JMLR: Workshop and Conference Proceedings vol 23 (2012) 28.1–28.20 25th Annual Conference on Learning Theory

Divergences and Risks for Multiclass Experiments

Darío García-García DARIO.GARCIA@ANU.EDU.AU
Australian National University and NICTA, Canberra ACT 0200, Australia

Robert C. Williamson BOB.WILLIAMSON@ANU.EDU.AU
Australian National University and NICTA, Canberra ACT 0200, Australia

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

Csiszár's f-divergence is a way to measure the similarity of two probability distributions. We study the extension of f-divergence to more than two distributions to measure their joint similarity. By exploiting classical results from the comparison of experiments literature we prove the resulting divergence satisfies all the same properties as the traditional binary one. Considering the multidistribution case actually makes the proofs simpler. The key to these results is a formal bridge between these multidistribution f-divergences and Bayes risks for multiclass classification problems.

Keywords: f-divergence, Bayes risk, comparison of experiments, multiclass losses, affinity, information distance, similarity.

1. Introduction

Machine learning variously analyses single objects, compares two objects, or sometimes compares multiple objects simultaneously. Depending whether one wishes to work with finite objects directly or to assume a probabilistic framework, one can describe how complex a single object is either in terms of its Kolmogorov complexity (Kolmogorov, 1965) or Shannon entropy (Shannon, 1948), where one represents objects using probability distributions. Two objects can be compared in terms of the information distance (Bennett et al., 1998) (the maximum relative Kolmogorov complexity — the length of the shortest program to describe a second object given the first, or the first given the second) or relative entropy, also known as Kullback-Leibler information or KL divergence (Kullback and Leibler, 1951). Comparing multiple (≥ 2) objects is less well studied. The multi-object generalisation of information distance (Long et al., 2008) has been formally studied only recently (Li, 2011; Vitányi, 2011, 2012). Others such generalisations are mentioned later in this introduction.

The f-divergence (Csiszár, 1967, 1963) $\mathbb{I}_f(P, Q)$ (defined formally later) is a generalisation of KL-divergence which measures the "distance" between two probability distributions P and Q. It is parametrized by a convex function $f\colon [0, \infty) \to \mathbb{R}$. It is sometimes argued that f-divergences other than KL-divergence are rarely used and have limited value. However by considering P and Q as the class conditional distributions of a binary experiment, $\mathbb{I}_f(P, Q)$ can be expressed in terms of the Bayes risk of the experiment with respect to a loss function ℓ that depends upon f (Gutenbrunner, 1990); see also (Österreicher and Vajda, 1993). With hindsight one can see many of these results in (Torgersen, 1981). This "bridge" between the two perspectives allows insight and simplification of many results (Reid and Williamson, 2011), and extends the results of (Liese and Vajda, 2006); see also (Liese and Vajda, 2008). Crucially, this bridge means that if one accepts the utility-theoretic foundations of statistical decision theory (Wald, 1950; Berger, 1985; French and Insua, 2000), which

form the basis of statistical learning theory, and one is thus willing to countenance a wide variety of loss functions as abstractions of various end-use problems, then one logically must accept the need and use of different f-divergences. This is hardly surprising and no different to the need to work with a wide variety of metrics in the theory of function spaces (Kolmogorov and Fomin, 1970) or the importance of the choice of similarity measure in clustering (Jain et al., 1999, Section 4).

The theory of binary losses (Reid and Williamson, 2010) has been extended to multiclass losses (Vernet et al., 2011). The extension actually simplifies and elucidates the binary results. This raises the question (which we answer in this paper): *what is the analogous extension of* $\mathbb{I}_f(P, Q)$ *to "multiclass" f-divergences* $\mathbb{I}_f(P_1, \ldots, P_k)$*?*

There are several possible extensions of f-divergences to multiple distributions, so it is not obvious at first whether one can talk of "the" extension. The extension which we study (which is not new) is the "natural" extension because it has the same properties as the traditional binary f-divergence.

The multidistribution f-divergence $\mathbb{I}_f(P_1, \ldots, P_k)$ (formally defined in §3) is known as the f-dissimilarity (Györfi and Nemetz, 1975, 1978). This was originally presented as a generalisation of Matusita's affinity (Matusita, 1971, 1967). Generalisations of particular divergences to several distributions are the *information radius* (Sibson, 1969) $R(P_1, \ldots, P_k) = \frac{1}{k} \sum_{i=1}^{k} \mathrm{KL}(P_i, \frac{P_1 + P_2 + \cdots + P_k}{k})$ where $\mathrm{KL}(P, Q)$ is the Kullback-Leibler divergence and the *average divergence* (Sgarro, 1981) $K(P_1, \ldots, P_k) = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j=1}^{k} \mathrm{KL}(P_i, P_j)$. Some other approaches to generalising f-divergences to more than two distributions are summarised by Basseville (2010).

The f-divergence $\mathbb{I}_f(P_1, \ldots, P_k)$ has been used in hypothesis testing (Menéndez et al., 2005; Zografos, 1998). Györfi and Nemetz (1975) bounded the minimal probability of error in terms of the f-affinity; see also (Glick, 1973; Toussaint, 1978). These results are analogous to surrogate regret bounds because there is in fact an *exact* relationship between $\mathbb{I}_f$ and the Bayes risk of an associated multiclass classification problem. Multidistribution f-divergences have also been used to extend rate-distortion theory (primarily as a technical means to get better bounds) (Zakai and Ziv, 1975) and to unify information theory with the second law of thermodynamics (Merhav, 2011). The *estimation* of $\mathbb{I}_f(P_1, \ldots, P_k)$ has been studied by Morales et al. (1998). As we will see, the connection to Bayes risk suggests alternate estimation schemes.

We utilise the theory of comparison of experiments, originally due to Blackwell (1951, 1953); see (Goel and Ginebra, 2003) for a gentle introduction, or (Torgersen, 1991b; Shiryaev and Spokoiny, 2000) for more complete treatments. We summarise the results we need in §2. The theory has been applied to multiclass decision problems by Torgersen (1970), but without drawing the connections that we do to multidistribution f-divergences.

The rest of the paper is organised as follows. We formally introduce multidistribution f-divergences (§3), relate divergences to Bayes risks (§4), prove the multidistribution f-divergence satisfies the same properties as the binary divergence $\mathbb{I}_f(P, Q)$ (§5), present some examples (§6) and conclusions (§7). Proofs missing from the main text can be found in Appendix A, while Appendix B develops examples of multidistribution divergences.

The *key point* of the paper is that by viewing f-divergences as a transformation of Bayes risks we can develop clear insight into what seems to be a complex notion (a divergence between multiple distributions). By exploiting this bridge all of the proofs in the paper are simple (especially when compared to those in the literature). We see this as a virtue! Analogously to what was observed by Vitányi (2011, Section I.B) "the new notation and proofs for the general case are simpler than the ... existing proofs for the particular case of pairwise" divergences.

Standard notation we use is as follows. Vectors are typeset bold: $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)'$ where the $'$ denotes transpose. $\mathbf{1} := (1, \dots, 1)'$, the simplex $\Delta^k := \{\boldsymbol{\pi} \in \mathbb{R}^k_+ \colon \boldsymbol{\pi}'\mathbf{1} = 1\}$, $[k] := \{1, \dots, k\}$, $\mathbf{P}_{[k]} := (P_1, \dots, P_k)'$ and $\mathbf{e}_k$ is a vector $\mathbf{e}$ such that $e_k = 1$, $e_i = 0$ if $i \neq k$. If P is a distribution on Θ, the *support* of P, $\operatorname{supp} P := \{\theta \in \Theta \colon P(\theta) > 0\}$. For distributions P and Q we write $P \ll Q$ if P is absolutely continuous with respect to Q, i.e. $P(A) = 0$ for every set A such that $Q(A) = 0$.

2. Statistical Experiments

We introduce statistical experiments and their comparison (Shiryaev and Spokoiny, 2000). A *statistical experiment* $E = (\Omega, \mathcal{F}, \{P_\theta \colon \theta \in \Theta\})$ is given by a set of probability spaces $(\Omega, \mathcal{F}, P_\theta)$, $\theta \in \Theta$, with the family of measures $\mathcal{P} = \{P_\theta \colon \theta \in \Theta\}$ being parameterized by a parameter θ.

It is natural to think of the elements $\theta \in \Theta$ as "theories" or hypotheses that, due to the experiment, manifest themselves through probability distributions on some observable space. For our purposes, we will assume that Θ is a finite set $\Theta = \{1, \dots, k\}$.

An experiment is said to be *totally non-informative* if the distributions $P_1 = \cdots = P_k$. This means that nothing can be learned about θ by observing samples from the experiment. On the other hand, an experiment is *totally informative* if $\operatorname{supp}(P_{\theta_i}) \neq \operatorname{supp}(P_{\theta_j})$ which we write $P_i \perp P_j$ $\forall i \neq j$. This directly implies that the hypothesis "generating" a sample can be known with certainty. In order to compare two experiments it is merely necessary that the set of hypothesis Θ is the same for both. Most notably this implies that it is perfectly possible to compare experiments with different sample spaces.

