Logique & Analyse **228** (2014), 563–590

PROBABILISTIC MERGING OPERATORS

MARTIN ADAMČÍK and GEORGE WILMERS

ABSTRACT

The present work presents a general theoretical framework for the study of operators which merge *partial* probabilistic knowledge from different sources which are individually consistent, but may be collectively inconsistent. We consider a number of principles for such an operator to satisfy including a set of principles derived from those of Konieczny and Pino Pérez [14] which were formulated for the different context of propositional merging. Finally we investigate two specific such merging operators derived from the Kullback-Leibler notion of informational distance: the social entropy operator, and its dual, the linear entropy operator. The first of these is strongly related to both the multi-agent normalised geometric mean pooling operator and the single agent maximum entropy inference process, **ME**. By contrast the linear entropy operator is similarly related to both the arithmetic mean pooling operator and the limit centre of mass inference process, $\mathbf{CM}^{\infty}$.

KEYWORDS

Uncertain reasoning, probability function, merging, Kullback-Leibler, divergence, probabilistic merging, merging operator, Konieczny and Pino Pérez, social entropy process, inference process, aggregation of probabilities, pooling operator, probabilistic inference, maximum entropy.

1. Introduction

This work studies some of the global logical desiderata which a well-defined process for merging partial probabilistic knowledge should satisfy. The probabilistic knowledge is thought of as arising from a finite set of agents (or sources), each of which declares her own consistent probabilistic knowledge base, while the probabilistic knowledge from all the agents together is typically inconsistent. The objective of such a merging process is to combine the probabilistic knowledge from a set of such agents into a single consistent probabilistic knowledge base, which best represents the *declared* knowledge of all the agents, *on the assumption that each agent has incorporated all of her relevant knowledge into her declared knowledge base*.

In the present work we shall confine ourselves to considering the simplest notion of a probabilistic knowledge base, along lines previously formulated

doi: 10.2143/LEA.228.0.3078175.

by Paris and Vencovská in their foundational axiomatic approach to single agent probabilistic reasoning under uncertainty (see [18], [19], [20], [25]), as will be explained in more detail in the next section. In short, the probabilistic knowledge base of an individual agent will be assumed to consist of a set of mathematically well-behaved constraints on the possible probability functions on a fixed finite set of atomic events; such a knowledge base will determine a non-empty closed convex set of probability functions, which represents the set of possible subjective belief functions which the agent may hold on the basis solely of her own knowledge.

We should emphasize the crucial importance in our general problematic of the assumption above that *all* the relevant knowledge of an individual agent is incorporated in the formal representation of her probabilistic knowledge base. This assumption is often referred to as the *Principle of Total Evidence*[1]. As was pointed out forcefully by Jaynes [12] in his work justifying the use of maximum entropy inference, in order to avoid hopeless confusion, it is essential that this assumption be respected in any discussion of the general axiomatic or logical characteristics of a mode of probabilistic inference, since otherwise the nature of the underlying problem can be surreptitiously changed in an arbitrary manner, resulting in the generation of an inexhaustible supply of phony paradoxes or inconsistencies. However when applied to the formalisation of any real life problem considered by a *human* agent, the Principle of Total Evidence is never observed in practice. This banal fact of life has historically bedevilled theoretical discussion of probabilistic inference, because it is often extremely hard to give any real world example to illustrate an abstract principle of probabilistic inference without an opponent being tempted to challenge one's reasoning using implicit or intuitive background information concerning the example, which has *not* been included in its formal representation. In the context of multi-agent probabilistic inference this situation has resulted in a heavy concentration of research on computationally pragmatic approaches to specialised problems of probabilistic inference, and a notable neglect of the study of more abstract axiomatic or foundational frameworks. This neglect appears to the authors to be unfortunate, not least because the foundations of artificial intelligence would seem to demand that the Principle of Total Evidence be taken seriously.

The result of applying a merging process as above will not in general determine a single probability function, but rather a non-empty closed convex set of such functions which is intended to *represent the collective declared knowledge of the agents* as if from the standpoint of an unbiased external observer with no knowledge of her own. The aim is that the merging

[1] This terminology is due to Carnap [4]. It is also sometimes called *Bernoulli's Maxim* after an early formulation of the idea in [3]. In [18], [19], [20] it is also called the *Watts Assumption*.

process should have an *intersubjective* character; by this we mean that if we assume (i) that each agent is reporting exactly the same subjective probabilistic knowledge as the observer would have access to if in the agent's place, and (ii) that the external observer has no knowledge concerning the agents other than the probabilistic knowledge bases which the latter report, then the merging process itself should have an objective character which is justified by certain rational principles. Finally, if for pragmatic purposes a single probability function must be chosen as the "social" belief function of the collective, then this choice can be made at a second stage by applying whatever single-agent inference process is preferred to the merged probabilistic knowledge base.

Much axiomatic analysis has been done previously on the very special case of probabilistic pooling (or aggregation) operators[2], where each individual agent's probabilistic knowledge base determines a unique probability function. Furthermore many different algorithms have been suggested for sub-problems or variants of the far more complex case of probabilistic merging considered in the present work[3]. However few authors have considered the global desiderata which such general probabilistic merging should satisfy, and where such desiderata have been considered, many authors have not clearly distinguished the operation of merging the probabilistic knowledge bases from the goal of choosing a *unique* probability function to represent the merged knowledge. One exception in this respect is Williamson [22] who stresses the philosophical distinction between these two processes, and has sought to adapt[4] to the probabilistic context the norms for propositional merging which were first formulated by Konieczny and Pino Pérez in [14].

We believe that Williamson's distinction above is a useful one. In this paper we formulate a probabilistic adaptation of the Konieczny and Pino Pérez principles. We then investigate in this context the properties of two particular probabilistic merging operators, social entropy, and linear entropy, which are respectively generalisations of the normalised geometric mean and linear pooling operators. Social entropy was defined in [25], and was shown in [26] to bear a natural relationship to the well-known[5] maximum entropy inference process **ME**. On the other hand linear entropy, which is a dual merging operator to social entropy, bears a corresponding natural relationship to $\mathbf{CM}^{\infty}$, the limit centre of mass inference process[6].

[2] See e.g. [10] for a survey.
[3] See e.g. [13],[17],[21],[23],[24],[25].
[4] See also [23],[24].
[5] See [18] or [20] for a detailed characterisation of **ME**.
[6] See [18] for a definition of $\mathbf{CM}^{\infty}$.

Konieczny and Pino Pérez in [14] proposed an axiomatic framework, referred to below as KPP, for expressing the desiderata required of a merging operator in a *non-probabilistic* context. Such an operator Δ acts on a sequence of knowledge bases $T_1,\ldots,T_n$ to generate a single knowledge base. The resulting *merged* knowledge base $\Delta(T_1,\ldots,T_n)$ should be consistent, and the operator Δ should satisfy certain general principles. In [14] 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 there, 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[7].

According to the KPP framework described above a minimal set of desiderata which a merging operator Δ should satisfy is embodied in the following six principles:

For every $n,m \geq 1$ and every propositional language L and knowledge bases $K_1,\ldots,K_n, F_1,\ldots,F_m$ for L:

(A1) $\Delta(K_1,\ldots,K_n)$ is a (consistent) knowledge base,

(A2) If $K_1,\ldots,K_n$ are jointly consistent then $\Delta(K_1,\ldots,K_n)$ is logically equivalent to $\bigcup_{i=1}^{n} K_i$,

(A3) If $K_1,\ldots,K_n$ and $F_1,\ldots,F_n$ are such that there exist a permutation π of the index set $\{1,\ldots,n\}$ such that K_i is logically equivalent to $F_{\pi(i)}$ for $1 \leq i \leq n$, then $\Delta(K_1,\ldots,K_n)$ is logically equivalent to $\Delta(F_1,\ldots,F_n)$,

(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

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

In the next section we will reformulate the ideas behind the KPP principles above in order to apply them to the different context of the merging of probabilistic knowledge bases, or more explicitly, to the search for a rationally justified method of merging probabilistic knowledge from distinct sources as described above.

Before continuing our discussion we will first formulate the prerequisite concepts which we need in order to define precisely the general notion of a probabilistic merging operator.

[7] See [15] for a survey paper and bibliography.

2. From Non-Probabilistic to Probabilistic Merging

Let $L = \{p_1 \dots p_h\}$ be a finite propositional language where $p_1, \dots, p_h$ are propositional variables. 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 $\wedge_{i=1}^{h} \pm p_i$, and $\pm p_i$ denotes either p_i and $\neg p_i$. We denote an enumeration these 2^h 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 $\mathrm{At}(L)$. The atoms of $\mathrm{At}(L)$ are thus mutually exclusive and exhaustive.

A *probability function* $\mathbf{w}$ over L is defined as 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 $\varphi \in SL$ may then be defined by setting

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

Whenever a sentence $\varphi \in SL$ is not satisfiable we set $\mathbf{w}(\varphi) = 0$. 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 makes sense only for atoms. Given a probability function $\mathbf{w} \in \mathbb{D}^L$, a conditional probability is defined by Bayes's formula

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

for any $\varphi, \psi \in SL$ such that $\mathbf{w}(\psi) \neq 0$ and is left undefined otherwise.

