}^L, \ldots, \mathbf{w}^{(n)} \in V_{\mathbf{K}_n}^L$.*

[1] In the presentation in [53] the region $\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is only defined assuming that condition (1.4) holds, but this does not significantly affect the results.

*(ii) The L-probability functions* $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ *maximise* $\sum_{j=1}^{J}(\prod_{k=1}^{n} w_j^{(k)})^{\frac{1}{n}}$, *subject only to* $\mathbf{w}^{(1)} \in V_{\mathbf{K}_1}^{L}, \ldots, \mathbf{w}^{(n)} \in V_{\mathbf{K}_n}^{L}$, *and*

$$\mathbf{v} = \mathbf{LogOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}).$$

*Proof.* Directly by theorem 2.2.5. □

We will denote the set of all (ordered) $n$-tuples satisfying condition 2 of the previous lemma by $\Gamma_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. In the very special case when each expert $i$ specifies a single probability function $\mathbf{w}^{(i)}$ the set $\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is just the singleton $\{\mathbf{LogOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})\}$.

In this section we have defined several p-merging operators ($\Delta^{\mathrm{KL}}$, $\hat{\Delta}^{D}$ for any convex Bregman divergence $D$) which satisfy the consistency principle **(K4)** and, as we will see later, also the collegiality principle **(K5)**. In the next section we will investigate these promising operators from the point of view of knowledge updating.

## 3.2 Averaging projections and fixed points

As we have noted before, lemma 2.1.8 is a basic result which stands behind many probabilistic expert systems. Given a space $\mathbb{D}^L$ this lemma allows us to interpret a new probability function $\mathbf{v} \in \mathbb{D}^L$ in a given knowledge base $W \subseteq \mathbb{D}^L$ as the $\mathbf{w} \in W$ which minimises $D_f(\mathbf{w} \| \mathbf{v})$ for a given Bregman divergence $D_f$. We can say that by identifying the single $\mathbf{w} \in W$ we have *updated* our knowledge base $W$ by new knowledge $\mathbf{v}$. An example of updating is the following problem which dates back to 1940, see [13].

**Example 3.2.1.** Suppose that a random sample of citizens answered two questions regarding their education and age. The level of education is scaled by $J_1$ categories $\alpha_1, \ldots, \alpha_{J_1}$ and age by $J_2$ categories $\beta_1, \ldots, \beta_{J_2}$. Let the distribution of respondents over $J_1 \cdot J_2$ atomic events be $\{\mathbf{v}(\alpha_j \wedge \beta_i)\}_{j=1,i=1}^{J_1,J_2}$. However the distributions of population in the country according to the level of education and age are well known statistical data, say $\mathbf{w}^{(1)}(\alpha_1), \ldots, \mathbf{w}^{(1)}(\alpha_{J_1})$ for the level of education and $\mathbf{w}^{(2)}(\beta_1), \ldots, \mathbf{w}^{(2)}(\beta_{J_2})$ for the age. Assume that $\mathbf{v}$ does not correspond to the national data i.e. $\sum_{j=1}^{J_1} \mathbf{v}(\alpha_j \wedge \beta_i) \neq \mathbf{w}^{(2)}(\beta_i)$ for some $i$ and $\sum_{i=1}^{J_2} \mathbf{v}(\alpha_j \wedge \beta_i) \neq \mathbf{w}^{(1)}(\alpha_j)$ for some $j$. How can we adjust our distribution $\mathbf{v}$ in order to make it consistent with the national distributions for education and age but still make it as close as possible to our original data? □

While in the previous example we update our incomplete but reliable knowledge by a new statistically obtained piece of information, we can use updating also from the other perspective:

**Example 3.2.2.** Suppose that the distribution of patients with respect to a disease A and symptoms C, D is well known at the national level. Now a doctor moved to a specific city where she investigated several patients. She observed that all patients having the disease A have also the symptom C but not D. However, at the national level there is a significant proportion of patients having both the disease A and the symptom D. One can argue that there might be something specific about this city which prevents the occurrence of the symptom D, and therefore the doctor wishes to find the estimation of the proportions of patients in the city which are consistent with her observations rather than adopt the national proportions. Nevertheless, the national proportions can still be highly related to the illness spreading in the city. Therefore the doctor wishes to somehow update her knowledge using the national proportions. □

A common method of updating in probabilistic expert systems is by means of the KL-projection, see [22] or [48]. In the previous example we would take the KL-projection of the national distribution of patients to the set of all possible proportions which, according to doctor's observation, could apply for the city.

In particular, KL-projections appear to be justified in the 'model theoretic' sense for problems such as in example 3.2.2. More specifically, if there are two large sets of examples $V$ and $W$ such that $W$ is small relatively to $V$ (e.g., $W$ are patients in a specific city), we are given a distribution $\mathbf{v}$ of examples in $V$ and a set of constraints $\mathbf{K}$ on possible distributions of examples in $W$, then the KL-projection identify the most likely distribution of examples in $W$ if they are taken randomly from $V$. For more details see theorem 8.7 of [39].

Since the above 'model theoretic' justification of KL-updating is essentially the same as the one for the **ME**-inference process, it is not surprising that KL-updating extends the maximum entropy inference process in the sense that given a language $L$ the KL-update of $\mathbf{K} \in CL$ by the uniform probability function $\underbrace{\left(\frac{1}{J}, \dots, \frac{1}{J}\right)}_{J}$, $J =$

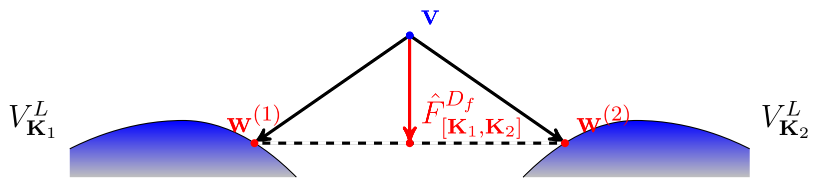


Figure 3.2: An averaging projective procedure.

$|\operatorname{At}(L)|$, is equal to $\mathbf{ME}_L(\mathbf{K})$ since

$$\arg\min_{\mathbf{w}\in V^L_{\mathbf{K}}} \sum_{j=1}^{J} w_j \log \frac{w_j}{\frac{1}{J}} = \arg\max_{\mathbf{w}\in V^L_{\mathbf{K}}} -\sum_{j=1}^{J} w_j \log w_j = \mathbf{ME}_L(\mathbf{K}).$$

Recently updating by a more general Bregman divergence has become popular, e.g. [5]. Remarkably, updating by a Bregman divergence is a unifying framework for a variety of techniques used in machine learning such as *logistic regression*, see [9].

There is a striking relation between an updating considered as a projection by means of a convex Bregman divergence $D_f$, and a $\hat{\Delta}^{D_f}$-merging operator. It is easy to see that given $\mathbf{K}_1, \dots, \mathbf{K}_n \in CL$ and $\mathbf{v} \in \hat{\Delta}^{D_f}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$, we have that those $\mathbf{w}^{(1)} \in V^L_{\mathbf{K}_1}, \dots, \mathbf{w}^{(n)} \in V^L_{\mathbf{K}_n}$ which globally minimise

$$\sum_{i=1}^{n} D_f(\mathbf{w}^{(i)} \| \mathbf{v})$$

are also $D_f$-projections of $\mathbf{v}$ into $V^L_{\mathbf{K}_1}, \dots, V^L_{\mathbf{K}_n}$ respectively. Every $\mathbf{v} \in \hat{\Delta}^{D_f}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ is then a *fixed point* of the mapping

$$\hat{F}^{D_f}_{[\mathbf{K}_1,\dots,\mathbf{K}_n]} : U \to \mathbb{D}^L,$$

where $U = \{\mathbf{v} \in \mathbb{D}^L : \text{ for all } 1 \le i \le n \ \exists \mathbf{w} \in V^L_{\mathbf{K}_i}(\mathbf{v} \gg_f \mathbf{w})\}$, defined by

$$\hat{F}^{D_f}_{[\mathbf{K}_1,\dots,\mathbf{K}_n]}(\mathbf{v}) = \mathbf{LinOp}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$$

and $\mathbf{w}^{(i)} = \arg\min_{\mathbf{w}\in V^L_{\mathbf{K}_i}} D_f(\mathbf{w}\|\mathbf{v})$ for all $1 \le i \le n$. A mapping such as above is called an *averaging projective procedure.* Figure 3.2 depicts the geometrical idea of this procedure.

If we denote the set of all fixed points of $\hat{F}^{D_f}_{[\mathbf{K}_1,\dots,\mathbf{K}_n]}$ by $\hat{\Theta}^{D_f}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ then we have already established the following.

**Theorem 3.2.3.** *For all $n$-tuples of knowledge bases $\mathbf{K}_1, \dots, \mathbf{K}_n \in CL$ and any convex Bregman divergence $D$*

$$\hat{\Delta}_L^D(\mathbf{K}_1, \dots, \mathbf{K}_n) \subseteq \hat{\Theta}_L^D(\mathbf{K}_1, \dots, \mathbf{K}_n).$$

*In particular $\hat{\Theta}_L^D(\mathbf{K}_1, \dots, \mathbf{K}_n)$ is always nonempty.*

The averaging projective procedure $\hat{F}^{\mathrm{KL}}$ was first investigated by Matúš in [34] and the idea of combining E2-updating separately applied to each expert with arithmetic averaging of the updates was first introduced by Predd et al. in [42].

The following example illustrates that in some contexts it makes a good sense to investigate the set of all fixed points $\hat{\Theta}_L^{D_f}(\mathbf{K}_1, \dots, \mathbf{K}_n)$.

**Example 3.2.4.** Assume that there are $n$ experts each with knowledge base $\mathbf{K}_1, \dots, \mathbf{K}_n \in CL$ respectively. Say that an independent chairman of the college has announced a probability function $\mathbf{v}$ to represent the agreement of the college of experts. Each expert then naturally updates her own knowledge base by what seems to be the right probability function. In other words, the expert '$i$' projects $\mathbf{v}$ to $V_{\mathbf{K}_i}^L$ obtaining the probability function $\mathbf{w}^{(i)}$. Each expert subsequently accepts $\mathbf{w}^{(i)}$ as her working hypothesis, but knowledge base $\mathbf{K}_i$ is not discarded, she only takes other people opinion into account. Then it is easy for the chairman to identify the average of the actual beliefs $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}$ of the experts. If he found that this average $\mathbf{v}'$ does not coincide with the originally announced probability function $\mathbf{v}$ then he would naturally feel unhappy about such a choice, so he would be tempted to iterate the process in a hope that eventually he will find a fixed point. □

**Theorem 3.2.5.** *Let $\mathbf{K}_1, \dots, \mathbf{K}_n \in CL$ and $D_f$ be a convex Bregman divergence. Assume that*

$$\mathbf{v} \in \hat{\Theta}_L^{D_f}(\mathbf{K}_1, \dots, \mathbf{K}_n) \setminus \hat{\Delta}_L^{D_f}(\mathbf{K}_1, \dots, \mathbf{K}_n).$$

*Then there is $1 \leq i \leq n$ such that the $D_f$-projection $\mathbf{w}^{(i)}$ of $\mathbf{v}$ to $V_{\mathbf{K}_i}^L$ does not $f$-dominate the $D_f$-projection of any $\mathbf{u} \in \hat{\Delta}_L^{D_f}(\mathbf{K}_1, \dots, \mathbf{K}_n)$ to $V_{\mathbf{K}_i}^L$.*

*Proof.* Assume that $\mathbf{v} \in \hat{\Theta}_L^{D_f}(\mathbf{K}_1, \dots, \mathbf{K}_n)$ and

$$\mathbf{u} \in \hat{\Delta}_L^{D_f}(\mathbf{K}_1, \dots, \mathbf{K}_n) \subseteq \hat{\Theta}_L^{D_f}(\mathbf{K}_1, \dots, \mathbf{K}_n).$$

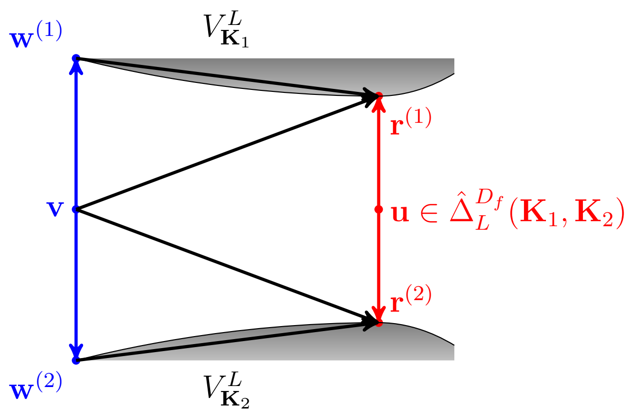


Figure 3.3: The situation in the proof of theorem 3.2.5 for $n = 2$.

Let us denote the $D_f$-projections of $\mathbf{v}$ to $V^L_{\mathbf{K}_1}, \dots, V^L_{\mathbf{K}_n}$ by $\mathbf{w}^{(1)} \dots, \mathbf{w}^{(n)}$ respectively. Accordingly let us denote the $D_f$-projections of $\mathbf{u}$ to $V^L_{\mathbf{K}_1}, \dots, V^L_{\mathbf{K}_n}$ by $\mathbf{r}^{(1)} \dots, \mathbf{r}^{(n)}$ respectively. We show that the assumption $\mathbf{w}^{(i)} \gg_f \mathbf{r}^{(i)}$ for all $1 \leq i \leq n$ leads to a contradiction.

First of all notice that since $\mathbf{v}, \mathbf{u} \in \hat{\Theta}^{D_f}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ then

$$\mathbf{v} = \mathbf{LinOp}(\mathbf{w}^{(1)} \dots, \mathbf{w}^{(n)}) \text{ and}$$

$$\mathbf{u} = \mathbf{LinOp}(\mathbf{r}^{(1)} \dots, \mathbf{r}^{(n)}).$$

Since $\mathbf{v} \notin \hat{\Delta}^{D_f}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$

$$\sum_{i=1}^{n} D_f(\mathbf{w}^{(i)} \| \mathbf{v}) > \sum_{i=1}^{n} D_f(\mathbf{r}^{(i)} \| \mathbf{u}).$$

Now by the theorem 2.1.20

$$\sum_{i=1}^{n} (\mathbf{r}^{(i)} - \mathbf{w}^{(i)}) \cdot ((\mathbf{w}^{(i)})^* - \mathbf{v}^*) < 0.$$

However, since $\mathbf{w}^{(1)} \dots, \mathbf{w}^{(n)}$ are $D_f$-projections of $\mathbf{v}$ to $V^L_{\mathbf{K}_1}, \dots, V^L_{\mathbf{K}_n}$ respectively, by the extended Pythagorean theorem (theorem 2.1.11)

$$\sum_{i=1}^{n} (\mathbf{r}^{(i)} - \mathbf{w}^{(i)}) \cdot ((\mathbf{w}^{(i)})^* - \mathbf{v}^*) \geq 0$$

which is a contradiction. $\square$

**Corollary 3.2.6.** [Of theorems 3.2.3 and 3.2.5.]

1. *If we restrict $CL$ to $BCL$ then for any $\mathbf{K}_1, \ldots, \mathbf{K}_n \in BCL$ and any convex Bregman divergence $D$*

$$\hat{\Delta}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \hat{\Theta}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

2. *If a convex Bregman divergence $D$ is such that $D(\mathbf{w}\|\mathbf{v}) \neq +\infty$ for all $\mathbf{w}, \mathbf{v} \in \mathbb{D}^L$ then for all $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$*

$$\hat{\Delta}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \hat{\Theta}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

*In particular for all $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ and $2 \geq r > 1$*

$$\hat{\Delta}_L^{D_r}(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \hat{\Theta}_L^{D_r}(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

This corollary says that the set $\hat{\Delta}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ in many cases represents the set of the fixed points $\hat{\Theta}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. In situations such as in example 3.2.4 this can be used as a justification of the $\hat{\Delta}^D$-merging operator.

The following example shows that the above restrictions are necessary:

$$\hat{\Theta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n) \setminus \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$$

can be nonempty and $\hat{\Theta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is not in general a convex set.

**Example 3.2.7.** Let $V_{\mathbf{K}_1}^L = \{\lambda(0, 0, \frac{1}{6}, \frac{5}{6}) + (1-\lambda)(0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}) : \lambda \in [0,1]\}$ and $V_{\mathbf{K}_2}^L = \{\lambda(0, 0, \frac{1}{3}, \frac{2}{3}) + (1-\lambda)(0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}) : \lambda \in [0,1]\}$. It is easy to check that $(0, 0, \frac{1}{4}, \frac{3}{4})$ and $(0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ are both fixed points but the former does not belong to the set $\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \mathbf{K}_2)$. The illustration is depicted in figure 3.4.

Now $\frac{1}{2}(0, 0, \frac{1}{4}, \frac{3}{4}) + \frac{1}{2}(0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}) = (0, \frac{1}{6}, \frac{7}{24}, \frac{13}{24})$ is in the convex hull of these two fixed points but we prove that it is not a fixed point itself and hence $\hat{\Theta}_L^{\mathrm{KL}}(\mathbf{K}_1, \mathbf{K}_2)$ is not a convex set. First of all notice that this point is equal to $\mathbf{LinOp}((0, \frac{1}{6}, \frac{1}{4}, \frac{7}{12}), (0, \frac{1}{6}, \frac{1}{3}, \frac{1}{2}))$ where $(0, \frac{1}{6}, \frac{1}{4}, \frac{7}{12}) \in V_{\mathbf{K}_1}^L$ and $(0, \frac{1}{6}, \frac{1}{3}, \frac{1}{2}) \in V_{\mathbf{K}_2}^L$ in both cases with the parameter $\lambda = \frac{1}{2}$. This is the only possible option if this point is to be an arithmetic mean of two points in $V_{\mathbf{K}_1}^L$ and $V_{\mathbf{K}_2}^L$ respectively.

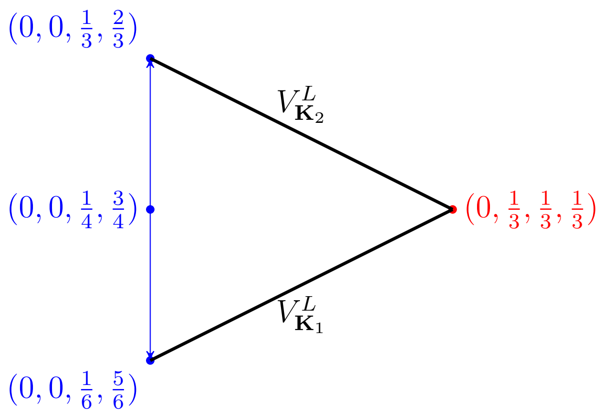


Figure 3.4: The illustration of example 3.2.7.

Now we verify whether the KL-projection of $(0, \frac{1}{6}, \frac{7}{24}, \frac{13}{24})$ to $V^L_{\mathbf{K}_2}$ is equal to $(0, \frac{1}{6}, \frac{1}{3}, \frac{1}{2})$. Consider $G(\lambda) = \frac{1}{3}(1-\lambda) \log \frac{\frac{1}{3}(1-\lambda)}{\frac{1}{6}} + \frac{1}{3} \log \frac{\frac{1}{3}}{\frac{7}{24}} + (\frac{1}{3}(1-\lambda) + \frac{2}{3}\lambda) \log \frac{\frac{1}{3}(1-\lambda)+\frac{2}{3}\lambda}{\frac{13}{24}}$. $\frac{\partial G(\lambda)}{\partial \lambda} = 0$ if

$$\frac{1}{3} \log(2(1-\lambda)) = \frac{1}{3} \log \frac{\frac{1}{3}(1-\lambda) + \frac{2}{3}\lambda}{\frac{13}{24}}$$

which holds when $\lambda = \frac{9}{17} \neq \frac{1}{2}$. Since $G(\lambda)$ is convex it follows that $(0, \frac{1}{6}, \frac{1}{3}, \frac{1}{2})$ is not the KL-projection of $(0, \frac{1}{6}, \frac{7}{24}, \frac{13}{24})$ to $V^L_{\mathbf{K}_2}$ and hence $(0, \frac{1}{6}, \frac{7}{24}, \frac{13}{24})$ can not be a fixed point. $\square$

Dually to what we have done in this section so far we obtain similar results when we relate the social entropy operator $\Delta^{\mathrm{KL}}$ to the conjugated KL-projection, provided that we confine ourselves to the class of weakly bounded knowledge bases $WBCL$.

**Lemma 3.2.8.** *Let $\mathbf{K}_1, \ldots, \mathbf{K}_n \in WBCL$. Suppose that $\mathbf{v} \in \Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in \Gamma^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ be such that $\mathbf{v} = \mathbf{LogOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$. Then $\mathbf{w}^{(1)} \in V^L_{\mathbf{K}_1}, \ldots, \mathbf{w}^{(n)} \in V^L_{\mathbf{K}_n}$ are the conjugated* KL*-projections of* $\mathbf{v}$ *into $V^L_{\mathbf{K}_1}, \ldots, V^L_{\mathbf{K}_n}$ respectively (which are unique).*

*Proof.* By definition we know that $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ globally minimise

$$\sum_{i=1}^{n} \mathrm{KL}(\mathbf{v} \| \mathbf{w}^{(i)}).$$

First of all, they exist with the above sum finite since whenever $\mathbf{K} \in WBCL$ then there is $\mathbf{w} \in V^L_{\mathbf{K}}$ such that $w_j \neq 0$ for all $j$ (by convexity of $V^L_{\mathbf{K}}$). Therefore, if $\mathbf{t}^{(i)}$ is the conjugated KL-projection of $\mathbf{v}$ into $V^L_{\mathbf{K}_i}$ then it cannot be that $t_j = 0$ while $v_j \neq 0$. By lemma 2.2.1 we have that for fixed $\mathbf{v}$ and for every $1 \leq i \leq n$ the sum $\sum_{j \in \mathrm{Sig}(\mathbf{x})} v_j \log \frac{v_j}{x_j}$ is a strictly convex function over the domain $\{\mathbf{x} : \ x_j > 0, \ 1 \leq j \leq J\}$ and therefore $\mathbf{t}^{(i)}$ coincides with $\mathbf{w}^{(i)}$ on $\mathrm{Sig}(\mathbf{v})$.

Now for any given $j$, $v_j = 0$ if and only if $w_j^{(1)} = 0, \ldots, w_j^{(n)} = 0$. This is proved in [53], theorem 3.6 (ii), and holds only under the assumption that for every $1 \leq i \leq n$ and every $1 \leq j \leq J = |\mathrm{At}(L)|$ there is $\mathbf{w} \in V^L_{\mathbf{K}_i}$ such that $w_j \neq 0$. (The restriction to $WBCL$ is therefore necessary to the proof.) This together with the fact that $\mathbf{t}^{(i)}$ coincides with $\mathbf{w}^{(i)}$ on $\mathrm{Sig}(\mathbf{v})$ gives that whenever $v_j = 0$ then $t_j^{(i)} = 0$ and the lemma is proved. $\square$

It follows that given $\mathbf{K}_1, \ldots, \mathbf{K}_n \in WBCL$ every $\mathbf{v} \in \Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is a *fixed point* of the *conjugated averaging projective procedure*

$$F^{\mathrm{KL}}_{[\mathbf{K}_1, \ldots, \mathbf{K}_n]} : \mathbb{D}^L \to \mathbb{D}^L,$$

defined by

$$F^{\mathrm{KL}}_{[\mathbf{K}_1, \ldots, \mathbf{K}_n]}(\mathbf{v}) = \mathbf{LogOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$$

where for all $1 \leq i \leq n$ $\mathbf{w}^{(i)}$ is the conjugated KL-projection of $\mathbf{v}$ into $V^L_{\mathbf{K}_i}$. This is well defined due to the restriction to $WBCL$ since then condition (1.4) holds.

If we denote the set of all fixed points of $F^{\mathrm{KL}}_{[\mathbf{K}_1, \ldots, \mathbf{K}_n]}$ by $\Theta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ then we have already established the following.

**Theorem 3.2.9.** *For all $n$-tuples of knowledge bases $\mathbf{K}_1, \ldots, \mathbf{K}_n \in WBCL$*

$$\Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \subseteq \Theta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

*In particular $\Theta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is always nonempty.*

If we restrict $WBCL$ to $BCL$ then we obtain:

**Theorem 3.2.10.** *For any $\mathbf{K}_1, \ldots, \mathbf{K}_n \in BCL$*

$$\Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \Theta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

*Proof.* The proof is similar to the proof of theorem 3.2.5. For a contradiction assume that $\mathbf{v} \in \Theta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \setminus \Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and

$$\mathbf{u} \in \Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \subseteq \Theta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

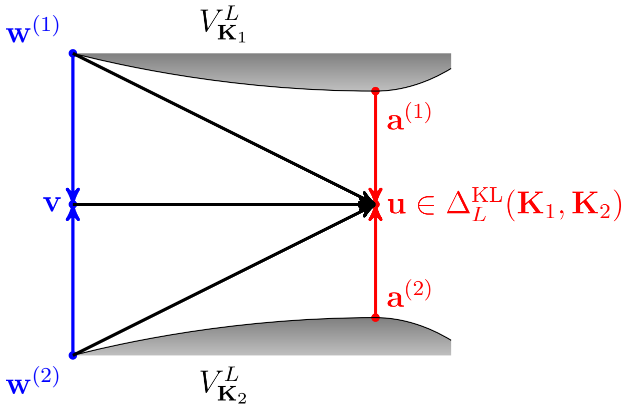


Figure 3.5: The situation in the proof of theorem 3.2.10 for $n = 2$.

Let us denote the conjugated KL-projections of $\mathbf{v}$ to $V^L_{\mathbf{K}_1}, \dots, V^L_{\mathbf{K}_n}$ by $\mathbf{w}^{(1)} \dots, \mathbf{w}^{(n)}$ respectively. Accordingly let us denote the conjugated KL-projections of $\mathbf{u}$ to $V^L_{\mathbf{K}_1}, \dots, V^L_{\mathbf{K}_n}$ by $\mathbf{a}^{(1)} \dots, \mathbf{a}^{(n)}$ respectively. By the assumption both $w_j \neq 0$ and $v_j \neq 0$ for all $1 \leq j \leq J$. Hence $\mathbf{w} \gg \mathbf{v}$ and $\mathbf{v} \gg \mathbf{w}$.

First of all notice that since $\mathbf{v}, \mathbf{u} \in \Theta^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ then

$$\mathbf{v} = \mathbf{LogOp}(\mathbf{w}^{(1)} \dots, \mathbf{w}^{(n)}) \text{ and}$$

$$\mathbf{u} = \mathbf{LogOp}(\mathbf{a}^{(1)} \dots, \mathbf{a}^{(n)}).$$

Since $\mathbf{v} \notin \Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$

$$\sum_{i=1}^{n} \mathrm{KL}(\mathbf{u} \| \mathbf{a}^{(i)}) < \sum_{i=1}^{n} \mathrm{KL}(\mathbf{v} \| \mathbf{w}^{(i)}). \tag{3.2}$$

By theorem 2.2.4 we have that

$$\sum_{i=1}^{n} \mathrm{KL}(\mathbf{u} \| \mathbf{w}^{(i)}) = \sum_{i=1}^{n} \mathrm{KL}(\mathbf{v} \| \mathbf{w}^{(i)}) + n \cdot \mathrm{KL}(\mathbf{u} \| \mathbf{v})$$

which by theorem 2.2.2 becomes

$$\sum_{i=1}^{n} \mathrm{KL}(\mathbf{u} \| \mathbf{a}^{(i)}) + n \cdot \mathrm{KL}(\mathbf{u} \| \mathbf{v}) \geq \sum_{i=1}^{n} \mathrm{KL}(\mathbf{v} \| \mathbf{w}^{(i)}) + n \cdot \mathrm{KL}(\mathbf{u} \| \mathbf{v})$$

and that contradicts (3.2). $\square$

The following lemma will be useful in the forthcoming proof.

**Lemma 3.2.11.** *Assume that we are given* $\mathbf{K} \in BCL$, $\mathbf{v}^{[i]} \in \mathbb{D}^L$, $i = 1, 2, \ldots$, *and* $\mathbf{w}^{[i]} \in \mathbb{D}^L$, $i = 1, 2, \ldots$, *such that* $v_j^{[i]} > 0$ *for all* $1 \leq j \leq J$ *and* $\mathbf{w}^{[i]}$ *is the* KL*-projection of* $\mathbf{v}^{[i]}$ *into* $V_{\mathbf{K}}^L$ *for all* $i = 1, 2, \ldots$. *Assume that* $\{\mathbf{v}^{[i]}\}_{i=1}^{\infty}$ *converges to* $\mathbf{v} \in \mathbb{D}^L$, *where* $v_j > 0$, $1 \leq j \leq J$, *and* $\{\mathbf{w}^{[i]}\}_{i=1}^{\infty}$ *converges to* $\mathbf{w} \in \mathbb{D}^L$. *Then* $\mathbf{w}$ *is the* KL*-projection of* $\mathbf{v}$ *into* $V_{\mathbf{K}}^L$.

*Proof.* For a contradiction assume that the KL-projection of $\mathbf{v}$ into $V_{\mathbf{K}}^L$ denoted by $\bar{\mathbf{w}}$ is distinct from $\mathbf{w}$. Then by the extended Pythagorean theorem $\mathrm{KL}(\mathbf{w}^{[i]} \| \mathbf{v}^{[i]}) + \mathrm{KL}(\bar{\mathbf{w}} \| \mathbf{w}^{[i]}) \leq \mathrm{KL}(\bar{\mathbf{w}} \| \mathbf{v}^{[i]})$. Due to the continuity of $\mathrm{KL}(\cdot \| \cdot)$ (we are confined to $BCL$)

$$\lim_{i \to \infty} \mathrm{KL}(\mathbf{w}^{[i]} \| \mathbf{v}^{[i]}) = \mathrm{KL}(\mathbf{w} \| \mathbf{v}),$$

$$\lim_{i \to \infty} \mathrm{KL}(\bar{\mathbf{w}} \| \mathbf{w}^{[i]}) = \mathrm{KL}(\bar{\mathbf{w}} \| \mathbf{w}) \text{ and}$$

$$\lim_{i \to \infty} \mathrm{KL}(\bar{\mathbf{w}} \| \mathbf{v}^{[i]}) = \mathrm{KL}(\bar{\mathbf{w}} \| \mathbf{v}).$$

Therefore $\mathrm{KL}(\mathbf{w} \| \mathbf{v}) + \mathrm{KL}(\bar{\mathbf{w}} \| \mathbf{w}) \leq \mathrm{KL}(\bar{\mathbf{w}} \| \mathbf{v})$ which contradicts the assumption that $\bar{\mathbf{w}}$ is the KL-projection of $\mathbf{v}$ into $V_{\mathbf{K}}^L$. $\square$

We conclude this section by stating remarkable results of Matúš [34][2] concerning convergence of sequences generated by the averaging projective procedures $\hat{F}^{\mathrm{KL}}$ and $F^{\mathrm{KL}}$. Matúš proved these results by applying the well known theorem of Csiszár and Tusnády, see [12], theorem 3.

**Theorem 3.2.12.** 1. *Let* $U = \{\mathbf{v} \in \mathbb{D}^L : \ v_j > 0, 1 \leq j \leq J\}$ *and* $\mathbf{K}_1, \ldots, \mathbf{K}_n \in BCL$. *Then for any* $\mathbf{v} \in U$ *the sequence*

$$\{\mathbf{v}^{[i]}\}_{i=0}^{\infty},$$

*where* $\mathbf{v}^{[0]} = \mathbf{v}$ *and* $\mathbf{v}^{[i+1]} = \hat{F}^{\mathrm{KL}}_{[\mathbf{K}_1, \ldots, \mathbf{K}_n]}(\mathbf{v}^{[i]})$, *converges to some probability function in* $\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$.

2. *Let* $\mathbf{K}_1, \ldots, \mathbf{K}_n \in BCL$. *For any* $\mathbf{v} \in U$, *where* $U$ *is defined above, the sequence*

$$\{\mathbf{v}^{[i]}\}_{i=0}^{\infty},$$

*where* $\mathbf{v}^{[0]} = \mathbf{v}$ *and* $\mathbf{v}^{[i+1]} = F^{\mathrm{KL}}_{[\mathbf{K}_1, \ldots, \mathbf{K}_n]}(\mathbf{v}^{[i]})$, *converges to some probability function in* $\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$.

---

[2]This is an unpublished manuscript, for a published reference to this theorem see [48].

*Proof for part 1.* With respect to our corollary 3.2.6 and with respect to the way we use this theorem later we only need to establish this theorem for $\mathbf{K}_1, \dots, \mathbf{K}_n \in BCL$. For a proof in full generality see [34].

Denote the KL-projections of $\mathbf{v}^{[i]}$ into $V^L_{\mathbf{K}_1}, \dots, V^L_{\mathbf{K}_n}$ by $\pi_1\mathbf{v}^{[i]}, \dots, \pi_n\mathbf{v}^{[i]}$ respectively. Then it is easy to observe (see theorem 2.1.13) that

$$\sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}^{[i]}\|\mathbf{v}^{[i]}) \geq \sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}^{[i]}\|\mathbf{v}^{[i+1]}) \geq \sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}^{[i+1]}\|\mathbf{v}^{[i+1]}),$$

for all $i = 1, 2, \dots$ . Due to the monotonicity of this sequence the limit $\lim_{i\to\infty}\sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}^{[i]}\|\mathbf{v}^{[i]})$ exists. Thanks to the compactness of $V^L_{\mathbf{K}_1}, \dots, V^L_{\mathbf{K}_n}$ the sequence $\{(\pi_1\mathbf{v}^{[i]}, \dots, \pi_n\mathbf{v}^{[i]}, \mathbf{v})\}_{i=1}^{\infty}$ has a convergent subsequence. Let us denote the limit of this subsequence $(\pi_1\mathbf{v}, \dots, \pi_n\mathbf{v}, \mathbf{v})$. Due to lemma 3.2.11 $\pi_k\mathbf{v}$ is really the KL-projection of $\mathbf{v}$ into $V^L_{\mathbf{K}_k}$ for all $1 \leq k \leq n$. Moreover

$$\lim_{i\to\infty}\sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}^{[i]}\|\mathbf{v}^{[i]}) = \sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}\|\mathbf{v}).$$

By theorem 2.2.3

$$\sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}\|\mathbf{v}^{[i]}) \leq \sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}\|\mathbf{v}) + \sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}\|\pi_k\mathbf{v}^{[i-1]}). \tag{3.3}$$

This is because $\mathbf{v}^{[i]} = \mathbf{LinOp}(\pi_1\mathbf{v}^{[i-1]}, \dots, \pi_n\mathbf{v}^{[i-1]})$. Moreover by the extended Pythagorean theorem

$$\sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}^{[i]}\|\mathbf{v}^{[i]}) + \sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}\|\pi_k\mathbf{v}^{[i]}) \leq \sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}\|\mathbf{v}^{[i]}). \tag{3.4}$$

An illustration of the situation is depicted in figure 3.6.

Now since $\lim_{i\to\infty}\sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}^{[i]}\|\mathbf{v}^{[i]}) = \sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}\|\mathbf{v})$ two equations (3.3) and (3.4) give that

$$\sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}\|\pi_k\mathbf{v}^{[i]}) \leq \sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}\|\pi_k\mathbf{v}^{[i-1]}) \tag{3.5}$$

for all $i = 1, 2, \dots$. We conclude that this is possible only if

$$\lim_{i\to\infty}\sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}\|\pi_k\mathbf{v}^{[i]})$$

exists.

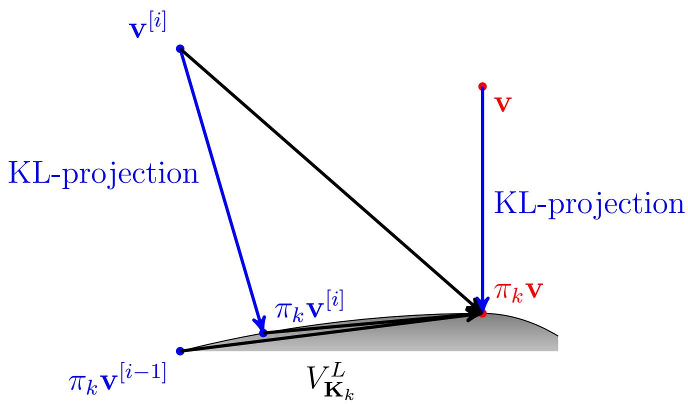


Figure 3.6: The situation in the proof of theorem 3.2.12, part 1.