A typical machine learning task comprises an experiment plus a *loss function* which takes into account the terminal consequences of the statistical decisions made based on the data observed from it: $\ell \colon \Theta \times D \to \mathbb{R}$, where D is the decision space. Some experiments E are always better (has lower risk) than others $\tilde{E}$, regardless of the loss function. We say that in that case experiment E is *sufficient* for $\tilde{E}$. A formal definition of risk is given in §3.

Sufficiency is a very stringent condition and for most pairs of experiments no such relation will hold. It thus defines a *partial order* between experiments.

Let $P^{\boldsymbol{\pi}} := \sum_{i=1}^k \pi_i P_i$, $\boldsymbol{\pi} \in \Delta^k$, be a probability measure which dominates all the distributions in the experiment (for example, this is always true if $\pi_i > 0$, $\forall i \in [k]$). It is well known that the vector-valued statistic

$$\mathbf{t}^{\boldsymbol{\pi}}(\omega) = \frac{1}{dP^{\boldsymbol{\pi}}(\omega)} \left(dP_1(\omega), \dots, dP_k(\omega)\right)',$$

taking values in $K^{\boldsymbol{\pi}} := \left\{\mathbf{u} \in \mathbb{R}^k_+ \colon \boldsymbol{\pi}'\mathbf{u} = 1\right\}$ is sufficient for E, so its distribution characterizes the statistical properties of E. Thus statistic is called the *likelihood ratio*. Consider now the *posterior probability* vector for a given prior probability $\boldsymbol{\pi} \in \Delta^k$:

$$\boldsymbol{\eta}^{\boldsymbol{\pi}}(\omega) = \frac{1}{dP^{\boldsymbol{\pi}}(\omega)} \left(\pi_1 dP_1(\omega), \dots, \pi_k dP_k(\omega)\right). \tag{1}$$

It can be interpreted as a normalized version of the likelihood ratio, so it takes values in the standard simplex Δ^k. In Bayesian terms talking about posterior probabilities implies that $P^{\boldsymbol{\pi}}$ has a meaning of prior probability, instead of just a base measure. However, we will not make such a distinction here.

Under this setup, a totally non-informative experiment will yield constant $\mathbf{t}^{\boldsymbol{\pi}}(\omega)$, while totally informative experiments result in $\mathbf{t}^{\boldsymbol{\pi}}(\omega)$ whose elements are all strictly 0 except for the one corresponding to the generating hypothesis. Intuitively, how informative an experiment is should be

related to how much the corresponding $\mathbf{t}^{\boldsymbol{\pi}}(\omega)$ deviates from the constant vector. How should that deviation be measured? An answer is given by the *Blackwell-Sherman-Stein theorem*. Before stating the theorem, a few definitions are in order. Let $\ell : \Theta \times \Delta^k \to \mathbb{R}$ be a loss function and $\boldsymbol{\pi} \in \Delta^k$. The *Bayes risk* of the (multiclass) classification problem with class-conditional distributions $\mathbf{P}_\Theta$ and prior probability $\boldsymbol{\pi}$ for loss ℓ is given by

$$\underline{\mathbb{L}}_\ell(\boldsymbol{\pi}, \mathbf{P}_\Theta) = \min_{T\colon \Omega \to \Delta^k} \sum_{i=1}^{k} \pi_i \mathbb{E}_{\omega \sim P_i}[\ell(i, T(\omega))]$$

Let $(\Omega, \mathcal{F})$ and $(\tilde{\Omega}, \tilde{\mathcal{F}})$ be two measurable spaces. A *Markov kernel* from $(\Omega, \mathcal{F})$ to $(\tilde{\Omega}, \tilde{\mathcal{F}})$ is a function mapping each point $\omega \in \Omega$ to a probability measure M_ω on $(\tilde{\Omega}, \tilde{\mathcal{F}})$ so that for each $B \in \tilde{\mathcal{F}}$ the map $\omega \mapsto M_\omega(B)$ is $\mathcal{F}$-measurable. Intuitively, Markov kernels can be understood as the equivalents of stochastic matrices in the continuous case. In general they introduce *randomization* by transforming a distribution concentrated on ω into a more spread-out distribution.

A Markov kernel M also defines a map between probability measures P on $\mathcal{F}$ and probability measures MP on $\tilde{\mathcal{F}}$ given by

$$MP(A) = \int_\Omega M(\omega, B) P(d\omega), \qquad B \in \tilde{\mathcal{F}}.$$

Theorem 1 (Blackwell-Sherman-Stein) *Suppose that $E = (\Omega, \mathcal{F}, \{P_\theta\}_{\theta\in\Theta})$ and $\tilde{E} = (\tilde{\Omega}, \tilde{\mathcal{F}}, \{\tilde{P}_\theta\}_{\theta\in\Theta})$ are two statistical experiments. Then the following conditions are all equivalent*

(C1) The Bayes risks satisfy $\underline{\mathbb{L}}_\ell(\boldsymbol{\pi}, \mathbf{P}_\Theta) \leq \underline{\mathbb{L}}_\ell(\boldsymbol{\pi}, \tilde{\mathbf{P}}_\Theta)$ for any loss function ℓ and prior probability vector $\boldsymbol{\pi}$.

(C2) There is a Markov kernel M such that $\tilde{P}_\theta = MP_\theta$ for all $\theta \in \Theta$, i.e. $\tilde{P}_\theta$ can be obtained from P_θ via randomization.

(C3) For some strictly positive set of weights $\boldsymbol{\pi}$, $\mathbb{E}_{P^{\boldsymbol{\pi}}}\left[\phi(\mathbf{t}^{\boldsymbol{\pi}}(\omega))\right] \geq \mathbb{E}_{\tilde{P}_{\boldsymbol{\pi}}}\left[\phi(\tilde{\mathbf{t}}^{\boldsymbol{\pi}}(\tilde{\omega}))\right]$ for every *convex $\phi(\cdot)$ defined on $K^{\boldsymbol{\pi}}$. Moreover, if this holds for any strictly positive weights, it also holds for any other set of weights resulting in a dominating measure.*

If any of the above conditions holds, all of them hold and experiment E is said to be sufficient *for $\tilde{E}$.*

This important theorem links the sufficiency ordering of experiments with the *convex* or *Bishop-de-Leeuw ordering* of the corresponding likelihood ratio statistics (Shaked and Shanthikumar, 1994). This convex ordering is once again a partial order. For the sake of comparing arbitrary experiments or random variables, it would be desirable to have mappings into a completely ordered set (such as $\mathbb{R}$) which are consistent with these important orderings. This is the main idea behind f-divergences; the f-divergence of $\mathbf{P}_{[k]}$ is a real number that measures the joint dissimilarity of the k distributions that make up $\mathbf{P}_{[k]}$.

3. f-divergences and their multidistribution extension

Assume that we have two binary experiments (dichotomies) $E = (\Omega, \mathcal{F}, \{P_1, P_2\})$ and $\tilde{E} = (\tilde{\Omega}, \tilde{\mathcal{F}}\{\tilde{P}_1, \tilde{P}_2\})$ such that E is sufficient for $\tilde{E}$. Let $\pi \in (0,1)$, $P^\pi = \pi P_1 + (1-\pi)P_2$, $\tilde{P}^\pi = \pi\tilde{P}_1 + (1-\pi)\tilde{P}_2$ and $\phi : K^{(\pi,1-\pi)'} \to \mathbb{R}$ be any convex function. Then, by Theorem 1,

$$\mathbb{E}_{P^\pi}\left[\phi\left(\frac{dP_1}{dP^\pi}, \frac{dP_2}{dP^\pi}\right)\right] \geq \mathbb{E}_{\tilde{P}^\pi}\left[\phi\left(\frac{d\tilde{P}_1}{d\tilde{P}^\pi}, \frac{d\tilde{P}_2}{d\tilde{P}^\pi}\right)\right].$$

Assume now that $P_1 \ll P_2$ and $\tilde{P}_1 \ll \tilde{P}_2$. Then, we can use P_2 and $\tilde{P}_2$ as dominating measures and write

$$\mathbb{E}_{P_2}\left[\phi\left(\frac{dP_1}{dP_2}, \frac{dP_2}{dP_2}\right)\right] \geq \mathbb{E}_{\tilde{P}_2}\left[\phi\left(\frac{d\tilde{P}_1}{d\tilde{P}_2}, \frac{d\tilde{P}_2}{d\tilde{P}_2}\right)\right]$$

Since ϕ is a convex function, $f(t) := \phi(t, 1)$ is also convex and thus the *binary f-divergence*

$$\mathbb{I}_f(P_1, P_2) := \mathbb{E}_{P_2}\left[f\left(\frac{dP_1}{dP_2}\right)\right] \geq \mathbb{E}_{\tilde{P}_2}\left[f\left(\frac{d\tilde{P}_1}{d\tilde{P}_2}\right)\right] = \mathbb{I}_f(\tilde{P}_1, \tilde{P}_2).$$

It is thus obvious that f-divergences $\mathbb{I}_f$ are consistent with the convex ordering for the likelihood ratio $\frac{dP_1}{dP_2}$ and thus for the sufficiency ordering of experiments. Hence f-divergences provide a proper relaxation of the partial sufficiency ordering into a total order.

