On Information Divergence Measures, Surrogate Loss Functions and Decentralized Hypothesis Testing

XuanLong Nguyen[*] Martin J. Wainwright[*†] Michael I. Jordan[*†]

Electrical Engineering & Computer Science[*]
UC Berkeley, CA 94720

Department of Statistics[†]
UC Berkeley, CA 94720

Abstract

We establish a general correspondence between two classes of statistical functions: *Ali-Silvey distances* (also known as *f-divergences*) and *surrogate loss functions*. Ali-Silvey distances play an important role in signal processing and information theory, for instance as error exponents in hypothesis testing problems. Surrogate loss functions (e.g., hinge loss, exponential loss) are the basis of recent advances in statistical learning methods for classification (e.g., the support vector machine, AdaBoost). We provide a connection between these two lines of research, showing how to determine the unique f-divergence induced by a given surrogate loss, and characterizing all surrogate loss functions that realize a given f-divergence. The correspondence between f-divergences and surrogate loss functions has applications to the problem of designing quantization rules for decentralized hypothesis testing in the framework of statistical learning (i.e., when the underlying distributions are unknown, but the learner has access to labeled samples).

1 Introduction

The class of *Ali-Silvey distances* or *f-divergences* plays a fundamental role in statistics, signal processing, information theory and related fields [1, 7]. Many of these divergences arise naturally as error exponents in an asymptotic setting. For instance, the Kullback-Leibler divergence specifies the exponential rate of decay of error probability in the Neyman-Pearson setting, and the Chernoff distance appears as the corresponding error exponent in a Bayesian setting [6, 4]. Motivated by such connections, various researchers from the 1960's onwards—studying problems such as signal selection or quantizer design in hypothesis testing—advocated the maximization of various types of f-divergences so as to sidestep the intractable problem of minimizing the probability of error directly [e.g., 9, 12, 17].

A similar set of issues has arisen in recent years in the field of statistical learning theory. Consider, for example, the binary classification problem, in which the learner is given access to samples from two underlying distributions and is asked to find a discriminant function that effectively classifies future samples from these distributions. This problem can be formulated in terms of minimizing the *Bayes risk*—the expectation of the 0-1 loss. This minimization problem is intractable, and the recent literature on statistical learning has focused on the notion of a computationally-tractable *surrogate loss function*—a convex upper bound on the 0-1 loss. Many practical and widely-used algorithms for learning classifiers can be formulated in terms of minimizing empirical averages of such surrogate loss functions. Well-known examples include the support vector machine based on the hinge loss [5], and the AdaBoost algorithm based on exponential loss [8]. A number of researchers [e.g., 2, 14, 19] have investigated the statistical consequences of using such surrogate loss functions— for instance, in characterizing the properties of loss functions required for consistent learning methods (methods that approach the minimum of the Bayes risk as the size of the sample grows).

The main contribution of this paper is to establish a connection between these two lines of research. More specifically, we elucidate a general correspondence between the class of

f-divergences, and the family of surrogate loss functions.[1] Our methods are constructive—we describe how to determine the unique f-divergence associated with any surrogate loss, and conversely we specify how to construct all surrogate losses that realize a given f-divergence. This correspondence has a number of interesting consequences. First, it partitions the set of surrogate loss functions into a set of equivalence classes, defined by the relation of inducing the same f-divergence measure. Second, it allows various well-known inequalities between f-divergences [15] to be leveraged in analyzing surrogate loss functions and learning procedures.

This work was partially motivated by the problem of designing local quantization rules for performing decentralized detection in a sensor network. Our previous work [10] addressed this decentralized detection problem in the learning setting, in which the underlying hypotheses are unknown but the learner has access to labeled samples. We developed a practical algorithm for learning local decision rules at each sensor as well as the fusion center rule, using surrogate loss functions. The correspondence developed here turns out to be useful in analyzing the statistical consistency of our procedures, as we discuss briefly in this paper. More broadly, it is interesting to note that the choice of decentralized decision rules can be viewed as a particular type of problem in experimental design, which motivated the classical research on f-divergences.

2 Background

We begin with necessary background on f-divergences and surrogate loss functions. We discuss the role of f-divergences and surrogate loss functions in experimental design problems, using decentralized detection as an illustrative example.

