KYBERNETIKA — VOLUME *55* (2019), NUMBER 4, PAGES 605–617

A NOTE ON HOW RÉNYI ENTROPY CAN CREATE A SPECTRUM OF PROBABILISTIC MERGING OPERATORS

Martin Adamčík

In this paper we present a result that relates merging of closed convex sets of discrete probability functions respectively by the squared Euclidean distance and the Kullback–Leibler divergence, using an inspiration from the Rényi entropy. While selecting the probability function with the highest Shannon entropy appears to be a convincingly justified way of representing a closed convex set of probability functions, the discussion on how to represent several closed convex sets of probability functions is still ongoing. The presented result provides a perspective on this discussion. Furthermore, for those who prefer the standard minimisation based on the squared Euclidean distance, it provides a connection to a probabilistic merging operator based on the Kullback–Leibler divergence, which is closely connected to the Shannon entropy.

Keywords: probabilistic merging, information geometry, Kullback–Leibler divergence, Rényi entropy

Classification: 52A99, 52C99

1. INTRODUCTION

Information geometry is a branch of mathematics that explores geometry of probability functions (also called distributions). In this paper, we will discuss sequences of probability functions, and show a limit theorem that gives us a perspective on an ongoing discussion in the field. Over the last 70 years [18] Kullback, Leibler, Shannon, Jaynes, Amari, Cichocki and others explored applications to communication, uncertain reasoning, neuroscience and cryptography. Resolving the debate around representing several sets of probability functions could further support a recently introduced application of this field to meta–analysis [4].

Formally, we work here with the set $\mathbb{D}^J$ of positive discrete probability functions, which consists of ordered J–tuples $\mathbf{v} = (v_1, \dots, v_J)$ that satisfy $\sum_{j=1}^{J} v_j = 1$ and $v_1 > 0, \dots, v_J > 0$. In other words, $\mathbb{D}^J$ is a $(J-1)$–dimensional open simplex. J will be a fixed constant greater than two throughout the paper. We say that a subset W of $\mathbb{D}^J$ is convex if for any two $\mathbf{v}, \mathbf{w} \in W$

$$(\lambda \cdot v_1 + (1-\lambda) \cdot w_1, \dots, \lambda \cdot v_J + (1-\lambda) \cdot w_J) \in W,$$

DOI: 10.14736/kyb-2019-4-0605

for all $\lambda \in [0,1]$. Furthermore, we say that a subset W of $\mathbb{D}^J$ is closed if the limit point of every convergent sequence constructed from the elements of W has its limit inside W.

Now, a question as to how to represent what is essentially a closed convex set of probability functions by a single probability function has been extensively investigated, see for example [9, 13, 14, 15]. Such a set to represent was usually defined in an application motivated way, and the representation was to satisfy some rational criteria. For example, the theoretical model discussed in [14] was inspired by a problem of a physician differentiating between different types of tumor from sample slides. In this problem, the physician's knowledge was sufficient only to restrict the distribution of possible combinations of visual features and tumor types, which effectively yielded a closed convex set of possible probability functions. Rational criteria could involve for example a principle that irrelevant information must not influence the chosen representation.

A similar question concerning several closed convex sets of probability functions representing inconsistent knowledge was in various settings discussed in [1, 2, 5, 11, 12, 19]. Even a single physician may eventually make a series of mutually contradicting claims, and there is need to address how to represent opinion of a panel of experts and to merge findings of contradicting medical studies.

For the purposes stated above, all those referenced papers had to in some way introduce a way of measuring 'distance' between two probability functions. This was, it seems, without exception based on some convex Bregman divergence [8]. Note that a divergence takes as its two arguments two probability functions and outputs a single real number.

Prominent examples of such a divergence are the squared Euclidean distance

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

and the Kullback–Leibler divergence from $\mathbf{v}$ to $\mathbf{w}$

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

where in our consideration 'log' denotes the natural logarithm. Although using the logarithm to base 2 is common in information theory, this would merely scale the divergence by a constant and thus would have no influence on optimisation problems. Our specific choice is based on [3].

The squared Euclidean distance, exceptionally a symmetric divergence, is an attractive choice [16], but it is the Kullback–Leibler divergence that is related to a famous notion of entropy. In the context of information theory, maximising the Shannon entropy $-\sum_{j=1}^{J} w_j \log w_j$ [18], or alternatively minimising the KL–divergence from the uniform $(\frac{1}{J}, \dots, \frac{1}{J})$, gives us the choice that carries the least additional information beyond to that contained in a given piece of knowledge. When we talk about representing a single closed convex set of probability functions, maximising the Shannon entropy appears to be the most rational option [10].

One way of representing several closed convex sets of probability functions $W_1, \ldots, W_n$ weighted respectively by $\lambda_1 > 0, \ldots, \lambda_n > 0$, $\sum_{i=1}^n \lambda_i = 1$, is to minimise the sum

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

subject to $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in W_1 \times \ldots \times W_n$ and $\mathbf{v} \in \mathbb{D}^J$. To shorten the notation and to be more explicit, we will denote the set of these representatives by

$$\Theta^{\mathrm{KL}}_{\vec{\lambda}}(W_1, \ldots, W_n) = \Big\{ \arg\min_{\mathbf{v} \in \mathbb{D}^J} \sum_{i=1}^{n} \lambda_i \, \mathrm{KL}(\mathbf{w}^{(i)} \| \mathbf{v}); \, (\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in W_1 \times \ldots \times W_n \Big\},$$

which is a nonempty closed convex set [1, Theorem 12].