But we already know that a subsequence of $\{(\pi_1\mathbf{v}^{[i]}, \dots, \pi_n\mathbf{v}^{[i]})\}_{i=1}^{\infty}$ converges to $(\pi_1\mathbf{v}, \dots, \pi_n\mathbf{v})$ hence a subsequence of the sequence $\{\sum_{k=1}^{n} \mathrm{KL}(\pi_k\mathbf{v}\|\pi_k\mathbf{v}^{[i]})\}_{i=1}^{\infty}$ decreases to zero which by (3.5) forces the whole sequence to converge to zero. Due to the fact that $\mathrm{KL}(\mathbf{x}\|\mathbf{y}) = 0$ only if $\mathbf{x} = \mathbf{y}$ and by the continuity we get

$$\lim_{i\to\infty} \pi_k\mathbf{v}^{[i]} = \pi_k\mathbf{v}.$$

It follows that $\lim_{i\to\infty}\mathbf{v}^{[i]}$ exists and is equal to $\mathbf{v}$. Moreover $\mathbf{v} = \lim_{i\to\infty}\mathbf{v}^{[i+1]} = \lim_{i\to\infty}\mathbf{LinOp}(\pi_1\mathbf{v}^{[i]}, \dots, \pi_n\mathbf{v}^{[i]}) = \mathbf{LinOp}(\pi_1\mathbf{v}, \dots, \pi_n\mathbf{v})$ and therefore $\mathbf{v}$ is a fixed point of the mapping $\hat{F}^{\mathrm{KL}}_{[\mathbf{K}_1,\dots,\mathbf{K}_n]}$ and by corollary 3.2.6 we have that $\mathbf{v} \in \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$.

The proof for part 2 is similar although some additional lemmas need to be proved. We omit the proof and refer to [34]. □

The problem of characterising both limits above more precisely remains open. On the other hand, the theorem above suggests a way to compute at least some points in $\Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ and $\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$, in particular the single points if these sets are singletons. Examples of how the theorem above can be used for a computation can be found in chapter 5.

Now consider the special case when knowledge bases $\mathbf{K}_1, \dots, \mathbf{K}_n \in CL$ are given by a partition $L_1 \cup \dots \cup L_n$ of the propositional language $L$ together with the marginal probability functions $\mathbf{w}^{(1)} \in \mathbb{D}^{L_1}, \dots, \mathbf{w}^{(n)} \in \mathbb{D}^{L_n}$ by the following:

$$V^L_{\mathbf{K}_i} = \{\mathbf{w} \in \mathbb{D}^L \colon \mathbf{w}|_{L_i} = \mathbf{w}^{(i)}\},$$

$1 \leq i \leq n$. Hence, each knowledge base is generated by a marginal probability function of the given language. Such knowledge bases are jointly consistent so let $W = \bigcap_{i=1}^{n} V_{\mathbf{K}_i}^{L}$.

In [34] the following is proved.

**Theorem 3.2.13.** *Let $\mathbf{K}_1, \ldots, \mathbf{K}_n$ be as above and bounded. Then the limit of the sequence*

$$\{\mathbf{v}^{[i]}\}_{i=0}^{\infty},$$

*where $\mathbf{v}^{[0]} \in \mathbb{D}^L$, $v_j > 0$ for $1 \leq j \leq J$, and $\mathbf{v}^{[i+1]} = F^{\mathrm{KL}}_{[\mathbf{K}_1,\ldots,\mathbf{K}_n]}(\mathbf{v}^{[i]})$, is in fact the* KL*-projection of $\mathbf{v}^{[0]}$ into $W$.*

However, as shown in [48], if the sequence is defined by means of the mapping $\hat{F}^{\mathrm{KL}}_{[\mathbf{K}_1,\ldots,\mathbf{K}_n]}$ then that limit is not the KL-projection of $\mathbf{v}^{[0]}$.

We note that the averaging projective procedure defined by means of the mapping $F^{\mathrm{KL}}_{[\mathbf{K}_1,\ldots,\mathbf{K}_n]}$ gives in the special case of knowledge bases generated by marginal probability functions the same answer as the well known *iterative projective fitting procedure* IPFP which we now define.

Let $L = L_0 \cup \ldots \cup L_{n-1}$ be a partition of a propositional language $L$ and let $\mathbf{w}^{(0)} \in \mathbb{D}^{L_0}, \ldots, \mathbf{w}^{(n-1)} \in \mathbb{D}^{L_{n-1}}$. IPFP is the sequence of $L$-probability functions $\mathbf{v}^{[0]}, \mathbf{v}^{[1]}, \ldots$ such that for all $i \in \mathbb{N}$ and all $\alpha \in \mathrm{At}(L)$ we have that

$$\mathbf{v}^{[i]}(\alpha) = \begin{cases} 0, & \text{if } \mathbf{v}^{[i-1]}|_{L_j}(\beta) = 0, \\ \mathbf{v}^{[i-1]}(\alpha)\frac{\mathbf{w}^{(j)}(\beta)}{\mathbf{v}^{[i-1]}|_{L_j}(\beta)}, & \text{otherwise,} \end{cases}$$

where $j = i (\mathrm{mod}\, n)$ and $\beta \in \mathrm{At}(L_j)$ is such that $\alpha \models \beta$. The convergence of this procedure was proved by Csiszár in [11]:

**Theorem 3.2.14.** *Let $L = L_0 \cup \ldots \cup L_{n-1}$ be a partition of a propositional language $L$, $\mathbf{w}^{(0)} \in \mathbb{D}^{L_0}, \ldots, \mathbf{w}^{(n-1)} \in \mathbb{D}^{L_{n-1}}$ and $\mathbf{v} \in \mathbb{D}^L$. Assume that $W = \{\mathbf{w} \in \mathbb{D}^L : \mathbf{w}|_{L_i} = \mathbf{w}^{(i)}, 0 \leq i \leq n-1\}$ and $\mathbf{v} \in \mathbb{D}^L$. If there is a probability function $\mathbf{w} \in W$ such that $\mathbf{v} \gg \mathbf{w}$ then* IPFP *converges for $\mathbf{v}^{[0]} := \mathbf{v}$ and the limit is equal to*

$$\lim_{i\to\infty} \mathbf{v}^{[i]} = \arg\min_{\mathbf{w}\in W} \mathrm{KL}(\mathbf{w}\|\mathbf{v}).$$

IPFP is popular for its computational speed and numerical stability. For a proof and more details on IPFP see [48].

The above theorem shows that IPFP can be effectively employed to find a KL-projection in example 3.2.1 whenever the KL-projection is the preferred method of updating. The reason why we have mentioned IPFP here is to demonstrate that for some problems KL-projections can be computed effectively. And computing KL-projections is the crucial part of the averaging projective procedure.

**Example 3.2.15.** Let $L = \{p\} \cup \{q\}$, $\mathbf{w}^{(1)}(p) = \frac{1}{3}$, $\mathbf{w}^{(2)}(q) = \frac{1}{4}$ and $\mathbf{v}^{[0]} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Then IPFP gives the sequence which is constant after two iterations:

$$\mathbf{v}^{[1]}(p \wedge q) = \frac{1}{4}\frac{\frac{1}{3}}{\frac{1}{2}} = \frac{1}{6}, \mathbf{v}^{[1]}(p \wedge \neg q) = \frac{1}{4}\frac{\frac{1}{3}}{\frac{1}{2}} = \frac{1}{6},$$

$$\mathbf{v}^{[1]}(\neg p \wedge q) = \frac{1}{4}\frac{\frac{2}{3}}{\frac{1}{2}} = \frac{2}{6}, \mathbf{v}^{[1]}(\neg p \wedge \neg q) = \frac{1}{4}\frac{\frac{2}{3}}{\frac{1}{2}} = \frac{2}{6},$$

and

$$\mathbf{v}^{[2]}(p \wedge q) = \frac{1}{6}\frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{12}, \mathbf{v}^{[2]}(p \wedge \neg q) = \frac{1}{6}\frac{\frac{3}{4}}{\frac{1}{2}} = \frac{3}{12},$$

$$\mathbf{v}^{[2]}(\neg p \wedge q) = \frac{2}{6}\frac{\frac{1}{4}}{\frac{1}{2}} = \frac{2}{12}, \mathbf{v}^{[2]}(\neg p \wedge \neg q) = \frac{2}{6}\frac{\frac{3}{4}}{\frac{1}{2}} = \frac{6}{12}.$$

It is easy to check that the KL-projection of $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ to $W = \{\mathbf{v} \in \mathbb{D}^L \colon \mathbf{v}|_{\{p\}} = \frac{1}{3}, \mathbf{v}|_{\{q\}} = \frac{1}{4}\}$ is in fact $\mathbf{ME}_L(W) = (\frac{1}{12}, \frac{3}{12}, \frac{2}{12}, \frac{6}{12})$, where the probability function is listed in the following order of atoms: $p \wedge q$, $p \wedge \neg q$, $\neg p \wedge q$ and $\neg p \wedge \neg q$. $\square$

In this chapter we have introduced several appealing p-merging operators and we have proved several technical results. In the next chapter we will go back and study some principles to be satisfied by these operators.

# Chapter 4

# Some principles for p-merging operators

## 4.1 Agreement and disagreement

This section is based on paper [2] by Wilmers and the author.

Konieczny and Pino Pérez in [30] proposed an axiomatic framework, referred to below as KPP, for expressing the desiderata required of a *non-probabilistic propositional* merging operator. Such an operator $\Delta$ acts on a multiset $K_1, \ldots, K_n$ of sets of knowledge bases to generate a single knowledge base. Each knowledge base $K_i$ is assumed to be consistent[1], but the union of two or more knowledge bases may not be consistent. The resulting *merged* knowledge base $\Delta(K_1, \ldots, K_n)$ should be consistent, and the operator $\Delta$ should at the minimum satisfy the principles listed below. In [30] the case was considered where a knowledge base is interpreted to mean a consistent set of sentences of a given finite propositional language $L$. However, as noted in [30], the general idea of a merging operator can easily be applied to other types of knowledge base, and there exists a large literature concerning such generalisations[2]. In this section we consider whether KPP can be applied to probabilistic merging.

In the KPP framework as formulated for consistent sets of sentences of a given finite propositional language $L$, a propositional merging operator $\Delta$ should satisfy the

[1] This is a slight restriction of the KPP formulation which is more appropriate for our present considerations.

[2] See [31] for a survey paper and bibliography.

following principles:

For every $n, m \geq 1$, every propositional language $L$ and consistent subsets $K_1, \ldots, K_n, F_1, \ldots, F_m \subseteq SL$

**(A1)** $\Delta(K_1, \ldots, K_n)$ is a nonempty consistent subset of $SL$,

**(A2)** if $K_1, \ldots, K_n$ and $F_1, \ldots, F_n$ are such that there exist a permutation $\sigma$ of the index set $\{1, \ldots, n\}$ such that $K_i$ is logically equivalent to $F_{\sigma(i)}$ for $1 \leq i \leq n$, then $\Delta(K_1, \ldots, K_n)$ is logically equivalent to $\Delta(F_1, \ldots, F_n)$,

**(A3)** if $K_1, \ldots, K_n$ are jointly consistent then $\Delta(K_1, \ldots, K_n)$ is logically equivalent to $\bigcup_{i=1}^{n} K_i$,

**(A4)** if $K_1$ and $F_1$ are jointly inconsistent then $\Delta(K_1, F_1) \not\models K_1$,

**(A5)** $\Delta(K_1, \ldots, K_n) \cup \Delta(F_1, \ldots, F_m) \models \Delta(K_1, \ldots, K_n, F_1, \ldots, F_m)$,

**(A6)** if $\Delta(K_1, \ldots, K_n) \cup \Delta(F_1, \ldots, F_m)$ is consistent then

$$\Delta(K_1, \ldots, K_n, F_1, \ldots, F_m) \models \Delta(K_1, \ldots, K_n) \cup \Delta(F_1, \ldots, F_m).$$

The idea, that these axioms of Konieczny and Pino-Pérez should be considered in relation to the merging of probabilistic information from different sources, was originally put forward by Williamson, see [49]. Notice that **(K1)** is a natural counterpart to **(A1)**; just as **(A1)** ensures that a propositional merging operator yields a consistent subset of $SL$, so **(K1)** ensures that a p-merging operator applied to a multiset of knowledge bases yields a knowledge base. Also **(K2)** corresponds to **(A2)** while **(K4*)** corresponds to **(A3)** which can be interpreted as saying that if the knowledge bases of a set of experts are collectively consistent then the merged knowledge base should simply consist of all the knowledge of the experts collected together.

On the other hand principles **(A4)**, **(A5)** and **(A6)** do not yet have probabilistic counterparts. Taking inspiration from them we formulate the following two principles for p-merging operators.

**(K6) Disagreement Principle.** Let $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ and $\mathbf{F}_1, \ldots, \mathbf{F}_m \in CL$. Assume that $\bigcap_{i=1}^{m} V_{\mathbf{F}_i}^{L} \neq \emptyset$. Then $\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \Delta_L(\mathbf{F}_1, \ldots, \mathbf{F}_m) = \emptyset$ implies that

$$\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m) \cap \Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \emptyset.$$

**(K6)** represents a significant but natural strengthening of **(A4)**, adapted to the p-merging context. Intuitively the principle says that if the merged knowledge base $\mathbf{K}$ of a set of experts is inconsistent with the merged knowledge $\mathbf{F}$ of a distinct set of experts, where the knowledge bases of the latter set are collectively consistent, then the result of merging the knowledge bases of all the experts together is also inconsistent with $\mathbf{K}$. Expressed more pithily, but less exactly, we could say that a coherent group who disagree with another group and then merge with them can be sure that they have influenced the opinions of the combined group.

**(K7) Agreement Principle.** Let $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ and $\mathbf{F}_1, \ldots, \mathbf{F}_m \in CL$. If $\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \Delta_L(\mathbf{F}_1, \ldots, \mathbf{F}_m) \neq \emptyset$ then

$$\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \Delta_L(\mathbf{F}_1, \ldots, \mathbf{F}_m) = \Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m).$$

**(K7)** combines the ideas of **(A5)** and **(A6)** into a single principle adapted to the probabilistic context. In particular **(K7)** implies that if each of two distinct sets of experts arrive at the same set of possible conclusions then the result of considering the knowledge bases of all the experts together should result in the same set of possible conclusions.

**Theorem 4.1.1.** *Whenever a p-merging operator $\Delta$ satisfies the strong consistency principle* **(K4*)** *and the agreement principle* **(K7)** *then it also satisfies the collegiality principle* **(K5)**.

*Proof.* Let $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ and $\mathbf{F}_1, \ldots, \mathbf{F}_m \in CL$ be such that

$$\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \subseteq \bigcap_{i=1}^{m} V^L_{\mathbf{F}_i} \neq \emptyset.$$

By **(K4*)** $\Delta_L(\mathbf{F}_1, \ldots, \mathbf{F}_m) = \bigcap_{i=1}^m V^L_{\mathbf{F}_i}$ so then $\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \Delta_L(\mathbf{F}_1, \ldots, \mathbf{F}_m) = \Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and by **(K7)**

$$\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m).$$

□

Our reformulation of the KPP principles into a probabilistic framework is a fairly straightforward translation. We should note however that whereas Williamson [49] previously advocated the relevance of the KPP principles in relation to the merging of knowledge bases, in a more recent paper [51] he rejects the KPP principles **(A3)**, **(A4)**, and **(A6)** as representing norms which are too strong to be applicable in this context. However in order to arrive at this conclusion Williamson uses a particular interpretation of the epistemological status of an expert's knowledge base, which he calls "granting". The point he makes is that several experts may grant the same piece of knowledge for inconsistent reasons. However, as granting violates our principle of total evidence we do not find Williamson's arguments against these principles to be applicable in our framework. Furthermore, as we will show later, some p-merging operators do in fact satisfy all the principles **(K1)**, **(K2)**, **(K3)**, **(K4*)** and satisfy **(K6)** and **(K7)** at least when their application is restricted to bounded knowledge bases $BCL$.

Williamson in [50] considered the following p-merging operator as possible. For any $n \geq 1$, any $L$ and $\mathbf{K}_1, \dots, \mathbf{K}_n \in CL$ let

$$W = \Big\{V\colon \exists I \subseteq \{1 \dots n\}\Big[V = \bigcap_{i \in I} V^L_{\mathbf{K}_i} \neq \emptyset \;\&\; (\forall j \in \{1 \dots n\} \backslash I)\Big(\bigcap_{i \in I} V^L_{\mathbf{K}_i}\Big) \cap V^L_{\mathbf{K}_j} = \emptyset\Big]\Big\}.$$

The *Convex Hull Operator* HULL is defined by

$$\mathrm{HULL}_L(\mathbf{K}_1, \dots, \mathbf{K}_n) = \Big\{\mathbf{v} \in \mathbb{D}^L\colon \exists \lambda_1, \dots, \lambda_{|W|} \in [0,1]$$

$$\exists \mathbf{w}^{(1)} \in V_1, \dots, \mathbf{w}^{(|W|)} \in V_{|W|} \,\Big(\mathbf{v} = \sum_{i=1}^{|W|} \lambda_i \mathbf{w}^{(i)} \;\&\; \sum_{i=1}^{|W|} \lambda_i = 1\Big)\Big\},$$

where $V_1, \dots, V_{|W|}$ is an enumeration of all sets in $W$ in some fixed order. In other words, the convex hull operator takes the convex hull of all the sets in $W$. Clearly $\mathrm{HULL}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ is closed convex and nonempty and hence it satisfies **(K1)**.

**Example 4.1.2.** Let $|L| = 1$, $V^L_{\mathbf{K}_1} = \{(x, 1-x), x \in [\frac{1}{4}, \frac{1}{2}]\}$, $V^L_{\mathbf{K}_2} = \{(x, 1-x), x \in [\frac{1}{2}, \frac{2}{3}]\}$ and $V^L_{\mathbf{K}_3} = \{(x, 1-x), x \in [\frac{3}{4}, 1]\}$ then $W = \Big\{\{(\frac{1}{2}, \frac{1}{2})\}, V^L_{\mathbf{K}_3}\Big\}$ and

$$\mathrm{HULL}_L(\mathbf{K}_1, \mathbf{K}_2, \mathbf{K}_3) = \Big\{(x, 1-x), x \in \Big[\frac{1}{2}, 1\Big]\Big\}. \qquad \square$$

**Theorem 4.1.3.** HULL *satisfies* **(K5)**.

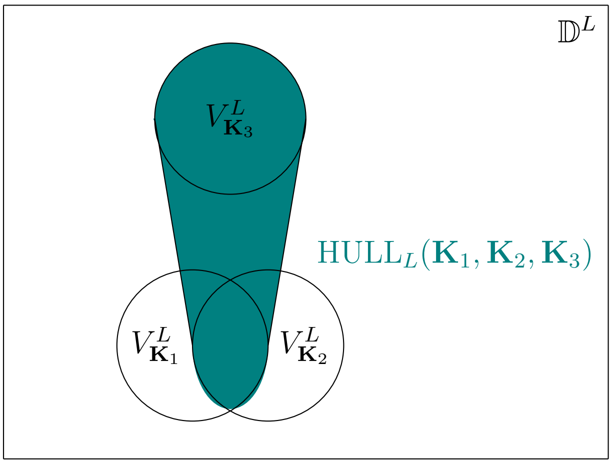


Figure 4.1: The illustration how the operator HULL works.

*Proof.* Let $\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m \in CL$. Assume that $\mathrm{HULL}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \subseteq \bigcap_{i=1}^m V^L_{\mathbf{F}_i} \neq \emptyset$. Now consider

$$W = \Big\{V\colon \exists I \subseteq \{1 \ldots n\} \Big[V = \bigcap_{i \in I} V^L_{\mathbf{K}_i} \neq \emptyset \;\&\; (\forall j \in \{1 \ldots n\} \setminus I) \Big(\bigcap_{i \in I} V^L_{\mathbf{K}_i}\Big) \cap V^L_{\mathbf{K}_j} = \emptyset\Big]\Big\}.$$

If $V \in W$ then

$$V \subseteq \mathrm{HULL}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \subseteq \bigcap_{i=1}^m V^L_{\mathbf{F}_i}.$$

Therefore $\mathrm{HULL}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m)$ is also the convex hull of all the sets from $W$ and hence $\mathrm{HULL}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \mathrm{HULL}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m)$. $\square$

HULL also obviously satisfies **(K2)**, **(K3)** and **(K4*)**. However it satisfies neither the principle of disagreement **(K6)** nor the principle of agreement **(K7)**. To see this consider the following example.

**Example 4.1.4.** Let $\mathbf{K}, \mathbf{F} \in CL$ be such that $V^L_{\mathbf{K}} \cap V^L_{\mathbf{F}} = \emptyset$. It follows that $\mathrm{HULL}_L(\mathbf{K}) \cap \mathrm{HULL}_L(\mathbf{F}) = \emptyset$ however $\mathrm{HULL}_L(\mathbf{K}) \subseteq \mathrm{HULL}_L(\mathbf{K}, \mathbf{F})$ and hence HULL does not satisfy the principle of disagreement. Moreover $\mathrm{HULL}_L(\mathbf{K}, \mathbf{F}) \cap \mathrm{HULL}_L(\mathbf{K}) \neq \emptyset$ while

$$\mathrm{HULL}_L(\mathbf{K}, \mathbf{F}) \cap \mathrm{HULL}_L(\mathbf{K}) \subsetneq \mathrm{HULL}_L(\mathbf{K}, \mathbf{F}, \mathbf{K})$$

and hence it does not satisfy the principle of agreement. $\square$

On the other hand it is not hard to see that obdurate merging operators based either on the **LinOp**-pooling operator or the **LogOp**-pooling operator (in particular

**OSEP** and **OLEP**) satisfy both **(K6)** and **(K7)**. In view of the fact that unlike **OSEP** and **OLEP** the p-merging operator HULL satisfies **(K4)** one might perhaps doubt that the linear entropy operator $\hat{\Delta}^{\mathrm{KL}}$ or the social entropy operator $\Delta^{\mathrm{KL}}$ which both satisfy **(K4)** also satisfy **(K6)** or **(K7)**. We will see however that once we confine ourselves to $BCL$ they do satisfy them. Before proceeding to that we introduce the following natural strengthening of the disagreement principle **(K6)**.

**(K6*) Strong Disagreement Principle.** Let $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ and $\mathbf{F}_1, \ldots, \mathbf{F}_m \in CL$. Then $\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \Delta_L(\mathbf{F}_1, \ldots, \mathbf{F}_m) = \emptyset$ implies that

$$\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m) \cap \Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \emptyset.$$

Trivially the strong disagreement principle implies the disagreement principle.

**Theorem 4.1.5.** *Let $D$ be a convex Bregman divergence, $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ and $\mathbf{F}_1, \ldots, \mathbf{F}_m \in BCL$. Then $\hat{\Delta}^D_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \hat{\Delta}^D_L(\mathbf{F}_1, \ldots, \mathbf{F}_m) = \emptyset$ implies*

$$\hat{\Delta}^D_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m) \cap \hat{\Delta}^D_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \emptyset.$$

*Proof.* Assume that $\mathbf{v} \in \hat{\Delta}^D_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and $\mathbf{v} \in \hat{\Delta}^D_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m)$. Let $(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}) \in \hat{\Gamma}^D_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ be an $n$-tuple associated with $\mathbf{v}$; in particular $\mathbf{v} = \mathbf{LinOp}(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)})$. Recall that the symbol $\hat{\Gamma}$ was defined in section 3.1. Let

$$(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}, \mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(m)}) \in \hat{\Gamma}^D_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m)$$

be an $(n+m)$-tuple associated with $\mathbf{v}$; then

$$\mathbf{v} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}, \mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(m)}).$$

This can only happen when

$$\mathbf{w}^{(i)} = \mathbf{v}^{(i)} \text{ for all } 1 \leq i \leq n$$

since the $D$-projections of the fixed $\mathbf{v}$ to each $V^L_{\mathbf{K}_i}$ are unique. Since in that case

$$\mathbf{v} = \mathbf{LinOp}(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)})$$

and

$$\mathbf{v} = \mathbf{LinOp}(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}, \mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(m)}).$$

we have that

$$\mathbf{v} = \mathbf{LinOp}(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(m)}).$$

It follows that $\mathbf{v} \in \hat{\Theta}^D_L(\mathbf{F}_1, \dots, \mathbf{F}_m)$. Since we have restricted $\mathbf{F}_1, \dots, \mathbf{F}_m$ to $BCL$ by corollary 3.2.6 $\mathbf{v} \in \hat{\Delta}^D_L(\mathbf{F}_1, \dots, \mathbf{F}_m)$ which concludes the proof. □

**Corollary 4.1.6.** *If the knowledge bases to which the linear entropy operator $\hat{\Delta}^{\mathrm{KL}}$ is applied are restricted to $BCL$ then $\hat{\Delta}^{\mathrm{KL}}$ satisfies* **(K6*)**.

*Proof.* Although theorem 4.1.5 holds with only $\mathbf{F}_1, \dots, \mathbf{F}_m$ restricted to $BCL$, to get the full principle of strong disagreement ('both ways') we need to restrict all knowledge bases to $BCL$. □

**Theorem 4.1.7.** *The linear Renyi operator $\hat{\Delta}^{D_r}$, for $2 \geq r > 1$, satisfies* **(K6*)**.

*Proof.* The proof is similar to the one for theorem 4.1.5. However, since in the case of the Renyi-B divergence $D_r$ the zero points cause no discontinuity, the strong disagreement principle holds without any restriction on the class $CL$. □

A closer look at the whole proof of the above corollary reveals that rather complicated geometrical properties of the $\hat{\Delta}^{\mathrm{KL}}$-merging operator can be expressed by something that everyone can understand. Something as simple as the principle that if someone disagrees with you then you should take this into an account.

The following counterexample shows that the assumption of the corollary above restricting knowledge bases to $BCL$ is necessary even if we reformulate it using the weaker disagreement principle **(K6)** in place of **(K6*)**.

**Example 4.1.8.** Assume that $|L| = 2$, $V^L_{\mathbf{K}_1} = \{(1,0,0,0)\}$, $V^L_{\mathbf{K}_2} = \{(0,1,0,0)\}$, $V^L_{\mathbf{F}_1} = \{(2x, 0, 1-2x, 0), x \in [0, \frac{1}{2}]\}$ and $V^L_{\mathbf{F}_2} = \{(0, 2x, 1-2x, 0), x \in [0, \frac{1}{2}]\}$. Clearly $\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \mathbf{K}_2) = \{(\frac{1}{2}, \frac{1}{2}, 0, 0)\}$ and $\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{F}_1, \mathbf{F}_2) = \{(0,0,1,0)\}$. Therefore $\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \mathbf{K}_2) \cap \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{F}_1, \mathbf{F}_2) = \emptyset$. Now we prove that also

$$\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \mathbf{K}_2, \mathbf{F}_1, \mathbf{F}_2) = \left\{\left(\frac{1}{2}, \frac{1}{2}, 0, 0\right)\right\},$$

i.e.

$$\left\{\left(\frac{1}{2}, \frac{1}{2}, 0, 0\right)\right\} = \{\arg\min_{\mathbf{v} \in \mathbb{D}^L} (\mathrm{KL}((1,0,0,0)\|\mathbf{v}) + \mathrm{KL}((0,1,0,0)\|\mathbf{v}) + \mathrm{KL}((\mathbf{w}^{(1)}\|\mathbf{v}) +$$
$$+ \mathrm{KL}(\mathbf{w}^{(2)}\|\mathbf{v})) : \mathbf{w}^{(1)} \in V^L_{\mathbf{F}_1}, \mathbf{w}^{(2)} \in V^L_{\mathbf{F}_2}\}.$$

Let

$$\mathbf{v}(x,y) = \mathbf{LinOp}((1,0,0,0),(0,1,0,0),(2x,0,1-2x,0),(0,2y,1-2y,0)) =$$
$$= \Big(\frac{x}{2}+\frac{1}{4},\frac{y}{2}+\frac{1}{4},\frac{1}{2}-\frac{x}{2}-\frac{y}{2},0\Big),$$

where $x \in [0,\frac{1}{2}]$ and $y \in [0,\frac{1}{2}]$. It is sufficient to prove that

$$2\cdot \mathrm{KL}\left(\Big(1,0,0,0\Big)\Big\|\Big(\frac{1}{2},\frac{1}{2},0,0\Big)\right) + 2\cdot \mathrm{KL}\left(\Big(0,1,0,0\Big)\Big\|\Big(\frac{1}{2},\frac{1}{2},0,0\Big)\right) <$$
$$< \mathrm{KL}((1,0,0,0)\|\mathbf{v}(x,y)) + \mathrm{KL}((0,1,0,0)\|\mathbf{v}(x,y)) + \mathrm{KL}((2x,0,1-2x,0)\|\mathbf{v}(x,y)) +$$
$$+ \mathrm{KL}((0,2y,1-2y,0)\|\mathbf{v}(x,y)) \tag{4.1}$$

for all $x \in [0,\frac{1}{2}]$ and $y \in [0,\frac{1}{2}]$ where either $x \neq \frac{1}{2}$ or $y \neq \frac{1}{2}$.

Assume that there are $x \in [0,\frac{1}{2}]$ and $y \in [0,\frac{1}{2}]$, where either $x \neq \frac{1}{2}$ or $y \neq \frac{1}{2}$, such that

$$2\cdot \mathrm{KL}\left(\Big(1,0,0,0\Big)\Big\|\Big(\frac{1}{2},\frac{1}{2},0,0\Big)\right) + 2\cdot \mathrm{KL}\left(\Big(0,1,0,0\Big)\Big\|\Big(\frac{1}{2},\frac{1}{2},0,0\Big)\right) \geq$$
$$\geq \mathrm{KL}((1,0,0,0)\|\mathbf{v}(x,y)) + \mathrm{KL}((0,1,0,0)\|\mathbf{v}(x,y)) + \mathrm{KL}((2x,0,1-2x,0)\|\mathbf{v}(x,y)) +$$
$$+ \mathrm{KL}((0,2y,1-2y,0)\|\mathbf{v}(x,y)).$$

Since $\mathrm{KL}(\mathbf{w}\|\mathbf{v})$ is a convex function whenever $\mathbf{v} \gg \mathbf{w}$, by symmetry and by the Jensen inequality we have that

$$4\cdot \mathrm{KL}\left(\Big(1,0,0,0\Big)\Big\|\Big(\frac{1}{2},\frac{1}{2},0,0\Big)\right) + 4\cdot \mathrm{KL}\left(\Big(0,1,0,0\Big)\Big\|\Big(\frac{1}{2},\frac{1}{2},0,0\Big)\right) \geq$$
$$\geq 2\cdot \mathrm{KL}\left(\Big(1,0,0,0\Big)\Big\|\mathbf{v}\Big(\frac{x+y}{2},\frac{y+x}{2}\Big)\right) + 2\cdot \mathrm{KL}\left(\Big(0,1,0,0\Big)\Big\|\mathbf{v}\Big(\frac{x+y}{2},\frac{y+x}{2}\Big)\right) +$$
$$+2\cdot \mathrm{KL}\left(\Big(2\frac{x+y}{2},0,1-2\frac{x+y}{2},0\Big)\Big\|\mathbf{v}\Big(\frac{x+y}{2},\frac{y+x}{2}\Big)\right) +$$
$$+2\cdot \mathrm{KL}\left(\Big(0,2\frac{y+x}{2},1-2\frac{y+x}{2},0\Big)\Big\|\mathbf{v}\Big(\frac{x+y}{2},\frac{y+x}{2}\Big)\right).$$

Therefore, for a contradiction, it is sufficient to show that (4.1) holds for $x = y$.

Notice that (4.1) is equivalent to

$$4\log 2 - \log\frac{1}{\frac{x}{2}+\frac{1}{4}} - \log\frac{1}{\frac{y}{2}+\frac{1}{4}} - 2x\log\frac{2x}{\frac{x}{2}+\frac{1}{4}} -$$

$$-2y\log\frac{2y}{\frac{y}{2}+\frac{1}{4}}-(1-2x)\log\frac{1-2x}{\frac{1}{2}-\frac{x}{2}-\frac{y}{2}}-(1-2y)\log\frac{1-2y}{\frac{1}{2}-\frac{x}{2}-\frac{y}{2}}<0.$$

For $x=y$ this becomes

$$-2\log\frac{1}{\frac{x}{2}+\frac{1}{4}}-4x\log\frac{2x}{\frac{x}{2}+\frac{1}{4}}+2\log 2+4x\log 2<0$$

which is further

$$(4x+2)\log\left(x+\frac{1}{2}\right)<4x\log(2x).$$

One can see that this inequality holds by considering the Jensen inequality

$$\frac{2x+1}{2}\log\frac{2x+1}{2}\leq\frac{2x\log(2x)+\log 1}{2}.$$

Since $x\log x$ is a strictly convex function the equality holds only for $x=\frac{1}{2}$. □

The following example shows that the social entropy operator $\Delta^{\mathrm{KL}}$ does not in general satisfy the principle of disagreement.

**Example 4.1.9.** Let $V^L_{\mathbf{K}}=\{(0,0,\frac{1}{3},\frac{2}{3})\}$ and $V^L_{\mathbf{F}}=\{(0,\frac{1}{3},\frac{2}{9},\frac{4}{9})\}$. Obviously $\Delta^{\mathrm{KL}}_L(\mathbf{K})\cap\Delta^{\mathrm{KL}}_L(\mathbf{F})=\emptyset$. However

$$\Delta^{\mathrm{KL}}_L(\mathbf{K},\mathbf{F})=\mathbf{LogOp}\left(\left(0,0,\frac{1}{3},\frac{2}{3}\right),\left(0,\frac{1}{3},\frac{2}{9},\frac{4}{9}\right)\right)=\left(0,0,\frac{1}{3},\frac{2}{3}\right).$$

□

On the other hand $\Delta^{\mathrm{KL}}$ does satisfy the principle of disagreement **(K6)** if the knowledge bases to which $\Delta^{\mathrm{KL}}$ is applied are restricted to $WBCL$. The proof is in [2]. If we restrict $WBCL$ further to $BCL$ then $\Delta^{\mathrm{KL}}$ also satisfies the strong principle of disagreement **(K6*)**. The following theorem actually states something a bit stronger.

**Theorem 4.1.10.** *Let* $\mathbf{K}_1,\dots,\mathbf{K}_n\in WBCL$ *and* $\mathbf{F}_1,\dots,\mathbf{F}_m\in BCL$. *Then*

$$\Delta^{\mathrm{KL}}_L(\mathbf{K}_1,\dots,\mathbf{K}_n)\cap\Delta^{\mathrm{KL}}_L(\mathbf{F}_1,\dots,\mathbf{F}_m)=\emptyset$$

*implies*

$$\Delta^{\mathrm{KL}}_L(\mathbf{K}_1,\dots,\mathbf{K}_n,\mathbf{F}_1,\dots,\mathbf{F}_m)\cap\Delta^{\mathrm{KL}}_L(\mathbf{K}_1,\dots,\mathbf{K}_n)=\emptyset.$$

*Proof.* Suppose that $\mathbf{K}_1,\dots,\mathbf{K}_n\in WBCL$ and $\mathbf{F}_1,\dots,\mathbf{F}_m\in BCL$ as above. Assume that for some fixed $\mathbf{v}$ we have that $\mathbf{v}\in\Delta^{\mathrm{KL}}_L(\mathbf{K}_1,\dots,\mathbf{K}_n)$ and $\mathbf{v}\in$

$\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \dots, \mathbf{K}_n, \mathbf{F}_1, \dots, \mathbf{F}_m)$. Let $(\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(n)}) \in \Gamma_L^{\mathrm{KL}}(\mathbf{K}_1, \dots, \mathbf{K}_n)$ be an $n$-tuple associated with $\mathbf{v}$. Then

$$\mathbf{v} = \mathbf{LogOp}(\mathbf{v}^{(1)} \dots, \mathbf{v}^{(n)}). \tag{4.2}$$

Similarly let $(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}, \mathbf{u}^{(1)}, \dots, \mathbf{u}^{(m)}) \in \Gamma_L^{\mathrm{KL}}(\mathbf{K}_1, \dots, \mathbf{K}_n, \mathbf{F}_1, \dots, \mathbf{F}_m)$ be associated with $\mathbf{v}$. Also

$$\mathbf{v} = \mathbf{LogOp}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}, \mathbf{u}^{(1)}, \dots, \mathbf{u}^{(m)}). \tag{4.3}$$

The above can only happen when for all $1 \leq j \leq J$, $J = |\operatorname{At}(L)|$,

$$w_j^{(i)} = v_j^{(i)} \text{ for all } 1 \leq i \leq n. \tag{4.4}$$

This is because by lemma 3.2.8 $\mathbf{w}^{(i)}$ and $\mathbf{v}^{(i)}$ are both conjugated KL-projections of $\mathbf{v}$ into $V_{\mathbf{K}_i}^L$, and such a projection is unique when we are confined to $WBCL$.

