Entropy **2014**, *16*, 6338-6381; doi:10.3390/e16126338

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

The Information Geometry of Bregman Divergences and Some Applications in Multi-Expert Reasoning

Martin Adamčík

Martin de Tours School of Management and Economics, Assumption University, MSME Building,
4th Floor, 88 Moo 8 Bang Na-Trad Km. 26 Bangsaothong, 10540 Samuthprakarn, Thailand;
E-Mail: maths38@gmail.com; Tel.: +66-991311270

External Editors: Juergen Landes and Jon Williamson

Received: 19 October 2014; in revised form: 24 November 2014 / Accepted: 25 November 2014 / Published: 1 December 2014

Abstract: The aim of this paper is to develop a comprehensive study of the geometry involved in combining Bregman divergences with pooling operators over closed convex sets in a discrete probabilistic space. A particular connection we develop leads to an iterative procedure, which is similar to the alternating projection procedure by Csiszár and Tusnády. Although such iterative procedures are well studied over much more general spaces than the one we consider, only a few authors have investigated combining projections with pooling operators. We aspire to achieve here a comprehensive study of such a combination. Besides, pooling operators combining the opinions of several rational experts allows us to discuss possible applications in multi-expert reasoning.

Keywords: Bregman divergence; information geometry; pooling operator; discrete probability function; probabilistic merging; multi-expert reasoning

1. Background

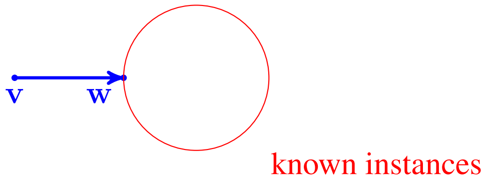
1.1. Introduction

Information geometry has been studied as a powerful tool for tackling various problems. It has been applied in neuroscience [1], expert systems [2], logistic regression [3], clustering [4] and probabilistic merging [5]. In this paper, we aim to present a comprehensive study of information geometry over a discrete probabilistic space in order to provide some specialized tools for researchers working in the area of multi-expert reasoning.

In the context of this paper, the domain of information geometry is the Euclidean space $\mathbb{R}^J$, for some fixed natural number $J \geq 2$, where we measure a divergence from one point to another one. A divergence is, in general asymmetric, a notion of distance, and we will represent it here by an arrow. A divergence can represent a cost function having various constraints, so many engineering problems correspond to the minimization of a divergence.

For example, in the areas of neuroscience and expert systems, given evidence $\mathbf{v}$ and a training set of known instances W, we may search for an instance $\mathbf{w} \in W$, which is "closest" to the evidence $\mathbf{v}$, so as to represent it in the given training set W. An illustration is depicted in Figure 1.

Figure 1. An illustration of a divergence.



A similar pattern of minimization appears also in the areas of clustering and regression. The aim of the former is to categorize several points into a given number of nodes in such a way that the sum of divergences from each point to its associated node is minimal. The aim of regression is to predict an unknown distribution of events based on the previously obtained statistical data by defining a function whose values minimize a sum of divergences to the data.

While several domains for divergences are considered in the literature, in the current presentation of information geometry, however, we will confine ourselves to the domain of positive discrete probability functions $\mathbb{D}^J$, where $\mathbb{D}^J$ is the set of all $\mathbf{w} \in \mathbb{R}^J$ restricted by $\sum_{j=1}^J w_j = 1$ and $w_1 > 0, \ldots, w_J > 0$. In our presentation, $J \geq 2$ will be always fixed, but otherwise arbitrary.

Although in information geometry, it does not make sense to talk about beliefs, applications in multi-expert reasoning are often developed from that perspective. It is then argued that rational beliefs should obey the laws of probability, for example the Dutch book argument by Ramsey and de Finetti [6] is perhaps the most compelling argument. It is therefore of a particular interest to develop information geometry over a probabilistic space if we wish to eventually apply it to multi-expert reasoning.

In addition to our restriction to discrete probability functions, we will confine ourselves to a special type of divergence, called a Bregman divergence [7], which has recently attracted attention in machine learning and plays a major role in optimization; *cf.* [3]. A Bregman divergence over a discrete probabilistic space is defined by a given strictly convex function $f : (0,1)^J \to \mathbb{R}$, which is differentiable over $\mathbb{D}^J$. For any $\mathbf{v}, \mathbf{w} \in \mathbb{D}^J$, the Bregman divergence generated by the function f is given by:

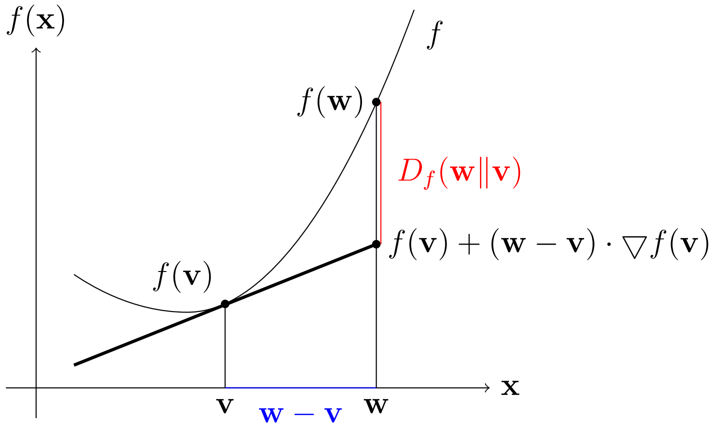
$$D_f(\mathbf{w}\|\mathbf{v}) = f(\mathbf{w}) - f(\mathbf{v}) - (\mathbf{w} - \mathbf{v}) \cdot \bigtriangledown f(\mathbf{v}),$$

where $\bigtriangledown f(\mathbf{v})$ is the gradient of f and $\cdot$ denotes the inner (dot) product of two vectors, *i.e.*,

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

We say that $D_f(\mathbf{w}\|\mathbf{v})$ is a Bregman divergence from $\mathbf{v} \in \mathbb{D}^J$ to $\mathbf{w} \in \mathbb{D}^J$. Figure 2 depicts a geometrical interpretation of a Bregman divergence.

Figure 2. A Bregman divergence.



By the first convexity condition applied to the (convex and differentiable) function f (see, e.g., [8]), $D_f(\mathbf{w}\|\mathbf{v}) \geq 0$ with equality holding only if $\mathbf{w} = \mathbf{v}$. This is the condition that makes $D_f(\cdot\|\cdot)$ a divergence as defined in information geometry. Note that, since a differentiable convex function is necessarily continuously differentiable (see [9]), $D_f(\mathbf{w}\|\mathbf{v})$ is a continuous function. However, note that this is not sufficient to establish the differentiability of D_f.

It is worth mentioning that the restriction $w_1 > 0, \ldots, w_J > 0$ for a probability function $\mathbf{w}$ that we have adopted here is important for the definition of a Bregman divergence. Some Bregman divergences do not have their generating function f differentiable over the whole space of probability functions. However, it is possible to define the notion of a Bregman divergence even if this condition is left out, but at the cost of some restrictions on f. We kindly refer the interested reader to [10] for further details. Nonetheless, the setting developed in [10] uses a rather complicated notation, which could prove to be impenetrable at first glance if it were adopted in the current paper.

In this paper, we study mainly Bregman divergences $D_f(\cdot\|\cdot)$, which are convex, *i.e.*, for all $\lambda \in [0,1]$ and all $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \mathbf{v}^{(1)}, \mathbf{v}^{(2)} \in \mathbb{D}^J$:

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

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

The following are examples of a convex Bregman divergence.

Example 1 (Squared Euclidean Distance)**.** *For any $J \geq 2$ let $f(\mathbf{x}) = \sum_{j=1}^{J}(x_j)^2$. Then, the divergence:*

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

will be denoted by E2*, and exceptionally, this divergence is symmetric.*

Example 2 (Kullback–Leibler Divergence)**.** *For any $J \geq 2$, let $f(\mathbf{x}) = \sum_{j=1}^{J} x_j \log x_j$, where* log *denotes the natural logarithm. (Note that in the information theory literature, this logarithm is often* *taken with base two. However, this does not affect the results of this paper in any way.) The well-known divergence:*

$$D_f(\mathbf{w}\|\mathbf{v}) = \sum_{j=1}^{J} w_j \log \frac{w_j}{v_j}$$

will be denoted by KL*.*

The convexity of the KL-divergence is easy to observe and is well known; see, e.g., [10].

1.2. Projections

For given $\mathbf{v} \in \mathbb{D}^J$, a Bregman divergence $D_f(\mathbf{w}\|\mathbf{v})$ is a strictly convex function in the first argument. This can be easily seen by considering $D_f(\mathbf{w}\|\mathbf{v}) = f(\mathbf{w}) - f(\mathbf{v}) - \sum_{j=1}^{J}(w_j - v_j)\frac{\partial f(\mathbf{v})}{\partial v_j}$ where $\mathbf{v}$ is constant. $f(\mathbf{v})$ is therefore constant, as well, and the claim follows, since strict convexity of f is not affected by adding the linear term $-\sum_{j=1}^{J}(w_j - v_j)\frac{\partial f(\mathbf{v})}{\partial v_j}$.

Owing to the observation above, if $\mathbf{v} \in \mathbb{D}^J$ is given and $W \subseteq \mathbb{D}^J$ is a closed convex nonempty set, we can define the D_f-projection of $\mathbf{v}$ into W. It is that unique point $\mathbf{w} \in W$ that minimizes $D_f(\mathbf{w}\|\mathbf{v})$ subject only to $\mathbf{w} \in W$. This property is crucial for the applicability of Bregman divergences. Note, however, that $D_f(\cdot\|\cdot)$ is not necessarily convex in its second argument; for a counterexample, consider the case $f(\mathbf{x}) = \sum_{j=1}^{4}(x_j)^3$.

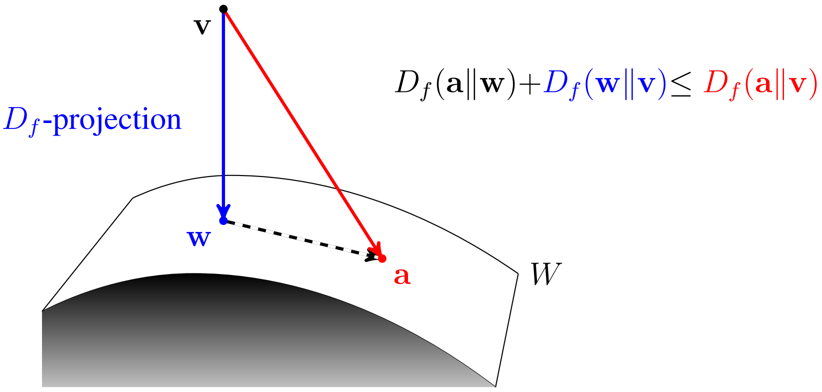
Perhaps the most useful property that a D_f-projection has is the extended Pythagorean property:

Theorem 1 (Extended Pythagorean Property)**.** *Let D_f be a Bregman divergence. Let $\mathbf{w}$ be the D_f-projection of $\mathbf{v} \in \mathbb{D}^J$ into a closed convex nonempty set $W \subseteq \mathbb{D}^J$. Let $\mathbf{a} \in W$. Then:*

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

This property, in the case of the Kullback–Leibler divergence, was proven first by Csiszár in [11]. The proof of the generalized theorem above is given in [1,12], where the interested reader can find a comprehensive study of Bregman divergences within the context of differential geometry. We illustrate the theorem in Figure 3.

Figure 3. The extended Pythagorean property.


$\mathbf{v}$
$D_f(\mathbf{a}\|\mathbf{w})+D_f(\mathbf{w}\|\mathbf{v})\leq D_f(\mathbf{a}\|\mathbf{v})$
D_f-projection
$\mathbf{w}$
$\mathbf{a}$
W

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

It is well-known that the Kullback–Leibler divergence is closely connected to the Shannon entropy defined for any $\mathbf{w} \in \mathbb{D}^J$ by:

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

where $\log$ denotes the natural logarithm. The importance of the Shannon entropy is that it could be described as a measure of the level of disorder, which in the context of information theory, can be interpreted as a measure of informational content. The higher the entropy of $\mathbf{w}$ is, the less information is carried by $\mathbf{w}$. In some contexts, one can then argue that given several seemingly equally probable choices of a probability function, one should choose the one that carries the least additional information [13]. Given a closed convex nonempty set W, the most entropic point in W will be denoted by $\mathbf{ME}(W)$.

Now, trying to find the most entropic point in a closed convex nonempty set $W \subseteq \mathbb{D}^J$ is, in fact, equivalent to finding a special KL-projection (the KL-projection of the uniform probability function $\underbrace{\left(\frac{1}{J}, \ldots, \frac{1}{J}\right)}_{J}$) since:

$$\arg\min_{\mathbf{w}\in W} \sum_{j=1}^{J} w_j \log \frac{w_j}{\frac{1}{J}} = \arg\max_{\mathbf{w}\in W} - \sum_{j=1}^{J} w_j \log w_j = \mathbf{ME}(W),$$

where $\arg\min_{x\in X} f(x)$ denotes that unique argument $x \in X$, where f has its global minimum, whenever such a unique point exists. The expression $\arg\max$ is defined accordingly.

Given the extensive justification of the Shannon entropy in various frameworks (see, e.g., [14,15]), it is perhaps not surprising that a common method of projecting in probabilistic expert systems is by means of the KL-projection; see [2,16]. In connection to the Shannon entropy, the KL-divergence is often referred to as the cross-entropy, and the projecting is called updating.

The above may perhaps be also an appealing reason to use projections in general to "represent" a given closed convex set of probability functions by a single point, in particular in expert reasoning. Moreover, recent use of projections by a Bregman divergence has become popular in other contexts; see, e.g., [4]. Remarkably, projections by a Bregman divergence also provide a unifying framework for a variety of techniques used in expert systems, such as logistic regression; see [3]. It is therefore of particular interest to investigate the geometry of Bregman divergences.

1.3. Pooling

In this subsection, we introduce probabilistic pooling, which is a method of aggregating several probability functions. Formally, a pooling operator $\mathbf{Pool}$ is defined for each $n \geq 1$ as a mapping:

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

Recall that J is a fixed natural number greater than or equal to two, which is otherwise arbitrary.

One possibility for choosing a pooling operator is to define one by means of a Bregman divergence. In particular, given a Bregman divergence D_f, $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)} \in \mathbb{D}^J$ and $\mathbf{a} \in \mathbb{D}^n$, we can ask which point

$\mathbf{v} \in \mathbb{D}^J$ has the least sum of Bregman divergences D_f from $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}$ weighted by $a_1, \dots, a_n$, respectively. It turns out that the resulting probability function is unique, and in each coordinate, it is simply the weighted arithmetic mean of the corresponding coordinates of $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)} \in \mathbb{D}^J$. In other words:

$$\arg\min_{\mathbf{v}\in\mathbb{D}^J} \sum_{i=1}^{n} a_i D_f(\mathbf{w}^{(i)} \| \mathbf{v}) = \Big(\sum_{i=1}^{n} a_i w_1^{(i)}, \dots, \sum_{i=1}^{n} a_i w_J^{(i)}\Big). \tag{1}$$

For a given family $\mathcal{A} = \{\mathbf{a}_n : \ \mathbf{a}_n \in \mathbb{D}^n, \ n = 1, 2, \dots\}$ of weighting vectors, we define the pooling operator $\mathbf{LinOp}_{\mathcal{A}}$ by Equation (1) for every $\mathbf{a} \in \mathcal{A}$. Instead of the right-hand side of Equation (1), we will simply write $\mathbf{LinOp_a}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$ if $\mathbf{a} \in \mathcal{A}$. A special choice for $\mathcal{A}$ is the family $\mathcal{N} = \{\mathbf{a}_n = (\frac{1}{n}, \dots, \frac{1}{n}) : \ n = 1, 2, \dots\}$, and the pooling operator $\mathbf{LinOp}_{\mathcal{N}}$ is well known in the literature as the $\mathbf{LinOp}$-pooling operator.

The fact that Equation (1) actually holds can be observed by employing the following theorem, which is folklore in information theory.

Theorem 2 (Parallelogram Theorem)**.** *Let D_f be a Bregman divergence, $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}, \mathbf{v} \in \mathbb{D}^J$ and $\mathbf{a} \in \mathbb{D}^n$. Then:*

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

$$+ D_f(\mathbf{LinOp_a}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}) \| \mathbf{v}).$$

Proof. Let $\mathbf{w} = \mathbf{LinOp_a}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$. The equality is easy to observe by:

$$\sum_{i=1}^{n} a_i \Big[f(\mathbf{w}^{(i)}) - f(\mathbf{v}) - \sum_{j=1}^{J} (w_j^{(i)} - v_j) \frac{\partial f(\mathbf{v})}{\partial v_j} \Big] =$$

$$= \sum_{i=1}^{n} a_i \Big[f(\mathbf{w}^{(i)}) - f(\mathbf{w}) - (\mathbf{w}^{(i)} - \mathbf{w}) \cdot \nabla f(\mathbf{w}) \Big] +$$

$$+ \Big[f(\mathbf{w}) - f(\mathbf{v}) - \sum_{j=1}^{J} (w_j - v_j) \frac{\partial f(\mathbf{v})}{\partial v_j} \Big]$$

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

Since $D_f(\mathbf{w} \| \mathbf{v}) = 0$, only if $\mathbf{w} = \mathbf{v}$, and otherwise, it is positive, the unique minimum of the left-hand side of Equation (1) is at the point $\mathbf{v} = \mathbf{LinOp_a}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$.

The situation above can be naturally interpreted in terms of random variables. Assume that X is a random variable taking values in $\{\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}\} \subseteq \mathbb{D}^J$ with the probability distribution $\mathbf{a} \in \mathbb{D}^n$, and we are given the problem of finding a random variable Y, such that the expected value:

$$\mathrm{E}(D_f(X \| Y))$$

is minimal. The unique answer to this question is then $Y = \mathrm{E}(X) = \sum_{i=1}^{n} a_i \mathbf{w}^{(i)}$. This underlines the reason why the $\mathbf{LinOp}_{\mathcal{A}}$-pooling operator is so popular in the decision theory literature, where several experts, each with his own probability function $\mathbf{w}^{(i)}$ representing his beliefs, seek to find a single probability function to represent their joint beliefs. The $\mathbf{LinOp}_{\mathcal{A}}$-pooling operator simply yields the expected value as if expert's beliefs were statistically obtained.

It is certainly interesting that the result above holds for any Bregman divergence, but as is shown in [17], Theorem 4, it is even more remarkable that Bregman divergences are the only divergences with such a property. However, we note that in order to establish this claim, a slightly more general setting was considered and that we have restricted the formulation of the original theorem to the only domain considered here $(0,1)^J$:

Theorem 3 (Banerjee, Guo, Wang)**.** *Let $F : (0,1)^J \times (0,1)^J \to \mathbb{R}$ be a divergence. Assume that $F(\mathbf{x}\|\mathbf{y})$, $\frac{\partial^2 F(\mathbf{x}\|\mathbf{y})}{\partial x_i \partial x_j}$, $1 \leq i,j \leq J$ are all continuous. Let $(\Omega, \mathrm{P}, \mathcal{F})$ be an arbitrary probability space, and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. For all random variables X taking values in $(0,1)^J$, if:*

$$\arg\min_{Y \in \mathcal{G}} F(X\|Y) = \mathrm{E}(X|\mathcal{G})$$

then $F(\mathbf{x}\|\mathbf{y}) = D_f(\mathbf{x}\|\mathbf{y})$ for some strictly convex and differentiable function $f : (0,1)^J \to \mathbb{R}$.

While in the statistical sense, the $\mathbf{LinOp}_{\mathcal{A}}$-pooling operator, where $\mathcal{A}$ is a family of weighting vectors, seems to be well placed, in the fields of multi-expert reasoning and probabilistic merging, the so-called $\mathbf{LogOp}_{\mathcal{A}}$-pooling operator often appeals more. For every $n \geq 1$ and every $\mathbf{a} \in \mathcal{A}$, it is defined by:

$$\mathbf{LogOp}_{\mathbf{a}}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) = \left(\frac{\prod_{i=1}^n (w_1^{(i)})^{a_i}}{\sum_{j=1}^J \prod_{i=1}^n (w_j^{(i)})^{a_i}}, \ldots, \frac{\prod_{i=1}^n (w_J^{(i)})^{a_i}}{\sum_{j=1}^J \prod_{i=1}^n (w_j^{(i)})^{a_i}} \right).$$

If $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ are considered to be beliefs of n-experts, respectively, then the $\mathbf{LogOp}_{\mathcal{A}}$-pooling operator appears to favor agreement over the expected value. For instance, consider the following example from utility theory. Say that Eleanor and George are looking for a film to watch and they have three options, A, B and C. Eleanor hates Movie A and under no circumstances would agree to watch it, while George absolutely loves it. Now, consider that the situation with respect to Film C is swapped: George hates it, while Eleanor would prefer to see it. They both consider Movie B uninteresting, but are willing to see it. The following probability functions could represent the preferences of Eleanor and George towards Movies A, B and C: $(0, 0.1, 0.9)$ and $(0.9, 0.1, 0)$, respectively. Moreover, we value the opinions of both of them equally, *i.e.*, $\mathcal{A} = \mathcal{N}$. Now, while the $\mathbf{LinOp}_{\mathcal{N}}$-pooling operator gives inconclusive $(0.45, 0.1, 0.45)$ by the $\mathbf{LogOp}_{\mathcal{N}}$-pooling operator (in the literature, this operator is simply known as the $\mathbf{LogOp}$-pooling operator), we obtain $(0, 1, 0)$. If we take the advice, then Eleanor and George should see the only film that is acceptable for both of them.

