On divergences, surrogate loss functions, and decentralized detection

XuanLong Nguyen
Computer Science Division
University of California, Berkeley
xuanlong@eecs.berkeley.edu

Martin J. Wainwright
Statistics Department and EECS Department
University of California, Berkeley
wainwrig@stat.berkeley.edu

Michael I. Jordan
Statistics Department and EECS Department
University of California, Berkeley
jordan@stat.berkeley.edu

October 25, 2005

Technical Report No. 695
Department of Statistics
University of California, Berkeley

Abstract

We develop a general correspondence between a family of loss functions that act as surrogates to 0-1 loss, and the class of Ali-Silvey or f-divergence functionals. This correspondence provides the basis for choosing and evaluating various surrogate losses frequently used in statistical learning (e.g., hinge loss, exponential loss, logistic loss); conversely, it provides a decision-theoretic framework for the choice of divergences in signal processing and quantization theory. We exploit this correspondence to characterize the statistical behavior of a nonparametric decentralized hypothesis testing algorithms that operate by minimizing convex surrogate loss functions. In particular, we specify the family of loss functions that are equivalent to 0-1 loss in the sense of producing the same quantization rules and discriminant functions.

1 Introduction

Over the past several decades, the classical topic of discriminant analysis has undergone significant and sustained development in various scientific and engineering fields. Much of this development has been driven by the physical, informational and computational constraints imposed by specific problem domains. Incorporating such constraints leads to interesting extensions of the basic discriminant analysis paradigm that involve aspects of experimental design. As one example, research in the area of "decentralized detection" focuses on problems in which measurements are collected by a collection of devices distributed over space (e.g., arrays of cameras, acoustic sensors, wireless nodes). Due to power and bandwidth limitations, these devices cannot simply relay their measurements to the common site where a hypothesis test is to be performed; rather, the measurements must be compressed prior to transmission, and the statistical test at the central site is performed on the transformed data (Tsitsiklis, 1993b, Blum et al., 1997). The

problem of designing such compression rules is of substantial current interest in the field of sensor networks (Chong and Kumar, 2003, Chamberland and Veeravalli, 2003). A closely related set of "signal selection" problems, arising for instance in radar array processing, also blend discriminant analysis with aspects of experimental design (Kailath, 1967).

The standard formulation of these problems—namely, as hypothesis-testing within either a Neyman-Pearson or Bayesian framework—rarely leads to computationally tractable algorithms. The main source of difficulty is the intractability of minimizing the probability of error, whether as a functional of the discriminant function or of the compression rule. Consequently, it is natural to consider loss functions that act as surrogates for the probability of error, and lead to practical algorithms. For example, the Hellinger distance has been championed for decentralized detection problems (Longo et al., 1990), due to the fact that it yields a tractable algorithm both for the experimental design aspect of the problem (i.e., the choice of compression rule) and the discriminant analysis aspect of the problem. More broadly, a class of functions known as *Ali-Silvey distances* or *f-divergences* (Ali and Silvey, 1966, Csiszaŕ, 1967)—which includes not only the Hellinger distance, but also the variational distance, Kullback-Leibler (KL) divergence and Chernoff distance—have been explored as surrogate loss functions for the probability of error in a wide variety of applied discrimination problems.

Theoretical support for the use of f-divergences in discrimination problems comes from two main sources. First, a classical result of Blackwell (1951) asserts that if procedure A has a smaller f-divergence than procedure B (for some particular f-divergence), then there exists some set of prior probabilities such that procedure A has a smaller probability of error than procedure B. This fact, though a relatively weak justification, has nonetheless proven useful in designing signal selection and quantization rules (Kailath, 1967, Poor and Thomas, 1977, Longo et al., 1990). Second, f-divergences often arise as exponents in asymptotic (large-deviation) characterizations of the optimal rate of convergence in hypothesis-testing problems; examples include Kullback-Leibler divergence for the Neyman-Pearson formulation, and the Chernoff distance for the Bayesian formulation (Cover and Thomas, 1991).

A parallel and more recent line of research in the field of statistical machine learning has also focused on computationally-motivated surrogate functions in discriminant analysis. In statistical machine learning, the formulation of the discrimination problem (also known as *classification*) is decision-theoretic, with the Bayes error interpreted as risk under a 0-1 loss. The algorithmic goal is to design discriminant functions by minimizing the empirical expectation of 0-1 loss, wherein empirical process theory provides the underlying analytic framework. In this setting, the non-convexity of the 0-1 loss renders intractable a direct minimization of probability of error, so that various researchers have studied algorithms based on replacing the 0-1 loss with "surrogate loss functions." These alternative loss functions are convex, and represent upper bounds or approximations to the 0-1 loss (see Figure 2 for an illustration). A wide variety of practically successful machine learning algorithms are based on such a strategy, including support vector machines (Cortes and Vapnik, 1995, Schölkopf and Smola, 2002), the AdaBoost algorithm (Freund and Schapire, 1997), the X4 method (Breiman, 1998), and logistic regression Friedman et al. (2000). More recent work by Bartlett et al. (2005), Steinwart (2005), Zhang (2004) and others provides theoretical support for these algorithms, in particular by characterizing statistical consistency and convergence rates of the resulting estimation procedures in terms of the properties of surrogate loss functions.

1.1 Our contributions

As mathematical objects, the f-divergences studied in information theory and the surrogate loss functions studied in statistical machine learning are fundamentally different: the former are functions on pairs of measures, whereas the latter are functions on values of discriminant functions and class labels. However, their

underlying role in obtaining computationally-tractable algorithms for discriminant analysis suggests that they should be related. Indeed, Blackwell's result hints at such a relationship, but its focus on 0-1 loss does not lend itself to developing a general relationship between f-divergences and surrogate loss functions. The primary contribution of this paper is to provide a detailed analysis of the relationship between f-divergences and surrogate loss functions, developing a full characterization of the connection, and explicating its consequences. We show that for any surrogate loss, regardless of its convexity, there exists a corresponding convex f such that minimizing the expected loss is equivalent to maximizing the f-divergence. We also provide necessary and sufficient conditions for an f-divergence to be realized from some (decreasing) convex loss function. More precisely, given a convex f, we provide a constructive procedure to generate *all* decreasing convex loss functions for which the correspondence holds.