Equation (4.3) can be by (4.4) rewritten for all $j \in \operatorname{Sig}(\mathbf{v})$ as

$$v_j = \frac{\left[\frac{\sum_{j=1}^J [\prod_{i=1}^n v_j^{(i)}]^{\frac{1}{n}}}{\sum_{j=1}^J [\prod_{i=1}^n v_j^{(i)}]^{\frac{1}{n}}} (\prod_{i=1}^n v_j^{(i)})^{\frac{1}{n}} (\prod_{i=1}^m u_j^{(i)})^{\frac{1}{n}}\right]^{\frac{n}{m+n}}}{\sum_{j=1}^J [\prod_{i=1}^n v_j^{(i)} \prod_{i=1}^m u_j^{(i)}]^{\frac{1}{n+m}}}$$

and since by (4.2) $v_j = \frac{(\prod_{i=1}^n v_j^{(i)})^{\frac{1}{n}}}{\sum_{j=1}^J [\prod_{i=1}^n v_j^{(i)}]^{\frac{1}{n}}}$ we have that

$$v_j = \frac{[\prod_{i=1}^m u_j^{(i)}]^{\frac{1}{m}}}{\frac{(\sum_{j=1}^J [\prod_{i=1}^n v_j^{(i)} \prod_{i=1}^m u_j^{(i)}]^{\frac{1}{n+m}})^{\frac{m+n}{m}}}{(\sum_{j=1}^J [\prod_{i=1}^n v_j^{(i)}]^{\frac{1}{n}})^{\frac{n}{m}}}}. \tag{4.5}$$

In order to obtain (4.5) above the cancelation of a term $[\prod_{i=1}^n v_j^{(i)}]^{\frac{1}{n}}$ is required; however this is permissible since we know by the discussion preceding theorem 3.2.9 that this term is non-zero for $j \in \operatorname{Sig}(\mathbf{v})$. For similar reasons the denominator on the right of (4.5) is finite and non-zero. On the other hand (4.5) holds even if $j \notin \operatorname{Sig}(\mathbf{v})$ in which case both sides are equal to zero.

Notice that the denominator of (4.5) is independent of $j$ so

$$\mathbf{v} = \mathbf{LogOp}(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(m)}).$$

It follows that $\mathbf{v} \in \Theta_L^{\mathrm{KL}}(\mathbf{F}_1, \dots, \mathbf{F}_m)$. Since we have restricted $\mathbf{F}_1, \dots, \mathbf{F}_m$ to $BCL$ by theorem 3.2.10 $\mathbf{v} \in \Delta_L^{\mathrm{KL}}(\mathbf{F}_1, \dots, \mathbf{F}_m)$ which concludes the proof. $\square$

**Theorem 4.1.11.**

*(i) For any convex Bregman divergence $D$ the $\hat{\Delta}^D$-operator satisfies* **(K7)**.

*(ii) The social entropy operator $\Delta^{\mathrm{KL}}$ satisfies* **(K7)** *for all knowledge bases in $WBCL$.*

*Proof.* (i) Since we are assuming that $\hat{\Delta}^D_L(\mathbf{K}_1,\ldots,\mathbf{K}_n)\cap\hat{\Delta}^D_L(\mathbf{F}_1,\ldots,\mathbf{F}_m)\neq\emptyset$, there is some $\mathbf{v}\in\hat{\Delta}^D_L(\mathbf{K}_1,\ldots,\mathbf{K}_n)\cap\hat{\Delta}^D_L(\mathbf{F}_1,\ldots,\mathbf{F}_m)$. For any such $\mathbf{v}$ this is equivalent to the assertion that for some $(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)})\in\hat{\Gamma}^D_L(\mathbf{K}_1,\ldots,\mathbf{K}_n)$ and some $(\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(m)})\in\hat{\Gamma}^D_L(\mathbf{F}_1,\ldots,\mathbf{F}_m)$

$$\sum_{i=1}^{n} D(\mathbf{w}^{(i)}\|\mathbf{v})=\hat{\mathrm{M}}^D_L(\mathbf{K}_1,\ldots,\mathbf{K}_n)\ \text{ and }\ \sum_{i=1}^{m} D(\mathbf{u}^{(i)}\|\mathbf{v})=\hat{\mathrm{M}}^D_L(\mathbf{F}_1,\ldots,\mathbf{F}_m).$$

Recall that $\hat{\mathrm{M}}^D_L(\mathbf{K}_1,\ldots,\mathbf{K}_n)$ was in section 3.1 defined as the minimal value of the sum $\sum_{i=1}^n D(\mathbf{w}^{(i)}\|\mathbf{v})$ subject only to $\mathbf{w}^{(1)}\in V^L_{\mathbf{K}_1},\ldots,\mathbf{w}^{(n)}\in V^L_{\mathbf{K}_n}$ and $\mathbf{v}\in\mathbb{D}^L$.

Then by definition

$$\hat{\mathrm{M}}^D_L(\mathbf{K}_1,\ldots,\mathbf{K}_n)+\hat{\mathrm{M}}^D_L(\mathbf{F}_1,\ldots,\mathbf{F}_m)\ \leq\ \hat{\mathrm{M}}^D_L(\mathbf{K}_1,\ldots,\mathbf{K}_n,\mathbf{F}_1,\ldots,\mathbf{F}_m)$$

and the same vectors $\mathbf{v}$, $\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)}$, $\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(m)}$ globally minimise the sum

$$\sum_{i=1}^{n} D(\mathbf{w}^{(i)}\|\mathbf{v})+\sum_{i=1}^{m} D(\mathbf{u}^{(i)}\|\mathbf{v})$$

subject to $\mathbf{w}^{(i)}\in V^L_{\mathbf{K}_i}$, $1\leq i\leq n$ and $\mathbf{u}^{(i)}\in V^L_{\mathbf{F}_i}$, $1\leq i\leq m$. Thus $\mathbf{v}\in\hat{\Delta}^D_L(\mathbf{K}_1,\ldots,\mathbf{K}_n,\mathbf{F}_1,\ldots,\mathbf{F}_m)$, and

$$\hat{\mathrm{M}}^D_L(\mathbf{K}_1,\ldots,\mathbf{K}_n)+\hat{\mathrm{M}}^D_L(\mathbf{F}_1,\ldots,\mathbf{F}_m)\ =\ \hat{\mathrm{M}}^D_L(\mathbf{K}_1,\ldots,\mathbf{K}_n,\mathbf{F}_1,\ldots,\mathbf{F}_m). \tag{4.6}$$

Since $\mathbf{v}$ was arbitrary we have proved that

$$\hat{\Delta}^D_L(\mathbf{K}_1,\ldots,\mathbf{K}_n)\cap\hat{\Delta}^D_L(\mathbf{F}_1,\ldots,\mathbf{F}_m)\ \subseteq\ \hat{\Delta}^D_L(\mathbf{K}_1,\ldots,\mathbf{K}_n,\mathbf{F}_1,\ldots,\mathbf{F}_m).$$

Now suppose $\mathbf{x}\in\hat{\Delta}^D_L(\mathbf{K}_1,\ldots,\mathbf{K}_n,\mathbf{F}_1,\ldots,\mathbf{F}_m)$. Then for some $(\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(n)},\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(m)})\in\hat{\Gamma}^D_L(\mathbf{K}_1,\ldots,\mathbf{K}_n,\mathbf{F}_1,\ldots,\mathbf{F}_m)$ and

$$\sum_{i=1}^{n} D(\mathbf{y}^{(i)}\|\mathbf{x})+\sum_{i=1}^{m} D(\mathbf{z}^{(i)}\|\mathbf{x})=\hat{\mathrm{M}}^D_L(\mathbf{K}_1,\ldots,\mathbf{K}_n,\mathbf{F}_1,\ldots,\mathbf{F}_m).$$

In view of (4.6) if we did not now have that $\sum_{i=1}^{n} D(\mathbf{y}^{(i)} \| \mathbf{x}) = \hat{\mathrm{M}}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and $\sum_{i=1}^{m} D(\mathbf{z}^{(i)} \| \mathbf{x}) = \hat{\mathrm{M}}_L^D(\mathbf{F}_1, \ldots, \mathbf{F}_m)$ then this would contradict the minimality of either $\hat{\mathrm{M}}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ or $\hat{\mathrm{M}}_L^D(\mathbf{F}_1, \ldots, \mathbf{F}_m)$. Hence $\mathbf{x} \in \hat{\Delta}_L^D(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \hat{\Delta}_L^D(\mathbf{F}_1, \ldots, \mathbf{F}_m)$ and the result is proved.

(ii) The proof for $\Delta^{\mathrm{KL}}$ is similar except that the final argument involving the equation corresponding to (4.6) fails if either of the quantities $\mathrm{M}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ or $\mathrm{M}_L^{\mathrm{KL}}(\mathbf{F}_1, \ldots, \mathbf{F}_m)$ is $+\infty$, which is the reason for the restriction of knowledge bases to $WBCL$ (which implies condition (1.4)). □

The fact that the p-merging operator $\hat{\Delta}^D$, where $D$ is a convex Bregman divergence, satisfies the collegiality principle **(K5)** is a direct consequence of the result above together with theorem 4.1.1. Similarly, in the case when we confine ourselves to $WBCL$, the same argument shows that the social entropy operator $\Delta^{\mathrm{KL}}$ satisfies the collegiality principle **(K5)**. However the following argument shows that the restriction to $WBCL$ is not necessary in this last case.

**Lemma 4.1.12.** *The social entropy operator $\Delta^{\mathrm{KL}}$ satisfies the collegiality principle* **(K5)**.

*Proof.* Let $\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m \in CL$ are such that $\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n) \subseteq \bigcap_{i=1}^{m} V_{\mathbf{F}_i}^L \neq \emptyset$. We need to prove that

$$\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m).$$

Since $\bigcap_{i=1}^{m} V_{\mathbf{F}_i}^L \neq \emptyset$ we have that $\mathrm{M}_L^{\mathrm{KL}}(\mathbf{F}_1, \ldots, \mathbf{F}_m) = 0 \neq +\infty$. Given the result above, it is sufficient to consider the case when $\mathrm{M}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is $+\infty$. However in this case both $\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and $\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m)$ are equal to $\mathbb{D}^L$. □

The principles of agreement and disagreement are variants of principles which appeared naturally in epistemology of merging as developed by Konieczny, Pino-Pérez, Williamson and others. The reason why we investigate them is because if a merging operator satisfies them, then for a consumer of an expert system based on such a merging operator it is easy to understand the recommendations of the expert system in the same way as we would understand the argumentation of another human. The beauty of this idea lies in the shift from a black-box expert system to a transparent one with a human-like argumentation.

## 4.2 Language invariance and Irrelevant Information

This section extends the ideas first introduced by Wilmers and the author in [1], which are themselves generalisations of classical notions of Paris and Vencovská in [39] and [41].

An obvious question we need to ask regarding probabilistic merging operators is whether they depend on the choice of a particular propositional language $L = \{a_1, \ldots, a_h\}$. For a fixed p-merging operator $\Delta$, language $L$, $\varphi \in SL$ and $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ consider $\mathbf{w}(\varphi)$ where $\mathbf{w} \in \Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. 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. This motivates the following principle.

**Language Invariance Principle (LI).** A probabilistic merging operator $\Delta$ 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

1. for any $\mathbf{w} \in \Delta_{L_2}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ we have that $\mathbf{w}|_{L_1} \in \Delta_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$, and
2. for any $\mathbf{v} \in \Delta_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ there is a $\mathbf{w}$ such that $\mathbf{w}|_{L_1} = \mathbf{v}$ and $\mathbf{w} \in \Delta_{L_2}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$.

Recall that if $L_1 \subseteq L_2$ and $\Delta$ is a set of probability functions we write $\Delta|_{L_1}$ to denote the set $\{\mathbf{w}|_{L_1} \colon \mathbf{w} \in \Delta\}$. Hence if the two conditions above hold we write

$$\Delta_{L_2}(\mathbf{K}_1, \ldots, \mathbf{K}_n)|_{L_1} = \Delta_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

The notion of language invariance was first defined by Paris for inference processes. In [39] he proved several results concerning language invariance and inference processes. In particular, both **ME** and $\mathbf{CM}^\infty$ satisfy the obvious reformulation of **LI** for an inference process $\mathcal{S}$: for any languages $L_1 \subseteq L_2$ and any knowledge base $\mathbf{K} \in CL_1$ we have that $\mathcal{S}_{L_2}(\mathbf{K})|_{L_1} = \mathcal{S}_{L_1}(\mathbf{K})$.

The language invariance principle appears to be a crucial principle in any framework — how could we possibly trust an expert system which changes its answers by merely expanding the language describing the problem? An inaccurate but nicely illustrative analogy would be if the validity of mathematical theorems depended on the unused symbols in which they are written in. It is perhaps surprising how overlooked this principle is in the literature, in particular in the field of computer science.

Now we ask a stronger 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? It would seem to be irrational if a probabilistic merging operator applied to such extended knowledge produced a probability function which, when restricted to the original language, would not be produced by the probabilistic merging operator applied to the original knowledge. This leads us to the:

**Irrelevant Information Principle (IIP).**

A probabilistic merging operator $\Delta$ satisfies the irrelevant information principle if whenever $L = L_1 \cup L_2$ are such that $L_1$ and $L_2$ are disjoint propositional languages, and $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL_1$ and $\mathbf{F}_1, \ldots, \mathbf{F}_n \in CL_2$, then

$$\Delta_L(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)|_{L_1} = \Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)|_{L_1}.$$

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

$$\Delta_L(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)|_{L_1} = \Delta_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

Both probabilistic merging operators $\hat{\Delta}^{\mathrm{KL}}$ and $\Delta^{\mathrm{KL}}$ are language invariant, as we show later in the corollaries 4.2.7 and 4.2.9 respectively. On the other hand the next two examples shows that neither of them satisfies **IIP**.

In this section, and particularly in the following example, we will adopt the following useful notation. For any $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ satisfying condition (1.4) the expression $C_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ will denote the maximum value of

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

subject to $\mathbf{w}^{(k)} \in V_{\mathbf{K}_k}^L$, for all $1 \le k \le n$. Recall that any n-tuple $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ which maximises the above sum belongs to $\Gamma_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and so generates a probability function

$$\mathbf{LogOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in \Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

And conversely, any n-tuple $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in \Gamma_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ maximises the above sum.

The following example shows that the social entropy operator $\Delta^{\mathrm{KL}}$ does not satisfy **IIP**.

**Example 4.2.1.** Let $L_1 = \{a_1\}$, $L_2 = \{a_2\}$ and $L = L_1 \cup L_2$. In the following we consider just two experts. Assume that the first expert possesses knowledge $\mathbf{K}_1 = \{Bel(a_1) = 0.2\}$, $\mathbf{F}_1 = \{Bel(a_2) = 0.2\}$ and the second has knowledge $\mathbf{K}_2 = \{Bel(a_1) = 0.3\}$, $\mathbf{F}_2 = \{Bel(a_2) = 0.4\}$. Suppose that $\mathbf{w}^{(1)} \in V^{L_1}_{\mathbf{K}_1 \cup \mathbf{F}_1}$ and $\mathbf{w}^{(2)} \in V^{L_2}_{\mathbf{K}_2 \cup \mathbf{F}_2}$. We can identify $C_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \mathbf{K}_2 \cup \mathbf{F}_2)$ by maximizing the following expression for parameters $\mathbf{w}^{(1)}(a_1 \wedge a_2) = x \in [0, 0.2]$ and $\mathbf{w}^{(2)}(a_1 \wedge a_2) = y \in [0, 0.3]$:

$$\begin{aligned} \mathrm{M}(x, y) = \quad & \sqrt{xy} + \sqrt{(0.2 - x)(0.3 - y)} + \sqrt{(0.2 - x)(0.4 - y)} + \\ & + \sqrt{(0.6 + x)(0.3 + y)}. \end{aligned}$$

It is the matter of elementary analysis to prove that the above is strictly maximal for $x = 0.12$ and $y = 0.24$. It follows that there is only one point $\mathbf{w} \in \Delta_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \mathbf{K}_2 \cup \mathbf{F}_2)$. Since

$$C_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \mathbf{K}_2 \cup \mathbf{F}_2) = \mathrm{M}(0.12, 0.24) = (\sqrt{0.12 \cdot 0.24} + \sqrt{0.08 \cdot 0.16}) +$$

$$+(\sqrt{0.08 \cdot 0.06} + \sqrt{0.72 \cdot 0.54}) = \sqrt{0.08} + \sqrt{0.48}$$

we evaluate

$$\mathbf{w}|_{L_1}(a_1) = \frac{\sqrt{0.12 \cdot 0.24} + \sqrt{0.08 \cdot 0.06}}{\sqrt{0.08} + \sqrt{0.48}}. \tag{4.7}$$

There is also only one point $\mathbf{v} \in \Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \mathbf{K}_2)|_{L_1}$. Notice that

$$\mathbf{v}(a_1) = \frac{\sqrt{0.06}}{\sqrt{0.06} + \sqrt{0.56}}. \tag{4.8}$$

Since the quantities (4.7) and (4.8) are not equal it follows that the social entropy operator $\Delta^{\mathrm{KL}}$ does not satisfy **IIP**. □

We now give an example to show that the $\hat{\Delta}^{\mathrm{KL}}$-operator does not satisfy **IIP**.

**Example 4.2.2.** Now consider $L_1 = \{p, q\}$, $L_2 = \{r\}$ and $L = L_1 \cup L_2$. We put $\mathbf{K}_1 = \{Bel(p) = 0.4, Bel(q) = 0.7\}$, $\mathbf{F}_1 = \{Bel(r) = 0\}$ and $\mathbf{K}_2 = \{Bel(p \wedge q) = 0, Bel(p \wedge \neg q) = 0.3, Bel(\neg p \wedge q) = 0.5, Bel(\neg p \wedge \neg q) = 0.2\}$, $\mathbf{F}_2 = \{Bel(r) = 1\}$.

First we show that there is only one point in $\hat{\Delta}^{\mathrm{KL}}_{L_1}(\mathbf{K}_1, \mathbf{K}_2)$.

$$V^{L_1}_{\mathbf{K}_1} = \{\mathbf{t} \in \mathbb{D}^L \colon\ (\mathbf{t}(p \wedge q), \mathbf{t}(p \wedge \neg q), \mathbf{t}(\neg p \wedge q), \mathbf{t}(\neg p \wedge \neg q)) =$$

$$= (x, 0.4 - x, 0.7 - x, x - 0.1),\ x \in [0.1, 0.4]\}.$$

Using the same order of atoms as above

$$V^{L_2}_{\mathbf{K}_2} = \{(0, 0.3, 0.5, 0.2)\}.$$

Now

$$\hat{\Delta}^{\mathrm{KL}}_{L_1}(\mathbf{K}_1, \mathbf{K}_2) = \Big\{ \arg\min_{\mathbf{v} \in \mathbb{D}^L} \Big( x \log \frac{x}{\frac{x}{2}} + (0.4 - x) \log \frac{0.4 - x}{\frac{0.4 - x + 0.3}{2}} +$$

$$+ (0.7 - x) \log \frac{0.7 - x}{\frac{0.7 - x + 0.5}{2}} + (x - 0.1) \log \frac{x - 0.1}{\frac{x - 0.1 + 0.2}{2}} + 0.3 \log \frac{0.3}{\frac{0.3 + 0.4 - x}{2}} + 0.5 \log \frac{0.5}{\frac{0.5 + 0.7 - x}{2}} +$$

$$+ 0.2 \log \frac{0.2}{\frac{0.2 - 0.1 + x}{2}} \Big), \mathbf{v} = \Big( \frac{x}{2}, \frac{0.4 - x + 0.3}{2}, \frac{0.7 - x + 0.5}{2}, \frac{x - 0.1 + 0.2}{2} \Big) \Big\}.$$

After some algebraic manipulation we get that the term to minimise is equivalent to

$$x \log x + (0.4 - x) \log (0.4 - x) + (0.7 - x) \log (0.7 - x) +$$

$$+ (x - 0.1) \log (x - 0.1) - x \log x - (0.7 - x) \log (0.7 - x) -$$

$$- (1.2 - x) \log (1.2 - x) - (x + 0.1) \log (x + 0.1).$$

By the first derivative test the critical points satisfy

$$\frac{(1.2 - x)(x - 0.1)}{(0.1 + x)(0.4 - x)} = 1$$

which gives us the only critical point for $x = 0.16$. By the second derivative test this continuous function has the global minimum at this point. This concludes our proof that there is only one point in $\hat{\Delta}^{\mathrm{KL}}_{L_1}(\mathbf{K}_1, \mathbf{K}_2)$.

On the other hand there are many points in the set $\hat{\Delta}^{\mathrm{KL}}_{L}(\mathbf{K}_1 \cup \mathbf{F}_1, \mathbf{K}_2 \cup \mathbf{F}_2)|_{L_1}$. If $\mathbf{t} \in V^{L}_{\mathbf{K}_1 \cup \mathbf{F}_1}$ then $\mathbf{t}(p \wedge q \wedge r) = \mathbf{t}(p \wedge \neg q \wedge r) = \mathbf{t}(\neg p \wedge q \wedge r) = \mathbf{t}(\neg p \wedge \neg q \wedge r) = 0$, $\mathbf{t}(p \wedge q \wedge \neg r) = x$, $\mathbf{t}(p \wedge \neg q \wedge \neg r) = 0.4 - x$, $\mathbf{t}(\neg p \wedge q \wedge \neg r) = 0.7 - x$ and $\mathbf{t}(\neg p \wedge \neg q \wedge \neg r) = x - 0.1$ for some $x \in [0.1, 0.4]$. There is only one $\mathbf{w} \in V^{L}_{\mathbf{K}_2 \cup \mathbf{F}_2}$; $\mathbf{w}(p \wedge q \wedge r) = 0$, $\mathbf{w}(p \wedge \neg q \wedge r) = 0.3$, $\mathbf{w}(\neg p \wedge q \wedge r) = 0.5$, $\mathbf{w}(\neg p \wedge \neg q \wedge r) = 0.2$ and $\mathbf{w}(p \wedge q \wedge \neg r) = \mathbf{w}(p \wedge \neg q \wedge \neg r) = \mathbf{w}(\neg p \wedge q \wedge \neg r) = \mathbf{w}(\neg p \wedge \neg q \wedge \neg r) = 0$. By lemma 3.1.1 if $\mathbf{v} \in \hat{\Delta}^{\mathrm{KL}}_{L}(\mathbf{K}_1 \cup \mathbf{F}_1, \mathbf{K}_2 \cup \mathbf{F}_2)$ then $\mathbf{v} = \mathbf{LinOp}(\mathbf{t}, \mathbf{w})$. Hence, regardless on the value of the parameter $x$, the sum of KL-divergences from the definition of the $\hat{\Delta}^{\mathrm{KL}}$-merging operator is

$$x\log\frac{x}{\frac{x}{2}}+(0.4-x)\log\frac{0.4-x}{\frac{0.4-x}{2}}+(0.7-x)\log\frac{0.7-x}{\frac{0.7-x}{2}}+(x-0.1)\log\frac{x-0.1}{\frac{x-0.1}{2}}+$$
$$+0.3\log\frac{0.3}{\frac{0.3}{2}}+0.5\log\frac{0.5}{\frac{0.5}{2}}+0.2\log\frac{0.2}{\frac{0.2}{2}}=2\log 2.$$

Therefore for each $x\in[0.1,0.4]$ we have that $\mathbf{LinOp}(\mathbf{t},\mathbf{w})\in\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1\cup\mathbf{F}_1,\mathbf{K}_2\cup\mathbf{F}_2)$.

Clearly

$$\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1\cup\mathbf{F}_1,\mathbf{K}_2\cup\mathbf{F}_2)|_{L_1}\neq\hat{\Delta}^{\mathrm{KL}}_{L_1}(\mathbf{K}_1,\mathbf{K}_2)$$

and, since we will later in corollary 4.2.9 prove that $\hat{\Delta}^{\mathrm{KL}}$ satisfies **LI**, we can conclude that the linear entropy operator $\hat{\Delta}^{\mathrm{KL}}$ does not satisfy **IIP**. □

Moreover, the convex hull operator HULL also does not satisfy **IIP** as the next example shows.

**Example 4.2.3.** Let $L=L_1\cup L_2$ and $L_1\cap L_2=\emptyset$. Let $\mathbf{K}_1,\mathbf{K}_2\in CL_1$ and $\mathbf{F}_1,\mathbf{F}_2\in CL_2$ be such that $V^{L_1}_{\mathbf{K}_1}\cap V^{L_1}_{\mathbf{K}_2}\neq\emptyset$, $V^{L_1}_{\mathbf{K}_1}\cap V^{L_1}_{\mathbf{K}_2}\neq V^{L_1}_{\mathbf{K}_1}$ and $V^{L_2}_{\mathbf{F}_1}\cap V^{L_2}_{\mathbf{F}_2}=\emptyset$. Then

$$\mathrm{HULL}_{L_1}(\mathbf{K}_1,\mathbf{K}_2)=V^{L_1}_{\mathbf{K}_1}\cap V^{L_1}_{\mathbf{K}_2}.$$

However since $V^{L}_{\mathbf{K}_1\cup\mathbf{F}_1}\cap V^{L}_{\mathbf{K}_2\cup\mathbf{F}_2}=\emptyset$, in the convex hull

$$\mathrm{HULL}_{L}(\mathbf{K}_1\cup\mathbf{F}_1,\mathbf{K}_2\cup\mathbf{F}_2)|_{L_1}$$

there are also probability functions from $V^{L_1}_{\mathbf{K}_1}$ which are not in $V^{L_1}_{\mathbf{K}_1}\cap V^{L_1}_{\mathbf{K}_2}$. Therefore

$$\mathrm{HULL}_{L_1}(\mathbf{K}_1,\mathbf{K}_2)\neq\mathrm{HULL}_{L}(\mathbf{K}_1\cup\mathbf{F}_1,\mathbf{K}_2\cup\mathbf{F}_2)|_{L_1}.$$

Later in theorem 4.2.4 we prove that HULL satisfies **LI** so we can conclude that HULL does not satisfies **IIP**. □

In view of the counterexamples above **IIP** appears hard for a probabilistic merging operator to satisfy. However we might argue that this principle is just too strong. If knowledge provided by experts for the language $L_2$ is itself inconsistent then the addition of such new knowledge may provide us with more information on how strongly the experts 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 experts is jointly consistent, then the principle of irrelevant information is arguably more justified. Accordingly we formulate the:

**Consistent Irrelevant Information Principle (CIIP).**

A probabilistic merging operator $\Delta$ satisfies the consistent irrelevant information principle if whenever $L = L_1 \cup L_2$ are such that $L_1$ and $L_2$ are disjoint propositional languages, $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL_1$ and $\mathbf{F}_1, \ldots, \mathbf{F}_n \in CL_2$, and moreover $\mathbf{F}_1, \ldots, \mathbf{F}_n$ are jointly consistent, then

$$\Delta_L(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)|_{L_1} = \Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)|_{L_1}.$$

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

$$\Delta_L(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)|_{L_1} = \Delta_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

For instance, in the toy example of the section 1.1 the information of both experts about a fault on the electronic switch is both consistent and *a priori* irrelevant to the probability that there will be a fault on the valve. Hence if we want to know only the probability that there will be a fault on the valve, then applying the **CIIP** we need consider only the fact that the first expert states that this probability is 4% and the second states that this probability is 8%.

**Theorem 4.2.4.** HULL *satisfies* **CIIP** *and* **LI**.

*Proof.* 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. In particular $\mathbf{F}_1, \ldots, \mathbf{F}_n$ could be empty.

Consider

$$W_{L_1} = \Big\{V \colon \exists I \subseteq \{1 \ldots n\} \Big[V = \bigcap_{i \in I} V^{L_1}_{\mathbf{K}_i} \neq \emptyset \ \&\ (\forall j \in \{1 \ldots n\} \backslash I) \Big(\bigcap_{i \in I} V^{L_1}_{\mathbf{K}_i}\Big) \cap V^{L_1}_{\mathbf{K}_j} = \emptyset\Big]\Big\}$$

and

$$W_L = \Big\{V \colon \exists I \subseteq \{1 \ldots n\} \Big[V = \bigcap_{i \in I} V^{L}_{\mathbf{K}_i \cup \mathbf{F}_i} \neq \emptyset \ \&$$
$$(\forall j \in \{1 \ldots n\} \setminus I) \Big(\bigcap_{i \in I} V^{L}_{\mathbf{K}_i \cup \mathbf{F}_i}\Big) \cap V^{L}_{\mathbf{K}_j \cup \mathbf{F}_j} = \emptyset\Big]\Big\}.$$

Recall that $\mathrm{HULL}_L(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)$ is defined as the convex hull of sets in $W_L$. Notice that if

$$\mathbf{v} \in V^{L}_{\mathbf{K}_i \cup \mathbf{F}_i} \cap V^{L}_{\mathbf{K}_j \cup \mathbf{F}_j}$$

then $\mathbf{v}|_{L_1}$ satisfies both $\mathbf{K}_i$ and $\mathbf{K}_j$ i.e. $\mathbf{v}|_{L_1} \in V^{L_1}_{\mathbf{K}_i} \cap V^{L_1}_{\mathbf{K}_j}$. On the other hand if $\mathbf{w} \in V^{L_1}_{\mathbf{K}_i} \cap V^{L_1}_{\mathbf{K}_j}$ then for any $\mathbf{t} \in V^{L_2}_{\mathbf{F}_i} \cap V^{L_2}_{\mathbf{F}_j}$

$$\mathbf{w} \cdot \mathbf{t} \in V^{L}_{\mathbf{K}_i \cup \mathbf{F}_i} \cap V^{L}_{\mathbf{K}_j \cup \mathbf{F}_j}$$

and $(\mathbf{w} \cdot \mathbf{t})|_{L_1} = \mathbf{w}$.

Therefore $V \in W_L$ if and only if $V|_{L_1} \in W_{L_1}$. Notice that this is not true when $\mathbf{F}_1, \ldots, \mathbf{F}_n$ are not jointly consistent, for instance see example 4.2.3. We can conclude that the convex hull of the sets in $W_L$ restricted to $L_1$ is the same as the convex hull of the sets in $W_{L_1}$, and the theorem follows. $\square$

Now we prove that the linear entropy operator $\hat{\Delta}^{\mathrm{KL}}$ and the social entropy operator $\Delta^{\mathrm{KL}}$ also satisfy **CIIP**. In the rest of this section we fix two distinct propositional languages $L_1 = \{a_1, \ldots, a_{h_1}\}$ and $L_2 = \{b_1, \ldots, b_{h_2}\}$. Moreover we fix $L = L_1 \cup L_2$ and $\mathrm{At}(L_1) = \{\alpha_1, \ldots, \alpha_J\}$ and $\mathrm{At}(L_2) = \{\beta_1, \ldots, \beta_I\}$.

For $\mathbf{r} \in \mathbb{D}^L$, in order to simplify the notation, we often denote $\mathbf{r}|_{L_1}(\alpha_j)$ by $r_{j\cdot}$. We 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 4.2.5.** *Let $w^{(k)}_j \geq 0$ 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*

$$\sum_{j=1}^{J} \Big[ \prod_{k=1}^{n} w^{(k)}_j \Big]^{\frac{1}{n}} \leq \Big[ \prod_{k=1}^{n} \sum_{j=1}^{J} w^{(k)}_j \Big]^{\frac{1}{n}}.$$

*Equality holds if and only if either there are real constants $l^{(1)} > 0, \ldots, l^{(n)} > 0$ such that $l^{(1)}(w^{(1)}_1, \ldots, w^{(1)}_J) = l^{(2)}(w^{(2)}_1, \ldots, w^{(2)}_J) = \ldots = l^{(n)}(w^{(n)}_1, \ldots, w^{(n)}_J)$ or $\sum_{j=1}^{J} w^{(k)}_j = 0$ for some $k$.*

This lemma is Hölder's inequality, see [23], and it will be very useful in the following proof.

**Lemma 4.2.6.** *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 and condition (1.4) holds for $\mathbf{K}_1, \ldots, \mathbf{K}_n$. Recall that $L = L_1 \cup L_2$. Then:*

*(a) If $\mathbf{v} \in \Delta^{\mathrm{KL}}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and $\mathbf{t}$ is an $L_2$-probability function such that $\mathbf{t} \in \bigcap_{k=1}^{n} V^{L_2}_{\mathbf{F}_k}$ then $\mathbf{v} \cdot \mathbf{t} \in \Delta^{\mathrm{KL}}_{L}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)$. In particular $\mathbf{F}_1, \ldots, \mathbf{F}_n$ could be empty in which case $\mathbf{t}$ can be arbitrary.*

*(b) Let* $\mathbf{r} \in \Delta_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)$*. Then* $\mathbf{r}|_{L_1} \in \Delta_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$*. Moreover* $C_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n) = C_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$.[3]

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

$$(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in \Gamma_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)$$

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

Let us consider probability functions $\mathbf{w}^{(1)}|_{L_1}, \ldots, \mathbf{w}^{(n)}|_{L_1}$. Let $M = \sum_{j=1}^J (\prod_{k=1}^n w_{j\cdot}^{(k)})^{\frac{1}{n}}$. Then $M \leq C_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ since $C_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is maximal. But by the lemma 4.2.5 also $C_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n) \leq M$, hence

$$C_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n) \leq C_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n). \tag{4.9}$$

(a) Let $\mathbf{t} \in \bigcap_{k=1}^n V_{\mathbf{F}_k}^{L_2}$. We are going to prove that

$$(\mathbf{p}^{(1)} \cdot \mathbf{t}, \ldots, \mathbf{p}^{(n)} \cdot \mathbf{t}) \in \Gamma_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n). \tag{4.10}$$

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,

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

since $\sum_{i=1}^I t_i = 1$. But from (4.9) we already know that $C_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n) \leq C_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ hence (4.10) is proved.
(b) By part (a) and by (4.9) we have

$$C_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n) = M = C_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n). \tag{4.11}$$

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

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

[3] *The symbol* $C^{\mathrm{KL}}$ *was defined just above example 4.2.1.*

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

Let us consider the 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

$$w_{ji}^{(k)} = w_{j\cdot}^{(k)} q_{i|j}. \tag{4.12}$$

By (4.11) the projections to $L_1$ satisfy

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

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 $\mathbf{v} \in \Delta_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$.

Moreover,

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

Then $r_{j\cdot} = \sum_i v_j q_{i|j} = v_j$ 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 $\mathbf{r}|_{L_1} \in \Delta_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. □

**Corollary 4.2.7.** *The social entropy operator $\Delta^{\mathrm{KL}}$ satisfies* **LI** *and* **CIIP**.

*Proof.* Let $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL_1$ satisfy (1.4) and $L_1 \cap L_2 = \emptyset$. By the previous lemma, part (b), if $\mathbf{r} \in \Delta_{L_1 \cup L_2}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ then $\mathbf{r}|_{L_1} \in \Delta_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. This together with part (a) gives **LI**; $\Delta_{L_1 \cup L_2}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)|_{L_1} = \Delta_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. Applying this identity to the lemma above we get the formulation of **CIIP**.

Now assume that $\mathbf{K}_1, \ldots, \mathbf{K}_n$ do not satisfy (1.4). By definition $\Delta^{\mathrm{KL}}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \mathbb{D}^{L_1}$. It is easy to see that $\Delta^{\mathrm{KL}}_{L_1 \cup L_2}(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \mathbb{D}^{L_1 \cup L_2}$. For instance, if $\mathbf{K}_1$ force $\mathbf{w} \in V^{L_1}_{K_1}$ to $\mathbf{w}(\alpha) = 0$ for some $\alpha \in \mathrm{At}(L_1)$ then $\mathbf{w}(\alpha \wedge \beta) = 0$ for all $\beta \in \mathrm{At}(L_2)$. Moreover adding a new knowledge base formulated in the language $L_2$ does not change this so $\Delta^{\mathrm{KL}}_{L_1 \cup L_2}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n) = \mathbb{D}^{L_1 \cup L_2}$ for any $\mathbf{F}_1, \ldots, \mathbf{F}_n \in CL_2$. $\mathbb{D}^{L_1 \cup L_2}|_{L_1} = \mathbb{D}^{L_1}$ and therefore $\Delta^{\mathrm{KL}}$ satisfies both **LI** and **CIIP**. □

Recall that in section 3.1 we have defined $\hat{\mathrm{M}}^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ the minimum value of

$$\sum_{k=1}^{n} \sum_{j \in \mathrm{Sig}(\mathbf{v})} w_j^{(k)} \log \frac{w_j^{(k)}}{v_j}$$

subject to $\mathbf{v} \in \mathbb{D}^L$ and $\mathbf{w}^{(k)} \in V^L_{\mathbf{K}_k}$, for all $1 \leq k \leq n$.