Looking at the definition above, $\Theta^{\mathrm{KL}}_{\vec{\lambda}}$ can be considered as a probabilitic merging operator acting on closed convex sets $W_1, \ldots, W_n \subseteq \mathbb{D}^J$ given a fixed $\vec{\lambda} = (\lambda_1, \ldots, \lambda_n)$. Note that the operator produces a single closed convex set as the representation of the given $W_1, \ldots, W_n$. The properties of this operator were investigated in [5], where it was denoted by $\hat{\Delta}^{\mathrm{KL}}_{\vec{\lambda}}$. The notation we use in this paper is from [1].

Additional probabilistic merging operators can be defined by replacing KL with a different convex Bregman divergence, or by swapping the two arguments of the divergence. Unlike in the case of representing a single closed convex set of probability functions where maximising the Shannon entropy is widely preferred, exploring various probabilistic merging operators is still an ongoing effort into which the present paper contributes.

Among a few existing results, we mention here a practical application of $\Theta^{\mathrm{KL}}_{\vec{\lambda}}$: In [3] it was argued that this operator is optimal when we combine several heterogeneous studies that present us with complex knowledge, assuming that the heterogeneity cannot be explained. A particular implementation discussing meta–analysis of several studies concerning the one–year incidence of cancer among patients with unprovoked venous thromboembolism was introduced in [4].

An idea to connect $\Theta^{\mathrm{KL}}_{\vec{\lambda}}$ with similarly defined $\Theta^{\mathrm{E}}_{\vec{\lambda}}$ (where the KL–divergence is replaced by the squared Euclidean distance) using a spectrum of probabilistic merging operators, originally suggested by George Wilmers in 2013, was mentioned in [2]. This was to be done by using a divergence from $\mathbf{v}$ to $\mathbf{w}$ given by

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

where $2 \geq r > 1$, but with no progress reported since then. This special Bregman divergence, which is illustrated in Figure 1, is related to the Rényi entropy $\frac{1}{1-r} \log \sum_{j=1}^{J} (w_j)^r$ [17], although the corresponding Rényi divergence is usually defined differently [6]. On the other hand, it is essentially the β divergence [7], which is defined as

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

where $2 \geq r > 1$. The β divergence is already known to satisfy a limiting theorem that relates it to the Kullback–Leibler divergence, and it is well established and investigated in information theory [6].

The proof that

$$\Theta_{\vec{\lambda}}^{D_r}(W_1,\ldots,W_n) = \Big\{ \arg\min_{\mathbf{v}\in\mathbb{D}^J} \sum_{i=1}^{n} \lambda_i D_r(\mathbf{w}^{(i)}\|\mathbf{v});\ (\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(n)}) \in W_1\times\ldots\times W_n \Big\},$$

is well defined; in particular, that D_r, $2 \geq r > 1$, is a convex function, is in [2, Theorem 2.1.18]. Note that $D_2 = \mathrm{E}$, the squared Euclidean distance. In this paper we will prove the following theorem, which suggests that there could be a spectrum of probabilistic merging operators between operators $\Theta_{\vec{\lambda}}^{\mathrm{E}}$ and $\Theta_{\vec{\lambda}}^{\mathrm{KL}}$. This provides a perspective on the ongoing discussion about probabilistic merging operators, and, for those who prefer the standard minimisation based on the squared Euclidean distance, it provides a connection to what seems to be (in a context of meta–analysis) a somewhat justified probabilistic merging operator.

Theorem 1.1. For any $W_1,\ldots,W_n \subseteq \mathbb{D}^J$ that are closed, convex and nonempty, and for any $\vec{\lambda}$ such that $\sum_{i=1}^{n}\lambda_i = 1$, $\lambda_i > 0$ for $1 \leq i \leq n$, we have that

$$\emptyset \neq \lim_{r\searrow 1} \Theta_{\vec{\lambda}}^{D_r}(W_1,\ldots,W_n) \subseteq \Theta_{\vec{\lambda}}^{\mathrm{KL}}(W_1,\ldots,W_n).$$

To better explain the statement of the theorem, it does not mean that every sequence $\{\mathbf{v}^{(k)}\}_{k=1}^{\infty}$, $\mathbf{v}^{(k)} \in \Theta_{\vec{\lambda}}^{D_{r_k}}(W_1,\ldots,W_n)$, where $r_k \searrow 1$ as $k\to\infty$, is convergent. In fact, one can come up with a simple counterexample where $W_1 = \ldots = W_n$ are not singletons, in which case $\Theta_{\vec{\lambda}}^{D_r}(W_1,\ldots,W_n) = \Theta_{\vec{\lambda}}^{\mathrm{KL}}(W_1,\ldots,W_n) = W_i$, for all $r > 1$, and simply let $\mathbf{v}^{(k)}$ alternate. Instead, we define $\lim_{r\searrow 1}\Theta_{\vec{\lambda}}^{D_r}(W_1,\ldots,W_n)$ as the set of all existing limits of such sequences: $\lim_{k\to\infty}\mathbf{v}^{(k)}$ exists if and only if

$$\lim_{k\to\infty}\mathbf{v}^{(k)} \in \lim_{r\searrow 1}\Theta_{\vec{\lambda}}^{D_r}(W_1,\ldots,W_n).$$

Theorem 1.1 can be then reformulated:

1. There exists a convergent sequence $\{\mathbf{v}^{(k)}\}_{k=1}^{\infty}$, $\mathbf{v}^{(k)} \in \Theta_{\vec{\lambda}}^{D_{r_k}}(W_1,\ldots,W_n)$, where $r_k \searrow 1$ as $k\to\infty$.