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 $\mathbb{D}^L$. We shall abbreviate the term probabilistic knowledge base by *p–knowledge base*. $V_{\mathbf{K}}^L$ may be thought of as the set of possible probability functions in $\mathbb{D}^L$ of a particular agent which are consistent with her p–knowledge base $\mathbf{K}$. We shall generally write $V_{\mathbf{K}}$ instead of $V_{\mathbf{K}}^L$ unless there is any ambiguity about which language is referred to. Note that this standard formulation ensures that linear constraint conditions such as $\mathbf{w}(\theta) = a$, $\mathbf{w}(\phi \mid \psi) = b$, and $\mathbf{w}(\psi \mid \theta) \leq c$, where $a, b, c \in [0,1]$ and $\theta, \phi, \psi \in SL$ are satisfiable L-sentences, are all permissible in a p–knowledge base $\mathbf{K}$ provided that the resulting constraint set $\mathbf{K}$ is consistent with the laws of probability. Note that a constraint such as $\mathbf{w}(\psi \mid \theta) \leq c$ is interpreted as $\mathbf{w}(\psi \wedge \theta) \leq c \cdot \mathbf{w}(\theta)$ which makes sense as a linear constraint even though $\mathbf{w}(\theta)$ may take the value zero (see [18] for details).

If $\mathbf{K}_1$ and $\mathbf{K}_2$ are such that $V_{\mathbf{K}_1} = V_{\mathbf{K}_2}$ we shall say that $\mathbf{K}_1$ and $\mathbf{K}_2$ are *equivalent*. In practice we shall only be interested in constraint sets up to

equivalence, and consequently we will informally identify a p–knowledge base $\mathbf{K}$ with its extension $V_{\mathbf{K}}$, and with slight abuse of language we may also refer to a non-empty closed subset of $\mathbb{D}^L$ as a p–knowledge base. Note that the non-emptiness of $V_{\mathbf{K}}$ corresponds to the assumption that $\mathbf{K}$ is consistent with the laws of probability, while if $\mathbf{K}$ and $\mathbf{F}$ are p–knowledge bases then the set of constraints $\mathbf{K} \cup \mathbf{F}$ corresponds to $V_{\mathbf{K}\cup\mathbf{F}} = V_{\mathbf{K}} \cap V_{\mathbf{F}}$, and so forms a p–knowledge base provided that the latter intersection is non-empty.

The set of all p–knowledge bases $V_{\mathbf{K}}$ over L is denoted by CL. A more restricted notion of p–knowledge base is a p–knowledge base *which bounds probability functions away from zero*. This is a p–knowledge base $\mathbf{K} \in CL$ such that $V_{\mathbf{K}}$ satisfies a set of constraints on $\mathbf{w} \in V_{\mathbf{K}}$ of the form

$$\{a_j \leq w_j : 1 \leq j \leq J\}$$

where $0 < a_j < 1$ for all $j = 1 \ldots J$. We call such a p–knowledge base *bounded*, and we will denote the set of all bounded p–knowledge bases for a given language L by BCL. A somewhat more general notion is that of a p–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 $\mathbf{w} \in V_{\mathbf{K}}$ such that $w_j \neq 0$. The set of weakly bounded p–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 $\mathbf{w} \in V_{\mathbf{K}}$ such that $w_j \neq 0$ for all $j = 1 \ldots J$.

There are at several possible motivations for studying p–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 arise from treating zero probabilities in certain contexts. In this paper we will confine ourselves to stating and proving some theorems concerning particular merging operators for certain classes of p–knowledge bases, but will not consider further the epistemological status of the various notions of p–knowledge base.

Let Δ 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 $\mathbb{D}^L$. We will call such a Δ 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 $\Delta_L(\mathbf{K}_1, \ldots, \mathbf{K}_n) \in CL$.

Note that **(K1)** is a natural counterpart to **(A1)**; just as **(A1)** ensures that a propositional merging operator applied to a sequence of knowledge bases

yields a knowledge base, so **(K1)** ensures that a p–merging operator applied to a sequence of p–knowledge bases yields a p–knowledge base.

In general we shall suppress the subscript L in Δ_L except where an ambiguity could be caused by such an omission. We may sometimes slightly abuse the above terminology by referring to an operator Δ as a p–merging operator even though the domain over which Δ is properly defined may be a certain subclass of the $\underbrace{CL \times \ldots \times CL}_{n}$. Whenever we do this however the correct restriction of the domain of application will always be made apparent.

We now set about reformulating the remaining KPP principles so as to make them applicable to the context of a p–merging operator Δ. We express the remaining five principles as follows:

For every $n \geq 1$ and every propositional language L

(K2) Consistency Principle. For all $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ if $\bigcap_{i=1}^{n} V_{\mathbf{K}_i}^{L} \neq \varnothing$ then $\Delta(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \bigcap_{i=1}^{n} V_{\mathbf{K}_i}^{L}$.

(K2) can be interpreted as saying that if the p–knowledge bases of a set of agents are collectively consistent then the merged p–knowledge base should simply consist of all the knowledge of the agents collected together. If there is only one agent, with p–knowledge base $\mathbf{K}$, the principle just asserts that $\Delta(\mathbf{K}) = V_{\mathbf{K}}^{L}$.

(K3) 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 π of the index set $\{1, \ldots, n\}$ such that $V_{\mathbf{K}_i}^{L} = V_{\mathbf{F}_{\pi(i)}}^{L}$ for $1 \leq i \leq n$, then $\Delta(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \Delta(\mathbf{F}_1, \ldots, \mathbf{F}_n)$.

Notice that **(K3)** has the effect that for any Δ which satisfies it, the order in which the p–knowledge bases occur when Δ is applied is immaterial, and therefore we can loosely refer to Δ as being applied to a multiset of p–knowledge bases instead of a sequence of such p–knowledge bases. On the other hand repetitions of p–knowledge bases *will* in general be significant, so the sequence (or multiset) of p–knowledge bases cannot be considered simply as a set; the Δ we consider behave somewhat analogously to the *majority* merging operators of the KPP framework [14] in the sense that adding further agents whose p–knowledge bases are copies of the p–knowledge base of some particular agent generally has the effect of increasing the influence of that p–knowledge base on the resulting merged p–knowledge base.

(K4) Disagreement Principle. Let $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ and $\mathbf{F}_1, \ldots, \mathbf{F}_m \in CL$. Assume that $\bigcap_{i=1}^{m} V_{\mathbf{F}_i}^{L} \neq \varnothing$.

Then $\Delta(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \Delta(\mathbf{F}_1, \ldots, \mathbf{F}_m) = \varnothing$ implies that

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

(K4) represents a significant but natural strengthening of **(A4)**, adapted to the p–merging context. Intuitively the principle says that if the merged p–knowledge base **K** of a set of agents is inconsistent with the merged p–knowledge **F** of a distinct set of agents, where the p–knowledge bases of the latter set are collectively consistent, then the result of merging the p–knowledge bases of all the agents together is also inconsistent with **K**. Expressed more pithily, if less exactly, we could say that a consistent group who disagree with another group and then merge with them can be sure that they have influenced the opinions of the combined group.

(K5) Agreement Principle. If $\Delta(\mathbf{K}_1,\ldots,\mathbf{K}_n) \cap \Delta(\mathbf{F}_1,\ldots,\mathbf{F}_m) \not\sim \varnothing$ then

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

(K5) combines the ideas of **(A5)** and **(A6)** into a single principle adapted to the probabilistic context. In particular **(K5)** implies that if each of two distinct sets of agents arrive at the same set of possible conclusions then the result of considering the p–knowledge bases of all the agents together should result in the same set of possible conclusions.

The intuitive idea behind p–merging is that the probabilistic knowledge from a set of agents should be shared by some objective *collaborative* process, which takes full account of the declared p–knowledge base of each agent, including the *implicit ignorance* of an agent whenever she has not specified a singleton probability function as constituting her p–knowledge base. The result of this process should be a new "social" or merged p–knowledge base, which represents the collective knowledge of the set of agents, just as if the set had merged to form a single agent. It is clear that if **(K2)** is to be satisfied then this merged p–knowledge base will not in general be a singleton.

This general intersubjective approach to probabilistic merging was expounded in a slightly different form by the second author in [25], and accords well with certain philosophical ideas elaborated independently by Williamson [22], [23]. Both stress the advantages of initially merging the p–knowledge bases of a set of agents into a single p–knowledge base, as opposed to merging the *default* belief functions of the individual agents into a single probability function, where by the default belief function of an agent we mean the unique probability function which that agent may hypothetically arrive at solely by considering her own p–knowledge base and applying to it a standard inference process[8] such as the maximum entropy inference process **ME**.

[8] See e.g. [18] for a comprehensive account of single agent inference processes, including **ME**.