**Lemma 4.2.8.** *Let $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL_1$ and $\mathbf{F}_1, \ldots, \mathbf{F}_n \in CL_2$ be such that $\mathbf{F}_1, \ldots, \mathbf{F}_n$ are jointly consistent. Recall that $L = L_1 \cup L_2$. Then:*

*(a) If $\mathbf{v} \in \hat{\Delta}^{\mathrm{KL}}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and $\mathbf{t}$ is an $L_2$-probability function such that $\mathbf{t} \in \bigcap_{k=1}^n V^{L_2}_{\mathbf{F}_k}$ then $\mathbf{v} \cdot \mathbf{t} \in \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)$. In particular $\mathbf{F}_1, \ldots, \mathbf{F}_n$ could be empty in which case $\mathbf{t}$ can be arbitrary.*

*(b) Let $\mathbf{r} \in \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)$. Then $\mathbf{r}|_{L_1} \in \hat{\Delta}^{\mathrm{KL}}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$.*

*Proof.* For a given $\mathbf{v} \in \hat{\Delta}^{\mathrm{KL}}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ let $(\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(n)}) \in \hat{\Gamma}^{\mathrm{KL}}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ be such that $\mathbf{v} = \mathbf{LinOp}(\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(n)})$. For a given

$$\mathbf{r} \in \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)$$

let $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in \hat{\Gamma}^{\mathrm{KL}}_L(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)$ be such that

$$\mathbf{r} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}).$$

Let us consider probability functions $\mathbf{w}^{(1)}|_{L_1}, \ldots, \mathbf{w}^{(n)}|_{L_1}$. Recall that by lemmas 2.1.16 and 2.1.19 and since $\mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \gg \mathbf{w}^{(k)}$ for all $1 \leq k \leq n$ (note that we leave out all indices such that $w_{ji}^{(1)} = \ldots = w_{ji}^{(n)} = 0$ and we implicitly assume this in all the summations below) we have that

$$\sum_{j,i} \sum_{k=1}^{n} x_{ji}^{(k)} \log \frac{x_{ji}^{(k)}}{\frac{\sum_{k=1}^{n} x_{ji}^{(k)}}{n}}$$

is a convex function. Then by the Jensen inequality

$$\hat{\mathrm{M}}_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n) \geq \sum_{k=1}^{n} \sum_{j} w_{j\cdot}^{(k)} \log \frac{n \cdot w_{j\cdot}^{(k)}}{\sum_{k=1}^{n} w_{j\cdot}^{(k)}}. \tag{4.13}$$

On the other hand by lemma 2.1.13 and the definition of $\hat{\mathrm{M}}$

$$\sum_{k=1}^{n} \sum_{j} w_{j\cdot}^{(k)} \log \frac{n \cdot w_{j\cdot}^{(k)}}{\sum_{k=1}^{n} w_{j\cdot}^{(k)}} \geq \hat{\mathrm{M}}_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

Hence

$$\hat{\mathrm{M}}_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n) \geq \hat{\mathrm{M}}_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n). \tag{4.14}$$

(a) Let $\mathbf{t} \in \bigcap_{k=1}^{n} V_{\mathbf{F}_k}^{L_2}$. We are going to prove that

$$(\mathbf{p}^{(1)} \cdot \mathbf{t}, \ldots, \mathbf{p}^{(n)} \cdot \mathbf{t}) \in \hat{\Gamma}_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n). \tag{4.15}$$

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,

$$\sum_{k=1}^{n} \sum_{j,i} p_j^{(k)} \cdot t_i \log \frac{n \cdot p_j^{(k)} \cdot t_i}{\sum_{k=1}^{n} p_j^{(k)} \cdot t_i} =$$

$$= \sum_i t_i \sum_{k=1}^{n} \sum_{j} p_j^{(k)} \log \frac{n \cdot p_j^{(k)}}{\sum_{k=1}^{n} p_j^{(k)}} = \hat{\mathrm{M}}_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n),$$

since $\sum_i t_i = 1$. But from (4.14) we already know that

$$\hat{\mathrm{M}}_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n) \geq \hat{\mathrm{M}}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$$

hence (4.15) is proved.

(b) By part (a) and by (4.14) we have

$$\hat{\mathrm{M}}_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n) = \hat{\mathrm{M}}_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

This together with (4.13) gives $\mathbf{w}^{(1)}|_{L_1}, \ldots, \mathbf{w}^{(n)}|_{L_1} \in \hat{\Gamma}_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and

$$\mathbf{LinOp}(\mathbf{w}^{(1)}|_{L_1}, \ldots, \mathbf{w}^{(n)}|_{L_1}) \in \hat{\Delta}_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

Since $\mathbf{r} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$ clearly $\mathbf{r}|_{L_1} = \mathbf{LinOp}(\mathbf{w}^{(1)}|_{L_1}, \ldots, \mathbf{w}^{(n)}|_{L_1})$ and the part (b) of the lemma is proved. □

**Corollary 4.2.9.** *The linear entropy operator* $\hat{\Delta}^{\mathrm{KL}}$ *satisfies* **LI** *and* **CIIP**.

*Proof.* Let $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL_1$ and $L_1 \cap L_2 = \emptyset$. By the previous lemma, the part (b), if $\mathbf{r} \in \hat{\Delta}^{\mathrm{KL}}_{L_1 \cup L_2}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ then $\mathbf{r}|_{L_1} \in \hat{\Delta}^{\mathrm{KL}}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. This together with the part (a) gives **LI**; $\hat{\Delta}^{\mathrm{KL}}_{L_1 \cup L_2}(\mathbf{K}_1, \ldots, \mathbf{K}_n)|_{L_1} = \hat{\Delta}^{\mathrm{KL}}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. Applying this identity to the lemma above we get the formulation of **CIIP**. $\square$

For the linear Renyi operator $\hat{\Delta}^{D_r}$, $2 \geq r > 1$, we can prove a similar, although weaker, result comparable to lemmas 4.2.6 and 4.2.8. Recall that we have defined $\hat{\mathrm{M}}^{D_r}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ to be the minimum value of

$$\sum_{k=1}^{n} \sum_{j=1}^{J} [(w_j^{(k)})^r - (v_j)^r - r(w_j^{(k)} - v_j)(v_j)^{r-1}]$$

subject to $\mathbf{v} \in \mathbb{D}^L$ and $\mathbf{w}^{(k)} \in V^L_{\mathbf{K}_k}$, for all $1 \leq k \leq n$.

**Lemma 4.2.10.** *Let $2 \geq r > 1$. Recall that $I = |\operatorname{At}(L_2)|$ and $L = L_1 \cup L_2$. Let $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL_1$, $\mathbf{F}_1, \ldots, \mathbf{F}_n \in CL_2$ be such that $\mathbf{t} = \underbrace{\left(\frac{1}{I}, \ldots, \frac{1}{I}\right)}_{I} \in \bigcap_{k=1}^n V^{L_2}_{\mathbf{F}_k}$. In particular $\mathbf{F}_1, \ldots, \mathbf{F}_n$ could be empty. Then:*

*(a) If $\mathbf{v} \in \hat{\Delta}^{D_r}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ then $\mathbf{v} \cdot \mathbf{t} \in \hat{\Delta}^{D_r}_{L}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)$.*

*(b) If $\mathbf{r} \in \hat{\Delta}^{D_r}_{L}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)$ then $\mathbf{r}|_{L_1} \in \hat{\Delta}^{D_r}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$.*

*Proof.* For a given $\mathbf{v} \in \hat{\Delta}^{D_r}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ let $(\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(n)}) \in \hat{\Gamma}^{D_r}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ be such that $\mathbf{v} = \mathbf{LinOp}(\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(n)})$. For a given $\mathbf{r} \in \hat{\Delta}^{D_r}_{L}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)$ let

$$(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in \hat{\Gamma}^{D_r}_{L}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n)$$

be such that $\mathbf{r} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$.

Let us consider probability functions $\mathbf{w}^{(1)}|_{L_1}, \ldots, \mathbf{w}^{(n)}|_{L_1}$. Recall that by lemmas 2.1.18 and 2.1.19 we have that

$$\sum_{j,i} \sum_{k=1}^{n} \left[ (x_{ji}^{(k)})^r - \left( \frac{\sum_{k=1}^n x_{ji}^{(k)}}{n} \right)^r - r\left( x_{ji}^{(k)} - \frac{\sum_{k=1}^n x_{ji}^{(k)}}{n} \right) \left( \frac{\sum_{k=1}^n x_{ji}^{(k)}}{n} \right)^{r-1} \right]$$

is a convex function. Then by the Jensen inequality

$$\hat{\mathrm{M}}^{D_r}_L(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n) \geq$$

$$\geq \sum_{j} \sum_{k=1}^{n} \left[ (w_{j\cdot}^{(k)})^r - \left( \frac{\sum_{k=1}^n w_{j\cdot}^{(k)}}{n} \right)^r - r\left( w_{j\cdot}^{(k)} - \frac{\sum_{k=1}^n w_{j\cdot}^{(k)}}{n} \right) \left( \frac{\sum_{k=1}^n w_{j\cdot}^{(k)}}{n} \right)^{r-1} \right] \cdot \frac{I}{I^r}. \quad (4.16)$$

On the other hand by lemma 2.1.13 and the definition of $\hat{\mathrm{M}}^{D_r}$

$$\sum_j \sum_{k=1}^n \Big[(w_{j\cdot}^{(k)})^r - \Big(\frac{\sum_{k=1}^n w_{j\cdot}^{(k)}}{n}\Big)^r - r\Big(w_{j\cdot}^{(k)} - \frac{\sum_{k=1}^n w_{j\cdot}^{(k)}}{n}\Big)\Big(\frac{\sum_{k=1}^n w_{j\cdot}^{(k)}}{n}\Big)^{r-1}\Big] \geq$$

$$\geq \hat{\mathrm{M}}_{L_1}^{D_r}(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

Hence

$$\hat{\mathrm{M}}_L^{D_r}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n) \geq \frac{I}{I^r} \cdot \hat{\mathrm{M}}_{L_1}^{D_r}(\mathbf{K}_1, \ldots, \mathbf{K}_n). \tag{4.17}$$

(a) We are going to prove that

$$(\mathbf{p}^{(1)} \cdot \mathbf{t}, \ldots, \mathbf{p}^{(n)} \cdot \mathbf{t}) \in \hat{\Gamma}_L^{D_r}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n). \tag{4.18}$$

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,

$$\sum_{k=1}^n \sum_{j,i} \Big[(p_j^{(k)} \cdot t_i)^r - \Big(\frac{\sum_{k=1}^n p_j^{(k)} \cdot t_i}{n}\Big)^r - r\Big(p_j^{(k)} \cdot t_i - \frac{\sum_{k=1}^n p_j^{(k)} \cdot t_i}{n}\Big)\Big(\frac{\sum_{k=1}^n p_j^{(k)} \cdot t_i}{n}\Big)^{r-1}\Big] =$$

$$= \sum_i (t_i)^r \sum_{k=1}^n \sum_j \Big[(p_j^{(k)})^r - \Big(\frac{\sum_{k=1}^n p_j^{(k)}}{n}\Big)^r - r\Big(p_j^{(k)} - \frac{\sum_{k=1}^n p_j^{(k)}}{n}\Big)\Big(\frac{\sum_{k=1}^n p_j^{(k)}}{n}\Big)^{r-1}\Big] =$$

$$= \frac{I}{I^r} \cdot \hat{\mathrm{M}}_{L_1}^{D_r}(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

But from (4.17) we already know that

$$\hat{\mathrm{M}}_L^{D_r}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n) \geq \frac{I}{I^r} \cdot \hat{\mathrm{M}}_{L_1}^{D_r}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$$

hence (4.18) is proved.

(b) By part (a) and by (4.17) we have

$$\hat{\mathrm{M}}_L^{D_r}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n) = \frac{I}{I^r} \cdot \hat{\mathrm{M}}_{L_1}^{D_r}(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

This together with (4.16) gives $\mathbf{w}^{(1)}|_{L_1}, \ldots, \mathbf{w}^{(n)}|_{L_1} \in \hat{\Gamma}_{L_1}^{D_r}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and

$$\mathbf{LinOp}(\mathbf{w}^{(1)}|_{L_1}, \ldots, \mathbf{w}^{(n)}|_{L_1}) \in \hat{\Delta}_{L_1}^{D_r}(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

Since $\mathbf{r} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$ clearly $\mathbf{r}|_{L_1} = \mathbf{LinOp}(\mathbf{w}^{(1)}|_{L_1}, \ldots, \mathbf{w}^{(n)}|_{L_1})$ and the part (b) of the lemma is proved. □

**Corollary 4.2.11.** *The linear Renyi operator* $\hat{\Delta}^{D_r}$, $2 \geq r > 1$, *satisfies* **LI**.

*Proof.* Let $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL_1$ and $L_1 \cap L_2 = \emptyset$. By the previous lemma, the part (b), if $\mathbf{r} \in \hat{\Delta}^{D_r}_{L_1 \cup L_2}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ then $\mathbf{r}|_{L_1} \in \hat{\Delta}^{D_r}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. This together with the part (a) gives **LI**; $\hat{\Delta}^{D_r}_{L_1 \cup L_2}(\mathbf{K}_1, \ldots, \mathbf{K}_n)|_{L_1} = \hat{\Delta}^{D_r}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. □

So far we have not seen any p-merging operator which satisfies the irrelevant information principle **IIP**. It might be a bit surprising that two p-merging operators, which until now we have ignored because they satisfy neither the consistency principle **(K4)** nor the collegiality principle **(K5)**, do in fact satisfy **IIP**. These operators are the obdurate social entropy operator **OSEP** and the obdurate linear entropy operator **OLEP**. Recall that these operators always produce a single probability function and therefore they are thus far the only instances of, what we call, social inference processes which we have considered. In the following section we prove that they satisfy **IIP** and we discuss whether these operators satisfy a further strengthening of this principle.

## 4.3 Independence principle

In this section we investigate the property which is called *independence*. In probability theory two events are independent if the probability that they both occur is the product of their individual probabilities. In multi-expert probabilistic reasoning the question has been studied as to whether there are pooling operators which preserve independence. We say that a pooling operator **Pool** *preserves independence* if whenever $\theta$, $\varphi \in SL$ and $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(h)} \in \mathbb{D}^L$ are such that $\theta \wedge \varphi$ is satisfiable and $\mathbf{w}^{(i)}(\theta \wedge \varphi) = \mathbf{w}^{(i)}(\theta) \cdot \mathbf{w}^{(i)}(\varphi)$, for all $1 \leq i \leq n$, then

$$\mathbf{Pool}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(h)})(\theta \wedge \varphi) = \mathbf{Pool}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(h)})(\theta) \cdot \mathbf{Pool}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(h)})(\varphi).$$

Unfortunately, Genest and Wagner in [17] proved that in a fairly general case the pooling operators which preserve independence are *dictatorships* — for some $i$

$$\mathbf{Pool}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(h)}) = \mathbf{w}^{(i)}.$$

More explicitly the above happens when a pooling operator **Pool** is of the form

$$\mathbf{Pool}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(h)})(\alpha_j) = \frac{F_j(\mathbf{w}^{(1)}(\alpha_j), \ldots, \mathbf{w}^{(n)}(\alpha_j))}{\sum_{k=1}^{J} F_k(\mathbf{w}^{(1)}(\alpha_k), \ldots, \mathbf{w}^{(n)}(\alpha_k))}, \tag{4.19}$$

for all $1 \leq j \leq J = |\operatorname{At}(L)|$, where $F_j : [0,1]^n \to [0,1]$, $\alpha_1, \ldots, \alpha_J \in \operatorname{At}(L)$ and $|L| > 2$. **LinOp** and **LogOp** pooling operators are both of the form (4.19) and consequently they do not preserve independence.

On the other hand, as was pointed out in [17], one can easily see that if $|L| = 2$ then **LogOp** pooling operator does preserve independence. (Note that for $|L| = 1$ the situation is trivial.) It is therefore of a particular interest to examine what particular properties present in the $|L| = 2$ case which ensure that the **LogOp** pooling operator preserves independence. In the following we propose a strengthening of one such property.

We say that probability functions $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)} \in \mathbb{D}^L$ are *strongly independent* for $L$-sentences $\theta$ and $\varphi$ if $\theta \wedge \varphi$ is satisfiable and there exist two sets of $L$-sentences $\{\alpha_1, \ldots, \alpha_{J_1}\}$ and $\{\beta_1, \ldots, \beta_{J_2}\}$ such that

1. $J_1 \cdot J_2 = J = |\operatorname{At}(L)|$,

2. every atom of $\operatorname{At}(L)$ is logically equivalent to one and only one sentence of the form $\alpha_j \wedge \beta_i$, where $1 \leq j \leq J_1$ and $1 \leq i \leq J_2$,

3. $\theta \equiv \bigvee_{j \in S_\theta} \alpha_j$, $\varphi \equiv \bigvee_{j \in S_\varphi} \beta_j$ for some index sets $S_\theta$ and $S_\varphi$ and

4. $\mathbf{w}^{(k)}(\alpha_j \wedge \beta_i) = \mathbf{w}^{(k)}(\alpha_j) \cdot \mathbf{w}^{(k)}(\beta_i)$, for all $1 \leq j \leq J_1$, $1 \leq i \leq J_2$ and $1 \leq k \leq n$.

Note that above conditions imply that

$$\mathbf{w}^{(k)}(\theta \wedge \varphi) = \mathbf{w}^{(k)}(\theta) \cdot \mathbf{w}^{(k)}(\varphi)$$

since $\mathbf{w}^{(k)}(\theta \wedge \varphi) = \sum_{j \in S_\theta, i \in S_\varphi} \mathbf{w}^{(k)}(\alpha_j \wedge \beta_i) = \sum_{j \in S_\theta, i \in S_\varphi} \mathbf{w}^{(k)}(\alpha_j)\mathbf{w}^{(k)}(\beta_i) = \sum_{j \in S_\theta} \mathbf{w}^{(k)}(\alpha_j) \sum_{i \in S_\varphi} \mathbf{w}^{(k)}(\beta_i) = \mathbf{w}^{(k)}(\theta)\mathbf{w}^{(k)}(\varphi)$.

Conditions 1–3 above hold for instance when $L_1 \cup L_2 = L$ is a partition of $L$, $\theta \in SL_1$, $\varphi \in SL_2$ and the two sets are $\operatorname{At}(L_1)$ and $\operatorname{At}(L_2)$. The following example shows that this is not the only case.

**Example 4.3.1.** Consider $L = \{a, b\}$ and $\theta = a$, $\varphi = a \leftrightarrow b$. Then for sets of $L$-sentences $\{a, \neg a\}$ and $\{a \leftrightarrow b, \neg(a \leftrightarrow b)\}$ conditions 1–3 hold.

The following lemma will help us to understand the notion of strong independence a little bit better.

**Lemma 4.3.2.** *Let $\{\alpha_1, \dots, \alpha_{J_1}\}$ and $\{\beta_1, \dots, \beta_{J_2}\}$ be such that conditions 1–4 hold for some $\mathbf{w} \in \mathbb{D}^L$ and $\theta$, $\varphi \in SL$. Then $\sum_{j=1}^{J_1} \mathbf{w}(\alpha_j) = 1$ and $\sum_{i=1}^{J_2} \mathbf{w}(\beta_i) = 1$.*

*Proof.* We prove only that $\sum_{i=1}^{J_2} \mathbf{w}(\beta_i) = 1$; the other case is analogous. Let $\alpha_j$ be such that $\mathbf{w}(\alpha_j) \neq 0$. Such $\alpha_j$ exists because there is at least one atom $\gamma \in \mathrm{At}(L)$ such that $\mathbf{w}(\gamma) \neq 0$ and we can choose $\alpha_j$ in a way that $\gamma \models \alpha_j$. Then

$$\mathbf{w}(\alpha_j) = \mathbf{w}(\bigvee_{\substack{\gamma \in \mathrm{At}(L)\\ \gamma \models \alpha_j}} \gamma) = \mathbf{w}(\bigvee_{i=1}^{J_2} (\alpha_j \wedge \beta_i)) = \mathbf{w}(\alpha_j) \sum_{i=1}^{J_2} \mathbf{w}(\beta_i).$$

□

The following theorem shows that if $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)} \in \mathbb{D}^L$ are strongly independent for $L$-sentences $\theta$ and $\varphi$ then the **LogOp**-pooling operator preserves independence restricted to sentences $\theta$ and $\varphi$.

**Theorem 4.3.3.** *Assume that $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)} \in \mathbb{D}^L$ are strongly independent for $L$-sentences $\theta$ and $\varphi$ and that they satisfy (1.3). Then*

$$\mathbf{LogOp}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})(\theta) \cdot \mathbf{LogOp}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})(\varphi) =$$

$$= \mathbf{LogOp}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})(\theta \wedge \varphi).$$

*Proof.* Let $A = \{\alpha_1, \dots, \alpha_J\}$ and $B = \{\beta_1, \dots, \beta_I\}$ be the two sets of $L$-sentences from the definition of strong independence. First of all note that

$$\mathbf{LogOp}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})(\theta \wedge \varphi) = \sum_{\substack{\gamma \in \mathrm{At}(L)\\ \gamma \models \theta \wedge \varphi}} \mathbf{LogOp}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})(\gamma).$$

From the definition every atom $\gamma \in \mathrm{At}(L)$ is logically equivalent to an unique $\alpha_\gamma \wedge \beta_\gamma$ where $\alpha_\gamma \in A$ and $\beta_\gamma \in B$. Therefore the above is equal to

$$\frac{\sum_{\substack{\gamma \models \theta \wedge \varphi\\ \gamma \in \mathrm{At}(L)}} \left[\prod_{k=1}^n \mathbf{w}^{(k)}(\alpha_\gamma \wedge \beta_\gamma)\right]^{\frac{1}{n}}}{\sum_{\gamma \in \mathrm{At}(L)} \left[\prod_{k=1}^n \mathbf{w}^{(k)}(\alpha_\gamma \wedge \beta_\gamma)\right]^{\frac{1}{n}}} = \frac{\sum_{\substack{\gamma \models \theta \wedge \varphi\\ \gamma \in \mathrm{At}(L)}} \left[\prod_{k=1}^n \mathbf{w}^{(k)}(\alpha_\gamma)\mathbf{w}^{(k)}(\beta_\gamma)\right]^{\frac{1}{n}}}{\sum_{\gamma \in \mathrm{At}(L)} \left[\prod_{k=1}^n \mathbf{w}^{(k)}(\alpha_\gamma)\mathbf{w}^{(k)}(\beta_\gamma)\right]^{\frac{1}{n}}}.$$

Let $\alpha \models \theta$, $\alpha \in A$. Then for every $\beta \models \varphi$, $\beta \in B$, there is an atom $\gamma \in \mathrm{At}(L)$ such that $\gamma \models \theta \wedge \varphi$ and $\gamma \equiv \alpha \wedge \beta$. On the other hand if $\alpha \not\models \theta$ then $\alpha \wedge \beta_i \not\models \theta \wedge \varphi$ for all $\beta_i$, $1 \leq i \leq I$. We can conclude that

$$\sum_{\substack{\gamma \models \theta \wedge \varphi\\ \gamma \in \mathrm{At}(L)}} [\prod_{k=1}^n \mathbf{w}^{(k)}(\alpha_\gamma)\mathbf{w}^{(k)}(\beta_\gamma)]^{\frac{1}{n}} = \sum_{\substack{\alpha \models \theta\\ \alpha \in A}} [\prod_{k=1}^n \mathbf{w}^{(k)}(\alpha)]^{\frac{1}{n}} \cdot \sum_{\substack{\beta \models \varphi\\ \beta \in B}} [\prod_{k=1}^n \mathbf{w}^{(k)}(\beta)]^{\frac{1}{n}}.$$

Applying the same procedure as above we prove that

$$\sum_{\gamma\in \mathrm{At}(L)} [\prod_{k=1}^{n} \mathbf{w}^{(k)}(\alpha_\gamma)\mathbf{w}^{(k)}(\beta_\gamma)]^{\frac{1}{n}} = \sum_{\alpha\in A}[\prod_{k=1}^{n} \mathbf{w}^{(k)}(\alpha)]^{\frac{1}{n}} \cdot \sum_{\beta\in B}[\prod_{k=1}^{n} \mathbf{w}^{(k)}(\beta)]^{\frac{1}{n}},$$

$$\sum_{\substack{\gamma\models\theta \\ \gamma\in \mathrm{At}(L)}} [\prod_{k=1}^{n} \mathbf{w}^{(k)}(\alpha_\gamma)\mathbf{w}(\beta_\gamma)]^{\frac{1}{n}} = \sum_{\substack{\alpha\models\theta \\ \alpha\in A}}[\prod_{k=1}^{n} \mathbf{w}^{(k)}(\alpha)]^{\frac{1}{n}} \cdot \sum_{\beta\in B}[\prod_{k=1}^{n} \mathbf{w}^{(k)}(\beta)]^{\frac{1}{n}}$$

and

$$\sum_{\substack{\gamma\models\varphi \\ \gamma\in \mathrm{At}(L)}} [\prod_{k=1}^{n} \mathbf{w}^{(k)}(\alpha_\gamma)\mathbf{w}^{(k)}(\beta_\gamma)]^{\frac{1}{n}} = \sum_{\alpha\in A}[\prod_{k=1}^{n} \mathbf{w}^{(k)}(\alpha)]^{\frac{1}{n}} \cdot \sum_{\substack{\beta\models\varphi \\ \beta\in B}}[\prod_{k=1}^{n} \mathbf{w}^{(k)}(\beta)]^{\frac{1}{n}}.$$

It follows that

$$\mathbf{LogOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})(\theta\wedge\varphi) =$$

$$= \frac{\sum_{\substack{\gamma\models\theta \\ \gamma\in \mathrm{At}(L)}} \left[\prod_{k=1}^{n} \mathbf{w}^{(k)}(\alpha_\gamma)\mathbf{w}^{(k)}(\beta_\gamma)\right]^{\frac{1}{n}}}{\sum_{\gamma\in \mathrm{At}(L)} \left[\prod_{k=1}^{n} \mathbf{w}^{(k)}(\alpha_\gamma)\mathbf{w}^{(k)}(\beta_\gamma)\right]^{\frac{1}{n}}} \cdot \frac{\sum_{\substack{\gamma\models\varphi \\ \gamma\in \mathrm{At}(L)}} \left[\prod_{k=1}^{n} \mathbf{w}^{(k)}(\alpha_\gamma)\mathbf{w}^{(k)}(\beta_\gamma)\right]^{\frac{1}{n}}}{\sum_{\gamma\in \mathrm{At}(L)} \left[\prod_{k=1}^{n} \mathbf{w}^{(k)}(\alpha_\gamma)\mathbf{w}^{(k)}(\beta_\gamma)\right]^{\frac{1}{n}}} =$$

$$= \mathbf{LogOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})(\theta) \cdot \mathbf{LogOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})(\varphi).$$

□

Now consider the following theorem.

**Theorem 4.3.4.** *Let* $L_1 \cap L_2 = \emptyset$, $\mathbf{K} \in CL_1$ *and* $\mathbf{F} \in CL_2$. *Then*

$$\mathbf{ME}_{L_1\cup L_2}(\mathbf{K}\cup\mathbf{F}) = \mathbf{ME}_{L_1}(\mathbf{K}) \cdot \mathbf{ME}_{L_2}(\mathbf{F}).$$

The proof is in [39]. The property described in this theorem is called the *independence property*. It is worth mentioning that **ME** is the only known inference process which satisfies this independence property, see [24]. (Note that this property is stronger than the independence principle for inference processes due to Paris and Vencovská mentioned in section 1.3.) Since **ME** satisfies the independence property and a corollary of theorem 4.3.3 is that **LogOp** preserves the independence of strongly independent probability functions for sentences formulated in distinct languages, we will see that the obdurate social entropy operator **OSEP** also satisfies the following:

**Independence Principle (IP).** A p-merging operator $\Delta$ satisfies the independence principle if whenever $L = L_1 \cup L_2$ is such that $L_1$ and $L_2$ are disjoint propositional languages and $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL_1$ and $\mathbf{F}_1, \ldots, \mathbf{F}_n \in CL_2$, then

$$\Delta_L(\mathbf{K}_1\cup\mathbf{F}_1, \ldots, \mathbf{K}_n\cup\mathbf{F}_n) = \Delta_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cdot \Delta_{L_2}(\mathbf{F}_1, \ldots, \mathbf{F}_n),$$

where $\Delta_{L_1}(\mathbf{K}_1,\ldots,\mathbf{K}_n)\cdot\Delta_{L_2}(\mathbf{F}_1,\ldots,\mathbf{F}_n)=\{\mathbf{v}\cdot\mathbf{w}\colon\ \mathbf{v}\in\Delta_{L_1}(\mathbf{K}_1,\ldots,\mathbf{K}_n),\mathbf{w}\in\Delta_{L_2}(\mathbf{F}_1,\ldots,\mathbf{F}_n)\}$.

Note that **IP** implies both the language invariance principle **LI** and the irrelevant information principle **IIP**.

The next theorem states that **OSEP** satisfies **IP**.

**Theorem 4.3.5.** *Let $L_1\cap L_2=\emptyset$, $\mathbf{K}_1,\ldots,\mathbf{K}_n\in CL_1$, $\mathbf{F}_1,\ldots,\mathbf{F}_n\in CL_2$ and $\varphi\in SL_1$, $\psi\in SL_2$. Assume that $\mathbf{K}_1,\ldots,\mathbf{K}_n$; $\mathbf{F}_1,\ldots,\mathbf{F}_n$ and $\mathbf{K}_1\cup\mathbf{F}_1,\ldots,\mathbf{K}_n\cup\mathbf{F}_n$ respectively satisfy (1.4). Then*

$$\mathbf{OSEP}_{L_1\cup L_2}(\mathbf{K}_1\cup\mathbf{F}_1,\ldots,\mathbf{K}_n\cup\mathbf{F}_n)(\varphi\wedge\psi)=$$

$$=\mathbf{OSEP}_{L_1}(\mathbf{K}_1,\ldots,\mathbf{K}_n)(\varphi)\cdot\mathbf{OSEP}_{L_2}(\mathbf{F}_1,\ldots,\mathbf{F}_n)(\psi).$$

*Proof.* By theorem 4.3.4 $\mathbf{ME}_{L_1\cup L_2}(\mathbf{K}_k\cup\mathbf{F}_k)=\mathbf{ME}_{L_1}(\mathbf{K}_k)\cdot\mathbf{ME}_{L_2}(\mathbf{F}_k)$ for all $1\le k\le n$. Therefore the probability functions $\mathbf{ME}_{L_1\cup L_2}(\mathbf{K}_k\cup\mathbf{F}_k)$, $1\le k\le n$, are strongly independent for $\varphi\in SL_1$ and $\psi\in SL_2$. Then by theorem 4.3.3 we have that

$$\mathbf{LogOp}(\mathbf{ME}_{L_1\cup L_2}(\mathbf{K}_1\cup\mathbf{F}_1),\ldots,\mathbf{ME}_{L_1\cup L_2}(\mathbf{K}_n\cup\mathbf{F}_n))(\varphi\wedge\psi)=$$

$$=\mathbf{LogOp}(\mathbf{ME}_{L_1}(\mathbf{K}_1),\ldots,\mathbf{ME}_{L_1}(\mathbf{K}_n))(\varphi)\cdot\mathbf{LogOp}(\mathbf{ME}_{L_2}(\mathbf{F}_1),\ldots,\mathbf{ME}_{L_2}(\mathbf{F}_n))(\psi).$$

□

This principle is obviously a strengthening of the irrelevant information principle and **OSEP** is so far the only natural p-merging operator which is known to satisfy it. This underlines how hard it is for this principle to be satisfied. Unfortunately, **OSEP** does not satisfy the consistency principle and the collegiality principle, and so from our point of view can hardly be a preferred choice for multi-expert reasoning.

**Theorem 4.3.6.** **OLEP** *satisfies* **IIP** *and* **LI**. *On the other hand* **OLEP** *does not satisfy* **IP**.

*Proof.* Let $L_1\cup L_2=L$, $L_1\cap L_2=\emptyset$, $\mathbf{K}_1,\ldots,\mathbf{K}_n\in CL_1$ and $\mathbf{F}_1,\ldots,\mathbf{F}_n\in CL_2$. Let $\mathbf{ME}_L(\mathbf{K}_k\cup\mathbf{F}_k)=\mathbf{w}^{(k)}$. By theorem 4.3.4 $w^{(k)}_{ji}=w^{(k)}_{j\cdot}w^{(k)}_{\cdot i}$, where $\mathbf{w}^{(k)}|_{L_1}=\mathbf{ME}_{L_1}(\mathbf{K}_k)$ and $\mathbf{w}^{(k)}|_{L_2}=\mathbf{ME}_{L_2}(\mathbf{F}_k)$. Hence

$$\mathbf{OLEP}_L(\mathbf{K}_1\cup\mathbf{F}_1,\ldots,\mathbf{K}_n\cup\mathbf{F}_n)(\alpha_j)|_{L_1}=\mathbf{LinOp}(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)})|_{L_1}=$$

$$= \sum_i \frac{\sum_{k=1}^n w_{j\cdot}^{(k)} w_{\cdot i}^{(k)}}{n} = \frac{\sum_{k=1}^n w_{j\cdot}^{(k)}}{n} = \mathbf{OLEP}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n)(\alpha_j)$$

for any $\alpha_j \in \mathrm{At}(L_1)$ which gives us the formulation of **IIP**. Since we can set $\mathbf{F}_1 = \ldots = \mathbf{F}_n = \emptyset$ it also follows that **OLEP** satisfies **LI**.

Notice that $\frac{\sum_{k=1}^n w_{j\cdot}^{(k)} w_{\cdot i}^{(k)}}{n} \neq \sum_j \frac{\sum_{k=1}^n w_{j\cdot}^{(k)} w_{\cdot i}^{(k)}}{n} \cdot \sum_i \frac{\sum_{k=1}^n w_{j\cdot}^{(k)} w_{\cdot i}^{(k)}}{n}$ and therefore **OLEP** does not satisfy **IP**. □

## 4.4 Merging of Kern-Isberner and Röder

In this section we mention an original idea how to merge knowledge bases due to Kern-Isberner and Röder, see [29]. Let $L^+ = L \cup \{s_1, \ldots, s_n\}$ where $s_1, \ldots, s_n$ are propositional variables all distinct from the variables in $L$. The atoms of $L^+$ are of the form $\alpha_j \wedge s_{i_1} \wedge \ldots s_{i_m} \wedge \neg s_{i_{m+1}} \wedge \ldots \wedge \neg s_{i_n}$, where $\alpha_j \in \mathrm{At}(L)$. Given $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ we define $\mathbf{K}_i(s_i)$ as $\mathbf{K}_i$ where every $L$-sentence $\theta$ is replaced by $L^+$-sentence $\theta | s_i$.

Now consider the knowledge base $\mathbf{K}$ defined by

$$\bigcup_{i=1}^n \mathbf{K}_i(s_i) \cup \bigcup_{\substack{i_1, i_2 \\ i_1 \neq i_2}} \{Bel(s_{i_1} \wedge s_{i_2}) = 0\} \cup \{Bel(\bigvee_{i=1}^n s_i) = 1\}.$$

$\mathbf{K}$ is obviously well defined (constraints are consistent). Kern-Isberner and Röder argue that this set naturally represents original possibly inconsistent $n$-tuple of knowledge bases $\mathbf{K}_1, \ldots, \mathbf{K}_n$. Following the justification of the **ME**-inference process, they define their probabilistic merging operator as the most entropic probability function in $V_{\mathbf{K}}^{L^+}$. Notice that this operator satisfies the defining principle (**K1**) since it always produces a single probability function. It is therefore another instance of a social inference process. In the following we describe how we can actually compute the result of this process.

Let us denote $\mathbf{ME}_{L^+}(\mathbf{K}) = \mathbf{v}$. In [29] it is proved that

$$\mathbf{v}(\alpha_j) = \sum_{i=1}^n \mathbf{v}(s_i)\, \mathbf{ME}_L(\mathbf{K}_i)(\alpha_j) \tag{4.20}$$

for all $\alpha_j \in \mathrm{At}(L)$.