2. If a sequence $\{\mathbf{v}^{(k)}\}_{k=1}^{\infty}$, $\mathbf{v}^{(k)} \in \Theta_{\vec{\lambda}}^{D_{r_k}}(W_1,\ldots,W_n)$, where $r_k \searrow 1$ as $k\to\infty$, is convergent then
$$\lim_{k\to\infty}\mathbf{v}^{(k)} \in \Theta_{\vec{\lambda}}^{\mathrm{KL}}(W_1,\ldots,W_n).$$

It is still an open problem whether or not we can place the equality in the expression;

$$\lim_{r\searrow 1}\Theta_{\vec{\lambda}}^{D_r}(W_1,\ldots,W_n) \overset{?}{=} \Theta_{\vec{\lambda}}^{\mathrm{KL}}(W_1,\ldots,W_n),$$

which would be necessary for the spectrum of probabilistic merging operators to be well defined. An additional open problem is to create a similar spectrum of probabilitic merging operators related to the social entropy operator [5] originally introduced by George Wilmers [19].

Before proving Theorem 1.1, some examples and several auxiliary results now follow.

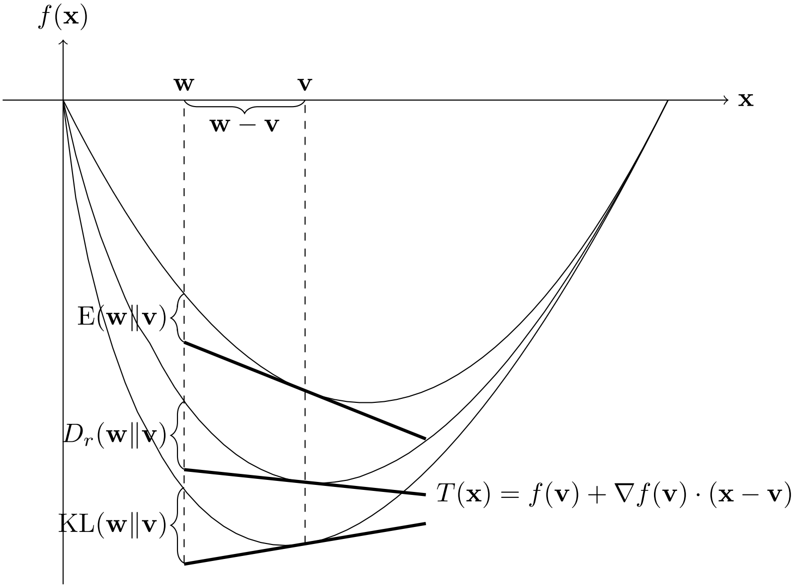
$f(\mathbf{x})$
$\mathbf{w}$
$\mathbf{v}$
$\mathbf{x}$
$\mathbf{w}-\mathbf{v}$
$\mathrm{E}(\mathbf{w}\|\mathbf{v})$
$D_r(\mathbf{w}\|\mathbf{v})$
$\mathrm{KL}(\mathbf{w}\|\mathbf{v})$
$T(\mathbf{x}) = f(\mathbf{v}) + \nabla f(\mathbf{v}) \cdot (\mathbf{x}-\mathbf{v})$

Fig. 1. A single–dimensional impression of divergences E, D_r and KL is depicted below. Each divergence is defined as a Bregman divergence $f(\mathbf{w}) - [f(\mathbf{v}) + \nabla f(\mathbf{v}) \cdot (\mathbf{w}-\mathbf{v})]$, where the strictly convex and differentiable function $f(\mathbf{x})$ is respectively $\sum_{i=1}^n [(x_j)^2 - x_j]$, $\frac{1}{r-1}\sum_{i=1}^n [(x_j)^r - x_j]$ and $\sum_{i=1}^n x_j \log x_j$. The hyperplane $T(\mathbf{x}) = f(\mathbf{v}) + \nabla f(\mathbf{v}) \cdot (\mathbf{x}-\mathbf{v})$, where $\nabla f(\mathbf{v})$ is the gradient and '$\cdot$' is the dot product, is in this single–dimensional impression the tangent line to f at $\mathbf{v}$. The divergence from $\mathbf{v}$ to $\mathbf{w}$ is then given by the difference $f(\mathbf{w}) - T(\mathbf{w})$, which is explicitly shown below for E, D_r and KL, respectively.

2. EXAMPLES

A probabilistic merging operator is usually expected to represent closed convex sets of probability functions $W_1, \dots, W_n$ weighted respectively by $\lambda_1 > 0, \dots, \lambda_n > 0$, $\sum_{i=1}^n \lambda_i = 1$, in a rational way; in [5] several principles were suggested. Among those, the consistency principle asserts that whenever $\bigcap_{i=1}^n W_i \neq \emptyset$ then the resulting set of representatives is equal to $\bigcap_{i=1}^n W_i$. It is obvious directly from their definitions that $\Theta_{\vec{\lambda}}^{\mathrm{E}}$, $\Theta_{\vec{\lambda}}^{D_r}$ and $\Theta_{\vec{\lambda}}^{\mathrm{KL}}$ satisfy the consistency principle. Having this in mind, we can introduce the following trivial example illustrated in Figure 2.