Our reformulation of the KPP principles into a probabilistic framework is a fairly straightforward translation with the exception, as noted above, of **(K4)**. In the sequel we will show that the hitherto known p–merging operators which satisfy perhaps the most attractive desiderata other than **(K1)-(K5)** do in fact satisfy the above principles **(K1)**, **(K2)**, **(K3)**, and satisfy **(K4)** and **(K5)** at least when their application is restricted to bounded p–knowledge bases[9]. In particular, in [25] and [26] a specific p–merging operator is defined, which we will here call the *social entropy operator*, denoted by Δ^{KL}, which is strongly related to Kullback-Leibler divergence. This was introduced as the first stage of a merging process called the *social entropy process*, **SEP**, which for any multiset of p–knowledge bases chooses a *unique* merged probability function. The first stage of **SEP** consists of applying Δ^{KL}, while the second stage of **SEP** simply chooses the unique maximum entropy point in the resulting merged p–knowledge base.

In the following section we will examine in detail some of the properties of the operator Δ^{KL} together with those of its dual, the *linear entropy operator* $\hat{\Delta}^{\mathrm{KL}}$.

3. Two Probabilistic Merging Operators

3.1. The Social Entropy Operator Δ^{KL}

In order to define the social entropy operator we first need to define Kullback-Leibler divergence $\mathrm{KL}\colon \mathbb{D}^L \times \mathbb{D}^L \to [0,+\infty]$. This may be thought of as a function which measures the (asymmetric) "informational distance" from one probability function to another, and returns a value in the interval $[0,+\infty]$. The asymmetry of this notion is the reason for the use of the term "divergence" rather than "distance". The Kullback-Leibler divergence from $\mathbf{w} \in \mathbb{D}^L$ to $\mathbf{v} \in \mathbb{D}^L$ is defined as $+\infty$ whenever $v_j \not\approx 0$ and $w_j = 0$ for some atom α_j. If this is not the case we say that $\mathbf{w}$ *dominates* $\mathbf{v}$ and write $\mathbf{w} \gg \mathbf{v}$. Let $\mathrm{Sig}(\mathbf{w}) = \{j : w_j \not\approx 0\}$. Then the Kullback-Leibler divergence from $\mathbf{w}$ to $\mathbf{v}$ is defined by

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

[9] We remark here that whereas, as noted above, Williamson previously advocated the relevance of the KPP principles in relation to probabilistic merging, in a recent paper [24] he criticises the KPP principles **(A2)**, **(A4)**, and **(A6)** as representing norms which are too strong to be applicable in this context. However in reaching this conclusion it appears that he is using an informal notion of *evidence base* which is very different from our restricted, but more formally defined, notion of p–knowledge base, and we do not feel that his criticisms are justified if applied to our framework.

with the usual convention that $x \log x$ is defined to take the value zero at $x = 0$. It is easy to show that $\mathrm{KL}(\mathbf{v}\|\mathbf{w})$ is always non-negative, that $\mathrm{KL}(\mathbf{v}\|\mathbf{w}) = 0$ if and only if $\mathbf{v} = \mathbf{w}$, and that $\mathrm{KL}(\mathbf{v}\|\mathbf{w})$ is finite if and only if $\mathbf{w} \gg \mathbf{v}$. (See e.g. [18].)

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

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

It is easy to see that this is well-defined (see [26]). Note that this value lies in the interval $[0, +\infty]$. Also $C_{\mathbf{K}_1,\ldots,\mathbf{K}_n} = 0$ 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 $C_{\mathbf{K}_1,\ldots,\mathbf{K}_n}$ is finite if and only if the following holds:

There is some atom α_j such that for no i is it the case that for all $\mathbf{w} \in V^L_{\mathbf{K}_i}$ $\mathbf{w}(\alpha_j) = 0$. (1)

The p–merging operator Δ^{KL} is now defined as follows: for any L and any $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ $\Delta^{\mathrm{KL}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is defined as

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

In [26] it is shown that for any $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ this set $\Delta^{KL}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is always a non-empty closed convex region of $\mathbb{D}^L$, and hence it follows that the p–merging operator Δ^{KL} satisfies **(K1)**. We note however that although Δ^{KL} is everywhere defined[10] 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 (3.1) above is satisfied, since otherwise applying Δ^{KL} simply returns the whole space $\mathbb{D}^L$. The fact that the social entropy operator Δ^{KL} satisfies **(K2)** follows at once from the fact noted above that $C_{\mathbf{K}_1,\ldots,\mathbf{K}_n} = 0$ if and only if $\mathbf{v} = \mathbf{w}^{(1)} = \ldots = \mathbf{w}^{(n)}$ in the definition of $C_{\mathbf{K}_1,\ldots,\mathbf{K}_n}$. Moreover Δ^{KL} satisfies **(K3)** trivially by definition.

Δ^{KL} turns out to have many other desirable properties, some of which closely resemble the axiomatic properties which have been used to characterise the **ME** inference process in [20], and [18]. (See [1], [25], [26] for details.) In particular we mention the following:

[10] In the presentation in [26] the region $\Delta^{KL}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is only defined assuming that condition (1) holds, but this does not significantly affect the results.

1. Language Invariance

Suppose $L \subset L'$ and $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$. $\mathbf{K}_1, \ldots, \mathbf{K}_n$ may also be regarded as p–knowledge bases in CL'. For any $\mathbf{w}' \in \mathbb{D}^{L'}$ denote by $\mathbf{w}' \restriction L$ the marginalisation of $\mathbf{w}'$ to $\mathbb{D}^L$. Then

$$\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \{\mathbf{w}' \restriction L : \mathbf{w}' \in \Delta_{L'}^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)\}$$

Language Invariance means that if we change a multiset of p–knowledge bases only by adding propositional variables to the language in which they are formulated but add no new knowledge, then the restriction of the new merged p–knowledge base to the original language is the same as the original merged p–knowledge base. The fact that Δ^{KL} satisfies language invariance is proved in [1].

2. The Consistent Irrelevant Information Principle.

Let $L = L_1 \cup L_2$ where L_1 and L_2 are disjoint propositional languages. Let $\mathbf{K}_1, \ldots, \mathbf{K}_n$ and $\mathbf{F}_1, \ldots, \mathbf{F}_n$ be knowledge bases formulated for the languages L_1 and L_2 respectively, and suppose that $\mathbf{F}_1, \ldots, \mathbf{F}_n$ are jointly consistent. Then

$$\Delta_L^{\mathrm{KL}}(\mathbf{K}_1 \cup \mathbf{F}_1, \ldots, \mathbf{K}_n \cup \mathbf{F}_n) \restriction L_1 = \Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n) \restriction L_1.$$

where $T \restriction L_1$ denotes the set of marginalisations to L_1 of probability functions over L belonging to a given set T of such probability functions. The above property of Δ^{KL} follows from Lemma 5.2 of [1]. Together with language invariance, it ensures that if a set of agents have p–knowledge bases formulated in the language L_1 then their merged p–knowledge base remains the same if each agent acquires additional new knowledge formulated in a disjoint language L_2 and the newly merged p–knowledge base of all the agents is then restricted to the language L_1, *provided that* all the new knowledge in the language L_2 is jointly consistent.

3. Δ^{KL} Generalises the LogOp Pooling Operator

In [25] the following equivalence between (i) and (ii) below is given, which provides an alternative characterisation of Δ^{KL} in the case when condition (1) above is satified:

(i) The L-probability functions $\mathbf{v}$, $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ minimize

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

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

(ii) The L-probability functions $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ maximize

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

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

$$v_j = \frac{\left(\prod_{k=1}^{n} w_j^{(k)}\right)^{\frac{1}{n}}}{\sum_{j=1}^{J} \left(\prod_{k=1}^{n} w_j^{(k)}\right)^{\frac{1}{n}}} \text{ for all } j = 1 \ldots J. \tag{2}$$

Whenever (2) holds we write $\mathbf{v} = \mathbf{LogOp}\,(\mathbf{w}^{(1)} \ldots \mathbf{w}^{(n)})$. **LopOp** is of course just the normalised geometric mean, or "logarithmic", pooling operator familiar to decision theorists. Thus we see that for $\mathbf{v}$ to be in Δ^{KL} there must exist some $\mathbf{w}^{(i)} \in V_{\mathbf{K}_i}$ which maximise the normalising factor in the definition of logarithmic pooling, and for which $\mathbf{v} = \mathbf{LogOp}\,(\mathbf{w}^{(1)} \ldots \mathbf{w}^{(n)})$. In the very special case when each each agent i specifies a single probability function $\mathbf{w}^{(i)}$ then $\Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is just the singleton $\{\mathbf{LogOp}\,(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})\}$. Notice that condition (1) is exactly the condition required to ensure that the **LopOp** pooling operator is defined.

