Risk-Based Generalizations of f-divergences

Darío García-García DARIO.GARCIA@ANU.EDU.AU
Research School of Computer Science, Australian National University and NICTA, Canberra, Australia

Ulrike von Luxburg ULRIKE.LUXBURG@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems, Tübingen, Germany

Raúl Santos-Rodríguez RSRODRIGUEZ@TSC.UC3M.ES
Dpto. Teoría de la Señal y Comunicaciones, Universidad Carlos III de Madrid, Leganés, Spain

Abstract

We derive a generalized notion of f-divergences, called (f, l)-divergences. We show that this generalization enjoys many of the nice properties of f-divergences, although it is a richer family. It also provides alternative definitions of standard divergences in terms of surrogate risks. As a first practical application of this theory, we derive a new estimator for the Kulback-Leibler divergence that we use for clustering sets of vectors.

1. Introduction

In this paper we are interested in defining functions that measure the divergence between probability distributions, motivated by the problem of clustering sets of points. This scenario arises mainly when working with sequences of data whose dynamical information can be discarded. Typical examples include speaker recognition, bag-of-words models for images or language processing. In order to leverage standard algorithms (such as spectral clustering) in this scenario, it is necessary to define a similarity function between sets $\mathbf{X}, \mathbf{Y}$ of points in some input space. This similarity should be a function of how separated the two sets $\mathbf{X}$ and $\mathbf{Y}$ are. It is a natural idea to measure the amount of separation between two sets with the help of a classifier. In the simplest case, we simply train a classifier that is supposed to separate the points $\mathbf{X}$ from $\mathbf{Y}$ and use its error rate as a similarity score. Intuitively, if the sets $\mathbf{X}$ and $\mathbf{Y}$ "overlap a lot", then the classifier will have a high error rate, which we interpret as a high similarity. If, on the other hand, $\mathbf{X}$ and $\mathbf{Y}$ are well separated from each other, the classifier will achieve a low error, leading to a low similarity score.

More abstractly, consider two probability distributions P, Q on the same space and their convex combination $\pi P + (1 - \pi)Q$ for some weight parameter $\pi \in [0, 1]$. Assign labels $+1$ to all points that have been drawn from P, and labels -1 to all points drawn from Q. The classification task (π, P, Q, l) consists in finding the optimal classification function for this setting under a given loss function l. We can now define a similarity score between P and Q as the overall expected loss for this task. In case we allow any possible classification function and choose the 0-1 loss, this similarity score becomes the Bayes error of the classification task. This approach immediately rises a couple of questions: what is the value of π we should use, and what is the best loss function l in order to obtain a meaningful similarity score? Moreover, using the Bayes error as a similarity measure is problematic, since it is hard to estimate. Estimation of risks might become easier if we restrict the classification function to a simple (e.g. parametric) family, like linear classifiers. However, this effectively imposes limitations on the features of the distributions that are being taken into account by the similarity measures. This can be beneficial if there is some domain knowledge substantiating that limitations, but detrimental in general. Moreover, optimizing the 0-1 loss is still a complex problem as it cannot be handled analytically. Instead, *surrogate losses* (Bartlett et al., 2006) are usually employed. They are functions that share some features of the 0-1 loss while being well-behaved.

A first pragmatic approach for defining risk-based affinities is to use the nearest neighbor (NN) rule. Being a non-parametric method, it can capture arbitrary "shapes" of the distributions, making it flexible

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

enough to define similarities between sets of points. From a theoretical point of view, we show that the NN risk is closely related to the Bayes risk for the well-known square loss (Sec. 2). From a practical point of view, there exist several efficient alternatives for obtaining error estimates with good distribution-free performance guarantees (Sec. 5.1).

We then derive a more general approach using the framework of f-divergences (Ali & Silvey, 1966) as a starting point. Many of the best-known divergence functions are members of this flexible family, all of which admit an integral representation in terms of Bayes errors. In this paper we deal with the effects of surrogating the 0-1 loss in such integral representations to define (f, l)-divergences. The class of (f, l)-divergences shares many properties with the f-divergences, but is richer. It also provides alternative representations of well-known divergences. As a first application of this general framework, we show how (f, l)-divergences can be used to obtain new estimators and bounds for the Kullback-Leibler divergence.

Notation and definitions

Let P, Q be a pair of probability distributions, and M their convex combination $M := \pi P + (1-\pi)Q$ for $\pi \in [0,1]$. Given a classification task (π, P, Q) whose goal is to assign labels $Y = 1$ to points coming from P and $Y = 0$ to points from Q, we denote by $\eta = P(Y = 1|X = x)$ and $\hat{\eta}$ the posterior class probability and its estimate, respectively. The representations (π, P, Q) and (η, M) are interchangeable.

