International Journal of Approximate Reasoning 52 (2011) 659–671

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Combining marginal probability distributions via minimization of weighted sum of Kullback–Leibler divergences

Jan Kracík

Department of Adaptive Systems, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague, Czech Republic

ARTICLE INFO

Article history:
Received 3 March 2010
Revised 14 January 2011
Accepted 14 January 2011
Available online 22 January 2011

Keywords:
Combining probabilities
Kullback–Leibler divergence
Maximum likelihood
Expert opinions
Linear opinion pool

ABSTRACT

This paper deals with the problem of combining marginal probability distributions as a means for aggregating pieces of expert information. A novel approach, which takes the combining problem as an analogy of statistical estimation, is proposed and discussed. The combined distribution is then searched as a minimizer of a weighted sum of Kullback–Leibler divergences of the given marginal distributions and corresponding marginals of the searched one. Necessary and sufficient conditions for a distribution to be a minimizer are stated. For discrete random variables an iterative algorithm for approximate solution of the minimization problem is proposed and its convergence is proved.

1. Introduction

Any problem of decision making under uncertainty inevitably relies on some kind of expert information but frequently it happens that several inconsistent pieces of expert information are available to the decision maker. In such cases it is typically preferred to aggregate the information pieces into a single one before they enter the decision making procedure. Such kind of an aggregation problem is considered in this paper. Namely, the expert information pieces are supposed to be expressed in a form of probability distributions. The problem of aggregating expert information then transforms to the problem of combining several probability distributions into a single one.

1.1. State of the art

Not surprisingly, plenty of various combining procedures exist in the literature. Extensive bibliographies can be found, e.g., in the review papers [1,2]. Clemen and Winkler in [2] distinguish two types of combining procedures: mathematical and behavioral ones. The behavioral procedures attempt to reach the aggregated information through some kind of interaction among experts. In what follows we restrict our attention to the mathematical procedures. According to [2], two major classes of mathematical procedures for combining probability distributions can be further identified: axiomatic and Bayesian ones.

The specific feature of axiomatic procedures is that they produce probability distributions possessing certain characteristic properties. A typical examples of this type of combining procedures are the well known linear opinion pool [3] and logarithmic opinion pool [4]. The linear opinion pool, which constructs the combined distribution as a convex combination of the given distributions, satisfies, e.g., the strong setwise function property, the zero preservation property, and the marginalization property. For details see, e.g., [5] or [1]. The logarithmic opinion pool, which constructs the combined distribution as a normalized geometric average of the given distributions, satisfies the property of external Bayesianity. For detail see [4].

E-mail address: kracik@utia.cas.cz

0888-613X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2011.01.002

In [6] both the linear and logarithmic opinion pools are considered as a solution of decision making problem with scoring functions based on the Kullback–Leibler divergence.

The Bayesian procedures for combining probability distributions, see, e.g., [7–11], follow the approach introduced in [12]. The core idea here is that the experts' probability distributions are to be taken as data and processed by a decision maker in a standard Bayesian way [13]. The combined distribution is then represented by a posterior probability distribution. A key element of all Bayesian methods is the decision maker's likelihood function for experts' opinions. However, in spite of its significance, a choice of a suitable likelihood is addressed rather shallowly in the literature.

In practice it is quite common that individual experts are able to provide information related only to some aspects of the considered problem. In such cases the experts' distributions can be specified only partially. In [14] a procedure is proposed which allows to combine probability assessments given across different partitionings of a sample space. The sample space is assumed to be discrete and the combined distribution is evaluated as a posterior probability in a form of an extended Dirichlet distribution. In [15] information sources in a form of incoherent partial conditional probability assessments are considered. The aim is to find a coherent conditional probability assessment. For this a discrepancy measure between partial conditional probability assessments and a joint probability distribution is introduced. A coherent probability assessment is then derived from the joint distribution minimizing the proposed discrepancy measure.

1.2. Short problem description

The approach to the combining probability distributions used in this paper differs from those commonly considered in the literature: The problem of combining probability distributions is taken as an analogy of statistical estimation. Namely, for expert information pieces in a form of marginal probability distributions we derive a combining procedure which can be seen as an analogy of maximum likelihood estimation. In such case the combined procedure is searched as a distribution minimizing a weighed sum of Kullback–Leibler divergences (A.1) of the given marginal distributions and corresponding marginals of the searched distribution. The aim of the paper is to analyze the proposed minimization problem and provide an algorithm for its solution.

1.3. Structure of the paper

The paper is structured as follows: In Section 2 the statistical view on combining probability distributions, which stands behind the proposed minimization task, is discussed. The notation and formal definition of the problem is in Section 3. In Section 4 the theoretical results, including necessary and sufficient conditions for a distribution to be a minimizer of the weighted sum of the Kullback–Leibler divergences, are presented. In Section 5 an iterative algorithm for approximate computing of the minimizer is proposed. The appendix contains definitions and basic properties of the Kullback–Leibler divergence and the cross-entropy.

2. Statistical view on combining probability distributions

The following, rather informal, exposition should clarify the ideas that form our view on the problem of combining probability distributions. It is vital for understanding of the meaning of the distributions resulting from the proposed combining procedure as well as for identifying the assumptions under which the method is applicable.

In general, it seems that in the existing methods for combining probability distributions the statistical essence of the addressed problem is neglected. Namely, an expert opinion, forecast, or any other kind of probabilistically described information is always an outcome of some more or less explicitly specified inference procedure. The experts' distributions can be then seen as some kind of statistics and the combined distribution can be naturally found as an estimate based on these statistics. The combining procedures thus can be designed as analogues of common estimation methods.

This approach provides a new insight into mechanisms of the combining methods and also allows the combining methods to be naturally generalized to the case in which the expert distributions are specified only partially, e.g., as marginal distributions. There is also another reason which supports the statistical approach to the combining probability distributions: In the literature a lot of attention is paid to various aspects of combining probability distributions but it seems that an integrated approach is missing. The presented statistical approach could fill the gap.

Assume, temporarily, that each expert distribution is a statistical estimate in a common sense, i.e., it is a probability distribution from a statistical model assigned by an estimator to observed data. The complete data observed by the experts are, however, not available to the decision maker. Instead, the experts provide the estimated distributions, which can be taken as values of the statistics represented by the expert's estimators. Here it is important to realize that the statistical models and estimators used by the individual experts can be mutually different due to different prior knowledge of the experts (domain knowledge, physical models, etc.) and different external factors (limited computational resources, required properties of the estimators, etc.). For the same reasons the statistical model and estimator used in the combining procedure can differ from those used by the individual experts. The estimators of individual experts thus need not be sufficient statistics for the statistical model used in the combining procedure. In this sense the combining task can be seen as an analogy of statistical estimation with incomplete observations.

The special case discussed above is not only instructive but is also of some practical importance. Nevertheless, an expert distribution need not be necessarily an outcome of a precisely specified statistical inference procedure; Often it is rather a result of an intuitive assessment. In such case it could be expected that the process through which the expert's distribution is selected is to some extent analogous to a common estimation task: From a statistical model the expert selects the distribution which fits his knowledge best. The statistical model is again selected with respect to expert's prior knowledge and external factors. It is natural to require that a procedure which allows to combine experts' distributions of this kind should be a generalization of a combining procedure based on statistical estimation. In other words, if an expert distribution can be taken as a result of a statistical estimation procedure, then the mere fact that the distribution is an estimate or not should not affect the result of the combining procedure. Such generalizations can based on a fact that certain estimation procedures can be taken as approximation problems.