4. Δ^{KL} is a Natural Companion to the ME Inference Process

At first sight this assertion might seem strange, since if Δ^{KL} is applied to the p–knowledge base $\mathbf{K}$ of a single agent X it simply returns the same p–knowledge base in the form $V_{\mathbf{K}}$, which does not help X to choose a single preferred point in $V_{\mathbf{K}}$. However let us imagine that X now appoints a fanatically unbiased oracle Y with p–knowledge base $\mathbf{F} = \{(\frac{1}{J}, \frac{1}{J} \ldots \frac{1}{J})\}$, in order to help her to choose a preferred point in her p–knowledge base. Y advises X to imagine cloning herself n times, for some large n, and forming a committee of $n+1$ members consisting of the n clones of X, together with Y as chairman. Finally Y advises X to compute the result of applying Δ^{KL} to the $n+1$ p–knowledge bases of the members of A_n and then to let $n \to \infty$. The result of this procedure is that the merged p–knowledge bases converge towards a single point, the maximum entropy point of $V_{\mathbf{K}}$. (See [26] for a proof[11]).

The following theorem is our first main result of the present work.

Theorem 3.1. *The p–merging operator* Δ^{KL} *satisfies the principles* **(K1)**, **(K2)** *and* **(K3)**. *Furthermore* Δ^{KL} *satisfies* **(K4)** *and* **(K5)** *provided that the p–knowledge bases to which* Δ^{KL} *is applied are restricted to WBCL.*

[11] In [26] a similar more general result is proved which holds for any number of agents.

The fact that **(K1)**, **(K2)** and **(K3)** hold for Δ^{KL} has been established above. The rest of the theorem will be proved in section 4.

□

3.2. The Linear Entropy Operator $\hat{\Delta}^{\mathrm{KL}}$

The Linear Entropy operator $\hat{\Delta}^{\mathrm{KL}}$ is a p–merging operator which may naturally be considered as the dual of the Δ^{KL} p–merging operator defined above.

In brief, whereas $\Delta^{\mathrm{KL}}(\mathbf{K}_1,\ldots,\mathbf{K}_n)$ comprises those $\mathbf{v}$ which globally minimise

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

$\hat{\Delta}^{\mathrm{KL}}(\mathbf{K}_1,\ldots,\mathbf{K}_n)$ comprises those $\mathbf{v}$ which globally minimise

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

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

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

As in 3.1 it is easy to see that this is well-defined, non-negative, and zero if and only if $\mathbf{v}=\mathbf{w}^{(1)}=\ldots=\mathbf{w}^{(n)}$ in the definition of $\hat{C}_{\mathbf{K}_1,\ldots,\mathbf{K}_n}$. However unlike the case for Δ^{KL} we may note that $\hat{C}_{\mathbf{K}_1,\ldots,\mathbf{K}_n}$ is always finite since any $\mathbf{v}$ all of whose coordinates are non-zero will always give a finite non-zero value to $\sum_{i=1}^{n}\mathrm{KL}(\mathbf{w}^{(i)}\|\,\mathbf{v})$.

The p–merging operator $\hat{\Delta}^{\mathrm{KL}}$ is now defined as follows: for any L and any $\mathbf{K}_1,\ldots,\mathbf{K}_n\in CL$ $\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1,\ldots,\mathbf{K}_n)$ is defined as

$$\{\mathbf{v}\in\mathbb{D}^L : \exists\mathbf{w}^{(1)}\in V^L_{\mathbf{K}_1},\ldots,\mathbf{w}^{(n)}\in V^L_{\mathbf{K}_n}\ \text{s.t.}\sum_{i=1}^{n}\mathrm{KL}(\mathbf{w}^{(i)}\|\,\mathbf{v}) = \hat{C}_{\mathbf{K}_1,\ldots,\mathbf{K}_n}\}.$$

It is easy to show (cf. section 4) that whenever

$$\sum_{i=1}^{n}\mathrm{KL}(\mathbf{w}^{(i)}\|\,\mathbf{v}) = \hat{C}_{\mathbf{K}_1,\ldots,\mathbf{K}_n}$$

then $\mathbf{v}=\mathbf{LinOp}(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)})$ where $\mathbf{LinOp}(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)})$ just returns the arithmetic mean of $\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)}$. Hence $\hat{\Delta}^{\mathrm{KL}}$ is a generalisation of the

arithmetic pooling operator **LinOp**, and indeed coincides with that operator in the special case when each $\mathbf{K}_i$ specifies a unique probability function.

It is straightforward to prove that $\sum_{j\in \mathrm{Sig}(\mathbf{y})} x_j \log x_j y_j$ is a convex function over the domain $\{(\mathbf{x},\mathbf{y}) \in \mathbb{D}^L \times \mathbb{D}^L : \mathbf{y} \gg \mathbf{x}\}$. It follows that the set $\hat{\Delta}^{\mathrm{KL}}_L(\mathbf{K}_1,\ldots,\mathbf{K}_n)$ is nonempty, closed and convex for all $\mathbf{K}_1,\ldots,\mathbf{K}_n \in CL$ and hence that the p–merging operator $\hat{\Delta}^{\mathrm{KL}}$ satisfies **(K1)**. As in the case of Δ^{KL}, the fact that the operator $\hat{\Delta}^{\mathrm{KL}}$ satisfies **(K2)** follows at once from the remark above that $\hat{C}_{\mathbf{K}_1,\ldots,\mathbf{K}_n} = 0$ if and only if $\mathbf{v} = \mathbf{w}^{(1)} = \ldots = \mathbf{w}^{(n)}$ in the definition of $\hat{C}_{\mathbf{K}_1,\ldots,\mathbf{K}_n}$. Similarly $\hat{\Delta}^{\mathrm{KL}}$ satisfies **(K3)** trivially by definition.

It can also be shown that, as in the case of Δ^{KL}, Language Invariance and the Consistent Irrelevant Information Principle of section 3.1 also hold for the p–merging operator $\hat{\Delta}^{\mathrm{KL}}$. Finally if the "chairman" procedure of section 3.1, which related Δ^{KL} to **ME** is instead applied using $\hat{\Delta}^{\mathrm{KL}}$ then the point chosen in $V_{\mathbf{K}}$ is not the maximum entropy point, but the CM^{∞} point, or limit centre of mass point, of $V_{\mathbf{K}}$. These last results have been proved in [0].

Before stating our second theorem of this article we introduce the following natural strengthening of the Disagreement Principle **(K4)**.

(K4*) Strong Disagreement Principle.

Let $\mathbf{K}_1,\ldots,\mathbf{K}_n \in CL$ and $\mathbf{F}_1,\ldots,\mathbf{F}_m \in CL$.
Then $\Delta(\mathbf{K}_1,\ldots,\mathbf{K}_n) \cap \Delta(\mathbf{F}_1,\ldots,\mathbf{F}_m) = \varnothing$ implies that

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

Trivially the Strong Disagreement Principle implies the Disagreement Principle.

Theorem 3.2. *The p–merging operator* $\hat{\Delta}^{\mathrm{KL}}$ *satisfies* **(K1)**, **(K2)**, **(K3)** *and* **(K5)**. *Furthermore if the p–knowledge bases to which* $\hat{\Delta}^{\mathrm{KL}}$ *is applied are restricted to BCL, then* $\hat{\Delta}^{\mathrm{KL}}$ *satisfies the Strong Principle of Disagreement* **(K4*)**.

The fact that **(K1)**, **(K2)** and **(K3)** hold for $\hat{\Delta}^{\mathrm{KL}}$ has been established above. The proofs for **(K4*)** and **(K5)** will be given in the next section.

□

Historical Remarks.

Minimising Kullback-Leibler divergence from a convex set to a given probability function, or KL-projection, has long been used for updating and in machine learning algorithms (see e.g. [2], [5], [7], [8], [11] and [18]). Connections between the minimisation of sums of Kullback-Leibler divergences and the operators **LinOp** and **LogOp** have also been noted previously by several authors within somewhat different frameworks. In particular we should

mention the work of Matúš [16] who proved a number of convergence theorems covering the iteration of alternating operations of KL–projection or its dual to several compact convex sets followed by **LinOp** or, respectively, **LogOp**, and showing that under certain conditions these iterations converge to fixed points. These fixed points correspond respectively to particular points of $\hat{\Delta}^{\mathrm{KL}}$ or Δ^{KL}.

4. Proofs of Results

In this section we prove the two main results of this paper – the theorems 3.1 and 3.2. Since the properties **(K1)**, **(K2)** and **(K3)** have already been established for the two p–merging processes it remains to deal with the agreement and disagreement principles. The proofs of the Agreement Principle **(K5)** are straightforward and are given in 4.5 below. However the proofs for the Disagreement Principle (**(K4)** or **(K4*)**) are more complex and are different in flavour for Δ^{KL} and for $\hat{\Delta}^{\mathrm{KL}}$. The result for $\hat{\Delta}^{\mathrm{KL}}$ is proved in 4.6 and that for Δ^{KL} in 4.8.