The example above illustrates why taking products rather than the arithmetic mean is popular when considering utilities. However, recently, the $\mathbf{LogOp}_{\mathcal{N}}$-pooling operator attracted attention also in multi-expert probabilistic reasoning; a prominent example here is the social entropy process by Wilmers [18]. An intriguing idea that originates in the social entropy process is to swap the direction of the Kullback–Leibler projections and establish the corresponding conjugated KL-projection of $\mathbf{w} \in \mathbb{D}^J$ into $V \subseteq \mathbb{D}^J$ as $\arg\min_{\mathbf{v} \in V} \mathrm{KL}(\mathbf{w}\|\mathbf{v})$ (it is easy to check that $\mathrm{KL}(\cdot\|\cdot)$ is strictly convex in its second argument) and the conjugated parallelogram theorem [10]:

Theorem 4. *Let* $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}, \mathbf{v} \in \mathbb{D}^J$ *and* $\mathbf{a} \in \mathbb{D}^n$. *Then:*

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

$$+ \mathrm{KL}(\mathbf{v} \| \,\mathbf{LogOp_a}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})).$$

Proof. Let $\mathbf{w} = \mathbf{LogOp_a}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$. First note that:

$$\sum_{i=1}^{n} a_i \sum_{j=1}^{J} v_j \log \frac{v_j}{w_j^{(i)}} = \sum_{j=1}^{J} v_j \log \frac{v_j}{\prod_{i=1}^{n} (w_j^{(i)})^{a_i}}.$$

Now:

$$\sum_{j=1}^{J} v_j \log \frac{v_j}{\prod_{i=1}^{n} (w_j^{(i)})^{a_i}} = \sum_{j=1}^{J} v_j \log \frac{v_j}{w_j} -$$

$$- \Big(\sum_{j=1}^{J} v_j \Big) \log \Big(\sum_{j=1}^{J} \prod_{i=1}^{n} (w_j^{(i)})^{a_i} \Big) =$$

$$= \sum_{j=1}^{J} v_j \log \frac{v_j}{w_j} + \sum_{i=1}^{n} a_i \sum_{j=1}^{J} w_j \log \frac{w_j}{w_j^{(i)}},$$

where we have used the fact that $\sum_{j=1}^{J} v_j = 1$. □

As a consequence, for given $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)} \in \mathbb{D}^J$, we get:

$$\arg \min_{\mathbf{v} \in \mathbb{D}^J} \sum_{i=1}^{n} a_i \,\mathrm{KL}(\mathbf{v} \| \mathbf{w}^{(i)}) = \mathbf{LogOp_a}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}).$$

Therefore, the $\mathbf{LogOp}_{\mathcal{A}}$-pooling operator can be naturally interpreted in information geometry. The question now arises as to whether this can be done with some other Bregman divergences. We will investigate this later.

The reader perhaps wonders which are the main practical differences in using different pooling operators. The $\mathbf{LinOp}_{\mathcal{A}}$-pooling operator, for example, satisfies the marginalization property, that is the values on the coordinates of the resulting probability function depend only on the corresponding coordinates of the probability functions that are pooled. The $\mathbf{LogOp}_{\mathcal{A}}$-pooling operator does not have this property. On the other hand, the $\mathbf{LogOp}_{\mathcal{A}}$-pooling operator, unlike the $\mathbf{LinOp}_{\mathcal{A}}$-pooling operator, is externally Bayesian. That is the order in which we combine pooling and Bayesian updating is irrelevant. See [19] for more details.

We, however, do not seek any conclusive answer as to which pooling operator to use in any particular context. In this paper, we only aim to provide geometric tools that can be used in multi-expert reasoning. For elaborate work on pooling operators, we refer to the literature, e.g., [19] for a survey, [20] for a classical problem of the relationship between pooling and probabilistic independence or [18] for a modern account on $\mathbf{LinOp}_{\mathcal{N}}$ and $\mathbf{LogOp}_{\mathcal{N}}$-pooling operators in probabilistic knowledge merging.

2. Projections and Pooling Combined

2.1. Averaging Projective Procedures

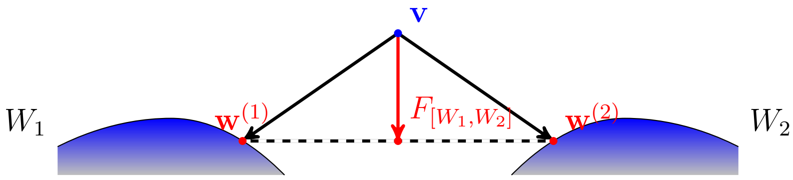
While the geometry of projections and the theory of pooling operators have been extensively studied in the literature (see the previous section), much less attention, however, was been devoted to the combination of them. A detailed study of this problem and a comprehensive analysis of the geometry involved is the main aim of this paper.

The central geometrical notion connecting projections and pooling in this paper is an averaging projective procedure F, which consists of a family of mappings $F_{[W_1,\dots,W_n]} : \mathbb{D}^J \to \mathbb{D}^J$, where sets $W_1, \dots, W_n \subseteq \mathbb{D}^J$ are closed convex and nonempty. A particular F is given by a family of strictly convex functions $d_{\mathbf{v}}$, $\mathbf{v} \in \mathbb{D}^J$ and a pooling operator $\mathbf{Pool}$ and is defined by the following two-stage process.

1. For an argument $\mathbf{v} \in \mathbb{D}^J$, put $\mathbf{w}^{(i)} = \arg\min_{\mathbf{w}\in W_i} d_{\mathbf{v}}(\mathbf{w})$, $1 \leq i \leq n$.
2. Set $F_{[W_1,\dots,W_n]}(\mathbf{v}) = \mathbf{Pool}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$.

For instance, the function $d_{\mathbf{v}}(\cdot)$ can be $D_f(\cdot\|\mathbf{v})$ for some Bregman divergence D_f and in such a particular case $F_{[W_1,\dots,W_n]}(\mathbf{v})$ first D_f-projects the argument $\mathbf{v}$ into each of $W_1, \dots, W_n$, and then, it "averages" the resulting probability functions by a pooling operator $\mathbf{Pool}$. Hence, the name: an averaging projective procedure. An illustration of F is depicted in Figure 4.

Figure 4. An illustration of an averaging projective procedure F.



Note that $W_1, \dots, W_n$ play dual roles in the definition above, which may perhaps appear clumsy. When they are fixed, $F_{[W_1,\dots,W_n]}$ is a mapping $\mathbb{D}^J \to \mathbb{D}^J$. However, the option to consider them also as variables will be the key to our following investigation and to the applicability of an averaging projective procedure in multi-expert reasoning, where $W_1, \dots, W_n$ will represent the respective knowledge of n experts. A straightforward interpretation is that the first stage simplifies sets to single probability functions, which then are being merged to a final social belief function of the college of experts.

With regard to previous research, the cases of $d_{\mathbf{v}}(\cdot)$ being $\mathrm{KL}(\cdot\|\mathbf{v})$ and $\mathrm{KL}(\mathbf{v}\|\cdot)$ with $\mathbf{Pool}$ be taken to the $\mathbf{LinOp}_{\mathcal{A}}$-pooling operator and the $\mathbf{LogOp}_{\mathcal{A}}$-pooling operator, respectively, were introduced and investigated by Matúš in [21]. The idea of combining the projections by means of the squared Euclidean distance E2 with the $\mathbf{LinOp}_{\mathcal{A}}$-pooling operator was first introduced by Predd *et al.* in [22].

Example 3. *In the definition of an averaging projective procedure, take $d_{\mathbf{v}}$ to be $\mathrm{KL}(\cdot\|\mathbf{v})$ and $\mathbf{Pool}$ to be the $\mathbf{LinOp}_{\mathcal{N}}$-pooling operator. Now, F is the mapping $\mathbb{D}^J \to \mathbb{D}^J$ for every $n \geq 1$ and all closed convex nonempty sets $W_1, \dots, W_n \subseteq \mathbb{D}^J$ given by $F_{[W_1,\dots,W_n]}(\mathbf{v})$ above.*

In particular, take $J = 3$, $n = 2$, $W_1 = \{(x, \frac{1}{2} - x, \frac{1}{2}),\ \frac{1}{10} \leq x \leq \frac{2}{5}\}$, $W_2 = \{(x, \frac{1}{4}, \frac{3}{4} - x),\ \frac{1}{10} \leq x \leq \frac{13}{20}\}$ and $\mathbf{v} = (\frac{1}{3}, \frac{1}{6}, \frac{1}{2})$. Then, the KL*-projection of* $\mathbf{v}$ *into* W_1 *is actually* $\mathbf{v}$ *itself, since* $\mathbf{v} \in W_1$ *and the* KL*-projection of* $\mathbf{v}$ *into* W_2 *is* $(\frac{3}{10}, \frac{1}{4}, \frac{9}{20})$*. Therefore:*

$$F_{[W_1,W_2]}(\mathbf{v}) = \mathbf{LinOp}_{(\frac{1}{2},\frac{1}{2})}\Big(\Big(\frac{1}{3}, \frac{1}{6}, \frac{1}{2}\Big), \Big(\frac{3}{10}, \frac{1}{4}, \frac{9}{20}\Big)\Big) = \Big(\frac{\frac{1}{3} + \frac{3}{10}}{2}, \frac{\frac{1}{6} + \frac{1}{4}}{2}, \frac{\frac{1}{2} + \frac{9}{20}}{2}\Big).$$

2.2. Obdurate Operators

In this section, we approach averaging projective procedures using the framework of probabilistic knowledge merging as defined in [5]. A probabilistic merging operator:

$$\Delta : \underbrace{\mathcal{P}(\mathbb{D}^J) \times \ldots \times \mathcal{P}(\mathbb{D}^J)}_{n} \to \mathcal{P}(\mathbb{D}^J),$$

is a mapping that maps a finite collection of closed convex nonempty subsets of $\mathbb{D}^J$, say $W_1, \ldots, W_n$, to a single closed convex nonempty subset of $\mathbb{D}^J$. In the area of multi-expert reasoning, we can perhaps interpret $\Delta(W_1, \ldots, W_n)$ as a representation of $W_1, \ldots, W_n$, which themselves individually represent knowledge bases of n experts.

A merging operator $\mathbf{O}$ is obdurate if, for every $n \geq 1$ and any $W_1, \ldots, W_n \subseteq \mathbb{D}^J$, we have that $\mathbf{O}(W_1, \ldots, W_n) = \{F_{[W_1,\ldots,W_n]}(\mathbf{v})\}$, where $\mathbf{v}$ is some fixed argument and F is an averaging projective procedure. Note that this operator always produces a singleton. Obdurate processes thus first represent sets as single probability functions, and then, they pool them by a pooling operator.

Although this may sound like a fairly restrictive setting, many existing natural probabilistic merging operators are of this form. The prominent example is the merging operator of Kern-Isberner and Rödder (**KIRP**) [23]. In this particular case, $\mathbf{v}$ is the uniform probability function, $d_{\mathbf{v}}(\cdot)$ is $\mathrm{KL}(\cdot\|\mathbf{v})$ and **Pool** is given by:

$$\mathbf{Pool}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) = \Big(\sum_{k=1}^{n} \frac{H(\mathbf{w}^{(k)})}{\sum_{i=1}^{n} H(\mathbf{w}^{(n)})} w_1^{(k)}, \ldots, \sum_{k=1}^{n} \frac{H(\mathbf{w}^{(k)})}{\sum_{i=1}^{n} H(\mathbf{w}^{(n)})} w_J^{(k)}\Big).$$

Recall that $H(\mathbf{w}^{(i)})$ is the Shannon entropy of $\mathbf{w}^{(i)}$, which is, in fact, the most entropic point in W_i.

In [23], Kern-Isberner and Rödder argue that $W_1, \ldots, W_n \subseteq \mathbb{D}^J$ can by considered as marginal probabilities in a subset $U \subseteq \mathbb{D}^{J+n}$, such that every probability function $\mathbf{v} \in U$ marginalizes to a $\mathbb{D}^J$-probability function belonging to one and only one set W_i. Since then, the point which **KIRP** produces is, in fact, the $\mathbb{D}^J$-marginal of the most entropic point in U, following the justification of the Shannon entropy, they conclude that such a point is a natural interpretation of $W_1, \ldots, W_n$ by a single probability function. **KIRP** thus maps the uniform probability function to the $\mathbb{D}^J$-marginal of the most entropic point in U. To date, **KIRP** has received much attention in the area of probabilistic knowledge merging.

However, any obdurate merging operator seems to be challenged by its violation of the following principle.

(CP) Consistency Principle. Let Δ be a probabilistic merging operator. Then, we say that Δ satisfies the consistency principle if, for every $n \geq 1$ and all $W_1, \ldots, W_n \subseteq \mathbb{D}^J$:

$$\bigcap_{i=1}^{n} W_i \neq \emptyset \text{ implies } \Delta(W_1, \ldots, W_n) \subseteq \bigcap_{i=1}^{n} W_i.$$

(CP) 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 the experts agree on.

This principle often falls under the following philosophical criticism. One might imagine a situation where several experts consider a large set of probability functions as admissible, while one believes 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 the knowledge of the whole group into that single probability function.

More rigorously, Williamson [24] introduces a particular interpretation of the epistemological status of an expert's knowledge base, which he calls "granting". He rejects (CP), as several experts may grant the same piece of knowledge for inconsistent reasons.

On the other hand, Adamčík and Wilmers in [5] assume that the way in which the knowledge was obtained is considered irrelevant, and each expert has incorporated all of his relevant knowledge into what he is declaring, contrary to Williamson's granting. This is sometimes referred to as the principle of total evidence [25] or the Watts assumption [26]. They argue that, although overall knowledge of any human expert can never be fully formalized, as a formalization is always an abstraction from reality, 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 produced by a philosophical opponent challenging one's reasoning using implicit background information, which is not included in the formal representation of a knowledge base.

However, in this paper, we do not wish to probe further into this philosophical argument, and instead, we present the following rather surprising theorem, which appeared for the first time in [10].

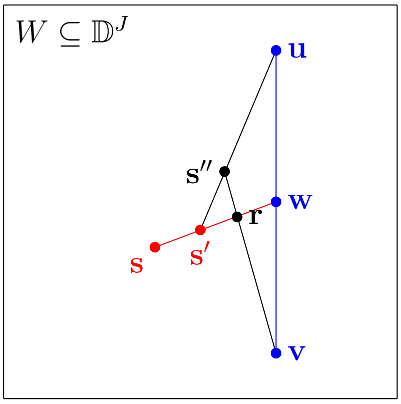
Theorem 5. *There is no obdurate merging operator* $\mathbf{O}$ *that satisfies the consistency principle (CP).*

Proof. Suppose that $J \geq 3$. Let d be the function to minimize from the definition of $\mathbf{O}$, where, for simplicity, we suppress the constant superscript. Let $\mathbf{v} \in \mathbb{D}^J$ be the unique minimizer of d over some sufficiently large closed convex subset W of $\mathbb{D}^J$. Let $\mathbf{w}, \mathbf{u} \in W$ be such that $d(\mathbf{v}) < d(\mathbf{w}) < d(\mathbf{u})$ 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 $\mathrm{s} \in W$ be such that $d(\mathbf{v}) < d(\mathrm{s}) < d(\mathbf{w})$ and s is not a linear combination of $\mathbf{v}$ and $\mathbf{u}$. Then, there is s', such that $\mathrm{s}' = \lambda\mathrm{s} + (1-\lambda)\mathbf{w}$ for some $0 < \lambda \leq 1$, and d is strictly increasing along the line from s' to $\mathbf{w}$. This is because d is strictly convex and $d(\mathrm{s}) < d(\mathbf{w})$. Note that if $J = 2$, then s would be always a linear combination of $\mathbf{v}$ and $\mathbf{u}$. Moreover, for sufficiently large $W \subseteq \mathbb{D}^3$, we can always choose $\mathbf{w}$, $\mathbf{u}$, s and s' in W as above.

Now, we show that d is also strictly increasing along the line from s' to $\mathbf{u}$. Assume this is not the case. Then, by the same argument as before, there is s'', such that $d(\mathrm{s}'') < d(\mathrm{s}')$. Due to the construction, the line from $\mathbf{v}$ to s'' intersects the line from s' to $\mathbf{w}$; let us denote the point of intersection as $\mathbf{r}$. Since d is strictly increasing along the line from s' to $\mathbf{w}$, we have that $d(\mathbf{r}) > d(\mathrm{s}') > d(\mathrm{s}'') > d(\mathbf{v})$. This, however, contradicts the convexity of d. The situation is depicted in Figure 5.

Figure 5. The situation in the proof of Theorem 5.



Now, assume that $W_1 = \{\lambda\mathbf{v} + (1-\lambda)\mathbf{w} : \lambda \in [0,1]\}$, $W_2 = \{\lambda\mathbf{s}' + (1-\lambda)\mathbf{w} : \lambda \in [0,1]\}$, $V_1 = \{\lambda\mathbf{v} + (1-\lambda)\mathbf{u} : \lambda \in [0,1]\}$ and $V_2 = \{\lambda\mathbf{s}' + (1-\lambda)\mathbf{u} : \lambda \in [0,1]\}$. Since $\mathbf{v}$ minimizes d and along the lines from $\mathbf{s}'$ to $\mathbf{w}$ and from $\mathbf{s}'$ to $\mathbf{u}$, the function d is strictly increasing, we have that:

$$\mathbf{O}(W_1, W_2) = \{\mathbf{Pool}(\mathbf{v}, \mathbf{s}')\} = \mathbf{O}(V_1, V_2), \tag{2}$$

where $\mathbf{Pool}$ is a pooling operator used in the second stage of $\mathbf{O}$. Suppose that $\mathbf{O}$ satisfies (CP). Then, $\mathbf{O}(W_1, W_2) = \{\mathbf{w}\}$ and $\mathbf{O}(V_1, V_2) = \{\mathbf{u}\}$, which contradicts Equation (2). □

The theorem above in some philosophical contexts can be used as an argument against the consistency principle, while from another perspective, it casts a shadow on the notion of an obdurate merging operator. This unfortunately includes the natural merging operator $\mathbf{OSEP}$, or obdurate social entropy process, defined as follows. For every $n \geq 1$ and all $W_1, \ldots, W_n \subseteq \mathbb{D}^J$:

$$\mathbf{OSEP}(W_1, \ldots, W_n) = \{\mathbf{LogOp}_{\mathcal{N}}(\mathbf{ME}(W_1), \ldots, \mathbf{ME}(W_n))\}.$$

Recall that $\mathbf{ME}(W_i)$ denotes the most entropic point in W_i or equivalently the KL-projection of the uniform probability function into W_i, and $\mathcal{N}$ is the family of weighting vectors $(\frac{1}{n}, \ldots, \frac{1}{n})$, one for every $n \geq 1$. It is easy to observe that $\mathbf{OSEP}$ is really an obdurate merging operator.

In [10], it is proven that $\mathbf{OSEP}$ is (thus far, the only known) probabilistic merging operator satisfying a particular version of the independence principle, a principle that is an attempt to resurrect the notion of the independence preservation of pooling operators [20] in the context of probabilistic merging operators.

One may say that the reason behind an obdurate merging operator not satisfying (CP) is its "forgetting" nature. In the first stage, it transforms sets $W_1, \ldots, W_n$ into $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ individually without taking into account other sets, thus "forgetting" any existing connections, such as the consistency. However, instead of changing the definition of an averaging projective procedure so as to make it not "forgetting", we will take a different viewpoint on the procedure itself in the following subsection.