This connection suggests a natural way to extend f-divergences to multiple distributions. Consider k probability distributions $P_1, \ldots, P_k$. In order to define a f-divergence between them it is necessary to specify the index j of the distribution to be used as a reference measure and a convex function $f_j \in \mathcal{C}^{k-1}$, where $\mathcal{C}^k := \{\phi : [0,\infty)^k \to \mathbb{R}, \phi \text{ convex}\}$. Alternatively, the characterization can be carried out in terms of a function $\phi \in \mathcal{C}^k$. Let

$$\mathbf{t}^j = \frac{1}{dP_j}\mathbf{P}_{[k]} \text{ and } \tilde{\mathbf{t}}^j = \left(\frac{dP_1}{dP_j}, \ldots, \frac{dP_{j-1}}{dP_j}, \frac{dP_{j+1}}{dP_j}, \ldots, \frac{dP_k}{dP_j}\right)'.$$

With these ingredients we define

$$\mathbb{I}_{\phi,j}(\mathbf{P}_{[k]}) = \mathbb{E}_{P_j}\left[\phi(\mathbf{t}^j)\right] = \mathbb{E}_{P_j}\left[f_j\left(\tilde{\mathbf{t}}^j\right)\right],$$

where $f_j \colon \mathbb{R}^{k-1} \to \mathbb{R}$ is defined to satisfy $f_j(\tilde{\mathbf{t}}^j) = \phi(\mathbf{t}_j)$. The notation in terms of $k-1$ dimensional functions links nicely with the usual definition of f-divergences for the binary case, but makes the multiclass derivations cumbersome, so in the following we will work only with k dimensional functions. Moreover, for notational simplicity we will assume in the following that $j = k$, so we can drop the indices. We can also think of the divergences as functions of two arguments: a vector $\mathbf{P}_{[k-1]}$ of probability distributions and a single distribution P_k, so

$$\boxed{\mathbb{I}_\phi(\mathbf{P}_{[k]}) := \mathbb{E}_{P_k}\left[\phi(\mathbf{t}^k)\right] \equiv \mathbb{E}_{P_k}\left[\phi(\mathbf{t})\right].} \tag{2}$$

The definition of an f-divergence can thus be seen as a two-step process: first we *relativise* the probability distributions by taking Radon-Nikodym derivatives with respect to the reference distribution. After that, the dispersion of the resulting likelihood ratio is measured using the desired convex function.

Since $\mathbf{t} \in K^k := \{\mathbf{x} \in [0,\infty)^k \colon x_k = 1\}$, the behaviour of $\phi(\mathbf{x})$ whenever $x_k \neq 1$ does not affect the divergence. As we show in Theorem 5, Jensen's inequality implies $\mathbb{I}_\phi(\mathbf{P}_{[k]}) \geq \phi(\mathbf{1})$. In the binary case f-divergences are usually defined using functions such that $f(1) = 0$, so the divergence is lower-bounded by 0. We can do the same thing in the general case and thus require $\phi \in \mathcal{C}_1^k$, where $\mathcal{C}_1^k := \{\phi \in \mathcal{C}^k : \phi(\mathbf{1}) = 0\}$. From now on when we write $\mathbb{I}_\phi(\mathbf{P}_{[k]})$ we will presume that $\phi \in \mathcal{C}_1^k$.

4. Relating Multidistribution f-Divergences and Bayes risks

We now relate multidistribution f-divergences to Bayes risks. There is a well-known relationship in the binary case relating posterior probabilities and likelihood ratios (Reid and Williamson, 2011, §4.1). This relationship is the key that bridges f-divergences and Bayes risks written in their typical form involving the posterior probability function. Here we present a general version of this relationship. Let $K_k^j := \{\mathbf{x} \in [0,\infty)^k \colon x_j = 1\}$ and define the mapping $R_{j,\boldsymbol{\pi}} : \Delta^k \to K_k^j$

$$R_{j,\boldsymbol{\pi}}(\boldsymbol{\eta}) := \frac{\pi_j}{\eta_j} \frac{\boldsymbol{\eta}}{\boldsymbol{\pi}} \tag{3}$$

where vector division is interpreted element-wise, and its inverse mapping $R_{j,\boldsymbol{\pi}}^{-1} : K_k^j \to \Delta^k$

$$R_{j,\boldsymbol{\pi}}^{-1}(\mathbf{t}) = \frac{\boldsymbol{\pi} \circ \mathbf{t}^j}{\boldsymbol{\pi}' \mathbf{t}^j}, \tag{4}$$

where $\circ$ denotes the element-wise or Hadamard product and calling this the inverse is justified by the following lemma.

Lemma 2 *If $\boldsymbol{\eta}^{\boldsymbol{\pi}}$ is defined via (1) and $P_1, \ldots, P_k \ll P_j$, then $\boldsymbol{\eta}^{\boldsymbol{\pi}} = R_{j,\boldsymbol{\pi}}^{-1}(\mathbf{t}^j)$ and $\mathbf{t}^j = R_{j,\boldsymbol{\pi}}(\boldsymbol{\eta}^{\boldsymbol{\pi}})$.*

As long as the densities are absolutely continuous with respect to the chosen base measure, that base measure can be changed using the Radon-Nikodym theorem, and the resulting likelihood ratio vector normalized to yield a posterior probability vector. Then, anything that can be written in terms of posterior probabilities (as is usually the case with the Bayes risk) can also be written as a function of likelihood ratios with respect to an arbitrary dominating measure. We can thus think of f-divergences as a reparametrization of Bayes risks when the dominating measure is selected from the set of class conditional distributions.

The Bayes risk for a loss $\ell \colon [k] \times \Delta^k \to \mathbb{R}$ can be rewritten as

$$\begin{aligned}
\underline{\mathbb{L}}_\ell(\pi, \mathbf{P}_{[k]}) &= \min_{T \colon \Omega \to \Delta^k} \sum_{i=1}^k \pi_i \mathbb{E}_{\omega \sim P_i}[\ell(i, T(\omega))] = \min_{T \colon \Omega \to \Delta^k} \int_\Omega \sum_{i=1}^k \pi_i \ell(i, T(\omega)) dP_i(\omega) \\
&= \min_{T \colon \Omega \to \Delta^k} \int_\Omega \sum_{i=1}^k \ell(i, T(\omega)) \frac{\pi_i dP_i}{dP^{\boldsymbol{\pi}}} dP^{\boldsymbol{\pi}} = \min_{T \colon \Omega \to \Delta^k} \mathbb{E}_{\omega \sim P^{\boldsymbol{\pi}}} \left[\mathbb{E}_{y \sim \boldsymbol{\eta}^{\boldsymbol{\pi}}(\omega)} \ell(y, T(\omega)) \right] \\
&= \mathbb{E}_{\omega \sim P^{\boldsymbol{\pi}}} \left[\min_{T(\omega) \in \Delta^k} \mathbb{E}_{y \sim \eta^\pi(\omega)} \ell(y, T(\omega)) \right] = \mathbb{E}_{\omega \sim P^{\boldsymbol{\pi}}} \left[\underline{L}_\ell(\boldsymbol{\eta}^{\boldsymbol{\pi}}(\omega)) \right],
\end{aligned} \tag{5}$$

where $\underline{L}_\ell \colon \Delta^k \to \mathbb{R}$ is a concave[1] function which we call the *point-wise Bayes risk*. The interchange of the min and the expectation is justified by (Rockafellar and Wets, 2004, Theorem 14.10).

There are results relating f-divergences and Bayes risks in the binary classification setting (Reid and Williamson, 2011, Theorem 9). We now generalise these to the multiclass setting. These results relate $\mathbb{I}_f$ to a Bayes risk $\underline{L}$.[2] We make use of the notion of *Statistical Information* due to DeGroot (1962, 1970) and defined as the difference between the prior and posterior Bayes risks:

$$\Delta\underline{\mathbb{L}}_\phi(\boldsymbol{\pi}, \mathbf{P}_{[k]}) := \underline{\mathbb{L}}(\boldsymbol{\pi}) - \underline{\mathbb{L}}(\boldsymbol{\pi}, \mathbf{P}_{[k]}). \tag{6}$$

Note that $\underline{\mathbb{L}}(\boldsymbol{\pi}) = \underline{L}(\boldsymbol{\pi})$.