Example 2.1. Let $W_1, \dots, W_n$ be such that $\bigcap_{i=1}^n W_i \neq \emptyset$, and $\lambda_1 > 0, \dots, \lambda_n > 0$, $\sum_{i=1}^n \lambda_i = 1$, be arbitrary. Then in this case

$$\lim_{r \searrow 1} \Theta_{\vec{\lambda}}^{D_r}(W_1, \dots, W_n) = \lim_{r \searrow 1} \bigcap_{i=1}^n W_i = \bigcap_{i=1}^n W_i = \Theta_{\vec{\lambda}}^{\mathrm{KL}}(W_1, \dots, W_n).$$

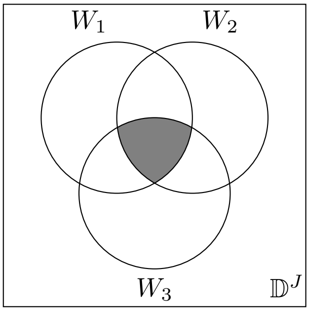

Fig. 2. W_1, W_2 and W_3, which are closed and convex sets in a simplex $\mathbb{D}^J$, are represented below as circles to make an impression of convexity. The shaded area $W_1 \cap W_2 \cap W_3$ corresponds to both $\Theta^{D_r}_{\vec{\lambda}}(W_1, W_2, W_3)$, $2 \geq r > 1$, and $\Theta^{\mathrm{KL}}_{\vec{\lambda}}(W_1, W_2, W_3)$. This is because $D_r(\mathbf{v}\|\mathbf{v}) = \mathrm{KL}(\mathbf{v}\|\mathbf{v}) = 0$, and $D_r(\mathbf{w}\|\mathbf{v}) > 0$ and $\mathrm{KL}(\mathbf{w}\|\mathbf{v}) > 0$ whenever $\mathbf{w} \neq \mathbf{v}$.

Although $\Theta^{D_r}_{\vec{\lambda}}$, $2 \geq r > 1$, and $\Theta^{\mathrm{KL}}_{\vec{\lambda}}$ in general produce sets, if at least one of $W_1, \dots, W_n \subseteq \mathbb{D}^J$ is a singleton then so are $\Theta^{D_r}_{\vec{\lambda}}(W_1, \dots, W_n)$ and $\Theta^{\mathrm{KL}}_{\vec{\lambda}}(W_1, \dots, W_n)$ [1, Theorem 18]. We will use this in the following non–trivial example illustrated in Figure 3.

Example 2.2. Let $J = 3$, $W_1 = \{(\frac{3}{6}, \frac{1}{6}, \frac{2}{6})\}$ be a singleton, and $W_2 = \{(\frac{4}{6} - x, \frac{1}{6} + x, \frac{1}{6}); x \in (0, \frac{1}{6})\}$ be a line segment. Let us weight the two closed convex sets equally: $\lambda_1 = \lambda_2 = \frac{1}{2}$. Then, after rounding to four decimal places,

$$\Theta^{D_2}_{\vec{\lambda}}(W_1, W_2) = \Big\{\Big(0.5417, 0.2083, 0.25\Big)\Big\},$$

$$\Theta^{D_{1.5}}_{\vec{\lambda}}(W_1, W_2) = \Big\{\Big(0.5518, 0.1982, 0.25\Big)\Big\},$$

$$\Theta^{D_{1.1}}_{\vec{\lambda}}(W_1, W_2) = \Big\{\Big(0.5604, 0.1896, 0.25\Big)\Big\},$$

$$\Theta^{D_{1.01}}_{\vec{\lambda}}(W_1, W_2) = \Big\{\Big(0.5623, 0.1877, 0.25\Big)\Big\},$$

$$\Theta^{D_{\mathrm{KL}}}_{\vec{\lambda}}(W_1, W_2) = \Big\{\Big(0.5625, 0.1875, 0.25\Big)\Big\}.$$

This illustrates that in this case

$$\lim_{r \searrow 1} \Theta^{D_r}_{\vec{\lambda}}(W_1, W_2) = \Theta^{\mathrm{KL}}_{\vec{\lambda}}(W_1, W_2).$$

Finally, let us repeat that it is an open question whether or not there is an example showing that the inclusion discussed in the theorem could be strict;

$$\lim_{r \searrow 1} \Theta^{D_r}_{\vec{\lambda}}(W_1, \dots, W_n) \overset{?}{\subset} \Theta^{\mathrm{KL}}_{\vec{\lambda}}(W_1, \dots, W_n).$$