2.1 Divergence measures

The class of f-divergences or Ali-Silvey distances [1, 7] provides a notion of distance between probability distributions μ and π. Specializing to discrete distributions for simplicity, we have

Definition 1. *Given any continuous convex function $f : [0, +\infty) \to \mathbb{R} \cup \{+\infty\}$, the f-divergence between measures μ and π is given by*

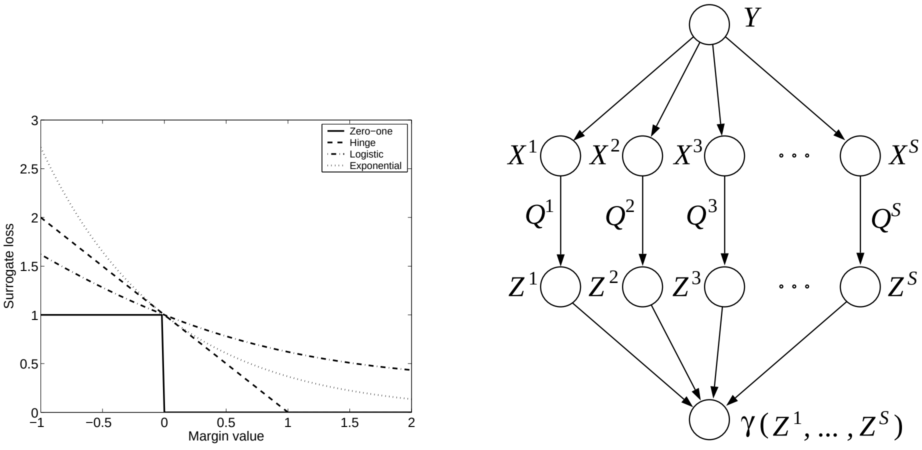
$$I_f(\mu, \pi) := \sum_z \pi(z) f\left(\frac{\mu(z)}{\pi(z)}\right). \tag{1}$$

The Kullback-Leibler divergence is an f-divergence; if we set $f(u) = u \log u$, then definition (1) yields $I_f(\mu, \pi) = \sum_z \mu(z) \log \frac{\mu(z)}{\pi(z)}$. Other examples include the variational distance $I_f(\mu, \pi) := \sum_z |\mu(z) - \pi(z)|$ generated by $f(u) := |u - 1|$, and the (squared) Hellinger distance $I_f(\mu, \pi) := \sum_{z \in \mathcal{Z}} (\sqrt{\mu(z)} - \sqrt{\pi(z)})^2$ generated by $f(u) = \frac{1}{2}(\sqrt{u} - 1)^2$,

2.2 Learning methods based on surrogate loss functions

The standard binary classification problem in statistical learning can be stated formally as follows. Let X be a random vector taking values in some Euclidean space $\mathcal{X}$, and let Y be a random variable taking values in the label space $\mathcal{Y} := \{-1, +1\}$. A standard assumption is that the Cartesian product space $\mathcal{X} \times \mathcal{Y}$ is endowed with some Borel regular probability measure P, but that P is unknown to the learner. The learner is instead given a collection of i.i.d. samples from P of the form $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, and the goal is to learn a decision rule.

[1] Proofs are omitted from this manuscript for lack of space; see [11] for proofs of all of our results.



(a) (b)

Figure 1. (a) Illustration of various convex loss functions that act as surrogates to the 0-1 loss (solid line). Shown are the hinge loss $\phi(t) := \max(0, 1 - t)$, the exponential loss $\phi(t) = \exp(-t)$ and the logistic loss $\phi(t) = \log[1 + \exp(-t)]$. (b) Decentralized detection system with S sensors, in which Y is the unknown hypothesis, $X = (X^1, \ldots, X^S)$ is the vector of sensor observations, and $Z = (Z^1, \ldots, Z^S)$ are the quantized messages transmitted from the sensors to the fusion center.

More formally, we consider measurable functions γ mapping from $\mathcal{X}$ to the real line $\mathbb{R}$; for any $x \in \mathcal{X}$, the associated decision is given by $\text{sign}(\gamma(x))$. The goal is to choose γ so as to minimize the *Bayes risk*—the probability of misclassification $P(Y \neq \gamma(X))$.