We start by reviewing some geometrical properties of the space of probability functions $\mathbb{D}^L$ with respect to the divergence KL. First of all notice that for given $\mathbf{v} \in \mathbb{D}^L$ the Kullback-Leibler divergence $\mathrm{KL}(\mathbf{w} \| \mathbf{v})$ is a strictly convex function in the first argument over the domain specified by $\mathbf{v} \gg \mathbf{w}$. Owing to this if $\mathbf{v} \in \mathbb{D}^L$ is given and $W \subseteq \mathbb{D}^L$ is a closed convex set such that there is at least one probability function in W which $\mathbf{v}$ dominates, then we can define the KL-*projection* of $\mathbf{v}$ to W. This is defined as that unique point $\mathbf{w} \in W$ which minimizes $\mathrm{KL}(\mathbf{w} \| \mathbf{v})$. For more details see [2].

The following theorem is due to Csiszár [7].

Theorem 4.1. (Extended Pythagorean Theorem) *Let* $\mathbf{w}$ *be the* KL*-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* $\mathbf{v} \gg \mathbf{w} \gg \mathbf{a}$. *Then*

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

□

The following theorem is well known in information theory, see for instance [6].

Theorem 4.2. (Parallelogram Theorem). *Let* $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}, \mathbf{v} \in \mathbb{D}^L$ *be such that* $\mathbf{v} \gg \mathbf{w}^{(i)}$ *for all* $1 \leq i \leq n$. *Then*

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

□

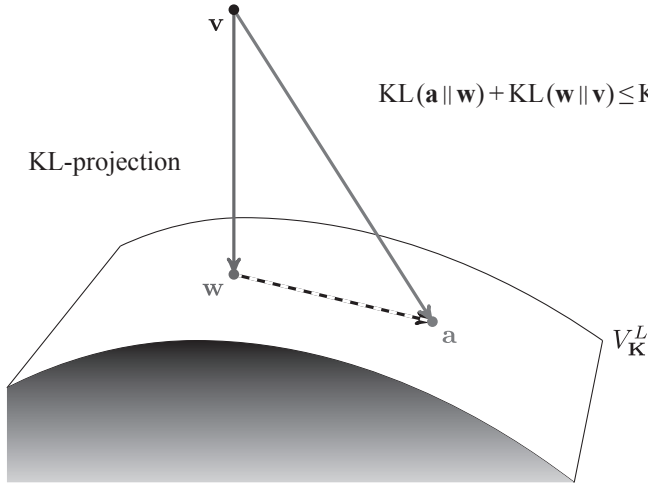

Figure 1. The illustration of the extended Pythagorean theorem.

Lemma 4.3. *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*

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

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* $\mathbf{u}^{(i)} \gg \mathbf{a}^{(i)}$ *for all* $1 \leq i \leq n$. *Then*

$$\sum_{i=1}^{n} \sum_{j \in \mathrm{Sig}(\mathbf{v})} (a_j^{(i)} - u_j^{(i)}) \cdot (\log u_j^{(i)} - \log v_j) < 0.$$

Proof. First of all notice that by the assumption $\mathbf{u}^{(i)} \gg \mathbf{a}^{(i)}$ for all $1 \leq i \leq n$ we have that

$$\mathrm{KL}(\mathbf{a}^{(i)} \| \mathbf{v}) - \mathrm{KL}(\mathbf{u}^{(i)} \| \mathbf{v}) - \mathrm{KL}(\mathbf{a}^{(i)} \| \mathbf{u}^{(i)}) = \sum_{j \in \mathrm{Sig}(\mathbf{v})} (a_j^{(i)} - u_j^{(i)}) \cdot (\log u_j^{(i)} - \log v_j). \tag{3}$$

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

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

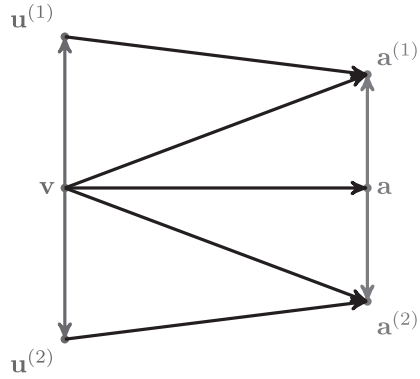

Figure 2. The illustration of lemma 4.3 for $n = 2$.
Arrows indicate corresponding Kullback-Leibler divergences.

Hence

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

$$-\sum_{i=1}^{n}\mathrm{KL}(\mathbf{a}^{(i)}\|\,\mathbf{u}^{(i)})=\sum_{i=1}^{n}\sum_{j\in\mathrm{Sig}(\mathbf{v})}(a_j^{(i)}-u_j^{(i)})\cdot(\log u_j^{(i)}-\log v_j). \tag{4}$$

Since $\mathrm{KL}(\mathbf{w}\,\|\,\mathbf{v})$ is a convex function in both arguments whenever $\mathbf{v}\gg\mathbf{w}$, by the Jensen inequality

$$n\cdot\mathrm{KL}(\mathbf{a}\,\|\,\mathbf{v})-\sum_{i=1}^{n}\mathrm{KL}(\mathbf{a}^{(i)}\|\,\mathbf{u}^{(i)})\le 0. \tag{5}$$

The inequality (5) together with the assumption that

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

gives that left-hand side of the equality (4) is negative and so the right-hand side is too, whence

$$\sum_{i=1}^{n}\sum_{j\in\mathrm{Sig}(\mathbf{v})}(a_j^{(i)}-u_j^{(i)})\cdot(\log u_j^{(i)}-\log v_j)<0$$

as required. □

Lemma 4.4. *Let* $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)} \in \mathbb{D}^L$ *be fixed. Then*

(i) $\sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{v})$ *is strictly minimal for* $\mathbf{v} = \mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$.

(ii) $\sum_{i=1}^{n} \mathrm{KL}(\mathbf{v} \| \mathbf{w}^{(i)})$ *is strictly minimal for* $\mathbf{v} = \mathbf{LogOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$ *provided that* $\mathbf{LogOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$ *is defined, i.e. provided that for some j for all i* $\mathbf{w}_j^{(i)} \neq 0$.

Proof. (i) By the parallelogram theorem the minimality of $\sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{v})$ for fixed $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ is equivalent to the minimality of $\mathrm{KL}(\mathbf{LinOp}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \| \mathbf{v})$. Since for any fixed $\mathbf{w}$ $\mathrm{KL}(\mathbf{w} \| \mathbf{v})$ is strictly minimal when $\mathbf{v} = \mathbf{w}$, the first part of the lemma follows.

(ii) The proof is straightforward, see e.g. [25].

□

We will denote by $\hat{\Gamma}_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ the set of all n-tuples $\mathbf{w}^{(1)} \in V_{\mathbf{K}_1}^L, \ldots, \mathbf{w}^{(n)} \in V_{\mathbf{K}_n}^L$ such that for some $\mathbf{v} \in \Delta_L^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n)$

$$\sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{v}) = \hat{C}_{\mathbf{K}_1, \ldots, \mathbf{K}_n}.$$

This notation will be useful in the following two proofs.

Theorem 4.5. *(i) The* $\hat{\Delta}^{\mathrm{KL}}$ *p–merging operator satisfies* **(K5)**.

(ii) The Δ^{KL} *p–merging operator satisfies* **(K5)** *for all p–knowledge bases in WBCL.*

Proof. The proofs are very similar in both cases, so we shall just give the proof for $\hat{\Delta}^{\mathrm{KL}}$ below.

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

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

Then since by definition $\hat{C}_{\mathbf{K}_1, \ldots, \mathbf{K}_n} + \hat{C}_{\mathbf{F}_1, \ldots, \mathbf{F}_m} \leq \hat{C}_{\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m}$ the same vectors $\mathbf{v}$, $\mathbf{w}^{(1)} \ldots \mathbf{w}^{(n)}$, $\mathbf{u}^{(1)} \ldots \mathbf{u}^{(m)}$ globally minimize the sum

$$\sum_{i=1}^{n} \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{v}) + \sum_{i=1}^{m} \mathrm{KL}(\mathbf{u}^{(i)} \| \mathbf{v}) \tag{6}$$

subject to $\mathbf{w}^{(i)} \in V_{\mathbf{K}_i}^L$, $1 \leq i \leq n$ and $\mathbf{u}^{(i)} \in V_{\mathbf{F}_i}^L$, $1 \leq i \leq m$.

Thus $\mathbf{v} \in \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1,\ldots,\mathbf{K}_n, \mathbf{F}_1,\ldots,\mathbf{F}_m)$, and

$$\hat{C}_{\mathbf{K}_1,\ldots,\mathbf{K}_n} + \hat{C}_{\mathbf{F}_1,\ldots,\mathbf{F}_m} = \hat{C}_{\mathbf{K}_1,\ldots,\mathbf{K}_n,\,\mathbf{F}_1,\ldots,\mathbf{F}_m}. \tag{7}$$

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

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

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

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

In view of (7) if we did not now have that $\sum_{i=1}^{n} \mathrm{KL}(\mathbf{y}^{(i)} \| \mathbf{x}) = \hat{C}_{\mathbf{K}_1,\ldots,\mathbf{K}_n}$ and $\sum_{i=1}^{m} \mathrm{KL}(\mathbf{z}^{(i)} \| \mathbf{x}) = \hat{C}_{\mathbf{F}_1,\ldots,\mathbf{F}_m}$ then this would contradict the minimality of either $\hat{C}_{\mathbf{K}_1,\ldots,\mathbf{K}_n}$ or $\hat{C}_{\mathbf{F}_1,\ldots,\mathbf{F}_m}$. Hence $\mathbf{x} \in \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1,\ldots,\mathbf{K}_n) \cap \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{F}_1,\ldots,\mathbf{F}_m)$ and the result is proved.