Theorem 3 *For an arbitrary function $\phi \in \mathcal{C}_1^k$ and prior probability $\boldsymbol{\pi} \in \Delta^k$ define*

$$\boxed{\underline{L}_\phi^{\boldsymbol{\pi}}(\boldsymbol{\eta}^{\boldsymbol{\pi}}) := -\frac{\eta_k^{\boldsymbol{\pi}}}{\pi_k}\phi\left(\frac{\pi_k}{\eta_k^{\boldsymbol{\pi}}}\frac{\boldsymbol{\eta}^{\boldsymbol{\pi}}}{\boldsymbol{\pi}}\right) + \underline{L}^{\boldsymbol{\pi}}(\boldsymbol{\pi}) \quad \forall \mathbf{P}_{[k]}.} \tag{7}$$

Then

$$\boxed{\Delta\underline{\mathbb{L}}_\phi(\boldsymbol{\pi}, \mathbf{P}_{[k]}) = \mathbb{I}_\phi(\mathbf{P}_{[k]}) \qquad \forall \mathbf{P}_{[k]}.} \tag{8}$$

Conversely, for an arbitrary point-wise Bayes risk $\underline{L}(\boldsymbol{\eta})$ and prior probability $\boldsymbol{\pi} \in \Delta^k$, if

$$\boxed{\phi_{\underline{L}}^{\boldsymbol{\pi}}(\mathbf{t}) := \underline{L}(\boldsymbol{\pi}) - \boldsymbol{\pi}'\mathbf{t}\,\underline{L}\left(\frac{\boldsymbol{\pi}\circ\mathbf{t}}{\boldsymbol{\pi}'\mathbf{t}}\right) \quad \forall \mathbf{t} \in \mathbb{R}_+^k} \tag{9}$$

then

$$\boxed{\mathbb{I}_{\phi_{\underline{L}}^{\boldsymbol{\pi}}}(\mathbf{P}_{[k]}) = \Delta\underline{\mathbb{L}}(\boldsymbol{\pi}, \mathbf{P}_{[k]}) \qquad \forall \mathbf{P}_{[k]}.} \tag{10}$$

Proof Manipulating the definition of an f-divergence in (2) we obtain

$$\int \phi(\mathbf{t}) dP_k = -\int -\phi(\mathbf{t})\frac{dP_k}{dP^{\boldsymbol{\pi}}}dP^{\boldsymbol{\pi}} = -\int -\frac{\eta_k^{\boldsymbol{\pi}}}{\pi_k}\phi\left(R_{k,\boldsymbol{\pi}}(\boldsymbol{\eta}^{\boldsymbol{\pi}})\right) dP^{\boldsymbol{\pi}}$$

where we have used $\eta_k^{\boldsymbol{\pi}} = \frac{\pi_k dP_k}{dP^{\boldsymbol{\pi}}}$. Then, by comparison with (5) we can define the point-wise risk $\underline{L}_\phi$ corresponding to the multiclass f-divergence parametrized by ϕ as

$$\underline{L}_\phi^{\boldsymbol{\pi}}(\boldsymbol{\eta}^{\boldsymbol{\pi}}) = -\frac{\eta_k^{\boldsymbol{\pi}}}{\pi_k}\phi\left(\frac{\pi_k}{\eta_k^{\boldsymbol{\pi}}}\frac{\boldsymbol{\eta}^{\boldsymbol{\pi}}}{\boldsymbol{\pi}}\right).$$

(Note that if $\phi \in \mathcal{C}_1^k$, then $\phi\left(R_{k,\boldsymbol{\pi}}(\boldsymbol{\pi})\right) = \phi(\mathbf{1}) = 0$ and $\underline{L}_\phi^{\boldsymbol{\pi}}(\boldsymbol{\pi}) = 0$. This is a rather unusual condition on Bayes risks, and that is why it is more natural to associate f-divergences to statistical informations.) Fix $\underline{L}^{\boldsymbol{\pi}}(\boldsymbol{\pi})$ to any desired value and define $\underline{L}_\phi^{\boldsymbol{\pi}}(\boldsymbol{\eta}^{\boldsymbol{\pi}})$ as in (7). Then the point-wise statistical information is $\Delta\underline{L}_\phi^{\boldsymbol{\pi}}(\boldsymbol{\eta}^{\boldsymbol{\pi}}) = \underline{L}^{\boldsymbol{\pi}}(\boldsymbol{\pi}) - \underline{L}_\phi^{\boldsymbol{\pi}}(\boldsymbol{\eta}^{\boldsymbol{\pi}})$. Obviously, $\Delta\underline{L}_\phi^{\boldsymbol{\pi}}(\boldsymbol{\pi}) = 0$. Taking expectations with respect to $P^{\boldsymbol{\pi}}$ we get the first result in the theorem.

The converse relation can be easily shown. Observe that

$$\int \underline{L}(\boldsymbol{\eta}) dP^{\boldsymbol{\pi}} = \int \underline{L}(\boldsymbol{\eta})\frac{dP^{\boldsymbol{\pi}}}{dP_k}dP_k$$

and $\frac{dP^{\boldsymbol{\pi}}}{dP_k}$ is simply $\frac{dP^{\boldsymbol{\pi}}}{dP_k} = \sum_{i=1}^k \pi_i \frac{dP_i}{dP_k} = \boldsymbol{\pi}'\mathbf{t}$. Applying the conversion between posterior probabilities and likelihood ratios in (3) completes the proof. ■

1. Since it is the pointwise infimum of linear functions
2. Such results can be combined with those that relate multiclass Bayes risks and proper losses (Vernet et al., 2012).

5. Properties of Multidistribution f-Divergences

In this section we present multidistribution analogs of the key properties of binary f-divergences. The connection with the comparison of experiments framework prove valuable here. It allows us to prove one of the most important properties of f-divergences, the information processing property, as a direct consequence of the Blackwell-Sherman-Stein theorem. Moreover, in the proofs and corollaries we emphasise what these properties mean for Bayes risks, taking advantage of the bridge between divergences and statistical informations.

5.1. Information Processing

The information processing property of binary f-divergences is one of their most defining properties (Pardo and Vajda, 1997; Harremoës and Tishby, 2007). It states that

$$\mathbb{I}_f(P, Q) \geq \mathbb{I}_f(\tilde{P}, \tilde{Q}),$$

where $\tilde{P} = MP$ and $\tilde{Q} = MQ$ are obtained from P and Q via some randomization mechanism (Markov kernel) M. The multidistribution version is as follows. We write $M\mathbf{P}_{[k]} = (MP_1, \ldots, MP_k)'$.

Theorem 4 *Let M be any Markov kernel. Then for any $\phi \in \mathcal{C}_1^k$,*

$$\mathbb{I}_\phi(\mathbf{P}_{[k]}) \geq \mathbb{I}_\phi(M\mathbf{P}_{[k]}) \qquad \forall \mathbf{P}_{[k]}.$$

Proof The experiment $E(\Omega, \mathcal{F}, \{P_i \colon i \in \Theta\})$, $\Theta = \{1, \ldots, k\}$ is sufficient for experiment $\tilde{E}(\Omega, \mathcal{F}, \{MP_i \colon i \in \Theta\})$ by Theorem 1 (C2). Then, (C3) of that same theorem yields the desired result since a multidistribution f-divergence is just the expectation of a convex function of the likelihood ratio. ■

5.2. Reflexivity

The property of reflexivity (Csiszár, 1967) for binary f-divergences also holds for multidistribution divergences.

Theorem 5 *For any function $\phi \in \mathcal{C}_1^k$ such that ϕ is strictly convex around $\mathbf{1}$*

$$\mathbb{I}_\phi(\mathbf{P}_{[k]}) = 0 \qquad \textit{if and only if } P_1 = \cdots = P_k.$$

Proof By Jensen's inequality,

$$\mathbb{I}_\phi(\mathbf{P}_{[k]}) = \int \phi\left(\frac{1}{dP_k}\mathbf{P}_{[k]}\right) dP_k \geq \phi\left(\left[\int dP_i\right]_{i=1}^k\right) = \phi(\mathbf{1}). \tag{11}$$

Furthermore, if ϕ is strictly convex around $\mathbf{1}$ then (11) becomes an equality if and only if $\frac{dP_i}{dP_k} = 1$ for all i. ■

Corollary 6 (Lower bound of multidistribution f-divergences) *For any $\phi \in \mathcal{C}^k$,*

$$\mathbb{I}_\phi\left(\mathbf{P}_{[k]}\right) \geq \phi(\mathbf{1}), \quad \forall \mathbf{P}_{[k]}.$$

5.3. Invariance to affine terms

Divergences do not change when an affine function is added to ϕ.

Theorem 7 *Let* $\tilde{\phi}(\mathbf{t}) := \phi(\mathbf{t}) + \mathbf{w}'(\mathbf{1} - \mathbf{t})$. *Then* $\mathbb{I}_\phi(\mathbf{P}_{[k]}) = \mathbb{I}_{\tilde{\phi}}(\mathbf{P}_{[k]}) \quad \forall \mathbf{P}_{[k]}$.

The following corollary comes from applying the transformation between divergences and statistical informations in (7).

Corollary 8 *If* $\underline{\tilde{L}}(\boldsymbol{\eta}) = \underline{L}(\boldsymbol{\eta}) + \mathbf{w}'\boldsymbol{\eta}$, $\mathbf{w} \in \mathbb{R}^k$, *then* $\Delta\underline{\mathbb{L}}(\boldsymbol{\pi}, \mathbf{P}_{[k]}) = \Delta\underline{\tilde{\mathbb{L}}}(\boldsymbol{\pi}, \mathbf{P}_{[k]})\ \forall \mathbf{P}_{[k]}$.

5.4. Uniqueness

If two divergences are equal (for all distributions), the corresponding ϕ are equal up to affine offsets.

Theorem 9 *If* $\mathbb{I}_\phi(\mathbf{P}_{[k]}) = \mathbb{I}_{\tilde{\phi}}(\mathbf{P}_{[k]})$ *for all* $\mathbf{P}_{[k]}$ *then* $\tilde{\phi}(\mathbf{t}) = \phi(\mathbf{t}) + \mathbf{w}'(\mathbf{1} - \mathbf{t})$ *for all* $\mathbf{t} \in K^k$.