2.3. Fixed Points

Our second approach to an averaging projective procedure F consists of considering the set of the fixed points of F. That is, for given $n \geq 1$ and given closed convex nonempty sets $W_1, \ldots, W_n \subseteq \mathbb{D}^J$, we are interested in whether there are any points $\mathbf{v} \in \mathbb{D}^J$, such that:

$$F_{[W_1,\ldots,W_n]}(\mathbf{v}) = \mathbf{v}.$$

Following the convincing justification for combining Bregman projections with the $\mathbf{LinOp}_{\mathcal{A}}$-pooling operator (see Section 1.3), for every convex Bregman divergence D_f and a family of weighting vectors $\mathcal{A}$, we consider here the averaging projective procedure $F^{D_f,\mathcal{A}}$ defined for every $n \geq 1$ and all closed convex nonempty sets $W_1, \ldots, W_n \subseteq \mathbb{D}^J$ by the following.

1. For an argument $\mathbf{v} \in \mathbb{D}^J$, take $\mathbf{w}^{(i)}$ the D_f-projection of $\mathbf{v}$ into W_i for all $1 \leq i \leq n$.
2. Set $F^{D_f,\mathcal{A}}_{[W_1,\ldots,W_n]}(\mathbf{v}) = \mathbf{LinOp}_{\mathbf{a}}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$, where $\mathbf{a} \in \mathcal{A}$.

The restriction to convex Bregman divergences is needed for some later theorems and is adopted *ad hoc*. Therefore, unfortunately, we cannot provide any elaborate justification for it.

Given closed convex nonempty sets $W_1, \ldots, W_n \subseteq \mathbb{D}^J$, we will denote the set of all fixed points of $F^{D_f,\mathcal{A}}$ defined above by $\Theta^{D_f}_{\mathbf{a}}(W_1, \ldots, W_n)$, where $\mathbf{a} \in \mathcal{A}$.

On the other hand, the conjugated parallelogram theorem (Theorem 4), suggesting the combination of the conjugated KL-projection with the **LogOp**-pooling operator, leads us to the consideration of those convex Bregman divergences, which are strictly convex also in the second argument. The squared Euclidean distance and the Kullback–Leibler divergence are instances of such divergences. A fairly general example is a Bregman divergence D_f, such that $f(\mathbf{v}) = \sum_{j=1}^{J} g(v_j)$, where g is a strictly convex function $(0,1) \to \mathbb{R}$, which is three times differentiable, and $g''(v_j) - (w_j - v_j)g'''(v_j) > 0$ for all $1 \leq j \leq J$ and all $\mathbf{w}, \mathbf{v} \in \mathbb{D}^J$ (this is easy to check by the Hessian matrix). Apart from the two divergences mentioned above, this condition is satisfied in particular if $g(v) = v^r$, $2 \geq r > 1$. Note that the Bregman divergence generated by such a function g is also convex in both arguments.

Assuming strict convexity in the second argument of D_f, we can define the conjugated D_f-projection of $\mathbf{v} \in \mathbb{D}^J$ into a closed convex nonempty set $W \subseteq \mathbb{D}^J$ as that unique $\mathbf{w} \in W$ that minimizes $\mathbb{D}_f(\mathbf{v}\|\mathbf{w})$ subject only to $\mathbf{w} \in W$. Moreover, since a sum of strictly convex functions is a strictly convex function, for any $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)} \in \mathbb{D}^J$, there exists a unique minimizer of:

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

which we denote $\mathbf{Pool}^{D_f}_{\mathbf{a}}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$. Thus, for a family of weighting vectors $\mathcal{A}$, we can define the $\mathbf{Pool}^{D_f}_{\mathcal{A}}$-pooling operator. Note that $\mathbf{Pool}^{\mathrm{KL}}_{\mathcal{A}} = \mathbf{LogOp}_{\mathcal{A}}$, $\mathbf{Pool}^{\mathrm{E2}}_{\mathcal{A}} = \mathbf{LinOp}_{\mathcal{A}}$ and that we do not need strict convexity in the second argument in these cases.

Theorem 6 (Conjugated Parallelogram Theorem)**.** *Let D_f be a Bregman divergence, $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}, \mathbf{v} \in \mathbb{D}^J$ and $\mathbf{a} \in \mathbb{D}^n$. Then:*

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

$$+D_f(\mathbf{v}\|\,\mathbf{Pool}_{\mathbf{a}}^{D_f}(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)})).$$

Proof. Let $\mathbf{w} = \mathbf{Pool}_{\mathbf{a}}^{D_f}(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)})$. We need to prove that:

$$\sum_{i=1}^{n} a_i\Big[f(\mathbf{v}) - f(\mathbf{w}^{(i)}) - \sum_{j=1}^{J}(v_j - w_j^{(i)})\frac{\partial f(\mathbf{w}^{(i)})}{\partial w_j^{(i)}}\Big] =$$

$$= \sum_{i=1}^{n} a_i\Big[f(\mathbf{w}) - f(\mathbf{w}^{(i)}) - \sum_{j=1}^{J}(w_j - w_j^{(i)})\frac{\partial f(\mathbf{w}^{(i)})}{\partial w_j^{(i)}}\Big] +$$

$$+\Big[f(\mathbf{v}) - f(\mathbf{w}) - \sum_{j=1}^{J}(v_j - w_j)\frac{\partial f(\mathbf{w})}{\partial w_j}\Big],$$

or equivalently:

$$\sum_{j=1}^{J}(v_j - w_j)\Big(\sum_{i=1}^{n} a_i\frac{\partial f(\mathbf{w}^{(i)})}{\partial w_j^{(i)}} - \frac{\partial f(\mathbf{w})}{\partial w_j}\Big) = 0. \tag{3}$$

Since $\mathbf{w} = \arg\min_{\mathbf{w}\in\mathbb{D}^J}\sum_{i=1}^{n} a_i\mathbb{D}_f(\mathbf{w}\|\mathbf{w}^{(i)})$, differentiation using the Lagrange multiplier method (since a differentiable convex function f is necessarily continuously differentiable (see [9]), the partial derivatives used above are all continuous and the Lagrange multiplier method is permissible) applied to the condition $\sum_{j=1}^{J} w_j = 1$ produces $\sum_{i=1}^{n} a_i\frac{\partial f(\mathbf{w}^{(i)})}{\partial w_j^{(i)}} - \frac{\partial f(\mathbf{w})}{\partial w_j} = \lambda$, $1 \leq j \leq J$, where λ is a constant independent of j. Therefore, Equation (3) is equal to $\sum_{j=1}^{J}(v_j - w_j)\lambda = 0$, and the theorem follows. □

The idea of defining a spectrum of pooling operators where the pooling operators **LinOp** and **LogOp** are special cases was developed previously in a similar manner, but in a slightly different framework of alpha-divergences; *cf.* [27].

Here, following [1,12], we will point out a geometrical relationship between pooling operators **LinOp** and $\mathbf{Pool}^{D_f}$, which will be helpful in illustrating some results of this paper.

Recall that the generator of a Bregman divergence D_f is a strictly convex function $f : (0,1)^J \to \mathbb{R}$, which is differentiable over $\mathbb{D}^J$. Let $\mathbf{w} \in \mathbb{D}^J$. We define $\mathbf{w}^* = \nabla f(\mathbf{w})$. Since f is a strictly convex function, the mapping $\mathbf{w} \to \nabla f(\mathbf{w})$ is injective; thus, the coordinates of $\mathbf{w}^*$ form a coordinate system. There are two kinds of affine structures in $\mathbb{D}^J$. $D_f(\mathbf{w}\|\mathbf{v})$ is convex in $\mathbf{w}$ with respect to the first structure and is convex in $\mathbf{v}^*$ with respect to the second structure.

Therefore, the proof above, in fact, gives $[\mathbf{v}]^* = [\mathbf{Pool}_{\mathbf{a}}^{D_f}(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)})]^* = \mathbf{LinOp}_{\mathbf{a}}([\mathbf{w}^{(1)}]^*,\ldots,[\mathbf{w}^{(n)}]^*) + \mathbf{c}$, where $\mathbf{c} = \underbrace{(\lambda,\ldots,\lambda)}_{J\text{-times}}$ is a normalizing vector induced by $\sum_{j=1}^{J} v_j = 1$.

The only other type of averaging projective procedure $\hat{F}^{D_f,\mathcal{A}}$ that we consider here will be generated by a convex differentiable Bregman divergence D_f, which is strictly convex in its second argument, and a family of weight $\mathcal{A}$ and is defined for every $n \geq 1$ and all closed convex nonempty sets $W_1,\ldots,W_n \subseteq \mathbb{D}^J$ by the following.

1. For an argument $\mathbf{v} \in \mathbb{D}^J$, take $\mathbf{w}^{(i)}$ the conjugated D_f-projection of $\mathbf{v}$ into W_i for all $1 \leq i \leq n$.
2. Set $\hat{F}^{D_f,\mathcal{A}}_{[W_1,\ldots,W_n]}(\mathbf{v}) = \mathbf{Pool}_{\mathbf{a}}^{D_f}(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)})$, where $\mathbf{a} \in \mathcal{A}$.

Given closed convex nonempty sets $W_1, \ldots, W_n \subseteq \mathbb{D}^J$, we will denote the set of all fixed points of $\hat{F}^{D_f,\mathcal{A}}$ defined above by $\hat{\Theta}_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$, where $\mathbf{a} \in \mathcal{A}$.

Note that we always require an additional assumption of D_f being differentiable for this type of averaging projective procedure. This assumption is essential to the proofs of some results concerning this procedure. We note that both divergences KL and E2 are differentiable.

Given a family of weighting vectors $\mathcal{A}$, our aim is to investigate $\Theta_{\mathcal{A}}^{D_f} = \{\Theta_{\mathbf{a}}^{D_f} : \mathbf{a} \in \mathcal{A}\}$ and $\hat{\Theta}_{\mathcal{A}}^{D_f} = \{\hat{\Theta}_{\mathbf{a}}^{D_f} : \mathbf{a} \in \mathcal{A}\}$ as operators acting on $\mathcal{P}(\mathbb{D}^J) \times \ldots \times \mathcal{P}(\mathbb{D}^J)$. In particular, we ask the following questions. Given any closed convex nonempty sets $W_1, \ldots, W_n \subseteq \mathbb{D}^J$ and $\mathbf{a} \in \mathcal{A}$:

- Are $\Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$ and $\hat{\Theta}_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$ always nonempty?
- Are these sets always closed and convex?

If both answers are positive, then we can consider $\Theta_{\mathcal{A}}^{D_f}$ and $\hat{\Theta}_{\mathcal{A}}^{D_f}$ as probabilistic merging operators. In such a case, the following question makes sense.

- As probabilistic merging operators, do they satisfy the consistency principle (CP)?

The fact that the answer to all three questions is "yes" is perhaps surprising, given that the much simpler obdurate merging operators do not satisfy (CP). We prove the above results in the following sequence of theorems, which conclude Section 2.

The following well-known lemma is a simple, but useful observation.

Lemma 1. *Let D_f be a Bregman divergence and $\mathbf{a}, \mathbf{v}, \mathbf{w} \in \mathbb{D}^J$. Then:*

$$D_f(\mathbf{a}\|\mathbf{v}) - D_f(\mathbf{a}\|\mathbf{w}) - D_f(\mathbf{w}\|\mathbf{v}) = (\mathbf{a} - \mathbf{w}) \cdot \Big(\nabla f(\mathbf{w}) - \nabla f(\mathbf{v})\Big).$$

Theorem 7. *Let D_f be a convex Bregman divergence, $W_1, \ldots, W_n \subseteq \mathbb{D}^J$ be closed convex nonempty sets and $\mathbf{a} \in \mathbb{D}^n$. Let $\mathbf{v}, \mathbf{w} \in \mathbb{D}^J$, $\mathbf{u}^{(1)} \in W_1, \ldots, \mathbf{u}^{(n)} \in W_n$ and $\mathbf{w}^{(1)} \in W_1, \ldots, \mathbf{w}^{(n)} \in W_n$ be such that $\mathbf{v} = \mathbf{LinOp}_{\mathbf{a}}(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)})$, $\mathbf{w} = \mathbf{LinOp}_{\mathbf{a}}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$ and $\mathbf{u}^{(i)}$ are the D_f-projection of $\mathbf{v}$ into W_i, $1 \leq i \leq n$. Then:*

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

Proof. First of all, by the extended Pythagorean property, we have that:

$$D_f(\mathbf{w}^{(i)}\|\mathbf{v}) - D_f(\mathbf{u}^{(i)}\|\mathbf{v}) - D_f(\mathbf{w}^{(i)}\|\mathbf{u}^{(i)}) \geq 0.$$

By the parallelogram theorem:

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

Hence:

$$\sum_{i=1}^{n} a_i D_f(\mathbf{w}^{(i)}\|\mathbf{w}) - \sum_{i=1}^{n} a_i D_f(\mathbf{u}^{(i)}\|\mathbf{v}) + D_f(\mathbf{w}\|\mathbf{v}) - \\ - \sum_{i=1}^{n} a_i D_f(\mathbf{w}^{(i)}\|\mathbf{u}^{(i)}) \geq 0. \tag{4}$$

Since we assume that $D_f(\cdot\|\cdot)$ is a convex function in both arguments by the Jensen inequality:

$$D_f(\mathbf{w}\|\mathbf{v}) - \sum_{i=1}^{n} a_i D_f(\mathbf{w}^{(i)}\|\mathbf{u}^{(i)}) \leq 0. \tag{5}$$

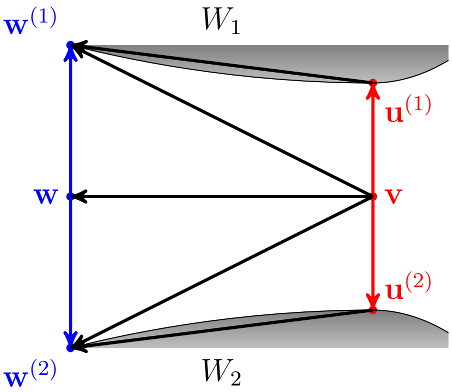
The Inequalities (4) and (5) give:

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

as required. □

Figure 6 depicts the situation in the proof above for $n = 2$. Arrows indicate corresponding divergences.

Figure 6. The situation in the proof of Theorem 7 for $n = 2$.

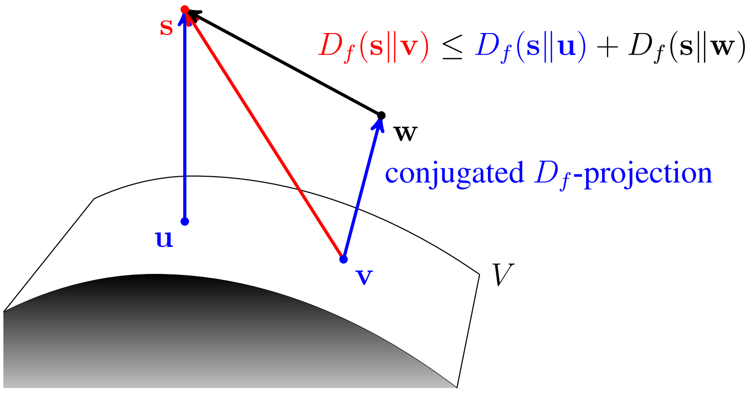


An interesting question related to conjugated Bregman projections arises as to whether a similar property to the Pythagorean property holds. It turns out that the corresponding property is the so-called four-point property, from to Csiszár and Tusnády. The following theorem in the case of the KL-divergence is a specific instance of a result in [28], Lemma 3, but the formulation using the term "conjugated KL-projection" first appeared in [21]. An illustration is depicted in Figure 7.

Theorem 8 (Four-Point Property)**.** *Let D_f be a convex differentiable Bregman divergence, which is strictly convex in its second argument. Let V be a convex closed nonempty subset of $\mathbb{D}^J$, and let $\mathbf{v}, \mathbf{u}, \mathbf{w}, \mathbf{s} \in \mathbb{D}^J$ be such that $\mathbf{v}$ is the conjugated D_f-projection of $\mathbf{w}$ into V and $\mathbf{u} \in V$ is arbitrary. Then:*

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

Figure 7. The illustration of the four-point property.



$\mathbf{s}$
$D_f(\mathbf{s}\|\mathbf{v}) \leq D_f(\mathbf{s}\|\mathbf{u}) + D_f(\mathbf{s}\|\mathbf{w})$
$\mathbf{w}$
conjugated D_f-projection
$\mathbf{u}$
$\mathbf{v}$
V

Proof. By Lemma 1, we have that:

$$D_f(\mathbf{s}\|\mathbf{w}) = D_f(\mathbf{s}\|\mathbf{v}) - D_f(\mathbf{w}\|\mathbf{v}) - (\mathbf{s}-\mathbf{w})\cdot(\nabla f(\mathbf{w}) - \nabla f(\mathbf{v})).$$

We can rewrite the above as:

$$\begin{aligned} D_f(\mathbf{s}\|\mathbf{w}) - D_f(\mathbf{s}\|\mathbf{v}) + D_f(\mathbf{s}\|\mathbf{u}) &= \\ = D_f(\mathbf{s}\|\mathbf{u}) - D_f(\mathbf{w}\|\mathbf{v}) - (\mathbf{s}-\mathbf{w})\cdot(\nabla f(\mathbf{w}) - \nabla f(\mathbf{v})). \end{aligned} \tag{6}$$

Since $D_f(\cdot\|\cdot)$ is a convex differentiable function, by applying the first convexity condition twice, we have that:

$$\begin{aligned} D_f(\mathbf{s}\|\mathbf{u}) \geq D_f(\mathbf{w}\|\mathbf{v}) + \\ + \sum_{j=1}^{J}(a_j - w_j)\frac{\partial}{\partial x_j}\Big[D_f(\mathbf{x}\|\mathbf{v})\Big]\Big|_{\mathbf{x}=\mathbf{w}} \\ + \sum_{j=1}^{J}(u_j - v_j)\frac{\partial}{\partial x_j}\Big[D_f(\mathbf{w}\|\mathbf{x})\Big]\Big|_{\mathbf{x}=\mathbf{v}}. \end{aligned} \tag{7}$$

Expressions (6) and (7) give that:

$$\begin{aligned} D_f(\mathbf{s}\|\mathbf{v}) \leq D_f(\mathbf{s}\|\mathbf{u}) + D_f(\mathbf{s}\|\mathbf{w}) - \\ - \sum_{j=1}^{J}(u_j - v_j)\frac{\partial}{\partial x_j}\Big[D_f(\mathbf{w}\|\mathbf{x})\Big]\Big|_{\mathbf{x}=\mathbf{v}}. \end{aligned}$$

However, since $\mathbf{v}$ is the conjugated D_f-projection of $\mathbf{w}$ into V, the gradient of $D_f(\mathbf{w}\|\cdot)$ at $(\mathbf{w},\mathbf{v})$ in the direction to $(\mathbf{w},\mathbf{u})$ must be greater than or equal to zero:

$$\sum_{j=1}^{J}(u_j - v_j)\frac{\partial}{\partial x_j}\Big[D_f(\mathbf{w}\|\mathbf{x})\Big]\Big|_{\mathbf{x}=\mathbf{v}} \geq 0$$

and the theorem follows. □

The following result appeared for the first time in [10], but without considering the weighting.

Theorem 9 (Characterization Theorem for $\Theta_{\mathbf{a}}^{D_f}$)**.** *Let D_f be a convex Bregman divergence, $\mathbf{a} \in \mathbb{D}^n$ and $W_1, \ldots, W_n \subseteq \mathbb{D}^J$ be closed convex nonempty sets. Then:*

$$\Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n) = \Big\{ \arg\min_{\mathbf{v}\in\mathbb{D}^J} \sum_{i=1}^{n} a_i D_f(\mathbf{w}^{(i)}\|\mathbf{v}) : \ \mathbf{w}^{(i)} \in W_i,\, 1 \leq i \leq n \Big\},$$

where the right hand-side denotes the set of all possible minimizers. That is the set of all probability functions $\mathbf{v} \in \mathbb{D}^J$, which globally minimize $\sum_{i=1}^{n} a_i D_f(\mathbf{w}^{(i)}\|\mathbf{v})$, subject only to $\mathbf{w}^{(1)} \in W_1, \ldots,$ $\mathbf{w}^{(n)} \in W_n$.