Defining the 0-1 loss as the function $\mathbb{I}(t)$ that is equal to 1 if $t \leq 0$ and 0 otherwise (see Figure 1 for an illustration), we note that the Bayes risk corresponds to the expectation of $\mathbb{I}(t)$: $P(Y \neq \gamma(X)) = \mathbb{E}[\mathbb{I}(Y\gamma(X))]$, where $Y\gamma(X)$ is known as the *margin*. This fact motivates the strategy of choosing γ by minimizing the empirical expectation of the 0-1 loss. Given the non-convexity of this minimization problem, it is natural to consider instead the empirical expectation of some convex function ϕ that upper bounds γ—this "convexifies" the problem. In fact, a variety of practical algorithms used for classification, including the support vector machine (SVM) [5] and the AdaBoost algorithm [8], can be understood in this general framework. The SVM corresponds to replacing 0-1 loss with the *hinge loss* $\phi(t) := \max(0, 1 - t)$, whereas the AdaBoost algorithm operates on the exponential loss $\phi(t) := \exp(-t)$. See Figure 1(a) for illustrations of these two convex surrogates, as well as the closely related logistic loss $\phi(t) = \log[1 + \exp(-t)]$.

2.3 Experimental design and decentralized detection

Our focus in this paper is the following extension of the standard binary classification problem. Suppose that the decision-maker, rather than having direct access to X, only observes some variable $Z \in \mathcal{Z}$ that is obtained via a (possibly stochastic) mapping $Q : \mathcal{X} \to \mathcal{Z}$. The mapping Q is referred to as an *experiment* in the statistical literature. We let $\mathcal{Q}$ denote the space of all stochastic experiments Q, and let $\mathcal{Q}_0$ denote its deterministic subset. Given a fixed experiment Q, we can then consider the standard binary classification problem associated with the space $\mathcal{Z}$–namely, to find a measurable function $\gamma \in \Gamma := \{\mathcal{Z} \to \mathbb{R}\}$ that minimizes the Bayes risk $P(Y \neq \text{sign}(\gamma(Z)))$. When the experiment Q is also allowed to vary, we are led to the broader question of determining both the classifier $\gamma \in \Gamma$, as well as the experiment $Q \in \mathcal{Q}$ so as to minimize the Bayes risk.

Decentralized hypothesis testing: An important example of such an experimental design problem is that of decentralized detection. This problem arises in a variety of applications, including human decision making, sensor networks, and distributed databases. Figure 1(b) provides a graphical representation of a binary decentralized detection problem. The system depicted in the figure consists of a set of S sensors that receive observations from the environment. The decentralized nature of the system arises from the fact the each sensor is permitted to relay only a summary message (as opposed to the full observation) back to the central fusion center. The goal is to design a local decision rule Q^i for each sensor $i \in \{1, \ldots, S\}$ so as to to minimize to overall probability of error. Note that the choice of these decision rules can be viewed as a choice of experiments in the statistical sense. There is a considerable literature on such problems when the distributions are known [18, 3, 16]. In contrast, our previous work [10] has focused on the statistical learning setting, in which the decision rules must be designed on the basis of labeled samples.

Approaches to experimental design: We now return to the problem of choosing both the classifier $\gamma \in \Gamma$, as well as the experiment choice $Q \in \mathcal{Q}$ so as to minimize the Bayes risk. Given priors $q = P(Y = -1)$ and $p = P(Y = 1)$, define nonnegative measures μ and π:

$$\begin{aligned} \mu(z) &= P(Y = 1, z) = p \int_x Q(z|x) dP(x|Y = 1) \\ \pi(z) &= P(Y = -1, z) = q \int_x Q(z|x) dP(x|Y = -1). \end{aligned}$$

As a consequence of Lyapunov's theorem, the space of $\{(\mu, \pi)\}$ obtained by varying $Q \in \mathcal{Q}$ (or $\mathcal{Q}_0$) is both compact and convex [cf. 17]. For simplicity, in this paper, we assume that the space $\mathcal{Q}$ is restricted such that both μ and π are strictly positive measures.

One approach to choosing Q is to define an f-divergence between μ and π; indeed this is the classical approach referred to earlier [e.g., 12]. Rather than following this route, however, we take an alternative path, setting up the problem in terms of the expectation of a surrogate loss ϕ, a quantity that we refer to as the "ϕ-risk":

$$R_\phi(\gamma, Q) = \mathbb{E}[\phi(Y\gamma(Z))] = \sum_z \phi(\gamma(z))\mu(z) + \phi(-\gamma(z))\pi(z). \tag{2}$$

This representation of the ϕ-risk in terms of μ and π allows us to compute the optimal value of $\gamma(z)$ for all $z \in \mathcal{Z}$, as well as the optimal ϕ-risk for a fixed Q. Let us define, for each Q, $R_\phi(Q) := \inf_{\gamma \in \Gamma} R_\phi(\gamma, Q)$.