**Theorem 4.4.1.** *Let* $\mathbf{v}$, $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ *and* $s_1 \ldots, s_n$ *be as above. Then*

$$\mathbf{v}(s_i) = \frac{e^{H(\mathbf{ME}_L(\mathbf{K}_i))}}{\sum_{i=1}^n e^{H(\mathbf{ME}_L(\mathbf{K}_i))}},$$

*for all* $1 \leq i \leq n$, *where* $H$ *denotes the Shannon entropy.*

*Proof.* By (4.20) and by the definition of $\mathbf{K}$ we have that

$$\mathbf{v}(\alpha_j \wedge s_i) = \mathbf{v}(s_i)\,\mathbf{ME}_L(\mathbf{K}_i)(\alpha_j)$$

and $\mathbf{v}(\alpha_j|s_i) = \mathbf{ME}_L(\mathbf{K}_i)(\alpha_j)$. Therefore $\mathbf{v} \in V_{\mathbf{K}}^{L^+}$ regardless on values $\mathbf{v}(s_i)$, $1 \leq i \leq n$, subject only to $\sum_{i=1}^n \mathbf{v}(s_i) = 1$. Hence, it is sufficient to identify for which $\mathbf{v}(s_i)$, $i = 1, \ldots, n$, the entropy

$$-\sum_{i=1}^{n}\sum_{j=1}^{J} \mathbf{v}(s_i)\,\mathbf{ME}_L(\mathbf{K}_i)(\alpha_j) \log \left[\mathbf{v}(s_i)\,\mathbf{ME}_L(\mathbf{K}_i)(\alpha_j)\right]$$

is maximal subject to $\sum_{i=1}^n \mathbf{v}(s_i) = 1$. Now

$$\frac{\partial}{\partial \mathbf{v}(s_i)}\Big[-\sum_{i=1}^{n}\sum_{j=1}^{J} \mathbf{v}(s_i)\,\mathbf{ME}_L(\mathbf{K}_i)(\alpha_j) \log \left[\mathbf{v}(s_i)\,\mathbf{ME}_L(\mathbf{K}_i)(\alpha_j)\right]+$$

$$+\lambda(\sum_{i=1}^{n} \mathbf{v}(s_i) - 1)\Big] = -1 - \log \mathbf{v}(s_i) + H(\mathbf{ME}_L(\mathbf{K}_i)) + \lambda.$$

Setting this equal to zero gives

$$\mathbf{v}(s_i) = e^{H(\mathbf{ME}_L(\mathbf{K}_i))} e^{\lambda - 1}.$$

Since $\sum_{i=1}^n \mathbf{v}(s_i) = 1$ we have that

$$e^{\lambda-1} = \frac{1}{\sum_{i=1}^n e^{H(\mathbf{ME}_L(\mathbf{K}_i))}}$$

and hence

$$\mathbf{v}(s_i) = \frac{e^{H(\mathbf{ME}_L(\mathbf{K}_i))}}{\sum_{i=1}^n e^{H(\mathbf{ME}_L(\mathbf{K}_i))}},\ i = 1, \ldots, n,$$

is the only critical point. Since the function is concave there is a global maximum at this point. $\square$

In particular, if $V_{\mathbf{K}_1}^L = \{\mathbf{w}^{(1)}\}, \ldots, V_{\mathbf{K}_n}^L = \{\mathbf{w}^{(n)}\}$ then $\mathbf{ME}_{L^+}(\mathbf{K})(\alpha_j)$ is equal to

$$\sum_{i=1}^{n} \frac{e^{H(\mathbf{w}^{(i)})}}{\sum_{i=1}^n e^{H(\mathbf{w}^{(i)})}} w_j^{(i)}$$

what is simply a *weighted* linear pooling operator.

Therefore the Kern-Isberner and Röder merging operator (**KIRP**) can be alternatively defined for any $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ by

$$\mathbf{KIRP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \sum_{i=1}^{n} \frac{e^{H(\mathbf{ME}_L(\mathbf{K}_i))}}{\sum_{i=1}^n e^{H(\mathbf{ME}_L(\mathbf{K}_i))}} \mathbf{ME}_L(\mathbf{K}_i),$$

where $H(\mathbf{w}^{(i)}) = -\sum_{j=1}^J w_j^{(i)} \log w_j^{(i)}$ and $\mathbf{w}^{(i)} = \mathbf{ME}_L(\mathbf{K}_i)$.

It is easy to see that **KIRP** satisfies principles **(K2)** and **(K3)**. Also:

**Theorem 4.4.2.** **KIRP** *satisfies* **(K6*)** *and* **(K7)**.

*Proof.* The theorem is easy to observe by considering the following identity.

$$\mathbf{KIRP}_L(\mathbf{K}_1,\dots,\mathbf{K}_n,\mathbf{F}_1,\dots,\mathbf{F}_m) =$$
$$= \frac{\sum_{i=1}^n e^{H(\mathbf{ME}_L(\mathbf{K}_i))}}{\sum_{i=1}^n e^{H(\mathbf{ME}_L(\mathbf{K}_i))} + \sum_{i=1}^m e^{H(\mathbf{ME}_L(\mathbf{F}_i))}} \mathbf{KIRP}_L(\mathbf{K}_1,\dots,\mathbf{K}_n) +$$
$$+ \frac{\sum_{i=1}^m e^{H(\mathbf{ME}_L(\mathbf{F}_i))}}{\sum_{i=1}^n e^{H(\mathbf{ME}_L(\mathbf{K}_i))} + \sum_{i=1}^m e^{H(\mathbf{ME}_L(\mathbf{F}_i))}} \mathbf{KIRP}_L(\mathbf{F}_1,\dots,\mathbf{F}_m).$$

□

Notice that **KIRP** gives higher weights to those experts whose knowledge bases are more entropic and hence less controversial. Therefore it is in spirit something in between obdurate merging operators and those which satisfy the principles of consistency and collegiality. However:

**Theorem 4.4.3.** **KIRP** *does not satisfy the consistency principle* **(K4)**.

*Proof.* Assume that $L$ has at least two propositional variables. In the proof of theorem 1.4.1 the following knowledge bases have been constructed: $V^L_{\mathbf{K}_1} = \{\lambda\mathbf{v} + (1-\lambda)\mathbf{w} : \lambda \in [0,1]\}$, $V^L_{\mathbf{K}_2} = \{\lambda\mathbf{a}' + (1-\lambda)\mathbf{w} : \lambda \in [0,1]\}$, $V^L_{\mathbf{F}_1} = \{\lambda\mathbf{v} + (1-\lambda)\mathbf{u} : \lambda \in [0,1]\}$ and $V^L_{\mathbf{F}_2} = \{\lambda\mathbf{a}' + (1-\lambda)\mathbf{u} : \lambda \in [0,1]\}$, where $\mathbf{w},\mathbf{u},\mathbf{a}',\mathbf{v}$ are distinct $L$-probability functions. If we take in place of a general concave function $f$ (in the proof of theorem 1.4.1) the Shannon entropy $H$ then we have that $\mathbf{ME}_L(\mathbf{K}_1) = \mathbf{v}$, $\mathbf{ME}_L(\mathbf{K}_2) = \mathbf{a}'$, $\mathbf{ME}_L(\mathbf{F}_1) = \mathbf{v}$ and $\mathbf{ME}_L(\mathbf{F}_2) = \mathbf{a}'$. Therefore

$$\mathbf{KIRP}_L(\mathbf{K}_1,\mathbf{K}_2) = \frac{e^{H(\mathbf{v})}}{e^{H(\mathbf{v})} + e^{H(\mathbf{a}')}}\mathbf{v} + \frac{e^{H(\mathbf{a}')}}{e^{H(\mathbf{v})} + e^{H(\mathbf{a}')}}\mathbf{a}' = \mathbf{KIRP}_L(\mathbf{F}_1,\mathbf{F}_2). \tag{4.21}$$

Now, if **KIRP** satisfied **(K4)** then $\mathbf{KIRP}_L(\mathbf{K}_1,\mathbf{K}_2) = \mathbf{w}$ and $\mathbf{KIRP}_L(\mathbf{F}_1,\mathbf{F}_2) = \mathbf{u}$ would contradict (4.21). □

One can also see that **KIRP** does not satisfy the collegiality principle **(K5)**:

**Example 4.4.4.** Let $\mathbf{K},\mathbf{F} \in CL$. Assume that $\mathbf{ME}_L(\mathbf{K}) \in V^L_{\mathbf{F}}$ but $\mathbf{ME}_L(\mathbf{F}) \neq \mathbf{ME}_L(\mathbf{K})$. Note that this is perfectly possible, for instance consider that the uniform probability function belongs to $V^L_{\mathbf{F}}$ but not to $V^L_{\mathbf{K}}$. Then clearly $\mathbf{KIRP}_L(\mathbf{K},\mathbf{F}) \neq \mathbf{ME}_L(\mathbf{K}) = \mathbf{KIRP}_L(\mathbf{K})$.

**Theorem 4.4.5. KIRP** *does satisfy* **LI**.

*Proof.* Let $L = L_1 \cup L_2$, $L_1 \cap L_2 = \emptyset$ and $\mathbf{K} \in CL_1$. Assume that $\mathbf{u} \in \mathbb{D}^{L_2}$ is the uniform probability function. Notice that by theorem 4.3.4 $\mathbf{ME}_{L_1}(\mathbf{K}) \cdot \mathbf{u} = \mathbf{ME}_L(\mathbf{K})$ so $H(\mathbf{ME}_{L_1}(\mathbf{K})) + H(\mathbf{u}) = H(\mathbf{ME}_L(\mathbf{K}))$. Therefore if $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL_1$ then

$$\mathbf{KIRP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)|_{L_1} = \sum_{k=1}^{n} \frac{e^{H(\mathbf{ME}_{L_1}(\mathbf{K}_k))} e^{H(\mathbf{u})}}{\sum_{k=1}^{n} e^{H(\mathbf{ME}_{L_1}(\mathbf{K}_k))} e^{H(\mathbf{u})}} \mathbf{ME}_L(\mathbf{K}_k)|_{L_1} =$$

$$= \sum_{k=1}^{n} \frac{e^{H(\mathbf{ME}_{L_1}(\mathbf{K}_k))}}{\sum_{k=1}^{n} e^{H(\mathbf{ME}_{L_1}(\mathbf{K}_k))}} \mathbf{ME}_{L_1}(\mathbf{K}_k) = \mathbf{KIRP}_{L_1}(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

□

On the other hand the following example shows that **KIRP** does not satisfy **IIP** and **CIIP**.

**Example 4.4.6.** Let $L = L_1 \cup L_2$, $L_1 = \{p\}$ and $L_2 = \{q\}$. Assume that $\mathbf{K}_1, \mathbf{K}_2$ are such that $V_{\mathbf{K}_1}^{L_1} = \{(\frac{1}{4}, \frac{3}{4})\}$ and $V_{\mathbf{K}_2}^{L_1} = \{(\frac{1}{2}, \frac{1}{2})\}$. Let $\mathbf{F}_1, \mathbf{F}_2 \in CL_2$ be such that $V_{\mathbf{F}_1}^{L_2} = \{(x, 1-x), \frac{1}{4} \leq x \leq \frac{1}{2}\}$ and $V_{\mathbf{F}_2}^{L_2} = \{(\frac{1}{4}, \frac{3}{4})\}$. Hence $\mathbf{ME}_{L_2}(\mathbf{F}_1) = (\frac{1}{2}, \frac{1}{2})$ and $H((\frac{1}{2}, \frac{1}{2})) = -\log \frac{1}{2} = h_1$ while $\mathbf{ME}_{L_2}(\mathbf{F}_2) = (\frac{1}{4}, \frac{3}{4})$ and $H((\frac{1}{4}, \frac{3}{4})) = -\frac{1}{4} \log \frac{1}{4} - \frac{3}{4} \log \frac{3}{4} = h_2$.

Now

$$\mathbf{KIRP}_L(\mathbf{K}_1 \cup \mathbf{F}_1, \mathbf{K}_2 \cup \mathbf{F}_2)(p) =$$
$$= \frac{e^{h_1} e^{h_2}}{2 \cdot e^{h_1} e^{h_2}} \frac{1}{4} \left( \frac{1}{2} + \frac{1}{2} \right) + \frac{e^{h_1} e^{h_2}}{2 \cdot e^{h_1} e^{h_2}} \frac{1}{2} \left( \frac{1}{4} + \frac{3}{4} \right) = \frac{3}{8},$$

where we have used theorem 4.3.4. On the other hand, since $h_1 \neq h_2$,

$$\mathbf{KIRP}_{L_1}(\mathbf{K}_1, \mathbf{K}_2)(p) =$$
$$= \frac{e^{h_2}}{e^{h_2} + e^{h_1}} \frac{1}{4} + \frac{e^{h_1}}{e^{h_2} + e^{h_1}} \frac{1}{2} \neq \frac{3}{8}.$$

□

Thus **KIRP**, like the obdurate merging operators previously considered does not satisfy the consistency principle, but unlike them it does not have the advantage of rationality when dealing with irrelevant information. Although for this reason one might find **KIRP** inconclusive, this process is an important example showing that not every naturally defined two stage process which applies an inference process at the first stage is necessary obdurate.

## 4.5 Relativisation

In section 1.3 we have mentioned the principle called *Relativisation* which was originally formulated in [39]. This is a fairly simple principle expressing how an inference process should treat a certain type of linear constraint. In this section we introduce a version for p-merging operators.

For $\theta \in SL$ and $\mathbf{w} \in \mathbb{D}^L$ we define $\mathbf{w}|_\theta$ by $\mathbf{w}|_\theta(\alpha) = \mathbf{w}(\alpha|\theta)$ for all $\alpha \in \mathrm{At}(L)$ if $\mathbf{w}(\theta) \neq 0$ and we leave it undefined otherwise. Note that given $\mathbf{w}(\theta) \neq 0$ and $\alpha \in \mathrm{At}(L)$ such that $\alpha \models \neg\theta$ we have that $\mathbf{w}|_\theta(\alpha) = 0$. Then for $\Delta \subseteq \mathbb{D}^L$ define

$$\Delta|_\theta = \{\mathbf{w}|_\theta \colon \mathbf{w} \in \Delta\}.$$

**Relativisation (REL).** Suppose that $\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_n \in CL$ are such that

$$\mathbf{K}_k = \{Bel(\theta) = c^{(k)}\} \cup \Big\{ \sum_s a_{si}^{(k)} Bel(\varphi_s|\theta) = d_i^{(k)},\ i = 1, \ldots, h^{(k)} \Big\},$$

$$\mathbf{F}_k = \mathbf{K}_k \cup \Big\{ \sum_s b_{si}^{(k)} Bel(\psi_s|\neg\theta) = e_i^{(k)},\ i = 1, \ldots, l^{(k)} \Big\},$$

where $0 < c^{(k)} < 1$, $\theta$ is fixed and $\varphi_s, \psi_s$ are $L$-sentences indexed over some index set. Let $\Delta$ be a p-merging operator. Then

$$\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)|_\theta = \Delta_L(\mathbf{F}_1, \ldots, \mathbf{F}_n)|_\theta.$$

This principle is fairly natural since if $\theta$ is true then $\mathbf{F}_1, \ldots, \mathbf{F}_n$ do not say anything more than $\mathbf{K}_1, \ldots, \mathbf{K}_n$ do respectively.

In the following we denote $\mathrm{At}(\theta) = \{j :\ \alpha_j \in \mathrm{At}(L), \alpha_j \models \theta\}$ and $\mathrm{At}(\neg\theta) = \{j : \alpha_j \in \mathrm{At}(L), \alpha_j \notin \mathrm{At}(\theta)\}$.

**Theorem 4.5.1.** $\hat{\Delta}^{\mathrm{KL}}$ *satisfies* **REL**.

*Proof.* Let $\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_n \in CL$ be as in the definition of **REL**. Assume that $\mathbf{w} \in \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and let $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in \hat{\Gamma}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ be such that $\mathbf{w} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$. By lemma 3.1.1 $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ minimise $\sum_{i=1}^n \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$ subject only to $\mathbf{w}^{(1)} \in V_{\mathbf{K}_1}^L, \ldots,\ \mathbf{w}^{(n)} \in V_{\mathbf{K}_n}^L$.

Now note that due to the constraints $\{Bel(\theta) = c^{(k)}\} \in \mathbf{K}_k$, $1 \leq k \leq n$, we have that $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ also minimise $\sum_{i=1}^n \sum_{j \in \mathrm{At}(\theta) \cap \mathrm{Sig}(\mathbf{w})} w_j^{(i)} \log \frac{n \cdot w_j^{(i)}}{\sum_{i=1}^n w_j^{(i)}}$, where $\mathbf{w} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$, subject to $\mathbf{w}^{(1)} \in V_{\mathbf{K}_1}^L, \ldots,\ \mathbf{w}^{(n)} \in V_{\mathbf{K}_n}^L$.

Let $\mathbf{v}^{(1)},\dots,\mathbf{v}^{(n)}$ be such that (i) they minimise

$$\sum_{k=1}^{n}\sum_{j\in \mathrm{At}(\neg\theta)\cap \mathrm{Sig}(\mathbf{v})} v_j^{(k)} \log \frac{n\cdot v_j^{(k)}}{\sum_{i=k}^{n} v_j^{(k)}},$$

where $\mathbf{v} = \mathbf{LinOp}(\mathbf{v}^{(1)},\dots,\mathbf{v}^{(n)})$, subject to $\mathbf{v}^{(1)} \in V^L_{\mathbf{F}_1},\dots,\ \mathbf{v}^{(n)} \in V^L_{\mathbf{F}_n}$, and (ii) $v_j^{(k)} = w_j^{(k)}$ for all $j \in \mathrm{At}(\theta)$ and $1 \le k \le n$. It follows that $\mathbf{v}^{(1)},\dots,\mathbf{v}^{(n)}$ minimise $\sum_{k=1}^{n} \mathrm{KL}(\mathbf{v}^{(k)} \| \mathbf{LinOp}(\mathbf{v}^{(1)},\dots,\mathbf{v}^{(n)})$ subject to $\mathbf{v}^{(1)} \in V^L_{\mathbf{F}_1},\dots,\mathbf{v}^{(n)} \in V^L_{\mathbf{F}_n}$ and hence $\mathbf{LinOp}(\mathbf{v}^{(1)},\dots,\mathbf{v}^{(n)}) \in \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{F}_1,\dots,\mathbf{F}_n)$.

We need to show that $\mathbf{LinOp}(\mathbf{w}^{(1)},\dots,\mathbf{w}^{(n)})|_\theta = \mathbf{LinOp}(\mathbf{v}^{(1)},\dots,\mathbf{v}^{(n)})|_\theta$. Let $j \in \mathrm{At}(\theta)$ and $\alpha_j$ be the corresponding $L$-atom. Then

$$\mathbf{LinOp}(\mathbf{w}^{(1)},\dots,\mathbf{w}^{(n)})|_\theta(\alpha_j) =$$

$$= \frac{\sum_{k=1}^{n} w_j^{(k)}}{\sum_{j\in\mathrm{At}(\theta)}\sum_{k=1}^{n} w_j^{(k)}} = \frac{\sum_{k=1}^{n} v_j^{(k)}}{\sum_{j\in\mathrm{At}(\theta)}\sum_{k=1}^{n} v_j^{(k)}} =$$

$$= \mathbf{LinOp}(\mathbf{v}^{(1)},\dots,\mathbf{v}^{(n)})|_\theta(\alpha_j).$$

Note that by $\{Bel(\theta) = c^{(k)}\} \in \mathbf{K}_k$, $1 \le k \le n$, $\sum_{j\in\mathrm{At}(\theta)}\sum_{k=1}^{n} w_j^{(k)} \neq 0$. Moreover, since $\mathbf{v}^{(k)}(\alpha_j|\theta) = \mathbf{w}^{(k)}(\alpha_j|\theta) = 0$ for $j \in \mathrm{At}(\neg\theta)$ and $1 \le k \le n$, we can conclude that $\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1,\dots,\mathbf{K}_n)|_\theta \subseteq \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{F}_1,\dots,\mathbf{F}_n)|_\theta$.

The proof for $\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{F}_1,\dots,\mathbf{F}_n)|_\theta \subseteq \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1,\dots,\mathbf{K}_n)|_\theta$ is similar. □

**Theorem 4.5.2.** *$\hat{\Delta}^{D_r}$, $2 \ge r > 1$, satisfies* **REL**.

*Proof.* The proof is analogous to the one above. The crucial difference is only that the quantity to minimise is

$$\sum_{k=1}^{n}\sum_{j\in\mathrm{At}(\neg\theta)}\Big[(w_j^{(k)})^r - \Big(\frac{\sum_{k=1}^{n} w_j^{(k)}}{n}\Big)^r - r\Big(w_j^{(k)} - \frac{\sum_{k=1}^{n} w_j^{(k)}}{n}\Big)\Big(\frac{\sum_{k=1}^{n} w_j^{(k)}}{n}\Big)^{r-1}\Big].$$

However, since the constraints $\{Bel(\theta) = c^{(k)}\} \in \mathbf{K}_k$, $1 \le k \le n$, have the same effect on this quantity as on the sum of KL-divergences, the theorem follows for $\hat{\Delta}^{D_r}$-operator as well. □

The following result is in fact a consequence of a more general theorem 3.9 in [53].

**Theorem 4.5.3.** *The $\Delta^{\mathrm{KL}}$-merging operator satisfies* **REL** *when its definition is restricted so that condition (1.4) is satisfied for any input of knowledge bases.*

*Proof.* Let $\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_n \in CL$ be as in the definition of **REL** and moreover assume that $\mathbf{K}_1, \ldots, \mathbf{K}_n$ and $\mathbf{F}_1, \ldots, \mathbf{F}_n$ satisfy condition (1.4) respectively. Assume that $\mathbf{w} \in \Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and let $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in \Gamma_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ be such that $\mathbf{w} = \mathbf{LogOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$. By lemma 3.1.4, $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ maximise $\sum_{j=1}^{J}(\prod_{k=1}^{n} w_j^{(k)})^{\frac{1}{n}}$ subject only to $\mathbf{w}^{(1)} \in V_{\mathbf{K}_1}^L, \ldots,\ \mathbf{w}^{(n)} \in V_{\mathbf{K}_n}^L$.

Now note that due to the constraints $\{Bel(\theta) = c^{(k)}\} \in \mathbf{K}_k$, $1 \le k \le n$, we have that $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ also maximise $\sum_{j \in \mathrm{At}(\theta)}(\prod_{k=1}^{n} w_j^{(k)})^{\frac{1}{n}}$ subject to $\mathbf{w}^{(1)} \in V_{\mathbf{K}_1}^L, \ldots,$ $\mathbf{w}^{(n)} \in V_{\mathbf{K}_n}^L$.

Let $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}$ be such that (i) they maximise $\sum_{j \in \mathrm{At}(\neg\theta)}(\prod_{k=1}^{n} v_j^{(k)})^{\frac{1}{n}}$ subject to $\mathbf{v}^{(1)} \in V_{\mathbf{F}_1}^L, \ldots,\ \mathbf{v}^{(n)} \in V_{\mathbf{F}_n}^L$, and (ii) $v_j^{(k)} = w_j^{(k)}$ for $j \in \mathrm{At}(\theta)$ and $1 \le k \le n$. It follows that $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}$ maximise $\sum_{j=1}^{J}(\prod_{k=1}^{n} v_j^{(k)})^{\frac{1}{n}}$ subject to $\mathbf{v}^{(1)} \in V_{\mathbf{F}_1}^L, \ldots, \mathbf{v}^{(n)} \in V_{\mathbf{F}_n}^L$ and hence $\mathbf{LogOp}(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}) \in \Delta_L^{\mathrm{KL}}(\mathbf{F}_1, \ldots, \mathbf{F}_n)$.

We need to show that $\mathbf{LogOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})|_\theta = \mathbf{LogOp}(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)})|_\theta$. Let $j \in \mathrm{At}(\theta)$ and $\alpha_j$ be the corresponding $L$-atom. Then

$$\mathbf{LogOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})|_\theta(\alpha_j) = \frac{\frac{(\prod_{k=1}^{n} w_j^{(k)})^{\frac{1}{n}}}{\sum_{j \in \mathrm{At}(\theta)}(\prod_{k=1}^{n} w_j^{(k)})^{\frac{1}{n}} + \sum_{j \in \mathrm{At}(\neg\theta)}(\prod_{k=1}^{n} w_j^{(k)})^{\frac{1}{n}}}}{\frac{\sum_{j \in \mathrm{At}(\theta)}(\prod_{k=1}^{n} w_j^{(k)})^{\frac{1}{n}}}{\sum_{j \in \mathrm{At}(\theta)}(\prod_{k=1}^{n} w_j^{(k)})^{\frac{1}{n}} + \sum_{j \in \mathrm{At}(\neg\theta)}(\prod_{k=1}^{n} w_j^{(k)})^{\frac{1}{n}}}} =$$

$$= \frac{(\prod_{k=1}^{n} w_j^{(k)})^{\frac{1}{n}}}{\sum_{j \in \mathrm{At}(\theta)}(\prod_{k=1}^{n} w_j^{(k)})^{\frac{1}{n}}}.$$

Note that by $\{Bel(\theta) = c^{(k)}\} \in \mathbf{K}_k$, $1 \le k \le n$, $\sum_{j \in \mathrm{At}(\theta)}(\prod_{k=1}^{n} w_j^{(k)})^{\frac{1}{n}} \neq 0$. Accordingly,

$$\mathbf{LogOp}(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)})|_\theta(\alpha_j) = \frac{(\prod_{k=1}^{n} v_j^{(k)})^{\frac{1}{n}}}{\sum_{j \in \mathrm{At}(\theta)}(\prod_{k=1}^{n} v_j^{(k)})^{\frac{1}{n}}}.$$

Since $v_j^{(k)} = w_j^{(k)}$ for $j \in \mathrm{At}(\theta)$ and $1 \le k \le n$, the two values above are equal. Moreover, since $\mathbf{v}^{(k)}(\alpha_j|\theta) = \mathbf{w}^{(k)}(\alpha_j|\theta) = 0$ for $j \in \mathrm{At}(\neg\theta)$ and $1 \le k \le n$, we can conclude that $\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)|_\theta \subseteq \Delta_L^{\mathrm{KL}}(\mathbf{F}_1, \ldots, \mathbf{F}_n)|_\theta$.

The proof for $\Delta_L^{\mathrm{KL}}(\mathbf{F}_1, \ldots, \mathbf{F}_n)|_\theta \subseteq \Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)|_\theta$ is similar. $\square$

The following example shows that condition (1.4) in theorem above is necessary.

**Example 4.5.4.** Let $L = \{a, b\}$. Assume that the knowledge of the first expert consists of $\mathbf{K}_1 = \{Bel(a) = \frac{1}{2},\ Bel(a \wedge b \mid a) = 0\}$ and $\mathbf{F}_1 = \{Bel(\neg a \wedge b \mid \neg a) = 0\}$, and the knowledge of the second expert consists of $\mathbf{K}_2 = \{Bel(a) = \frac{1}{2},\ Bel(a \wedge \neg b \mid a) = 0\}$ and $\mathbf{F}_2 = \{Bel(\neg a \wedge \neg b \mid \neg a) = 0\}$. Then $V^L_{\mathbf{K}_1} = \{(0, \frac{1}{2}, x, \frac{1}{2} - x),\ x \in [0, \frac{1}{2}]\}$, $V^L_{\mathbf{K}_1 \cup \mathbf{F}_1} = \{(0, \frac{1}{2}, 0, \frac{1}{2})\}$, $V^L_{\mathbf{K}_2} = \{(\frac{1}{2}, 0, x, \frac{1}{2} - x),\ x \in [0, \frac{1}{2}]\}$ and $V^L_{\mathbf{K}_2 \cup \mathbf{F}_2} = \{(\frac{1}{2}, 0, \frac{1}{2}, 0)\}$, where the atoms in probability functions are listed as follows: $a \wedge b$, $a \wedge \neg b$, $\neg a \wedge b$ and $\neg a \wedge \neg b$.

Now, while $\Delta^{\mathrm{KL}}_L(\mathbf{K}_1 \cup \mathbf{F}_1, \mathbf{K}_2 \cup \mathbf{F}_2) = \mathbb{D}^L$, for all $\mathbf{v} \in \Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \mathbf{K}_2)$ we have that $\mathbf{v}(a) = 0$ so the principle of relativisation does not even make sense for such knowledge bases. □

**Theorem 4.5.5.** **OLEP** *and* **OSEP** *satisfy* **REL**.

*Proof.* First, recall that **ME** satisfies the Relativisation principle. Therefore, if $\mathbf{K}_1, \dots, \mathbf{K}_n, \mathbf{F}_1, \dots, \mathbf{F}_n \in CL$ are as in the definition of **REL** and $\alpha \models \theta$ is an $L$-atom then

$$\frac{\mathbf{ME}_L(\mathbf{K}_i)(\alpha)}{\mathbf{ME}_L(\mathbf{K}_i)(\theta)} = \mathbf{ME}_L(\mathbf{K}_i)|_\theta(\alpha) = \mathbf{ME}_L(\mathbf{F}_i)|_\theta(\alpha) = \frac{\mathbf{ME}_L(\mathbf{F}_i)(\alpha)}{\mathbf{ME}_L(\mathbf{F}_i)(\theta)}. \tag{4.22}$$

Due to $\{Bel(\theta) = c^{(i)}\} \in \mathbf{K}_i$ and $\{Bel(\theta) = c^{(i)}\} \in \mathbf{F}_i$ we have that $\mathbf{ME}_L(\mathbf{K}_i)(\theta) = \mathbf{ME}_L(\mathbf{F}_i)(\theta) \neq 0$ and (4.22) gives $\mathbf{ME}_L(\mathbf{K}_i)(\alpha) = \mathbf{ME}_L(\mathbf{F}_i)(\alpha)$ for all $1 \leq i \leq n$.

Now,

$$\mathbf{OLEP}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)|_\theta(\alpha) = \mathbf{LinOp}(\mathbf{ME}_L(\mathbf{K}_1), \dots, \mathbf{ME}_L(\mathbf{K}_n))|_\theta(\alpha) =$$
$$= \frac{\sum_{i=1}^n \mathbf{ME}_L(\mathbf{K}_i)(\alpha)}{\sum_{\alpha \models \theta} \sum_{i=1}^n \mathbf{ME}_L(\mathbf{K}_i)(\alpha)} = \frac{\sum_{i=1}^n \mathbf{ME}_L(\mathbf{F}_i)(\alpha)}{\sum_{\alpha \models \theta} \sum_{i=1}^n \mathbf{ME}_L(\mathbf{F}_i)(\alpha)} =$$
$$= \mathbf{LinOp}(\mathbf{ME}_L(\mathbf{F}_1), \dots, \mathbf{ME}_L(\mathbf{F}_n))|_\theta(\alpha) = \mathbf{OLEP}_L(\mathbf{F}_1, \dots, \mathbf{F}_n)|_\theta(\alpha).$$

Similarly, given that $\mathbf{K}_1, \dots, \mathbf{K}_n$ and $\mathbf{F}_1, \dots, \mathbf{F}_n$ satisfy condition (1.4) respectively,

$$\mathbf{OSEP}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)|_\theta = \mathbf{LogOp}(\mathbf{ME}_L(\mathbf{K}_1), \dots, \mathbf{ME}_L(\mathbf{K}_n))|_\theta$$

and

$$\mathbf{OSEP}_L(\mathbf{F}_1, \dots, \mathbf{F}_n)|_\theta = \mathbf{LogOp}(\mathbf{ME}_L(\mathbf{F}_1), \dots, \mathbf{ME}_L(\mathbf{F}_n))|_\theta.$$

Let $\alpha \models \theta$ be an $L$-atom. Then

$$\mathbf{LogOp}(\mathbf{ME}_L(\mathbf{K}_1), \dots, \mathbf{ME}_L(\mathbf{K}_n))|_\theta(\alpha) = \frac{\frac{(\prod_{i=1}^n \mathbf{ME}_L(\mathbf{K}_i)(\alpha))^{\frac{1}{n}}}{\sum_{\alpha \in \mathrm{At}(L)} (\prod_{i=1}^n \mathbf{ME}_L(\mathbf{K}_i)(\alpha))^{\frac{1}{n}}}}{\frac{\sum_{\substack{\alpha \models \theta \\ \alpha \in \mathrm{At}(L)}} (\prod_{i=1}^n \mathbf{ME}_L(\mathbf{K}_i)(\alpha))^{\frac{1}{n}}}{\sum_{\alpha \in \mathrm{At}(L)} (\prod_{i=1}^n \mathbf{ME}_L(\mathbf{K}_i)(\alpha))^{\frac{1}{n}}}}$$

and

$$\mathbf{LogOp}(\mathbf{ME}_L(\mathbf{F}_1),\ldots,\mathbf{ME}_L(\mathbf{F}_n))|_\theta(\alpha) = \frac{(\prod_{i=1}^n \mathbf{ME}_L(\mathbf{F}_i)(\alpha))^{\frac{1}{n}}}{\sum_{\substack{\alpha\models\theta\\ \alpha\in \mathrm{At}(L)}} (\prod_{i=1}^n \mathbf{ME}_L(\mathbf{F}_i)(\alpha))^{\frac{1}{n}}}.$$

Since $\mathbf{ME}_L(\mathbf{K}_i)(\alpha) = \mathbf{ME}_L(\mathbf{F}_i)(\alpha)$ for all $1 \le i \le n$ the theorem follows. □

The following example shows that **KIRP** does not satisfy relativisation.

**Example 4.5.6.** Let $L = \{a, b\}$. Assume that the knowledge of the first expert consists of $\mathbf{K}_1 = \{Bel(a) = \frac{1}{2},\ Bel(a \wedge b \mid a) = \frac{1}{2}\}$ and $\mathbf{F}_1 = \{Bel(\neg a \wedge b \mid \neg a) = \frac{1}{4}\}$, and the knowledge of the second expert consists of $\mathbf{K}_2 = \{Bel(a) = \frac{1}{2},\ Bel(a\wedge b \mid a) = \frac{1}{4}\}$. Then $\mathbf{ME}_L(\mathbf{K}_1) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, $\mathbf{ME}_L(\mathbf{K}_1 \cup \mathbf{F}_1) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{8}, \frac{3}{8})$ and $\mathbf{ME}_L(\mathbf{K}_2) = (\frac{1}{8}, \frac{3}{8}, \frac{1}{4}, \frac{1}{4})$, where the atoms in probability functions are listed as follows: $a\wedge b$, $a\wedge\neg b$, $\neg a \wedge b$ and $\neg a \wedge \neg b$. Let $H((\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}) = h_1$ and $H((\frac{1}{8}, \frac{3}{8}, \frac{1}{4}, \frac{1}{4})) = H((\frac{1}{4}, \frac{1}{4}, \frac{1}{8}, \frac{3}{8})) = h_2$.

Now

$$\mathbf{KIRP}_L(\mathbf{K}_1, \mathbf{K}_2)(a \wedge b \mid a) =$$

$$\frac{\frac{e^{h_1}}{e^{h_1}+e^{h_2}}\mathbf{ME}_L(\mathbf{K}_1)(a\wedge b) + \frac{e^{h_2}}{e^{h_1}+e^{h_2}}\mathbf{ME}_L(\mathbf{K}_2)(a\wedge b)}{\frac{e^{h_1}}{e^{h_1}+e^{h_2}}\mathbf{ME}_L(\mathbf{K}_1)(a) + \frac{e^{h_2}}{e^{h_1}+e^{h_2}}\mathbf{ME}_L(\mathbf{K}_2)(a)} =$$

$$= 2\frac{e^{h_1}}{e^{h_1}+e^{h_2}}\frac{1}{4} + 2\frac{e^{h_2}}{e^{h_1}+e^{h_2}}\frac{1}{8} \tag{4.23}$$

and

$$\mathbf{KIRP}_L(\mathbf{K}_1 \cup \mathbf{F}_1, \mathbf{K}_2)(a \wedge b \mid a) =$$

$$\frac{\frac{e^{h_2}}{e^{h_2}+e^{h_2}}\mathbf{ME}_L(\mathbf{K}_1\cup\mathbf{F}_1)(a\wedge b) + \frac{e^{h_2}}{e^{h_2}+e^{h_2}}\mathbf{ME}_L(\mathbf{K}_2)(a\wedge b)}{\frac{e^{h_2}}{e^{h_2}+e^{h_2}}\mathbf{ME}_L(\mathbf{K}_1\cup\mathbf{F}_1)(a) + \frac{e^{h_2}}{e^{h_2}+e^{h_2}}\mathbf{ME}_L(\mathbf{K}_2)(a)} = \frac{1}{4} + \frac{1}{8}. \tag{4.24}$$

Since $e^{h_2} \neq e^{h_1}$ the quantities (4.23) and (4.24) are not equal. □

As it is shown in the following example, also HULL does not satisfy **REL**.

**Example 4.5.7.** Let $L = \{a, b\}$. Assume that the knowledge of the first expert consists of $\mathbf{K}_1 = \{Bel(a) = \frac{1}{2},\ Bel(a \wedge b \mid a) = \frac{1}{2}\}$ and $\mathbf{F}_1 = \{Bel(\neg a \wedge b \mid \neg a) = \frac{1}{2}\}$, and the knowledge of the second expert consists of $\mathbf{K}_2 = \{Bel(a) = \frac{1}{2}\}$ and $\mathbf{F}_2 = \{Bel(\neg a \wedge b \mid \neg a) = \frac{1}{4}\}$. Then $V^L_{\mathbf{K}_1} = \{(\frac{1}{4}, \frac{1}{4}, x, \frac{1}{2} - x),\ x \in [0, \frac{1}{2}]\}$, $V^L_{\mathbf{K}_1\cup\mathbf{F}_1} = \{(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})\}$, $V^L_{\mathbf{K}_2} = \{(y, \frac{1}{2} - y, x, \frac{1}{2} - x),\ y \in [0, \frac{1}{2}], x \in [0, \frac{1}{2}]\}$ and $V^L_{\mathbf{K}_2\cup\mathbf{F}_2} = \{(y, \frac{1}{2} - y, \frac{1}{8}, \frac{3}{8}),\ y \in [0, \frac{1}{2}]\}$, where the atoms in probability functions are listed as follows: $a \wedge b$, $a \wedge \neg b$, $\neg a \wedge b$ and $\neg a \wedge \neg b$.