2.1. Statistical estimation as an approximation problem

A common formulation of statistical estimation is based on an assumption that a sample of observed data is generated by an unknown "true" distribution which is known to be within a certain family of distribution – a statistical model. However, this assumption is mostly unrealistic in practice. First of all, the probability is just a mathematical model that allows us to treat uncertainty caused by incomplete knowledge and the concept of the true distribution is only ancillary. Moreover, even if we accept the assumption that the true distribution exists, it is practically impossible for it to be, e.g., within an a priori selected statistical model which forms a finite-dimensional subspace of an infinite-dimensional space of all probability distributions of the considered random variable.

A more natural view on the statistical estimation is that the estimate is a distribution from a statistical model which is in some sense closest to the observed behavior of the considered phenomenon, i.e., to the observed data. Some of the common estimation methods clearly fit into this concept. For example, the logarithmic likelihood function is equal to a negative cross-entropy (A.7) of an empirical distribution from the observed data and the statistical model multiplied by the number of observed data; see Appendix A.2. The maximum likelihood estimation is then equivalent to the minimization of the cross-entropy between the empirical distribution and the statistical model. Nevertheless, this approximation problem is well defined for any probability distribution in place of the empirical one whereas the mere fact that the distribution to be approximated is an empirical one or not plays no role. Note that despite its usefulness the connection between statistical estimation and approximation is rather rarely employed in the literature; For examples of the usage, see, e.g., [16,17].

2.2. Combining probability distributions as an approximation problem

Taking the statistical estimation as an approximation problem allows us to derive the combining procedures from common estimation methods. In this paper a combining procedure derived from the maximum likelihood estimation is considered. Namely, for a multivariate random variable each expert is supposed to provide partial information in a form of a marginal distribution. These marginal distributions are supposed to be selected from statistical models consisting of all probability distributions of corresponding random variables. In this sense these distribution can be taken as analogies of empirical distributions. Similarly, the combined distribution is searched within a class of all probability distributions. Through the special case in which the experts' information pieces can be equivalently expressed as sequences of partially observed data we get the following condition for the combined distribution: The combined distribution is a minimizer of the weighted sum of negative cross-entropies of the experts' distributions and the corresponding marginals of the searched one. The analogy with an estimation task indicates that the meaning of the weights is similar to numbers of observed data. Furthermore, it is obvious that the information pieces must be "independent". Again, the meaning of the independence could be specified only through the analogy with the estimation task, in which the data are supposed to be partially observed realizations of independent and identically distributed random variables.

For completely specified experts' distributions, i.e., not marginal ones, it can be easily proved that the combined distribution is a weighted sum of the experts' distributions with weights equal to those in the minimization task. In other words, the resulting combining procedure is the well known linear opinion pool [3]. Nevertheless, for expert information pieces given as marginal distributions the approximation problem becomes significantly more difficult. Its algorithmic solution supported by theoretical results is the main contribution of this paper. In what follows the approximation problem is formulated in a slightly different form: The Kullback–Leibler divergence (A.1) is used instead of the cross-entropy. From the relation (A.11) between the Kullback–Leibler divergence and the cross-entropy it is clear that the modified approximation problem has the same solution as the original one except the case in which some of the expert distributions have infinite differential entropy. The reason for using the Kullback–Leibler divergence is that it is well established measure of proximity of probability distributions. Furthermore, the Kullback–Leibler divergence plays a crucial role in the field of information geometry [18], which is intended to be used for further extensions of the proposed combining procedure.

Remark 2.1. The prior knowledge of the individual experts could be naturally exploited to assemble prior knowledge for the combining task, which is then used to select a suitable statistical model for the combining procedure. Nevertheless, the expert distributions do not mediate this kind of knowledge. Firstly, a single distribution provides only little evidence about the complete statistical model. Secondly, the choices of experts' statistical models can be affected by the external conditions.

In summary, exploiting the prior knowledge of individual experts is a problem different from the discussed combining probability distributions and should be treated separately.

Remark 2.2. The weights in the approximation task are supposed to be provided by the experts. They can be derived from number of data observed by the experts in case that the experts' distributions are based on real observations. In the opposite case, e.g., the device of imaginary results [19] can be employed to elicit the weights.

Remark 2.3. It may seem that the proposed statistical view on combining probability distributions makes the problem extremely complex. In fact, it just reflects the real complexity of the problem. Any reasonable combining procedure must necessarily take into account various factors such as dependence of the expert information pieces, their relevance, or various constraints under which the expert distributions are assessed. The statistical approach allows us to concretize these factors at least through the analogy with statistical estimation. Moreover, it is clear that the problem of combining probability distributions into a single one easily becomes an ill conditioned problem, if no supporting information is available.

3. Notation and problem formulation

The following general notational conventions are used throughout the text:

- $\int \cdots \mathrm{d}x$ without explicitly specified integration domain is to be taken as a definite integral over the range of x. Analogously, $\sum_x$ denotes the sum over the range of x.
- $\operatorname{argmin}_{x\in M} f(x)$ denotes a set of $x \in M$ for which $f(x)$ attains its minimum on M.
- A shortcut *pdf* stands for a probability density function.

3.1. Problem formulation

Let, for some $n \in \mathbb{N}, X = (X_1, \ldots, X_n)$ be a multivariate random variable with values in $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n \subset \mathbb{R}^n$. Let $P \in \mathbb{N}$ and for $p = 1, \ldots, P$ let ${}^p\mathcal{I} \subset \{1, \ldots, n\}$ be nonempty sets. For $p = 1, \ldots, P$ we define random variables ${}^pX = (X_i)_{i\in {}^p\mathcal{I}}$ and $\bar{}^pX = (X_i)_{i\in\{1,\ldots,n\}\setminus {}^p\mathcal{I}}$. Values of X, ${}^pX$, and $\bar{}^pX$ are denoted by small letters x, ${}^px$, and $\bar{}^px$, respectively. Throughout the text the random variables and their values are not explicitly distinguished and are commonly denoted by small letters x, ${}^px$, and $\bar{}^px$. The ranges of x, ${}^px$, and $\bar{}^px$ are denoted by $\mathcal{X}$, ${}^p\mathcal{X}$, and $\bar{}^p\mathcal{X}$, respectively. The set of all pdfs of x is denoted by $\mathcal{P}$. For $f(x) \in \mathcal{P}$, $f({}^px)$ refer to marginal pdfs of $f(x)$; $f(\bar{}^px|{}^px)$ denotes a conditional pdf of $\bar{}^px$ given ${}^px$, whereas for ${}^p\mathcal{I} = \{1, \ldots, n\}$ it is, by convention, identically 1. For $f(x) \in \mathcal{P}$ a short notation f without the argument (x) is occasionally used, if no confusion arises.