We write $\mathbb{E}_P[f]$ for the expectation of a function $f(x)$ of a random variable $x \sim P$. Let l be a loss function $l : \{0,1\} \times [0,1] \to \mathbb{R}$. The point-wise risk L_l associated to l is given by $L_l(\eta(x), \hat{\eta}(x)) = \eta(x)l(1, \hat{\eta}(x)) + (1-\eta(x))l(0, \hat{\eta}(x))$, and the (expected) risk $\mathbb{L}_l$ is thus $\mathbb{L}_l(\eta, M) = \mathbb{E}_M[L_l(\eta(x), \hat{\eta}(x))]$. Optimal or *Bayes* risks are denoted by an underline, so $\underline{L}_l(\eta(x)) = \inf_{\hat{\eta}(x)} L_l(\eta(x), \hat{\eta}(x))$ and $\underline{\mathbb{L}}_l(\pi, P, Q) = \underline{\mathbb{L}}_l(\eta, M) = \mathbb{E}_M[\underline{L}_l(\eta(x))]$. The *prior* Bayes risk is the optimal risk when only the prior class probability π is known $\underline{\mathbb{L}}_l(\pi) = \underline{L}_l(\pi)$.

2. NN error and surrogate Bayes risks

In this section we show how the asymptotic error rate of the NN rule relates to the Bayes risk for a certain loss. Consider the *square loss* $l_{SQ}(y, \hat{y})$ defined over $\{0,1\} \times [0,1]$ as

$$l_{SQ}(y, \hat{y}) = \begin{cases} (1-\hat{y})^2 & , y = 1 \\ \hat{y}^2 & , y = 0 \end{cases} \tag{1}$$

It induces the following point-wise risk.

$$L_{SQ}(\eta(x), \hat{\eta}(x)) = \eta(x)(1-\hat{\eta}(x))^2 + (1-\eta(x))\hat{\eta}(x)^2.$$

It is very easy to check that, for a given $\eta(x)$, the minimum of $L_{SQ}(\eta(x), \hat{\eta}(x))$ is achieved whenever $\hat{\eta}(x) = \eta(x)$. Losses that induce a point-wise risk satisfying this intuitive property are known as *proper losses* (Buja et al., 2005). The corresponding optimal (Bayes) point-wise risk is then given simply by

$$\underline{L}_{SQ}(\eta(x)) = \eta(x)(1-\eta(x)).$$

It is well-known (see e.g. Devroye et al. (1996), Chap. 5) that the asymptotic error rate for the NN rule $\mathbb{L}^{NN}_{0-1}$ can be written as

$$\mathbb{L}^{NN}_{0-1}(\eta, M) = \mathbb{E}_M[2\eta(x)(1-\eta(x))].$$

The following theorem is then obvious from the above discussion and the definition of expected Bayes risk.

Theorem 2.1. *For any pair of distributions P, Q and prior probability $\pi \in [0,1]$, the asymptotic error rate of the NN rule satisfies*

$$\mathbb{L}^{NN}_{0-1}(\pi, P, Q) = 2\underline{\mathbb{L}}_{SQ}(\pi, P, Q).$$

So the error probability of the NN rule provides a way to estimate the Bayes risk for the square loss. It is worth mentioning that the NN rule is not minimizing the risk under the square loss, since there is a factor of 2 in the formula. However, both magnitudes are in a one-to-one correspondence.

3. Background on f-divergences

In this section we recapitulate definitions and known facts about f-divergences. Given a convex function $f : (0, \infty) \to \mathbb{R}$, with $f(1) = 0$, the corresponding f-divergence (Ali & Silvey, 1966) between two probability distributions P, Q over an input space $\mathcal{X}$ is defined as

$$\mathbb{I}_f(P, Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] = \int_{\mathcal{X}} dQ f\left(\frac{dP}{dQ}\right),$$

if P is absolutely continuous with respect to Q, and ∞ otherwise. Many well-known divergences can be cast into this framework by adequately choosing the generating function f. Some important examples include the variational, Kullback-Leibler (KL) and Pearson's χ^2 divergences.

Our discussion will be based mainly on a classical result (see e.g. Österreicher & Vajda (1993)) that shows how f-divergences can be represented by a weighted integral of *statistical informations* $\Delta\underline{\mathbb{L}}_{0-1}(\pi, P, Q)$ under

the 0-1 loss. These informations can be intuitively interpreted as the risk reduction provided by the knowledge of the exact posterior probability η instead of just the prior probability π. They are defined as

$$\begin{aligned}\Delta\underline{\mathbb{L}}_{0-1}(\pi, P, Q) &= \underline{\mathbb{L}}_{0-1}(\pi) - \underline{\mathbb{L}}_{0-1}(\pi, P, Q) \\ &= \min(\pi, 1-\pi) - \underline{\mathbb{L}}_{0-1}(\pi, P, Q).\end{aligned}$$

The integral representation of f-divergences is given by

$$\mathbb{I}_f(P, Q) = \int_0^1 \Delta\underline{\mathbb{L}}_{0-1}(\pi, P, Q)\gamma_f(\pi)d\pi, \qquad (2)$$

where the weight function $\gamma_f(\pi)$ is related to the curvature of the function f defining the divergence

$$\gamma_f(\pi) = \frac{1}{\pi^3} f''\left(\frac{1-\pi}{\pi}\right). \qquad (3)$$

Since f is a convex function, the weights $\gamma_f(\pi)$ are non-negative. For a comprehensive list of well-known f-divergences and their associated f and weight functions please refer to Reid & Williamson (2011).