Since $\mathbf{K}_1$ and $\mathbf{K}_2$ are jointly consistent we have that

$$\mathrm{HULL}_L(\mathbf{K}_1, \mathbf{K}_2) = V^L_{\mathbf{K}_1} \cap V^L_{\mathbf{K}_2} = \Big\{\Big(\frac{1}{4}, \frac{1}{4}, x, \frac{1}{2} - x\Big),\ x \in \Big[0, \frac{1}{2}\Big]\Big\}.$$

On the other hand $\mathbf{K}_1 \cup \mathbf{F}_1$ and $\mathbf{K}_2 \cup \mathbf{F}_2$ are inconsistent and hence

$$\Big(y, \frac{1}{2} - y, \frac{1}{8}, \frac{3}{8}\Big) \in V^L_{\mathbf{K}_2 \cup \mathbf{F}_2} \subset \mathrm{HULL}_L(\mathbf{K}_1 \cup \mathbf{F}_1, \mathbf{K}_2 \cup \mathbf{F}_2),$$

say for $y = 0$. But certainly $(0, \frac{1}{2}, \frac{1}{8}, \frac{3}{8})|_a \notin \{(\frac{1}{4}, \frac{1}{4}, x, \frac{1}{2} - x),\ x \in [0, \frac{1}{2}]\}|_a$. $\square$

# Chapter 5

# Making a merging process single-valued

## 5.1 Chairman theorems

We started our quest for multi-expert merging procedures by observing that the most naive generalisations of single-expert inference processes do not appear very appealing once the principle of consistency is considered. In chapter 3 we examined the linear entropy operator $\hat{\Delta}^{\mathrm{KL}}$ and the social entropy operator $\Delta^{\mathrm{KL}}$ and argued that, when restricted to $BCL$, they naturally occur as the fixed points of processes of knowledge updating and pooling. In chapter 4 we have seen that when confined to $BCL$ both $\hat{\Delta}^{\mathrm{KL}}$ and $\Delta^{\mathrm{KL}}$ possess relatively attractive properties when compared to other known p-merging operators.

In this section we will focus only on these two operators $\hat{\Delta}^{\mathrm{KL}}$ and $\Delta^{\mathrm{KL}}$ and we investigate their susceptibility to a small bias by the uniform probability function — the most uninformative point of $\mathbb{D}^L$. The study of this problem first occurred in [52] where Wilmers argued that an independent adjudicator, whose only knowledge consists of what is related to him by the given college of experts, can rationally bias the agreement procedure by including himself as an additional expert, whose personal probability function is the uniform one, in order to calculate a single social probability function; and then find what would happen to this social probability function if his contribution were happen to be infinitesimally small relative to that of the other experts. He showed that in the case of the $\Delta^{\mathrm{KL}}$-merging operator this point of agreement is characterised by applying the **ME**-inference process to the region defined by $\Delta^{\mathrm{KL}}$. In what follows we adapt this result and show that in the case of the $\hat{\Delta}^{\mathrm{KL}}$-merging operator the corresponding point can be characterised by applying the $\mathbf{CM}^{\infty}$-inference process to the region defined by $\hat{\Delta}^{\mathrm{KL}}$.

**Theorem 5.1.1.** *Let $\mathbf{K}_1, \dots, \mathbf{K}_n \in BCL$ be such that for at least one $i$ $V^L_{\mathbf{K}_i}$ is a singleton. Then $\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ is a singleton.*

*Proof.* Without loss of generality assume that $V^L_{\mathbf{K}_1} = \{\mathbf{v}\}$. For a contradiction suppose that $\mathbf{w}, \mathbf{r} \in \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ and $\mathbf{w} \neq \mathbf{r}$. Denote $\mathbf{w}^{(2)}, \dots, \mathbf{w}^{(n)}$ the KL-projections of $\mathbf{w}$ into $V^L_{\mathbf{K}_2}, \dots, V^L_{\mathbf{K}_n}$ respectively and $\mathbf{r}^{(2)}, \dots, \mathbf{r}^{(n)}$ the KL-projections of $\mathbf{r}$ into $V^L_{\mathbf{K}_2}, \dots, V^L_{\mathbf{K}_n}$ respectively. By corollary 3.2.6 $\mathbf{w} = \mathbf{LinOp}(\mathbf{v}, \mathbf{w}^{(2)}, \dots, \mathbf{w}^{(n)})$ and $\mathbf{r} = \mathbf{LinOp}(\mathbf{v}, \mathbf{r}^{(2)}, \dots, \mathbf{r}^{(n)})$.

Now consider $\mathbf{x} = \lambda\mathbf{w} + (1-\lambda)\mathbf{r}$ for some $\lambda \in (0,1)$. By theorem 3.1.2 we have that $\mathbf{x} \in \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$. Since $\mathrm{KL}(\cdot\|\cdot)$ is a convex function, by the Jensen inequality we have that

$$\mathrm{KL}(\mathbf{v}\|\mathbf{x}) + \sum_{i=2}^{n} \mathrm{KL}(\lambda\mathbf{w}^{(i)} + (1-\lambda)\mathbf{r}^{(i)}\|\mathbf{x}) \leq$$

$$\leq \lambda\Big(\mathrm{KL}(\mathbf{v}\|\mathbf{w}) + \sum_{i=2}^{n} \mathrm{KL}(\mathbf{w}^{(i)}\|\mathbf{w})\Big) + (1-\lambda)\Big(\mathrm{KL}(\mathbf{v}\|\mathbf{r}) + \sum_{i=2}^{n} \mathrm{KL}(\mathbf{r}^{(i)}\|\mathbf{r})\Big) =$$

$$= \hat{\mathrm{M}}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n). \tag{5.1}$$

However, since $\mathbf{w}, \mathbf{r}, \mathbf{x} \in \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ and $\lambda\mathbf{w}^{(i)} + (1-\lambda)\mathbf{r}^{(i)} \in V^L_{\mathbf{K}_i}$, $1 \leq i \leq n$, the above is possible only with the equality.

On the other hand, by lemma 2.2.1, the following Jensen inequality is strict:

$$\mathrm{KL}(\mathbf{v}\|\mathbf{x}) < \lambda\,\mathrm{KL}(\mathbf{v}\|\mathbf{w}) + (1-\lambda)\,\mathrm{KL}(\mathbf{v}\|\mathbf{r}).$$

Note that the border points $\lambda = 0, 1$ are excluded. Therefore (5.1) yields

$$\sum_{i=2}^{n} \mathrm{KL}(\lambda\mathbf{w}^{(i)} + (1-\lambda)\mathbf{r}^{(i)}\|\mathbf{x}) >$$

$$> \lambda\Big(\sum_{i=2}^{n} \mathrm{KL}(\mathbf{w}^{(i)}\|\mathbf{w})\Big) + (1-\lambda)\Big(\sum_{i=2}^{n} \mathrm{KL}(\mathbf{r}^{(i)}\|\mathbf{r})\Big).$$

However this contradicts the Jensen inequality. □

Note, that example 4.2.2 shows that in the theorem above it is not possible to replace $BCL$ by $CL$.

**Theorem 5.1.2** (Chairman Theorem for $\hat{\Delta}^{\mathrm{KL}}$)**.** *Let $\mathbf{I} \in BCL$ be such that $V^L_{\mathbf{I}} = \{\mathbf{t}\}$, where $\mathbf{t} = (\frac{1}{J}, \dots, \frac{1}{J})$ and $J = |\operatorname{At}(L)|$. Let $\mathbf{K}_1, \dots, \mathbf{K}_n \in BCL$. Define*

$$\{\mathbf{v}^{[m]}\} = \hat{\Delta}^{\mathrm{KL}}_L(\underbrace{\mathbf{K}_1, \dots, \mathbf{K}_1}_{m}, \dots, \underbrace{\mathbf{K}_n, \dots, \mathbf{K}_n}_{m}, \mathbf{I}),$$

*for all $m = 1, 2, \dots$ . Then $\{\mathbf{v}^{[m]}\}_{m=1}^{\infty}$ converges and*

$$\lim_{m\to\infty} \mathbf{v}^{[m]} = \mathbf{CM}^{\infty}_L(\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)).$$

*Proof.* First of all recall that $\hat{\mathrm{M}}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ denotes the minimal value of

$$\sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{v})$$

subject to $\mathbf{w}^{(i)} \in V^L_{\mathbf{K}_i}$, $1 \leq i \leq n$, and $\mathbf{v} \in \mathbb{D}^L$. Furthermore, we denote by $E_m$ the minimal value of

$$m \sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{v}) - m\hat{\mathrm{M}}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n) + \mathrm{KL}(\mathbf{t} \| \mathbf{v}) \tag{5.2}$$

subject to $\mathbf{w}^{(i)} \in V^L_{\mathbf{K}_i}$, $1 \leq i \leq n$, and $\mathbf{v} \in \mathbb{D}^L$. By the definition of $\hat{\mathrm{M}}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ we have that $0 \leq E_m$ for all $m = 1, 2, \dots$ . Notice that for $(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}) \in \hat{\Gamma}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ and $\mathbf{v} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$ (5.2) becomes $\mathrm{KL}(\mathbf{t} \| \mathbf{v})$.

Note that for a fixed $m$ if $\mathbf{v} \in \mathbb{D}^L$ globally minimises (5.2) subject to $\mathbf{w}^{(i)} \in V^L_{\mathbf{K}_i}$, $1 \leq i \leq n$, then $\mathbf{v} \in \hat{\Delta}^{\mathrm{KL}}_L(\underbrace{\mathbf{K}_1, \dots, \mathbf{K}_1}_{m}, \dots, \underbrace{\mathbf{K}_n, \dots, \mathbf{K}_n}_{m}, \mathbf{I})$ (by theorem 5.1.1 such a $\mathbf{v}$ is unique), and conversely if $\mathbf{v} \in \hat{\Delta}^{\mathrm{KL}}_L(\underbrace{\mathbf{K}_1, \dots, \mathbf{K}_1}_{m}, \dots, \underbrace{\mathbf{K}_n, \dots, \mathbf{K}_n}_{m}, \mathbf{I})$ then $\mathbf{v}$ minimises (5.2) subject to above constraints.

Now let $\mathbf{r} = \mathbf{CM}^{\infty}_L(\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n))$. Since $\mathbf{K}_1, \dots, \mathbf{K}_n \in BCL$ this means that

$$\mathbf{r} = \arg \max_{\mathbf{v} \in \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)} \sum_{j=1}^{J} \log v_j = \arg \min_{\mathbf{v} \in \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)} \mathrm{KL}(\mathbf{t} \| \mathbf{v}).$$

Since $\mathbf{r} \in \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ it follows that for all $m \in \mathbb{N}$

$$E_m \leq \mathrm{KL}(\mathbf{t} \| \mathbf{r}). \tag{5.3}$$

Since $\mathbb{D}^L \subseteq \mathbb{R}^J$ is a compact space there exist a convergent subsequence of the sequence $\{\mathbf{v}^{[m]}\}_{m=1}^{\infty}$. Let $\{\mathbf{v}^{[m_k]}\}_{k=1}^{\infty}$ be any such convergent subsequence. Let

$$(\underbrace{\mathbf{w}^{(1)[m_k]}, \dots, \mathbf{w}^{(1)[m_k]}}_{m_k} \dots \underbrace{\mathbf{w}^{(n)[m_k]}, \dots, \mathbf{w}^{(n)[m_k]}}_{m_k}, \mathbf{t}) \in \hat{\Gamma}^{\mathrm{KL}}_L(\underbrace{\mathbf{K}_1, \dots, \mathbf{K}_1}_{m_k} \dots \underbrace{\mathbf{K}_n, \dots, \mathbf{K}_n}_{m_k}, \mathbf{I})$$

be associated with $\mathbf{v}^{[m_k]}$ for all $k \in \mathbb{N}$. Recall that these are unique for every $\mathbf{v}^{[m_k]}$ because $\mathbf{w}^{(i)[m_k]}$ corresponds to the KL-projection of $\mathbf{v}^{[m_k]}$ to $V^L_{\mathbf{K}_i}$, $1 \leq i \leq n$.

Now the sequence

$$\{\sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)[m_k]} \| \mathbf{v}^{[m_k]})\}_{k=1}^{\infty} \tag{5.4}$$

is bounded in $\mathbb{R}$. That is because we are operating in $BCL$ and therefore the coordinates of all $\mathbf{w}^{(i)[m_k]}$, $1 \leq i \leq n$ and $k = 1, 2, \dots$, are bounded away from zero by some constants. And, since $\mathbf{v}^{[m_k]} = \mathbf{LinOp}(\mathbf{w}^{(1)[m_k]}, \dots, \mathbf{w}^{(n)[m_k]})$, we can find some global bound away from zero for the coordinates of all $\mathbf{v}^{[m_k]}$, $k = 1, 2, \dots$. Consequently the sequence (5.4) also has a convergent subsequence. If the limit of any such convergent subsequence was not equal to $\hat{\mathrm{M}}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ then there would be eventually an $m$ which would contradict (5.3).

Now we show that existence of a convergent subsequence and the fact that every convergent subsequence has the limit $\hat{\mathrm{M}}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ implies that $\{\sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)[m_k]} \| \mathbf{v}^{[m_k]})\}_{k=1}^{\infty}$ converges and that

$$\lim_{k \to \infty} \sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)[m_k]} \| \mathbf{v}^{[m_k]}) = \hat{\mathrm{M}}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n). \tag{5.5}$$

Assume that (5.5) is not the case. Then there is an open neighbourhood of the point $\hat{\mathrm{M}}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ outside of which there is an infinite number of the members of the sequence $\{\sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)[m_k]} \| \mathbf{v}^{[m_k]})\}_{k=1}^{\infty}$. By the same compactness argument as before we have that there is a convergent subsequence with the limit distinct from $\hat{\mathrm{M}}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ which contradicts our previous claim.

Note that we already know that $\lim_{k\to\infty} \mathbf{v}^{[m_k]}$ exists and we can denote it by $\mathbf{v}$. However we do not know whether the same is true for $\lim_{k\to\infty} \mathbf{w}^{(i)[m_k]}$, $1 \leq i \leq n$. On the other hand since $V^L_{\mathbf{K}_1}, \dots, V^L_{\mathbf{K}_n}$ are compact there is a subsequence $\{m_l\}_{l=1}^{\infty}$ of the sequence of indices $\{m_k\}_{k=1}^{\infty}$ such that for all $1 \leq i \leq n$ the subsequence $\{\mathbf{w}^{(i)[m_l]}\}_{l=1}^{\infty}$ of the sequence $\{\mathbf{w}^{(i)[m_k]}\}_{k=1}^{\infty}$ is convergent. Let us denote the corresponding limits $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}$. Note that also $\lim_{l\to\infty} \mathbf{v}^{[m_l]} = \mathbf{v}$. Since $\mathrm{KL}(\cdot\|\cdot)$ is a continuous function in both variables (we are confined to $BCL$) the value of $\sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{v})$ is equal to the limit $\lim_{l\to\infty} \sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)[m_l]} \| \mathbf{v}^{[m_l]})$ which, by (5.5), is $\hat{\mathrm{M}}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$. But this means that we have found a minimiser of $\sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{v})$ subject to $\mathbf{w}^{(i)} \in V^L_{\mathbf{K}_i}$, $1 \leq i \leq n$, and $\mathbf{v} \in \mathbb{D}^L$.

It follows that $\mathbf{v} = \lim_{k\to\infty} \mathbf{v}^{[m_k]} \in \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \dots, \mathbf{K}_n)$ and therefore $\lim_{k\to\infty} E_{m_k} = \mathrm{KL}(\mathbf{t} \| \lim_{k\to\infty} \mathbf{v}^{[m_k]})$ which, by the definition of $\mathbf{CM}^\infty$, is greater or equal to $\mathrm{KL}(\mathbf{t}\|\mathbf{r})$. However, by (5.3) this is possible only if $\lim_{k\to\infty} \mathbf{v}^{[m_k]} = \mathbf{r}$.

In fact we have proved that every convergent subsequence of $\{\mathbf{v}^{[m]}\}_{m=1}^\infty$ has $\mathbf{r}$ as the limit. Now exactly as several lines above we prove that this implies convergence of whole sequence $\{\mathbf{v}^{[m]}\}_{m=1}^\infty$ and that $\lim_{m\to\infty} \mathbf{v}^{[m]} = \mathbf{r}$. Assume that this is not the case. Then there is an open neighbourhood of the point $\mathbf{r}$ outside of which there is an infinite number of the members of the sequence $\{\mathbf{v}^{[m]}\}_{m=1}^\infty$. Since then these lie in a compact space there is a convergent subsequence among them with limit distinct from $\mathbf{r}$ which contradicts our previous claim. $\square$

The following theorem is proved in [53].

**Theorem 5.1.3.** *Let $\mathbf{K}_1, \dots, \mathbf{K}_n \in CL$ be such that condition (1.4) holds and for at least one $i$ $V_{\mathbf{K}_i}^L$ is a singleton. Then $\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \dots, \mathbf{K}_n)$ is a singleton.*

The following theorem is a significant result of Wilmers, see [53].

**Theorem 5.1.4** (Chairman Theorem for $\Delta^{\mathrm{KL}}$). *Let $\mathbf{I} \in CL$ be such that $V_{\mathbf{I}}^L = \{\mathbf{t}\}$, where $\mathbf{t} = (\frac{1}{J}, \dots, \frac{1}{J})$ and $J = |\mathrm{At}(L)|$. Let $\mathbf{K}_1, \dots, \mathbf{K}_n \in CL$ be such that condition (1.4) holds. Define*

$$\{\mathbf{v}^{[m]}\} = \Delta_L^{\mathrm{KL}}(\underbrace{\mathbf{K}_1, \dots, \mathbf{K}_1}_{m}, \dots, \underbrace{\mathbf{K}_n, \dots, \mathbf{K}_n}_{m}, \mathbf{I}),$$

*for all $m = 1, 2, \dots$ . Then $\{\mathbf{v}^{[m]}\}_{m=1}^\infty$ converges and*

$$\lim_{m\to\infty} \mathbf{v}^{[m]} = \mathbf{ME}_L(\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \dots, \mathbf{K}_n)).$$

We prove this theorem here only for the restriction to $BCL$. We follow the same lines as in the proof of theorem 5.1.2. For the proof in full generality see [53].

*Proof for the restriction to $BCL$.* First of all recall that $\mathrm{M}_L^{\mathrm{KL}}(\mathbf{K}_1, \dots, \mathbf{K}_n)$ denotes the minimal value of

$$\sum_{i=1}^n \mathrm{KL}(\mathbf{v}\|\mathbf{w}^{(i)})$$

subject to $\mathbf{w}^{(i)} \in V^L_{\mathbf{K}_i}$, $1 \leq i \leq n$, and $\mathbf{v} \in \mathbb{D}^L$. Furthermore, we denote by $E_m$ the minimal value of

$$m \sum_{i=1}^{n} \mathrm{KL}(\mathbf{v} \| \mathbf{w}^{(i)}) - m\, \mathrm{M}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n) + \mathrm{KL}(\mathbf{v} \| \mathbf{t}) \tag{5.6}$$

subject to $\mathbf{w}^{(i)} \in V^L_{\mathbf{K}_i}$, $1 \leq i \leq n$, and $\mathbf{v} \in \mathbb{D}^L$. By the definition of $\mathrm{M}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ we have that $0 \leq E_m$ for all $m = 1, 2, \dots$ .

Note, that for a fixed $m$ if $\mathbf{v} \in \mathbb{D}^L$ globally minimises (5.6) subject to $\mathbf{w}^{(i)} \in V^L_{\mathbf{K}_i}$, $1 \leq i \leq n$, then $\mathbf{v} \in \Delta^{\mathrm{KL}}_L(\underbrace{\mathbf{K}_1, \dots, \mathbf{K}_1}_{m}, \dots, \underbrace{\mathbf{K}_n, \dots, \mathbf{K}_n}_{m}, \mathbf{I})$ (by theorem 5.1.3 such a $\mathbf{v}$ is unique) and conversely if $\mathbf{v} \in \Delta^{\mathrm{KL}}_L(\underbrace{\mathbf{K}_1, \dots, \mathbf{K}_1}_{m}, \dots, \underbrace{\mathbf{K}_n, \dots, \mathbf{K}_n}_{m}, \mathbf{I})$ then $\mathbf{v}$ minimises (5.6) subject to above constraints.

Now let $\mathbf{r} = \mathbf{ME}_L(\Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n))$, that is

$$\mathbf{r} = \arg \max_{\mathbf{v} \in \Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)} - \sum_{j=1}^{J} v_j \log v_j = \arg \min_{\mathbf{v} \in \Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)} \mathrm{KL}(\mathbf{v} \| \mathbf{t}).$$

Since $\mathbf{r} \in \Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ it follows that for all $m \in \mathbb{N}$

$$E_m \leq \mathrm{KL}(\mathbf{r} \| \mathbf{t}). \tag{5.7}$$

Since $\mathbb{D}^L \subseteq \mathbb{R}^J$ is a compact space there exist a convergent subsequence of the sequence $\{\mathbf{v}^{[m]}\}_{m=1}^{\infty}$. Let $\{\mathbf{v}^{[m_k]}\}_{k=1}^{\infty}$ be any such convergent subsequence. Let

$$(\underbrace{\mathbf{w}^{(1)[m_k]}, \dots, \mathbf{w}^{(1)[m_k]}}_{m_k} \dots \underbrace{\mathbf{w}^{(n)[m_k]}, \dots, \mathbf{w}^{(n)[m_k]}}_{m_k}, \mathbf{t}) \in \Gamma^{\mathrm{KL}}_L(\underbrace{\mathbf{K}_1, \dots, \mathbf{K}_1}_{m_k} \dots \underbrace{\mathbf{K}_n, \dots, \mathbf{K}_n}_{m_k}, \mathbf{I})$$

be associated with $\mathbf{v}^{[m_k]}$ for all $k \in \mathbb{N}$. Recall that by lemma 3.2.8 these are unique for every $\mathbf{v}^{[m_k]}$ because $\mathbf{w}^{(i)[m_k]}$ corresponds to the KL-projection of $\mathbf{v}^{[m_k]}$ to $V^L_{\mathbf{K}_i}$, $1 \leq i \leq n$.

Now the sequence

$$\{\sum_{i=1}^{n} \mathrm{KL}(\mathbf{v}^{[m_k]} \| \mathbf{w}^{(i)[m_k]})\}_{k=1}^{\infty} \tag{5.8}$$

is bounded in $\mathbb{R}$. That is because we are operating in $BCL$ and therefore the coordinates of all $\mathbf{w}^{(i)[m_k]}$, $1 \leq i \leq n$ and $k = 1, 2, \dots$ , are bounded away from zero by some constants. And, since $\mathbf{v}^{[m_k]} = \mathbf{LogOp}(\mathbf{w}^{(1)[m_k]}, \dots, \mathbf{w}^{(n)[m_k]})$, we can find some global bound away from zero for the coordinates of all $\mathbf{v}^{[m_k]}$, $k = 1, 2, \dots$ . Consequently the sequence (5.8) also has a convergent subsequence. If the limit of any such convergent subsequence was not equal to $\mathrm{M}^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)$ then there would be eventually an $m$ which would contradict (5.7).

A similar argument to the one used in the proof of chairman theorem for $\hat{\Delta}^{\mathrm{KL}}$ gives that $\{\sum_{i=1}^{n} \mathrm{KL}(\mathbf{v}^{[m_k]} \| \mathbf{w}^{(i)[m_k]})\}_{k=1}^{\infty}$ converges and that

$$\lim_{k\to\infty} \sum_{i=1}^{n} \mathrm{KL}(\mathbf{v}^{[m_k]} \| \mathbf{w}^{(i)[m_k]}) = \mathrm{M}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n). \tag{5.9}$$

Note that we already know that $\lim_{k\to\infty} \mathbf{v}^{[m_k]}$ exists and we can denote it by $\mathbf{v}$. However we do not know whether the same is true for $\lim_{k\to\infty} \mathbf{w}^{(i)[m_k]}$, $1 \leq i \leq n$. On the other hand since $V_{\mathbf{K}_1}^L, \ldots, V_{\mathbf{K}_n}^L$ are compact there is a subsequence $\{m_l\}_{l=1}^{\infty}$ of the sequence of indices $\{m_k\}_{k=1}^{\infty}$ such that for all $1 \leq i \leq n$ the subsequence $\{\mathbf{w}^{(i)[m_l]}\}_{l=1}^{\infty}$ of the sequence $\{\mathbf{w}^{(i)[m_k]}\}_{k=1}^{\infty}$ is convergent. Let us denote the corresponding limits $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$. Note that also $\lim_{l\to\infty} \mathbf{v}^{[m_l]} = \mathbf{v}$. Since $\mathrm{KL}(\cdot\|\cdot)$ is a continuous function in both variables (we are confined to $BCL$) the value of $\sum_{i=1}^{n} \mathrm{KL}(\mathbf{v}\|\mathbf{w}^{(i)})$ is equal to the limit $\lim_{l\to\infty} \sum_{i=1}^{n} \mathrm{KL}(\mathbf{v}^{[m_l]} \| \mathbf{w}^{(i)[m_l]})$ which, by (5.9), is $\mathrm{M}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$.

It follows that $\lim_{k\to\infty} \mathbf{v}^{[m_k]} \in \Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and therefore $\lim_{k\to\infty} E_{m_k} = \mathrm{KL}(\lim_{k\to\infty} \mathbf{v}^{[m_k]} \| \mathbf{t})$ which, by the definition of $\mathbf{ME}$, is greater or equal to $\mathrm{KL}(\mathbf{r}\|\mathbf{t})$. However, by (5.7) this is possible only if $\lim_{k\to\infty} \mathbf{v}^{[m_k]} = \mathbf{r}$. As we have seen in the proof of chairman theorem for $\hat{\Delta}^{\mathrm{KL}}$ this implies convergence of whole sequence $\{\mathbf{v}^{[m]}\}_{m=1}^{\infty}$ and that $\lim_{m\to\infty} \mathbf{v}^{[m]} = \mathbf{r}$. $\square$

Both chairman theorems suggest that if we want to create from $\hat{\Delta}^{\mathrm{KL}}$ and $\Delta^{\mathrm{KL}}$ merging operators which are stable with respect to an infinitesimal bias by the most uninformative point — a fairly natural condition — we should consider the following two p-merging operators:

1. The *Linear Entropy Process* (**LEP**) to be for any $L$ and any $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ defined by

$$\mathbf{LEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \mathbf{CM}_L^{\infty}(\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)) \text{ and}$$

2. the *Social Entropy Process* (**SEP**) to be for any $L$ and any $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ defined by

$$\mathbf{SEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \mathbf{ME}_L(\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)).$$

However the question now arises as to how these operators behave with respect to the principles which we have investigated in the previous chapter. In the next section we take a look on this problem.

Note that by requiring the stability induced by the chairman theorems we have obtained the pleasing property that both **LEP** and **SEP** always return a single probability function and therefore they belong to the special type of p-merging operator — social inference processes. Social inference processes are particularly interesting for the implementation of practical expert system, where in general we may expect a single probability function to be produced.

Also note that **SEP** was first defined by Wilmers in [52] and it was further investigated in [53] and [1].

## 5.2 Principles for LEP and SEP

In this section we investigate whether **LEP** and **SEP** satisfy all the previously considered principles for probabilistic merging. We take an advantage of our results for $\hat{\Delta}^{\mathrm{KL}}$ and $\Delta^{\mathrm{KL}}$ merging operators.

Since the $\hat{\Delta}^{\mathrm{KL}}$ and $\Delta^{\mathrm{KL}}$ merging operators satisfy the principle of equivalence **(K2)** and both **ME** and $\mathbf{CM}^{\infty}$ satisfy the equivalence principle, see section 1.3, we have that **LEP** and **SEP** also satisfy **(K2)**.

Similarly since the $\hat{\Delta}^{\mathrm{KL}}$ and $\Delta^{\mathrm{KL}}$ merging operators satisfy the principle of atomic renaming **(K3)** and both **ME** and $\mathbf{CM}^{\infty}$ satisfy the atomic renaming principle, see section 1.3, we have that **LEP** and **SEP** also satisfy **(K3)**.

Assume that $\mathbf{K}_1, \dots, \mathbf{K}_n \in CL$. $\hat{\Delta}^{\mathrm{KL}}$ and $\Delta^{\mathrm{KL}}$ satisfy the principle of consistency **(K4)**. Hence, if $\bigcap_{i=1}^n V_{\mathbf{K}_i}^L \neq \emptyset$ then $\mathbf{LEP}_L(\mathbf{K}_1, \dots, \mathbf{K}_n) \in \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \dots, \mathbf{K}_n) \subseteq \bigcap_{i=1}^n V_{\mathbf{K}_i}^L$ and $\mathbf{SEP}_L(\mathbf{K}_1, \dots, \mathbf{K}_n) \in \Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \dots, \mathbf{K}_n) \subseteq \bigcap_{i=1}^n V_{\mathbf{K}_i}^L$. Therefore **LEP** and **SEP** satisfy **(K4)**. On the other hand note that the principle of strong consistency **(K4*)** does not really have a sense for social inference processes.

Suppose that $\mathbf{K}_1, \dots, \mathbf{K}_n$, $\mathbf{F}_1, \dots, \mathbf{F}_m \in CL$, $\bigcap_{i=1}^m V_{\mathbf{F}_i}^L \neq \emptyset$ and $\mathbf{LEP}_L(\mathbf{K}_1, \dots, \mathbf{K}_n) \in \bigcap_{i=1}^m V_{\mathbf{F}_i}^L$. Then, since the $\hat{\Delta}^{\mathrm{KL}}$-merging operator satisfies the principle of agreement, we have that

$$\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \dots, \mathbf{K}_n, \mathbf{F}_1, \dots, \mathbf{F}_m) = \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \dots, \mathbf{K}_n) \cap \bigcap_{i=1}^m V_{\mathbf{F}_i}^L.$$

Consequently

$$\mathbf{LEP}_L(\mathbf{K}_1, \dots, \mathbf{K}_n) \in \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \dots, \mathbf{K}_n, \mathbf{F}_1, \dots, \mathbf{F}_m) \subseteq \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \dots, \mathbf{K}_n)$$

and therefore $\mathbf{LEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \mathbf{LEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m)$. We conclude that **LEP** satisfies the principle of collegiality **(K5)**. A proof that **SEP** satisfies **(K5)** can be constructed analogously.

We note that all these results in the case of **SEP** have been observed by Wilmers in [52].

It is a simple consequence of the agreement principle **(K7)**, which is satisfied by $\hat{\Delta}^{\mathrm{KL}}$ and $\Delta^{\mathrm{KL}}$, that if $\mathbf{SEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \mathbf{SEP}_L(\mathbf{F}_1, \ldots, \mathbf{F}_m)$ and $\mathbf{LEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \mathbf{LEP}_L(\mathbf{F}_1, \ldots, \mathbf{F}_m)$ for some $\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots \mathbf{F}_m \in CL$ then

$$\Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \Delta^{\mathrm{KL}}_L(\mathbf{F}_1, \ldots, \mathbf{F}_m) = \Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m)$$

and

$$\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{F}_1, \ldots, \mathbf{F}_m) = \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m).$$

Therefore

$$\mathbf{SEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \in \Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m) \subseteq \Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$$

and

$$\mathbf{LEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \in \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m) \subseteq \hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

So it follows that $\mathbf{SEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \mathbf{SEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m)$, $\mathbf{LEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \mathbf{LEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m)$ and consequently both **SEP** and **LEP** satisfy **(K7)**.

Unfortunately, neither **SEP** nor **LEP** satisfy the disagreement principle **(K6)**. This is because the principles of consistency and disagreement lead to a contradiction for the social inference processes in general. Let $\mathcal{S}$ be a social inference process. Consider $\mathcal{S}_L(\mathbf{K}_1) \neq \mathcal{S}_L(\mathbf{K}_2) \in V^L_{\mathbf{K}_1}$ and $\{\mathcal{S}_L(\mathbf{K}_2)\} = V^L_{\mathbf{K}_2}$. This can be always achieved given that $V^L_{\mathbf{K}_1}$ contains at least two distinct points. Then the principle of consistency gives $\mathcal{S}_L(\mathbf{K}_1, \mathbf{K}_2) = \mathcal{S}_L(\mathbf{K}_2)$ what contradicts the principle of disagreement.

Another principle with a flavour of the disagreement principle has been introduced by Savage in [46] as a converse to the collegiality principle. The following is a reformulation of it in our framework.

**Savage Principle.** Let $\Delta$ be a p-merging operator. Let $\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m \in CL$ and assume that $\bigcap_{i=1}^m V_{\mathbf{F}_i}^L \neq \emptyset$. If $\Delta(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \bigcap_{i=1}^m V_{\mathbf{F}_i}^L = \emptyset$ then

$$\Delta(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \Delta(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m) = \emptyset.$$

We can make the following trivial observation.

**Lemma 5.2.1.** *Let $\Delta$ be a p-merging operator which satisfies the consistency principle* **(K4)** *and the disagreement principle* **(K6)**. *Then $\Delta$ satisfies the Savage principle.*

A consequence is that both $\Delta^{\mathrm{KL}}$ and $\hat{\Delta}^{\mathrm{KL}}$ merging operators satisfy the Savage principle once we have confined ourselves to $BCL$.

**Theorem 5.2.2. LEP** *and* **SEP** *satisfy the Savage principle when restricted to $BCL$.*

*Proof.* Let $\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m \in BCL$ and $\bigcap_{i=1}^m V_{\mathbf{F}_i}^L \neq \emptyset$. Assume that

$$\mathbf{LEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \notin \bigcap_{i=1}^m V_{\mathbf{F}_i}^L. \tag{5.10}$$

If $\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \bigcap_{i=1}^m V_{\mathbf{F}_i}^L = \emptyset$ then by the disagreement principle for $\hat{\Delta}^{\mathrm{KL}}$-operator

$$\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m) = \emptyset.$$

If $\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \bigcap_{i=1}^m V_{\mathbf{F}_i}^L \neq \emptyset$ then by the agreement principle for $\hat{\Delta}^{\mathrm{KL}}$-operator

$$\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \bigcap_{i=1}^m V_{\mathbf{F}_i}^L = \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m)$$

which by assumption (5.10) does not contain $\mathbf{LEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$.

In any case $\mathbf{LEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \notin \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m)$ which concludes the proof for **LEP**. A proof for **SEP** can be constructed similarly. □

Note that the result above for **SEP** above was proved differently by Savage in [46].

One would perhaps guess that the Savage principle is only a weaker version of the disagreement principle. However, while **OLEP** satisfies the disagreement principle, the following example shows that **OLEP** does not satisfy the Savage principle.

**Example 5.2.3.** Let $V_{\mathbf{K}}^L = \{(\frac{2}{8}, \frac{6}{8})\}$, $V_{\mathbf{F}_1}^L = \{(\frac{1}{8}, \frac{7}{8})\}$ and $V_{\mathbf{F}_2}^L = \{(x, 1-x), x \in [\frac{1}{8}, \frac{3}{8}]\}$. $\mathbf{OLEP}_L(\mathbf{K}) = (\frac{2}{8}, \frac{6}{8}) \notin V_{\mathbf{F}_1}^L \cap V_{\mathbf{F}_2}^L = (\frac{1}{8}, \frac{7}{8})$, however

$$\mathbf{OLEP}_L(\mathbf{K}, \mathbf{F}_1, \mathbf{F}_2) = \mathbf{LinOp}(\mathbf{ME}_L(\mathbf{K}), \mathbf{ME}_L(\mathbf{F}_1), \mathbf{ME}_L(\mathbf{F}_2)) =$$

$$= \mathbf{LinOp}\left(\left(\frac{2}{8}, \frac{6}{8}\right), \left(\frac{1}{8}, \frac{7}{8}\right), \left(\frac{3}{8}, \frac{5}{8}\right)\right) = \left(\frac{2}{8}, \frac{6}{8}\right).$$ □

**Example 5.2.4.** By setting

$$V^L_{\mathbf{K}} = \left\{ \mathbf{LogOp}\left(\left(\frac{1}{8}, \frac{7}{8}\right), \left(\frac{3}{8}, \frac{5}{8}\right)\right)\right\}$$

in the example 5.2.3 we get a counterexample to the Savage principle for **OSEP**. □

Now we investigate whether **LEP** and **SEP** satisfy the crucial principle of language invariance.