The proof for Δ^{KL} is similar except that the final argument involving the equation corresponding to (7) fails if either of the quantities $C_{\mathbf{K}_1,\ldots,\mathbf{K}_n}$ or $C_{\mathbf{F}_1,\ldots,\mathbf{F}_m}$ is $+\infty$, which is the reason for the restriction of p–knowledge bases to *WBCL*, since with that restriction these quantities are necessarily finite. □

The following theorem proves that the $\hat{\Delta}^{\mathrm{KL}}$ p–merging operator satisfies the Strong Disagreement Principle **(K4*)** if the p–knowledge bases are restricted to *BCL*. This together with theorem 4.5 above and the results of section 3.2 is sufficient to establish our theorem 3.2.

Theorem 4.6. *Let* $\mathbf{K}_1,\ldots,\mathbf{K}_n \in CL$ *and* $\mathbf{F}_1,\ldots,\mathbf{F}_m \in CL$ *be such that for every*

$$(\mathbf{v}^{(1)},\ldots,\mathbf{v}^{(n)}, \mathbf{u}^{(1)},\ldots,\mathbf{u}^{(m)}) \in \hat{\Gamma}_L^{\mathrm{KL}}(\mathbf{K}_1,\ldots,\mathbf{K}_n, \mathbf{F}_1,\ldots,\mathbf{F}_m)$$

there is $(\mathbf{a}^{(1)},\ldots,\mathbf{a}^{(m)}) \in \hat{\Gamma}_L^{\mathrm{KL}}(\mathbf{F}_1,\ldots,\mathbf{F}_m)$ *such that*

$$\mathbf{u}^{(i)} \gg \mathbf{a}^{(i)} \text{ for all } 1 \leq i \leq m.$$

Then $\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1,\ldots,\mathbf{K}_n) \cap \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{F}_1,\ldots,\mathbf{F}_m) = \varnothing$ *implies*

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

In particular the above condition that

$$\mathbf{u}^{(i)} \gg \mathbf{a}^{(i)} \text{ for all } 1 \leq i \leq m$$

holds trivially if $\mathbf{F}_1,\ldots,\mathbf{F}_m \in BCL$, so the theorem suffices to show that the Strong Disagreement Principle **(K4*)** holds in that case.

Proof. Assume that $\mathbf{v} \in \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1,\ldots,\mathbf{K}_n)$ and

$$\mathbf{v} \in \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1,\ldots,\mathbf{K}_n, \mathbf{F}_1,\ldots,\mathbf{F}_m).$$

Let $(\mathbf{v}^{(1)},\ldots,\mathbf{v}^{(n)}) \in \hat{\Gamma}_L^{\mathrm{KL}}(\mathbf{K}_1,\ldots,\mathbf{K}_n)$ be an n-tuple associated with $\mathbf{v}$; in particular then $\mathbf{v} = \mathbf{LinOp}\,(\mathbf{v}^{(1)},\ldots,\mathbf{v}^{(n)})$.

Let

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

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 projections of the fixed $\mathbf{v}$ to each $V_{\mathbf{K}_i}$ 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

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

Now let $\mathbf{a} \in \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{F}_1,\ldots,\mathbf{F}_m)$ and $(\mathbf{a}^{(1)},\ldots,\mathbf{a}^{(n)}) \in \hat{\Gamma}_L^{\mathrm{KL}}(\mathbf{F}_1,\ldots,\mathbf{F}_m)$ an m-tuple associated with $\mathbf{a}$ be such that $\mathbf{u}^{(i)} \gg \mathbf{a}^{(i)}$ for all $1 \leq i \leq m$. This is possible by the assumption of the theorem.

If $\mathbf{v} \in \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{F}_1,\ldots,\mathbf{F}_m)$ then

$$\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1,\ldots,\mathbf{K}_n) \cap \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{F}_1,\ldots,\mathbf{F}_m) \neq \varnothing$$

and we are done. On the other hand we show that $\mathbf{v} \notin \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{F}_1,\ldots,\mathbf{F}_m)$ leads to a contradiction. First of all notice that from this assumption it follows that

$$\sum_{i=1}^{m} \mathrm{KL}(\mathbf{u}^{(i)} \| \mathbf{v}) > \sum_{i=1}^{m} \mathrm{KL}(\mathbf{a}^{(i)} \| \mathbf{a}).$$

Then by the lemma 4.3

$$\sum_{i=1}^{m} \sum_{j \in \mathrm{Sig}(\mathbf{v})} (a_j^{(i)} - u_j^{(i)}) \cdot (\log u_j^{(i)} - \log v_j) < 0.$$

On the other hand by the extended Pythagorean theorem (theorem 4.1) and by equation (3)

$$0 \leq \sum_{i=1}^{m} \mathrm{KL}(\mathbf{a}^{(i)} \| \mathbf{v}) - \mathrm{KL}(\mathbf{u}^{(i)} \| \mathbf{v}) - \mathrm{KL}(\mathbf{a}^{(i)} \| \mathbf{u}^{(i)}) =$$

$$= \sum_{i=1}^{m} \sum_{j \in \mathrm{Sig}(\mathbf{v})} (a_j^{(i)} - u_j^{(i)}) \cdot (\log u_j^{(i)} - \log v_j) < 0.$$

which is a contradiction.

□

The following counterexample shows that the restriction on the p–knowledge bases in theorem 4.6 is necessary even if we are considering only the weaker Disagreement Principle **(K4)** in place of **(K4*)**.

Example 4.7. *Assume that* $|L| = 2$, $V_{\mathbf{K}_1} = \{(1,0,0,0)\}$, $V_{\mathbf{K}_2} = \{(0,1,0,0)\}$, $V_{\mathbf{F}_1} = \{(x,0,1-x,0) : x \in [0,1]\}$ *and* $V_{\mathbf{F}_2} = \{(0,x,1-x,0) : x \in [0,1]\}$. *Clearly* $\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1,\mathbf{K}_2) = \{(\frac{1}{2},\frac{1}{2},0,0)\}$ *and* $\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{F}_1,\mathbf{F}_2) = \{(0,0,1,0)\}$. *Therefore* $\hat{\Delta}_L^{\mathrm{KL}}(\mathbf{K}_1,\mathbf{K}_2) \cap \hat{\Delta}_L^{\mathrm{KL}}(\mathbf{F}_1,\mathbf{F}_2) = \varnothing$. *It can now be shown that*

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

which suffices to contradict the Disagreement Principle.

□

Before leaving the discussion of $\hat{\Delta}^{\mathrm{KL}}$ we note that this p–merging operator is one of a large class of p–merging operators which all satisfy the same properties as $\hat{\Delta}^{\mathrm{KL}}$ does in theorem 4.6. These are formed by the class of

operators generated by substituting any convex Bregman divergence[12] in place of Kullback–Leibler divergence in the definition of $\hat{\Delta}^{\mathrm{KL}}$. This holds primarily because the well-known geometric properties of Bregman divergences, such as the extended Pythagorean theorem above, are exactly what is required for the proof of **(K4*)**. Amongst such Bregman divergences is the very special case of squared Euclidean distance E2, which has previously been considered in the context of probabilistic merging (see for instance [17]), and is defined by

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

Since in the case of this divergence the zero points cause no discontinuity, the Strong Disagreement Principle holds without any restriction on the class CL for the $\hat{\Delta}^{\mathrm{E2}}$ p–merging operator defined for any $\mathbf{K}_1, \ldots, \mathbf{K}_n \in CL$ as the set $\hat{\Delta}^{\mathrm{E2}}_L(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ of probability functions $\mathbf{v} \in \mathbb{D}^L$ which globally minimise the sum of squared Euclidean distances

$$\sum_{i=1}^{n} \mathrm{E2}(\mathbf{w}^{(i)} \,\|\, \mathbf{v}) \tag{8}$$

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

However what all the p–merging operators defined by convex Bregman divergences have in common is that they are generalisations of **LinOp** and reduce to **LinOp** when marginalised as pooling operators, see [0]. The social entropy operator Δ^{KL}, which marginalises to the **LogOp** pooling operator therefore has very different characteristics from p–merging operators defined in this way.

Finally the theorem below proves that the Δ^{KL}-merging operator satisfies the Disagreement Principle **(K4)** if the p–knowledge bases are restricted to *WBCL*. This together with theorem 4.5 above, and the earlier results of section 3.1, is sufficient to establish our theorem 3.1.

Theorem 4.8. *For all p–knowledge bases* $\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m \in WBCL$ *the social entropy operator* Δ^{KL} *satisfies* **(K4)**.