4. (f, l)-divergences

We propose a risk-based generalization of the family of f-divergences, based on the integral representation in Eq. (2). The main idea is to substitute the 0-1 loss for an arbitrary loss function l. This way, we can express this new generalization as follows.

Definition For a convex function $f : (0, \infty) \to \mathbb{R}$ with $f(1) = 0$, and a loss $l : \{0, 1\} \times [0, 1] \to \mathbb{R}_+$, we define the corresponding (f, l)-*divergence* $\mathbb{I}_{f,l}$ as

$$\mathbb{I}_{f,l} = \int_0^1 \Delta\underline{\mathbb{L}}_l(\pi, P, Q)\gamma_f(\pi)d\pi, \qquad (4)$$

where $\gamma_f(\pi)$ is given by Eq. (3) and

$$\Delta\underline{\mathbb{L}}_l(\pi, P, Q) = \underline{\mathbb{L}}_l(\pi) - \underline{\mathbb{L}}_l(\pi, P, Q). \qquad (5)$$

Obviously, the original f-divergences can be obtained as a particular case of (f, l)-divergences by setting $l = l_{0-1}$. Note that the idea of substituting 0-1 for more general losses is at the core of almost every practical classifier. This is the idea of *surrogate losses* (Bartlett et al., 2006): Since the 0-1 loss is not very well behaved and thus hard to handle, most learning algorithms use, explicitly or implicitly, other kind of losses that approximate the 0-1 loss while being much more amenable to theoretical analysis and numerical optimization. These surrogates are almost always[1] proper losses whose second term is mapped from $[0, 1]$ to $\mathbb{R}$. Thus, if the goal is to define divergences that can be nicely estimated using classification risks it is very natural to work with surrogate/proper losses, since they are what most practical classifiers optimize.

[1]The most important exception being the hinge loss

4.1. Some properties of (f, l)-divergences

In this section we will study how we can get interesting properties for (f, l)-divergences by adequately choosing the loss l. We will implicitly assume all losses to be proper (see Sec. 2).

As we will show in Sec. 4.2, (f, l) and f-divergences are deeply connected, so it is natural to recover most properties of standard f-divergences with a sensible selection of the loss function l. For an overview of the most important properties of f-divergences, please refer to Österreicher (2002). Due to space constraints, we show in the form of short theorems a small representative selection of such properties, along with the conditions that the losses must satisfy in order for those properties to hold. We sketch the proofs, which are quite straight-forward.

Theorem 4.1 (Non-negativity and identity of indiscernibles)**.** *For any convex f and any proper loss l, $\mathbb{I}_{f,l}(P, Q) \geq 0$ for all P, Q. Moreover, if f is non-trivial ($\exists \pi \in (0, 1) \mid \gamma_f(\pi) > 0$) and l is such that $\underline{L}_l$ is strictly concave, then equality holds iff $P = Q$.*

This theorem can be easily proved by applying Jensen's inequality, noting that point-wise Bayes risks $\underline{L}_l$ induced by proper losses are always concave (Savage, 1971). It is easy to check that most common proper losses, such as square or log-losses, induce strictly concave point-wise Bayes risks $\underline{L}_l$, so the condition is not very restrictive.

Theorem 4.2 (Symmetry)**.** *If l is a proper loss such that $l(0, \hat{\eta}) = l(1, 1-\hat{\eta})$, then $\mathbb{I}_{f,l}(P, Q) = \mathbb{I}_{f,l}(Q, P)$ if $f(t) = f^*(t) + c(t-1)$, $c \in \mathbb{R}$, where f^* is the Csiszar's dual (or $*$-conjugate) of function f.*

This is analogous to the standard symmetry property of f-divergences. The proof uses the fact that the condition on f implies $\gamma_f(\pi) = \gamma_f(1-\pi)$, and then it mainly involves showing that $\Delta\underline{\mathbb{L}}_l(\pi, P, Q) = \Delta\underline{\mathbb{L}}_l(1-\pi, Q, P)$ for $\pi \in [0, 1]$. Once again, standard losses satisfy the simple and natural condition imposed on l for the symmetry property to hold.

Theorem 4.3 (Information Processing)**.** $\mathbb{I}_{f,l}(P, Q) \geq \mathbb{I}_{f,l}(\Phi(P), \Phi(Q))$*, where Φ is any transformation.*

This is also analogous to a standard f-divergences property. The proof relies on the non-decreasing property of Bayes risks under arbitrary transformations.

4.2. Connecting f and (f,l)-divergences

In this section we show how some (f,l)-divergences are equivalent to standard f-divergences via a transformation of the weight function depending on the loss l. This will provide insight into the effect of using a surrogate loss for divergence definition, as well as motivating surprising ways of estimating some well-known divergences.

The discussion is based on the one-to-one relationship between statistical informations and f-divergences, as stated in the following classical result

Theorem 4.4 (Österreicher & Vajda (1993), Thm. 2)**.**
Given an arbitrary loss l, then defining

$$f_l^\pi(t) = \underline{L}_l(\pi) - (\pi t + 1 - \pi)\underline{L}_l\left(\frac{\pi t}{\pi t + 1 - \pi}\right) \quad (6)$$

for $\pi \in [0,1]$ implies f_l^π is convex and $f_l^\pi(1) = 0$, and

$$\Delta\underline{\mathbb{L}}_l(\pi, P, Q) = \mathbb{I}_{f_l^\pi}(P, Q) \quad (7)$$

for all distributions P and Q.