Proof. It is easy to see that, given closed convex nonempty sets $W_1, \ldots, W_n \subseteq \mathbb{D}^J$, we have that those $\mathbf{w}^{(1)} \in W_1, \ldots, \mathbf{w}^{(n)} \in W_n$, which together with $\mathbf{v} \in \mathbb{D}^J$, globally minimize:

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

are also the D_f-projections of $\mathbf{v}$ into $W_1, \ldots, W_n$ respectively. This, together with Equation (1) (the equation preceding Theorem 2), gives:

$$\Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n) \supseteq \Big\{ \arg\min_{\mathbf{v} \in \mathbb{D}^J} \sum_{i=1}^{n} a_i D_f(\mathbf{w}^{(i)} \| \mathbf{v}) : \mathbf{w}^{(i)} \in W_i, 1 \le i \le n \Big\}.$$

Now, assume that $\mathbf{v} \in \Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$ and

$$\mathbf{u} \in \Big\{ \arg\min_{\mathbf{v} \in \mathbb{D}^J} \sum_{i=1}^{n} a_i D_f(\mathbf{w}^{(i)} \| \mathbf{v}) : \mathbf{w}^{(i)} \in W_i, 1 \le i \le n \Big\}.$$

Let us denote the D_f-projections of $\mathbf{v}$ into $W_1, \ldots, W_n$ by $\mathbf{w}^{(1)} \ldots, \mathbf{w}^{(n)}$, respectively. Accordingly, let us denote the D_f-projections of $\mathbf{u}$ into $W_1, \ldots, W_n$ by $\mathbf{r}^{(1)} \ldots, \mathbf{r}^{(n)}$, respectively. Suppose that $\sum_{i=1}^{n} a_i D_f(\mathbf{w}^{(i)} \| \mathbf{v}) > \sum_{i=1}^{n} a_i D_f(\mathbf{r}^{(i)} \| \mathbf{u})$, *i.e.*,

$$\mathbf{v} \notin \Big\{ \arg\min_{\mathbf{v} \in \mathbb{D}^J} \sum_{i=1}^{n} a_i D_f(\mathbf{w}^{(i)} \| \mathbf{v}) : \mathbf{w}^{(i)} \in W_i, 1 \le i \le n \Big\}.$$

This contradicts Theorem 7, and therefore:

$$\Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n) \subseteq \Big\{ \arg\min_{\mathbf{v} \in \mathbb{D}^J} \sum_{i=1}^{n} a_i D_f(\mathbf{w}^{(i)} \| \mathbf{v}) : \mathbf{w}^{(i)} \in W_i, 1 \le i \le n \Big\}.$$

□

Let us now deviate for a while from the goals of this subsection and stress the importance of the restriction to the positive discrete probability functions, which was detailed in Section 1.1. The problem with the KL-divergence is that the function $f(\mathbf{x}) = \sum_{j=1}^{J} x_j \log x_j$ is not differentiable if some $x_j = 0$. Without the adopted restriction, the KL-divergence is therefore usually defined by:

$$\mathrm{KL}(\mathbf{w} \| \mathbf{v}) = \begin{cases} \sum_{j:\, v_j \neq 0} w_j \log \frac{w_j}{v_j}, & \text{if } v_j = 0 \text{ implies } w_j = 0 \text{ for all } 1 \le j \le J, \\ +\infty, & \text{otherwise.} \end{cases}$$

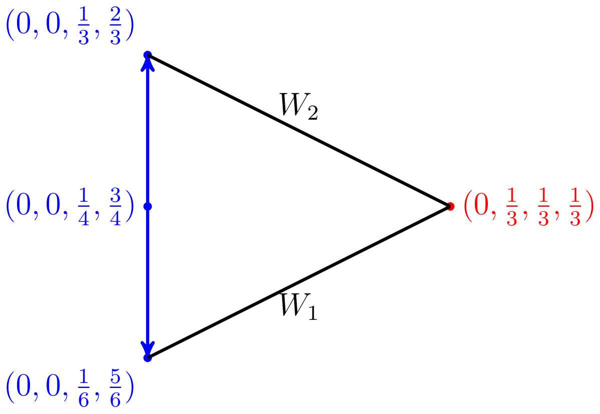
If $v_j = 0$ implies $w_j = 0$ for all $1 \le j \le J$, we say that $\mathbf{v}$ dominates $\mathbf{w}$ and write $\mathbf{v} \gg \mathbf{w}$.

The first problem we would face with this definition is whether the notion of the KL-projection makes sense. For given $\mathbf{v} \in \mathbb{D}^J$ and closed convex nonempty set $W \subseteq \mathbb{D}^J$, the KL-projection of $\mathbf{v}$ into W makes sense only if there is at least one $\mathbf{w} \in W$, such that $\mathbf{v} \gg \mathbf{w}$.

However, even if adding this condition to all of the discussion concerning the KL-projection above (this is perfectly possible, as seen in [10]), Theorem 9 still could not hold, as the following example demonstrates.

Example 4. *Let* $W_1 = \{\lambda(0, 0, \frac{1}{6}, \frac{5}{6}) + (1 - \lambda)(0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}) : \lambda \in [0, 1]\}$ *and* $W_2 = \{\lambda(0, 0, \frac{1}{3}, \frac{2}{3}) + (1 - \lambda)(0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}) : \lambda \in [0, 1]\}$. *Assume that* $\mathbf{a} = (\frac{1}{2}, \frac{1}{2})$. *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 of global minimizers* $\mathbf{v}$ *of* $\mathrm{KL}(\mathbf{w}^{(1)} \| \mathbf{v}) + \mathrm{KL}(\mathbf{w}^{(2)} \| \mathbf{v})$ *subject to* $\mathbf{w}^{(1)} \in W_1$ *and* $\mathbf{w}^{(2)} \in W_2$. *An illustration is depicted in Figure 8.*

Figure 8. The illustration of Example 4.


$(0,0,\frac{1}{3},\frac{2}{3})$
W_2
$(0,0,\frac{1}{4},\frac{3}{4})$
$(0,\frac{1}{3},\frac{1}{3},\frac{1}{3})$
W_1
$(0,0,\frac{1}{6},\frac{5}{6})$

Moreover, some variant of the above example would show that the set $\Theta_{\mathbf{a}}^{\mathrm{KL}}(W_1, W_2)$ is not convex, which would wreck our aims; more details are given in [10].

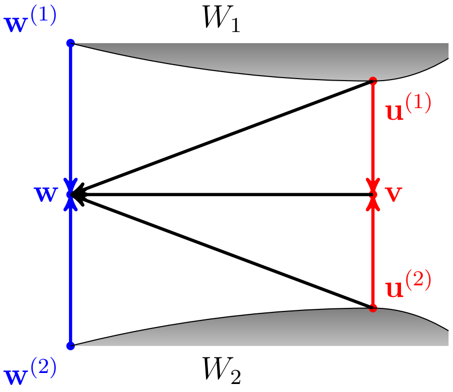
On the other hand, neither of those Bregman divergences, which generate functions, are differentiable over the whole space of discrete probability functions (e.g., the squared Euclidean distance) and would encounter the difficulties of the KL-divergence. In particular, Theorem 9 formulated over the whole space of discrete probability functions (as opposed to only the positive ones) would still hold for such Bregman divergences.

Now, we shall go back and prove a theorem similar to Theorem 9 for the $\hat{\Theta}_{\mathcal{A}}^{D_f}$-operator. In order to do that, we will need the following analogue of Theorem 7.

Theorem 10. *Let D_f be a convex differentiable Bregman divergence, which is strictly convex in its second argument, and let $W_1, \ldots, W_n \subseteq \mathbb{D}^J$ be closed convex nonempty sets and $\mathbf{a} \in \mathbb{D}^n$. Let $\mathbf{v}, \mathbf{w} \in \mathbb{D}^J$ and $\mathbf{u}^{(1)} \in W_1, \ldots, \mathbf{u}^{(n)} \in W_n$ and $\mathbf{w}^{(1)} \in W_1, \ldots, \mathbf{w}^{(n)} \in W_n$ be such that $\mathbf{v} = \mathbf{Pool}_{\mathbf{a}}^{D_f}(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)})$, $\mathbf{w} = \mathbf{Pool}_{\mathbf{a}}^{D_f}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$ and $\mathbf{u}^{(i)}$ are the conjugated D_f-projection of $\mathbf{v}$ into W_i, $1 \le i \le n$. Then:*

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

Figure 9. The situation in the proof of Theorem 10 for $n = 2$.



Proof. By Theorem 6, we have that:

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

which by the four-point property (notice that we need the differentiability of D_f to employ the four-point property) (Theorem 8) becomes:

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

and hence:

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

as required, see Figure 9. □

The theorem above is fairly similar to Theorem 7. Let us use the dual affine structure in $\mathbb{D}^J$ defined after the proof of Theorem 6 to analyze this more closely. For $W \subset \mathbb{D}^J$, define $W^* = \{\mathbf{w}^*;\ \mathbf{w} \in W\}$ and define the dual divergence D_f^* to the divergence D_f by $D_f^*(\mathbf{v}^*\|\mathbf{w}^*) = D_f(\mathbf{w}\|\mathbf{v})$. Since, by Theorem 6, we have that $[\mathbf{v}]^* = [\mathbf{Pool}_{\mathbf{a}}^{D_f}(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)})]^* = \mathbf{LinOp}_{\mathbf{a}}([\mathbf{w}^{(1)}]^*,\ldots,[\mathbf{w}^{(n)}]^*) + \mathbf{c}_{\mathbf{v}}$, where $\mathbf{c}_{\mathbf{v}} = \underbrace{(\lambda,\ldots,\lambda)}_{J\text{-times}}$ is a normalizing vector induced by $\sum_{j=1}^{J} v_j = 1$, the theorem above can be rewritten as follows.

Let D_f be a convex differentiable Bregman divergence, which is strictly convex in its second argument, and let $W_1,\ldots,W_n \subseteq \mathbb{D}^J$ be closed convex nonempty sets and $\mathbf{a} \in \mathbb{D}^n$. Let $\mathbf{v},\mathbf{w} \in \mathbb{D}^J$, $\mathbf{u}^{(1)} \in W_1,\ldots,\mathbf{u}^{(n)} \in W_n$ and $\mathbf{w}^{(1)} \in W_1,\ldots,\mathbf{w}^{(n)} \in W_n$ be such that $\mathbf{v}^* = \mathbf{LinOp}_{\mathbf{a}}([\mathbf{u}^{(1)}]^*,\ldots,[\mathbf{u}^{(n)}]^*) + \mathbf{c}_{\mathbf{v}}$, $\mathbf{w}^* = \mathbf{LinOp}_{\mathbf{a}}([\mathbf{w}^{(1)}]^*\ldots,[\mathbf{w}^{(n)}]^*) + \mathbf{c}_{\mathbf{w}}$ and $[\mathbf{u}^{(i)}]^*$ are the D_f^*-projection of $\mathbf{v}^*$ into W_i^*, $1 \leq i \leq n$. Then:

$$\sum_{i=1}^{n} a_i D_f^*([\mathbf{u}^{(i)}]^*\|\mathbf{v}^*) \leq \sum_{i=1}^{n} a_i D_f^*([\mathbf{w}^{(i)}]^*\|\mathbf{w}^*).$$

This illustrates that if D_f is a convex differentiable Bregman divergence that is strictly convex in its second argument, then Theorems 7 and 10 are dual with respect to *.

Theorem 11 (Characterization Theorem for $\hat{\Theta}_{\mathbf{a}}^{D_f}$)**.** *Let D_f be a convex differentiable Bregman divergence, which is strictly convex in its second argument, and let $W_1,\ldots,W_n \subseteq \mathbb{D}^J$ be closed convex nonempty sets and $\mathbf{a} \in \mathbb{D}^n$. Then:*

$$\hat{\Theta}_{\mathbf{a}}^{D_f}(W_1,\ldots,W_n) = \Big\{ \arg\min_{\mathbf{v}\in\mathbb{D}^J} \sum_{i=1}^{n} a_i D_f(\mathbf{v}\|\mathbf{w}^{(i)}) :\ \mathbf{w}^{(i)} \in W_i,\ 1 \leq i \leq n \Big\},$$

where the right hand-side denotes the set of all possible minimizers.

Proof. The proof is similar to the proof of Theorem 9. First, given closed convex nonempty sets $W_1,\ldots,W_n \subseteq \mathbb{D}^J$, we have that those $\mathbf{w}^{(1)} \in W_1,\ldots,\mathbf{w}^{(n)} \in W_n$, which together with $\mathbf{v} \in \mathbb{D}^J$, globally minimize:

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

that are also the conjugated D_f-projections of $\mathbf{v}$ into $W_1, \dots, W_n$, respectively. This together with the definition of $\mathbf{Pool}_{\mathbf{a}}^{D_f}$ gives:

$$\hat{\Theta}_{\mathbf{a}}^{D_f}(W_1, \dots, W_n) \supseteq \Big\{ \arg\min_{\mathbf{v}\in\mathbb{D}^J} \sum_{i=1}^{n} a_i D_f(\mathbf{v}\|\mathbf{w}^{(i)}) : \; \mathbf{w}^{(i)} \in W_i, 1 \leq i \leq n \Big\}.$$

Second, assume that $\mathbf{v} \in \hat{\Theta}_{\mathbf{a}}^{D_f}(W_1, \dots, W_n)$ and:

$$\mathbf{u} \in \Big\{ \arg\min_{\mathbf{v}\in\mathbb{D}^J} \sum_{i=1}^{n} a_i D_f(\mathbf{v}\|\mathbf{w}^{(i)}) : \; \mathbf{w}^{(i)} \in W_i, 1 \leq i \leq n \Big\}.$$

Let us denote the conjugated D_f-projections of $\mathbf{v}$ into $W_1, \dots, W_n$ by $\mathbf{w}^{(1)} \dots, \mathbf{w}^{(n)}$, respectively. Accordingly, let us denote the conjugated D_f-projections of $\mathbf{u}$ into $W_1, \dots, W_n$ by $\mathbf{r}^{(1)} \dots, \mathbf{r}^{(n)}$, respectively. Suppose that:

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

i.e., $\mathbf{v} \notin \Big\{ \arg\min_{\mathbf{v}\in\mathbb{D}^J} \sum_{i=1}^{n} a_i D_f(\mathbf{v}\|\mathbf{w}^{(i)}) : \; \mathbf{w}^{(i)} \in W_i, 1 \leq i \leq n \Big\}$. This contradicts Theorem 10, and therefore:

$$\hat{\Theta}_{\mathbf{a}}^{D_f}(W_1, \dots, W_n) \subseteq \Big\{ \arg\min_{\mathbf{v}\in\mathbb{D}^J} \sum_{i=1}^{n} a_i D_f(\mathbf{v}\|\mathbf{w}^{(i)}) : \; \mathbf{w}^{(i)} \in W_i, 1 \leq i \leq n \Big\}.$$

□

The following simple observation originally from [10] based on Equation (1) (alternatively on the parallelogram theorem) will be used in the proof of the forthcoming theorem.

Lemma 2. *Let D_f be a convex Bregman divergence and $\mathbf{a} \in \mathbb{D}^n$. Then, the following are equivalent:*

1. *The probability functions $\mathbf{v}, \mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)} \in \mathbb{D}^J$ minimize the quantity:*

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

   subject to $\mathbf{w}^{(1)} \in W_1, \dots, \mathbf{w}^{(n)} \in W_n$.

2. *The probability functions $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)} \in \mathbb{D}^J$ minimize the quantity:*

$$\sum_{i=1}^{n} a_i D_f(\mathbf{w}^{(i)}\| \mathbf{LinOp}_{\mathbf{a}}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$$

   subject to $\mathbf{w}^{(1)} \in W_1, \dots, \mathbf{w}^{(n)} \in W_n$ and $\mathbf{v} = \mathbf{LinOp}_{\mathbf{a}}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$.

Theorem 12. *Let D_f be a convex Bregman divergence. Then, for all nonempty closed convex sets $W_1, \dots, W_n \subseteq \mathbb{D}^J$ and $\mathbf{a} \in \mathbb{D}^n$, the set $\Big\{ \arg\min_{\mathbf{v}\in\mathbb{D}^J} \sum_{i=1}^{n} a_i D_f(\mathbf{w}^{(i)}\|\mathbf{v}) : \; \mathbf{w}^{(i)} \in W_i, 1 \leq i \leq n \Big\}$ is a nonempty closed convex region of $\mathbb{D}^J$.*

Proof. This proof is from [10]. Let $\mathbf{v}, \mathbf{s} \in \left\{ \arg\min_{\mathbf{v}\in\mathbb{D}^J} \sum_{i=1}^n a_i D_f(\mathbf{w}^{(i)}\|\mathbf{v}) : \mathbf{w}^{(i)} \in W_i, 1 \leq i \leq n \right\}$, as the set is clearly nonempty. For convexity, we need to show that $\lambda\mathbf{v} + (1-\lambda)\mathbf{s} \in \left\{ \arg\min_{\mathbf{v}\in\mathbb{D}^J} \sum_{i=1}^n a_i D_f(\mathbf{w}^{(i)}\|\mathbf{v}) : \mathbf{w}^{(i)} \in W_i, 1 \leq i \leq n \right\}$ for any $\lambda \in [0,1]$.

Assume that $\mathbf{w}^{(1)} \in W_1, \dots, \mathbf{w}^{(n)} \in W_n$ are such that $\mathbf{v} = \mathbf{LinOp_a}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$ and $\mathbf{u}^{(1)} \in W_1, \dots, \mathbf{u}^{(n)} \in W_n$ are such that $\mathbf{s} = \mathbf{LinOp_a}(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(n)})$. It is easy to observe that the convexity of $D_f(\cdot\|\cdot)$ implies convexity of:

$$g(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) = \sum_{i=1}^n a_i D_f(\mathbf{x}^{(i)}\| \, \mathbf{LinOp_a}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}))$$

over the convex region specified by constraints $\mathbf{x}^{(i)} \in W_i$, $1 \leq i \leq n$. Moreover, the function g attains its minimum over this convex region at points $(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$ and $(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(n)})$. We need to show that g also attains its minimum at the point:

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

for any $\lambda \in [0,1]$. Since g is convex by the Jensen inequality, we have that:

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

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

Since $g(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}) = g(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(n)})$, the inequality above can only hold with equality, and therefore, by Lemma 2,

$$\lambda\mathbf{v} + (1-\lambda)\mathbf{s} \in \left\{ \arg\min_{\mathbf{v}\in\mathbb{D}^J} \sum_{i=1}^n a_i D_f(\mathbf{w}^{(i)}\|\mathbf{v}) : \mathbf{w}^{(i)} \in W_i, 1 \leq i \leq n \right\}$$

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

Moreover, since convexity implies continuity, the minimization of a convex function over a closed convex region produces a closed convex set. Therefore, the fact that $W_1, \dots, W_n$ are all closed and convex implies that the set of n-tuples $(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$, which are global minimizers of g over the region specified by $\mathbf{w}^{(i)} \in W_i$, $1 \leq i \leq n$, is closed. Additionally, since closed regions are preserved by projections in the Euclidean space, the set given by $\mathbf{LinOp_a}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$ is closed, as well. □

The following observation immediately follows by the definition of $\mathbf{Pool}_\mathbf{a}^{D_f}$.

Lemma 3. *Let D_f be a convex Bregman divergence and $\mathbf{a} \in \mathbb{D}^n$. Then, the following are equivalent:*

1. *The probability functions $\mathbf{v}, \mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)} \in \mathbb{D}^J$ minimize the quantity:*

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

subject to $\mathbf{w}^{(1)} \in W_1, \dots, \mathbf{w}^{(n)} \in W_n$.

2. *The probability functions* $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)} \in \mathbb{D}^J$ *minimize the quantity:*

$$\sum_{i=1}^{n} a_i D_f(\mathbf{Pool}_{\mathbf{a}}^{D_f}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \| \mathbf{w}^{(i)})$$

subject to $\mathbf{w}^{(1)} \in W_1, \ldots, \mathbf{w}^{(n)} \in W_n$ *and* $\mathbf{v} = \mathbf{Pool}_{\mathbf{a}}^{D_f}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$.

Theorem 13. *Let* D_f *be a convex Bregman divergence. Then, for all nonempty closed convex sets* $W_1, \ldots, W_n \subseteq \mathbb{D}^J$ *and* $\mathbf{a} \in \mathbb{D}^n$*, the set* $\left\{ \arg\min_{\mathbf{v} \in \mathbb{D}^J} \sum_{i=1}^{n} a_i D_f(\mathbf{v} \| \mathbf{w}^{(i)}) : \ \mathbf{w}^{(i)} \in W_i,\, 1 \leq i \leq n \right\}$ *is a nonempty closed convex region of* $\mathbb{D}^J$*.*

Proof. Let $\mathbf{v}, \mathbf{s} \in \left\{ \arg\min_{\mathbf{v} \in \mathbb{D}^J} \sum_{i=1}^{n} a_i D_f(\mathbf{v} \| \mathbf{w}^{(i)}) : \ \mathbf{w}^{(i)} \in W_i,\, 1 \leq i \leq n \right\}$, as the set is clearly nonempty. For convexity, we need to show that $\lambda \mathbf{v} + (1-\lambda)\mathbf{s} \in \left\{ \arg\min_{\mathbf{v} \in \mathbb{D}^J} \sum_{i=1}^{n} a_i D_f(\mathbf{v} \| \mathbf{w}^{(i)}) : \mathbf{w}^{(i)} \in W_i,\, 1 \leq i \leq n \right\}$ for any $\lambda \in [0, 1]$.