Assume that we are given pdfs ${}^pf({}^px)$ and weights ${}^p\alpha > 0, p = 1, \ldots, P$. Without loss of generality, it is supposed that $\bigcup_{p\in\{1,\ldots,P\}} {}^p\mathcal{I} = \{1, 2, \ldots, n\}$ and $\sum_p {}^p\alpha = 1$. For fixed pdfs ${}^pf({}^px)$ and weights ${}^p\alpha$ we define a function $\mathcal{D}(f)$ acting on the set $\mathcal{P}$,

$$\mathcal{D}(f) = \sum_p {}^p\alpha \mathsf{D}\left({}^pf({}^px)||f({}^px)\right), \tag{3.1}$$

where $\mathsf{D}\left(f(x)||g(x)\right)$ denotes the Kullback–Leibler divergence of pdfs $f(x)$ and $g(x)$ defined by (A.1). The problem to be considered is then formulated as follows: *Find a joint pdf* $f(x) \in \mathcal{P}$ *so that*

$$f(x) \in \operatorname*{argmin}_{f\in\mathcal{P}} \mathcal{D}(f). \tag{3.2}$$

Remark 3.1. From the basic properties of the Kullback–Leibler divergence, see Appendix A.1, it immediately follows that if the given pdfs ${}^pf({}^px)$ are marginals of a common joint pdf, then (for arbitrary positive weights ${}^p\alpha$) pdf $f(x)$ minimizing $\mathcal{D}$ fulfills $f({}^px) = {}^pf({}^px)$, for all $p \in \{1, \ldots, P\}$.

Remark 3.2. The problem formulation as well as the theoretical results in the next section are proposed for continuous random variables. For x being a discrete random variable with values in $\mathbb{N}^n$, both the appropriate problem formulation and the results can be acquired by taking the densities with respect to the counting measure and substituting the integrals with sums. This approach is adopted in Section 5.

4. Theoretical results

The minimization problem (3.2) can be solved easily using the basic properties of the Kullback–Leibler divergence, see Appendix A.1, e.g., if ${}^p\mathcal{I} = \{1, 2, \ldots, n\}$ for all $p \in \{1, \ldots, P\}$, or if $P = 2$. However, in a general case the solution of (3.2) is not so straightforward.

First, let us consider a set

$$\mathcal{F} = \{f(x) \in \mathcal{P} | \mathcal{D}(f) < +\infty\}\,. \tag{4.1}$$

The set $\mathcal{F}$ has the following properties:

- $\mathcal{F}$ is nonempty. To prove it, consider an arbitrary $h(x) \in \mathcal{P}$ such that $h(x) > 0$ on $\mathcal{X}$. For a pdf

$$f(x) = \sum_p {}^p\alpha\, {}^pf({}^px)\, h(\bar{}^px|\,{}^px)$$

it holds

$$\begin{aligned}\mathcal{D}(f) &= \sum_p {}^p\alpha \int {}^pf({}^px) \ln \frac{{}^pf({}^px)}{\int \sum_{r=1}^P {}^r\alpha\, {}^rf({}^rx)\, h(\bar{}^rx|\,{}^rx)\mathrm{d}\,\bar{}^px}\mathrm{d}\,{}^px \\ &\le \sum_p {}^p\alpha \int {}^pf({}^px) \ln \frac{{}^pf({}^px)}{\int {}^p\alpha\, {}^pf({}^px)\, h(\bar{}^px|\,{}^px)\mathrm{d}\,\bar{}^px}\mathrm{d}\,{}^px = -\sum_p {}^p\alpha \ln {}^p\alpha,\end{aligned}$$

and thus $f(x) \in \mathcal{F}$.
- $\mathcal{F}$ contains $\operatorname{argmin}_{f\in\mathcal{P}} \mathcal{D}(f)$. This property follows directly from the definition (4.1) of the set $\mathcal{F}$ and the fact that $\mathcal{F} \neq \emptyset$. For this reason, it is sufficient to search for

$$f(x) \in \operatorname*{argmin}_{f\in\mathcal{F}} \mathcal{D}(f)$$

instead of (3.2).
- $\mathcal{F}$ is a convex set. The convexity follows from the convexity of the Kullback–Leibler divergence (A.2).

A crucial role for derivation of properties of $f(x)$ minimizing the function $\mathcal{D}$ has the operator $A : \mathcal{F} \to \mathcal{P}$ defined, for fixed ${}^pf({}^px)$ and ${}^p\alpha$, $p = 1, \ldots, P$, by

$$Af = \sum_{p=1}^P {}^p\alpha f(\bar{}^px|{}^px)\, {}^pf({}^px)\,. \tag{4.2}$$

Note that the operator A is well defined in the sense that if for some p it holds $f({}^px)\,|_{{}^px={}^p\tilde{x}} = 0$ for some ${}^p\tilde{x} \in {}^p\mathcal{X}$, then it holds ${}^pf({}^px)\,|_{{}^px={}^p\tilde{x}} = 0$, because $f(x) \in \mathcal{F}$. The ambiguity in $f(\bar{}^px|{}^px)|_{{}^px={}^p\tilde{x}}$ is then irrelevant.

A key property of the operator A is given by the following proposition.

Proposition 4.1. *For all $f(x) \in \mathcal{F}$ it holds*

$$\mathcal{D}(f) - \mathcal{D}(Af) \ge \mathsf{D}\,(Af||f)\,.$$

Proof. The proof is straightforward and is based on definitions of the cross-entropy (A.7) and the Kullback–Leibler divergence (A.1), properties of the cross-entropy (A.10), (A.8), and on definition (4.2) of the operator A.

From the definition (3.1) of the function $\mathcal{D}$ we get by multiplying the pdfs of ${}^px$ with $f(\bar{}^px|{}^px)$

$$\mathcal{D}(f) - \mathcal{D}(Af) = \sum_{p=1}^P {}^p\alpha \int {}^pf({}^px)\, f(\bar{}^px|{}^px) \ln \frac{(Af)({}^px) f(\bar{}^px|{}^px)}{f({}^px)\, f(\bar{}^px|{}^px)}\mathrm{d}x, \tag{4.3}$$

which can be rewritten using the cross-entropy (A.7) as

$$\mathcal{D}(f) - \mathcal{D}(Af) = \sum_{p=1}^P {}^p\alpha \left[\mathsf{K}\left({}^pf({}^px)\, f(\bar{}^px|{}^px), f(x)\right) - \mathsf{K}\left({}^pf({}^px)\, f(\bar{}^px|{}^px),\, (Af)({}^px) f(\bar{}^px|{}^px)\right)\right]. \tag{4.4}$$

From (A.10) and (A.8) it follows that

$$\mathsf{K}\left({}^pf({}^px)\, f(\bar{}^px|{}^px),\, (Af)({}^px) f(\bar{}^px|{}^px)\right) \le \mathsf{K}\left({}^pf({}^px)\, f(\bar{}^px|{}^px),\, (Af)(x)\right). \tag{4.5}$$

Inserting (4.5) into (4.4) and using (A.7) and (4.2) we get

$$\mathcal{D}(f) - \mathcal{D}(Af) \ge \sum_{p=1}^P {}^p\alpha \int {}^pf({}^px)\, f(\bar{}^px|{}^px) \ln \frac{(Af)(x)}{f(x)}\mathrm{d}x = \mathsf{D}\,(Af||f)\,. \qquad \square$$

Remind that the function $\mathcal{D}$ is defined using the Kullback–Leibler divergence. On that account, the convention $0 \ln 0 = 0$, adopted in its definition (A.1), is employed in the above expressions. Then, e.g., the equality (4.3) holds also in case that $f({}^{\tilde{p}}x|{}^{p}x) = 0$ for some $x \in \mathcal{X}$.

Corollary 4.1. *For all $f(x) \in \mathcal{F}$ it holds $Af \in \mathcal{F}$.*

A direct consequence of Proposition 4.1 gives a necessary condition for $f(x)$ to be a minimizer of the function $\mathcal{D}$.