This may seem at odds with the result in Nguyen et al. (2009) which establish a many-to-one relationship between losses and f-divergences. However, note that in that work they are concerned with margin classification losses, while here we work with proper losses. The many link functions that can be coupled with a given proper loss to yield classification losses introduce that extra degree of freedom (Reid & Williamson, 2011).

Exploiting this representation of statistical information for arbitrary losses, Eq. (4) can be rewritten as $\mathbb{I}_{f,l} = \int_0^1 \mathbb{I}_{f_l^\pi}(P,Q)\gamma_f(\pi)d\pi$. Now we can leverage the weighted integral representation of $\mathbb{I}_{f_l^\pi}$ as given by Eq. (2), yielding

$$\begin{aligned}
\mathbb{I}_{f,l} &= \int_0^1 \left(\int_0^1 \Delta\underline{\mathbb{L}}_{0-1}(\pi', P, Q)\varphi_{l,\pi}(\pi')d\pi'\right)\gamma_f(\pi)d\pi \\
&= \int_0^1 \Delta\underline{\mathbb{L}}_{0-1}(\pi', P, Q)\left(\int_0^1 \varphi_{l,\pi}(\pi')\gamma_f(\pi)d\pi\right)d\pi' \\
&= \int_0^1 \Delta\underline{\mathbb{L}}_{0-1}(\pi, P, Q)\gamma_{f,l}(\pi)d\pi, \quad (8)
\end{aligned}$$

where $\varphi_{l,\pi}(\pi')$ is the weight function corresponding to f_l^π, as given by Eq. (3)

$$\varphi_{l,\pi}(\pi') = \frac{1}{\pi^3} f_l^{\pi\,\prime\prime}\left(\frac{1-\pi}{\pi}\right). \quad (9)$$

So we get the following theorem.

Theorem 4.5. *Assume a (f,l)-divergence with weight function $\gamma_f(\pi)$ and loss function l. Let $\varphi_{l,\pi}$ be given by Eq. (9). Whenever*

$$\gamma_{f,l}(\pi) = (T_l\gamma_f)(\pi) = \int_0^1 \varphi_l(\pi,\pi')\gamma_f(\pi')d\pi'$$

converges, then that (f,l)-divergence is equivalent to a standard f-divergence with weight function $\gamma_{f,l}(\pi)$.

In this case, both divergences are intrinsically the same one, but expressed on different bases. The relationships between the weight functions is given by a linear operator T_l with kernel $\varphi_l(\pi,\pi') \equiv \varphi_{l,\pi}(\pi')$. This connection has the important effect of allowing the estimation of standard f-divergences by using statistical informations under adequate proper/surrogate losses.

Note that Reid & Williamson (2011) connect losses and f-divergences by associating a loss l with a divergence with $f = f_l^{\frac{1}{2}}$ (see Thm. 4.4). That can be seen to be a particular case of (f,l)-divergences when f is chosen to represent the *variational divergence* V, since $\gamma_V \propto \delta(\pi - \frac{1}{2})$.

4.3. A worked-out example: Square loss

Here we will show how the above results particularize to the square loss defined in Section 2. Using Eq. (6) we can get the f function associated to the statistical information under that loss,

$$f_{SQ}^\pi(t) = \pi(1-\pi) - \frac{\pi(1-\pi)t}{\pi t + 1 - \pi}. \quad (10)$$

The weight function of the integral representation of $\mathbb{I}_{f_{SQ}^\pi}$ can be obtained by plugging in the above result into Eq. (9). With a little algebra we get

$$\varphi_{SQ}(\pi,\pi') = \frac{2(1-\pi')^2\pi'^2}{(\pi'(1-2\pi)+\pi)^3}. \quad (11)$$

Let us now apply this kernel to find the equivalent f-divergence of some (f,SQ)-divergences. With some hindsight, we start with Jeffreys (J) divergence, which is a symmetrized version of KL. The weight function corresponding to the integral representation of the J divergence is given by $\gamma_J(\pi) = \frac{1}{\pi^2(1-\pi)^2}$. We then get this very simple and interesting expression for the final weights

$$\gamma_{J,SQ}(\pi) = (T_{SQ}\gamma_J)(\pi) = \frac{1}{\pi^2(1-\pi)^2} \quad (12)$$

that is to say, the weight function associated with the f-divergence equivalent of the (J,SQ)-divergence is exactly the same weight function of the standard Jeffreys divergence. An analogous result holds for the

KL divergence, whose weights are given by $\gamma_{KL}(\pi) = \gamma_{KL,SQ}(\pi) = \frac{1}{\pi^2(1-\pi)}$. The weight functions for both KL and Jeffreys divergences are eigenfunctions of the integral operator T_{SQ} with eigenvalue 1. In some sense, they are *eigendivergences* of the square loss. This is summarized in the following corollary.