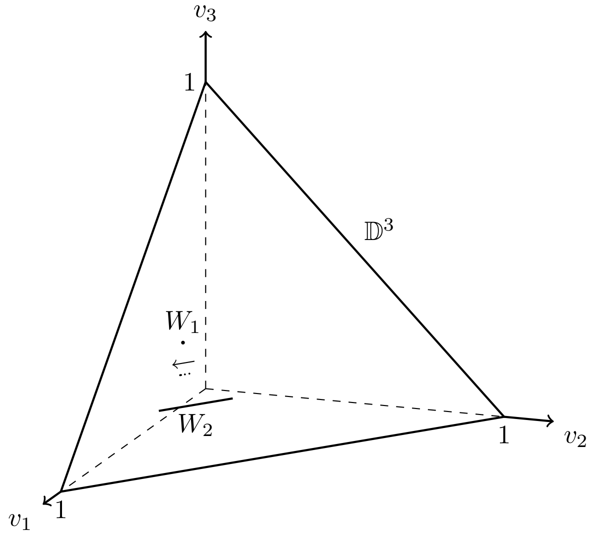

Fig. 3. Two sets $W_1 = \{(\frac{3}{6}, \frac{1}{6}, \frac{2}{6})\}$ and $W_2 = \{(\frac{4}{6} - x, \frac{1}{6} + x, \frac{1}{6}); x \in (0, \frac{1}{6})\}$ to represent are shown in the two–dimensional simplex $\mathbb{D}^3$ below, which is an object in a three–dimensional space $\mathbb{R}^3$. The arrow indicates the direction in which the unique points inside $\Theta^{D_r}_{\vec{\lambda}}(W_1, W_2)$, $2 \geq r > 1$, converge to the unique point inside $\Theta^{\mathrm{KL}}_{\vec{\lambda}}(W_1, W_2)$, with the actual points shown directly below the arrow. This figure is to scale.

3. AUXILIARY RESULTS

First, we will need the following lemma from [1, Lemma 2].

Lemma 3.1. Let D be either D_r or KL, $W_1, \ldots, W_n$ be closed convex and nonempty sets of probability functions, and $\lambda_1 > 0, \ldots, \lambda_n > 0$ satisfy $\sum_{i=1}^n \lambda_i = 1$. Then the following are equivalent:

1. Probability functions $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}, \mathbf{v}$ minimise the sum

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

   subject to $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in W_1 \times \ldots \times W_n$ and $\mathbf{v} \in \mathbb{D}^J$.

2. Probability functions $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}$ minimise the sum

$$\sum_{i=1}^n \lambda_i D\Big(\mathbf{w}^{(i)} \Big\| \sum_{k=1}^n \lambda_k \mathbf{w}^{(k)}\Big),$$

subject to $(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}) \in W_1 \times \ldots \times W_n$, where $\sum_{k=1}^n \lambda_k \mathbf{w}^{(k)} \in \mathbb{D}^J$ is in every coordinate defined by a weighted arithmetic mean $\sum_{k=1}^n \lambda_k w_j^{(k)}$, $1 \leq j \leq J$; and the corresponding $\mathbf{v}$ from Part 1 of this lemma is uniquely determined by

$$\mathbf{v} = \sum_{k=1}^{n} \lambda_k \mathbf{w}^{(k)}.$$

Second, Peter Hawes has shown [9, Page 62] that $\frac{x^r-1}{r}$ uniformly converges to $\log x$ as $r \searrow 0$ if $0 < \delta \leq x \leq 1$. This is equivalent to saying that $y\frac{x^{r-1}-1}{r-1}$ uniformly converges to $y \log x$ as $r \searrow 1$ if $0 < \delta \leq x \leq 1$. (In Figure 1 one could imagine this as the middle curve being transformed uniformly from the top curve to the bottom curve.) His proof was the inspiration for the following straightforward modification, which is the key to prove Theorem 1.1.

Lemma 3.2. Let there be a $\delta > 0$ such that v and w are confined to $[\delta, 1]$. Then

$$\frac{w^r}{r-1} - \frac{v^r}{r-1} - \frac{r}{r-1}(w-v)v^{r-1} - v + w,$$

where $r \in (1, 2]$, uniformly converges to

$$w \log \frac{w}{v}$$

as $r \searrow 1$.

Proof. We can rewrite the first expression (with the parameter r) using

$$-\int_w^1 wx^{r-2}\,\mathrm{d}x + \int_v^1 wry^{r-2}\,\mathrm{d}y - \int_v^1 vry^{r-2}dy + \int_v^1 vy^{r-2}\,\mathrm{d}y, \tag{1}$$

which is equal to

$$\begin{aligned} &-\Big[w\frac{x^{r-1}}{r-1}\Big]_{x=w}^{x=1} + \Big[wr\frac{y^{r-1}}{r-1}\Big]_{y=v}^{y=1} - \Big[vr\frac{y^{r-1}}{r-1}\Big]_{y=v}^{y=1} + \Big[v\frac{y^{r-1}}{r-1}\Big]_{y=v}^{y=1} \\ &= \Big[-\frac{w}{r-1} + \frac{w^r}{r-1}\Big] + \Big[\frac{r}{r-1}w - \frac{r}{r-1}wv^{r-1}\Big] \\ &+\Big[-\frac{r}{r-1}v + \frac{r}{r-1}v^r\Big] + \Big[\frac{v}{r-1} - \frac{v^r}{r-1}\Big] \\ &= \frac{w^r}{r-1} - \frac{v^r}{r-1} - \frac{r}{r-1}wv^{r-1} + \frac{r}{r-1}v^r \\ &-\frac{r}{r-1}v + \frac{v}{r-1} - \frac{w}{r-1} + \frac{r}{r-1}w. \end{aligned}$$

Since $-\frac{r}{r-1}v + \frac{v}{r-1} = -v$ and $-\frac{w}{r-1} + \frac{r}{r-1}w = w$, the above becomes

$$\frac{w^r}{r-1} - \frac{v^r}{r-1} - \frac{r}{r-1}(w-v)v^{r-1} - v + w.$$

Our aim is to show that (1), as $r \searrow 1$, uniformly converges to

$$-\int_w^1 wx^{-1}\,\mathrm{d}x + \int_v^1 wy^{-1}\,\mathrm{d}y - \int_v^1 vy^{-1}dy + \int_v^1 vy^{-1}\,\mathrm{d}y, \tag{2}$$

which then would be equal to

$$-[w\log x]_{x=w}^{x=1} + [w\log y]_{y=v}^{y=1} = w\log w - w\log v = w\log\frac{w}{v}.$$