Assume that $\mathbf{w}^{(1)} \in W_1, \ldots, \mathbf{w}^{(n)} \in W_n$ are such that $\mathbf{v} = \mathbf{Pool}_{\mathbf{a}}^{D_f}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$ and $\mathbf{u}^{(1)} \in W_1, \ldots, \mathbf{u}^{(n)} \in W_n$ are such that $\mathbf{s} = \mathbf{Pool}_{\mathbf{a}}^{D_f}(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)})$. Now, for any $\lambda \in [0, 1]$,

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

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

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

where the first inequality follows by convexity of $D_f(\cdot \| \cdot)$ and the second by the definition of $\mathbf{Pool}_{\mathbf{a}}^{D_f}$ as the unique minimizer. However, the inequality above can only hold with equality and, by Lemma 3,

$$\lambda \mathbf{v} + (1-\lambda)\mathbf{s} \in \left\{ \arg\min_{\mathbf{v} \in \mathbb{D}^J} \sum_{i=1}^{n} a_i D_f(\mathbf{w}^{(i)} \| \mathbf{v}) : \ \mathbf{w}^{(i)} \in W_i,\, 1 \leq i \leq n \right\}$$

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

Moreover, since convexity implies continuity, the minimization of a convex function over a closed convex region produces a closed convex set. Therefore, the fact that $W_1, \ldots, W_n$ are all closed and convex implies that the set of n-tuples $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$, which are global minimizers of $\sum_{i=1}^{n} a_i D_f(\mathbf{Pool}_{\mathbf{a}}^{D_f}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \| \mathbf{w}^{(i)})$ over the region specified by $\mathbf{w}^{(i)} \in W_i$, $1 \leq i \leq n$, is closed. Additionally, since closed regions are preserved by projections in the Euclidean space, the set given by $\mathbf{Pool}_{\mathbf{a}}^{D_f}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$ is closed, as well. □

Finally, we can establish our initial claims:

Theorem 14. *Let* $\mathcal{A}$ *be a family of weighting vectors. The operator* $\Theta_{\mathcal{A}}^{D_f}$*, where* D_f *is a convex Bregman divergence, and the operator* $\hat{\Theta}_{\mathcal{A}}^{D_f}$*, where* D_f *is a convex differentiable Bregman divergence, which is strictly convex in its second argument, are well defined probabilistic merging operators that satisfy (CP).*

Proof. First, the fact that $\Theta_{\mathcal{A}}^{D_f}$ is well defined as a probabilistic merging operator follows Theorems 9 and 12. Accordingly, $\hat{\Theta}_{\mathcal{A}}^{D_f}$ is a well-defined probabilistic merging operator by Theorems 11 and 13.

Second, let $\mathbf{a} \in \mathcal{A}$ (in particular $\mathbf{a} \in \mathbb{D}^n$) and $W_1, \ldots, W_n \subseteq \mathbb{D}^J$ be closed, convex, nonempty and have a nonempty intersection. Clearly, every point in that intersection minimizes $\sum_{i=1}^n a_i D_f(\mathbf{w}^{(i)} \| \mathbf{v})$ and $\sum_{i=1}^n a_i D_f(\mathbf{v} \| \mathbf{w}^{(i)})$ subject to $\mathbf{w}^{(1)} \in W_1, \ldots, \mathbf{w}^{(n)} \in W_n$ with both expressions attaining the zero value. Since $D_f(\mathbf{w} \| \mathbf{v}) = 0$ only if $\mathbf{w} = \mathbf{v}$, those points in the intersection are the only points minimizing the above quantities. □

It turns out that, given closed convex nonempty sets $W_1, \ldots, W_n \subseteq \mathbb{D}^J$ and weighting $\mathbf{a}$, the sets of fixed points $\Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$ and $\hat{\Theta}_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$ posses attractive properties, which make the operators $\Theta_{\mathcal{A}}^{D_f}$ and $\hat{\Theta}_{\mathcal{A}}^{D_f}$ suitable for probabilistic merging. The following example taken from [10] illustrates a possible philosophical justification for considering the set of all fixed points of a mapping consisting of a convex Bregman projection and a pooling operator.

Example 5. *Assume that there are n experts, each with his own knowledge represented by closed convex nonempty sets $W_1, \ldots, W_n \subseteq \mathbb{D}^J$, 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 his own knowledge by what seems to be the right probability function. In other words, the expert "i" projects $\mathbf{v}$ to W_i, obtaining the probability function $\mathbf{w}^{(i)}$. Each expert subsequently accepts $\mathbf{w}^{(i)}$ as his working hypothesis, but he does not discard his knowledge base W_i; he only takes into account other people's opinions. Then, it is easy for the chairman to identify the average of the actual beliefs $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ of the experts. If he found that this average $\mathbf{v}'$ did 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 the hope that, eventually, he will find a fixed point.*

It seems that, in a broad philosophical setting, such as in the example above, we ought to study any possible combination of Bregman projections with pooling operators. The question as to which other combination produces a well-defined probabilistic merging operator satisfying the consistency principle (CP) is open to investigation.

3. Convergence

3.1. Iterative Processes

In this section, we continue the investigation of the averaging projective procedures $F^{D_f, \mathcal{A}}$ and $\hat{F}^{D_f, \mathcal{A}}$. Recall that, given a convex Bregman divergence D_f and a family of weighting vectors $\mathcal{A}$, $F^{D_f, \mathcal{A}}$, was defined in the previous section for every $n \geq 1$ and all closed convex nonempty sets $W_1, \ldots, W_n \subseteq \mathbb{D}^J$ by the following.

1. For an argument $\mathbf{v} \in \mathbb{D}^J$, take $\mathbf{w}^{(i)}$ as the D_f-projection of $\mathbf{v}$ into W_i for all $1 \leq i \leq n$.
2. Set $F_{[W_1, \ldots, W_n]}^{D_f, \mathcal{A}}(\mathbf{v}) = \mathbf{LinOp_a}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$, where $\mathbf{a} \in \mathcal{A}$.

For D_f, which is moreover differentiable and strictly convex in the second argument, $\hat{F}^{D_f, \mathcal{A}}$ was defined analogously by conjugated projections and the $\mathbf{Pool}_{\mathcal{A}}^{D_f}$-pooling operator.

Our current aim is to find out what will happen if we iterate the application of averaging projective procedures $F^{D_f,\mathcal{A}}$ and $\hat{F}^{D_f,\mathcal{A}}$. In particular:

- Will the resulting sequences converge?

We shall find the answer in this subsection.

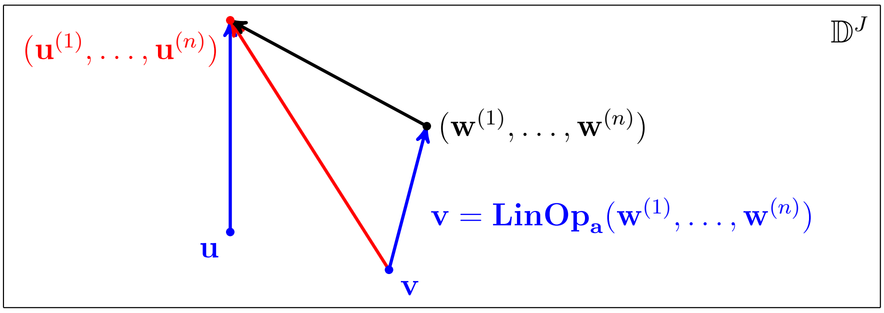
It is intriguing that we can abstractly define a "conjugated projection" with respect to a summation of a convex differentiable Bregman divergence D_f. Let $\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)} \in \mathbb{D}^J$ and $\mathbf{a} \in \mathbb{D}^n$. Then, the "conjugated projection" of $(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)})$ into $\mathbb{D}^J$ is defined by the global minimizer of $\sum_{i=1}^n a_i D_f(\mathbf{w}^{(i)}\|\mathbf{v})$, which, by Equation (1), is $\mathbf{v} = \mathbf{LinOp_a}(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)})$.

The claim that this behaves as a "conjugated projection" is supported by the following analogue of the four-point property illustrated in Figure 10.

Theorem 15. *Let D_f be a convex differentiable Bregman divergence. Let $\mathbf{a} \in \mathbb{D}^n$, $\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)} \in \mathbb{D}^J$ and $\mathbf{v} = \mathbf{LinOp_a}(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)})$. Let $\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(n)} \in \mathbb{D}^J$ and $\mathbf{u} \in \mathbb{D}^J$. Then:*

$$\sum_{i=1}^n a_i D_f(\mathbf{u}^{(i)}\|\mathbf{v}) \le \sum_{i=1}^n a_i D_f(\mathbf{u}^{(i)}\|\mathbf{u}) + \sum_{i=1}^n a_i D_f(\mathbf{u}^{(i)}\|\mathbf{w}^{(i)}).$$

Figure 10. The illustration of Theorem 15.


$\mathbb{D}^J$
$(\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(n)})$
$(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)})$
$\mathbf{v} = \mathbf{LinOp_a}(\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)})$
$\mathbf{u}$
$\mathbf{v}$

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

Proof. The proof is similar to the one of the actual four-point property (Theorem 8) only with a slightly different argument at the end: after obtaining:

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

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

we proceed with:

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

$$= \sum_{j=1}^J (u_j - v_j)\Big[\sum_{k=1}^J (\sum_{i=1}^n a_i w_k^{(i)} - v_k)\frac{\partial \frac{\partial f(\mathbf{x})}{\partial x_k}}{\partial x_j}\Big|_{\mathbf{x}=\mathbf{v}}\Big] = 0,$$

since $\sum_{i=1}^n a_i w_k^{(i)} = v_k$ for all $1 \le k \le J$, and the theorem follows. □

Similarly, given $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)} \in \mathbb{D}^J$, $\mathbf{a} \in \mathbb{D}^n$ and a convex differentiable Bregman divergence D_f, which is strictly convex in its second argument, we can consider $\mathbf{Pool}^{D_f}_{\mathbf{a}}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$ the "projection" of $(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$ into $\mathbb{D}^J$, since Theorem 6 resembles (a special case of) the extended Pythagorean property: for any $\mathbf{u} \in \mathbb{D}^J$:

$$\sum_{i=1}^{n} a_i D_f(\mathbf{u} \| \mathbf{Pool}^{D_f}_{\mathbf{a}}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})) + \sum_{i=1}^{n} a_i D_f(\mathbf{Pool}^{D_f}_{\mathbf{a}}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}) \| \mathbf{w}^{(i)}) =$$

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

The two observations above and the following lemma will be essential to the proofs of the two main theorems of this subsection.

Lemma 4. *Let D_f be a convex Bregman divergence. Assume that we are given a closed convex nonempty set W, $\mathbf{v}^{[i]} \in \mathbb{D}^L$, $i = 1, 2, \dots$ and $\mathbf{w}^{[i]} \in \mathbb{D}^J$, $i = 1, 2, \dots$, such that $\mathbf{w}^{[i]}$ is the D_f-projection of $\mathbf{v}^{[i]}$ into W for all $i = 1, 2, \dots$. Assume that $\{\mathbf{v}^{[i]}\}_{i=1}^{\infty}$ converges to $\mathbf{v} \in \mathbb{D}^J$ and $\{\mathbf{w}^{[i]}\}_{i=1}^{\infty}$ converges to $\mathbf{w} \in \mathbb{D}^J$. Then, $\mathbf{w}$ is the D_f-projection of $\mathbf{v}$ into W.*

Proof. For a contradiction, assume that the D_f-projection of $\mathbf{v}$ into W denoted by $\bar{\mathbf{w}}$ is distinct from $\mathbf{w}$. Then, by the extended Pythagorean property, $D_f(\mathbf{w}^{[i]} \| \mathbf{v}^{[i]}) + D_f(\bar{\mathbf{w}} \| \mathbf{w}^{[i]}) \leq D_f(\bar{\mathbf{w}} \| \mathbf{v}^{[i]})$. Since $D_f(\cdot \| \cdot)$ is continuous (see Section 1.1), we have that:

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

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

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

Therefore: $D_f(\mathbf{w} \| \mathbf{v}) + D_f(\bar{\mathbf{w}} \| \mathbf{w}) \leq D_f(\bar{\mathbf{w}} \| \mathbf{v})$, which contradicts the assumption that $\bar{\mathbf{w}}$ is the D_f-projection of $\mathbf{v}$ into W. □

Finally, we are going to answer the question about whether the iteration of the averaging projective procedures $F^{D_f,\mathcal{A}}$ and $\hat{F}^{D_f,\mathcal{A}}$ converges; however, the result for $F^{D_f,\mathcal{A}}$ will be limited only to the case when D_f is differentiable. Both results below should be attributed to a number of people. First, the results are applications of well-known alternative projections due to Csiszár and Tusnády; see [28], Theorem 3. In a particular case of the Kullback–Leibler divergence, the theorems were observed and proven by Matúš in [21]. Last, but not least, Eggermont and LaRiccia reformulated original alternative projections in terms of Bregman divergences in [29].

Theorem 16. *Let D_f be a convex differentiable Bregman divergence, $\mathcal{A}$ be a family of weighting vectors and $\mathbf{a} \in \mathcal{A}$ be such that $\mathbf{a} \in \mathbb{D}^n$ and $W_1, \dots, W_n \subseteq \mathbb{D}^J$ are closed, convex and nonempty. Then, for any $\mathbf{v} \in \mathbb{D}^J$, the sequence:*

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

where $\mathbf{v}^{[0]} = \mathbf{v}$ and $\mathbf{v}^{[i+1]} = F^{D_f,\mathcal{A}}_{[W_1,\dots,W_n]}(\mathbf{v}^{[i]})$ converge to some probability function in $\Theta^{D_f}_{\mathbf{a}}(W_1, \dots, W_n)$. (Recall that $\Theta^{D_f}_{\mathbf{a}}(W_1, \dots, W_n)$ is the set of the fixed points of $F^{D_f,\mathcal{A}}_{[W_1,\dots,W_n]}$, i.e., all points $\mathbf{v}$, such that $F^{D_f,\mathcal{A}}_{[W_1,\dots,W_n]}(\mathbf{v}) = \mathbf{v}$.)

Proof. This proof is inspired by [21].

Denote the D_f-projections of $\mathbf{v}^{[i]}$ into $W_1, \dots, W_n$ by $\pi_1\mathbf{v}^{[i]}, \dots, \pi_n\mathbf{v}^{[i]}$, respectively. Then, it is easy to observe that:

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

for all $i = 1, 2, \dots$. Due to the monotonicity of this sequence, the limit $\lim_{i\to\infty}\sum_{k=1}^{n} a_k D_f(\pi_k\mathbf{v}^{[i]}\|\mathbf{v}^{[i]})$ exists. Thanks to the compactness of $W_1, \dots, W_n$, the sequence $\{(\pi_1\mathbf{v}^{[i]}, \dots, \pi_n\mathbf{v}^{[i]}, \mathbf{v}^{[i]})\}_{i=1}^{\infty}$ 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 4, $\pi_k\mathbf{v}$ is really the D_f-projection of $\mathbf{v}$ into W_k for all $1 \leq k \leq n$. Moreover

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

By Theorem 15:

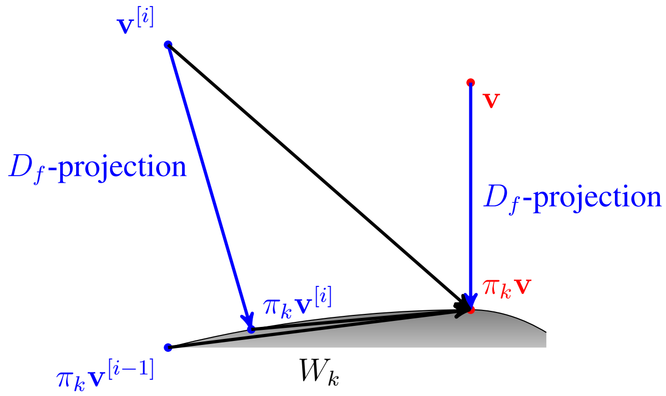
$$\sum_{k=1}^{n} a_k D_f(\pi_k\mathbf{v}\|\mathbf{v}^{[i]}) \leq \sum_{k=1}^{n} a_k D_f(\pi_k\mathbf{v}\|\mathbf{v}) + \sum_{k=1}^{n} a_k D_f(\pi_k\mathbf{v}\|\pi_k\mathbf{v}^{[i-1]}). \tag{8}$$

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

$$\sum_{k=1}^{n} a_k D_f(\pi_k\mathbf{v}^{[i]}\|\mathbf{v}^{[i]}) + \sum_{k=1}^{n} a_k D_f(\pi_k\mathbf{v}\|\pi_k\mathbf{v}^{[i]}) \leq \sum_{k=1}^{n} a_k D_f(\pi_k\mathbf{v}\|\mathbf{v}^{[i]}). \tag{9}$$

An illustration of the situation is depicted in Figure 11.

Figure 11. The situation in the proof of Theorem 16.



Now, since:

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

and $\sum_{k=1}^{n} a_k D_f(\pi_k\mathbf{v}^{[i]}\|\mathbf{v}^{[i]}) \geq \sum_{k=1}^{n} a_k D_f(\pi_k\mathbf{v}\|\mathbf{v})$ for all $i = 1, 2, \dots$, Equations (8) and (9) give that:

$$\sum_{k=1}^{n} a_k D_f(\pi_k\mathbf{v}\|\pi_k\mathbf{v}^{[i]}) \leq \sum_{k=1}^{n} a_k D_f(\pi_k\mathbf{v}\|\pi_k\mathbf{v}^{[i-1]}) \tag{10}$$

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

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

exists.

However, we already know that a subsequence of $\{(\pi_1 \mathbf{v}^{[i]}, \ldots, \pi_n \mathbf{v}^{[i]})\}_{i=1}^{\infty}$ converges to $(\pi_1 \mathbf{v}, \ldots, \pi_n \mathbf{v})$; hence, a subsequence of the sequence $\{\sum_{k=1}^{n} a_k D_f(\pi_k \mathbf{v} \| \pi_k \mathbf{v}^{[i]})\}_{i=1}^{\infty}$ decreases to zero, which by Equation (10), forces the whole sequence to converge to zero. Due to the fact that $D_f(\mathbf{x} \| \mathbf{y}) = 0$, 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, $\mathbf{v} = \lim_{i\to\infty} \mathbf{v}^{[i+1]} = \lim_{i\to\infty} \mathbf{LinOp_a}(\pi_1 \mathbf{v}^{[i]}, \ldots, \pi_n \mathbf{v}^{[i]}) = \mathbf{LinOp_a}(\pi_1 \mathbf{v}, \ldots, \pi_n \mathbf{v})$, and therefore, $\mathbf{v}$ is a fixed point of the mapping $F^{D_f,\mathcal{A}}_{[W_1,\ldots,W_n]}$; hence, $\mathbf{v} \in \Theta^{D_f}_{\mathbf{a}}(W_1, \ldots, W_n)$. □

The following analogue of Lemma 4 will be needed in the forthcoming theorem.