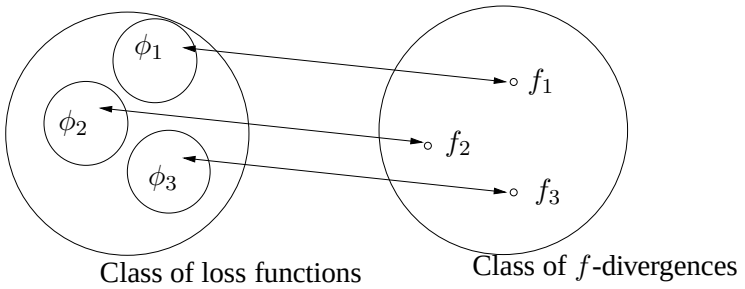

Figure 1. Illustration of the correspondence between f-divergences and loss functions. For each loss function ϕ, there exists exactly one corresponding f-divergence (induced by some underlying convex function f) such that the ϕ-risk is equal to the negative f-divergence. Conversely, for each f-divergence, there exists a whole set of surrogate loss functions ϕ for which the correspondence holds. Within the class of convex loss functions and the class of f-divergences, one can construct equivalent loss functions and equivalent f-divergences, respectively. For the class of classification-calibrated decreasing convex loss functions, we can characterize the correspondence precisely.

The relationship is illustrated in Figure 1; whereas each surrogate loss ϕ induces only one f-divergence, note that in general there are many surrogate loss functions that correspond to the same f-divergence. As particular examples of the general correspondence established in this paper, we show that the hinge loss corresponds to the variational distance, the exponential loss corresponds to the Hellinger distance, and the logistic loss corresponds to the capacitory discrimination distance.

This correspondence—in addition to its intrinsic interest as an extension of Blackwell's work—has several specific consequences. First, there are numerous useful inequalities relating the various f-divergences (Topsoe, 2000); our theorem allows these inequalities to be exploited in the analysis of loss functions. Second, the minimizer of the Bayes error and the maximizer of f-divergences are both known to possess certain extremal properties (Tsitsiklis, 1993a); our result provides a natural connection between these properties. Third, our theorem allows a notion of equivalence to be defined among loss functions: in particular, we say that loss functions are equivalent if they induce the same f-divergence. We then exploit the constructive nature of our theorem to exhibit all possible convex loss functions that are equivalent (in the sense just defined) to the 0-1 loss. Finally, we illustrate the application of this correspondence to the problem of decentralized detection. Whereas the more classical approach to this problem is based on f-divergences (Kailath, 1967, Poor and Thomas, 1977), our method instead builds on the framework of statistical machine learning. The correspondence allows us to establish consistency results for a novel algorithmic framework for decentralized detection: in particular, we prove that for any surrogate loss function equivalent to 0-1 loss,

our estimation procedure is consistent in the strong sense that it will asymptotically choose Bayes-optimal quantization rules.

The remainder of the paper is organized as follows. In Section 2, we define a version of discriminant analysis that is suitably general so as to include problems that involve a component of experiment design (such as in decentralized detection, and signal selection). We also provide a formal definition of surrogate loss functions, and present examples of optimized risks based on these loss functions. In Section 3, we state and prove the correspondence theorem between surrogate loss functions and f-divergences. Section 4 illustrates the correspondence using well-known examples of loss functions and their f-divergence counterparts. In Section 5, we discuss connections between the choice of quantization designs and Blackwell's classic results on comparisons of experiments. We introduce notions of equivalence among surrogate loss functions, and explore their properties. In Section 6, we establish the consistency of schemes for choosing Bayes-optimal classifiers based on surrogate loss functions that are equivalent to 0-1 loss. We conclude with a discussion in Section 7.

2 Background and problem set-up

2.1 Binary classification and its extension

We begin by defining a classical discriminant analysis problem, in particular the *binary classification problem*. Let X be a covariate taking values in a compact topological space $\mathcal{X}$, and let $Y \in \mathcal{Y} := \{-1, +1\}$ be a binary random variable. The product space $(X \times Y)$ is assumed to be endowed with a Borel regular probability measure $\mathbb{P}$. A *discriminant function* is a measurable function f mapping from $\mathcal{X}$ to the real line, whose sign is used to make a classification decision. The goal is to choose the discriminant function f so as to minimize the probability of making the incorrect classification, also known as the *Bayes risk*. This risk is defined as follows

$$\mathbb{P}(Y \neq \operatorname{sign}(f(X))) = \mathbb{E}\big[\mathbb{I}[Y \neq \operatorname{sign}(f(X))\big], \tag{1}$$

where $\mathbb{I}$ is a 0-1-valued indicator function.

The focus of this paper is an elaboration of this basic problem in which the decision-maker, rather than having direct access to X, observes a random variable Z with range $\mathcal{Z}$ that is obtained via a (possibly stochastic) mapping $Q : \mathcal{X} \to \mathcal{Z}$. In a statistical context, the choice of the mapping Q can viewed as choosing a particular *experiment*; in the signal processing literature, where $\mathcal{Z}$ is generally taken to be discrete, the mapping Q is often referred to as a *quantizer*. In any case, the mapping Q can be represented by conditional probabilities $Q(z|x)$.