Proposition 4.2. *If $f(x) \in \operatorname{argmin}_{f\in\mathcal{F}} \mathcal{D}(f)$, then it holds*

$$Af = f. \tag{4.6}$$

Proof. Suppose that $f(x) \in \operatorname{argmin}_{f\in\mathcal{F}} \mathcal{D}(f)$ and $Af \neq f$. Then it holds $\mathsf{D}(Af||f) > 0$ and from Proposition 4.1 it follows that $\mathcal{D}(Af) < \mathcal{D}(f)$, which is in a contradiction with $f(x) \in \operatorname{argmin}_{f\in\mathcal{F}} \mathcal{D}(f)$. □

The opposite implication to Proposition 4.2 does not hold in general. However, under an additional assumption, the equality $Af = f$ provides also a sufficient condition for $f(x)$ to be a minimizer of the function $\mathcal{D}$.

Proposition 4.3. *Let for $f(x) \in \mathcal{F}$ it holds $f(x) > 0$ on $\mathcal{X}$ and $Af = f$. Then $f(x)$ satisfies $f(x) \in \operatorname{argmin}_{f\in\mathcal{F}} D(f)$.*

Proof. Assume that $f(x) \in \mathcal{F}$ satisfies $f(x) > 0$ on $\mathcal{X}$ and $Af = f$. For $h(x) \in \mathcal{P}$, let us define a function $q_{f,h} : [0, 1] \to \mathbb{R}$,

$$q_{f,h}(\omega) = \mathcal{D}((1-\omega)f + \omega h) = \sum_p \int {}^{p}f({}^{p}x) \ln \frac{{}^{p}f({}^{p}x)}{(1-\omega)f({}^{p}x) + \omega h({}^{p}x)} \mathrm{d}\,{}^{p}x.$$

First, we prove that $q_{f,h}(\omega)$ has a derivative on a (right) neighbourhood of 0. For all $p \in \{1, \ldots, P\}$ it holds

$$\left|\frac{\partial}{\partial\omega}\left({}^{p}f({}^{p}x) \ln \frac{{}^{p}f({}^{p}x)}{(1-\omega)f({}^{p}x) + \omega h({}^{p}x)}\right)\right| \tag{4.7}$$

$$= \left|{}^{p}f({}^{p}x) \frac{h({}^{p}x) - f({}^{p}x)}{(1-\omega)f({}^{p}x) + \omega h({}^{p}x)}\right| \leq {}^{p}f({}^{p}x) \frac{h({}^{p}x) + f({}^{p}x)}{(1-\omega)f({}^{p}x)} \leq \frac{h({}^{p}x) + f({}^{p}x)}{{}^{p}\alpha(1-\omega)},$$

where the last inequality follows from the fact that $Af = f$, which implies $f({}^{p}x) \geq {}^{p}\alpha\, {}^{p}f({}^{p}x)$. Thus, for all $p \in \{1, \ldots, P\}$, the expression (4.7) has an integrable upper bound independent of ω on $[0, \omega_0]$, for some $\omega_0 > 0$, which ensures that the derivative of $q_{f,h}(\omega)$ exists on some right neighbourhood of 0. For the derivative of $q_{f,h}(\omega)$ at $\omega = 0$ we get

$$\left.\frac{\partial q_{f,h}(\omega)}{\partial\omega}\right|_{\omega=0} = \sum_p {}^{p}\alpha \int {}^{p}f({}^{p}x) \frac{f({}^{p}x) - h({}^{p}x)}{f({}^{p}x)} \mathrm{d}\,{}^{p}x = 1 - \int \frac{Af(x)}{f(x)} h(x) \mathrm{d}x. \tag{4.8}$$

Now, assume that $\mathcal{D}(\tilde{f}) < \mathcal{D}(f)$ for some $\tilde{f}(x) \in \mathcal{F}$. Then, because $\mathcal{D}$ is a convex function on $\mathcal{F}$, it holds

$$\left.\frac{\partial q_{f,\tilde{f}}(\omega)}{\partial\omega}\right|_{\omega=0} = \lim_{\varepsilon\to 0+} \frac{\mathcal{D}((1-\varepsilon)f + \varepsilon\tilde{f}) - \mathcal{D}(f)}{\varepsilon} \leq \mathcal{D}(\tilde{f}) - \mathcal{D}(f) < 0. \tag{4.9}$$

Simultaneously, according to (4.8) it holds

$$\left.\frac{\partial q_{f,\tilde{f}}(\omega)}{\partial\omega}\right|_{\omega=1} = 0,$$

which is in a contradiction with (4.9). □

Remark 4.1. Without the assumption that $f(x) > 0$ on $\mathcal{X}$ the implication in Proposition 4.3 need not hold, as it is illustrated by the following example. On the other hand, this assumption is not necessary even for ${}^{p}f({}^{p}x)$ being positive for all $p \in \{1, \ldots, P\}$.

Example 4.1. Let $x = (x_1, x_2)$, $\mathcal{X}_1 = \mathcal{X}_2 = \{0, 1\}$, $P = 2$, ${}^1x = x_1$, ${}^2x = x_2$, and

$${}^1f(x_1) = \begin{cases} \kappa & \text{for } x_1 = 0, \\ 1-\kappa & \text{for } x_1 = 1, \end{cases} \quad {}^2f(x_2) = \begin{cases} \lambda & \text{for } x_2 = 0, \\ 1-\lambda & \text{for } x_2 = 1, \end{cases}$$

for some $\kappa, \lambda \in (0, 1)$.

For any pdf $f(x_1, x_2) \in \mathcal{P}$ it holds that $f(x_1, x_2) \in \operatorname{argmin}_{f\in\mathcal{F}} \mathcal{D}(f)$ iff its marginals $f(x_1)$ and $f(x_2)$ are equal to pdfs ${}^1f(x_1)$ and ${}^2f(x_2)$, respectively.

Now, consider a pdf $\tilde{f}(x_1, x_2)$ defined by

$$\tilde{f}(x_1, x_2) = \begin{cases} {}^1\alpha\kappa + {}^2\alpha\lambda & \text{for } x_1 = 0, x_2 = 0, \\ {}^1\alpha(1-\kappa) + {}^2\alpha(1-\lambda) & \text{for } x_1 = 1, x_2 = 1, \\ 0 & \text{otherwise.} \end{cases}$$

The pdf $\tilde{f}(x)$ satisfies $A\tilde{f} = \tilde{f}$ for any $\kappa, \lambda \in (0, 1)$, but $\tilde{f}(x) \in \operatorname{argmin}_{f\in\mathcal{F}} \mathcal{D}(f)$ only if $\kappa = \lambda$.

Remark 4.2. Using the definition of the operator A the necessary condition (4.6) for $f(x)$ to be a minimizer of the function D has a form

$$f(x) = \sum_{p=1}^{P} {}^p\alpha f({}^{\bar{p}}x|{}^px)\, {}^pf({}^px)\,.$$

Obviously, this relation is the defining equation of the linear opinion pool in which the parts of pdfs which are not specified by individual experts are substituted by the corresponding parts of the resulting pdf. This fact illustrates that the combining procedure based on (3.2) can be seen as a natural extension of the linear opinion pool for incompletely specified expert information pieces. Contrary to the case of completely specified information pieces, here the combined pdf is to be searched as a solution of an implicit equation.