**Theorem 5.2.5.** **SEP** *satisfies* **LI**.

*Proof.* Let $L = L_1 \cup L_2$ where $L_1 \cap L_2 = \emptyset$. Let $\mathbf{K}_1, \dots, \mathbf{K}_n \in CL_1$. We prove that

$$\mathbf{ME}_L(\Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n)|_{L_1} = \mathbf{ME}_{L_1}(\Delta^{\mathrm{KL}}_{L_1}(\mathbf{K}_1, \dots, \mathbf{K}_n))$$

which is sufficient to establish the claim.

Let $\mathbf{v} = \mathbf{ME}_L(\Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \dots, \mathbf{K}_n))$ and $\mathbf{r} = \mathbf{ME}_{L_1}(\Delta^{\mathrm{KL}}_{L_1}(\mathbf{K}_1, \dots, \mathbf{K}_n))$. Since the $\Delta^{\mathrm{KL}}$-merging operator is language invariant we have that $\mathbf{v}|_{L_1} \in \Delta^{\mathrm{KL}}_{L_1}(\mathbf{K}_1, \dots, \mathbf{K}_n)$ and hence

$$H(\mathbf{v}|_{L_1}) \leq H(\mathbf{r}). \tag{5.11}$$

On the other hand

$$H(\mathbf{v}|_{L_2}) \leq H(\mathbf{t}), \tag{5.12}$$

where $\mathbf{t} \in \mathbb{D}^{L_2}$ is the uniform probability function.

Let $|\operatorname{At}(L_1)| = J$ and $|\operatorname{At}(L_2)| = I$. Recall that we write $v_{ji}$, $1 \leq j \leq J$ and $1 \leq i \leq I$, to denote values of $\mathbf{v} \in \mathbb{D}^L$ on all $L$-atoms, and we write $v_{j\cdot} = \sum_{i=1}^I v_{ji}$ and $v_{\cdot i} = \sum_{j=1}^J v_{ji}$ to simplify the notation.

By the Jensen inequality applied to the concave function $-x \log x$

$$H(\mathbf{v}) - H(\mathbf{v}|_{L_1}) - H(\mathbf{v}|_{L_2}) =$$

$$= -\sum_{j=1,i=1}^{J,I} v_{ji} \log v_{ji} + \sum_{j=1}^{J} v_{j\cdot} \log v_{j\cdot} + \sum_{i=1}^{I} v_{\cdot i} \log v_{\cdot i} \leq \log(JI) - \log(J) - \log(I) = 0.$$

This together with (5.11) and (5.12) gives

$$H(\mathbf{v}) \leq H(\mathbf{v}|_{L_1}) + H(\mathbf{v}|_{L_2}) \leq H(\mathbf{r}) + H(\mathbf{t}) = H(\mathbf{r} \cdot \mathbf{t}).$$

Note that the last equality can be observed by a straightforward algebraic manipulation.

By lemma 4.2.6 we have that $\mathbf{r} \cdot \mathbf{t} \in \Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$, which is possible only if $\mathbf{v} = \mathbf{r} \cdot \mathbf{t}$ since $\mathbf{v}$ is the unique most entropic point in $\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. Therefore $\mathbf{v}|_{L_1} = \mathbf{r}$ and the theorem follows. $\square$

The proof above is essentially a combination of the fact that the $\Delta^{\mathrm{KL}}$-operator satisfies **LI** and the proof that the **ME**-inference process is language invariant. Accordingly, we combine the fact that the $\hat{\Delta}^{\mathrm{KL}}$-merging operator satisfies **LI** and the proof that the $\mathbf{CM}^{\infty}$-inference process is language invariant into the following proof.

**Theorem 5.2.6.** **LEP** *satisfies* **LI**.

*Proof.* Let $L = L_1 \cup L_2$ where $L_1 \cap L_2 = \emptyset$. Let $|\operatorname{At}(L_1)| = J$ and $|\operatorname{At}(L_2)| = I$. Let $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL_1$. We prove that

$$\mathbf{CM}_L^{\infty}(\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)|_{L_1} = \mathbf{CM}_{L_1}^{\infty}(\hat{\Delta}_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)).$$

Notice that whenever $\alpha \in \operatorname{At}(L_1)$ is such that for all $\mathbf{w} \in \hat{\Delta}_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ we have $\mathbf{w}(\alpha) = 0$ then for all $\mathbf{v} \in \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ we have $\mathbf{v}(\alpha \wedge \beta) = 0$ for $\beta \in \operatorname{At}(L_2)$. Therefore, in order to simplify the notation in the formulation of the $\mathbf{CM}^{\infty}$-inference process, we may assume that there is no $\alpha \in \operatorname{At}(L_1)$ such that for all $\mathbf{w} \in \hat{\Delta}_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ we have $\mathbf{w}(\alpha) = 0$.

Let $\mathbf{v} = \mathbf{CM}_L^{\infty}(\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n))$ and $\mathbf{r} = \mathbf{CM}_{L_1}^{\infty}(\hat{\Delta}_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n))$. Since the $\hat{\Delta}^{\mathrm{KL}}$-merging operator is language invariant we have that $\mathbf{v}|_{L_1} \in \hat{\Delta}_{L_1}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and hence

$$\sum_{j=1}^{J} \log(v_{j\cdot}) \leq \sum_{j=1}^{J} \log(r_j). \tag{5.13}$$

Since log is concave by the Jensen inequality

$$\sum_{j=1}^{J} \log(v_{j\cdot}) = J \log I + \sum_{j=1}^{J} \log\left(\frac{v_{j\cdot}}{I}\right) \geq J \log I + \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} \log(v_{ji}). \tag{5.14}$$

Now let $\mathbf{t}$ be the uniform $L_2$-probability function $(\frac{1}{I}, \ldots, \frac{1}{I})$. Then, by lemma 4.2.8, $\mathbf{r} \cdot \mathbf{t} \in \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and therefore

$$J \log I + \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} \log(v_{ji}) \geq J \log I + \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} \log\left(\frac{r_j}{I}\right) = \sum_{j=1}^{J} \log(r_j).$$

This together with (5.14) gives

$$\sum_{j=1}^{J} \log(v_{j\cdot}) \geq \sum_{j=1}^{J} \log(r_j)$$

which together with (5.13) and uniqueness of the maximum point implies $\mathbf{r} = \mathbf{v}|_{L_1}$. $\square$

With respect to the stronger principles concerning irrelevant information there are couple of immediate negative results: **SEP** does not satisfy the irrelevant information principle due to example 4.2.1 and **LEP** cannot satisfy even the consistent irrelevant information principle since it is known that $\mathbf{CM}^\infty$ does not satisfy the irrelevant information principle for inference processes, see [24] or [39], (consider only one expert). However, the remaining question whether **SEP** satisfies the consistent irrelevant information principle is still open.

Finally, the fact, that **SEP** satisfies the principle of relativisation **REL** given condition (1.4), is a consequence of a more general result proved by Wilmers in [53] (theorem 3.9). We prove here only the following result:

**Theorem 5.2.7.** **LEP** *satisfies* **REL**.

*Proof.* We proceed similarly to the proof of theorem 4.5.1:
Let $\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_n \in CL$ be such that

$$\mathbf{K}_k = \{Bel(\theta) = c^{(k)}\} \cup \{\sum_s a_{si}^{(k)} Bel(\varphi_s|\theta) = d_i^{(k)} | i = 1, \ldots, h^{(k)}\},$$

$$\mathbf{F}_k = \mathbf{K}_k \cup \{\sum_s b_{si}^{(k)} Bel(\psi_s|\neg\theta) = e_i^{(k)} | i = 1, \ldots, l^{(k)}\},$$

where $0 < c^{(k)} < 1$, $\theta$ is fixed and $\varphi_s, \psi_s \in SL$.

Suppose that $\mathbf{w} = \mathbf{LEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. Therefore $\mathbf{w} \in \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ and let $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in \hat{\Gamma}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ be such that $\mathbf{w} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$. By lemma 3.1.1 $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ minimise $\sum_{k=1}^n \mathrm{KL}(\mathbf{w}^{(k)} \| \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}))$ subject only to $\mathbf{w}^{(1)} \in V_{\mathbf{K}_1}^L, \ldots,\ \mathbf{w}^{(n)} \in V_{\mathbf{K}_n}^L$.

Now, note that due to the constraints $\{Bel(\theta) = c^{(k)}\} \in \mathbf{K}_k$, $1 \leq k \leq n$, we have that $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ also minimise

$$\sum_{k=1}^n \sum_{j \in \mathrm{At}(\theta) \cap \bigcup_{k=1}^n \mathrm{Sig}(\mathbf{w}^{(k)})} w_j^{(k)} \log \frac{n \cdot w_j^{(k)}}{\sum_{k=1}^n w_j^{(k)}}$$

subject to $\mathbf{w}^{(1)} \in V_{\mathbf{K}_1}^L, \ldots,\ \mathbf{w}^{(n)} \in V_{\mathbf{K}_n}^L$.

While the above follows by $\mathbf{w} \in \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$, the fact that actually $\mathbf{w} = \mathbf{CM}_L^\infty(\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n))$ gives that $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ also maximise

$$\sum_{j \in \mathrm{Sig}_L(\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n))} \log \frac{\sum_{k=1}^n w_j^{(k)}}{n},$$

where $\mathrm{Sig}_L(\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1,\dots,\mathbf{K}_n)) = \{j\colon \exists \mathbf{w} \in \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1,\dots,\mathbf{K}_n),\, w_j \neq 0\}$, subject to $\mathbf{w}^{(1)} \in V_{\mathbf{K}_1}^L,\dots,\mathbf{w}^{(n)} \in V_{\mathbf{K}_n}^L$ and to the equation $\hat{\mathrm{M}}_L^{\mathrm{KL}}(\mathbf{K}_1,\dots,\mathbf{K}_n) = \sum_{k=1}^n \sum_{j\in\bigcup_{k=1}^n \mathrm{Sig}(\mathbf{w}^{(k)})} w_j^{(k)} \log \frac{n\cdot w_j^{(k)}}{\sum_{k=1}^n w_j^{(k)}}$.

Therefore, again due to the constraints $\{Bel(\theta) = c^{(k)}\} \in \mathbf{K}_k$, $1 \leq k \leq n$, we have that $\mathbf{w}^{(1)},\dots,\mathbf{w}^{(n)}$ maximise

$$\sum_{j\in\mathrm{At}(\theta)\cap\mathrm{Sig}_L(\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1,\dots,\mathbf{K}_n))} \log \frac{\sum_{k=1}^n w_j^{(k)}}{n}, \tag{5.15}$$

subject to the above conditions.

Let $\mathbf{u} = \mathbf{CM}_L^\infty(\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{F}_1,\dots,\mathbf{F}_n))$ and $(\mathbf{u}^{(1)},\dots,\mathbf{u}^{(n)}) \in \hat{\Gamma}_L^{\mathrm{KL}}(\mathbf{F}_1,\dots,\mathbf{F}_n)$ be such that $\mathbf{u} = \mathbf{LinOp}(\mathbf{u}^{(1)},\dots,\mathbf{u}^{(n)})$. It follows that $\mathbf{u}^{(1)},\dots,\mathbf{u}^{(n)}$ minimise $\sum_{k=1}^n \mathrm{KL}(\mathbf{u}^{(k)} \| \mathbf{LinOp}(\mathbf{u}^{(1)},\dots,\mathbf{u}^{(n)}))$ subject to $\mathbf{u}^{(1)} \in V_{\mathbf{F}_1}^L,\dots,\mathbf{u}^{(n)} \in V_{\mathbf{F}_n}^L$. Due to the constraints $\{Bel(\theta) = c^{(k)}\} \in \mathbf{F}_k$, $1 \leq k \leq n$, we have that $\mathbf{u}^{(1)},\dots,\mathbf{u}^{(n)}$ also minimise $\sum_{k=1}^n \sum_{j\in\mathrm{At}(\theta)\cap\bigcup_{k=1}^n \mathrm{Sig}(\mathbf{u}^{(k)})} u_j^{(k)} \log \frac{n\cdot u_j^{(k)}}{\sum_{k=1}^n u_j^{(k)}}$ subject to $\mathbf{u}^{(1)} \in V_{\mathbf{F}_1}^L,\dots,\mathbf{u}^{(n)} \in V_{\mathbf{F}_n}^L$ and they also maximise

$$\sum_{j\in\mathrm{At}(\theta)\cap\mathrm{Sig}_L(\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{F}_1,\dots,\mathbf{F}_n))} \log \frac{\sum_{k=1}^n u_j^{(k)}}{n}, \tag{5.16}$$

subject to $\mathbf{u}^{(1)} \in V_{\mathbf{F}_1}^L,\dots,\mathbf{u}^{(n)} \in V_{\mathbf{F}_n}^L$ and to the equation $\hat{\mathrm{M}}_L^{\mathrm{KL}}(\mathbf{F}_1,\dots,\mathbf{F}_n) = \sum_{k=1}^n \sum_{j\in\bigcup_{k=1}^n \mathrm{Sig}(\mathbf{u}^{(k)})} u_j^{(k)} \log \frac{n\cdot u_j^{(k)}}{\sum_{k=1}^n u_j^{(k)}}$.

Since the maximisation problem for $\sum_{j\in\mathrm{At}(\theta)} \log x_j$ over a closed convex area has a unique solution and KL-projections are unique, both (5.15) and (5.16) must give that $w_j^{(k)} = u_j^{(k)}$ for all $j \in \mathrm{At}(\theta)$ and $1 \leq k \leq n$.

The remaining proof of the second equality in $\mathbf{LEP}_L(\mathbf{K}_1,\dots,\mathbf{K}_n)|_\theta = \mathbf{LinOp}(\mathbf{w}^{(1)},\dots,\mathbf{w}^{(n)})|_\theta = \mathbf{LinOp}(\mathbf{u}^{(1)},\dots,\mathbf{u}^{(n)})|_\theta = \mathbf{LEP}_L(\mathbf{F}_1,\dots,\mathbf{F}_n)|_\theta$ is the same as the one for theorem 4.5.1. □

## 5.3 Computability

In the previous section we have argued that **LEP** and **SEP** are attractive merging operators. However computing the result of applying these procedures proves to be tricky. First of all, there are some serious computational issues with the whole setting which we have introduced in this thesis. In particular it is not even possible to find a random Turing machine running in polynomial time which on input given by a set of constraints $\mathbf{K} \in CL$ verifies the consistency of $\mathbf{K}$ (given that the problems solvable in a randomized polynomial time cannot be solved in a polynomial time), see theorem 10.7 of [39].

However, some closely related computational problems have been extensively studied in the literature. As we have noted in section 3.2 this includes procedures for finding a KL-projection to a closed convex set of probability functions. These show that in many particular practical implementations the problem of intractability does not arise, e.g. as in the case when knowledge bases are generated by marginal probability functions and where the IPFP-procedure can be applied to effectively find a KL-projection, see section 3.2. Therefore, instead of trying to find algorithms to compute **LEP** and **SEP** from scratch, we will assume in this section that some effective procedures for KL-projections and conjugated KL-projections are given.

Let $\mathbf{K}_1, \ldots, \mathbf{K}_n \in BCL$. Recall the averaging projective procedure $\hat{F}^{D_f}_{[\mathbf{K}_1,\ldots,\mathbf{K}_n]}$ defined in section 3.2. Here we consider only the Kullback-Leibler divergence $D_f = \mathrm{KL}$ in which case

$$\hat{F}^{\mathrm{KL}}_{[\mathbf{K}_1,\ldots,\mathbf{K}_n]}(\mathbf{v}) = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$$

and $\mathbf{w}^{(i)} = \arg\min_{\mathbf{w} \in V^L_{\mathbf{K}_i}} \mathrm{KL}(\mathbf{w} \| \mathbf{v})$ for all $1 \leq i \leq n$. Note that the domain of $\hat{F}^{\mathrm{KL}}_{[\mathbf{K}_1,\ldots,\mathbf{K}_n]}$ is restricted to $\{\mathbf{v} \in \mathbb{D}^L : v_j > 0,\ 1 \leq j \leq J\}$, $J = |\operatorname{At}(L)|$, due to our initial restriction to $BCL$.

By theorem 3.2.12 we know that the sequence

$$\{\mathbf{v}^{[i]}\}_{i=0}^{\infty}, \tag{5.17}$$

where $\mathbf{v}^{[0]} = (\frac{1}{J}, \ldots, \frac{1}{J})$, $J = |\operatorname{At}(L)|$, is the uniform $L$-probability function and $\mathbf{v}^{[i+1]} = \hat{F}^{\mathrm{KL}}_{[\mathbf{K}_1,\ldots,\mathbf{K}_n]}(\mathbf{v}^{[i]})$, converges to some probability function in $\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. This procedure can be immediately used to compute $\mathbf{LEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ in the case when $\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is a singleton. By theorem 5.1.1 this happens, for instance, when at least one of $V^L_{\mathbf{K}_1}, \ldots, V^L_{\mathbf{K}_n}$ is a singleton.

Let $\mathbf{I} \in CL$ be such that $V^L_{\mathbf{I}} = \{\mathbf{t}\}$, where $\mathbf{t} = (\frac{1}{J}, \ldots, \frac{1}{J})$. For every $m = 1, 2, \ldots$

we define the sequence $\{\mathbf{v}_{[m]}^{[i]}\}_{i=0}^{\infty}$ by $\mathbf{v}_{[m]}^{[0]} = (\frac{1}{J}, \dots, \frac{1}{J})$ and

$$\mathbf{v}_{[m]}^{[i+1]} = \hat{F}^{\mathrm{KL}}_{[\mathbf{I},\underbrace{\mathbf{K}_1 \dots \mathbf{K}_1}_{m},\dots,\underbrace{\mathbf{K}_n \dots \mathbf{K}_n}_{m}]}(\mathbf{v}_{[m]}^{[i]}).$$

By theorem 3.2.12

$$\lim_{i\to\infty} \mathbf{v}_{[m]}^{[i]} = \mathbf{LEP}_L(\mathbf{I}, \underbrace{\mathbf{K}_1 \dots \mathbf{K}_1}_{m}, \dots, \underbrace{\mathbf{K}_n \dots \mathbf{K}_n}_{m}).$$

By the chairman theorem for $\hat{\Delta}^{\mathrm{KL}}$

$$\lim_{m\to\infty} \lim_{i\to\infty} \mathbf{v}_{[m]}^{[i]} = \mathbf{LEP}_L(\mathbf{K}_1, \dots, \mathbf{K}_n). \tag{5.18}$$

It is interesting that

$$\lim_{m\to\infty} \mathbf{v}_{[m]}^{[i]} = \mathbf{v}^{[i]},$$

where $\{\mathbf{v}^{[i]}\}_{i=0}^{\infty}$ was defined in (5.17), as we show in lemma 5.3.2.

**Lemma 5.3.1.** *Let* $\mathbf{K}_1, \dots, \mathbf{K}_n \in BCL$. *Then* $\hat{F}^{\mathrm{KL}}_{[\mathbf{K}_1,\dots,\mathbf{K}_n]}$ *is a continuous function over the domain* $\{\mathbf{v} :\ v_j > 0, 1 \le j \le J\}$.

*Proof.* Assume that a sequence $\{\mathbf{v}_i\}_{i=0}^{\infty}$, $\mathbf{v}_i \in \mathbb{D}^L$, $\mathbf{v}_i(\alpha_j) > 0$, $1 \le j \le J$, converges to $\mathbf{v} \in \mathbb{D}^L$, $v_j > 0$, $1 \le j \le J$. Consider the sequence of the corresponding KL-projections $\mathbf{w}_i^{(k)} \in V^L_{\mathbf{K}_k}$, $1 \le k \le n$. Now due to compactness of $V^L_{\mathbf{K}_1}, \dots, V^L_{\mathbf{K}_n}$ the sequence $\{(\mathbf{v}_i, \mathbf{w}_i^{(1)}, \dots, \mathbf{w}_i^{(n)})\}_{i=0}^{\infty}$ has a convergent subsequence which, by lemma 3.2.11, converges to $(\mathbf{v}, \mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$, where $\mathbf{w}^{(k)}$ is the KL-projection of $\mathbf{v}$ into $V^L_{\mathbf{K}_k}$ for all $1 \le k \le n$. But then the whole sequence $\{(\mathbf{v}_i, \mathbf{w}_i^{(1)}, \dots, \mathbf{w}_i^{(n)})\}_{i=0}^{\infty}$ converges to $(\mathbf{v}, \mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$ because otherwise we could find a convergent subsequence which would contradict lemma 3.2.11.

Since the operator **LinOp** is clearly continuous the function $\hat{F}^{\mathrm{KL}}_{[\mathbf{K}_1,\dots,\mathbf{K}_n]}(\mathbf{v}_i)$ converges to $\hat{F}^{\mathrm{KL}}_{[\mathbf{K}_1,\dots,\mathbf{K}_n]}(\mathbf{v})$ and therefore the function $\hat{F}^{\mathrm{KL}}_{[\mathbf{K}_1,\dots,\mathbf{K}_n]}$ is continuous provided that $\mathbf{K}_1, \dots, \mathbf{K}_n \in BCL$. □

**Lemma 5.3.2.** *Let* $\mathbf{v}_{[m]}^{[i]}$ *and* $\mathbf{v}^{[i]}$, $i = 0, 1, \dots$, $m = 1, 2, \dots$, *be as above. Then*

$$\lim_{m\to\infty} \mathbf{v}_{[m]}^{[i]} = \mathbf{v}^{[i]}.$$

*Proof.* By induction we prove that $\lim_{m\to\infty} \mathbf{v}^{[i]}_{[m]} = \mathbf{v}^{[i]}$ for all $i = 0, 1, \ldots$ . First, for $i = 0$ the sequence is constant. Now assume that $\lim_{m\to\infty} \mathbf{v}^{[i]}_{[m]} = \mathbf{v}^{[i]}$ for all $i < h$. Since by lemma 5.3.1 $\hat{F}^{\mathrm{KL}}_{[\mathbf{K}_1,\ldots,\mathbf{K}_n]}$ is a continuous function we have that

$$\lim_{m\to\infty} \hat{F}^{\mathrm{KL}}_{[\mathbf{K}_1,\ldots,\mathbf{K}_n]}(\mathbf{v}^{[h-1]}_{[m]}) = \hat{F}^{\mathrm{KL}}_{[\mathbf{K}_1,\ldots,\mathbf{K}_n]}(\mathbf{v}^{[h-1]}) = \mathbf{v}^{[h]}. \tag{5.19}$$

Now as $m$ increases the difference between $\hat{F}^{\mathrm{KL}}_{[\mathbf{K}_1,\ldots,\mathbf{K}_n]}$ and $\hat{F}^{\mathrm{KL}}_{[\mathbf{I},\underbrace{\mathbf{K}_1\ldots\mathbf{K}_1}_{m},\ldots,\underbrace{\mathbf{K}_n\ldots\mathbf{K}_n}_{m}]}$ vanishes so

$$\lim_{m\to\infty} \mathbf{v}^{[h]}_{[m]} = \lim_{m\to\infty} \hat{F}^{\mathrm{KL}}_{[\mathbf{I},\underbrace{\mathbf{K}_1\ldots\mathbf{K}_1}_{m},\ldots,\underbrace{\mathbf{K}_n\ldots\mathbf{K}_n}_{m}]}(\mathbf{v}^{[h-1]}_{[m]}) = \lim_{m\to\infty} \hat{F}^{\mathrm{KL}}_{[\mathbf{K}_1,\ldots,\mathbf{K}_n]}(\mathbf{v}^{[h-1]}_{[m]})$$

and by (5.19) that all important inductive step follows. □

Now notice that if the limits in (5.18) were interchangeable then this would offer a nice algorithm to compute **LEP** in full generality (but with no claims to any theoretical results on the complexity of computation) and it would answer the question to closely characterise the limit $\lim_{i\to\infty} \mathbf{v}^{[i]}$. Unfortunately, the following simple example shows that these limits are not interchangeable since for some $\mathbf{K}_1, \ldots, \mathbf{K}_n$

$$\lim_{i\to\infty} \mathbf{v}^{[i]} \neq \mathbf{LEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n).$$

**Example 5.3.3.** Let $L = \{p, q\}$, $\mathbf{K}_1 = \{Bel(p) = \frac{1}{4}\}$ and $\mathbf{K}_2 = \{Bel(q) = \frac{1}{4}\}$ and assume the additional constraint that every atom is bounded by 0.01. Note that $V^L_{\mathbf{K}_1} = \{(x, \frac{1}{4} - x, y, \frac{3}{4} - y),\ x \in [0.01, \frac{1}{4} - 0.01], y \in [0.01, \frac{3}{4} - 0.01]\}$ and $V^L_{\mathbf{K}_2} = \{(x, y, \frac{1}{4}-x, \frac{3}{4}-y),\ x \in [0.01, \frac{1}{4}-0.01], y \in [0.01, \frac{3}{4}-0.01]\}$, where probability functions are listed in the following order of atoms: $p \wedge q, p \wedge \neg q, \neg p \wedge q, \neg p \wedge \neg q$.

Then the members of the sequence $\{\mathbf{v}^{[i]}\}_{i=0}^{\infty}$ can be computed by two minimisation problems: Find $x \in [0.01, \frac{1}{4} - 0.01]$ and $y \in [0.01, \frac{3}{4} - 0.01]$ which minimise

$$x \log \frac{x}{v^{[i]}_1} + \left(\frac{1}{4} - x\right) \log \frac{\frac{1}{4} - x}{v^{[i]}_2} + y \log \frac{y}{v^{[i]}_3} + \left(\frac{3}{4} - y\right) \log \frac{\frac{3}{4} - y}{v^{[i]}_4}$$

and other couple $\bar{x} \in [0.01, \frac{1}{4} - 0.01]$ and $\bar{y} \in [0.01, \frac{3}{4} - 0.01]$ which minimise

$$\bar{x} \log \frac{\bar{x}}{v^{[i]}_1} + \bar{y} \log \frac{\bar{y}}{v^{[i]}_2} + (\frac{1}{4} - \bar{x}) \log \frac{\frac{1}{4} - \bar{x}}{v^{[i]}_3} + (\frac{3}{4} - \bar{y}) \log \frac{\frac{3}{4} - \bar{y}}{v^{[i]}_4}.$$

Then $v^{[i+1]}_1 = \frac{x+\bar{x}}{2}$, $v^{[i+1]}_2 = \frac{\frac{1}{4}-x+\bar{y}}{2}$, $v^{[i+1]}_3 = \frac{\frac{1}{4}-\bar{x}+y}{2}$ and $v^{[i+1]}_4 = \frac{\frac{3}{2}-\bar{y}-y}{2}$. After setting $\mathbf{v}^{[0]} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ it turns out that in each iteration $\bar{x} = x$ and $\bar{y} = y$.

After performing a numerical computation for the first one hundred iterations we obtain

$$\{\mathbf{v}^{[100]}\} \approx (0.0488395,\ 0.2011605,\ 0.2011605,\ 0.5488394).$$

The rate of convergence for the first atom $p \wedge q$ is depicted in figure 5.1 by the bottom red line.

However, since $\mathbf{K}_1$ and $\mathbf{K}_2$ are jointly consistent, we have that

$$\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1, \mathbf{K}_2) = V^L_{\mathbf{K}_1} \cap V^L_{\mathbf{K}_2} = \left\{ \left(x,\ \frac{1}{4} - x,\ \frac{1}{4} - x,\ \frac{1}{2} + x\right),\ x \in \left[0.01, \frac{0.96}{4}\right] \right\}.$$

We compute that $\mathbf{CM}^\infty_L(\mathbf{K}_1 \cup \mathbf{K}_2)$ is approximately

$$(0.091506,\ 0.15849,\ 0.15849,\ 0.5915),$$

which is obviously not equal to the limit of the sequence $\{\mathbf{v}^{[i]}\}_{i=0}^\infty$.[1] $\square$

It seems that only achievable way to use (5.18) to estimate a result of applying **LEP** is to choose a sufficiently big $m$ and for this $m$ iterate the sequence $\{\mathbf{v}^{[i]}_{[m]}\}_{i=0}^\infty$. However the next theorem shows that the rate of convergence depends on $m$ and in fact this often materialises in a negative way for a practical computation.

**Theorem 5.3.4.** *$\{\mathbf{v}^{[i]}_{[m]}\}_{i=0}^\infty$ does not converge uniformly in $m$.*

*Proof.* Since we can treat probability functions as vectors in $\mathbb{R}^J$ the following variant of Moore-Osgood theorem is relevant here, see [44].

Assume that $v^{[i]}_{[m]} \in \mathbb{R}$ and $v_{[m]}$ and $v^{[i]}$ in $\mathbb{R}$ are given. If

$$\lim_{i\to\infty} v^{[i]}_{[m]} = v_{[m]} \text{ uniformly in } m$$

and

$$\lim_{m\to\infty} v^{[i]}_{[m]} = v^{[i]} \text{ pointwise for every } i$$

then the double limits exists and

$$\lim_{m\to\infty} \lim_{i\to\infty} v^{[i]}_{[m]} = \lim_{i\to\infty} \lim_{m\to\infty} v^{[i]}_{[m]}.$$

But then, if $\{\mathbf{v}^{[i]}_{[m]}\}_{i=0}^\infty$ was convergent uniformly in $m$ then the limits in (5.18) would be interchangeable which would contradict example 5.3.3. $\square$

[1] Just for comparison $\mathbf{ME}_L(\mathbf{K}_1 \cup \mathbf{K}_2)$ is approximately $(0.0625,\ 0.1875,\ 0.1875,\ 0.5625)$.

**Example 5.3.5.** Consider the situation from example 5.3.3. We compute numerically the first members of the sequence $\{\mathbf{v}^{[i]}_{[m]}(p \wedge q)\}_{i=0}^{\infty}$ for several values of $m$ and we compare them with the sequence $\{\mathbf{v}^{[i]}(p \wedge q)\}_{i=0}^{\infty}$. The algorithm we use is as follows. Note that due to the design of the knowledge bases only one minimisation problem is sufficient to solve in each iteration as we have pointed out in the previous example.

$v_1 := \frac{1}{4}$; $v_2 := \frac{1}{4}$; $v_3 := \frac{1}{4}$; $v_4 := \frac{1}{4}$;
**for** $i$ **from** 1 **by** 1 **to** 200 **do**
Minimise$\left(x \log \frac{x}{v_1} + \left(\frac{1}{4} - x\right) \log \frac{\frac{1}{4}-x}{v_2} + y \log \frac{y}{v_3} + \left(\frac{3}{4} - y\right) \log \frac{\frac{3}{4}-y}{v_4}, x = 0.01..\frac{0.96}{4}, y = 0.01..\frac{2.96}{4}\right)$;
$v_1 := \frac{x \cdot m + x \cdot m + \frac{1}{4}}{m+m+1}$; $v_2 := \frac{(\frac{1}{4}-x)\cdot m + y \cdot m + \frac{1}{4}}{m+m+1}$; $v_3 := \frac{(\frac{1}{4}-x)\cdot m + y \cdot m + \frac{1}{4}}{m+m+1}$; $v_4 := \frac{(\frac{3}{4}-y)\cdot m + (\frac{3}{4}-y)\cdot m + \frac{1}{4}}{m+m+1}$;
**end do**;

The numerical result for $m = 10,\ 20,\ 30$ is plotted in figure 5.1. We can see that as $m$ rises the limit points of sequences are converging to the $\mathbf{CM}^{\infty}_{L}(\mathbf{K}_1 \cup \mathbf{K}_2)(p \wedge q)$ which is denoted by the black dotted line. The red line denotes the sequence $\{\mathbf{v}^{[i]}\}_{i=0}^{\infty}(p \wedge q)$.

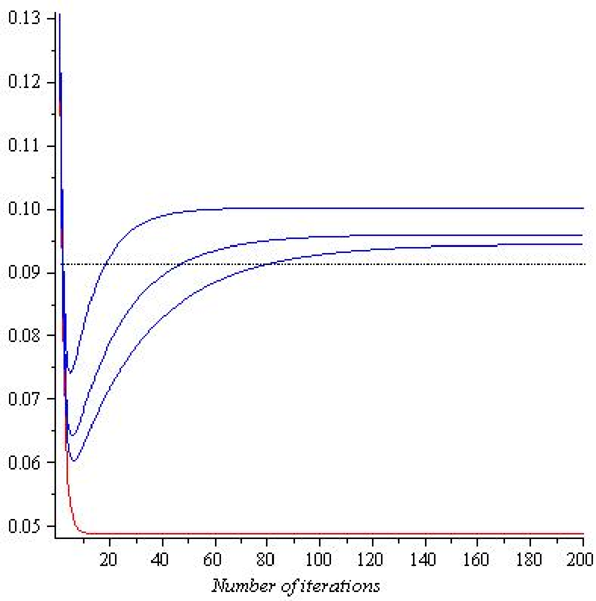


Figure 5.1: The numerical computation for example 5.3.5. Blue lines from top are for $m = 10$, $m = 20$ and $m = 30$.

The numerical result for $m = 30,\ 60,\ 90$ is plotted in figure 5.2. We can conclude that although the eventual precision rises as $m$ increases, the rate of convergence is affected severely. Therefore there is a significant trade-off between the precision and the number of iterations. □

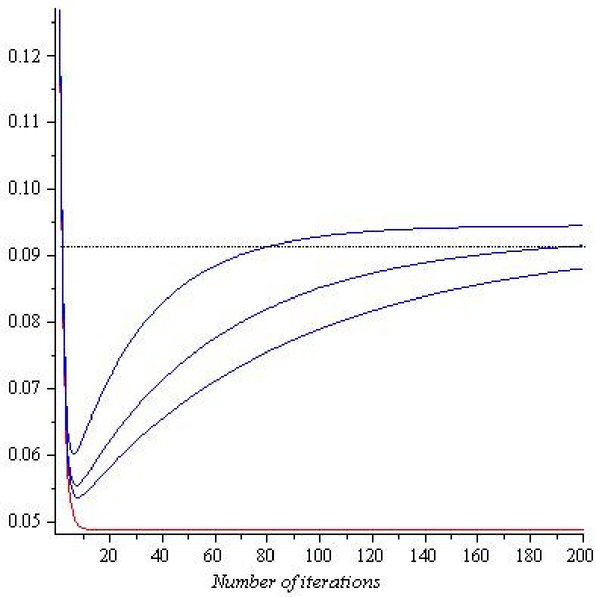


Figure 5.2: The numerical computation for example 5.3.5. Blue lines from top are for $m = 30$, $m = 60$ and $m = 90$.

Now consider the averaging projective procedure $F^{\mathrm{KL}}_{[\mathbf{K}_1,\ldots,\mathbf{K}_n]}$ defined in section 3.2 but also restricted to $BCL$. By theorem 3.2.12 we know that the sequence

$$\{\mathbf{u}^{[i]}\}_{i=0}^{\infty}, \tag{5.20}$$

where $\mathbf{u}^{[0]} = (\frac{1}{J}, \ldots, \frac{1}{J})$ and $\mathbf{u}^{[i+1]} = F^{\mathrm{KL}}_{[\mathbf{K}_1,\ldots,\mathbf{K}_n]}(\mathbf{u}^{[i]})$, converges to some probability function in $\Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. This procedure can be immediately used to compute $\mathbf{SEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ in a case when $\Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is a singleton. By theorem 5.1.3 this happen for instance when at least one of $V^L_{\mathbf{K}_1}, \ldots, V^L_{\mathbf{K}_n}$ is a singleton.

In the following example we will see that there are $\mathbf{K}_1, \ldots, \mathbf{K}_n \in BCL$ such that $\lim_{i\to\infty} \mathbf{u}^{[i]} \neq \mathbf{SEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. Note that we cannot use example 5.3.3 since in that case actually $\lim_{i\to\infty} \mathbf{u}^{[i]} = \mathbf{SEP}_L(\mathbf{K}_1, \mathbf{K}_2)$ due to theorem 3.2.13.

**Example 5.3.6.** Let $L = \{p, q, r\}$, $\mathbf{K}_1 = \{Bel(p \wedge r) = \frac{1}{12}, Bel(q \wedge r) = \frac{1}{12}, Bel(p \wedge \neg r) = \frac{1}{6}, Bel(\neg p \wedge q \wedge \neg r) = \frac{1}{6}, Bel(\neg p \wedge \neg q \wedge \neg r) = \frac{1}{6}\}$, $\mathbf{K}_2 = \{Bel(p \wedge r) = \frac{1}{12}, Bel(q \wedge r) = \frac{1}{12}, Bel(\neg p \wedge \neg r) = \frac{2}{6}, Bel(p \wedge q \wedge \neg r) = \frac{1}{12}, Bel(p \wedge \neg q \wedge \neg r) = \frac{1}{12}\}$ and assume the additional constraint that every atom is bounded by 0.01. Note that

$$V^L_{\mathbf{K}_1} = \left\{\left(x, \frac{1}{12} - x, \frac{1}{12} - x, \frac{2}{6} + x, y, \frac{1}{6} - y, \frac{1}{6}, \frac{1}{6}\right), x \in \left[0.01, \frac{0.88}{12}\right], y \in \left[0.01, \frac{0.94}{6}\right]\right\}$$

and

$$V^L_{\mathbf{K}_2} = \left\{\left(x, \frac{1}{12} - x, \frac{1}{12} - x, \frac{2}{6} + x, \frac{1}{12}, \frac{1}{12}, y, \frac{2}{6} - y\right), x \in \left[0.01, \frac{0.88}{12}\right], y \in \left[0.01, \frac{1.94}{6}\right]\right\}$$

where probability functions are listed in the following order of atoms: $p \wedge q \wedge r, p \wedge \neg q \wedge r, \neg p \wedge q \wedge r, \neg p \wedge \neg q \wedge r, p \wedge q \wedge \neg r, p \wedge \neg q \wedge \neg r, \neg p \wedge q \wedge \neg r, \neg p \wedge \neg q \wedge \neg r$.