Let $\mathcal{Q}$ denote the space of all stochastic Q, and let $\mathcal{Q}_0$ denote the subset of deterministic mappings. When the underlying experiment Q is fixed, then we simply have a binary classification problem on the space $\mathcal{Z}$: that is, our goal is to find a real-valued measurable function γ on $\mathcal{Z}$ so as to minimize the Bayes risk $\mathbb{P}(Y \neq \operatorname{sign}(\gamma(Z)))$. We use Γ to represent the space of all such possible discriminant functions on $\mathcal{Z}$. This paper is motivated by the problem of specifying the classifier $\gamma \in \Gamma$, as well as the experiment choice $Q \in \mathcal{Q}$, so as to minimize the Bayes risk.

Throughout the paper, we assume that $\mathcal{Z}$ is a discrete space for simplicity. We note in passing that this requirement is not essential to our analysis. It is only needed in Section 6, where we require that $\mathcal{Q}$ and $\mathcal{Q}_0$ be compact. This condition is satisfied when Z is discrete.

2.2 Surrogate loss functions

As shown in equation (1), the Bayes risk corresponds to the expectation of the 0-1 loss

$$\phi(y, \gamma(z)) = \mathbb{I}[y \neq \text{sign}(\gamma(z))]. \tag{2}$$

Given the nonconvexity of this loss function, it is natural to consider a surrogate loss function ϕ that we optimize in place of the 0-1 loss. In particular, we focus on loss functions of the form $\phi(y, \gamma(z)) = \phi(y\gamma(z))$, where $\phi : \mathbb{R} \to \mathbb{R}$ is typically a convex upper bound on the 0-1 loss. In the statistical learning literature, the quantity $y\gamma(z)$ is known as the *margin* and $\phi(y\gamma(z))$ is often referred to as a "margin-based loss function." Given a particular loss function ϕ, we denote the associated *ϕ-risk* by $R_\phi(\gamma, Q) := \mathbb{E}\phi(Y\gamma(Z))$.

A number of such loss functions are used commonly in the statistical learning literature. See Figure 2 for an illustration of some different surrogate functions, as well as the original 0-1 loss. First, the *hinge loss* function

$$\phi_{hinge}(y\gamma(z)) \quad := \quad \max\{1 - y\gamma(z), 0\} \tag{3}$$

underlies the so-called support vector machine (SVM) algorithm (Schölkopf and Smola, 2002). Second, the *logistic loss* function

$$\phi_{log}(y\gamma(z)) \quad := \quad \log\left(1 + \exp^{-y\gamma(z)}\right) \tag{4}$$

forms the basis of logistic regression (Friedman et al., 2000). As a third example, the Adaboost algo-

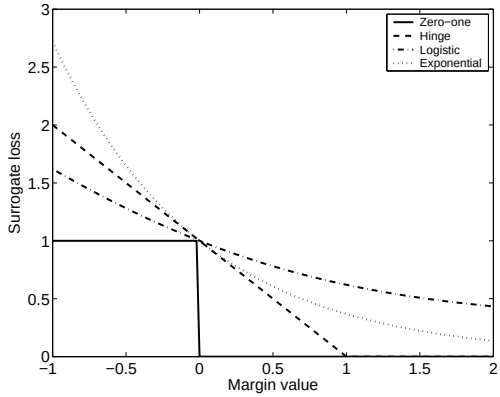

Figure 2. Illustrations of the 0-1 loss function, and three surrogate loss functions: hinge loss, logistic loss, and exponential loss.

rithm (Freund and Schapire, 1997) uses a *exponential loss* function:

$$\phi_{exp}(y\gamma(z)) \quad := \quad \exp(-y\gamma(z)). \tag{5}$$

Finally, another possibility (though less natural for a classification problem) is the *least squares* function:

$$\phi_{sqr}(y\gamma(z)) := (1 - y\gamma(z))^2. \tag{6}$$

Bartlett et al. (2005) have provided a general definition of surrogate loss functions. Their definition is crafted so as to permit the derivation of a general bound that links the ϕ-risk and the Bayes risk, thereby permitting an elegant general treatment of the consistency of estimation procedures based on surrogate losses. The definition is essentially a pointwise form of a Fisher consistency condition that is appropriate for the classification setting; in particular, it takes the following form:

Definition 1. *A loss function ϕ is* classification-calibrated *if for any $a, b \geq 0$ and $a \neq b$:*

$$\inf_{\{\alpha \in \mathbb{R} \,|\, \alpha(a-b)<0\}} \big[\phi(\alpha)a + \phi(-\alpha)b\big] \quad > \quad \inf_{\alpha \in \mathbb{R}} \big[\phi(\alpha)a + \phi(-\alpha)b\big]. \tag{7}$$

As will be clarified subsequently, this definition ensures that the decision rule γ behaves equivalently to the Bayes decision rule in the (binary) classification setting.

For our purposes we will find it useful to consider a somewhat more restricted definition of surrogate loss functions. In particular, we impose the following three conditions on any surrogate loss function $\phi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$:

A1: ϕ is classification-calibrated.

A2: ϕ is continuous and convex.

A3: Let $\alpha^* = \inf\big\{\alpha \in \mathbb{R} \cup \{+\infty\} \;\big|\; \phi(\alpha) = \inf \phi\big\}$. If $\alpha^* < +\infty$, then for any $\epsilon > 0$,

$$\phi(\alpha^* - \epsilon) \geq \phi(\alpha^* + \epsilon). \tag{8}$$

The interpretation of Assumption A3 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. Moreover, this assumption is satisfied by all of the loss functions commonly considered in the literature; in particular, any decreasing function ϕ (e.g., hinge loss, logistic loss, exponential loss) satisfies this condition, as does the least squares loss (which is not decreasing).