5. Algorithmic solution for discrete random variables

As an analytical solution of Eq. (4.6) is not known (except few trivial cases), Proposition 4.2 itself cannot be used to find potential minimizers of the function $\mathcal{D}$. However, under some additional assumptions, an approximation of $f(x) \in \operatorname{argmin}_{f\in\mathcal{P}} \mathcal{D}(f)$ can be found using an iterative algorithm based on the propositions stated in Section 4. A core of the algorithm consist in repetitive application of the operator A defined by (4.2). Namely, for an arbitrary pdf $\varphi_0(x) \in \mathcal{F}$, we consider a sequence of pdfs $(\varphi_k(x))_{k=0}^{+\infty}$ defined by a recursive relation

$$\varphi_{k+1} = A\varphi_k. \tag{5.1}$$

Proposition 4.1 ensures that $(\mathcal{D}(\varphi_k))_{k=0}^{+\infty}$ is a non-increasing sequence. Particularly, if it is guaranteed that $\varphi_k(x) > 0$ on $\mathcal{X}$, then it holds, according to Proposition 4.3, that

$$\begin{aligned} &\mathcal{D}(\varphi_{k+1}) < \mathcal{D}(\varphi_k) \quad \text{if } \varphi_k \neq \varphi_{k+1}, \\ &\varphi_k(x) \in \operatorname*{argmin}_{f\in\mathcal{P}} \mathcal{D}(f) \quad \text{if } \varphi_k = \varphi_{k+1}. \end{aligned}$$

However, to this point nothing guarantees that $\mathcal{D}(\varphi_k) - \mathcal{D}(\varphi_{k+1})$ being arbitrarily small, yet positive, for some positive $\varphi_k(x)$, implies that $\mathcal{D}(\varphi_k)$ is close to the minimum. In other words, still it is not assured that $\lim_{k\to\infty} \mathcal{D}(\varphi_k) = \min_{f\in\mathcal{P}} \mathcal{D}(f)$, even if it is provided that $\varphi_k(x) > 0$ on $\mathcal{X}$. For discrete random variables the convergence and some other issues are discussed in the following paragraphs.

Suppose that $\mathcal{X}_1, \ldots, \mathcal{X}_n$ are finite sets. In this case, the convergence of $\mathcal{D}(\varphi_k)$ to $\min_{f\in\mathcal{P}} \mathcal{D}(f)$ can be proved, e.g., if for some $\varepsilon > 0$ it holds $\varphi_k(x) > \varepsilon$ on $\mathcal{X}$, for all $k \in \mathbb{N}$. This property of $\varphi_k(x)$ is guaranteed, for example, if for some $p \in \{1, \ldots, P\}$ it holds ${}^px = x$ and ${}^pf(x) > 0$ on $\mathcal{X}$. The convergence is given by the following proposition.

Proposition 5.1. *Suppose that the sequence $(\varphi_k(x))_{k=0}^{+\infty}$ of pdfs defined by (5.1), for some $\varphi_0(x) \in \mathcal{F}$, has the property*

$$\exists \varepsilon > 0,\ \forall k \in \mathbb{N}, \quad \varphi_k(x) \geq \varepsilon \text{ on } \mathcal{X}.$$

Then it holds

$$\lim_{k\to\infty} \mathcal{D}(\varphi_k) = \min_{f\in\mathcal{P}} \mathcal{D}(f). \tag{5.2}$$

Proof. Suppose that (5.2) does not hold. Then, because from Proposition 4.1 it follows that $(\mathcal{D}(\varphi_k))_{k=0}^{+\infty}$ is non-increasing, it holds

$$\exists c > 0,\ \forall f \in \operatorname*{argmin}_{f\in\mathcal{P}} \mathcal{D}(f),\ \forall k \in \mathbb{N},\quad \mathcal{D}(\varphi_k) - \mathcal{D}(f) \geq c. \tag{5.3}$$

As stated in the proof of Proposition 4.3, it holds

$$\frac{\mathrm{d}}{\mathrm{d}\omega}\mathcal{D}((1-\omega)\varphi_k + \omega f)\Big|_{\omega=0} = 1 - \sum_x \left(\sum_p {}^p\alpha \frac{{}^pf({}^px)}{\varphi_k({}^px)}\right) f(x)\,. \tag{5.4}$$

From definition (5.1) of $\varphi_{k+1}(x)$, definition (4.2) of the operator A, and from relation (5.4) it follows that

$$\sum_x \frac{\varphi_{k+1}(x)}{\varphi_k(x)} f(x) = 1 - \frac{\mathrm{d}}{\mathrm{d}\omega}\mathcal{D}((1-\omega)\varphi_k + \omega f)\Big|_{\omega=0}.$$

Due to convexity of $\mathcal{D}(\cdot)$, it holds

$$\frac{\mathrm{d}}{\mathrm{d}\omega}\mathcal{D}((1-\omega)\varphi_k + \omega f)\Big|_{\omega=0} \leq \mathcal{D}(f) - \mathcal{D}(\varphi_k), \tag{5.5}$$

which, together with (5.3), implies that, for all $f(x) \in \operatorname{argmin}_{f\in\mathcal{P}} \mathcal{D}(f)$,

$$\sum_x \frac{\varphi_{k+1}(x)}{\varphi_k(x)} f(x) \geq 1 + c. \tag{5.6}$$

From (5.6) it then follows that for some $\tilde{x}_k \in \mathcal{X}$ it must hold

$$\frac{\varphi_{k+1}(\tilde{x}_k)}{\varphi_k(\tilde{x}_k)} \geq 1 + c. \tag{5.7}$$

Using Lemma A.1 we get a lower estimate

$$\mathsf{D}\,(\varphi_{k+1}||\varphi_k) \geq \varphi_{k+1}(\tilde{x}_k) \ln \frac{\varphi_{k+1}(\tilde{x}_k)}{\varphi_k(\tilde{x}_k)} + (1 - \varphi_{k+1}(\tilde{x}_k)) \ln \frac{1-\varphi_{k+1}(\tilde{x}_k)}{1-\varphi_k(\tilde{x}_k)}. \tag{5.8}$$

Lemma A.2 applied to (5.8) together with inequality (5.7) then implies that, for all $k \in \mathbb{N}$, it holds

$$\mathsf{D}\,(\varphi_{k+1}||\varphi_k) \geq \varepsilon(1+c) \ln \frac{\varepsilon(1+c)}{\varepsilon} + (1 - \varepsilon(1+c)) \ln \frac{1-\varepsilon(1+c)}{1-\varepsilon}, \tag{5.9}$$

which is positive, as it represents a Kullback–Leibler divergence of two non-equal pdfs of a binary random variable. Because, according to Proposition 4.1,

$$\mathcal{D}(\varphi_k) - \mathcal{D}(\varphi_{k+1}) \geq \mathsf{D}\,(\varphi_{k+1}||\varphi_k)\,,$$

it follows from (5.9) that $\lim_{k\to+\infty} \mathcal{D}(\varphi_k) = -\infty$, which is in a contradiction with the non-negativity of the function $\mathcal{D}$. □