Lemma 5. *Let D_f be a convex differentiable Bregman divergence, which is strictly convex in its second argument. Assume that we are given a closed convex nonempty set W, $\mathbf{v}^{[i]} \in \mathbb{D}^L$, $i = 1, 2, \ldots$ and $\mathbf{w}^{[i]} \in \mathbb{D}^J$, $i = 1, 2, \ldots$, such that $\mathbf{w}^{[i]}$ is the conjugated D_f-projection of $\mathbf{v}^{[i]}$ into W for all $i = 1, 2, \ldots$. Assume that $\{\mathbf{v}^{[i]}\}_{i=1}^{\infty}$ converges to $\mathbf{v} \in \mathbb{D}^J$ and $\{\mathbf{w}^{[i]}\}_{i=1}^{\infty}$ converges to $\mathbf{w} \in \mathbb{D}^J$. Then, $\mathbf{w}$ is the conjugated D_f-projection of $\mathbf{v}$ into W.*

Proof. For a contradiction, assume that the conjugated D_f-projection of $\mathbf{v}$ into W denoted by $\bar{\mathbf{w}}$ is distinct from $\mathbf{w}$. Then, by the four-point property, $D_f(\mathbf{v}^{[i]} \| \mathbf{w}^{[i]}) \leq D_f(\mathbf{v}^{[i]} \| \bar{\mathbf{w}}) + D_f(\mathbf{v}^{[i]} \| \mathbf{v})$. Since $D_f(\cdot \| \cdot)$ is continuous, we have that:

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

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

$$\lim_{i\to\infty} D_f(\mathbf{v}^{[i]} \| \mathbf{v}) = D_f(\mathbf{v} \| \mathbf{v}) = 0.$$

Therefore: $D_f(\mathbf{v} \| \mathbf{w}) \leq D_f(\mathbf{v} \| \bar{\mathbf{w}})$, which contradicts the assumption that $\bar{\mathbf{w}}$ is the conjugated D_f-projection of $\mathbf{v}$ into W. □

Theorem 17. *Let D_f be a convex differentiable Bregman divergence, which is strictly convex in its second argument, $\mathcal{A}$ be a family of weighting vectors and $\mathbf{a} \in \mathcal{A}$ be such that $\mathbf{a} \in \mathbb{D}^n$ and $W_1, \ldots, W_n \subseteq \mathbb{D}^J$ are closed, convex and nonempty. Then, for any $\mathbf{v} \in \mathbb{D}^J$, the sequence:*

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

where $\mathbf{v}^{[0]} = \mathbf{v}$ and $\mathbf{v}^{[i+1]} = \hat{F}^{D_f,\mathcal{A}}_{[W_1,\ldots,W_n]}(\mathbf{v}^{[i]})$, converges to some probability function in $\hat{\Theta}^{D_f}_{\mathbf{a}}(W_1, \ldots, W_n)$. (Recall that $\hat{\Theta}^{D_f}_{\mathbf{a}}(W_1, \ldots, W_n)$ is the set of the fixed points of $\hat{F}^{D_f,\mathcal{A}}_{[W_1,\ldots,W_n]}$, i.e., all points $\mathbf{v}$, such that $\hat{F}^{D_f,\mathcal{A}}_{[W_1,\ldots,W_n]}(\mathbf{v}) = \mathbf{v}$.)

Proof. Denote the conjugated D_f-projections of $\mathbf{v}^{[i]}$ into $W_1, \ldots, W_n$ by $\pi_1\mathbf{v}^{[i]}, \ldots, \pi_n\mathbf{v}^{[i]}$, respectively. Then, it is easy to observe that:

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

for all $i = 1, 2, \ldots$. Due to the monotonicity of this sequence, the limit $\lim_{i\to\infty} \sum_{k=1}^{n} a_k D_f(\mathbf{v}^{[i]} \| \pi_k \mathbf{v}^{[i]})$ exists. Thanks to the compactness of $W_1, \ldots, W_n$, the sequence $\{(\pi_1\mathbf{v}^{[i]}, \ldots, \pi_n\mathbf{v}^{[i]}, \mathbf{v}^{[i]})\}_{i=1}^{\infty}$ has a convergent subsequence. Let us denote the limit of this subsequence $(\pi_1\mathbf{v}, \ldots, \pi_n\mathbf{v}, \mathbf{v})$. Due to Lemma 5, $\pi_k\mathbf{v}$ is really the conjugated D_f-projection of $\mathbf{v}$ into W_k for all $1 \leq k \leq n$. Moreover:

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

By the four-point property:

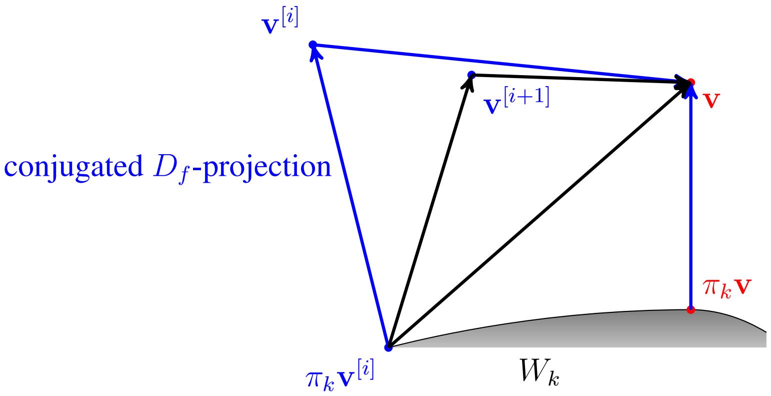
$$\sum_{k=1}^{n} a_k D_f(\mathbf{v} \| \pi_k \mathbf{v}^{[i]}) \leq \sum_{k=1}^{n} a_k D_f(\mathbf{v} \| \pi_k \mathbf{v}) + D_f(\mathbf{v} \| \mathbf{v}^{[i]}). \tag{11}$$

Moreover, by Theorem 6:

$$\sum_{k=1}^{n} a_k D_f(\mathbf{v}^{[i+1]} \| \pi_k \mathbf{v}^{[i]}) + D_f(\mathbf{v} \| \mathbf{v}^{[i+1]}) = \sum_{k=1}^{n} a_k D_f(\mathbf{v} \| \pi_k \mathbf{v}^{[i]}). \tag{12}$$

That is because $\mathbf{v}^{[i+1]} = \mathbf{Pool}_{\mathbf{a}}^{D_f}(\pi_1\mathbf{v}^{[i]}, \ldots, \pi_n\mathbf{v}^{[i]})$. An illustration of the situation is depicted in Figure 12.

Figure 12. The situation in the proof of Theorem 16.



Now, since:

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

and $\sum_{k=1}^{n} a_k D_f(\mathbf{v}^{[i+1]} \| \pi_k \mathbf{v}^{[i]}) \geq \sum_{k=1}^{n} a_k D_f(\mathbf{v} \| \pi_k \mathbf{v})$ for all $i = 1, 2, \ldots$, the expressions (11) and (12) give that:

$$D_f(\mathbf{v} \| \mathbf{v}^{[i+1]}) \leq D_f(\mathbf{v} \| \mathbf{v}^{[i]}) \tag{13}$$

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

$$\lim_{i\to\infty} D_f(\mathbf{v}\|\mathbf{v}^{[i]})$$

exists.

However, we already know that a subsequence of $\{\mathbf{v}^{[i]}\}_{i=1}^{\infty}$ converges to $\mathbf{v}$; hence, a subsequence of the sequence $\{D_f(\mathbf{v}\|\mathbf{v}^{[i]})\}_{i=1}^{\infty}$ decreases to zero, which by Equation (13), forces the whole sequence to converge to zero. Due to the fact that $D_f(\mathbf{x}\|\mathbf{y}) = 0$ only if $\mathbf{x} = \mathbf{y}$ and by the continuity, we get:

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

and, subsequently, $\lim_{i\to\infty} \pi_k\mathbf{v}^{[i]} = \pi_k\mathbf{v}$, $1 \le k \le n$ (the subsequence of $\{\pi_k\mathbf{v}^{[i]}\}_{i=1}^{\infty}$ has $\pi_k\mathbf{v}$ as a limit, and $\{D_f(\mathbf{v}^{[i]}\|\pi_k\mathbf{v}^{[i]})\}_{i=1}^{\infty}$ is monotonic).

Moreover, $\mathbf{v} = \lim_{i\to\infty} \mathbf{v}^{[i+1]} = \lim_{i\to\infty} \mathbf{Pool}_{\mathbf{a}}^{D_f}(\pi_1\mathbf{v}^{[i]}, \ldots, \pi_n\mathbf{v}^{[i]}) = \mathbf{Pool}_{\mathbf{a}}^{D_f}(\pi_1\mathbf{v}, \ldots, \pi_n\mathbf{v})$, since $\mathbf{Pool}_{\mathbf{a}}^{D_f}$ is continuous ($\sum_{k=1}^{n} a_k D_f(\cdot\|\cdot)$ is continuous and strictly convex in the first argument). Therefore, $\mathbf{v}$ is a fixed point of the mapping $\hat{F}^{D_f,\mathcal{A}}_{[W_1,\ldots,W_n]}$, and hence, $\mathbf{v} \in \hat{\Theta}_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$. □

The problem of characterizing the limits of Theorems 16 and 17 more precisely remains open. On the other hand, the theorems suggest a way to compute at least some points in $\Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$ and $\hat{\Theta}_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$, although we have not investigated how fast the sequences converge. Moreover, also the question of how effective it is to compute D_f-projections and conjugated D_f-projections was left unanswered. This latter problem was nevertheless addressed in the literature, at least in the case of the KL-divergence and sets $W_1, \ldots, W_n$ generated by finite collections of marginal probability functions. In such a case, the well-known iterative projective fitting procedure IPFP can be effectively employed [16].

3.2. Chairmen Theorems

In this section, for a convex differentiable Bregman divergence D_f, which is strictly convex in its second argument, and a family of weighting vectors $\mathcal{A}$, we investigate the susceptibility of $\Theta_{\mathcal{A}}^{D_f}$ and $\hat{\Theta}_{\mathcal{A}}^{D_f}$-merging operators to a small bias by an arbitrary probability function in $\mathbb{D}^J$. The study of this problem first occurred in [18], 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 (not arbitrary), in order to calculate a single social probability function and then find what would happen to this social probability function if his contribution happened to be infinitesimally small relative to that of the other experts. He showed that in the case of the $\hat{\Theta}_{\mathcal{N}}^{\mathrm{KL}}$-merging operator, this point of agreement is characterized by the most entropic point in the region defined by $\hat{\Theta}_{\mathcal{N}}^{\mathrm{KL}}$. A similar theorem for the $\Theta_{\mathcal{N}}^{\mathrm{KL}}$-merging operator was proven in [10]. In what follows, we adapt these results to our general situation.

The following theorem will tell us that, in some particular case of $W_1, \ldots, W_n \subseteq \mathbb{D}^J$, we can always tell that the set $\Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$ is a singleton.

Theorem 18. *Let $W_1, \ldots, W_n \subseteq \mathbb{D}^J$ be closed, convex, nonempty and such that, for at least one i W_i is a singleton. Let D_f be a convex Bregman divergence, which is strictly convex in its second argument and $\mathbf{a} \in \mathbb{D}^n$. Then, $\Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$ is a singleton.*

Proof. Without loss of generality, assume that $W_1 = \{\mathbf{v}\}$. For a contradiction, suppose that $\mathbf{w}, \mathbf{r} \in \Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$ and $\mathbf{w} \neq \mathbf{r}$. Denote $\mathbf{w}^{(2)}, \ldots, \mathbf{w}^{(n)}$ the D_f-projections of $\mathbf{w}$ into $W_2, \ldots, W_n$, respectively, and $\mathbf{r}^{(2)}, \ldots, \mathbf{r}^{(n)}$ the D_f-projections of $\mathbf{r}$ into $W_2, \ldots, W_n$, respectively. By definition, $\mathbf{w} = \mathbf{LinOp_a}(\mathbf{v}, \mathbf{w}^{(2)}, \ldots, \mathbf{w}^{(n)})$ and $\mathbf{r} = \mathbf{LinOp_a}(\mathbf{v}, \mathbf{r}^{(2)}, \ldots, \mathbf{r}^{(n)})$.

Now, consider $\mathbf{x} = \lambda\mathbf{w} + (1-\lambda)\mathbf{r}$ for some $\lambda \in (0, 1)$. By Theorems 9 and 12, we have that $\mathbf{x} \in \Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$. Since $D_f(\cdot\|\cdot)$ is a convex function, by the Jensen inequality, we have that:

$$a_1 D_f(\mathbf{v}\|\mathbf{x}) + \sum_{i=2}^{n} a_i D_f(\lambda\mathbf{w}^{(i)} + (1-\lambda)\mathbf{r}^{(i)}\|\mathbf{x}) \le$$

$$\le \lambda\Big(a_1 D_f(\mathbf{v}\|\mathbf{w}) + \sum_{i=2}^{n} a_i D_f(\mathbf{w}^{(i)}\|\mathbf{w})\Big) + (1-\lambda)\Big(a_1 D_f(\mathbf{v}\|\mathbf{r}) + \sum_{i=2}^{n} a_i D_f(\mathbf{r}^{(i)}\|\mathbf{r})\Big). \tag{14}$$

However, since $\mathbf{w}, \mathbf{r}, \mathbf{x} \in \Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$ and $\lambda\mathbf{w}^{(i)} + (1-\lambda)\mathbf{r}^{(i)} \in W_i$, $1 \le i \le n$, the above is possible only with the equality.

On the other hand, since D_f is strictly convex in its second argument, the following Jensen inequality is strict:

$$D_f(\mathbf{v}\|\mathbf{x}) < \lambda D_f(\mathbf{v}\|\mathbf{w}) + (1-\lambda) D_f(\mathbf{v}\|\mathbf{r}).$$

Note that the border points $\lambda = 0, 1$ are excluded. Therefore, Equation (14) yields:

$$\sum_{i=2}^{n} a_i D_f(\lambda\mathbf{w}^{(i)} + (1-\lambda)\mathbf{r}^{(i)}\|\mathbf{x}) >$$

$$> \lambda\Big(\sum_{i=2}^{n} a_i D_f(\mathbf{w}^{(i)}\|\mathbf{w})\Big) + (1-\lambda)\Big(\sum_{i=2}^{n} a_i D_f(\mathbf{r}^{(i)}\|\mathbf{r})\Big).$$

However, this contradicts the Jensen inequality. □

Theorem 19 (Chairman Theorem for $\Theta_{\mathcal{A}}^{D_f}$)**.** *Let $I \subseteq \mathbb{D}^J$ be a singleton consisting of an arbitrary probability function $\mathbf{t} \in \mathbb{D}^J$. Let $W_1, \ldots, W_n \subseteq \mathbb{D}^J$ be closed, convex and nonempty, $\mathbf{a} \in \mathcal{A}$ be such that $\mathbf{a} \in \mathbb{D}^n$ and D_f be a convex Bregman divergence, which is strictly convex in its second argument. For $1 > \lambda > 0$, define (by the previous theorem, the following set is a singleton):*

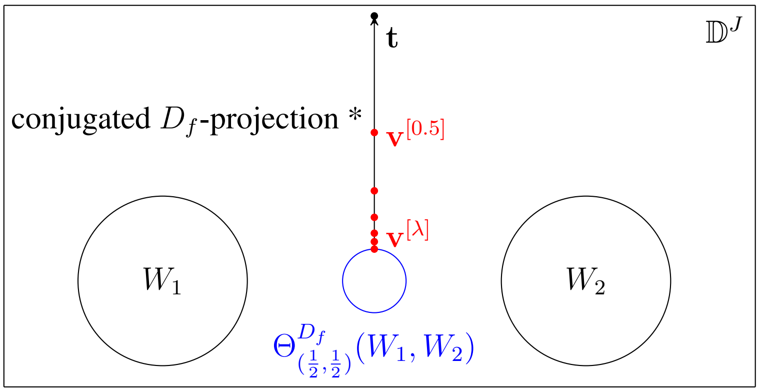
$$\{\mathbf{v}^{[\lambda]}\} = \Theta^{D_f}_{(\lambda, a_1 - \lambda a_1, \ldots, a_n - \lambda a_n)}(I, W_1, \ldots, W_n).$$

Then, $\lim_{\lambda \searrow 0} \mathbf{v}^{[\lambda]}$ exists and equals

$$\arg \min_{\mathbf{v} \in \Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)} D_f(\mathbf{t}\|\mathbf{v}),$$

i.e., it equals the conjugated D_f-projection of the probability function $\mathbf{t}$ into $\Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$.

Figure 13. The illustration of the chairman theorem for $\Theta_{\mathcal{N}}^{D_f}$.


t
$\mathbb{D}^J$
conjugated D_f-projection *
$\mathbf{v}^{[0.5]}$
$\mathbf{v}^{[\lambda]}$
W_1
W_2
$\Theta^{D_f}_{(\frac{1}{2},\frac{1}{2})}(W_1, W_2)$

* Note that the fact that $\mathbf{v}^{[\lambda]}$-s lie on the arrow does not have any meaning.

Proof. This proof is inspired by [30], where a slightly stronger result is proven for the special case of $\Theta_{\mathcal{N}}^{\mathrm{KL}}$. We note that Theorem 9 from Section 2.3 is implicitly used in what follows.

First, denote $\mathrm{M}_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$ as the minimal value of:

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

subject to $\mathbf{w}^{(i)} \in W_i$, $1 \leq i \leq n$ and $\mathbf{v} \in \mathbb{D}^J$. Furthermore, we denote E_λ as the minimal value of:

$$(1-\lambda)\Big[\sum_{i=1}^{n} a_i D_f(\mathbf{w}^{(i)} \| \mathbf{v}) - \mathrm{M}_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)\Big] + \lambda D_f(\mathbf{t} \| \mathbf{v}) \tag{15}$$

subject to $\mathbf{w}^{(i)} \in W_i$, $1 \leq i \leq n$ and $\mathbf{v} \in \mathbb{D}^J$. By the definition of $\mathrm{M}_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$, we have that $0 \leq E_\lambda$ for all $1 > \lambda > 0$.

Note that for a fixed λ, if $\mathbf{v} \in \mathbb{D}^J$ globally minimizes Equation (15) subject to $\mathbf{w}^{(i)} \in W_i$, $1 \leq i \leq n$, then $\mathbf{v} \in \Theta^{D_f}_{(\lambda, a_1 - \lambda a_1, \ldots, a_n - \lambda a_n)}(I, W_1, \ldots, W_n)$ (by Theorem 18, such a $\mathbf{v}$ is unique), and conversely, if $\mathbf{v} \in \Theta^{D_f}_{(\lambda, a_1 - \lambda a_1, \ldots, a_n - \lambda a_n)}(I, W_1, \ldots, W_n)$, then $\mathbf{v}$ minimizes Equation (15), subject to the above constraints.

Now, let $\mathbf{r} = \arg\min_{\mathbf{v} \in \Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)} D_f(\mathbf{t} \| \mathbf{v})$. Since $\mathbf{r} \in \Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$, it follows that for all $1 > \lambda > 0$, we have that:

$$E_\lambda \leq \lambda D_f(\mathbf{t} \| \mathbf{r}). \tag{16}$$

Since $\mathbb{D}^J \subseteq \mathbb{R}^J$ is a compact space, there exists a sequence $\{\lambda_m\}_{m=1}^{\infty}$, $0 < \lambda_m < 1$, $\lim_{m\to\infty} \lambda_m = 0$, such that $\{\mathbf{v}^{[\lambda_m]}\}_{m=1}^{\infty}$ converges. Let $\mathbf{w}^{(i)[\lambda_m]}$ be the D_f-projection of $\mathbf{v}^{[\lambda_m]}$ into W_i for all $1 \leq i \leq n$ and $m = 1, 2, \ldots$. By Equation (16), the sequence:

$$\Big\{\sum_{i=1}^{n} a_i D_f(\mathbf{w}^{(i)[\lambda_m]} \| \mathbf{v}^{[\lambda_m]})\Big\}_{m=1}^{\infty}$$

converges to $\mathrm{M}_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$.

Note that we already know that $\lim_{m\to\infty} \mathbf{v}^{[\lambda_m]}$ exists, and we denote it by $\mathbf{v}$. However, we do not know whether the same is true for $\lim_{m\to\infty} \mathbf{w}^{(i)[\lambda_m]}$, $1 \leq i \leq n$. On the other hand, since

$W_1, \dots, W_n$ are compact, the considered sequences have convergent subsequences. Let us denote the corresponding limits $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}$. Since $D_f(\cdot\|\cdot)$ is a continuous function in both variables, the value of $\sum_{i=1}^n a_i D_f(\mathbf{w}^{(i)}\|\mathbf{v})$ must be equal to $\mathrm{M}_{\mathbf{a}}^{D_f}(W_1, \dots, W_n)$. However, this means that we have found a global minimizer $(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}, \mathbf{v})$ of $\sum_{i=1}^n a_i D_f(\mathbf{w}^{(i)}\|\mathbf{v})$ subject to $\mathbf{w}^{(i)} \in W_i$, $1 \le i \le n$, and $\mathbf{v} \in \mathbb{D}^J$.

It follows that $\mathbf{v} = \lim_{m\to\infty} \mathbf{v}^{[\lambda_m]} \in \Theta_{\mathbf{a}}^{D_f}(W_1, \dots, W_n)$. By Equation (16):

$$0 \le (1-\lambda_m)\Big[\sum_{i=1}^n a_i D_f(\mathbf{w}^{(i)[\lambda_m]}\|\mathbf{v}^{[\lambda_m]}) - \mathrm{M}_{\mathbf{a}}^{D_f}(W_1, \dots, W_n)\Big] + \lambda_m D_f(\mathbf{t}\|\mathbf{v}^{[\lambda_m]}) \le \lambda_m D_f(\mathbf{t}\|\mathbf{r}).$$

Hence, $0 \le \lambda_m[D_f(\mathbf{t}\|\mathbf{r}) - D_f(\mathbf{t}\|\mathbf{v}^{[\lambda_m]})]$ for all $m = 1, 2, \dots$. However, by definition of $\mathbf{r}$, this is possible only if $\mathbf{r} = \mathbf{v}$.