Bartlett et al. (2005) also derived a simple lemma that characterizes classification-calibration for convex functions:

Lemma 2. *Let ϕ be a convex function. Then ϕ is classification-calibrated if and only if it is differentiable at 0 and $\phi'(0) < 0$.*

Consequently, Assumption A1 is equivalent to requiring that ϕ be differentiable at 0 and $\phi'(0) < 0$. These facts also imply that the quantity α^* defined in Assumption A3 is strictly positive. Finally, although ϕ is not defined for $-\infty$, we shall use the convention that $\phi(-\infty) = +\infty$.

2.3 Examples of optimum ϕ-risks

For each fixed experiment Q, we define the *optimal ϕ-risk* (a function of Q) as follows:

$$R_\phi(Q) := \inf_{\gamma \in \Gamma} R_\phi(\gamma, Q). \tag{9}$$

Let $p = \mathbb{P}(Y = 1)$ and $q = \mathbb{P}(Y = -1)$, where $p, q > 0$ and $p + q = 1$, define a prior on the hypothesis space. Any fixed experiment Q induces positive measures μ and π over $\mathcal{Z}$ as follows:

$$\mu(z) \quad := \quad \mathbb{P}(Y = 1, Z = z) = p \int_x Q(z|x) d\mathbb{P}(x|Y = 1) \tag{10a}$$

$$\pi(z) \quad := \quad \mathbb{P}(Y = -1, Z = z) = q \int_x Q(z|x) d\mathbb{P}(x|Y = -1). \tag{10b}$$

The integrals are defined with respect to a dominating measure, e.g., $\mathbb{P}(x|Y = 1) + \mathbb{P}(x|Y = -1)$. It can be shown using Lyapunov's theorem that the space of $\{(\mu, \pi)\}$ by varying $Q \in \mathcal{Q}$ (or $\mathcal{Q}_0$) is both convex and compact under an appropriately defined topology (see Tsitsiklis, 1993a).

For simplicity, in this paper, we assume that the spaces $\mathcal{Q}$ and $\mathcal{Q}_0$ are restricted such that both μ and π are strictly positive measures. Note that the measures μ and π are constrained by the following simple relations:

$$\sum_{z\in\mathcal{Z}} \mu(z) \;=\; \mathbb{P}(Y=1), \qquad \sum_{z\in\mathcal{Z}} \pi(z) \;=\; \mathbb{P}(Y=-1), \quad \text{and} \quad \mu(z)+\pi(z) = \mathbb{P}(z) \quad \text{for each } z \in \mathcal{Z}\,.$$

Note that Y and Z are independent conditioned on X. Therefore, letting $\eta(x) = \mathbb{P}(Y=1|x)$, we can write

$$R_\phi(\gamma, Q) = \mathbb{E}_X\big[\sum_z \phi(\gamma(z))\eta(X)Q(z|X) + \phi(-\gamma(z))(1-\eta(X))Q(z|X)\big]. \tag{11}$$

On the basis of this equation, the ϕ-risk can be written in the following way:

$$\begin{aligned} R_\phi(\gamma, Q) &= \mathbb{E}\phi(Y\gamma(Z)) = \sum_z \phi(\gamma(z))\mathbb{E}_X\big[\eta(X)Q(z|X)\big] + \phi(-\gamma(z))\mathbb{E}_X\big[(1-\eta(X))Q(z|X)\big] \\ &= \sum_z \phi(\gamma(z))\mu(z) + \phi(-\gamma(z))\pi(z). \end{aligned} \tag{12}$$

This representation allows us to compute the optimal value for $\gamma(z)$ for all $z \in \mathcal{Z}$, as well as the optimal ϕ-risk for a fixed Q. We illustrate this procedure with some examples:

0-1 loss. In this case, it is straightforward to see from equation (12) that $\gamma(z) = \mathrm{sign}(\mu(z) - \pi(z))$. As a result, the optimal Bayes risk given a fixed Q takes the form:

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

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

Hinge loss. If ϕ is hinge loss, then equation (12) again yields that $\gamma(z) = \mathrm{sign}(\mu(z) - \pi(z))$. As a result, the optimal risk for hinge loss takes the form:

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

Least squares loss. If ϕ is least squares loss, then $\gamma(z) = \frac{\mu(z)-\pi(z)}{\mu(z)+\pi(z)}$, so that the optimal risk for least squares loss takes the form:

$$\begin{aligned} R_{sqr}(Q) &= \sum_{z\in\mathcal{Z}} \frac{4\mu(z)\pi(z)}{\mu(z)+\pi(z)} = 1 - \sum_{z\in\mathcal{Z}} \frac{(\mu(z)-\pi(z))^2}{\mu(z)+\pi(z)} \\ &= 1 - \Delta(\mu, \pi), \end{aligned}$$

where $\Delta(\mu,\pi)$ denotes the *triangular discrimination* distance defined by $\Delta(\mu,\pi) := \sum_{z\in\mathcal{Z}} \frac{(\mu(z)-\pi(z))^2}{\mu(z)+\pi(z)}$.

Logistic loss. If ϕ is logistic loss, then $\gamma(z) = \log\frac{\mu(z)}{\pi(z)}$. As a result, the optimal risk for logistic loss takes the form:

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

where $KL(U,V)$ denotes the Kullback-Leibler divergence between two measures U and V, and $C(U,V)$ denotes the *capacitory discrimination* distance defined by

$$C(U,V) \;:=\; KL(U||\frac{U+V}{2}) + KL(V||\frac{U+V}{2}).$$

Exponential loss. If ϕ is exponential loss, then $\gamma(z) = \frac{1}{2}\log\frac{\mu(z)}{\pi(z)}$. The optimal risk for exponential loss takes the form:

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

where $h(\mu,\pi) := \frac{1}{2}\sum_{z\in\mathcal{Z}}(\sqrt{\mu(z)} - \sqrt{\pi(z)})^2$ denotes the Hellinger distance between measures μ and π.

It is worth noting that in all of these cases, the optimum ϕ-risk takes the form of a well-known "distance" or "divergence" function. This observation motivates a more general investigation of the relationship between surrogate loss functions and the form of the optimum risk.