5.1. Stopping rule

Proposition 5.1 says that, under the given assumptions, for an arbitrary initial approximation $\varphi_0(x) \in \mathcal{F}$, an arbitrarily good approximation (in the sense of the value of $\mathcal{D}$) can be acquired by repetitive application of the operator A; However, to this point we are not able to evaluate the quality of the approximation. For this purpose, a lower estimate of $\min_{f\in\mathcal{P}} \mathcal{D}(f)$ based on (5.5) can be used. Namely, from (5.5), (5.4), and the definition (4.2) of the operator A it follows that for positive $\varphi_k(x)$ it holds

$$\mathcal{D}(f) \geq \mathcal{D}(\varphi_k) + 1 - \sum_x \frac{A\varphi_k(x)}{\varphi_k(x)} f(x)\,, \tag{5.10}$$

for all $f(x) \in \mathcal{P}$. The lower estimate of $\min_{f\in\mathcal{P}} \mathcal{D}(f)$ is then acquired by substituting $\sum_x \frac{A\varphi_k(x)}{\varphi_k(x)} f(x)$ in (5.10) by its upper estimate independent of the unknown $f(x)$. For $\mathcal{X}$ being finite, the simplest estimate is

$$\sum_x \frac{A\varphi_k(x)}{\varphi_k(x)} f(x) \leq \max_{x\in\mathcal{X}} \frac{A\varphi_k(x)}{\varphi_k(x)}, \tag{5.11}$$

which gives a lower bound for $\min_{f\in\mathcal{P}} \mathcal{D}(f)$:

$$\min_{f\in\mathcal{P}} \mathcal{D}(f) \geq \mathcal{D}(\varphi_k) + 1 - \max_{x\in\mathcal{X}} \frac{A\varphi_k(x)}{\varphi_k(x)}. \tag{5.12}$$

For (5.11) to be a suitable estimate for a stopping rule, it is necessary to show that the right-hand side of (5.12) converges to $\min_{f\in\mathcal{P}} \mathcal{D}(f)$. Under the assumptions of Proposition 5.1, the convergence is guaranteed by the following proposition.

Proposition 5.2. *Suppose that the sequence $(\varphi_k(x))_{k=0}^{+\infty}$ of pdfs defined by (5.1), for some $\varphi_0(x) \in \mathcal{F}$, has the property*

$$\exists \varepsilon > 0,\ \forall k \in \mathbb{N}, \quad \varphi_k(x) \geq \varepsilon \text{ on } \mathcal{X}.$$

Then, it holds

$$\max_{x\in\mathcal{X}} \frac{A\varphi_k(x)}{\varphi_k(x)} \to 1. \tag{5.13}$$

Proof. Suppose that (5.13) does not hold. Then, because $\sum_x \varphi_k(x) = \sum_x A\varphi_k(x) = 1$, there exist a strictly increasing sequence $(k_j)_{j=1}^{\infty}$, $k_j \in \mathbb{N}$, and $c > 0$ so that, for all $j \in \mathbb{N}$,

$$\frac{A\varphi_{k_j}(\tilde{x}_j)}{\varphi_{x_j}(\tilde{x}_j)} \geq 1 + c,$$

for some $\tilde{x}_j \in \mathcal{X}$. The rest of the proof is an analogy of the proof of Proposition 5.1. □

A stopping rule for the recursive evaluation of approximations $\varphi_k(x)$ based on the estimate (5.11) has a form

$$\text{stop if } \max_{x\in\mathcal{X}} \frac{A\varphi_k(x)}{\varphi_k(x)} - 1 \leq \zeta, \tag{5.14}$$

where $\zeta > 0$ is a predefined threshold specifying a precision of the resulting approximation. If the condition (5.14) is fulfilled for some $\varphi_k(x)$, then, according to (5.12), it holds that $\mathcal{D}(\varphi_k) - \min_{f\in\mathcal{P}} \mathcal{D}(f) \leq \zeta$. Proposition 5.2 guarantees that, under the given assumptions, the stopping condition (5.14) is fulfilled within a finite number of iterations.

The estimate (5.11) is too rough for (5.14) to be an efficient stopping rule. A more efficient, but computationally more expensive, stopping rule can be obtained from (5.10) by employing a more accurate estimate of $\sum_x \frac{A\varphi_k(x)}{\varphi_k(x)} f(x)$. For example, using (4.6) and the definition (4.2) of the operator A, we get for $f(x) \in \operatorname{argmin}_{f\in\mathcal{P}} \mathcal{D}(f)$

$$\begin{aligned}\sum_x \frac{A\varphi_k(x)}{\varphi_k(x)} f(x) &= \sum_{p=1}^{P} {}^p\alpha \sum_{{}^p x\in {}^p\mathcal{X}} {}^p f({}^p x) \sum_{\bar{p}x\in\bar{p}\mathcal{X}} \frac{A\varphi_k({}^p x,\ \bar{p}x)}{\varphi_k({}^p x,\ \bar{p}x)} f(\bar{p}x|{}^p x) \\ &\leq \sum_{p=1}^{P} {}^p\alpha \sum_{{}^p\mathcal{X}\in{}^p\mathcal{X}} {}^p f({}^p x) \left(\max_{\bar{p}x\in\bar{p}\mathcal{X}} \frac{A\varphi_k({}^p x,\ \bar{p}x)}{\varphi_k({}^p x,\ \bar{p}x)} \right),\end{aligned} \tag{5.15}$$

where

$$\max_{\bar{p}x\in\bar{p}\mathcal{X}} \frac{A\varphi_k({}^p x,\ \bar{p}x)}{\varphi_k({}^p x,\ \bar{p}x)} = \frac{A\varphi_k({}^p x)}{\varphi_k({}^p x)}$$

by convention in case that ${}^p\mathcal{I} = \{1, \ldots, n\}$. As it holds that

$$\sum_{p=1}^{P} {}^p\alpha \sum_{{}^p x\in{}^p\mathcal{X}} {}^p f({}^p x) \left(\max_{\bar{p}x\in\bar{p}\mathcal{X}} \frac{A\varphi_k({}^p x,\ \bar{p}x)}{\varphi_k({}^p x,\ \bar{p}x)} \right) \leq \max_{x\in\mathcal{X}} \frac{A\varphi_k(x)}{\varphi_k(x)},$$

the inequality (5.15) provides a more accurate upper estimate of the sum $\sum_x \frac{A\varphi_k(x)}{\varphi_k(x)} f(x)$ then (5.11).

Remark 5.1. In general, the set $\operatorname{argmin}_{f\in\mathcal{P}} \mathcal{D}(f)$ is a convex set of more then one element. Though we have proven in Proposition 5.1 that, under appropriate assumptions, the sequence $(\mathcal{D}(\varphi_k))_{k=0}^{+\infty}$ converges to the minimum, it does not directly follow that the sequence $(\varphi_k(x))_{k=0}^{+\infty}$ converges. We propose a working hypothesis that the sequence $(\varphi_k(x))_{k=0}^{+\infty}$ converges, at least under appropriate assumptions. The convergence as well as the dependence of the limit pdf on the initial approximation $\varphi_0(x)$ is a subject of further study.

Remark 5.2. According to Proposition 5.1, the condition $\varphi_k(x) \geq \varepsilon$ for some $\varepsilon > 0$ is sufficient for the sequence $(\mathcal{D}(\varphi_k))_{k=0}^{+\infty}$ to converge to the minimum. Nevertheless, this condition is obviously not necessary. The iterative algorithm can be applied even if $\varphi_k(x) \geq \varepsilon$ cannot be guaranteed. Whenever the stopping condition in (5.14) is fulfilled for some $\varphi_k(x)$, the specified precision of the approximation is acquired. However, without the assumption that $\varphi_k(x) \geq \varepsilon$, for all $k \in \mathbb{N}$, it is not guaranteed that the stopping condition is fulfilled within a finite number of iterations.

Remark 5.3. Although the results stated in Section 4 are formulated for the continuous case, the algorithmic solution proposed in Section 5 is restricted to discrete random variables. A reason is that in the continuous case it is significantly more difficult to find reasonable sufficient conditions for the convergence of $(\mathcal{D}(\varphi_k))_{k=0}^{+\infty}$ to the minimum. Another reason is that the operator A defined in (4.2) employs both the conditioning and mixing operations which causes that the approximations $\varphi_k(x)$ do not possess a finite-dimensional parameterization (common to all $\varphi_k(x)$). In [20] a modification of the iterative algorithm is proposed which, for x being a continuous random variable, searches for an approximation of the minimizing pdf within a class of Gaussian mixtures (convex combinations of Gaussian pdfs).