2.3.1 Illustrative examples

We illustrate this calculation using the four loss functions shown in Figure 1.

0-1 loss. If ϕ is 0-1 loss, then $\gamma(z) = \text{sign}(\mu(z) - \pi(z))$. Thus the optimal Bayes risk given a fixed Q takes the form

$$R_{bayes}(Q) = \sum_{z \in \mathcal{Z}} \min\{\mu(z), \pi(z) \;=\; \frac{1}{2} - \frac{1}{2} \sum_{z \in \mathcal{Z}} |\mu(z) - \pi(z)| \;=\; \frac{1}{2}(1 - V(\mu, \pi)),$$

where $V(\mu, \pi) := \sum_{z \in \mathcal{Z}} |\mu(z) - \pi(z)|$ denotes the variational distance between two measures μ and π.

Hinge loss. Consider the hinge loss $\phi_{hinge}(y\gamma(z)) = (1-y\gamma(z))_+$, illustrated in Figure 1. In this case, the optimal decision rule is $\gamma(z) = \text{sign}(\mu(z) - \pi(z))$, so that that optimized ϕ-risk takes the form

$$R_{hinge}(Q) = \sum_{z\in\mathcal{Z}} 2\min\{\mu(z), \pi(z)\} = 1 - \sum_{z\in\mathcal{Z}} |\mu(z) - \pi(z)| = 1 - V(\mu, \pi) = 2R_{bayes}(Q). \tag{3}$$

Logistic loss. For the logistic loss $\phi_{log}(y\gamma(z)) := \log\left(1 + \exp^{-y\gamma(z)}\right)$, the optimal decision rule is given by $\gamma(z) = \log\frac{\mu(z)}{\pi(z)}$. Calculating the optimized risk, we obtain

$$R_{log}(Q) = \sum_{z\in\mathcal{Z}} \mu(z)\log\frac{\mu(z)+\pi(z)}{\mu(z)} + \pi(z)\log\frac{\mu(z)+\pi(z)}{\pi(z)} = \log 2 - KL(\mu||\frac{\mu+\pi}{2}) - KL(\pi||\frac{\mu+\pi}{2}),$$

where $KL(U,V)$ denotes the Kullback-Leibler (KL) divergence. (The quantity $C(U,V) := KL(U||\frac{U+V}{2}) + KL(V||\frac{U+V}{2})$ is known as the *capacitory discrimination* distance).

Exponential loss: Now considering the exponential loss $\phi_{exp}(y\gamma(z)) = \exp(-y\gamma(z))$, we calculate the optimal decision rule $\gamma(z) = \frac{1}{2}\log\frac{\mu(z)}{\pi(z)}$. Thus we have

$$R_{exp}(Q) = \sum_{z\in\mathcal{Z}} 2\sqrt{\mu(z)\pi(z)} = 1 - \sum_{z\in\mathcal{Z}}(\sqrt{\mu(z)} - \sqrt{\pi(z)})^2 = 1 - 2h^2(\mu, \pi),$$

where $h(\mu, \pi)$ denotes the Hellinger distance between measures μ and π.

Observe that all of the distances given above are particular instances of f-divergences. This fact hints at a deeper connection between optimized ϕ-risks and f-divergences that we elucidate in the following section.

3 Correspondence between divergence and surrogate loss

In this section, we develop the correspondence between f-divergences and loss functions. We begin by providing a more precise definition of the notion of a surrogate loss function.

3.1 Properties of surrogate loss functions

First, we require that any *surrogate loss function* ϕ is continuous and convex. Second, the function ϕ must be *classification-calibrated* [2], meaning that for any $a, b \geq 0$ and $a \neq b$, $\inf_{\alpha:\alpha(a-b)<0} \phi(\alpha)a + \phi(-\alpha)b > \inf_{\alpha\in\mathbb{R}} \phi(\alpha)a + \phi(-\alpha)b$. To gain intuition for these requirements, recall the representation of the ϕ-risk given in equation (2): it implies that given a fixed Q, the optimal $\gamma(z)$ takes a value α that minimizes $\phi(\alpha)\mu(z) + \phi(-\alpha)\pi(z)$. In order for the decision rule γ to behave equivalently to the Bayes decision rule, we require that the optimal value of α (which defines $\gamma(z)$) should have the same sign as the Bayes decision rule $\text{sign}(P(Y=1|z) - P(Y=-1|z)) = \text{sign}(\mu(z) - \pi(z))$.