3 Correspondence between surrogate loss functions and divergences

The correspondence exemplified in the previous section turns out to be quite general. So as to make this connection precise, we begin by defining the class of *f-divergence functions*, which includes all of the examples discussed above as well as numerous others (Csiszáŕ, 1967, Ali and Silvey, 1966):

Definition 3. *Given any continuous convex function $f : [0,+\infty) \to \mathbb{R}\cup\{+\infty\}$, the f-divergence between measures μ and π is given by*

$$I_f(\mu,\pi) := \sum_z \pi(z) f\left(\frac{\mu(z)}{\pi(z)}\right). \tag{13}$$

As particular cases, the variational distance is given by $f(u) = |u-1|$, Kullback-Leibler divergence by $f(u) = u\ln u$, triangular discrimination by $f(u) = (u-1)^2/(u+1)$, and Hellinger distance by $f(u) = \frac{1}{2}(\sqrt{u}-1)^2$. Other well-known f-divergences include the (negative) Bhattacharyya distance ($f(u) = -2\sqrt{u}$), and the (negative) harmonic distance ($f(u) = -\frac{4u}{u+1}$).

As discussed in the introduction, these functions are widely used in the engineering literature to solve problems in decentralized detection and signal selection. Specifically, for a pre-specified joint distribution $\mathbb{P}(X,Y)$ and a given quantizer Q, one defines an f-divergence on the class-conditional distributions

$\mathbb{P}(Z|Y=1)$ and $\mathbb{P}(Z|Y=-1)$. This f-divergence is then viewed as a function of the underlying Q, and the optimum quantizer is chosen by maximizing the f-divergence. Typically, the discriminant function γ—which acts on the quantized space $\mathcal{Z}$— has an explicit form in terms of the distributions $P(Z|Y=1)$ and $P(Z|Y=-1)$. As we have discussed, the choice of the class of f-divergences as functions to optimize is motivated both by Blackwell's classical theorem (Blackwell, 1951) on the design of experiments, as well as by the computational intractability of minimizing the probability of error, a problem rendered particularly severe in practice when X is high dimensional (Kailath, 1967, Poor and Thomas, 1977, Longo et al., 1990).

3.1 From ϕ-risk to f-divergence

In the following two sections, we develop a general relationship between optimal ϕ-risks and f-divergences. The easier direction, on which we focus in the current section, is moving from ϕ-risk to f-divergence. In particular, we begin with a simple result that shows that any ϕ-risk induces a corresponding f-divergence.

Proposition 4. *For each fixed Q, let γ_Q be the optimal decision rule for the fusion center. Then the ϕ-risk for (Q, γ_Q) is a f-*divergence *between μ and π, as defined in equation* (10)*, for some convex function f:*

$$R_\phi(Q) = -I_f(\mu, \pi). \tag{14}$$

Moreover, this relation holds whether or not ϕ is convex.

Proof. The optimal ϕ-risk has the form

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

For each z, define $u := \frac{\mu(z)}{\pi(z)}$. With this notation, the function $\min_\alpha(\phi(-\alpha) + \phi(\alpha)u)$ is concave as a function of u (since the minimum of a collection of linear functions is concave). Thus, if we define

$$f(u) := -\min_\alpha(\phi(-\alpha) + \phi(\alpha)u). \tag{15}$$

then the claim follows. Note that the argument does not require convexity of ϕ. □

Remark: We can also write $I_f(\mu, \pi)$ in terms of an f-divergence between the two conditional distributions $\mathbb{P}(Z|Y=1) \sim \mathbb{P}_1(Z)$ and $\mathbb{P}(Z|Y=-1) \sim \mathbb{P}_{-1}(Z)$. Recalling the notation $q = \mathbb{P}(Y=-1)$, we have:

$$I_f(\mu, \pi) = q\sum_z \mathbb{P}_{-1}(z) f\left(\frac{(1-q)\mathbb{P}_1(z)}{q\mathbb{P}_{-1}(z)}\right) = I_{f_q}(\mathbb{P}_1, \mathbb{P}_{-1}), \tag{16}$$

where $f_q(u) := qf((1-q)u/q)$. Although it is equivalent to study either form of divergences, we focus primarily on the representation (14) because the prior probabilities are absorbed in the formula. It will be convenient, however, to use the alternative (16) when the connection to the general theory of comparison of experiments is discussed.

3.2 From f-divergence to ϕ-risk

In this section, we develop the converse of Proposition 4. Given a divergence $I_f(\mu, \pi)$ for some convex function f, does there exists a loss function ϕ for which $R_\phi(Q) = -I_f(\mu, \pi)$? We establish that such a correspondence indeed holds for a general class of margin-based convex loss functions; in such cases, it is possible to construct ϕ to induce a given f-divergence.

3.2.1 Some intermediate functions

Our approach to establishing the desired correspondence proceeds via some intermediate functions, which we define in this section. First, let us define, for each β, the inverse mapping

$$\phi^{-1}(\beta) := \inf\{\alpha : \phi(\alpha) \leq \beta\}, \tag{17}$$

where $\inf \emptyset := +\infty$. The following result summarizes some useful properties of ϕ^{-1}:

Lemma 5. *(a) For all $\beta \in \mathbb{R}$ such that $\phi^{-1}(\beta) < +\infty$, the inequality $\phi(\phi^{-1}(\beta)) \leq \beta$ holds. Furthermore, equality occurs when ϕ is continuous at $\phi^{-1}(\beta)$.*

(b) The function $\phi^{-1} : \mathbb{R} \to \overline{\mathbb{R}}$ is strictly decreasing and convex.