5.5. Change of order

A standard property of binary f-divergences is that

$$\mathbb{I}_f(P, Q) = \mathbb{I}_{f^\diamond}(Q, P),$$

where $f^\diamond(t) = tf(\frac{1}{t})$ is known as the *Csiszár dual* of the function f. We now extend this change of order property to the multidistribution case.

The multidistribution analog of the Csiszár dual is $\phi^{\diamond_j}$ and is defined as follows for $j \in [k]$ and $\mathbf{t} \in \mathbb{R}^k$,

$$\phi^{\diamond_j}(\mathbf{t}) := t_j\, \phi\left(\frac{1}{t_j}\mathbf{t}^{j\leftrightarrow k}\right) \tag{12}$$

where $\mathbf{t}^{j\leftrightarrow k} = (t_1^{j\leftrightarrow k}, \ldots, t_k^{j\leftrightarrow k})$ and

$$t_c^{j\leftrightarrow k} = \begin{cases} t_c, & c \neq j, k \\ t_j, & c = k \\ t_k, & c = j. \end{cases}$$

Theorem 10 *For every* $\mathbf{P}_{[k]}$

$$\mathbb{I}_\phi(\mathbf{P}_{[k]}) = \mathbb{I}_{\phi^{\diamond_j}}(\mathbf{P}_{[k]}^{j\leftrightarrow k}).$$

Moreover, the mapping $\phi \to \phi^{\diamond_j}$ *is an involution, since* $(\phi^{\diamond_j})^{\diamond_j} = \phi$.

Corollary 11 *The divergence* $\mathbb{I}_\phi$ *induced by the function* $\tilde{\phi} = \phi + \sum_{i=1}^{k-1} \phi^{\diamond_i}$ *is* symmetric *in the sense that for all* $\mathbf{P}_{[k]}$, *for all* $j \in [k-1]$,

$$\mathbb{I}_{\tilde{\phi}}(\mathbf{P}_{[k]}) = \mathbb{I}_{\tilde{\phi}}(\mathbf{P}_{[k]}^{j\leftrightarrow k}).$$

Corollary 12 *For all* $\phi \in \mathcal{C}_1^k$, *if* $\mathbb{I}_\phi(\mathbf{P}_{[k]}) = \Delta\underline{\mathbb{L}}(\boldsymbol{\pi}, \mathbf{P}_{[k]})$ *then* $\mathbb{I}_{\phi^{\diamond_j}}(\mathbf{P}_{[k]}) = \Delta\underline{\mathbb{L}}(\boldsymbol{\pi}^{j\leftrightarrow k}, \mathbf{P}_{[k]}^{j\leftrightarrow k})$.

Divergences	**Risks**
$\mathbb{I}_\phi(\mathbf{P}_{[k]})$	$\Delta\underline{\mathbb{L}}(\boldsymbol{\pi}, \mathbf{P}_{[k]}) = \underline{\mathbb{L}}(\boldsymbol{\pi}) - \underline{\mathbb{L}}(\boldsymbol{\pi}, \mathbf{P}_{[k]})$
$\tilde{\phi} = \phi + \mathbf{w}'(\mathbf{1} - \mathbf{t})$ $\mathbb{I}_{\tilde{\phi}}(\mathbf{P}_{[k]}) = \mathbb{I}_\phi(\mathbf{P}_{[k]})$	$\underline{\tilde{L}}(\boldsymbol{\eta}) = \underline{L}(\boldsymbol{\eta}) + \left(\frac{\mathbf{w}}{\boldsymbol{\pi}}\right)' \boldsymbol{\eta}$ $\underline{\tilde{\mathbb{L}}}(\boldsymbol{\pi}, \mathbf{P}_{[k]}) = \underline{\mathbb{L}}(\boldsymbol{\pi}, \mathbf{P}_{[k]}) + \mathbf{w}'\mathbf{1}$
$\phi^{\diamond j}$	$\underline{\mathbb{L}}\left(\boldsymbol{\pi}^{j\leftrightarrow k}, \mathbf{P}_{[k]}^{j\leftrightarrow k}\right)$
$\phi(\mathbf{e}_k)$	$\underline{L}(\boldsymbol{\pi}) - \pi_k \underline{L}(\mathbf{e}_k)$
$\phi^{\diamond j}(\mathbf{e}_k)$	$-\pi_j \underline{L}(\mathbf{e}_j)$

Table 1: Summary of some relations between operations on ϕ and $\underline{L}$

5.6. Range of Values

The range of values theorem (Csiszár, 1963; Vajda, 1972) bounds the values of a binary f-divergence in terms of properties of the defining convex function. We generalize this theorem via the bridge between divergences and risks.

Theorem 13 *For any k-vector of probability distributions $\mathbf{P}_{[k]}$ and function $\phi \in \mathcal{C}^k$ such that ϕ is bounded below,*

$$\phi(\mathbf{1}) \;\leq\; \mathbb{I}_\phi\left(\mathbf{P}_{[k]}\right) \;\leq\; \phi(\mathbf{e}_k) + \sum_{j=1}^{k-1} \phi^{\diamond j}(\mathbf{e}_k).$$

6. Examples

In this section we particularize the formulae in the previous sections to obtain some multiclass generalizations of well-known f-divergences. With one exception, we will focus on divergences which are linked to a risk with a well-known multiclass generalization. Then, we will recover the generalized divergence based on the multiclass risk. We present the calculations for Variational divergence below. Other divergences are summarised in Table 2 with derivations relegated to Appendix B.

Variational Divergence Consider the 0-1 loss $\ell_{0-1}(i, \hat{i}) := [\![\hat{i} \neq i]\!]$ which has corresponding point-wise and expected Bayes risks

$$\underline{L}_{0-1}(\boldsymbol{\eta}) = 1 - \max_i \eta_i \quad \text{and} \quad \underline{\mathbb{L}}_{0-1}(\boldsymbol{\eta}, P^{\boldsymbol{\pi}}) = \mathbb{E}_{\omega\sim P^{\boldsymbol{\pi}}}[1 - \max_i \eta_i(\omega)].$$

In the binary case, the 0-1 loss is related to the well-known *variational divergence*, given by $V(P, Q) = \int \left|\frac{dQ}{d\lambda} - \frac{dP}{d\lambda}\right| d\lambda$, where λ is any measure dominating P and Q. This divergence corresponds with $\mathbb{I}_f$ where the convex function $f(t) = |t - 1|$. Then, $\underline{\mathbb{L}}_{0-1}\left(\frac{1}{2}, P, Q\right) = \frac{1}{2} - \frac{1}{4}V(P, Q)$. Strictly speaking, the f function defining the divergence corresponding to the statistical information for the 0-1 loss in a binary experiment (under a uniform prior) is given by $f(t) = \frac{1}{2}\max(0, 1 - t)$. Note that this hinge function is used as a primitive for the integral representation of f-divergences (Reid and Williamson, 2011).

Divergence	**Multidistribution** $\phi(\mathbf{t})$	$\underline{L}(\boldsymbol{\eta})$	**Loss**	**Binary** $f(t)$
Variational	$\frac{k-1}{k} - \frac{1}{k}\left(\sum_{i=1}^{k} t_i - \max_i(t_i)\right)$	$1 - \max_i \eta_i$	0-1	$\frac{1}{2}\max(0, 1-t)$
Triangular	$\frac{1}{2k}\left(k - 1 - \sum_{i=1}^{k} t_i + \frac{\sum_{i=1}^{k} t_i^2}{\sum_{i=1}^{k} t_i}\right)$	$\frac{1}{2}\left(1 - \sum_{i=1}^{k} \eta_i^2\right)$	Square	$\frac{1}{4}\frac{t-1}{t+1}$
Jensen-Shannon	$\ln(k) + \frac{1}{k}\sum_{i=1}^{k} t_i \ln\left(\frac{t_i}{\sum_{j=1}^{k} t_j}\right)$	$\sum_{i=1}^{k} \eta_i \ln \frac{1}{\eta_i}$	Log	$\frac{1}{2}\left[t \ln\left(\frac{t}{t+1}\right) + \ln\left(\frac{4}{t+1}\right)\right]$
Matusita	$1 - \left(\prod_{i=1}^{k} t_i\right)^{\frac{1}{k}}$	$k\left(\prod_{i=1}^{k} \eta_i\right)^{\frac{1}{k}}$	—	$1 - \sqrt{t}$

Table 2: Some multidistribution f-divergences and their corresponding Bayes risks.