To this end, and without loss of generality, we consider the difference

$$\begin{aligned}&\Big|\int_v^1 wry^{r-2}\,\mathrm{d}y - \int_v^1 wy^{-1}\,\mathrm{d}y\Big|\\ &= \Big|\int_v^1 wry^{r-2} - wy^{-1}\,\mathrm{d}y\Big|.\end{aligned} \tag{3}$$

Now, by the assumption of the lemma, there is δ such that $y \in [\delta, 1]$. Therefore,

$$\Big|wry^{r-2} - wy^{-1}\Big|,$$

as a function of y, has its maximum over $[\delta, 1]$ actually at δ, given $\delta > 0$ is sufficiently small. See Figure 4 for further explanation. Hence, (3) is less than the area of a rectangle having one side equal to this maximal value and the other side equal to one;

$$\Big|wr(\delta)^{r-2} - w(\delta)^{-1}\Big|,$$

and we may establish that

$$\begin{aligned}\lim_{r\searrow 1}\Big|\int_v^1 wry^{r-2}\,\mathrm{d}y - \int_v^1 wy^{-1}\,\mathrm{d}y\Big| &\le \lim_{r\searrow 1}\Big|wr(\delta)^{r-2} - w(\delta)^{-1}\Big|\\ &= \Big|w(\delta)^{-1} - w(\delta)^{-1}\Big| = 0.\end{aligned}$$

The argument above is nearly identical for all four integral pairs, so we have indeed that (1) uniformly converges to (2) as $r \searrow 1$, and the statement of the lemma follows. □

4. PROOF OF THEOREM 1.1

Notice that for a fixed r; $1 < r \le 2$, $\Theta_{\vec{\lambda}}^{D_r}(W_1, \ldots, W_n) =$

$$\Big\{\arg\min_{\mathbf{v}\in\mathbb{D}^J}\sum_{i=1}^n \lambda_i \sum_{j=1}^J [(w_j^{(i)})^r - (v_j)^r - r(w_j^{(i)} - v_j)(v_j)^{r-1}];$$

$$(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in W_1 \times \ldots \times W_n\Big\}$$

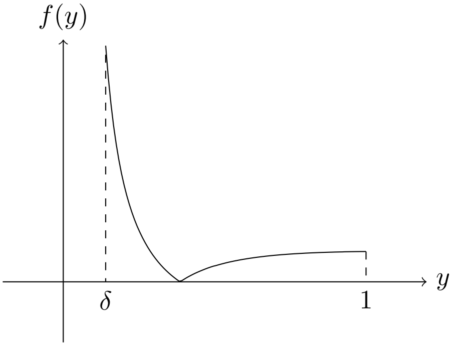

Fig. 4. $f(y) = |wry^{r-2} - wy^{-1}|$ for $r = 1.1$ is plotted below. The function has its maximum at $y = \delta$ as $y^{-1} - ry^{r-2}$ is actually strictly decreasing over $[\delta, 1]$. This can be observed by considering $\frac{d}{dy}[y^{-1} - ry^{r-2}] = -y^{-2} - r(r-2)y^{r-3} = -y^{-2}(1 + r(r-2)y^{r-1})$. The expression is negative since, for $r > 1$ and $y \in [\delta, 1]$, we have $r(r-2) > -1$ and $0 < y^{r-1} \leq 1$. Finally, for sufficiently small $\delta > 0$, clearly $f(\delta) > f(1) = w(r-1)$.

is equivalent to

$$\Big\{\arg\min_{\mathbf{v}\in\mathbb{D}^J} \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{J} \Big[\frac{(w_j^{(i)})^r}{r-1} - \frac{(v_j)^r}{r-1} - \frac{r}{r-1}(w_j^{(i)} - v_j)(v_j)^{r-1} - v_j + w_j^{(i)}\Big];$$

$$(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in W_1 \times \ldots \times W_n\Big\}. \tag{4}$$

This is because $\sum_{j=1}^{J} v_j = 1$ and $\sum_{j=1}^{J} w_j^{(i)} = 1$ for every $1 \leq i \leq n$ (see the definition of $\mathbb{D}^J$), and scaling by a fixed $r - 1 > 0$ does not make any difference to the resulting set of minimisers.

Since we assume that all $W_1, \ldots, W_n$ are closed sets, each $\mathbf{w}^{(i)} \in W_i$, $1 \leq i \leq n$, is bounded away from zero in every coordinate. Due to Lemma 3.1 every

$$\mathbf{v} \in \Theta_{\vec{\lambda}}^{D_r}(W_1, \ldots, W_n)$$

is given by

$$v_j = \sum_{i=1}^{n} \lambda_i w_j^{(i)},\ 1 \leq j \leq J,$$

for some $\mathbf{w}^{(1)} \in W_1, \ldots, \mathbf{w}^{(n)} \in W_n$ so $\Theta_{\vec{\lambda}}^{D_r}(W_1, \ldots, W_n)$ is likewise bounded away from zero. Therefore, given $W_1, \ldots, W_n$, there is a constant $\delta > 0$ such that confining (4) into $[\delta, 1]^J \subseteq \mathbb{D}^J$ makes no difference to the resulting set of minimisers. This, together with the fact that finite summing preserves uniform convergence, will allows us to use Lemma 3.2 in what follows.