Proof. See Appendix A. □

Using the function ϕ^{-1}, we then define a new function $\Psi : \mathbb{R} \to \overline{\mathbb{R}}$ by

$$\Psi(\beta) \quad := \quad \begin{cases} \phi(-\phi^{-1}(\beta)) & \text{if } \phi^{-1}(\beta) \in \mathbb{R}, \\ +\infty & \text{otherwise.} \end{cases} \tag{18}$$

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

$$\beta_1 \; := \; \inf\{\beta : \Psi(\beta) < +\infty\} \qquad \text{and} \qquad \beta_2 \; := \; \inf\{\beta : \Psi(\beta) = \inf \Psi\}. \tag{19}$$

It is simple to check that $\inf \phi = \inf \Psi = \phi(\alpha^*)$, and $\beta_1 = \phi(\alpha^*)$, $\beta_2 = \phi(-\alpha^*)$. Furthermore, by construction, we have $\Psi(\beta_2) = \phi(\alpha^*) = \beta_1$, as well as $\Psi(\beta_1) = \phi(-\alpha^*) = \beta_2$. The following properties of Ψ are particularly useful for our main results.

Lemma 6. *(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 a decreasing function, then Ψ is convex in $(-\infty, +\infty)$.

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

(d) For any $\alpha \geq 0$, $\phi(\alpha) = \Psi(\phi(-\alpha))$. In particular, there exists $u^* \in (\beta_1, \beta_2)$ such that $\Psi(u^*) = u^*$.

(e) The function Ψ satisfies $\Psi(\Psi(\beta)) \leq \beta$ for all $\beta \in$ Dom(Ψ). Moreover, if ϕ is a continuous function on its domain $\{\alpha \in \mathbb{R} \,|\, \phi(\alpha) < +\infty\}$, then $\Psi(\Psi(\beta)) = \beta$ for all $\beta \in (\beta_1, \beta_2)$.

Proof. See Appendix B. □

Remark: With reference to statement (b), if ϕ is not a decreasing function, then the function Ψ need not be convex on the entire real line. See Appendix B for an example.

The following result provides the necessary connection between the function Ψ and the f-divergence associated with ϕ, as defined in equation (15):

Proposition 7. *(a) Given a loss function ϕ, the associated f-divergence* (15) *satisfies the relation*

$$f(u) = \Psi^*(-u), \tag{20}$$

where Ψ^* denotes the conjugate dual of Ψ. If the surrogate loss ϕ is decreasing, then $\Psi(\beta) = f^*(-\beta)$.

(b) For a given Ψ, there exists a point $u^* \in (\beta_1, \beta_2)$ such that $\Psi(u^*) = u^*$. All loss functions ϕ that induce Ψ via (18) *take the form:*

$$\phi(\alpha) \quad = \quad \begin{cases} u^* & \text{if } \alpha = 0 \\ \Psi(g(\alpha + u^*)) & \text{if } \alpha > 0 \\ g(-\alpha + u^*) & \text{if } \alpha < 0, \end{cases} \tag{21}$$

where $g : [u^*, +\infty) \to \overline{\mathbb{R}}$ is some increasing continuous convex function such that $g(u^*) = u^*$, and g is right-differentiable at u^* with $g'(u^*) > 0$.

Proof. (a) From equation (15), we have

$$f(u) \quad = \quad -\inf_{\alpha \in \mathbb{R}} \Big(\phi(-\alpha) + \phi(\alpha) u) \Big) = - \inf_{\left\{ \alpha, \beta \,\middle|\, \phi^{-1}(\beta) \in \mathbb{R},\ \phi(\alpha) = \beta \right\}} \Big(\phi(-\alpha) + \beta u \Big).$$

For β such that $\phi^{-1}(\beta) \in \mathbb{R}$, there might be more than one α such that $\phi(\alpha) = \beta$. However, our assumption (8) ensures that $\alpha = \phi^{-1}(\beta)$ results in minimum $\phi(-\alpha)$. Hence,

$$\begin{aligned} f(u) \quad &= \quad -\inf_{\beta : \phi^{-1}(\beta) \in \mathbb{R}} \Big(\phi(-\phi^{-1}(\beta)) + \beta u \Big) = -\inf_{\beta \in \mathbb{R}} (\beta u + \Psi(\beta)) \\ &= \quad \sup_{\beta \in \mathbb{R}} (-\beta u - \Psi(\beta)) = \Psi^*(-u). \end{aligned}$$

If ϕ is decreasing, then Ψ is convex. By convex duality and the lower semicontinuity of Ψ (from Lemma 6), we can also write:

$$\Psi(\beta) = \Psi^{**}(\beta) = f^*(-\beta). \tag{22}$$

(b) From Lemma 6, we have $\Psi(\phi(0)) = \phi(0) \in (\beta_1, \beta_2)$. As a consequence, $u^* := \phi(0)$ satisfies the relation $\Psi(u^*) = u^*$. Since ϕ is decreasing and convex on the interval $(-\infty, 0]$, for any $\alpha \geq 0$, $\phi(-\alpha)$ can be written as the form:

$$\phi(-\alpha) = g(\alpha + u^*),$$

where g is some increasing convex function. From Lemma 6, we have $\phi(\alpha) = \Psi(\phi(-\alpha)) = \Psi(g(\alpha + u^*)$ for $\alpha \geq 0$. To ensure the continuity at 0, there holds $u^* = \phi(0) = g(u^*)$. To ensure that ϕ is classification-calibrated, we require that ϕ is differentiable at 0 and $\phi'(0) < 0$. These conditions in turn imply that g must be right-differentiable at u^* with $g'(u^*) > 0$. □

3.2.2 A converse theorem

One important aspect of Proposition 7(a) is that it suggests a route—namely via convex duality (Rockafellar, 1970)—to recover the function Ψ from f, assuming that Ψ is lower semi-continuous. We exploit this intuition in the following:

Theorem 8. *Given a lower semicontinuous convex function $f : \mathbb{R} \to \overline{\mathbb{R}}$, consider the function:*

$$\Psi(\beta) = f^*(-\beta). \tag{23}$$

Define $\beta_1 := \inf\{\beta : \Psi(\beta) < +\infty\}$ and $\beta_2 := \inf\{\beta : \Psi(\beta) \leq \inf \Psi\}$, and suppose that Ψ is decreasing, and satisfies $\Psi(\Psi(\beta)) = \beta$ for all $\beta \in (\beta_1, \beta_2)$.