Corollary 4.6.

$$\begin{aligned} \mathbb{I}_J(P,Q) &= \mathbb{I}_{J,SQ}(P,Q) \\ \mathbb{I}_{KL}(P,Q) &= \mathbb{I}_{KL,SQ}(P,Q) \end{aligned}$$

Note that many (f, SQ)-divergences cannot be realized as standard f-divergences. Consider for example the (χ^2, SQ)-divergence. Since $\gamma_{\chi^2}(\pi) = \frac{1}{\pi^3}$, applying the integral operator to try to express it as an f-divergence yields $\gamma_{\chi^2,SQ}(\pi) = 2\int_0^1 \frac{(1-\pi')^2}{\pi'} \frac{1}{(\pi'(1-2\pi)+\pi)^3}$, which diverges, showing that the (χ^2, SQ)-divergence is not an f-divergence. In particular, this negative example shows that the class of (f, l)-divergences is strictly larger than the class of f-divergences.

5. Application: divergence estimation via NN errors

Coupling Thm. 2.1 with Corollary 4.6 shows that it is possible to define KL divergences as weighted integrals of NN error rates. Without this result, the obvious way of using the integral representation to estimate an f-divergence would be to plug-in a consistent classifier (such as k-NN with an adequate election of k) or class probability estimator to obtain the 0-1 Bayes risks. Most recent proposals for KL divergence estimation relies on direct estimation of the likelihood ratio (Nguyen et al., 2008; Wang et al., 2009; Suzuki et al., 2009), and thus of the posterior class probabilities, avoiding individual density estimation. Our proposal avoids any explicit density and likelihood ratio or posterior estimation. Instead, we have shown that it is possible to use the risk of a simple, non-consistent classifier such as NN to obtain an error-rate based exact expression for the KL divergence (interestingly, using the exact same weight function as we would use with the Bayes errors).

5.1. Estimating the NN risk

From an empirical estimation point of view, the problem remains to obtain good estimates of the NN error rate for the whole range of prior probabilities $\pi \in [0, 1]$. The particularities of the NN rule can be exploited to obtain closed-form estimates of the error rate for a given π. One example is *complete stratified cross validation* (Mullin & Sukthankar, 2000). The idea behind complete cross-validation is to obtain the expectation of the risk over all the possible test/train partitions of data satisfying the desired proportions, instead of resorting to empirical resampling. Nonetheless, running this process for a large enough number of π's is a very time consuming process.

To speed things up we have devised a simple "closed-form sampling scheme", specially tailored for the task of estimating risks over the whole range of prior probabilities, which we now sketch. The main idea is to subsample just one of the sets, depending on π. Assume we are given two sets $\mathbf{X}$ and $\mathbf{Y}$, with n_X and n_Y elements coming from P and Q respectively, so the estimated prior probability is just $\pi_0 = \frac{n_X}{n_X+n_Y}$. The error for $\pi = \pi_0$ can be estimated using standard methods such as deleted estimate (Devroye et al. (1996) Chap. 24), yielding error estimates in $\{0, 1\}$ for each point $z \in (\mathbf{X} \cup \mathbf{Y})$. In order to obtain error estimates for $\pi \neq \pi_0$, our proposal is to calculate the expectation of the probability of error at each point z given that we are subsampling $\mathbf{X}$ if $\pi < \pi_0$ or $\mathbf{Y}$ if $\pi > \pi_0$. We can obtain the desired expectation just by knowing the order of the closest point to z in both $\mathbf{X}$ and $\mathbf{Y}$ and calculating the ratio of partitions that result in the point changing its label with respect to case $\pi = \pi_0$. For example, consider the case of a point $z \in \mathbf{X}$ which is correctly classified for $\pi = \pi_0$, and whose closest point in $\mathbf{Y}$ occupies the $k_Y(z)$ position in the ordered list of neighbors. Let $n_s(\pi)$ be the number of points from $\mathbf{X}$ that must be taken away for the desired π to hold. Point z will become incorrectly classified whenever its nearest neighbor after subsampling belongs to $\mathbf{Y}$. That is to say, whenever the $k_Y(z) - 1$ first neighbors of z are taken away from $\mathbf{X}$. This is a sampling-without-replacement scenario, and the probability of such an event is given by the hypergeometric distribution, yielding $P_e(z;\pi) = \binom{n_x-(k_Y(z)-1)}{n_s(\pi)-(k_Y(z)-1)} \cdot \binom{n_x}{n_s(\pi)}^{-1}$. The reasoning is similar for points which are originally incorrectly classified. Note that this method is based solely on the order of the neighbors of each point.

5.2. Risk-based bounds of KL and Jeffreys divergences

Finally, upper bounds on the NN error rate can be used to obtain lower bounds on the estimated divergences. For example, consider the following result.