We may apply the formulae in the previous sections to obtain a multiclass generalization of this divergence, starting from the multiclass 0-1 loss. The ϕ function defining the f-divergence corresponding to the multiclass 0-1 loss under a uniform prior is given by

$$\phi(\mathbf{t}) = \frac{k-1}{k} - \frac{1}{k}\left(\sum_{i\in[k]} t_i - \max_{i\in[k]}(t_i)\right).$$

When $k = 2$ this reduces to the standard case. Denoting $t_1 = t$ and since $t_2 = \frac{dQ}{dQ} = 1$,

$$\frac{1}{2} - \frac{1}{2}\left(t + 1 - \max(t, 1)\right) = -\frac{1}{2}\left(t - \max(t, 1)\right) = \frac{1}{2}\left(\max(t, 1) - t\right) = \frac{1}{2}\max(0, 1 - t).$$

By analogy with the binary case, we can define the multiclass variational divergence as the one resulting from the convex function

$$\phi_V(\mathbf{t}) = -\frac{4}{k}\left(\sum_{i=1}^{k} t_i - \max_i(t_i)\right).$$

In the binary case, adopting the standard notation for binary divergences, this reduces to

$$f_V(t) = -2(t + 1 - \max(t, 1)) = -2\min(t, 1),$$

which by Th. 7 results in the same divergence as the function $t \mapsto |t - 1|$ since $\min(t, 1) = \frac{t+1-|t-1|}{2}$, so $-2\min(t, 1) = |t - 1| - t - 1$ and both functions differ only by an affine term.

7. Conclusions

We have studied the f-affinity and shown it is indeed a natural generalisation of Csiszár's f-divergence. The justification of "natural" comes from its properties which mimic those of the classical binary f-divergence. We have proved (simply!) these properties via the bridge to Bayes risks

which generalises the existing bridge in the binary case. Viewing the experiment as the fundamental object is a basic tenet of the theory of comparison of experiments — one can view $\mathbb{I}_f(\mathbf{P}_{[k]})$ as a "measure of information" in an experiment; confer (Lindley, 1956; DeGroot, 1962). Thus we see the natural interpretation of $\mathbb{I}_\phi(\mathbf{P}_{[k]})$ is as a measure of *joint similarity* of $P_1, \ldots, P_k$; analogous to the notion of information distance between multiple objects (Li, 2011; Vitányi, 2011, 2012).

Given the bridge, we expect many other results can be transferred. Certainly we expect to be able to extend integral representations (general experiments are combinations of simple ones — confer (Vernet et al., 2011) and (Birnbaum, 1961) and the extension to k classes due to Torgersen (1970)). Integral representations for f-divergences are well known (Österreicher and Feldman, 1981; Feldman and Österreicher, 1989; Liese and Vajda, 2006; Reid and Williamson, 2011), but we are unaware of results for multidistribution divergences. It is also reasonable to expect one could extend the approximate comparison of dichotomies of Torgersen (1991a) (confer Liese and Vajda (2006)), surrogate regret bounds and their relation to Pinsker style inequalities (Reid and Williamson, 2011) and results on the joint range of two multidistribution f-divergences (Harremoës and Vajda, 2010).

Acknowledgments

This work was supported by the Australian Research Council (ARC). NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the ARC through the ICT Centre of Excellence program.

References

Michèlle Basseville. Divergence measures for statistical data processing. Technical Report PI 1961, IRISA, November 2010. URL http://hal.inria.fr/inria-00542337/fr/.

Charles H. Bennett, Péter Gács, Ming Li, Paul M.B. Vitanyi, and Wojciech H. Zurek. Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423, 1998.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 1985.

Allan Birnbaum. On the foundations of statistical inference: Binary experiments. *The Annals of Mathematical Statistics*, 32(2):414–435, June 1961.

David Blackwell. Comparison of Experiments. In Jerzy Neyman, editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages 93–102, Berkeley and Los Angeles, 31 July – 12 August 1951. University of California Press.

David Blackwell. Equivalent Comparisons of Experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

Imre Csiszár. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *A Magyar Tudományos Akadémia Matematikai és Fizikai Tudományok Osztályának Közleményei*, 8:85–108, 1963.

Imre Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:29–318, 1967.

Morris H. DeGroot. Uncertainty, Information, and Sequential Experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.

Morris H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill Book Company, 1970.

Dorian Feldman and Ferdinand Österreicher. A note on f-divergences. *Studia Scientiarum Mathematicarum Hungarica*, 24:191–200, 1989.

Simon French and David Ríos Insua. *Statistical Decision Theory*. Arnold, 2000.

Ned Glick. Separation and probability of correct classification among two or more distributions. *Annals of the Institute of Statistical Mathematics*, 25(1):373–382, 1973.

Prem K. Goel and Josep Ginebra. When is one experiment 'always better than' another? *Journal of the Royal Statistical Society Series D (The Statistician)*, 52(4):515–537, 2003.

Cornelius Gutenbrunner. On applications of the representation of f-divergences as averaged minimal Bayesian risk. In *Transactions of the 11th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pages 449–456, Dordrecht; Boston, 1990. Kluwer Academic Publishers.

László Györfi and Tibor Nemetz. f-dissimilarity: A general class of separation measures of several probability measures. In I. Csiszár and P. Elias, editors, *Topics in Information Theory*, volume 16 of *Colloquia Mathematica Societatis János Bolyai*, pages 309–321. North-Holland, 1975.

László Györfi and Tibor Nemetz. f-dissimilarity: A generalization of the affinity of several distributions. *Annals of the Institute of Statistical Mathematics*, 30:105–113, 1978.

Peter Harremoës and Naftali Tishby. The information bottleneck revisited or how to choose a good distortion measure. In *Proceedings of the IEEE International Symosium on Information Theory*, pages 566–570, 2007.

Peter Harremoës and Igor Vajda. On pairs of f-divergences and their joint range. Technical Report arXiv:1007.0097v1, ArXiV, July 2010.

Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

Andrey N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:1–7, 1965.

Andrey N. Kolmogorov and Sergei V. Fomin. *Introductory Real Analysis*. Dover, 1970.

Solomon Kullback and Richard Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951.

Ming Li. Information distance and its extensions. In T. Elomaa, J. Hollmén, and H. Mannila, editors, *Discovery Science*, LNCS 6926, pages 18–28. Springer, 2011.

Friederich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

Friederich Liese and Igor Vajda. f-divergences: Sufficiency, deficiency and testing of hypotheses. In Neil S. Barnett and Sever S. Dragomir, editors, *Advances in Inequalities from Probability Theory and Statistics*, pages 113–158. Nova Science Publishers, New York, 2008.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

Dennis V. Lindley. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Chong Long, Xiaoyan Zhu, Ming Li, and Bin Ma. Information shared by many objects. In *Proceedings of the 17th ACM conference on Information and Knowledge Management*, pages 1213–1220. ACM, 2008.

Kamei Matusita. On the notion of affinity of several distributions and some of its applications. *Annals of the Institute of Statistical Mathematics*, 19:181–192, 1967.

Kamei Matusita. Some properties of affinity and applications. *Annals of the Institute of Statistical Mathematics*, 23(1):137–155, 1971.

M. Luisa Menéndez, Julio A. Pardo, Leandro Pardo, and Konstantinos Zografos. A preliminary test in classification and probabilities of misclassification. *Statistics*, 39(3):183–205, 2005.

Neri Merhav. Data processing theorems and the second law of thermodynamics. *IEEE Transactions on Information Theory*, 57(8):4926–4939, August 2011.

Dominigo Morales, Leandro Pardo, and Konstantinos Zografos. Informational distances and related statistics in mixed continuous and categorical variables. *Journal of Statistical Planning and Inference*, 75:47–63, 1998.

Ferdinand Österreicher and Dorian Feldman. Divergenzen von Wahrscheinlichkeitsverteilungen — Integralgeometrisch Betrachtet. *Acta Mathematica Academaiae Scientiarum Hungarica*, 37(4): 329–337, 1981.

Ferdinand Österreicher and Igor Vajda. Statistical information and discrimination. *IEEE Transactions on Information Theory*, 39(3):1036–1039, 1993.

María del Carmen Pardo and Igor Vajda. About distances of discrete distributions satisfying the data processing theorem of information theory. *Information Theory, IEEE Transactions on*, 43 (4):1288–1293, 1997.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

R. Tyrrell Rockafellar and Roger J-B. Wets. *Variational Analysis*. Springer-Verlag, Berlin, 2004.

Andrea Sgarro. Informational divergence and the dissimilarity of probability distributions. *Calcolo*, 18(3):293–302, 1981.

Moshe Shaked and J. George Shanthikumar. *Stochastic Orders and their Applications*. Associated Press, 1994.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27: 379–423; 623–656, 1948.

Albert N. Shiryaev and Vladimir G. Spokoiny. *Statistical Experiments and Decisions: Asymptotic Theory*. World Scientific Publishing Company, 2000.

Robin Sibson. Information radius. *Probability Theory and Related Fields*, 14(2):149–160, 1969.

Erik N. Torgersen. Comparison of experiments when the parameter space is finite. *Probability Theory and Related Fields*, 16(3):219–249, 1970.

Erik N. Torgersen. Measures of Information Based on Comparison with Total Information and with Total Ignorance. *The Annals of Statistics*, 9(3):638–657, 1981.

Erik N. Torgersen. Stochastic orders and comparison of experiments. In *Stochastic Orders and Decision Under Risk*, pages 334–371. Institute of Mathematical Statistics, 1991a.

Erik N. Torgersen. *Comparison of Statistical Experiments*. Cambridge University Press, 1991b.

Godfried T. Toussaint. Probability of error, expected divergence and the affinity of several distributions. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6):482–485, 1978.

Igor Vajda. On the f-divergence and singularity of probability measures. *Periodica Mathematica Hungarica*, 2:223–234, 1972.