In fact, we have proven that for every sequence $\{\lambda_m\}_{m=1}^\infty$, such that $\lim_{m\to\infty} \lambda_m = 0$ and $\{\mathbf{v}^{[\lambda_m]}\}_{m=1}^\infty$ is convergent, $\{\mathbf{v}^{[\lambda_m]}\}_{m=1}^\infty$ must converge to $\mathbf{r}$. Therefore, assume that there is a sequence $\{\lambda_m\}_{m=1}^\infty$, such that $\lim_{m\to\infty} \lambda_m = 0$, but $\{\mathbf{v}^{[\lambda_m]}\}_{m=1}^\infty$ is not convergent. Then, there is an open neighborhood of the point $\mathbf{r}$ outside of which there are an infinite number of the members of the sequence $\{\mathbf{v}^{[\lambda_m]}\}_{m=1}^\infty$. Since $\mathbb{D}^J$ is compact, this sequence must have a convergent subsequence with a limit distinct from $\mathbf{r}$. That, however, contradicts our previous claim. □

The theorem above is illustrated in Figure 13. Indeed, if $\Theta_{\mathbf{a}}^{D_f}(W_1, \dots, W_n)$ is a singleton, then the limit in the theorem above is obvious. By Theorem 18, this happens in particular when at least one of $W_1, \dots, W_n$ is a singleton. However, it is not hard to observe an interesting case; consider that $W_1, \dots, W_n$ have a nonempty intersection, which is not a singleton. In this case, the limit above is, in fact, the conjugated D_f-projection of the probability function $\mathbf{t}$ into that intersection. Such a conjugated projection depends on $\mathbf{t}$. In particular, we can recover any point in the intersection by setting it to be the point $\mathbf{t}$.

The following analogue of Theorem 18 has a fairly similar proof.

Theorem 20. *Let $W_1, \dots, W_n \subseteq \mathbb{D}^J$ be closed, convex, nonempty and such that, for at least one i W_i is a singleton. Let D_f be a convex Bregman divergence, which is strictly convex in its second argument and $\mathbf{a} \in \mathbb{D}^n$. Then, $\hat{\Theta}_{\mathbf{a}}^{D_f}(W_1, \dots, W_n)$ is a singleton.*

Theorem 21 (Chairman Theorem for $\hat{\Theta}_{\mathcal{A}}^{D_f}$)**.** *Let $I \subseteq \mathbb{D}^J$ be a singleton consisting of an arbitrary probability function $\mathbf{t} \in \mathbb{D}^J$. Let $W_1, \dots, W_n \subseteq \mathbb{D}^J$ be closed, convex and nonempty, $\mathbf{a} \in \mathcal{A}$ be such that $\mathbf{a} \in \mathbb{D}^n$ and D_f be a convex differentiable Bregman divergence, which is strictly convex in its second argument. For $1 > \lambda > 0$, define:*

$$\{\mathbf{v}^{[\lambda]}\} = \hat{\Theta}_{(\lambda, a_1 - \lambda a_1, \dots, a_n - \lambda a_n)}^{D_f}(I, W_1, \dots, W_n).$$

Then, $\lim_{\lambda \searrow 0} \mathbf{v}^{[\lambda]}$ exists and equals:

$$\arg \min_{\mathbf{v} \in \hat{\Theta}_{\mathbf{a}}^{D_f}(W_1, \dots, W_n)} D_f(\mathbf{v}\|\mathbf{t}),$$

i.e., it equals the D_f-projection of the probability function $\mathbf{t}$ into $\hat{\Theta}_{\mathbf{a}}^{D_f}(W_1, \dots, W_n)$.

The proof is analogous to the one of Theorem 19, so we omit it.

4. Applications

4.1. Relationship to Inference Processes

In this subsection, we will discuss some striking relationships between the chairmen theorems and the framework of inference processes [26]. Inference processes are methods of reasoning by which an expert may select a single probability function from a nonempty closed convex set of possible options. In our framework, it is simply a problem of choosing a single probability function in a closed convex nonempty set $W \subseteq \mathbb{D}^J$. This selection is, however, not arbitrary, and it is expected to satisfy some rational principles based on symmetry and consistency, as discussed in [15]. The maximum entropy ($\mathbf{ME}$) inference process, which chooses the most entropic point in a given closed convex nonempty set, is uniquely justified by a list of such principles, as Paris and Vencovská showed [15].

As discussed in Section 1.2, the most entropic point in a closed convex nonempty set $W \subseteq \mathbb{D}^J$ coincides with the KL-projection of the uniform probability function into W. This can be immediately applied to the chairman theorem for $\hat{\Theta}^{\mathrm{KL}}_{\mathcal{A}}$, where $\mathcal{A}$ is a family of weighting vectors:

Let $I \subseteq \mathbb{D}^J$ be a singleton consisting of the uniform probability function $\mathbf{t} \in \mathbb{D}^J$. Let $W_1, \ldots, W_n \subseteq \mathbb{D}^J$ be closed, convex and nonempty and $\mathbf{a} \in \mathcal{A}$ be such that $\mathbf{a} \in \mathbb{D}^n$. For $1 > \lambda > 0$, define:

$$\{\mathbf{v}^{[\lambda]}\} = \hat{\Theta}^{\mathrm{KL}}_{(\lambda, a_1 - \lambda a_1, \ldots, a_n - \lambda a_n)}(I, W_1, \ldots, W_n).$$

Then:

$$\lim_{\lambda \searrow 0} \mathbf{v}^{[\lambda]} = \mathbf{ME}(\hat{\Theta}^{\mathrm{KL}}_{\mathbf{a}}(W_1, \ldots, W_n)).$$

For the family of weighting vectors:

$$\mathcal{N} = \Big\{ \big(\underbrace{\frac{1}{n}, \ldots, \frac{1}{n}}_{n} \big) : \; n = 1, 2, \ldots \Big\}$$

the operator that results by applying the $\mathbf{ME}$-inference process to the operator $\hat{\Theta}^{\mathrm{KL}}_{\mathcal{N}}$ is, in fact, a probabilistic merging operator, which was introduced and studied by Wilmers in [18] under the name "social entropy process" or $\mathbf{SEP}$, for short. In that paper, Wilmers argues that this merging operator is, to date, the most appealing with respect to symmetry and consistency; somehow, in the spirit of the original justification for the $\mathbf{ME}$-inference process, although the problem of finding a complete justification is still open.

Whether $\mathbf{SEP}$ will turn out to be the most appealing probabilistic merging operator or not, by the same manner as above, we can define several probabilistic merging operators related to several other classical inference processes.

For example, the conjugated KL-projection of the uniform probability function into a closed convex nonempty set $W \subseteq \mathbb{D}^J$ in fact generates the so-called $\mathbf{CM}^\infty$-inference process (a limit version of the central mass process [26]). We write simply $\mathbf{CM}^\infty(W)$ to denote the point of the projection, which is explicitly given by:

$$\mathbf{CM}^\infty(W) = \arg\min_{\mathbf{w} \in W} -\sum_{j=1}^{J} \log w_j.$$

The chairman theorem for $\Theta^{\mathrm{KL}}_{\mathcal{N}}$ then suggests considering the probabilistic merging operator defined for every $n \geq 1$ and all closed convex nonempty sets $W_1, \dots, W_n \subseteq \mathbb{D}^J$ by:

$$\mathbf{coSEP}(W_1, \dots, W_n) = \{\mathbf{CM}^{\infty}(\Theta^{\mathrm{KL}}_{\mathbf{a}}(W_1, \dots, W_n))\},$$

where $\mathbf{a} \in \mathbb{D}^n$ and $\mathbf{a} \in \mathcal{N}$. We will call this operator the conjugated social entropy process $\mathbf{coSEP}$.

What is really appealing about the operators $\mathbf{SEP}$ and $\mathbf{coSEP}$ is that there are singletons; we simply say that they satisfy the singleton principle (SP). Furthermore, the consistency principle (CP) is obviously satisfied by all of them. However, there is an interesting principle that can never be satisfied by a probabilistic merging operator that satisfies (CP) and is always a singleton: the disagreement principle introduced in [5].

(DP) Disagreement Principle. Let Δ be a probabilistic merging operator. Then, we say that Δ satisfies the disagreement principle if, for every $n, m \geq 1$ and all $W_1, \dots, W_n \subseteq \mathbb{D}^J$ and $V_1, \dots, V_m \subseteq \mathbb{D}^J$:

$$\Delta(W_1, \dots, W_n) \cap \Delta(V_1, \dots, V_m) = \emptyset$$

implies:

$$\Delta(W_1, \dots, W_n, V_1, \dots, V_m) \cap \Delta(W_1, \dots, W_n) = \emptyset.$$

We cite [5] on the desirability of this principle: the principle (informally) says "…that a consistent group who disagrees with another group and then merges with them can be sure that they have influenced the opinions of the combined group."

Theorem 22. *There is no probabilistic merging operator that satisfies all (SP), (CP) and (DP).*

Proof. Let Δ be a probabilistic merging operator. Assume that $V \subsetneq W \subseteq \mathbb{D}^J$ and that V is a singleton. Suppose that $\Delta(W) \neq V = \Delta(V)$. Then, by (CP), $\Delta(W, V) = V$, which contradicts (DP). □

Theorem 23. *The probabilistic merging operators $\Theta^{D_f}_{\mathcal{N}}$ and $\hat{\Theta}^{D_f}_{\mathcal{N}}$, where D_f is a convex Bregman divergence for the prior and is additionally differentiable and strictly convex in its second argument for the latter, satisfy (DP).*

Proof. We prove the theorem only for $\hat{\Theta}^{D_f}_{\mathcal{N}}$. The proof for $\Theta^{D_f}_{\mathcal{N}}$ is similar.

Let $W_1, \dots, W_n, V_1, \dots, V_m \subseteq \mathbb{D}^J$ be closed convex and nonempty. For a contradiction, assume that $\mathbf{v} \in \hat{\Theta}^{D_f}_{(\frac{1}{n}, \dots, \frac{1}{n})}(W_1, \dots, W_n)$, $\mathbf{v} \in \hat{\Theta}^{D_f}_{(\frac{1}{n+m}, \dots, \frac{1}{n+m})}(W_1, \dots, W_n, V_1, \dots, V_m)$ and, at the same time, $\mathbf{v} \notin \hat{\Theta}^{D_f}_{(\frac{1}{m}, \dots, \frac{1}{m})}(V_1, \dots, V_m)$.

Denote $\mathbf{v}^{(i)}$ the conjugated D_f-projection of $\mathbf{v}$ into V_i, $1 \leq i \leq m$. Then, there is $\mathbf{u} \in \mathbb{D}^J$, such that $\mathbf{u} = \mathbf{Pool}^{D_f}_{(\frac{1}{m}, \dots, \frac{1}{m})}(\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(m)})$, *i.e.*, $\sum_{i=1}^{m} \frac{1}{m} D_f(\mathbf{v} \| \mathbf{v}^{(i)}) > \sum_{i=1}^{m} \frac{1}{m} D_f(\mathbf{u} \| \mathbf{v}^{(i)})$. Since every Bregman divergence is strictly convex in its first argument, we have that:

$$\frac{\partial}{\partial \lambda} \Big[\sum_{i=1}^{m} D_f((1-\lambda)\mathbf{v} + \lambda \mathbf{u} \| \mathbf{v}^{(i)}) \Big] \Big|_{\lambda=0} < 0. \tag{17}$$

Now, denote $\mathbf{w}^{(i)}$ the conjugated D_f-projection of $\mathbf{v}$ into W_i, $1 \leq i \leq n$. Since $\mathbf{v} = \mathbf{Pool}^{D_f}_{(\frac{1}{n+m},\dots,\frac{1}{n+m})}(\mathbf{w}^{(1)},\dots,\mathbf{w}^{(n)},\mathbf{v}^{(1)},\dots,\mathbf{v}^{(m)})$ and $\mathbf{v} = \mathbf{Pool}^{D_f}_{(\frac{1}{n},\dots,\frac{1}{n})}(\mathbf{w}^{(1)},\dots,\mathbf{w}^{(n)})$, the strict convexity of Bregman divergences in their first argument gives also:

$$\frac{\partial}{\partial\lambda}\Big[\sum_{i=1}^{n} D_f((1-\lambda)\mathbf{v}+\lambda\mathbf{u}\|\mathbf{w}^{(i)}) + \sum_{i=1}^{m} D_f((1-\lambda)\mathbf{v}+\lambda\mathbf{u}\|\mathbf{v}^{(i)})\Big]\Big|_{\lambda=0} \geq 0$$

and:

$$\frac{\partial}{\partial\lambda}\Big[\sum_{i=1}^{n} D_f((1-\lambda)\mathbf{v}+\lambda\mathbf{u}\|\mathbf{w}^{(i)})\Big]\Big|_{\lambda=0} \geq 0.$$

However, this contradicts Equation (17). □

We can conclude that, before deciding which probabilistic merging operator to use, we need to establish which two of the three properties we want the operator to satisfy. In this paper, we have seen instances of all three options, as listed in Table 1.

Table 1. Examples for three saturated possibilities with respect to the consistency principle (CP), disagreement principle (DP) and singleton principle (SP). **KIRP**, Kern-Isberner and Rödder; **OSEP**, obdurate social entropy process; **SEP**, social entropy process; **coSEP**, conjugated social entropy process.

Principles	Probabilistic Merging Operators
(DP), (CP)	$\Theta^{D_f}_{\mathcal{N}}, \hat{\Theta}^{D_f}_{\mathcal{N}}$
(DP), (SP)	**KIRP**, **OSEP**
(CP), (SP)	**SEP**, **coSEP**

Recall that **KIRP** is the operator due to Kern-Isberner and Röder and **OSEP** is the obdurate social entropy process; see Section 2.2 for more details. A proof that **KIRP** and **OSEP** satisfy (DP) can be easily obtained as a modification of the proof of Theorem 23, so we omit it.

4.2. Computability

In this subsection, we would like to propose a method corresponding to the classical method of projection, but in the multi-expert context. The possible use could be similar; if the knowledge of a college of experts could be characterized by a closed convex nonempty set of probability functions, then we would like to find such a probability function in that set that is "closest" to a given piece of information represented by another probability function. We only need to specify a way to represent the knowledge of the college by such a single set and pair it with an appropriate method of projection.

Throughout this subsection, assume that we are given closed convex nonempty sets of probability functions $W_1,\dots,W_n \subseteq \mathbb{D}^J$ with weighting $\mathbf{a} \in \mathcal{A}$, where a_i is the weight of W_i and a probability function $\mathbf{v} \in \mathbb{D}^J$ to represent.

If the measure of "being closed" is quantified by a projection by means of a convex differentiable Bregman divergence D_f, which is strictly convex in its second argument, our proposed method consists of the following. First, represent $W_1, \ldots, W_n$ by a single, closed and convex set $\hat{\Theta}^{D_f}_{\mathbf{a}}(W_1, \ldots, W_n)$, and then, take the D_f-projection of $\mathbf{v}$ into $\hat{\Theta}^{D_f}_{\mathbf{a}}(W_1, \ldots, W_n)$.

On the other hand, if the measure of "being closed" is quantified by a conjugated projections by means of a convex differentiable Bregman divergence D_f, which is strictly convex in its second argument, we first represent $W_1, \ldots, W_n$ by a single, closed convex set $\Theta^{D_f}_{\mathbf{a}}(W_1, \ldots, W_n)$ and then take the conjugated D_f-projection of $\mathbf{v}$ into $\Theta^{D_f}_{\mathbf{a}}(W_1, \ldots, W_n)$.

The methods have two distinguishing features:

1. If all of the sets $W_1, \ldots, W_n$ are singletons, the methods reduce to $\mathbf{Pool}^{D_f}_{\mathcal{A}}$ and $\mathbf{LinOp}_{\mathcal{A}}$-pooling operators respectively.
2. If $W_1, \ldots, W_n$ have a nonempty intersection V, they reduce to D_f and conjugated D_f-projections into V, respectively.

In this subsection, we shall investigate how effective it is to compute the results of those two methods. Notice that **SEP** and **coSEP**, defined in Section 4.1, are specific instances of those procedures, respectively, in which case, we are interested in KL-projections and conjugated KL-projections of the uniform probability function.

There are indeed some serious computational issues. The most essential is the following. A closed convex nonempty set $W \subseteq \mathbb{D}^J$ is often given by a set of constraints on $\mathbb{D}^J$. How can we effectively verify that the resulting set W is nonempty? Unfortunately, it is not even possible to find a random Turing machine running in polynomial time that upon input given by a set of constraints on probability functions verifies the consistency of this set of constraints (given that the problems solvable in a randomized polynomial time cannot be solved in a polynomial time); see Theorem 10.7 of [26].

However, some computational problems closely related to projections have been extensively studied in the literature. As we have noted in Section 3.1, 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 given closed convex nonempty sets are generated by marginal probability functions and where the IPFP-procedure can be applied to effectively find a KL-projection; see [16]. Therefore, we will assume that some effective procedures for D_f-projections and conjugated D_f-projections are given.

Under such an assumption, the iterative processes from Section 3.1 and the Chairmen theorems offer a way how to compute (although possibly ineffectively) the results of the two methods above. We shall start with the latter.

By Theorem 16, we know that the sequence:

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

where $\mathbf{v}^{[0]} = \mathbf{t}$ is arbitrary in $\mathbb{D}^J$ and $\mathbf{v}^{[i+1]} = F^{D_f,\mathcal{A}}_{[W_1,\ldots,W_n]}(\mathbf{v}^{[i]})$, converges to some probability function in $\Theta^{D_f}_{\mathbf{a}}(W_1, \ldots, W_n)$. Notice that D_f is required to be differentiable in order to establish this conclusion.

Recall that by Theorem 18, $\Theta^{D_f}_{\mathbf{a}}(W_1, \ldots, W_n)$ is a singleton when at least one of $W_1, \ldots, W_n$ is a singleton. Let $I \in \mathbb{D}^J$ be such that $I = \{\mathbf{v}\}$. For every $1 > \lambda > 0$, we define the sequence $\{\mathbf{v}^{[i]}_{[\lambda]}\}_{i=0}^{\infty}$ by $\mathbf{v}^{[0]}_{[\lambda]} = \mathbf{t}$ ($\mathbf{t}$ can be arbitrary) and:

$$\mathbf{v}^{[i+1]}_{[\lambda]} = F^{D_f,(\lambda, a_1 - \lambda a_1, \ldots, a_n - \lambda a_n)}_{[I, W_1, \ldots, W_n]}(\mathbf{v}^{[i]}_{[\lambda]}).$$

By Theorem 16:

$$\{\lim_{i\to\infty}\mathbf{v}^{[i]}_{[\lambda]}\} = \Theta^{D_f}_{(\lambda, a_1-\lambda a_1,\ldots,a_n-\lambda a_n)}(I, W_1,\ldots,W_n).$$

By the chairman theorem for $\Theta^{D_f}_{\mathcal{A}}$:

$$\lim_{\lambda\searrow 0}\lim_{i\to\infty}\mathbf{v}^{[i]}_{[\lambda]} = \arg\min_{\mathbf{w}\in\Theta^{D_f}_{\mathbf{a}}(W_1,\ldots,W_n)} D_f(\mathbf{v}\|\mathbf{w}) \tag{18}$$

i.e., equals the conjugated D_f-projection of the probability function $\mathbf{v}$ into $\Theta^{D_f}_{\mathbf{a}}(W_1,\ldots,W_n)$.

Now, notice that if the limits in Equation (18) were interchangeable, then this would offer an answer to the question from Section 3.1 to closely characterize the limit $\lim_{i\to\infty}\mathbf{v}^{[i]}$ (but with no claims to any theoretical results on the complexity of the computation). Unfortunately, the following simple example introduced in [10] shows that these limits are not interchangeable.