Theorem 5.1. *For all distributions P, Q over $\mathcal{X}$ with finite second moment we have*

$$\begin{aligned} \mathbb{I}_{KL}(P,Q) &\geq \int_0^1 \frac{(1-\pi)\Delta^2(\pi,P,Q)}{1+\pi(1-\pi)\Delta^2(\pi,P,Q)} d\pi. \\ \mathbb{I}_J(P,Q) &\geq \int_0^1 \frac{\Delta^2(\pi,P,Q)}{1+\pi(1-\pi)\Delta^2(\pi,P,Q)} d\pi. \end{aligned}$$

where Δ stands for the Mahalanobis distance between P and Q with prior probability π

$$\Delta(\pi, P, Q) = \sqrt{(\mu_p - \mu_q)^T \Sigma^{-1} (\mu_p - \mu_q)},$$

with $\Sigma(\pi, P, Q) = \pi\Sigma_p + (1-\pi)\Sigma_q$, $\Sigma_p = \mathbb{E}[(x-\mu_p)(x-\mu_p)^T]$ and analogously for Σ_q.

The theorem is obtained by plugging into Eq. (4) the following bound on the NN error rate due to Devijver (see e.g. Devroye et al. (1996), Chap. 5).

$$\mathbb{L}_{0-1}^{NN}(\pi, P, Q) \leq \frac{2\pi(1-\pi)}{1 + \pi(1-\pi)\Delta^2(\pi, P, Q)}.$$

5.3. Experimental results

To bridge the gap between theory and practice, in this section we will study the square loss-based divergence measures in both synthetic and real-world applications. Given the huge flexibility of the (f, l)-divergence framework, we have to restrict ourselves to some particular case. Specifically, we will focus on using the Nearest Neighbor classifier to estimate KL divergences, since that is arguably the most straightforward application of the theoretical results.

Based on above results, KL divergence can be expressed in terms of NN errors.

$$\mathbb{I}_{KL}(P, Q) = \frac{1}{2} \int_0^1 \Delta\underline{\mathbb{L}}_{0-1}^{NN}(\pi, P, Q) \gamma_{KL}(\pi) d\pi,$$

where $\Delta\underline{\mathbb{L}}_{0-1}^{NN}(\pi, P, Q) = \mathbb{L}_{0-1}^{NN}(\pi) - \mathbb{L}_{0-1}^{NN}(\pi, P, Q) = 2\pi(1-\pi) - \mathbb{L}_{0-1}^{NN}(\pi, P, Q)$. We have devised a naive estimation procedure, consisting of quadrature integration with uniform sampling of $\pi \in [\pi_{min}, \pi_{max}]$. A more sophisticated approach could be taken by using some kind of importance sampling depending on the weight function γ_{KL}. The error rates $\mathbb{L}_{0-1}^{NN}$ at each π are estimated using our procedure sketched in Section 5.1. The thresholds on π can be used in a way akin to the usual assumption in divergence estimation that the likelihood ratio is bounded and falls within some given thresholds. Statistical informations outside these thresholds are assumed to be 0, effectively regularizing the divergence estimate. In our experiments we fix $\pi_{min} = 10^{-3}, \pi_{max} = 1 - 10^{-3}$. We denote this non-parametric estimator NN-KL. The same approach has been used for obtaining an estimator of the bound in Eq. (13), yielding algorithm NNbound-KL.