It can be shown [2] that in the convex case ϕ is classification-calibrated if and only if it is differentiable at 0 and $\phi'(0) < 0$. Lastly, let $\alpha^* = \inf_\alpha\{\phi(\alpha) = \inf\phi\}$. If $\alpha^* < +\infty$, then for any $\delta > 0$, we require that

$$\phi(\alpha^* - \delta) \geq \phi(\alpha^* + \delta) \tag{4}$$

The interpretation of condition (4) is that one should penalize deviations away from α^* in the negative direction at least as strongly as deviations in the positive direction; this requirement is intuitively reasonable given the margin-based interpretation of α.

3.2 From ϕ-risk to f-divergence

We begin with a simple result that formalizes how any ϕ-risk induces a corresponding f-divergence. More precisely, the following lemma proves that the optimal ϕ-risk for a fixed Q can be written as the negative of an f-divergence between μ and π.

Lemma 2. *For each fixed Q, let γ_Q denote the optimal decision rule. The ϕ-risk for (Q, γ_Q) is an f*-divergence *between μ and π for some convex function f:*

$$R_\phi(Q) = -I_f(\mu, \pi). \tag{5}$$

Proof. The optimal ϕ-risk takes the form:

$$R_\phi(Q) = \sum_{z \in \mathcal{Z}} \inf_\alpha (\phi(\alpha)\mu(z) + \phi(-\alpha)\pi(z)) = \sum_z \pi(z) \inf_\alpha \left(\phi(-\alpha) + \phi(\alpha)\frac{\mu(z)}{\pi(z)} \right).$$

For each z let $u = \frac{\mu(z)}{\pi(z)}$, then $\inf_\alpha(\phi(-\alpha) + \phi(\alpha)u)$ is a concave function of u (since minimization over a set of linear function is a concave function). Thus, the claim follows by defining

$$f(u) := -\inf_\alpha(\phi(-\alpha) + \phi(\alpha)u). \tag{6}$$

3.3 From f-divergence to ϕ-risk

In the remainder of this section, we explore the converse of Lemma 2. Given a divergence $I_f(\mu, \pi)$ for some convex function f, does there exists a function ϕ for which $R_\phi(Q) = -I_f(\mu, \pi)$? We provide a precise characterization of the set of f-divergences that can be realized in this way, as well as a constructive procedure for determining all ϕ that realize a given f-divergence.

Our method requires the introduction of several intermediate functions. First, let us define, for each β, the inverse mapping $\phi^{-1}(\beta) := \inf\{\alpha : \phi(\alpha) \leq \beta\}$, where $\inf \emptyset := +\infty$. Using the function ϕ^{-1}, we then define a new function $\Psi : \mathbb{R} \to \overline{\mathbb{R}}$ by

$$\Psi(\beta) \quad := \quad \begin{cases} \phi(-\phi^{-1}(\beta)) & \text{if } \phi^{-1}(\beta) \in \mathbb{R}, \\ +\infty & \text{otherwise.} \end{cases} \tag{7}$$

Note that the domain of Ψ is $\text{Dom}(\Psi) = \{\beta \in \mathbb{R} : \phi^{-1}(\beta) \in \mathbb{R}\}$. Define

$$\beta_1 := \inf\{\beta : \Psi(\beta) < +\infty\} \text{ and } \beta_2 := \inf\{\beta : \Psi(\beta) = \inf \Psi\}. \tag{8}$$

It is simple to check that $\inf \phi = \inf \Psi = \phi(\alpha^*)$, and $\beta_1 = \phi(\alpha^*)$, $\beta_2 = \phi(-\alpha^*)$. Furthermore, $\Psi(\beta_2) = \phi(\alpha^*) = \beta_1$, $\Psi(\beta_1) = \phi(-\alpha^*) = \beta_2$. With this set-up, the following lemma captures several important properties of Ψ:

Lemma 3. *(a) Ψ is strictly decreasing in (β_1, β_2). If ϕ is decreasing, then Ψ is also decreasing in $(-\infty, +\infty)$. In addition, $\Psi(\beta) = +\infty$ for $\beta < \beta_1$.*

(b) Ψ is convex in $(-\infty, \beta_2]$. If ϕ is decreasing, then Ψ is convex in $(-\infty, +\infty)$.

(c) Ψ is lower semi-continuous, and continuous in its domain.

(d) There exists $u^* \in (\beta_1, \beta_2)$ such that $\Psi(u^*) = u^*$.