Proof. Let $\mathbf{K}_1, \ldots, \mathbf{K}_n, \mathbf{F}_1, \ldots, \mathbf{F}_m \in WBCL$ be such that $\bigcap_{i=1}^{m} V_{\mathbf{F}_i} \neq \varnothing$.

We must show that

$$\Delta^{\mathrm{KL}}(\mathbf{K}_1, \ldots, \mathbf{K}_n) \cap \Delta^{\mathrm{KL}}(\mathbf{F}_1, \ldots, \mathbf{F}_m) = \varnothing$$

[12] For the definition of a Bregman divergence see [2].

implies that

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

We prove the contrapositive. Suppose that for some fixed $\mathbf{v}$ we have that $\mathbf{v}\in\Delta^{\mathrm{KL}}(\mathbf{K}_1,\ldots,\mathbf{K}_n)$ and $\mathbf{v}\in\Delta^{\mathrm{KL}}(\mathbf{K}_1,\ldots,\mathbf{K}_n,\mathbf{F}_1,\ldots,\mathbf{F}_m)$. Let $\mathbf{v}^{(1)},\ldots,\mathbf{v}^{(n)}$ be such that they minimize $\sum_{j=1}^{J}\sum_{i=1}^{n}v_j\log\frac{v_j}{v_j^{(i)}}$ subject to $\mathbf{v}^{(1)}\in V_{\mathbf{K}_1},\ldots,\mathbf{v}^{(n)}\in V_{\mathbf{K}_n}$. Then

$$\mathbf{v}=\mathbf{LogOp}\,(\mathbf{v}^{(1)}\ldots,\mathbf{v}^{(n)}) \tag{9}$$

and furthermore by theorem 3.6(ii) of [26], since the constraint sets $\mathbf{K}_1,\ldots,\mathbf{K}_n$ are weakly bounded, for each j with $1\le j\le J$ the coordinates $v_j,v_j^{(1)},\ldots v_j^{(n)}$ are either all zero or all non-zero.

Similarly let $\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)},\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(m)}$ be such that they minimize

$$\sum_{j=1}^{J}\sum_{i=1}^{n}v_j\log\frac{v_j}{w_j^{(i)}}+\sum_{j=1}^{J}\sum_{i=1}^{m}v_j\log\frac{v_j}{u_j^{(i)}} \tag{10}$$

subject to $\mathbf{w}^{(1)}\in V_{\mathbf{K}_1},\ldots,\mathbf{w}^{(n)}\in V_{\mathbf{K}_n}$ and $\mathbf{u}^{(1)}\in V_{\mathbf{F}_1},\ldots,\mathbf{u}^{(m)}\in V_{\mathbf{F}_m}$. Equivalently $\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)},\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(m)}$ are such as to maximize

$$\sum_{j=1}^{J}\left[\prod_{i=1}^{n}w_j^{(i)}\prod_{i=1}^{m}u_j^{(i)}\right]^{\frac{1}{n+m}} \tag{11}$$

subject to $\mathbf{w}^{(1)}\in V_{\mathbf{K}_1},\ldots,\mathbf{w}^{(n)}\in V_{\mathbf{K}_n}$ and $\mathbf{u}^{(1)}\in V_{\mathbf{F}_1},\ldots,\mathbf{u}^{(m)}\in V_{\mathbf{F}_m}$ and are such that

$$\mathbf{v}=\mathbf{LogOp}\,(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)},\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(m)}). \tag{12}$$

Again, as in the case of (9) above, the weak boundedness of the constraint sets ensures that for each j with $1\le j\le J$ the coordinates $v_j,w_j^{(1)},\ldots w_j^{(n)},u_j^{(1)},\ldots u_j^{(m)}$ are either all zero or all non-zero.

Now notice that the minimisation of (10) can only occur when for all j such that $v_j\neq 0$

$$w_j^{(i)}=v_j^{(i)}\ \text{ for all }\ 1\le i\le n \tag{13}$$

since, for fixed $\mathbf{v}$, $\sum_{j\in\mathrm{Sig}(\mathbf{v})}v_j\log\frac{v_j}{w_j^{(i)}}$ is a strictly convex function of $\mathbf{w}^{(i)}$ for those $\mathbf{w}^{(i)}$ such that $w_j^{(i)}$ is non-zero if and only if $j\in\mathrm{Sig}(\mathbf{v})$, and hence it has a unique minimizer subject to $\mathbf{w}^{(i)}\in V_{\mathbf{K}_i}$ for all $1\le i\le n$, and by the definition of $\mathbf{v}$ that minimiser must be $\mathbf{v}^{(i)}$.

Equation (12) can by (13) be rewritten as

$$v_j = \frac{\left[\prod_{i=1}^{n} v_j^{(i)}\right]^{\frac{1}{n+m}} \left[\prod_{i=1}^{m} u_j^{(i)}\right]^{\frac{1}{n+m}}}{\sum_{j'=1}^{J} \left[\prod_{i=1}^{n} v_{j'}^{(i)} \prod_{i=1}^{m} u_{j'}^{(i)}\right]^{\frac{1}{n+m}}} \tag{14}$$

while from (9)

$$v_j = \frac{\left[\prod_{i=1}^{n} v_j^{(i)}\right]^{\frac{1}{n}}}{\sum_{j'=1}^{J} \left[\prod_{i=1}^{n} v_{j'}^{(i)}\right]^{\frac{1}{n}}}. \tag{15}$$

Raising equation (14) to the power $\frac{n+m}{m}$ and (14) to the power $\frac{n}{m}$, and dividing the first by the second, we obtain for all $j \in \mathrm{Sig}(\mathbf{v})$,

$$v_j = \frac{\left[\prod_{i=1}^{m} u_j^{(i)}\right]^{\frac{1}{m}}}{\frac{\left[\sum_{j'=1}^{J} \left[\prod_{i=1}^{n} v_{j'}^{(i)} \prod_{i=1}^{m} u_{j'}^{(i)}\right]^{\frac{1}{n+m}}\right]^{\frac{n+m}{m}}}{\left[\sum_{j'=1}^{J} \left[\prod_{i=1}^{n} v_{j'}^{(i)}\right]^{\frac{1}{n}}\right]^{\frac{n}{m}}}}. \tag{16}$$

Notice that in order to obtain (16) above the cancelation of a term $[\prod_{i=1}^{n} v_j^{(i)}]^{\frac{1}{m}}$ is required; however this is permissible since we know by the remark following (9) that this term is non-zero for $j \in \mathrm{Sig}(\mathbf{v})$. For similar reasons the denominator on the right of (16) is finite and non-zero. On the other hand (16) holds even if $j \notin \mathrm{Sig}(\mathbf{v})$ since by remarks above both sides are then zero.

Now since the denominator on the right of (16) is independent of j it follows at once that

$$\mathbf{v} = \mathbf{LogOp}(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(m)}). \tag{17}$$

Now let $\mathbf{a}$ be consistent with $\mathbf{F}_i$ for all i so that in particular

$$\mathbf{a} \in \Delta^{\mathrm{KL}}(\mathbf{F}_1, \ldots, \mathbf{F}_m) = \bigcap_{i=1}^{m} V_{\mathbf{F}_i}.$$

Consider

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

We need to show that

$$\Delta^{\mathrm{KL}}(\mathbf{K}_1,\ldots,\mathbf{K}_n)\cap\Delta^{\mathrm{KL}}(\mathbf{F}_1,\ldots,\mathbf{F}_m)=\Delta^{\mathrm{KL}}(\mathbf{K}_1,\ldots,\mathbf{K}_n)\cap\bigcap_{i=1}^{m}V_{\mathbf{F}_i}\neq\varnothing.$$

For this it is sufficient to prove that $\frac{d}{d\lambda}F|_{\lambda=0}>0$ unless $\mathbf{u}^{(1)}=\ldots=\mathbf{u}^{(n)}=\mathbf{v}$. The idea here is that if the maximum value of F is obtained for $\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(n)}$ which are not all equal then from the existence of the point $\mathbf{a}\in\bigcap_{i=1}^{m}V_{\mathbf{F}_i}$ we can show that $F(0)<F(\lambda)$ for some $\lambda>0$; this is a contradiction since by the convexity of the $V_{\mathbf{F}_i}$ each $\mathbf{u}^{(i)}+\lambda(\mathbf{a}-\mathbf{u}^{(i)})\in V_{\mathbf{F}_i}$ so that $F(0)<F(\lambda)$ contradicts the maximality of (11) given (13).