Example 6. *Let* $J = 4$, $W_1 = \{(x, \frac{1}{4}-x, y, \frac{3}{4}-y),\ x\in[0.01,\frac{1}{4}-0.01], y\in[0.01,\frac{3}{4}-0.01]\}$ *and* $W_2 = \{(x, y, \frac{1}{4}-x, \frac{3}{4}-y),\ x\in[0.01,\frac{1}{4}-0.01], y\in[0.01,\frac{3}{4}-0.01]\}$*. Assume that the weighting is* $\mathcal{N}$*,* $D_f = \mathrm{KL}$ *and the probability function* $\mathbf{v}\in\mathbb{D}^4$ *to interpret is the uniform probability function. In other words, we are looking for* $\mathbf{coSEP}(W_1, W_2)$*.*

Then, the members of the sequence $\{\mathbf{v}^{[i]}\}_{i=0}^{\infty}$ *can be computed by two minimization problems: find* $x\in[0.01,\frac{1}{4}-0.01]$ *and* $y\in[0.01,\frac{3}{4}-0.01]$ *that minimize:*

$$x\log\frac{x}{v^{[i]}_1} + \left(\frac{1}{4}-x\right)\log\frac{\frac{1}{4}-x}{v^{[i]}_2} + y\log\frac{y}{v^{[i]}_3} + \left(\frac{3}{4}-y\right)\log\frac{\frac{3}{4}-y}{v^{[i]}_4}$$

and another couple $\bar{x}\in[0.01,\frac{1}{4}-0.01]$ *and* $\bar{y}\in[0.01,\frac{3}{4}-0.01]$ *that minimize:*

$$\bar{x}\log\frac{\bar{x}}{v^{[i]}_1} + \bar{y}\log\frac{\bar{y}}{v^{[i]}_2} + (\frac{1}{4}-\bar{x})\log\frac{\frac{1}{4}-\bar{x}}{v^{[i]}_3} + (\frac{3}{4}-\bar{y})\log\frac{\frac{3}{4}-\bar{y}}{v^{[i]}_4}.$$

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 the 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 coordinate of probability functions is depicted in Figure 14 by the bottom red line.

However, since W_1 *and* W_2 *are jointly consistent, we have that:*

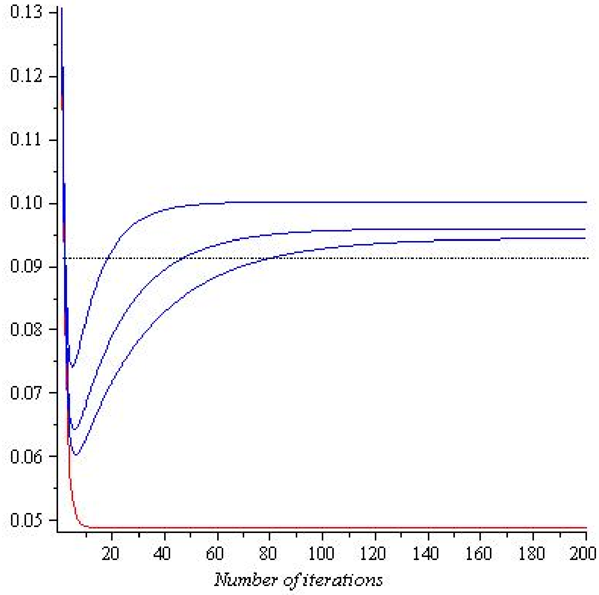
$$\Theta^{D_f}_{(\frac{1}{2},\frac{1}{2})}(W_1,W_2) = W_1\cap W_2 = \left\{\left(x,\ \frac{1}{4}-x,\ \frac{1}{4}-x,\ \frac{1}{2}+x\right),\ x\in\left[0.01,\ \frac{0.96}{4}\right]\right\}.$$

We compute that $\mathbf{CM}^{\infty}(W_1\cap W_2)$ *(the conjugated* KL*-projection of the uniform probability function) is approximately:*

$$(0.091506,\ 0.15849,\ 0.15849,\ 0.5915),$$

which is obviously not equal to the limit of the sequence $\{\mathbf{v}^{[i]}\}_{i=0}^{\infty}$.

Figure 14. The numerical computation for Example 7. Blue lines from the top are for $\lambda = \frac{1}{21}$, $\lambda = \frac{1}{41}$ and $\lambda = \frac{1}{61}$. This graph is taken from [10].



It seems that the only viable way to use Equation (18) to estimate a result of the conjugated D_f-projection into $\Theta_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$ is to choose a sufficiently small λ, and for this λ, iterate the sequence $\{\mathbf{v}_{[\lambda]}^{[i]}\}_{i=0}^{\infty}$. However, the rate of convergence heavily depends on λ, and in fact, this often materializes in a negative way for a practical computation [10]:

Example 7. *Consider the situation from Example 6. We compute numerically the first coordinate of initial members of the sequence $\{\mathbf{v}_{[\lambda]}^{[i]}\}_{i=0}^{\infty}$ for several values of λ, and we compare them with the first coordinate of the sequence $\{\mathbf{v}^{[i]}\}_{i=0}^{\infty}$. The algorithm we use is as follows. Note that due to the design of the sets, only one minimization problem is sufficient to solve in each iteration, as we have pointed out in the previous example.*

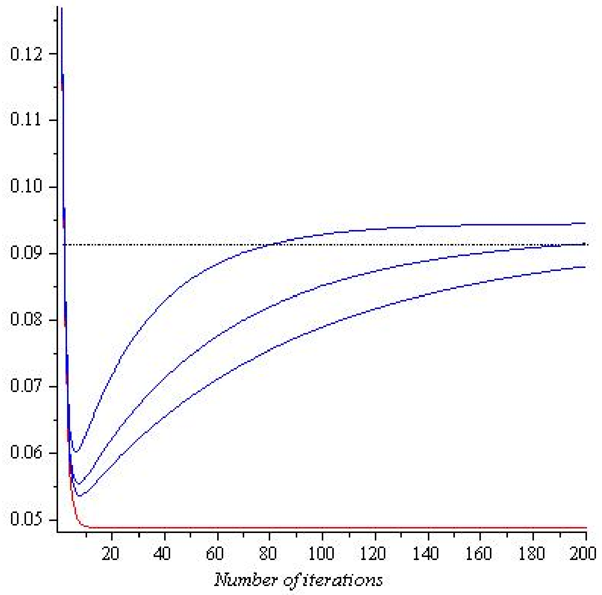
```
$v_1 := \frac{1}{4}$; $v_2 := \frac{1}{4}$; $v_3 := \frac{1}{4}$; $v_4 := \frac{1}{4}$;
for $i$ from $1$ by $1$ to $200$ do
Minimize$\left(x \log \frac{x}{v_1} + \left(\frac{1}{4} - x\right) \log \frac{\frac{1}{4}-x}{v_2} + y \log \frac{y}{v_3} + \left(\frac{3}{4} - y\right) \log \frac{\frac{3}{4}-y}{v_4}, x = 0.01..\frac{0.96}{4}, y = 0.01..\frac{2.96}{4}\right)$;
$v_1 := \frac{1}{4} \cdot \lambda + x \cdot (\frac{1}{2} - \frac{1}{2}\lambda) + x \cdot (\frac{1}{2} - \frac{1}{2}\lambda)$; $v_2 := \frac{1}{4} \cdot \lambda + (\frac{1}{4} - x) \cdot (\frac{1}{2} - \frac{1}{2}\lambda) + y \cdot (\frac{1}{2} - \frac{1}{2}\lambda)$; $v_3 :=$
$\frac{1}{4} \cdot \lambda + \frac{1}{4} - x) \cdot (\frac{1}{2} - \frac{1}{2}\lambda) + y \cdot (\frac{1}{2} - \frac{1}{2}\lambda)$; $v_4 := \frac{1}{4} \cdot \lambda + (\frac{3}{4} - y) \cdot (\frac{1}{2} - \frac{1}{2}\lambda) + (\frac{3}{4} - y) \cdot (\frac{1}{2} - \frac{1}{2}\lambda)$;
end do;
```

The numerical result for $\lambda = \frac{1}{21}, \frac{1}{41}, \frac{1}{61}$ is plotted in Figure 14. We can see that as λ decreases, the limit points of sequences are converging to the first coordinate of $\mathbf{CM}^{\infty}(W_1 \cap W_2)$, which is denoted by the black dotted line. The red line denotes the first coordinate of the sequence $\{\mathbf{v}^{[i]}\}_{i=0}^{\infty}$.

The numerical result for $\lambda = \frac{1}{61}, \frac{1}{121}, \frac{1}{181}$ is plotted in Figure 15. We can conclude that, although the eventual precision rises as λ decreases, the rate of convergence is affected severely. Therefore, there is a significant trade-off between the precision and the number of iterations.

Notice that, as λ decreases, the blue lines point-wise converge to the red line. This convergence is, however, obviously not uniform.

Figure 15. The numerical computation for Example 7. Blue lines from the top are for $\lambda = \frac{1}{61}$, $\lambda = \frac{1}{121}$ and $\lambda = \frac{1}{181}$. This graph is taken from [10].



Now, consider the prior method, which follows a fairly similar computation idea. By Theorem 17, we know that the sequence:

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

where $\mathbf{u}^{[0]} = \mathbf{t}$ is arbitrary in $\mathbb{D}^J$ and $\mathbf{u}^{[i+1]} = \hat{F}^{D_f,\mathcal{A}}_{[W_1,\ldots,W_n]}(\mathbf{u}^{[i]})$, converges to some probability function in $\hat{\Theta}^{D_f}_{\mathbf{a}}(W_1,\ldots,W_n)$. This procedure can be, for instance, immediately used to compute $\mathbf{SEP}(W_1,\ldots,W_n)$ in a case when $\hat{\Theta}^{\mathrm{KL}}_{(\frac{1}{n},\ldots,\frac{1}{n})}(W_1,\ldots,W_n)$ is a singleton. By Theorem 20, this happens when at least one of $W_1,\ldots,W_n$ is a singleton.

One may perhaps expect that if $\mathbf{u}^{[0]}$ is the uniform probability function, then $\{\lim_{i\to\infty}\mathbf{u}^{[i]}\} = \mathbf{SEP}(W_1,\ldots,W_n)$. In the following example from [10], we will, however, see that this is not true in general. Note that we cannot use Example 6, since in that case, actually, $\lim_{i\to\infty}\mathbf{u}^{[i]} = \mathbf{SEP}(W_1, W_2)$.

Example 8. *Let* $J = 8$*,*

$$W_1 = \left\{\left(x, \frac{1}{12}-x, \frac{1}{12}-x, \frac{2}{6}+x, y, \frac{1}{6}-y, \frac{1}{6}, \frac{1}{6}\right), x \in \left[0.01, \frac{0.88}{12}\right], y \in \left[0.01, \frac{0.94}{6}\right]\right\}$$

and:

$$W_2 = \left\{\left(x, \frac{1}{12}-x, \frac{1}{12}-x, \frac{2}{6}+x, \frac{1}{12}, \frac{1}{12}, y, \frac{2}{6}-y\right), x \in \left[0.01, \frac{0.88}{12}\right], y \in \left[0.01, \frac{1.94}{6}\right]\right\}.$$

W_1 *and* W_2 *have a nonempty intersection;* $W_1 \cap W_2 = \{(x, \frac{1}{12}-x, \frac{1}{12}-x, \frac{2}{6}+x, \frac{1}{12}, \frac{1}{12}, \frac{1}{6}, \frac{1}{6}),\ x \in [0.01, \frac{0.88}{12}]\}$*, and we can compute that* $\mathbf{SEP}(W_1, W_2)$ *is the most entropic probability function from the set above with* x *equal to approximately* 0.013888*.*

However, the sequence $\{\mathbf{u}^{[i]}\}_{i=0}^{\infty}$ *is already constant after one iteration and equals* $\mathbf{CM}^{\infty}(W_1) = \mathbf{CM}^{\infty}(W_2) = \mathbf{CM}^{\infty}(W_1 \cap W_2)$*, in which case,* $x \approx 0.029231$*.*

By the aid of the chairman theorem for $\hat{\Theta}^{D_f}_{\mathcal{A}}$, we also suggest a way to approximate the D_f-projection of $\mathbf{v}$ into $\hat{\Theta}^{D_f}_{\mathbf{a}}(W_1,\ldots,W_n)$, but we have no claims to any theoretical results on the complexity of computation. Let $I = \{\mathbf{v}\}$. For every $1 > \lambda > 0$, we define the sequence $\{\mathbf{u}_{[\lambda]}^{[i]}\}_{i=0}^{\infty}$ by $\mathbf{u}_{[\lambda]}^{[0]} = \mathbf{t}$, which is arbitrary, and:

$$\mathbf{u}_{[\lambda]}^{[i+1]} = \hat{F}_{[I,W_1,\ldots,W_n]}^{D_f,(\lambda,a_1-\lambda a_1,\ldots,a_n-\lambda a_n)}(\mathbf{u}_{[\lambda]}^{[i]}).$$

By Theorem 17:

$$\{\lim_{i\to\infty} \mathbf{u}_{[m]}^{[i]}\} = \hat{\Theta}_{(\lambda,a_1-\lambda a_1,\ldots,a_n-\lambda a_n)}^{D_f}(I, W_1, \ldots, W_n).$$

By the chairman theorem for $\hat{\Theta}_{\mathcal{A}}^{D_f}$:

$$\lim_{\lambda\searrow 0}\lim_{i\to\infty} \mathbf{u}_{[\lambda]}^{[i]} = \arg \min_{\mathbf{w}\in\hat{\Theta}_{\mathbf{a}}^{D_f}(W_1,\ldots,W_n)} D_f(\mathbf{w}\|\mathbf{v}) \tag{19}$$

i.e., equals the D_f-projection of the probability function $\mathbf{v}$ into $\hat{\Theta}_{\mathbf{a}}^{D_f}(W_1, \ldots, W_n)$.

In particular, to approximate $\mathbf{SEP}(W_1, \ldots, W_n)$ using Equation (19), one needs to choose a sufficiently small λ and then iterate the sequence $\{\mathbf{u}_{[\lambda]}^{[i]}\}_{i=0}^{\infty}$, where $\mathbf{u}_{[\lambda]}^{[0]} = \mathbf{v}$ is the uniform probability function, $\mathcal{A} = \mathcal{N}$ and $D_f = \mathrm{KL}$. However, the question of how to determine such an λ and i in order to achieve a specific level of accuracy merits further investigation.

The special case of the problem above when $W_1, \ldots, W_n$ have a nonempty intersection was extensively studied in the literature, and many scientific and engineering problems can be expressed as a problem of finding a point in such an intersection. Bregman in [7] showed the convergence of (what is now called) cyclic Bregman projections to a point in the intersection (the notion of a Bregman divergence is used only for the Euclidean space, but in [7], a more general topological vector space was considered). Many cyclic algorithms with appealing applications have been developed since then; see, e.g., [31,32].

Although the approach we propose offers the option of an empty intersection, it always leads to a meaningful point, and in particular, if the intersection is nonempty, it chooses a point inside the intersection; our study cannot be considered as an extension of the classical method of cyclic projections, which was developed over (possibly infinite) Banach spaces [33] in contrast to a limited discrete probabilistic space, which we are considering.

It is also worth mentioning that the method of cyclic projections, even in the case of an empty intersection, often provides more useful results than our method. An example is the noise reduction algorithm from [34].

One can perhaps conclude that the approach offered in this paper is at its best only another contribution to the problem of finding a point in a convex set by means of geometry, which, however, offers some interesting insights into the combination of Bregman projections with pooling operators.

Acknowledgments

The author is indebted to George Wilmers, whose support and wisdom allowed the creation of this paper. Thanks goes also to Alena Vencovská and František Matúš for sharing their ideas with me and to an anonymous reviewer for pointing out the connections to the dual affine structure in the probabilistic simplex.

The paper is an extension of some results that the author obtained as a Ph.D. student at the University of Manchester while supported by the (European Community's) Seventh Framework Programme (FP7/2007-2013) under Grant Agreement No. 238381.

Last, but not least, the author is grateful for the support received from the Assumption University in Thailand, without which the paper could not be finished.

Conflicts of Interest

The author declares no conflict of interest other than disclosed above in acknowledgments.

References

1. Amari, S. Divergence, Optimization and Geometry. In *Neural Information Processing: 16th International Conference*; Leung, C., Lee, M., Chan, J.H., Eds.; Iconip: Bangkok, Thailand, 2009; pp. 185–193.
2. Hájek, P.; Havránek, T.; Jiroušek, J. *Uncertain Information Processing in Expert Systems*; Raton B., Arbor, A., Eds.; CRC Press: London, UK, 1992.
3. Collins, M.; Schapire, R.E. Logistic Regression, AdaBoost and Bregman Distances. *Mach. Learn.* **2002**, *48*, 253–285.
4. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman Divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
5. Adamčík, M.; Wilmers, G.M. Probabilistic Merging Operators. *Log. Anal.* **2015**, in press.
6. De Finetti, B. Sul Significato Soggettivo della Probabilitá. *Fund. Math.* **1931**, *17*, 298–329.
7. Bregman, L.M. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 200–217.
8. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004; pp. 1–716.
9. Rockafeller, R.T. *Convex Analysis. Princeton Landmarks in Mathematics*; Princeton University Press: Princeton, NJ, USA, 1997; pp. 1–469.
10. Adamčík, M. Collective Reasoning under Uncertainty and Inconsistency. Ph.D. Thesis, The University of Manchester, Manchester, UK, 2014; pp. 1–150.
11. Csiszár, I. I-Divergence Geometry of Probability Distribution and Minimization Problems. *Ann. Probab.* **1975**, *3*, 146–158.
12. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; AMS and Oxford University Press: New York, NY, USA, 2000; pp. 1–206.
13. Jaynes, E.T. Where do we Stand on Maximum Entropy? In *The Maximum Entropy Formalism*; Levine, R.D., Tribus, M., Eds.; M.I.T. Press: Cambridge, MA, USA, 1979; pp. 15–118.
14. Paris, J.B.; Vencovská, A. On the Applicability of Maximum Entropy to Inexact Reasoning. *Int. J. Approx. Reason.* **1989**, *3*, 1–34.
15. Paris, J.B.; Vencovská, A. A Note on the Inevitability of Maximum Entropy. *Int. J. Approx. Reason.* **1990**, *4*, 183–224.

16. Vomlel, J. Methods of Probabilistic Knowledge Integration. Ph.D. Thesis, Czech Technical University: Prague, Czech, 1999; pp. 1–123.
17. Banerjee, A.; Guo, X.; Wang H. On the optimality of conditional expectation as a Bregman predictor. *IEEE Trans. Inf. Theory* **2005**, *51*, 2664–2669.
18. Wilmers, G.M. The Social Entropy Process: Axiomatising the Aggregation of Probabilistic Beliefs. In *Probability, Uncertainty and Rationality*; Hosni, H., Montagna, F., Eds.; CRM Series: Pisa, Italy, 2010; pp. 87–104.
19. Genest, C.; Zidek, J.V. Combining probability distributions: A critique and an annotated bibliography. *Stat. Sci.* **1986**, *1*, 114–135.
20. Genest, C.; Wagner, C.G. Further Evidence Against Independence Preservation in Expert Judgement Synthesis. *Aequ. Math.* **1986**, *32*, 74–86.
21. Matúš, F. On iterated averages of I-projections. In *Statistiek und Informatik*; Universität Bielefeld: Bielefeld, Germany, 2007; pp. 1–12.
22. Predd, J.B.; Osherson, D.N.; Kulkarni, S.R.; Poor, H.V. Aggregating Probabilistic Forecasts from Incoherent and Abstaining Experts. *Decis. Anal.* **2008**, *5*, 177–189.
23. Kern-Isberner, G.; Rödder, W. Belief Revision and Information Fusion on Optimum Entropy. *Int. J. Intel. Syst.* **2004**, *19*, 837–857.
24. Williamson, J. Deliberation, Judgement and the Nature of Evidence. *Econ. Philos.* **2014**, in press.
25. Carnap, R. On the application of inductive logic. *Philos. Phenomenol. Res.* **1947**, *8*, 133–148.
26. Paris, J.B. *The Uncertain Reasoner Companion*. Cambridge University Press: Cambridge, UK, 1994; pp. 1–224.
27. Amari, S. Integration of stochastic models by minimizing alpha-divergence. *Neural Comput.* **2007**, *19*, 2780–2796.
28. Csiszár, I.; Tusnády, G. Informational Geometry and Alternating Minimization Procedures. *Stat. Decis.* **1984**, *1*, 205–237.
29. Eggermont, P.P.B.; LaRiccia, V.N. On EM-like algorithms for minimum distance estimation. Preprint 1998, University of Delaware: Delaware, NC, USA; pp. 1–29.
30. Wilmers, G.M. Generalising the Maximum Entropy Inference Process to the Aggregation of Probabilistic Beliefs. Preprint 2011, Version 6; The University of Manchester: Manchester, UK; pp. 1–40.
31. Bauschke, H.H. Projection Algorithms and Monotone Operators. Ph.D. Thesis, Simon Fraser University: Burnaby, BC, Canada, 1996; pp. 1–223.
32. Censor, Y.; Zenios, S.A. *Parallel Optimization: Theory, Algorithms, and Applications*; Oxford University Press: New York, NY, USA, 1997; pp. 1–541.
33. Bauschke, H.H.; Borwein, J.M.; Combettes, P.L. Bregman monotone optimization algorithms. *SIAM J. Control Optim.* **2003**, *42*, 596–636.

34. Tofighi, M.; Kose, K.; Cetin, A.E. Denoising Using Projections Onto Convex Sets (POCS) Based Framework. **2013**, arXiv:1309.0700.