$\mathbf{K}_1$ and $\mathbf{K}_2$ are jointly consistent; $V^L_{\mathbf{K}_1} \cap V^L_{\mathbf{K}_2} = \{(x, \frac{1}{12} - x, \frac{1}{12} - x, \frac{2}{6} + x, \frac{1}{12}, \frac{1}{12}, \frac{1}{6}, \frac{1}{6}),\ x \in [0.01, \frac{0.88}{12}]\}$ and we can compute that $\mathbf{SEP}_L(\mathbf{K}_1, \mathbf{K}_2)$ is the most entropic probability function from the set above with $x$ equals approximately 0.013888.

However the sequence $\{\mathbf{u}^{[i]}\}_{i=0}^{\infty}$ is already constant after one iteration and equal to $\mathbf{CM}^{\infty}_L(\mathbf{K}_1) = \mathbf{CM}^{\infty}_L(\mathbf{K}_2) = \mathbf{CM}^{\infty}_L(V^L_{\mathbf{K}_1} \cap V^L_{\mathbf{K}_2})$ in which case $x \approx 0.029231$. □

By the aid of the chairman theorem for $\Delta^{\mathrm{KL}}$ (5.1.4) we also suggest a modification of the above procedure to approximate $\mathbf{SEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$, but we have no claims to any theoretical results on the complexity of computation. Let $\mathbf{I} \in CL$ be such that $V^L_{\mathbf{I}} = \{\mathbf{t}\}$, where $\mathbf{t} = (\frac{1}{J}, \ldots, \frac{1}{J})$ is the uniform probability function. For every $m = 1, 2, \ldots$ we define the sequence $\{\mathbf{u}^{[i]}_{[m]}\}_{i=0}^{\infty}$ by $\mathbf{u}^{[0]}_{[m]} = (\frac{1}{J}, \ldots, \frac{1}{J})$ and

$$\mathbf{u}^{[i+1]}_{[m]} = F^{\mathrm{KL}}_{[\mathbf{I}, \underbrace{\mathbf{K}_1 \ldots \mathbf{K}_1}_{m}, \ldots, \underbrace{\mathbf{K}_n \ldots \mathbf{K}_n}_{m}]}(\mathbf{u}^{[i]}_{[m]}).$$

By theorem 3.2.12

$$\lim_{i \to \infty} \mathbf{u}^{[i]}_{[m]} = \mathbf{SEP}_L(\mathbf{I}, \underbrace{\mathbf{K}_1 \ldots \mathbf{K}_1}_{m}, \ldots, \underbrace{\mathbf{K}_n \ldots \mathbf{K}_n}_{m}).$$

By the chairman theorem for $\Delta^{\mathrm{KL}}$

$$\lim_{m \to \infty} \lim_{i \to \infty} \mathbf{u}^{[i]}_{[m]} = \mathbf{SEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n). \tag{5.21}$$

In order to approximate $\mathbf{SEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ using (5.21) one needs to choose a sufficiently big $m$ and then iterate the sequence $\{\mathbf{u}^{[i]}_{[m]}\}_{i=0}^{\infty}$. However the question as to how to determine such an $m$ and $i$ in order to achieve a specific level of accuracy merits further investigation.

## 5.4 The toy example again

In this section we will go back to our motivation example from section 1.1. Taking an advantage of the previously suggested algorithms we compute what result the linear entropy process **LEP** and the social entropy process **SEP** give in this particular case.

Recall that in this example two safety experts are evaluating safety in a chemical factory. The first expert believes that there is a 4% chance that a mechanical problem will cause the valve to fail. The second expert comes up with a different opinion that there is an 8% chance that a mechanical problem will cause the valve to fail. Moreover, the first safety expert thinks that there is a 7% chance that the electronic switch will fail. We suppose that both experts have no other knowledge related to this problem.

To keep us confined to $BCL$ we will need to add additional constraints that the values of probability functions on atoms are bounded, say by 0.001. The knowledge bases of the two safety experts generate in this case the following two closed and convex sets of probability functions.

$$V^L_{\mathbf{K}_1} = \{(x, 0.04 - x, 0.07 - x, x + 0.89),\ x \in [0.001, 0.039]\} \text{ and}$$

$$V^L_{\mathbf{K}_2} = \{(x, 0.08 - x, y, 0.92 - y),\ x \in [0.001, 0.079], y \in [0.001, 0.919\},$$

where $L = \{p, q\}$ and $p$ stands for sentence "a fault on the valve" and $q$ stands for sentence "a fault on the electronic switch". Probability functions are listed in the following order of atoms: $p \wedge q,\ p \wedge \neg q,\ \neg p \wedge q,\ \neg p \wedge \neg q$.

First we approximate the value of $\mathbf{LEP}_L(\mathbf{K}_1, \mathbf{K}_2)$. To make sure that the approximation is accurate enough we choose $m = 10000$ and apply the following algorithm.

$v_1 := \frac{1}{4};\ v_2 := \frac{1}{4};\ v_3 := \frac{1}{4};\ v_4 := \frac{1}{4};$

**for** $i$ **from** 1 **by** 1 **to** 50 **do**

Minimise$\Big(x_1 \log \frac{x_1}{v_1} + (0.04 - x_1) \log \frac{0.04-x_1}{v_2} + (0.07 - x_1) \log \frac{0.07-x_1}{v_3} + (0.89 + x_1) \log \frac{0.89+x_1}{v_4}, x_1 = 0.001..0.039\Big)$;

Minimise$\Big(x_2 \log \frac{x_2}{v_1} + (0.08 - x_2) \log \frac{0.08-x_2}{v_2} + y_2 \log \frac{y_2}{v_3} + (0.92 - y_2) \log \frac{0.92-y_2}{v_4}, x_2 = 0.001..0.079, y_2 = 0.001..0.919\Big)$;

$v_1 := \frac{x_1 \cdot m + x_2 \cdot m + \frac{1}{4}}{m+m+1};\ v_2 := \frac{(0.04-x_1)\cdot m + (0.08-x_2)\cdot m + \frac{1}{4}}{m+m+1};\ v_3 := \frac{(0.07-x_1)\cdot m + y_2 \cdot m + \frac{1}{4}}{m+m+1};\ v_4 := \frac{(0.89+x_1)\cdot m + (0.92-y_2)\cdot m + \frac{1}{4}}{m+m+1};$

**end do**;

After 50 iterations the result is approximately as follows: $\mathbf{v}^{[50]}_{[10000]}(p \wedge q) \approx 0.0096514$, $\mathbf{v}^{[50]}_{[10000]}(p \wedge \neg q) \approx 0.0503705$, $\mathbf{v}^{[50]}_{[10000]}(p \wedge \neg q) \approx 0.0622752$ and $\mathbf{v}^{[50]}_{[10000]}(\neg p \wedge \neg q) \approx 0.8777027$. Figure 5.3 shows that 50 iterations are more than enough to get very close to the limit point of $\{\mathbf{v}^{[i]}_{[10000]}\}_{i=0}^{\infty}$.

Although **LEP** does not satisfy the consistent irrelevant information principle in this particular case we can use the fact that the linear entropy operator $\hat{\Delta}^{\mathrm{KL}}$ satisfies **CIIP**. If we denote $\mathbf{F}_1 = \{Bel(p) = 0.04\}$ and $\mathbf{F}_2 = \{Bel(p) = 0.08\}$ then

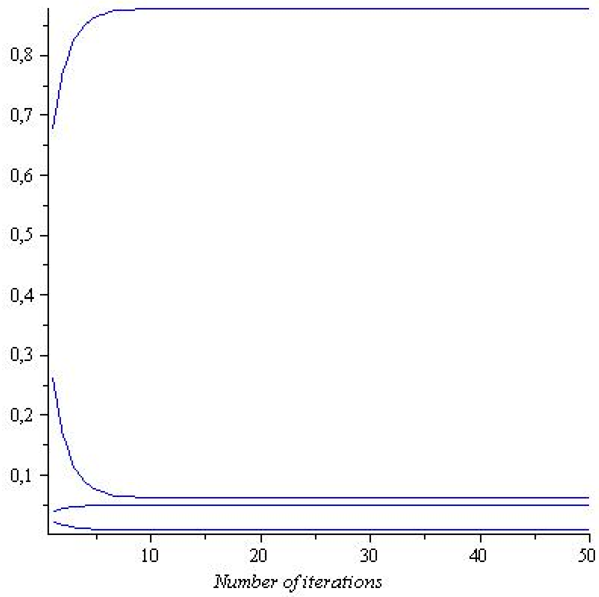


Figure 5.3: The rate of convergence when approximating $\mathbf{LEP}_L(\mathbf{K}_1, \mathbf{K}_2)$ for $m = 10000$ and for all four atoms.

there is only one point in $\hat{\Delta}^{\mathrm{KL}}_{\{p\}}(\mathbf{F}_1, \mathbf{F}_2) = \{(0.06, 0.94)\}$. By **CIIP** we have that $\mathbf{LEP}_L(\mathbf{K}_1, \mathbf{K}_2)|_{\{p\}} \in \hat{\Delta}^{\mathrm{KL}}_{\{p\}}(\mathbf{F}_1, \mathbf{F}_2)$ and consequently $\mathbf{LEP}_L(\mathbf{K}_1, \mathbf{K}_2)(p) = 0.06$. Now we can use this to check the accuracy of the computation above which gives

$$\mathbf{v}^{[50]}_{[10000]}(p) = \mathbf{v}^{[50]}_{[10000]}(p \wedge q) + \mathbf{v}^{[50]}_{[10000]}(p \wedge \neg q) \approx 0.0600219$$

and that is fairly accurate.

We perform a similar computation to approximate $\mathbf{SEP}_L(\mathbf{K}_1, \mathbf{K}_2)$. The dual algorithm to the one above is as follows.

$v_1 := \frac{1}{4}$; $v_2 := \frac{1}{4}$; $v_3 := \frac{1}{4}$; $v_4 := \frac{1}{4}$;

**for** $i$ **from** 1 **by** 1 **to** 50 **do**

Minimise$\left(v_1 \log \frac{v_1}{x_1} + v_2 \log \frac{v_2}{0.04-x_1} + v_3 \log \frac{v_3}{0.07-x_1} + v_4 \log \frac{v_4}{0.89+x_1}, x_1 = 0.001..0.039\right)$;

Minimise$\left(v_1 \log \frac{v_1}{x_2} + v_2 \log \frac{v_2}{0.08-x_2} + v_3 \log \frac{v_3}{y_2} + v_4 \log \frac{v_4}{0.92-y_2}, x_2 = 0.001..0.079, y_2 = 0.001..0.919\right)$;

$c := \left((x_1)^m (x_2)^m \frac{1}{4}\right)^{\frac{1}{2m+1}} + \left((0.04 - x_1)^m (0.08 - x_2)^m \frac{1}{4}\right)^{\frac{1}{2m+1}} + \left((0.07 - x_1)^m (y_2)^m \frac{1}{4}\right)^{\frac{1}{2m+1}} + \left((0.89 + x_1)^m (0.92 - y_2)^m \frac{1}{4}\right)^{\frac{1}{2m+1}}$;

$v_1 := \frac{\left((x_1)^m (x_2)^m \frac{1}{4}\right)^{\frac{1}{2m+1}}}{c}$; $v_2 := \frac{\left((0.04-x_1)^m (0.08-x_2)^m \frac{1}{4}\right)^{\frac{1}{2m+1}}}{c}$;

$v_3 := \frac{\left((0.07-x_1)^m (y_2)^m \frac{1}{4}\right)^{\frac{1}{2m+1}}}{c}$; $v_4 := \frac{\left((0.89+x_1)^m (0.92-y_2)^m \frac{1}{4}\right)^{\frac{1}{2m+1}}}{c}$;

**end do**;

After 50 iterations the result is approximately as follows: $\mathbf{u}^{[50]}_{[10000]}(p \wedge q) \approx 0.013163$, $\mathbf{u}^{[50]}_{[10000]}(p \wedge \neg q) \approx 0.0436207$, $\mathbf{u}^{[50]}_{[10000]}(p \wedge \neg q) \approx 0.0596818$ and $\mathbf{u}^{[50]}_{[10000]}(\neg p \wedge \neg q) \approx 0.8835343$. Figure 5.4 shows that 50 iterations are more than enough to get very close to the limit point of $\{\mathbf{u}^{[i]}_{[10000]}\}_{i=0}^{\infty}$.

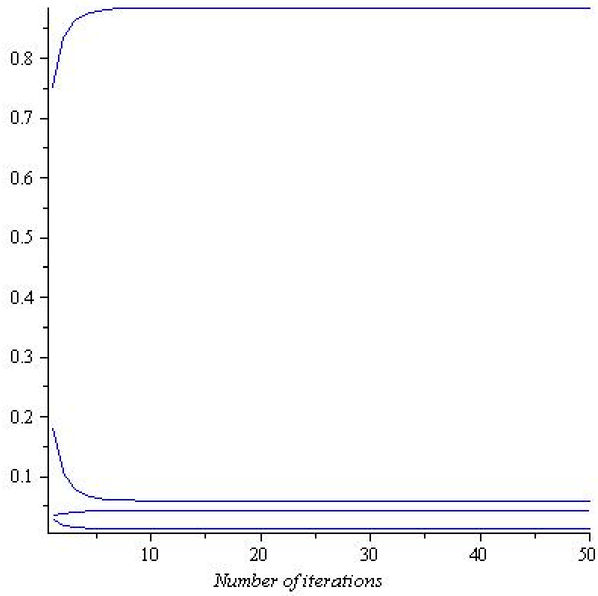


Figure 5.4: The rate of convergence when approximating $\mathbf{SEP}_L(\mathbf{K}_1, \mathbf{K}_2)$ for $m = 10000$ and for all four atoms.

Similarly as in the case of **LEP**, we take an advantage of the fact that $\Delta^{\mathrm{KL}}$ satisfies **CIIP**. For this reason we know that

$$\mathbf{SEP}_L(\mathbf{K}_1, \mathbf{K}_2)(p) = \mathbf{SEP}_{\{p\}}(\{Bel(p) = 0.04\}, \{Bel(p) = 0.08\})(p) \approx 0.0567754\ .$$

Our algorithm took us fairly close to that value:

$$\mathbf{u}^{[50]}_{[10000]}(p) = \mathbf{u}^{[50]}_{[10000]}(p \wedge q) + \mathbf{u}^{[50]}_{[10000]}(p \wedge \neg q) \approx 0.0567837\ .$$

Finally, we provide the answers which **LEP** and **SEP** give to the question which started this thesis: 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 electronic switch will be faulty, based only on the experts' subjective knowledge and without any other assumptions? By **LEP** this probability should be approximately 0.96% and by **SEP** it should be approximately 1.31%. But whether a rational adjudicator should adopt either of these procedures is a different question altogether. Throughout this thesis we have argued that both these p-merging operators have appealing properties in comparison to other operators considered. However there are many places where this argument fall short. A full axiomatic characterisation, similar to one of those used to justify the maximum entropy inference process, would be preferred.

# Chapter 6

# Conclusion

## 6.1 Summary

In this thesis we have pursued the problem of probabilistic knowledge integration by combining both the geometrical notion of projections by means of a Bregman divergence and the framework of pooling operators. Our first original result (theorem 1.4.1) showed that the most obvious approach of obdurate merging does not satisfy the natural principle of consistency.

In chapter 2 we have therefore studied the geometry of convex Bregman divergences and we have shown how Bregman divergences relate to pooling operators. We have proved several technical results which have been used later in the thesis. In particular theorem 2.1.20 is the main original contribution of the chapter. Several dual results for the Kullback-Leibler divergence have been also listed.

In chapter 3 we have defined probabilistic merging operators $\hat{\Delta}^D$ and $\Delta^{\mathrm{KL}}$, where $D$ is a convex Bregman divergence and KL is the Kullback-Leibler divergence, which satisfy the consistency principle and extend the framework of pooling operators. We have shown how these operators relate to the existing results in the literature. In particular we have shown that their images coincide with the sets of the fixed points of certain mappings under fairly general conditions. These results (essentially theorems 3.2.5 and 3.2.10) are perhaps the most important in the thesis. Moreover they have been used to prove Matúš's convergence theorems (3.2.12) and to prove results relating to the strong disagreement principle, namely theorems 4.1.5 and 4.1.10.

| | HULL | $\Delta^{\mathrm{KL}}$ | $\hat{\Delta}^{\mathrm{KL}}$ | $\hat{\Delta}^{D_r}$ | **OSEP** | **OLEP** | **KIRP** |
|---|---|---|---|---|---|---|---|
| **(K2)** | **YES** | **YES** | **YES** | **YES** | **YES** | **YES** | **YES** |
| **(K3)** | **YES** | **YES** | **YES** | **YES** | **YES** | **YES** | **YES** |
| **(K4)** | **YES** | **YES** | **YES** | **YES** | **NO**[1.4.1] | **NO**[1.4.1] | **NO**[4.4.3] |
| **(K4*)** | **YES** | **YES** | **YES** | **YES** | **NO**[1.4.1] | **NO**[1.4.1] | **NO**[4.4.3] |
| **(K5)** | **YES**[4.1.3] | **YES**[4.1.12] | **YES**[4.1.1] | **YES**[4.1.1] | **NO**[1.4.2] | **NO**[1.4.2] | **NO**[4.4.4] |
| **(K6)** | **NO**[4.1.4] | **NO**[4.1.9] | **NO**[4.1.8] | **YES**[4.1.7] | **YES** | **YES** | **YES**[4.4.2] |
| **(K6*)** | **NO**[4.1.4] | **NO**[4.1.9] | **NO**[4.1.8] | **YES**[4.1.7] | **YES** | **YES** | **YES**[4.4.2] |
| **(K7)** | **NO**[4.1.4] | **NO**[4.1.11] | **YES**[4.1.11] | **YES**[4.1.11] | **YES** | **YES** | **YES**[4.4.2] |
| **REL** | **NO**[4.5.7] | **NO**[4.5.4] | **YES**[4.5.1] | **YES**[4.5.2] | **YES**[4.5.5] | **YES**[4.5.5] | **NO**[4.5.6] |
| **LI** | **YES**[4.2.4] | **YES**[4.2.7] | **YES**[4.2.9] | **YES**[4.2.11] | **YES**[4.3.5] | **YES**[4.3.6] | **YES**[4.4.5] |
| **CIIP** | **YES**[4.2.4] | **YES**[4.2.7] | **YES**[4.2.9] | ? | **YES**[4.3.5] | **YES**[4.3.6] | **NO**[4.4.6] |
| **IIP** | **NO**[4.2.3] | **NO**[4.2.1] | **NO**[4.2.2] | ? | **YES**[4.3.5] | **YES**[4.3.6] | **NO**[4.4.6] |
| **IP** | **NO**[4.2.3] | **NO**[4.2.1] | **NO**[4.2.2] | ? | **YES**[4.3.5] | **NO**[4.3.6] | **NO**[4.4.6] |

Table 6.1: The list of principles and p-merging operators.

To summarise chapter 4 where we have compared our p-merging operators with those investigated elsewhere in the literature we list in table 6.1 the probabilistic merging operators defined over $CL$. On the other hand, if we confine ourselves only to $BCL$ then we get table 6.2. Each number refers to an example or a theorem where the corresponding result has been proved. Recall that **(K2)** is the principle of equivalence, **(K3)** is the principle of atomic renaming, **(K4)** is the principle of consistency, **(K4*)** is the principle of strong consistency, **(K5)** is the collegiality principle, **(K6)** is the principle of disagreement, **(K6*)** is the principle of strong disagreement, **(K7)** is the principle of agreement, **REL** is the principle of relativisation, **LI** is the principle of language invariance, **CIIP** is the principle of consistent irrelevant information, **IIP** is the principle of irrelevant information, and finally **IP** is the principle of independence.

We have seen that obdurate inference processes **OSEP** and **OLEP** possess some excellent properties. However it seems that the price we need to pay for this, in the form that they satisfy neither the consistency principle **(K4)** nor the very natural collegiality principle **(K5)**, is too high. We believe that these principles are absolutely necessary. The same criticism thus applies to **KIRP**.

The convex hull operator HULL has its weakness in terms of the principles of agreement, disagreement and relativisation. However it satisfies **(K4)**, it is language invariant and it satisfies the pleasing principle of consistent irrelevant information.

The social entropy operator $\Delta^{\mathrm{KL}}$ and the linear entropy operator $\hat{\Delta}^{\mathrm{KL}}$ seem to be

| | HULL | $\Delta^{\mathrm{KL}}$ | $\hat{\Delta}^{\mathrm{KL}}$ | $\hat{\Delta}^{D_r}$ | **OSEP** | **OLEP** | **KIRP** |
|---|---|---|---|---|---|---|---|
| **(K2)** | **YES** | **YES** | **YES** | **YES** | **YES** | **YES** | **YES** |
| **(K3)** | **YES** | **YES** | **YES** | **YES** | **YES** | **YES** | **YES** |
| **(K4)** | **YES** | **YES** | **YES** | **YES** | **NO**$^{1.4.1}$ | **NO**$^{1.4.1}$ | **NO**$^{4.4.3}$ |
| **(K4*)** | **YES** | **YES** | **YES** | **YES** | **NO**$^{1.4.1}$ | **NO**$^{1.4.1}$ | **NO**$^{4.4.3}$ |
| **(K5)** | **YES**$^{4.1.3}$ | **YES**$^{4.1.1}$ | **YES**$^{4.1.1}$ | **YES**$^{4.1.1}$ | **NO**$^{1.4.2}$ | **NO**$^{1.4.2}$ | **NO**$^{4.4.4}$ |
| **(K6)** | **NO**$^{4.1.4}$ | **YES**$^{4.1.10}$ | **YES**$^{4.1.5}$ | **YES**$^{4.1.7}$ | **YES** | **YES** | **YES**$^{4.4.2}$ |
| **(K6*)** | **NO**$^{4.1.4}$ | **YES**$^{4.1.10}$ | **YES**$^{4.1.5}$ | **YES**$^{4.1.7}$ | **YES** | **YES** | **YES**$^{4.4.2}$ |
| **(K7)** | **NO**$^{4.1.4}$ | **YES**$^{4.1.11}$ | **YES**$^{4.1.11}$ | **YES**$^{4.1.11}$ | **YES** | **YES** | **YES**$^{4.4.2}$ |
| **REL** | **NO**$^{4.5.7}$ | **YES**$^{4.5.3}$ | **YES**$^{4.5.1}$ | **YES**$^{4.5.2}$ | **YES**$^{4.5.5}$ | **YES**$^{4.5.5}$ | **NO**$^{4.5.6}$ |
| **LI** | **YES**$^{4.2.4}$ | **YES**$^{4.2.7}$ | **YES**$^{4.2.9}$ | **YES**$^{4.2.11}$ | **YES**$^{4.3.5}$ | **YES**$^{4.3.6}$ | **YES**$^{4.4.5}$ |
| **CIIP** | **YES**$^{4.2.4}$ | **YES**$^{4.2.7}$ | **YES**$^{4.2.9}$ | ? | **YES**$^{4.3.5}$ | **YES**$^{4.3.6}$ | **NO**$^{4.4.6}$ |
| **IIP** | **NO**$^{4.2.3}$ | **NO**$^{4.2.1}$ | ? | ? | **YES**$^{4.3.5}$ | **YES**$^{4.3.6}$ | **NO**$^{4.4.6}$ |
| **IP** | **NO**$^{4.2.3}$ | **NO**$^{4.2.1}$ | ? | ? | **YES**$^{4.3.5}$ | **NO**$^{4.3.6}$ | **NO**$^{4.4.6}$ |

Table 6.2: The list of principles and p-merging operators restricted to $BCL$.

attractive when compared to other p-merging operators hitherto considered. However $\Delta^{\mathrm{KL}}$ does not satisfy the most interesting principles concerning irrelevant information and independence, and the same is conjectured for $\hat{\Delta}^{\mathrm{KL}}$ even once we have confined ourselves to $BCL$.

The linear Renyi operator $\hat{\Delta}^{D_r}$, $2 \geq r > 1$, seems to be as interesting as the operators $\Delta^{\mathrm{KL}}$ and $\hat{\Delta}^{\mathrm{KL}}$. There are still some unanswered questions about this operator but it is promising as it generalises the linear Euclidean operator and there is no need to restrict it to $BCL$.

In chapter 5 we have investigated how to modify $\hat{\Delta}^{\mathrm{KL}}$ and $\Delta^{\mathrm{KL}}$-merging operators in order to obtain a single point as the result of a merging procedure. We have used the 'Chairman idea' due to Wilmers and similarly as he has defined operator **SEP** from the social entropy operator we have defined operator **LEP** from the linear entropy operator. We recapitulate the properties of **LEP** and **SEP** in table 6.3. Finally, by combining the 'Chairman idea' with Matúš's convergence theorems we have proposed possible algorithms to compute their results but with no claims on computational feasibility.

| | LEP | SEP | OSEP | OLEP |
|---|---|---|---|---|
| **(K2)** | **YES** | **YES** | **YES** | **YES** |
| **(K3)** | **YES** | **YES** | **YES** | **YES** |
| **(K4)** | **YES** | **YES** | **NO** | **NO** |
| **(K5)** | **YES** | **YES** | **NO** | **NO** |
| **(K6)** | **NO** | **NO** | **YES** | **YES** |
| **(K6*)** | **NO** | **NO** | **YES** | **YES** |
| **(K7)** | **YES** | **YES** | **YES** | **YES** |
| **Savage** | **YES** | **YES** | **NO** | **NO** |
| **REL** | **YES** | **YES** | **YES** | **YES** |
| **LI** | **YES** | **YES** | **YES** | **YES** |
| **CIIP** | **NO** | ? | **YES** | **YES** |
| **IIP** | **NO** | **NO** | **YES** | **YES** |
| **IP** | **NO** | **NO** | **YES** | **NO** |

Table 6.3: The list of principles to compare **LEP** with **SEP** over $BCL$.

## 6.2 Future research

Although the problem of finding a full axiomatic characterisation of particular p-merging operators may be hard, there are several obvious open questions regarding the principles for probabilistic merging which we have considered in this work. Perhaps the most interesting question is whether there exists a social inference process (or a naturally defined p-merging operator) which satisfies the principle of irrelevant information as well as the principles of consistency and collegiality.

In section 3.2 we have argued that fixed points of the averaging projective procedure and the conjugated averaging projective procedure (defined by KL-divergence) may be viewed as points of equilibrium for a group seeking agreement. If we add that in addition we expect this equilibrium to be stable with respect to infinitesimal bias by the most uninformative point then we are left with the **LEP** and **SEP** respectively.

Note however, that here we assume that the KL-divergence is the preferred choice for projections (or conjugated projections). As we have mentioned earlier, there are certainly arguments arising from Shannon entropy in favour of KL-divergence, and currently it is the most popular measure of information distance. Nevertheless, other convex Bregman divergences may perhaps be as appealing in some respects.

In particular we believe that an investigation of the Renyi-B divergence is promising since it is already known that a theorem analogous to the chairman theorem holds for the linear Renyi operator $\hat{\Delta}^{D_r}$, $2 \geq r > 1$. We then conjecture that for any $\mathbf{K}_1, \ldots, \mathbf{K}_n \in BCL$

$$\lim_{r \searrow 1} \arg \min_{\mathbf{v} \in \hat{\Delta}_L^{D_r}(\mathbf{K}_1, \ldots, \mathbf{K}_n)} D_r(\mathbf{t} \| \mathbf{v}) = \mathbf{LEP}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n),$$

where $\mathbf{t} \in \mathbb{D}^L$ is the uniform probability function. The proof should involve theorem 1.3.4. Considering conjugated projections of Renyi-B divergence for $2 \geq r > 1$ may give us a dual result for **SEP**.

Moreover, the result due to Csiszár and Tusnády, on which the proof of theorem 3.2.12 is based, was observed also for Bregman divergences, see theorem 2.17 of [14]. It may be therefore possible to extend the algorithm we proposed in section 5.3 also for the linear Renyi operator.

As we have seen in the examples in chapter 5 where **LEP** and **SEP** were computed, one of the problems appeared to be the computation of the KL-projection (respectively the conjugated KL-projection). As we have noted before, this problem was studied in the literature and there are some results to build on which can be used to develop further the framework presented in this thesis. However, for more general applications than our toy examples, the rate of convergence of the algorithm we proposed in section 5.3 needs to be carefully investigated.

# Bibliography

[1] Adamčík, M. and Wilmers, G.M. (2012), *The Irrelevant Information Principle for Collective Probabilistic Reasoning.* Proceedings of the 9th Workshop on Uncertain Processing, Mariánské Lázně, pp. 1-12. Awaiting publication in Kybernetika.

[2] Adamčík, M. and Wilmers, G.M. (2013), *Probabilistic Merging Operators.* Awaiting publication in Logique et Analyse.

[3] Amari, S. (2009), *Divergence, Optimization and Geometry.* Neural Information Processing: 16th International Conference, Iconip, pp. 185-193.

[4] Amari, S. and Cichocki, A. (2010), *Information geometry of divergence functions.* Bulletin of the Polish academy of sciences 58(1), pp. 183-195.

[5] Banerjee, A., Merugu, S., Dhillon, I.S. and Ghosh, J. (2005), *Clustering with Bregman Divergences.* Journal of Machine Learning Research 6, pp. 1705-1749.

[6] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization.* Cambridge University Press, Cambridge.

[7] Bregman, L.M. (1967). *The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming.* USSR Computational Mathematics and Mathematical Physics 7(3), pp. 200-217.

[8] Carnap, R. (1947) *On the application of inductive logic.* Philosophy and Phenomenological Research 8, pp. 133-148.

[9] Collins, M. and Schapire, R.E. (2002), *Logistic Regression, AdaBoost and Bregman Distances.* Machine Learning 48, pp. 253-285.

[10] Cover, T.M. and Thomas, J.A. (1991), *Elements of Information Theory.* Wiley Series in Telecommunications, John Wiley and Sons, New York.

[11] Csiszár, I. (1975), *I-Divergence Geometry of Probability Distribution and Minimization Problems.* The Annals of Probability 3(1), pp. 146-158.

[12] Csiszár, I. and Tusnády, G. (1984), *Informational Geometry and Alternating Minimization Procedures.* Statistic and Decisions 1, pp. 205-237.

[13] Deming, W. E. and Stephan, F. F. (1940), *On a least square adjustment of a sampled frequency table when the expected marginals totals are unknown.* Annals of Mathematical Statistics 11, pp. 427-444.

[14] Eggermont, P.P.B. and LaRiccia, V.N. (1998), *On EM-like algorithms for minimum distance estimation*, University of Delaware, Unpublished manuscript.

[15] de Finetti, B. (1931) *Sul Significato Soggettivo della Probabilitá.* Fundamenta Mathematicae 17, pp. 298-329.

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

[17] Genest, Ch. and Wagner, C.G. (1987), *Further Evidence Against Independence Preservation in Expert Judgement Synthesis.* Aequationes Mathematicae 32, pp. 74-86.

[18] Genest, C. and Zidek, J.V. (1986), *Combining probability distributions: A critique and an annotated bibliography.* Statistical Science 1(1), pp. 114-135.

[19] Gneiting, T. and Raftery, A.E. (2007), *Strictly Proper Scoring Rules, Prediction and Estimation.* Journal of American Statistical Association 102(477), pp. 359-377.

[20] Good, I.J. (1952), *Rational Decisions.* Journal of the Royal Statistical Society B-14, pp. 107-114.

[21] Grünwald, P. and Dawid, A.P. (2004), *Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory.* Annals of Statistics 32(4), pp. 1367-1433.

[22] Hájek, P., Havránek, T. and Jiroušek, J. (1992), *Uncertain Information Processing in Expert Systems.* CRC Press, Boca Raton, Ann Arbor, London, Tokyo.

[23] Hardy, G.H., Littlewood, J.E. and Pólya, G. (1934), *Inequalities.* Cambridge University Press, Cambridge.

[24] Hawes, P. (2007), *An Investigation of Properties of Some Inference Processes.* PhD Thesis, University of Manchester, Manchester.

[25] Jaynes, E.T. (1979), *Where do we Stand on Maximum Entropy?* The Maximum Entropy Formalism, edited by R.D. Levine and M. Tribus, M.I.T. Press, pp. 15-118.

[26] Jaynes, E.T. (2003), *Probability Theory: The Logic of Science.* Cambridge University Press, Cambridge.

[27] Johnson, R.W. and Shore, J.E. (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.

[28] Kahneman, D. and Tversky, A. (1979), *Prospect Theory: An Analysis of Decision under Risk.* Econometrica 47(2), pp. 263-291.

[29] Kern-Isberner, G. and Rödder, W. (2004), *Belief Revision and Information Fusion on Optimum Entropy.* International Journal of Inteligent System 19, pp. 837-857.

[30] Konieczny, S. and Pino-Pérez, R. (1998), *On the Logic of Merging.* Proceedings of the 6th International Conference on Principles of Knowledge Representation and Reasoning, San Francisco, pp. 488-498.

[31] Konieczny, S. and Pino-Pérez, R. (2011), *Logic Based Merging.* J. Philosophical Logic 40, pp. 239-270.

[32] König, H. (1992), *A general minimax theorem based on connectedness.* Archiv der Mathematik 59, pp. 55-64.

[33] Landes, J. and Williamson, J. (2013), *Objective Bayesianism and the Maximum Entropy Principle.* Entropy 15, pp. 3528-3591.

[34] Matúš, F. (1998), *On iterated averages of I-projections.* Unpublished manuscript.

[35] McCarthy, J. (1956), *Measures of the value of information.* Proceedings of the National Academy of Sciences 42, pp. 654-655.

[36] Mitchell, J. (2013), *The Recalibrated and Copula Opinion Pools.* Warwick Business School.

[37] Mohamed, I. (1998), *MSc Thesis Manchester University.*

[38] Osherson, D. and Vardi, M. (2006), *Aggregating disparate estimates of chance.* Games and Economic Behavior 56(1), pp. 148-173.

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

[40] Paris, J.B. and Vencovská, A. (1989), *On the Applicability of Maximum Entropy to Inexact Reasoning.* International Journal of Approximate Reasoning 3(1), pp. 1-34.

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

[42] Predd, J.B., Osherson, D.N., Kulkarni, S.R. and Poor H.V. (2008), *Aggregating Probabilistic Forecasts from Incoherent and Abstaining Experts.* Decision Analysis 5(4), pp. 177-189.

[43] Rényi, A. (1961), *On measures of information and entropy.* Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability, pp. 547-561.

[44] Rudin, W., (1976) *Principles Of Mathematical Analysis.* McGraw-Hill Higher Education, 3rd edition.

[45] Savage, L.J. (1971), *Elicitation of personal probabilities and expectations*. Journal of the American Statistical Association 66, pp. 783-801.

[46] Savage, S. (2010), *The Logical and Philosophical Foundations of Social Inference Processes*. MSc Dissertation, University of Manchester.

[47] Shao, Y.P. (1999), *Expert systems diffusion in British banking: Diffusion models and media factor*. Information and Management 35(1), pp. 1-8.

[48] Vomlel, J. (1999), *Methods of Probabilistic Knowledge Integration*. PhD Thesis, Czech Technical University, Prague.

[49] Williamson, J. (2009), *Aggregating Judgements by Merging Evidence*. J. Logic Computation 19(3), pp. 461-473.

[50] Williamson, J. (2010), *In Defense of Objective Bayesianism*. Oxford University Press, Oxford.

[51] Williamson, J. (2013), *Deliberation, Judgement and the Nature of Evidence*. Economics and Philosophy, in press.

[52] Wilmers, G.M. (2010), *The Social Entropy Process: Axiomatising the Aggregation of Probabilistic Beliefs*. Probability, Uncertainty and Rationality edited by H. Hosni and F. Montagna, 10 CRM series, pp. 87-104.

[53] Wilmers, G.M. (2011), *Generalising the Maximum Entropy Inference Process to the Aggregation of Probabilistic Beliefs*. Unpublished manuscript.

# Index