Note that by (17) $v_j=\frac{(\prod_{i=1}^{m}u_j^{(i)})^{\frac{1}{m}}}{M}$, where

$$M=\sum_{j=1}^{J}\left[\prod_{i=1}^{m}u_j^{(i)}\right]^{\frac{1}{m}}<1\ \text{ unless }\ \mathbf{u}^{(1)}=\ldots=\mathbf{u}^{(m)}. \tag{18}$$

Now

$$\frac{d}{d\lambda}F(\lambda)=\sum_{j=1}^{J}\left(\prod_{i=1}^{n}v_j^{(i)}\right)^{\frac{1}{n+m}}\left[\prod_{i=1}^{m}u_j^{(i)}+\lambda\sum_{i=1}^{m}[(a_j-u_j^{(i)})\prod_{i'\neq i,\,i'=1,\ldots,m}u_j^{(i')}]+\right.$$

$$\left.+O(\lambda^2)\right]^{\frac{1}{n+m}-1}\cdot\left[\sum_{i=1}^{m}[(a_j-u_j^{(i)})\prod_{i'\neq i,\,i'=1,\ldots,m}u_j^{(i')}]+O(\lambda)\right]\cdot\frac{1}{n+m}.$$

We obtained this by expanding

$$\prod_{i=1}^{m}[u_j^{(i)}+\lambda(a_j-u_j^{(i)})]=\prod_{i=1}^{m}u_j^{(i)}+\lambda\sum_{i=1}^{m}[(a_j-u_j^{(i)})\prod_{i'\neq i,\,i'=1,\ldots,m}u_j^{(i')}]+O(\lambda^2).$$

Furthermore

$$\frac{d}{d\lambda}F(\lambda)|_{\lambda=0}=\frac{1}{n+m}\cdot\sum_{j=1}^{J}\left[\prod_{i=1}^{n}v_j^{(i)}\right]^{\frac{1}{n+m}}\left[(Mv_j)^m\right]^{-1+\frac{1}{n+m}}\cdot$$

$$\cdot\left[\sum_{i=1}^{m}(a_j u_j^{(i)}M^m(v_j)^m-M^m(v_j)^m)\right]=C\cdot\sum_{j=1}^{J}(v_j)^{\frac{n}{n+m}-m+\frac{m}{n+m}+m}\left[\sum_{i=1}^{m}\frac{a_j}{u_j^{(i)}}-1\right]=$$

$$=C\cdot\left[\sum_{j=1}^{J}\sum_{i=1}^{m}\frac{v_j a_j}{u_j^{(i)}}-m\right]=C\cdot\left[\sum_{j=1}^{J}a_j\sum_{i=1}^{m}\frac{\left(\prod_{k=1}^{m}u_j^{(k)}\right)^{\frac{1}{m}}}{u_j^{(i)}M}-m\right],$$

where

$C = \frac{1}{n+m} \cdot M^{\frac{1}{n+m}} \cdot (\sum_{j=1}^{J} [\prod_{i=1}^{n} v_j^{(i)}]^{\frac{1}{n}})^{\frac{n}{n+m}}$ is a positive constant.

Note that if $u_j^{(i)} = 0$ and $a_j \neq 0$ for some $1 \leq i \leq m$ and some j then $F(\lambda) \to +\infty$ as $\lambda \to 0^+$. On the other hand if also $a_j = 0$ then we can just leave out that index j from the summation. Finally by the arithmetic–geometric inequality[13]

$$C \cdot \left[\sum_{j=1}^{J} a_j \sum_{i=1}^{m} \frac{1}{u_j^{(i)}} \frac{\left(\prod_{k=1}^{m} u_j^{(k)} \right)^{\frac{1}{m}}}{M} - m \right] \geq$$

$$\geq C \cdot \left[\sum_{j=1}^{J} a_j m \cdot \left(\prod_{i=1}^{m} \frac{1}{u_j^{(i)}} \right)^{\frac{1}{m}} \frac{\left(\prod_{k=1}^{m} u_j^{(k)} \right)^{\frac{1}{m}}}{M} - m \right] = Cm \frac{1-M}{M}.$$

By (18) the last term is greater than 0 unless $\mathbf{u}^{(1)} = \ldots = \mathbf{u}^{(m)}$, which concludes the proof.

The following counterexample shows that the theorem above fails if *WBCL* is replaced by *CL*.

Example 4.9. *Let* $V_{\mathbf{K}_1}^L = \{(0,0,\frac{1}{3},\frac{2}{3})\}$ *and* $V_{\mathbf{F}_1}^L = \{(0,\frac{1}{3},\frac{2}{9},\frac{4}{9})\}$.
Obviously $\Delta^{\mathrm{KL}}(\mathbf{K}_1) \cap \Delta^{\mathrm{KL}}(\mathbf{F}_1) = \varnothing$.
However $\Delta^{\mathrm{KL}}(\mathbf{K}_1, \mathbf{F}_1) = \mathbf{LogOp}\,[(0,0,\frac{1}{3},\frac{2}{3}),(0,\frac{1}{3},\frac{2}{9},\frac{4}{9})] = (0,0,\frac{1}{3},\frac{2}{3})$.

□

Acknowledgements

Both authors are indebted to Jon Williamson and Alena Vencovská for valuable discussions concerning the problematic of probabilistic merging, and to the anonymous referees for several helpful comments. The first author was supported by the [European Community's] Seventh Framework Programme [FP7/2007-2013] under grant agreement no 238381. The second author is grateful to the School of Mathematics of the University of Leeds for the technical support which he has received since his retirement from the University of Manchester.

[13] For all $x_1 \geq 0, \ldots, x_n \geq 0$ the inequality $(\prod_{i=1}^{n} x_i)^{\frac{1}{n}} \leq \frac{\sum_{i=1}^{n} x_i}{n}$ holds.

References

[0] Adamčík, M. (2014), *The Information Geometry of Bregman Divergences and Some Applications in Multi-Expert Reasoning*. Entropy 16(12), pp. 6338-6381.

[1] Adamčík, M. and Wilmers, G.M. (2012), *The Irrelevant Information Principle for Collective Probabilistic Reasoning*. Kybernetika 50(2), pp. 175-188.

[2] Amari, S. (2009), *Divergence, Optimization and Geometry*. Neural Information Processing: 16th International Conference, Iconip, pp. 185-193.

[3] Bernoulli, J. (1713), *Ars Conjectandi, opus posthumum: Accedit Tractatus de Seriebus Infinitis, et Epistola Gallicè Scripta de Ludo Pilae Reticularis*. Thurneysen Brothers, Basel. Translated to English by Edith Sylla as *The Art of Conjecturing, together with Letter to a Friend on Sets in Court Tennis*, Johns Hopkins University Press, Baltimore, MD, 2006.

[4] Carnap, R. (1947), *On the application of inductive logic*. Philosophy and Phenomenological Research 8, pp. 133-148.

[5] Collins, M. and Schapire, R.E. (2002), *Logistic Regression, AdaBoost and Bregman Distances*. Machine Learning 48, pp. 253-285.

[6] Cover, T.M. and Thomas, J.A. (1991), *Elements of Information Theory*. Wiley Series in Telecommunications, John Wiley and Sons, New York.

[7] Csiszár, I. (1975), *I-Divergence Geometry of Probability Distribution and Minimization Problems*. The Annals of Probability 3(1), pp. 146-158.

[8] Csiszár, I. and Tusnády, G. (1984), *Informational Geometry and Alternating Minimization Procedures*. Statistic and Decisions 1, pp. 205-237.

[9] 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.

[10] Genest, C. and Zidek, J.V. (1986), *Combining probability distributions: A critique and an annotated bibliography*. Statistical Science 1(1), pp. 114-135.

[11] Hájek, P., Havránek, T. and Jiroušek, J. (1992), *Uncertain Information Processing in Expert Systems*. CRC Press, Boca Raton, Ann Arbor, London, Tokyo.

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

[13] Kern-Isberner, G. and Rödder, W. (2004), *Belief Revision and Information Fusion on Optimum Entropy*. International Journal of Inteligent System 19, pp. 837-857.

[14] 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.

[15] Konieczny, S. and Pino Pérez, R. (2011), *Logic Based Merging*. J. Philosophical Logic 40, pp. 239-270.

[16] Matúš, F. (2007), *On iterated averages of I-projections*. Statistik und Informatik, Universität Bielefeld, Bielefeld, pp. 1-12.

[17] Osherson, D. and Vardi, M. (2006), *Aggregating disparate estimates of chance*. Games and Economic Behavior 56(1), pp. 148-173.

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

[19] PARIS, J.B. (1998), *Common sense and maxiumum entropy*. Synthese 117, pp. 75-93.
[20] PARIS, J.B. and VENCOVSKÁ, A. (1990), *A Note on the Inevitability of Maximum Entropy*. International Journal of Approximate Reasoning 4, pp. 183-224.
[21] 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.
[22] WILLIAMSON, J. (2009), *Aggregating Judgements by Merging Evidence*. J. Logic Computation 19(3), pp. 461-473.
[23] WILLIAMSON, J. (2010), *In Defense of Objective Bayesianism*. Oxford University Press, Oxford.
[24] WILLIAMSON, J. (2013), *Deliberation, Judgement and the Nature of Evidence*. Economics and Philosophy, in press.
[25] 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.
[26] WILMERS, G.M. (2015), *A Foundational Approach to Generalising the Maximum Entropy Inference Process to the Multi-Agent Context*. Entropy 17(2), pp. 594-645.

Martin ADAMČÍK
Martin de Tours School of
Management and Economics
Assumption University
maths38@gmail.com

George WILMERS
School of Mathematics
University of Leeds
george.wilmers@gmail.com