(e) There holds $\Psi(\Psi(\beta)) = \beta$ for all $\beta \in (\beta_1, \beta_2)$.

The connection between Ψ and an f-divergence arises from the following fact. Given the definition (7) of Ψ, it is possible to show that

$$f(u) = \sup_{\beta \in \mathbb{R}} (-\beta u - \Psi(\beta)) = \Psi^*(-u), \tag{9}$$

where Ψ^* denotes the conjugate dual of the function Ψ. Hence, if Ψ is a lower semicontinuous convex function, it is possible to recover Ψ from f by means of convex duality [13]: $\Psi(\beta) = f^*(-\beta)$. Thus, equation (7) provides means for recovering a loss function ϕ from Ψ. Indeed, the following theorem provides a constructive procedure for finding all such ϕ when Ψ satisfies necessary conditions specified in Lemma 3:

Theorem 4. *(a) Given a lower semicontinuous convex function $f : \mathbb{R} \to \overline{\mathbb{R}}$, define:*

$$\Psi(\beta) = f^*(-\beta). \tag{10}$$

If Ψ is a decreasing function satisfying the properties specified in parts (c), (d) and (e) of Lemma 3, then there exist convex continuous loss functions ϕ for which (5) *and* (6) *hold.*
(b) More precisely, all such functions ϕ are of the form: For any $\alpha \geq 0$,

$$\phi(\alpha) = \Psi(g(\alpha + u^*)), \quad \textit{and} \quad \phi(-\alpha) = g(\alpha + u^*), \tag{11}$$

where $u^* \in (\beta_1, \beta_2)$ satisfies $\Psi(u^*) = u^*$ and $g : [u^*, +\infty) \to \overline{\mathbb{R}}$ is any increasing continuous convex function such that $g(u^*) = u^*$. Moreover, g is differentiable at u^*+ and $g'(u^*+) > 0$.

One interesting consequence of Theorem 4 that any realizable f-divergence can be obtained from a fairly large set of ϕ loss functions. Indeed, Theorem 4(b) reveals that for $\alpha \leq 0$, we are free to choose a function g that must satisfy only mild conditions; given a choice of g, then ϕ is specified for $\alpha > 0$ by equation (11). We describe below how the Hellinger distance, for instance, is realized not only by the exponential loss (as described earlier), but also by many other surrogate loss functions.

3.4 Illustrative examples

We provide a few examples to illustrate Theorem 4; see [11] for additional examples.

Hellinger distance: As a first example, consider Hellinger distance, which is an f-divergence[2] with $f(u) = -2\sqrt{u}$. Augment the domain of f with $f(u) = +\infty$ for $u < 0$. Following the prescription of Theorem 4(a), we first recover Ψ from f:

$$\Psi(\beta) = f^*(-\beta) = \sup_{u \in \mathbb{R}} (-\beta u - f(u)) = \begin{cases} 1/\beta & \text{when } \beta > 0 \\ +\infty & \text{otherwise.} \end{cases}$$

Clearly, $u^* = 1$. Now if we choose $g(u) = e^{u-1}$, then we obtain the exponential loss $\phi(\alpha) = \exp(-\alpha)$. However, making the alternative choice $g(u) = u$, we obtain the function $\phi(\alpha) = 1/(\alpha + 1)$ and $\phi(-\alpha) = \alpha + 1$, which also realizes the Hellinger distance.

[2] We consider f-divergences for two convex functions f_1 and f_2 to be equivalent if f_1 and f_2 are related by a linear term, i.e., $f_1 = cf_2 + au + b$ where $c > 0$, because then I_{f_1} and I_{f_2} differ by a constant.

0-1 loss: Recall that we have shown previously that the 0-1 loss induces the variational distance, which can be expressed as an f-divergence with $f_{\text{var}}(u) = -2\min(u, 1)$ for $u \geq 0$. It is thus of particular interest to determine other loss functions that also lead to variational distance. If we augment the function f_{var} by defining $f_{\text{var}}(u) = +\infty$ for $u < 0$, then we can recover Ψ from f_{var} as follows:

$$\Psi(\beta) = f^*_{\text{var}}(-\beta) = \sup_{u \in \mathbb{R}}(-\beta u - f_{\text{var}}(u)) = \begin{cases} (2-\beta)_+ & \text{when } \beta \geq 0 \\ +\infty & \text{when } \beta < 0. \end{cases}$$

Clearly $u^* = 1$. Choosing $g(u) = u$ leads to the hinge loss $\phi(\alpha) = (1-\alpha)_+$, consistent with our earlier findings. Making the alternative choice $g(u) = e^{u-1}$ leads to a rather different loss—namely, $\phi(\alpha) = (2 - e^{\alpha})_+$ for $\alpha \geq 0$ and $\phi(\alpha) = e^{-\alpha}$ for $\alpha < 0$—that also realizes the variational distance.