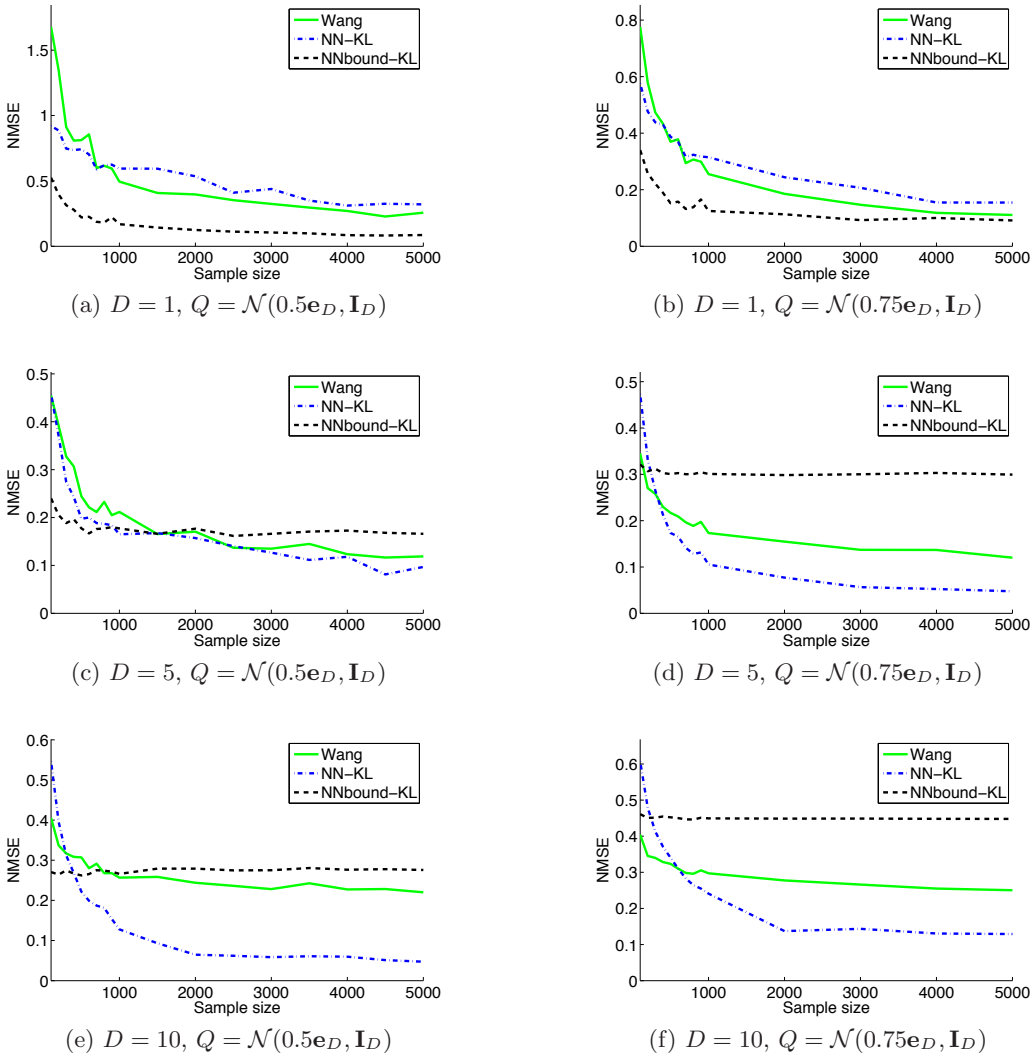
5.4. KL divergence estimation

Our benchmark for divergence estimation will be the proposal in Wang et al. (2009), which is arguably the state-of-the-art in non-parametric estimators for KL divergence. It is based on direct estimation of the likelihood ratio at each point using nearest-neighbor distances $\hat{\mathbb{I}}_{KL}(P, Q) = \frac{D}{n_X} \sum_{i=1}^{n_X} \log \frac{\nu_k(x_i)}{\rho_k(x_i)} + \log \frac{n_Y}{n_X - 1}$, where $\nu_k(z)$ and $\rho_k(z)$ are the distances from $x \in \mathbf{X}$ to its k-th nearest neighbor in $\mathbf{Y}$ and $\mathbf{X}$ respectively, and D is the dimension of the data. This algorithm was shown to outperform previous proposals, like data-dependent partitions or direct kernel plug-in estimates. In our experiments we have used $k = 1$.

We have run the algorithms in synthetic datasets comprised of samples from Gaussian distributions of different dimensionalities, with unit covariance matrices. Figure 1 show plots of normalized mean square error (NMSE) (averaged over 100 runs) using separations of $\mu_Q = 0.5\mathbf{e}_D$ (left) and $\mu_Q = 0.75\mathbf{e}_D$ (right), where $\mathbf{e}_D$ is the unit vector in $\mathbb{R}^D$, for different dimensionalities $D = \{1, 5, 10\}$. The NN-KL estimator improves its performance in comparison with both the Wang estimator and the risk-based lower bound as the dimensionality increases. Intuitively, in high-dimensional scenarios it may be easier to estimate error rates than likelihood ratios. The abrupt change in MSE slope of the NN-KL estimator in Fig.1f is due to the thresholds on π limiting the divergence estimate. Figure 2 shows the results for KL divergence estimation between samples from Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_3)$ and a uniform distribution $\text{Unif}[-3, 3]^3$. In this case, the Mahalanobis-based bound for the KL divergence is totally useless, since both distributions have the same mean. The Wang estimator achieves an impressive performance in this scenario. Nonetheless, the NN-error based estimator remains competitive. In general, our proposed estimator is competitive with the state of the art, showing that risk-based estimation of divergence measures is a promising line to explore.

5.4.1. Musical Genre Clustering

Following García-García et al. (2010), we define a clustering task on a subset of the *gareageband* dataset consisting of snippets of around 60s of songsbelonging to the following genres: "Punk", "Heavy Metal", "Classical", and "Reggae". There are 100 songs from each genre. For preprocessing, first Mel frequency cepstral coefficients (MFCCs) are extracted in overlapped windows of short duration. Then, a multivariate autoregressive model of lag 3 is fitted to each block of the sequence of MFCCs, using 2s windows and 1s hop-size. Data dimensionality is $D = 135$. Since spectral clustering works with symmetric affinity matrices, we choose to estimate Jeffreys divergence. We also introduce into the comparison the best results reported in García-García et al. (2010) (algorithm SSD) and



(a) $D = 1$, $Q = \mathcal{N}(0.5\mathbf{e}_D, \mathbf{I}_D)$

(b) $D = 1$, $Q = \mathcal{N}(0.75\mathbf{e}_D, \mathbf{I}_D)$

(c) $D = 5$, $Q = \mathcal{N}(0.5\mathbf{e}_D, \mathbf{I}_D)$

(d) $D = 5$, $Q = \mathcal{N}(0.75\mathbf{e}_D, \mathbf{I}_D)$

(e) $D = 10$, $Q = \mathcal{N}(0.5\mathbf{e}_D, \mathbf{I}_D)$

(f) $D = 10$, $Q = \mathcal{N}(0.75\mathbf{e}_D, \mathbf{I}_D)$

Figure 1. NMSE of the different estimators of $\mathrm{KL}(P, Q)$ divergence, $P = \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$.

the maximum mean discrepancy (algorithm MMD) (Gretton et al., 2007), which is a well-known measure of dissimilarity based on RKHS embeddings of distributions. We used Gaussian kernels for those embeddings. We swept the width parameter within a sensible range and report the best performance. Finally, we also report clustering results using an affinity matrix based on simple nearest neighbor risks (algorithm NN), obtained via complete cross-validation procedure (Mullin & Sukthankar, 2000) with a training set size parameter of 50%. For the actual clustering step, we use normalized-cut (Shi & Malik, 2000). Prior to the clustering, distance matrices are turned into affinities by using a Gaussian kernel. Its width is automatically selected as the one maximizing the eigengap (since the number of clusters is assumed to be known).