Elodie Vernet, Robert C. Williamson, and Mark D. Reid. Composite multiclass losses. In *NIPS*, 2011.

Elodie Vernet, Robert C. Williamson, and Mark D. Reid. Composite multiclass losses. Submitted to *Journal of Machine Learning Research*, June 2012.

Paul M.B. Vitányi. Information distance in multiples. *IEEE Transactions on Information Theory*, 57(4):2451–2456, 2011.

Paul M.B. Vitányi. Information distance: New developments. Technical report, Arxiv preprint arXiv:1201.1221, 2012.

Abraham Wald. *Statistical Decision Functions*. John Wiley & Sons, New York, 1950.

Moshe Zakai and Jacob Ziv. A generalization of the rate-distortion theory and applications. In Giuseppe Longo, editor, *Information Theory: New Trends and Open Problems*, pages 87–123. Springer, 1975.

Konstantinos Zografos. f-dissimilarity of several distributions in testing statistical hypotheses. *Annals of the Institute of Statistical Mathematics*, 50(2):295–310, 1998.

Appendix A. Proofs

Proof of Lemma 2 We can use the Radon-Nykodym theorem to write

$$\eta_i^{\boldsymbol{\pi}} = \frac{\pi_i dP_i}{\sum_{t=1}^k \pi_t dP_t} = \frac{\pi_i \frac{dP_i}{dP_j}}{\sum_{t=1}^k \pi_t \frac{dP_t}{dP_j}} = \frac{\pi_i \mathbf{t}_i^j}{\sum_{t=1}^k \pi_t \mathbf{t}_t^j}$$

■

Proof of Theorem 7

$$\begin{aligned}
\mathbb{I}_{\tilde{\phi}}(\mathbf{P}_{[k]}) &= \int \left(\phi \left(\left(\frac{dP_1}{dP_k}, \ldots, \frac{dP_{k-1}}{dP_k}, 1 \right)' \right) + \mathbf{w}' \left(1 - \frac{dP_1}{dP_k}, \ldots, 1 - \frac{dP_{k-1}}{dP_k}, 0 \right)' \right) dP_k \\
&= \mathbb{I}_{\phi}(\mathbf{P}_{[k]}) + \int \sum_{i=1}^k w_i \left(1 - \frac{dP_i}{dP_k} \right) dP_k \\
&= \mathbb{I}_{\phi}(\mathbf{P}_{[k]}) + \sum_{i=1}^k \int w_i \left(1 - \frac{dP_i}{dP_k} \right) dP_k \\
&= \mathbb{I}_{\phi}(\mathbf{P}_{[k]}) + \sum_{i=1}^k w_i \left(\int dP_k - \int dP_i \right) = \mathbb{I}_{\phi}(\mathbf{P}_{[k]}).
\end{aligned}$$

■

Proof of Theorem 9 Assume that two risk curves $\underline{L}(\boldsymbol{\eta})$ and $\tilde{\underline{L}}(\boldsymbol{\eta})$ result in the same statistical information for all $\mathbf{P}_{[k]}$ and denote $R(\boldsymbol{\eta}) = \underline{L}(\boldsymbol{\eta}) - \underline{L}'(\boldsymbol{\eta})$. From (8) this implies

$$R(\boldsymbol{\pi}) = \int R(\boldsymbol{\eta}^{\boldsymbol{\pi}}(\omega))dP^{\boldsymbol{\pi}}(\omega) =: c \quad \forall \mathbf{P}_{[k]}. \tag{13}$$

Consider the case where $P_1 \perp \ldots \perp P_k$ and denote $\tilde{\mathbf{w}} = \tilde{w}_{[k]}$, where $\tilde{w}_i = R(\mathbf{e}_i)$. Then

$$c = \tilde{\mathbf{w}}'\boldsymbol{\pi}. \tag{14}$$

Consider an arbitrary vector $\mathbf{u} \in [0,1]^k$ and k distributions over a finite set $\Omega = [k]$ given by $P_i(\omega) = u_i\delta(\omega) + (1-u_i)\delta(\omega - i)$, $i \in [k]$, where δ is the Kronecker delta function. Let $\boldsymbol{a}_{\mathbf{u}} := \frac{\boldsymbol{\pi}\circ\mathbf{u}}{\boldsymbol{\pi}'\mathbf{u}} \in \Delta^k$. Then from (13),

$$c = \int R(\boldsymbol{\eta}^{\boldsymbol{\pi}}(\omega))dP^{\boldsymbol{\pi}}(\omega) = R(\mathbf{a}_{\mathbf{u}})\mathbf{u}'\boldsymbol{\pi} + (\boldsymbol{\pi}\circ(\mathbf{1}-\mathbf{u}))'\,\tilde{\mathbf{w}}.$$

Together with (14) this implies that $R(\mathbf{a}_{\mathbf{u}})\mathbf{u}'\boldsymbol{\pi} = (\boldsymbol{\pi}\circ\mathbf{u})'\,\tilde{\mathbf{w}}$, so that $R(\mathbf{a}_{\mathbf{u}}) = \tilde{\mathbf{w}}'\mathbf{a}_{\mathbf{u}}$. This must hold for every $\mathbf{u} \in [0,1]^k$, and so for every $\mathbf{a}_{\mathbf{u}} \in \Delta^k$, and so

$$\Delta\underline{\mathbb{L}}(\boldsymbol{\pi}, \mathbf{P}_{[k]}) = \Delta\tilde{\underline{\mathbb{L}}}(\boldsymbol{\pi}, \mathbf{P}_{[k]}) \quad \forall \mathbf{P}_{[k]} \Rightarrow \tilde{\underline{L}}(\boldsymbol{\eta}) = \underline{L}(\boldsymbol{\eta}) + \tilde{\mathbf{w}}'\boldsymbol{\eta}.$$

The converse implication is given by Corollary 8, so we obtain a characterisation. Translating to divergences we use (9) to obtain

$$\phi^{\boldsymbol{\pi}}_{\underline{L}}(\mathbf{t}) - \phi^{\boldsymbol{\pi}}_{\tilde{\underline{L}}}(\mathbf{t}) = \boldsymbol{\pi}'\tilde{\mathbf{w}} - \boldsymbol{\pi}'\mathbf{t}\left(\frac{(\boldsymbol{\pi}\circ\mathbf{t})'\,\tilde{\mathbf{w}}}{(\boldsymbol{\pi}'\mathbf{t})\,\tilde{\mathbf{w}}}\right) = \boldsymbol{\pi}'\tilde{\mathbf{w}} - (\boldsymbol{\pi}\circ\tilde{\mathbf{w}})'\,\mathbf{t} = (\boldsymbol{\pi}\circ\tilde{\mathbf{w}})'\,(\mathbf{1}-\mathbf{t}).$$

Setting $\mathbf{w} = \boldsymbol{\pi}\circ\tilde{\mathbf{w}}$ gives the desired result. ■

Proof of Theorem 13 The lower bound is trivial and comes from Corollary 6. We prove the upper bound by using the link between divergences and statistical informations. It is easy to see that the statistical information $\Delta\underline{\mathbb{L}}(\boldsymbol{\pi}, \mathbf{P}_{[k]})$ is upper bounded by $\underline{L}(\boldsymbol{\pi}) - \sum_{i=1}^k \pi_i \underline{L}(\mathbf{e}_i)$, which equals $\Delta\underline{\mathbb{L}}(\boldsymbol{\pi}, \mathbf{P}_{[k]})$ for a totally informative experiment where $P_1 \perp \ldots \perp P_k$. We can write the convex function $\phi^{\boldsymbol{\pi}}_{\underline{L}}$ defining the equivalent divergence for a given $\boldsymbol{\pi}$ using (9). Then,

$$\phi^{\boldsymbol{\pi}}_{\underline{L}}(\mathbf{e}_k) = \underline{L}(\boldsymbol{\pi}) - \pi_k \lim_{\mathbf{t}\to\mathbf{e}_k} \underline{L}\left(\frac{\boldsymbol{\pi}\circ\mathbf{t}}{\boldsymbol{\pi}'\mathbf{t}}\right) - \sum_{i=1}^{k-1} \pi_i \lim_{\mathbf{t}\to\mathbf{e}_k} t_i \underline{L}\left(\frac{\boldsymbol{\pi}\circ\mathbf{t}}{\boldsymbol{\pi}'\mathbf{t}}\right),$$

where[3] $\lim_{\mathbf{t}\to\mathbf{t}_0} f(\mathbf{t}) = \lim_{||\mathbf{t}-\mathbf{t}_0||_2\to 0} f(\mathbf{t})$. If ϕ is bounded below then $\underline{L}$ is bounded above, so then $\phi^{\boldsymbol{\pi}}_L(\mathbf{e}_k) = \underline{L}(\boldsymbol{\pi}) - \pi_k\underline{L}(\mathbf{e}_k)$, since $\lim_{\mathbf{t}\to\mathbf{e}_k} t_i\underline{L}\left(\frac{\boldsymbol{\pi}\circ\mathbf{t}}{\boldsymbol{\pi}'\mathbf{t}}\right) \neq 0$ for any i implies that $\underline{L}(e_k)) \to -\infty$, so $\phi^{\boldsymbol{\pi}}_L(\mathbf{e}_k) = \infty$, as the equation predicts. Recalling the definition in (12) we have

$$\phi^{\diamond_j}(\mathbf{t}) = t_j\left[\underline{L}(\boldsymbol{\pi}) - \sum_{\substack{i=1\\ i\neq j}}^{k-1}\left(\pi_i\frac{t_i}{t_j} + \pi_j\frac{1}{t_j} + \pi_k\right)\underline{L}\left(\frac{\frac{1}{t_j}\boldsymbol{\pi}\circ\mathbf{t}^{j\leftrightarrow k}}{\frac{1}{t_j}\boldsymbol{\pi}'\mathbf{t}^{j\leftrightarrow k}}\right)\right].$$

3. Since $\mathbf{t} \in K^k \subset \mathbb{R}^k$ the choice of norm is actually irrelevant because all norms are equivalent in finite dimensions.