Remark: It is worth noting that not all f-divergences can be realized by a (margin-based) surrogate loss. The list of non-realizable f-divergences includes the KL divergence $KL(\mu||\pi)$ (as well as $KL(\pi||\mu)$). Interestingly, however, the *symmetric* KL divergence $KL(\mu||\pi) + KL(\pi||\mu)$ is a realizable f-divergence. One of the corresponding ϕ losses constructed via Theorem 4 turns out to have the simple closed-form expression $\phi(\alpha) = e^{-\alpha} - \alpha$; see [11].

4 Equivalence of loss functions and decentralized detection

The previous section was devoted to study of the correspondence between f-divergences and the optimal ϕ-risk $R_\phi(Q)$ for a fixed experiment Q. Recall that our ultimate goal is that of solving the experimental design problem of choosing an optimal Q. As discussed previously, the function Q might correspond to a local decision rule in a sensor network [17, 10].

Our approach to this problem is the natural one of jointly optimizing the ϕ-risk (or more precisely, its empirical version) over both the decision γ and the choice of experiment Q (hereafter referred to as a quantizer). This procedure raises the natural theoretical question: for what loss functions ϕ does such joint optimization lead to minimum Bayes risk? Note that the minimum here is taken over both the decision rule γ and the space of experiments Q, so that this question is not covered by standard consistency results [19, 14, 2]. To this end, we shall consider the comparison of loss functions and the comparison of quantization schemes. Then we describe how the results developed herein can be leveraged to resolve the issue of consistency of learning optimal quantizer design from empirical data.

Universal equivalence of loss functions: We begin by introducing a notion of equivalence between arbitrary loss functions ϕ_1 and ϕ_2, or alternatively between the corresponding divergences induced by f_1 and f_2.

Definition 5. *The surrogate loss functions ϕ_1 and ϕ_2 are* universally equivalent, *denoted by $\phi_1 \overset{u}{\approx} \phi_2$ (and $f_1 \overset{u}{\approx} f_2$), if for any $P(X,Y)$ and quantization rules Q_1, Q_2, there holds:*

$$R_{\phi_1}(Q_1) \leq R_{\phi_1}(Q_2) \Leftrightarrow R_{\phi_2}(Q_1) \leq R_{\phi_2}(Q_2). \tag{12}$$

The following result provides necessary and sufficient conditions for universal equivalence:

Theorem 6. *Suppose that f_1 and f_2 are differentiable a.e., convex functions that map $[0, +\infty)$ to $\mathbb{R}$. Then $f_1 \overset{u}{\approx} f_2$ if and only if $f_1(u) = cf_2(u) + au + b$ for constants $a, b \in \mathbb{R}$ and $c > 0$.*

If we restrict our attention to convex and differentiable a.e. functions f, then it follows that all f-divergences universally equivalent to the variational distance must have the form

$$f(u) = -c\min(u,1) + au + b \qquad \text{with } c > 0. \tag{13}$$

As a consequence, the only ϕ-loss functions universally equivalent to 0-1 loss are those that induce an f-divergence of this form. One well-known example of such a function is the hinge loss; more generally, Theorem 4 allows us to construct all such ϕ.

Consistency in experimental design: The notion of universal equivalence might appear quite restrictive because condition (12) must hold for *any* underlying probability measure $P(X,Y)$. However, this is precisely what we need when $P(X,Y)$ is unknown. Assume that the knowledge about $P(X,Y)$ comes from an empirical data sample $(x_i, y_i)_{i=1}^n$.