Table 1 shows the results for the clustering task. The best performance is obtained when the distance matrix is obtained using our proposed NN-risk based estimator of the Jeffreys divergence. It is remarkable how the other estimator of Jeffreys divergence results in much worse results. This is likely due to the high dimensionality of the feature vectors, coupled with small sample size. This kind of data is likely to present a manifold structure, so the explicit dependence of the Wang estimator on the data dimensionality hinders its performance, since the actual intrinsic dimensionality is surely much lower than the ambient space dimension.

5.4.2. Speaker clustering

We also simulate a speaker clustering scenario using the UCI Japanese Vowels dataset, comprised of 12-

Table 1. Clustering error for the 4-way music genre recognition task on the Garageband (GB) dataset and the 9-way speaker clustering task on the Japanese Vowels (JV) dataset

	SSD	NN	NN-J	Wang-J	NN Bound-J	MMD
GB	33.25%	39.50%	21.25%	47.75%	46.00%	31.50%
JV	12.07%	8.15%	10.00%	16.30%	7.41%	20.37%

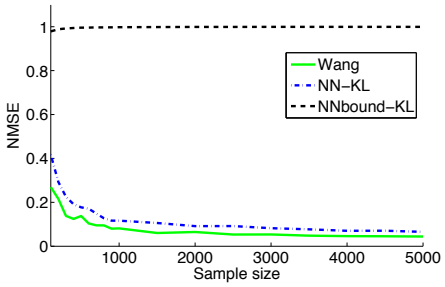


Figure 2. NMSE of the different estimators of $\mathrm{KL}(P, Q)$ divergence, $P = \mathcal{N}(\mathbf{0}, \mathbf{I_3})$, $Q = \mathrm{Unif}[-3, 3]^3$.

dimensional time series of LPC cepstrum coefficients coming from 9 different speakers. There are 30 sequences per speaker, with lengths ranging from 7 to 29 vectors. The results in Table 1 show that this is a much simpler task than the genre clustering one, as most algorithms perform quite well. It is remarkable how the three NN-error based algorithms give the best performance. In this case, the added expressiveness of KL divergence does not compensate the more complex (and noisier) estimation procedure, as can be seen by the Mahalanobis-based bound achieving the highest performance, followed by the simple NN risk.

6. Conclusions

(f, l)-divergences generalize standard f-divergences by surrogating the 0-1 loss by an arbitrary loss l. Many convenient properties are preserved if some simple conditions are imposed on l. (f, l)-divergences can also provide alternative representations of standard f-divergences. We applied this theory and a result linking the error of the nearest-neighbor classifier with the Bayes risk under the square loss to derive a novel order statistics-based estimator for KL and J divergences.

Acknowledgements

We thank the anonymous reviewers for their helpful suggestions and Bob Williamson for comments on the final manuscript. DGG was a member of the TSC Department of Universidad Carlos III when this work was done, and thanks Emilio Parrado for helpful discussions. RSR was funded by the Spanish Ministry of Science and Innovation under TEC2008-01348 grant.

References

Ali, S.M. and Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society*, 28:131–142, 1966.

Bartlett, P., Jordan, M., and McAuliffe, J. Convexity, classification and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

Buja, A., Stuetzle, W., and Shen, Y. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, University of Pennsylvania, 2005.

Devroye, L., Györfi, L., and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

García-García, D., Parrado, E., Arenas, J., and Díaz-de-María, F. Music genre classification using the temporal structure of songs. In *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing*, 2010.

Gretton, A., Borgwardt, K., Rasch, M., Schoelkopf, B., and Smola, A. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19*, 2007.

Mullin, Matthew and Sukthankar, Rahul. Complete cross-validation for nearest neighbor classifiers. In *Proceedings of the International Conference on Machine Learning*, 2000.

Nguyen, X.L., Wainwright, M.J., and Jordan, M.I. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems 20*, 2008.

Nguyen, X.L., Wainwright, M.J., and Jordan, M.I. On surrogate loss functions and f-divergences. *Annals of Statistics*, 37(2):876–904, 2009.

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

Savage, L.J. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Shi, J. and Malik, J. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

Österreicher, F. Csiszar's f-divergences-basic properties. Research report, Institute of Mathematics, University of Salzburg, Austria, 2002.

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

Suzuki, T., Sugiyama, M., and Tanaka, T. Mutual information approximation via maximum likelihood estimation of density ratio. In *Proceedings of the IEEE Symposium on Information Theory*, 2009.

Wang, Q., Kulkarni, S.R., and Verdú, S. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55:2392–2405, 2009.