6. Summary and conclusions

In this paper a problem of combining marginal probability distributions as a means for aggregating pieces of expert information is studied. For this purpose a novel approach, which takes the combining problem as an analogy of statistical estimation, is proposed and discussed; see Section 2. The combined distribution is then searched as a minimizer of a weighted sum of the Kullback–Leibler divergences from the given marginal distributions to the corresponding marginals of the searched one (relations (3.1) and (3.2)), which can be taken as an analogy of a maximum likelihood estimate from information represented by the experts' distributions.

The results achieved in the paper are following:

- A necessary condition for a distribution to be a solution of the proposed minimization task is stated (Proposition 4.2). It is also proved that under an additional assumption this condition is also a sufficient one (Proposition 4.3). These results cover both discrete and continuous case.
- For discrete random quantities an iterative algorithm for an approximate solution of the minimization task is presented (relation (5.1)) and its convergence is proved (Proposition 5.1). A stopping rule, which guarantees that a required precision of the approximation is achieved, is also derived (relation (5.14)).

The open problems are related especially to convergence issues of the iterative algorithm. Particularly, more detailed investigation of sufficient conditions for convergence is needed. Other problems are sketched in Remark 5.1.

In a long term, the combining procedure could be generalized to cover dependent expert information. The analogy of the discussed problem and statistical estimation seems to provide a good starting point for this direction. Nevertheless, identification of a form and extent of information dependence inevitably requires adequate additional information, which can be hardly available in practice. On that account, we expect that a rigorous treatment of dependent information pieces will lead to a shift from a purely probabilistic approach towards imprecise probabilities [21].

Acknowledgement

This work was supported by GAČR project 102/08/0567.

A. Discrepancy of pdfs

A.1. Kullback–Leibler divergence

The Kullback–Leibler divergence [22] is a member of a class of so called f-divergences [23,24] which are used to quantify discrepancy between pairs of probability distributions. For a pair of pdfs $f(x)$ and $g(x)$ of probability distributions F and G, respectively, of a random variable x, the Kullback–Leibler divergence is defined as

$$\mathsf{D}\left(f(x)||g(x)\right) = \begin{cases} \int f(x) \ln \frac{f(x)}{g(x)} \mathrm{d}x & \text{for } F \ll G, \\ +\infty & \text{otherwise}, \end{cases} \tag{A.1}$$

where $F \ll G$ denotes absolute continuity of F with respect to G, and the integrand is defined using the conventions $0 \ln 0 = 0$, $0 \ln \frac{0}{0} = 0$. In this paper the following basic properties of the Kullback–Leibler divergence are used:

- Non-negativity: For all pdfs $f(x)$, $g(x)$ it holds $\mathsf{D}\left(f(x)||g(x)\right) \geq 0$, where the equality holds iff $f(x) = g(x)$.
- Convexity in both arguments: For all pdfs $f(x)$, $g(x)$, $h(x)$ and arbitrary $\alpha \in [0, 1]$ it holds

$$\mathrm{D}\left(\alpha f(x)+(1-\alpha)h(x)||g(x)\right) \le \alpha \mathrm{D}\left(f(x)||g(x)\right)+(1-\alpha)\mathrm{D}\left(h(x)||g(x)\right)$$
$$\mathrm{D}\left(f(x)||\alpha g(x)+(1-\alpha)h(x)\right) \le \alpha \mathrm{D}\left(f(x)||g(x)\right)+(1-\alpha)\mathrm{D}\left(f(x)||h(x)\right) \tag{A.2}$$

Lemma A.1. *For arbitrary pdfs $f(x)$, $g(x)$ and a set $M \subset \mathcal{X}$, let $a = \int_M f(x)\mathrm{d}x$, $b = \int_M g(x)\mathrm{d}x$. Then*

$$\mathrm{D}\left(f(x)||g(x)\right) \ge a \ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}, \tag{A.3}$$

using the conventions $0\ln 0 = 0$, $0\ln\frac{0}{0} = 0$, $\ln\frac{c}{0} = +\infty$ *for* $c > 0$.

Proof. Suppose that $a, b \in (0,1)$. Then

$$\begin{aligned}
\mathrm{D}\left(f(x)||g(x)\right) &= a\int_M \frac{f(x)}{a}\left(\ln\frac{\frac{f(x)}{a}}{\frac{g(x)}{b}} + \ln\frac{a}{b}\right)\mathrm{d}x + (1-a)\int_{M^C}\frac{f(x)}{1-a}\left(\ln\frac{\frac{f(x)}{1-a}}{\frac{g(x)}{1-b}} + \ln\frac{1-a}{1-b}\right)\mathrm{d}x \\
&= a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b} + \mathrm{D}\left(\frac{1}{a}f(x)I_M(x)\middle\|\frac{1}{b}g(x)I_M(x)\right) \\
&\quad + \mathrm{D}\left(\frac{1}{1-a}f(x)I_{M^C}(x)\middle\|\frac{1}{1-b}g(x)I_{M^C}(x)\right) \\
&\ge a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b},
\end{aligned}$$

where $I_M(x)$ denotes the indicator function of the set M,

$$I_M(x) = \begin{cases} 1 & \text{if } x \in M, \\ 0 & \text{if } x \notin M. \end{cases}$$

Verification of (A.3) for $a \in \{0,1\}$ or $b \in \{0,1\}$ is trivial. □

Note that a proposition analogous to Lemma A.1 can be stated for any finite partition of $\mathcal{X}$; for more details see [25].

Lemma A.2. *Let $s, t \in (0,1)$ satisfy $\frac{s}{t} \ge C$ and $t \ge \varepsilon$, for some $C > 1$ and $\varepsilon > 0$. Then it holds*

$$s\ln\frac{s}{t} + (1-s)\ln\frac{1-s}{1-t} \ge C\varepsilon\ln C + (1-C\varepsilon)\ln\frac{1-C\varepsilon}{1-\varepsilon}.$$

Proof. Let us consider a function $u(a,b) = a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}$ for $a, b \in (0,1)$. As for $a > b > 0$, $\frac{\partial}{\partial a}u(a,b) = \ln\frac{a(1-b)}{b(1-a)} > 0$, it holds

$$u(s,t) \ge u(Ct,t). \tag{A.4}$$

Now, define a function

$$v(b) = u(Cb,b) = Cb\ln C + (1-Cb)\ln\frac{1-Cb}{1-b},$$

for $b \in (0,1)$. We prove that its derivative

$$\frac{\mathrm{d}}{\mathrm{d}b}v(b) = C\ln\frac{C-Cb}{1-Cb} + \frac{1-C}{1-b}$$

is positive for $b > 0$: For $b = 0$, it holds

$$\left(C\ln\frac{C-Cb}{1-Cb} + \frac{1-C}{1-b}\right)\Big|_{b=0} = C\ln C + 1 - C > 0, \tag{A.5}$$

because $(C\ln C + 1 - C)|_{C=1} = 0$ and $\frac{\mathrm{d}}{\mathrm{d}C}(C\ln C + 1 - C) = \ln C > 0$ for $C > 1$. For the second derivative of $v(b)$ it holds

$$\frac{\mathrm{d}^2}{\mathrm{d}b^2}v(b) = \frac{(C-1)^2}{(1-b)^2(1-Cb)} > 0 \tag{A.6}$$

for $b < \frac{1}{C}$. From (A.5) and (A.6) it follows that $\frac{\mathrm{d}}{\mathrm{d}b}v(b) > 0$ for $b \in (0, \frac{1}{C})$ and thus $v(t) \ge v(\varepsilon)$, which, together with (A.4), proves the lemma. □