First, let M be the minimal value of

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

subject to $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in W_1 \times \ldots \times W_n$ and $\mathbf{v} \in \mathbb{D}^J$. Let also $\mathbf{w}^{(i,k)}$, $1 \leq i \leq n$, minimise

$$\sum_{i=1}^{n} \lambda_i D_{r_k}(\mathbf{w}^{(i,k)} \| \mathbf{v}^{(k)})$$

subject to $(\mathbf{w}^{(1,k)}, \ldots, \mathbf{w}^{(n,k)}) \in W_1 \times \ldots \times W_n$, for a fixed $\mathbf{v}^{(k)} \in \Theta_{\vec{\lambda}}^{D_{r_k}}(W_1, \ldots, W_n)$. (They are unique since D_r is strictly convex in the first argument; as is every Bregman divergence.) Presume that there is a sequence $\{\mathbf{v}^{(k)}\}_{k=1}^{\infty}$,

$$\mathbf{v}^{(k)} \in \Theta_{\vec{\lambda}}^{D_{r_k}}(W_1, \ldots, W_n),$$

where $r_k \searrow 1$ as $k \to \infty$, such that for all k

$$\sum_{i=1}^{n} \lambda_i \operatorname{KL}(\mathbf{w}^{(i,k)} \| \mathbf{v}^{(k)}) > M + \epsilon,$$

for some fixed $\epsilon > 0$. Due to the uniform convergence discussed above, there is k_0 such that for all $k > k_0$

$$\left| \sum_{i=1}^{n} \lambda_i \operatorname{KL}(\mathbf{w}^{(i,k)} \| \mathbf{v}^{(k)}) - \sum_{i=1}^{n} \lambda_i D_{r_k}(\mathbf{w}^{(i,k)} \| \mathbf{v}^{(k)}) \right| < \frac{\epsilon}{4}$$

and

$$\left| M - \sum_{i=1}^{n} \lambda_i D_{r_k}(\mathbf{w}^{(i)} \| \mathbf{v}) \right| < \frac{\epsilon}{4},$$

where $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}, \mathbf{v}$ specified in the last inequality jointly minimise

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

subject to $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)}) \in W_1 \times \ldots \times W_n$ and $\mathbf{v} \in \mathbb{D}^J$ (they give the value M). Finally, considering the last three inequalities together,

$$\sum_{i=1}^{n} \lambda_i D_{r_k}(\mathbf{w}^{(i,k)} \| \mathbf{v}^{(k)}) > \sum_{i=1}^{n} \lambda_i D_{r_k}(\mathbf{w}^{(i)} \| \mathbf{v}) + \frac{\epsilon}{2},$$

which is a contradiction with $\mathbf{v}^{(k)} \in \Theta_{\vec{\lambda}}^{D_{r_k}}(W_1, \ldots, W_n)$.

Since the above lead to a contradiction, we have that

$$\lim_{k \to \infty} \sum_{i=1}^{n} \lambda_i \operatorname{KL}(\mathbf{w}^{(i,k)} \| \mathbf{v}^{(k)}) = M$$

for all sequences $\{\mathbf{v}^{(k)}\}_{k=1}^{\infty}$, $\mathbf{v}^{(k)} \in \Theta_{\vec{\lambda}}^{D_{r_k}}(W_1, \ldots, W_n)$, where $r_k \searrow 1$ as $k \to \infty$.

Assume that $\lim_{k\to\infty} \mathbf{v}^{(k)}$ exists; and hence $\lim_{k\to\infty} \mathbf{v}^{(k)} \in \lim_{r\searrow 1} \Theta_{\vec{\lambda}}^{D_r}(W_1, \ldots, W_n)$. Even if every limit $\lim_{k\to\infty} \mathbf{w}^{(i,k)}$ does not exist, since we operate in the compact space $W_1 \times \ldots \times W_n$, there is a convergent subsequence $\{(\mathbf{w}^{(1,k_p)}, \ldots, \mathbf{w}^{(n,k_p)})\}_{p=1}^{\infty}$, and we will denote its limit $(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(n)})$. Then, since $\mathrm{KL}(\cdot\|\cdot)$ is a continuous function over the considered domain (bounded away from zero in every coordinate),

$$\lim_{k\to\infty} \sum_{i=1}^{n} \lambda_i \,\mathrm{KL}(\mathbf{w}^{(i,k)} \| \mathbf{v}^{(k)}) = \sum_{i=1}^{n} \lambda_i \,\mathrm{KL}(\lim_{p\to\infty} \mathbf{w}^{(i,k_p)} \| \lim_{p\to\infty} \mathbf{v}^{(k_p)})$$

$$= \sum_{i=1}^{n} \lambda_i \,\mathrm{KL}(\mathbf{w}^{(i)} \| \lim_{k\to\infty} \mathbf{v}^{(k)}) = M.$$

This means that

$$\lim_{r\searrow 1} \Theta_{\vec{\lambda}}^{D_r}(W_1, \ldots, W_n) \subseteq \Theta_{\vec{\lambda}}^{\mathrm{KL}}(W_1, \ldots, W_n).$$