(a) Then any *continuous loss function ϕ of the form* (21) *must induce f-divergence with respect to f in the sense of* (14) *and* (15)*.*

(b) Moreover, if Ψ is differentiable at the point $u^* \in (\beta_1, \beta_2)$ such that $\Psi(u^*) = u^*$, then any such ϕ is classification-calibrated.

Proof. (a) Since f is lower semicontinuous by assumption, convex duality allows us to write

$$f(u) = f^{**}(u) = \Psi^*(-u) = \sup_{\beta \in \mathbb{R}}(-\beta u - \Psi(\beta)) = -\inf_{\beta \in \mathbb{R}}(\beta u + \Psi(\beta)).$$

Proposition 7(b) guarantees that all convex loss function ϕ for which equations (14) and (15) hold must have the form (21). Note that Ψ is lower semicontinuous and convex by definition. It remains to show that any convex loss function ϕ of form (21) must be linked to Ψ via the relation

$$\Psi(\beta) = \begin{cases} \phi(-\phi^{-1}(\beta)) & \text{if } \phi^{-1}(\beta) \in \mathbb{R}, \\ +\infty & \text{otherwise.} \end{cases} \tag{24}$$

Since Ψ is assumed to be a decreasing function, the function ϕ defined in equation (21) is also a decreasing function. By assumption, we also have $\Psi(\Psi(\beta)) = \beta$ for any $\beta \in (\beta_1, \beta_2)$. Therefore, it is straightforward to verify that there exists $u^* \in (\beta_1, \beta_2)$ such that $\Psi(u^*) = u^*$. Using the value u^*, we divide our analysis into three cases:

- For $\beta \geq u^*$, there exists $\alpha \geq 0$ such that $g(\alpha + u^*) = \beta$. Choose the largest α that is so. From our definition of ϕ, $\phi(-\alpha) = \beta$. Thus $\phi^{-1}(\beta) = -\alpha$. It follows that $\phi(-\phi^{-1}(\beta)) = \phi(\alpha) = \Psi(g(\alpha + u^*)) = \Psi(\beta)$.
- For $\beta < \beta_1 = \inf_{u \in \mathbb{R}} \Psi(u)$, we have $\Psi(\beta) = +\infty$.
- Lastly, for $\beta_1 \leq \beta < u^* < \beta_2$, then there exists $\alpha > 0$ such that $g(\alpha + u^*) \in (u^*, \beta_2)$ and $\beta = \Psi(g(\alpha + u^*))$, which implies that $\beta = \phi(\alpha)$ from our definition. Choose that smallest α that satisfies these conditions. Then $\phi^{-1}(\beta) = \alpha$, and it follows that $\phi(-\phi^{-1}(\beta)) = \phi(-\alpha) = g(\alpha + u^*) = \Psi(\Psi(g(\alpha + u^*))) = \Psi(\beta)$, where we have used the fact that $g(\alpha + u^*) \in (\beta_1, \beta_2)$).

The proof is complete.

(b) From Lemma 6(e), we have $\Psi(\Psi(\beta)) = \beta$ for $\beta \in (\beta_1, \beta_2)$. This fact, in conjunction with the assumption that Ψ is differentiable at u^*, implies that $\Psi'(u^*) = -1$. Therefore, by choosing g to be differentiable at u^* with $g'(u^*) > 0$, as dictated by Proposition 7(b), ensures that ϕ is also differentiable at 0 and $\phi'(0) < 0$. Thus, by Lemma 2, the function ϕ is classification-calibrated. □

Remark: One interesting consequence of Theorem 8 that any f-divergence can be obtained from a fairly large set of surrogate loss functions. More precisely, from the procedure (21), we see that any valid ϕ is specified by a function g that need satisfy only a mild set of conditions. It is important to note that not all ϕ losses of the form (21) are convex, but they still satisfy (15). We illustrate this flexibility with several examples in Section 4.

3.2.3 Some additional properties

Theorem 8 provides one set of conditions for an f-divergence to be realized by some surrogate loss ϕ, as well as a constructive procedure for finding all such loss functions. The following result provides a related set of conditions that can be easier to verify. We say that an f-divergence is *symmetric* if $I_f(\mu, \pi) = I_f(\pi, \mu)$ for any measures μ and π. With this definition, we have the following:

Corollary 9. *The following are equivalent:*

(a) f is realizable by some surrogate loss function ϕ (via Proposition 4).

(b) f-divergence I_f is symmetric.

(c) For any $u > 0$, $f(u) = uf(1/u)$.

Proof. (a) $\Rightarrow$ (b): From Proposition 4, we have the representation $R_\phi(Q) = -I_f(\mu, \pi)$. Alternatively, we can write:

$$R_\phi(Q) = \sum_z \mu(z) \min_\alpha \left(\phi(\alpha) + \phi(-\alpha) \frac{\pi(z)}{\mu(z)} \right) \;=\; -\sum_z \mu(z) f\left(\frac{\pi(z)}{\mu(z)} \right),$$

which is equal to $-I_f(\pi, \mu)$, thereby showing that the f-divergence is symmetric.