A.2. Cross-entropy

The cross-entropy is tightly related to the Kullback–Leibler divergence, though it does not belong among f-divergences. For a pair of pdfs $f(x), g(x)$ of probability distributions F and G, the cross-entropy is usually defined as

$$\mathsf{K}(f(x), g(x)) = \int f(x) \ln \frac{1}{g(x)} \mathrm{d}x, \tag{A.7}$$

where the integral is defined using the convention $0 \ln \frac{c}{0} = 0$. However, the definition (A.7) can be extended also to distributions F having a discrete component. In this case the corresponding "pdf" (the distribution F does not have a density in a rigorous sense) can be formally expressed as $f(x) = \alpha f_c(x) + (1-\alpha) f_d(x)$, where $f_c(x)$ is a pdf of the absolutely continuous component of F, $f_d(x)$ formally represents a pdf of the discrete component of F, and $\alpha \in [0, 1]$. $f_d(x)$ can be written as a weighted sum of the Dirac delta functions

$$f_d(x) = \sum_{k=1}^{K} \gamma_k \delta(x_k - x),$$

for some non-negative $\gamma_1, \ldots, \gamma_K$ satisfying $\sum_{k=1}^{K} \gamma_k = 1$ and $x_1, \ldots, x_k \in \mathcal{X}$. The defining relation (A.7) then can be used also in this more general case. The extension is justified by the fact that for a sequence of pdfs $f_i(x)$, $i = 1, \ldots$ of absolutely continuous distributions F_i weakly converging to F it holds $\mathsf{K}(f_i(x), g(x)) \to \mathsf{K}(f(x), g(x))$.

Note that for $f(x)$ being an empirical pdf from data $x_1, \ldots, x_T$, i.e.,

$$f(x) = \frac{1}{T} \sum_{t=1}^{T} \delta(x - x_i),$$

and a parametric model $g(x|\theta)$, $\theta \in \Theta$, the cross-entropy $\mathsf{K}(f(x), g(x|\theta))$ satisfies

$$\mathsf{K}(f(x), g(x|\theta)) = -\frac{1}{T} \prod_{t=1}^{T} g(x_i|\theta).$$

In words, it is proportional to the negative log-likelihood from the data $x_1, \ldots, x_T$.

Elementary properties of the cross-entropy are:

- Convexity in the second argument: For all pdfs $f(x), g(x), h(x)$ and $\alpha \in [0, 1]$ it holds
$$\mathsf{K}(f(x), \alpha g(x) + (1-\alpha) h(x)) \le \alpha \mathsf{K}(f(x), g(x)) + (1-\alpha) \mathsf{K}(f(x), h(x)).$$
- For all pdfs $f(x), g(x)$, it holds
$$\mathsf{K}(f(x), f(x)) \le \mathsf{K}(f(x), g(x)) \tag{A.8}$$
with equality iff $f(x) = g(x)$.

For a joint pdf $f(x, y)$ and a conditional pdf $g(y|x)$ of random variables x, y we define a conditional cross-entropy

$$\mathsf{K}(f(x, y), g(y|x)) = \int f(x) \mathsf{K}(f(y|x), g(y|x)) \, \mathrm{d}x, \tag{A.9}$$

where, for fixed x, $\mathsf{K}(f(y|x), g(y|x))$ is taken as the non-conditional cross-entropy defined by (A.7). Using (A.9), we get for a pair of joint pdfs $f(x, y)$ and $g(x, y)$

$$\mathsf{K}(f(x, y), g(x, y)) = \mathsf{K}(f(x), g(x)) + \mathsf{K}(f(x, y), g(y|x)). \tag{A.10}$$

The cross-entropy and the Kullback–Leibler divergences are related through the differential entropy $\mathsf{H}(f(x)) = \mathsf{K}(f(x), f(x))$ by the equality

$$\mathsf{K}(f(x), g(x)) = \mathsf{D}(f(x) || g(x)) + \mathsf{H}(f(x)), \tag{A.11}$$

if both sides exist.

References

[1] C. Genest, J.V. Zidek, Combining probability distributions: a critique and an annotated bibliography, Statistical Science 1 (1986) 114–148.
[2] R.T. Clemen, R.L. Winkler, Combining probability distributions from experts in risk analysis, Risk Analysis 19 (1999) 187–203.
[3] M. Stone, The opinion pool, The Annals of Mathematical Statistics 32 (1961) 1339–1342.
[4] C. Genest, A characterization theorem for externally bayesian groups, The Annals of Statistics 12 (1984) 1100–1105.
[5] K.J. Mcconway, Marginalization and linear opinion pools, Journal of the American Statistical Association 76 (1981) 410–414.
[6] A.E. Abbas, A Kullback–Leibler view of linear and log-linear pools, Decision Analysis 6 (2009) 25–37.

[7] R.F. Bordley, R.W. Wolff, On the aggregation of individual probability estimates, Management Science 27 (1981) 959–964.
[8] R.T. Clemen, Combining overlapping information, Management Science 33 (1987) 373–380.
[9] R.T. Clemen, R.L. Winkler, Aggregating point estimates: a flexible modeling approach, Management Science 39 (1993) 501–515.
[10] C. Genest, M.J. Schervish, Modeling expert judgments for bayesian updating, Annals of Statistics 13 (1985) 1198–1212.
[11] D. Lindley, Reconciliation of probability distributions, Operations Research 31 (1983) 866–880.
[12] P.A. Morris, Decision analysis expert use, Management Science 20 (1974) 1233–1241.
[13] J. Bernardo, A. Smith, Bayesian Theory, second ed., John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore, 1997.
[14] R.F. Bordley, Combining the opinions of experts who partition events differently, Decision Analysis 6 (2009) 38–46.
[15] A. Capotorti, G. Regoli, F. Vattari, Correction of incoherent conditional probability assessments, International Journal of Approximate Reasoning 51 (2010) 718–727.
[16] R. Kulhavý, Recursive Nonlinear Estimation: A Geometric Approach, Lecture Notes in Control and Information Sciences, vol. 216, Springer-Verlag, London, 1996.
[17] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (1974) 716–723.
[18] S.-I. Amari, H. Nagaoka, Methods of Information Geometry, American Mathematical Society, Providence, 2007.
[19] I. Good, Probability and the Weighing of Evidence, C. Griffin, London, 1950.
[20] J. Kracík, Cooperation Methods in Bayesian Decision Making with Multiple Participants, Ph.D. Thesis, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, 2009.
[21] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman & Hall, London, New York, 1991.
[22] S. Kullback, R.A. Leibler, On information and sufficiency, The Annals of Mathematical Statistics 22 (1951) 79–86.
[23] S.M. Ali, S.D. Silvey, A general class of coefficients of divergence of one distribution from another, Journal of the Royal Statistical Society. Series B (Methodological) 28 (1966) 131–142.
[24] I. Csiszar, Information-type measures of difference of probability distributions and indirect observations, Studia Scientiarum Mathematicarum Hungarica 2 (1967) 299–318.
[25] F. Liese, I. Vajda, On divergences and informations in statistics and information theory, IEEE Transactions on Information Theory 52 (2006) 4394–4412.