Since we operate in a compact space, some limits $\lim_{k\to\infty} \mathbf{v}^{(k)}$ indeed exist so we have that

$$\emptyset \neq \lim_{r\searrow 1} \Theta_{\vec{\lambda}}^{D_r}(W_1, \ldots, W_n).$$

Note that in the above proof we can in fact show that if $\lim_{k\to\infty} \mathbf{v}^{(k)}$ exists then every $\lim_{k\to\infty} \mathbf{w}^{(i,k)}$ must exist. If not, considering that an infinite number of members of the sequence $\{\mathbf{w}^{(i,k)}\}_{k=1}^{\infty}$ that do not belong to the above constructed convergent sequence, we can find another convergent sequence. If its limit, say $\mathbf{u}^{(i)}$, is distinct from $\mathbf{w}^{(i)}$ for some i, we have also

$$\sum_{i=1}^{n} \lambda_i \,\mathrm{KL}(\mathbf{u}^{(i)} \| \lim_{k\to\infty} \mathbf{v}^{(k)}) = M.$$

In other words, we obtain one $\lim_{k\to\infty} \mathbf{v}^{(k)} \in \Theta_{\vec{\lambda}}^{\mathrm{KL}}(W_1, \ldots, W_n)$, but multiple KL–projections of $\lim_{k\to\infty} \mathbf{v}^{(k)}$ into W_i, for some i. However, that is a contradiction as such projections are unique since $\mathrm{KL}(\cdot\|\cdot)$ is strictly convex in its first argument and W_i is closed and convex.

ACKNOWLEDGEMENT

The author is grateful for the support received from Assumption University of Thailand, and for the suggestions provided by anonymous reviewers.

(Received April 28, 2018)

REFERENCES

[1] M. Adamčík: The information geometry of Bregman divergences and some applications in multi–expert reasoning. Entropy *16* (2014), 6338–6381. DOI:10.3390/e16126338

[2] M. Adamčík: Collective Reasoning under Uncertainty and Inconsistency. Phd Thesis, University of Manchester, Manchester 2014, `http://eprints.ma.man.ac.uk/2110/`.

[3] M. Adamčík: On the applicability of the 'number of possible states' argument in multi–expert reasoning. J. Appl. Logic *19* (2016), 20–49. DOI:10.1016/j.jal.2016.10.001

[4] M. Adamčík: A logician's approach to meta–analysis with unexplained heterogeneity. J. Biomed. Inform. *71* (2017), 110–129. DOI:10.1016/j.jbi.2017.05.017

[5] M. Adamčík and G. M. Wilmers: Probabilistic merging operators. Logique Analyse *228* (2014), 563–590. DOI:10.2143/LEA.228.0.3078175

[6] S. Amari and A. Cichocki: Families of Alpha– Beta– and Gamma– divergences: Flexible and robust measures of similarities. Entropy *12* (2010), 1532–1568. DOI:10.3390/e12061532

[7] A. Basu, I. R. Harris, N. Hjort, and M. Jones: Robust and efficient estimation by minimising a density power divergence. Biometrika *85* (1998), 549–559. DOI:10.1093/biomet/85.3.549

[8] L. M. Bregman: The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. USSR Comput. Mathematics Math. Physics *1* (1967), 200–217. DOI:10.1016/0041-5553(67)90040-7

[9] P. Hawes: Investigation of Properties of Some Inference Processes. Phd Thesis, University of Manchester, Manchester 2007, `http://eprints.ma.man.ac.uk/1304/`.

[10] E. T. Jaynes: Where do we stand on maximum entropy? In: The Maximum Entropy Formalism (R. D. Levine, M. Tribus, eds.), M.I.T. Press, 1979, pp. 15–118.

[11] G. Kern-Isberner and W. Rödder: Belief revision and information fusion on optimum entropy. Int. J. Intell. Systems *19* (2004), 837–857. DOI:10.1002/int.20027

[12] D. Osherson and M. Vardi: Aggregating disparate estimates of chance. Games Econom. Behavior *56* (2006), 148–173. DOI:10.1016/j.geb.2006.04.001

[13] J. B. Paris: The Uncertain Reasoner Companion. Cambridge University Press, Cambridge 1994.

[14] J. B. Paris and A. Vencovská: On the applicability of maximum entropy to inexact reasoning. Int. J. Approx. Reason. *3* (1989), 1–34. DOI:10.1016/0888-613x(89)90012-1

[15] J. B. Paris and A. Vencovská: A note on the inevitability of maximum entropy. Int. J. Approx. Reason. *4* (1990), 183–224. DOI:10.1016/0888-613x(90)90020-3

[16] J. B. Predd, D. N. Osherson, S. R Kulkarni, and H. V. Poor: Aggregating probabilistic forecasts from incoherent and abstaining experts. Decision Analysis *5* (2008), 177–189. DOI:10.1287/deca.1080.0119

[17] A. Rényi: On measures of entropy and information. In: Proc. Fourth Berkeley Symposium on Mathematics, Statistics and Probability *1* (1961), 547–561.

[18] C. E. Shannon: A mathematical theory of communication. Bell System Techn. J. *27* (1948), 379–423, 623-656. DOI:10.1002/j.1538-7305.1948.tb00917.x

[19] G. M. Wilmers: A foundational approach to generalising the maximum entropy inference process to the multi–agent context. Entropy *17* (2015), 594–645. DOI:10.3390/e17020594

Martin Adamčík, Martin de Tours School of Management and Economics, Assumption University, 10540 Samut Prakan. Thailand.
e-mail: maths38@gmail.com