Similarly as above $(\phi^{\boldsymbol{\pi}})^{\diamond_j}(\mathbf{e}_k) = -\pi_j \underline{L}(\mathbf{e}_j)$, so

$$\phi^{\boldsymbol{\pi}}(\mathbf{e}_k) + \sum_{j=1}^{k-1} (\phi^{\boldsymbol{\pi}})^{\diamond_j}(\mathbf{e}_k) = \underline{L}(\pi) - \sum_{j=1}^{k} \pi_j \underline{L}(\mathbf{e}_j) \geq \Delta\underline{\mathbb{L}}(\pi, \mathbf{P}_{[k]}) = \mathbb{I}_{\phi^{\boldsymbol{\pi}}_{\underline{L}}}\left(\mathbf{P}_{[k]}\right) \forall \mathbf{P}_{[k]}.$$

■

Appendix B. Examples

Here we include the analyses of the the other examples in Table 2.

Triangular Discrimination Triangular discrimination is the f-divergence corresponding to the square loss

$$\underline{L}_{\mathrm{SQ}}(\boldsymbol{\eta}) = \frac{1}{2}\left(1 - \sum_{i=1}^{k} \eta_i^2\right)$$

in a binary setting when both classes are equally likely. Using the results from previous sections, we can find that the multiclass triangular discrimination is then defined by the following convex function

$$\phi_{\mathrm{SQ}}(\mathbf{t}) = \frac{1}{2k}\left(k - 1 - \sum_{i=1}^{k} t_i + \frac{\sum_{i=1}^{k} t_i^2}{\sum_{i=1}^{k} t_i}\right)$$

which in the binary case reduces to

$$f_{\mathrm{SQ}}(t) = \frac{1}{4}\frac{t-1}{t+1}$$

Jensen-Shannon divergence The binary Jensen-Shannon divergence

$$\mathrm{JS}(P, Q) = \frac{1}{2}\left(\mathrm{KL}\left(P, \frac{P+Q}{2}\right) + \mathrm{KL}\left(Q, \frac{P+Q}{2}\right)\right)$$

is well-known to correspond with the statistical information for the log-loss, whose point-wise Bayes risk is given by

$$\underline{L}_{\log}(\boldsymbol{\eta}) = \sum_{i=1}^{k} \eta_i \ln \frac{1}{\eta_i},$$

when $k = 2$ and both classes are equally likely ($\pi_1 = \pi_2 = \frac{1}{2}$). Using our multiclass f-divergence framework, we can get the following expression for the convex function defining the divergence associated to the log-loss in the general case (assuming also a uniform prior for notational simplicity).

$$\phi_{\mathrm{JS}}(\mathbf{t}) = \ln(k) + \frac{1}{k}\sum_{i=1}^{k} t_i \ln\left(\frac{t_i}{\sum_{j=1}^{k} t_j}\right)$$

In the binary case this reduces to $f_{\mathrm{JS}}(t) = \frac{1}{2}\left[t \ln\left(\frac{t}{t+1}\right) + \ln\left(\frac{4}{t+1}\right)\right]$, as in (Reid and Williamson, 2011, Table 1). The resulting divergence can be written in terms of standard, binary KL divergences

as follows

$$\mathbb{I}_{\mathrm{JS}}\left(\mathbf{P}_{[k]}\right) = \int \phi_{\mathrm{JS}}\left(\left(\frac{dP_1}{dP_k}, \ldots, \frac{dP_{k-1}}{dP_k}, 1\right)'\right)$$
$$= \int \left(\ln(k) - \frac{1}{k}\sum_{i=1}^{k} \frac{dP_i}{dP_k} \ln\left(\frac{\sum_{j=1}^{k} \frac{dP_j}{dP_k}}{\frac{dP_i}{dP_k}}\right)\right) dP_k =$$
$$= \ln(k) - \frac{1}{k}\left(\sum_{i=1}^{k} \int \ln \frac{\sum_{j=1}^{k} dP_j}{dP_i} dP_i\right) = \ln(k) - \frac{1}{k}\left(\sum_{i=1}^{k} \ln \frac{1}{k} + \int \ln \frac{\sum_{j=1}^{k} \frac{1}{k} dP_j}{dP_i} dP_i\right)$$
$$= \ln(k) - \ln k + \frac{1}{k}\sum_{i=1}^{k} \int \ln \frac{dP_i}{\sum_{j=1}^{k} \frac{1}{k} dP_j} dP_i dP_i = \frac{1}{k}\sum_{i=1}^{k} \mathrm{KL}\left(P_i, \frac{1}{k}\sum_{j=1}^{k} P_j\right),$$

which turns out to be a natural extension of the binary case in (B). In fact, this generalization of the Jensen-Shannon divergence coincides with the one which was proposed in Lin (1991): $\mathrm{JS}(\mathbf{P}_{[k]}) = H\left(\sum_{i=1}^{k} \pi_i P_i\right) - \sum_{i=1}^{k} \pi_i H(P_i)$, where $H(\cdot)$ denotes Shannon's entropy, since

$$H\left(\sum_{i=1}^{k} \pi_i P_i\right) - \sum_{i=1}^{k} \pi_i H(P_i) = \int \frac{1}{k}\sum_{i=1}^{k} dP_i \ln\left(\frac{1}{\frac{1}{k}\sum_{j=1}^{k} dP_j}\right) - \sum_{i=1}^{k} \frac{1}{k} \int \ln\left(\frac{1}{dP_i}\right) dP_i$$
$$= \frac{1}{k}\sum_{i} \left[\int \ln \frac{1}{\frac{1}{k}\sum_{j} dP_j} dP_i - \int \ln \frac{1}{dP_i} dP_i\right]$$
$$= \frac{1}{k}\sum_{i=1}^{k} H\left(P_i, \frac{1}{k}\sum_{j=1}^{k} P_j\right) - H(P_i) = \frac{1}{k}\sum_{i=1}^{k} \mathrm{KL}\left(P_i, \frac{1}{k}\sum_{j=1}^{k} P_j\right),$$

where $H(\cdot, \cdot)$ denotes the cross-entropy and we have also assumed uniform prior probabilities $\pi_1 = \cdots = \pi_k = \frac{1}{k}$. Hence our general framework for multiclass f-divergences naturally encompasses the existing multiclass Jensen-Shannon divergence.

Matusita divergence The Matusita affinity between distributions (Matusita, 1967) is given by

$$\rho(\mathbf{P}_{[k]}) = \int \left(\prod_{i=1}^{k} \frac{dP_i}{d\lambda}\right)^{\frac{1}{k}} d\lambda$$

where λ is any measure dominating $P_1, \ldots, P_k$. The f-divergence corresponding to the convex function $\phi_\rho \in \mathcal{C}_1^k$

$$\phi_\rho(\mathbf{t}) = 1 - \left(\prod_{i=1}^{k} t_i\right)^{\frac{1}{k}}$$

can be written as

$$\mathbb{I}_{\phi_\rho}\left(\mathbf{P}_{[k]}\right) = -\int \left(\prod_{i=1}^{k} \frac{dP_i}{dP_k}\right)^{\frac{1}{k}} dP_k = -\int \left(\prod_{i=1}^{k} \frac{dP_i}{d\lambda}\right)^{\frac{1}{k}} d\lambda = -\rho(\mathbf{P}_{[k]}),$$

so we refer to that divergence as Matusita's divergence (Györfi and Nemetz, 1975). Using (7), the risk corresponding to this divergence is given by

$$\underline{L}_\rho^{\boldsymbol{\pi}}(\boldsymbol{\eta}) = -\frac{\eta_k}{\pi_k} + \left(\prod_{i=1}^{k} \frac{\eta_i}{\pi_i}\right)^{\frac{1}{k}}$$

When $\pi_1 = \cdots = \pi_k = \frac{1}{k}$ this reduces to $\underline{L}_\rho(\boldsymbol{\eta}) = k\left[\left(\prod_{i=1}^{k} \eta_i\right)^{\frac{1}{k}} - \eta_k\right]$. By Corollary 8, this point-wise risk generates the same statistical information as

$$\underline{L}_\rho(\boldsymbol{\eta}) = k\left(\prod_{i=1}^{k} \eta_i\right)^{\frac{1}{k}},$$

which in the binary case corresponds to $[0,1] \ni \eta \mapsto \underline{L}(\eta) = 2\sqrt{\eta(1-\eta)}$.