Consider any algorithm (such as that proposed by Nguyen et al. [10]) that involves choosing a classifier-quantizer pair $(\gamma, Q) \in \Gamma \times \mathcal{Q}$ by minimizing an empirical version of ϕ-risk:

$$\hat{R}_\phi(\gamma, Q) := \frac{1}{n}\sum_{i=1}^{n}\sum_{z}\phi(y_i\gamma(z))Q(z|x_i).$$

More formally, suppose that $(\mathcal{C}_n, \mathcal{D}_n)$ is a sequence of increasing compact function classes such that $\mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq \ldots \subseteq \Gamma$ and $\mathcal{D}_1 \subseteq \mathcal{D}_2 \subseteq \ldots \subseteq \mathcal{Q}$. Let (γ_n^*, Q_n^*) be an optimal solution to the minimization problem $\min_{(\gamma,Q)\in(\mathcal{C}_n,\mathcal{D}_n)} \hat{R}_\phi(\gamma, Q)$, and let R_{bayes}^* denote the minimum Bayes risk achieved over the space of decision rules $(\gamma, Q) \in (\Gamma, \mathcal{Q})$. We call $R_{bayes}(\gamma_n^*, Q_n^*) - R_{bayes}^*$ the *Bayes error* of our estimation procedure. We say that such a procedure is *universally consistent* if the Bayes error tends to zero as $n \to \infty$, i.e., for any (unknown) Borel probability measure P on $X \times Y$,

$$\lim_{n\to\infty} R_{bayes}(\gamma_n^*, Q_n^*) - R_{bayes}^* = 0 \quad \text{in probability.}$$

When the surrogate loss ϕ is universally equivalent to 0-1 loss, we can prove that suitable learning procedures are indeed universally consistent. At a high level, our approach leverages the framework developed by various authors [19, 14, 2] for the case of ordinary classification: in particular, we exploit the strategy of decomposing the Bayes error into a combination of

(a) *approximation error* introduced by the bias of the function classes $\mathcal{C}_n \subseteq \Gamma$: $\mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) = \inf_{(\gamma,Q)\in(\mathcal{C}_n,\mathcal{D}_n)} R_\phi(\gamma, Q) - R_\phi^*$, where $R_\phi^* := \inf_{(\gamma,Q)\in(\Gamma,\mathcal{Q})} R_\phi(\gamma, Q)$.

(b) *estimation error* introduced by the variance of using finite sample size n, $\mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) = \mathbb{E}\sup_{(\gamma,Q)\in(\mathcal{C}_n,\mathcal{D}_n)} |\hat{R}_\phi(\gamma, Q) - R_\phi(\gamma, Q)|$, where the expectation is taken with respect to the (unknown) probability measure $P(X,Y)$.

Complete details can be found in the technical report [11].

5 Conclusions

We have presented a general theoretical connection between surrogate loss functions and f-divergences. As illustrated by our application to decentralized detection, this connection can provide new domains of application for statistical learning theory. We also expect that this connection will provide new applications for f-divergences within learning theory; note in particular that bounds among f-divergences (of which many are known; see, e.g., [15]) induce corresponding bounds among loss functions.

References

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

[2] P. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification and risk bounds. *Journal of the American Statistical Association*, In press, 2005.

[3] J. F. Chamberland and V. V. Veeravalli. Decentralized detection in sensor networks. *IEEE Trans. on Signal Processing*, 51(2):407–416, 2003.

[4] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations. *Annals of Statistics*, 23:493–507, 1952.

[5] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[6] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

[7] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungarica*, 2:299–318, 1967.

[8] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[9] T. Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. on Communication Technology*, 15(1):52–60, 1967.

[10] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Nonparametric decentralized detection using kernel methods. *IEEE Trans. on Signal Processing*, In press, 2004.

[11] X. Nguyen, M. J. Wainwright, and M. I. Jordan. On information divergence measures, surrogate loss functions and decentralized hypothesis testing. Technical report, University of California, Berkeley, 2005. [`www.cs.berkeley.edu/∼xuanlong/decon.pdf`].

[12] H. V. Poor and J. B. Thomas. Applications of Ali-Silvey distance measures in the design of generalized quantizers for binary decision systems. *IEEE Trans. on Comm.*, 25:893–900, 1977.

[13] G. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[14] I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *IEEE Trans. on Information Theory*, 51:128–142, 2005.

[15] F. Topsœ. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. on Information Theory*, 46:1602–1609, 2000.

[16] J. N. Tsitsiklis. Decentralized detection. In *Advances in Statistical Signal Processing*, pages 297–344. JAI Press, 1993.

[17] J. N. Tsitsiklis. Extremal properties of likelihood-ratio quantizers. *IEEE Trans. on Communication*, 41(4):550–558, 1993.

[18] V. V. Veeravalli, T. Basar, and H. V. Poor. Decentralized sequential detection with a fusion center performing the sequential test. *IEEE Trans. on Information Theory*, 39(2):433–442, 1993.

[19] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annal of Statistics*, 53:56–134, 2004.