(b) $\Rightarrow$ (c): By assumption, the following relation holds for any measures μ and π:

$$\sum_z \pi(z) f(\mu(z)/\pi(z)) = \sum_z \mu(z) f(\pi(z)/\mu(z)). \tag{25}$$

Take any instance of $z = l \in \mathcal{Z}$, and consider measures μ' and π', which are defined on the space $\mathcal{Z} - \{l\}$ such that $\mu'(z) = \mu(z)$ and $\pi'(z) = \pi(z)$ for all $z \in \mathcal{Z} - \{l\}$. Since Equation (25) also holds for μ' and π', it follows that

$$\pi(z) f(\mu(z)/\pi(z)) = \mu(z) f(\pi(z)/\mu(z))$$

for all $z \in \mathcal{Z}$ and any μ and π. Hence, $f(u) = uf(1/u)$ for any $u > 0$.

(c) $\Rightarrow$ (a): It suffices to show that all sufficient conditions specified by Theorem 8 are satisfied.

Since any f-divergence is defined by applying f to a likelihood ratio (see definition (13)), we can assume $f(u) = +\infty$ for $u < 0$ without loss of generality. Since $f(u) = uf(1/u)$ for any $u > 0$, it can be verified using subdifferential calculus (Rockafellar, 1970) that for any $u > 0$, there holds:

$$\partial f(u) = f(1/u) + \partial f(1/u) \frac{-1}{u}. \tag{26}$$

Given some $u > 0$, consider any $v_1 \in \partial f(u)$. Combined with (26), we have

$$f(u) - v_1 u \in \partial f(1/u). \tag{27}$$

By definition of conjugate duality,

$$f^*(v_1) \;=\; v_1 u - f(u).$$

Define $\Psi(\beta) = f^*(-\beta)$. Then,

$$\begin{aligned} \Psi(\Psi(-v_1)) &= \Psi(f^*(v_1)) = \Psi(v_1 u - f(u)) \\ &= f^*(f(u) - v_1 u) = \sup_{\beta \in \mathbb{R}} (\beta f(u) - \beta v_1 u - f(\beta)). \end{aligned}$$

Note that the supremum is achieved at $\beta = 1/u$ because of (27). Therefore, $\Psi(\Psi(-v_1)) = -v_1$ for any $v_1 \in \partial f(u)$ for $u > 0$. In other words, $\Psi(\Psi(\beta)) = \beta$ for any $\beta \in \{-\partial f(u), u > 0\}$. By convex duality, $\beta \in -\partial f(u)$ for some $u > 0$ if and only if $-u \in \partial\Psi(\beta)$ for some $u > 0$ (Rockafellar, 1970). This condition on β is equivalent to $\partial\Psi(\beta)$ containing some negative value. This is satisfied by any $\beta \in (\beta_1, \beta_2)$. Hence, $\Psi(\Psi(\beta)) = \beta$ for $\beta \in (\beta_1, \beta_2)$. In addition, $f(u) = +\infty$ for $u < 0$, Ψ is a decreasing function. Now, as an application of Theorem 8, I_f is realizable by some (decreasing) surrogate loss function.

□

Remarks. It is worth noting that not all f-divergences are symmetric; well-known cases of asymmetric divergences include the Kullback-Leibler divergences $KL(\mu, \pi)$ and $KL(\pi, \mu)$, which correspond to the functions $f(u) = -\log u$ and $f(u) = u \log u$, respectively. Corollary 9 establishes that such asymmetric f-divergences cannot be generated by *any* (margin-based) surrogate loss function ϕ. Therefore, margin-based surrogate losses can be considered as symmetric loss functions. It is important to note that our analysis can be extended to show that asymmetric f-divergences can be realized by general (asymmetric) loss functions. Finally, from the proof of Corollary 9, it can be deduced that if an f-divergence is realized by some surrogate loss function, it is also realized by some decreasing surrogate loss function.

Most surrogate loss functions ϕ considered in statistical learning are bounded from below (e.g., $\phi(\alpha) \geq 0$ for all $\alpha \in \mathbb{R}$). The following result establishes a link between (un)boundedness and the properties of the associated f:

Corollary 10. *Assume that ϕ is a decreasing (continuous convex) loss function corresponding to an f-divergence, where f is a continuous convex function that is bounded from below by an affine function. Then ϕ is* unbounded *from below if and only if f is 1-coercive, i.e., $f(x)/||x|| \to +\infty$ as $||x|| \to \infty$.*

Proof. ϕ is unbounded from below if and only if $\Psi(\beta) = \phi(-\phi^{-1}(\beta)) \in \mathbb{R}$ for all $\beta \in \mathbb{R}$, which is equivalent to the dual function $f(\beta) = \Psi^*(-\beta)$ being 1-coercive (cf. Hiriart-Urruty and Lemaréchal, 2001).

□

Therefore, for any decreasing and lower-bounded ϕ loss (which includes the hinge, logistic and exponential losses), the associated f-divergence is *not* 1-coercive. Other interesting f-divergences such as the *symmetric* KL divergence considered in Bradt and Karlin (1956) are 1-coercive, meaning that any associated surrogate loss ϕ cannot be bounded below. We illustrate such properties of f-divergences and their corresponding loss functions in the following section.

4 Examples of loss functions and f-divergences

In this section, we consider a number of specific examples in order to illustrate the correspondence developed in the previous section. As a preliminary remark, it is simple to check that if f_1 and f_2 are related by $f_1(u) = cf_2(u) + au + b$ for some constants $c > 0$ and a, b, then $I_{f_1}(\mu, \pi) = I_{f_2}(\mu, \pi) + a\mathbb{P}(Y = 1) + b\mathbb{P}(Y = -1)$. This relationship implies that the f-divergences I_{f_1} and I_{f_2}, when viewed as functions of Q, are equivalent (up to an additive constant). For this reason, in the following development, we consider divergences so related to be equivalent. We return to a more in-depth exploration of this notion of equivalence in Section 5.

Example 1 (Hellinger distance). The Hellinger distance is equivalent to the negative of the Bhattacharyya distance, which is an f-divergence with $f(u) = -2\sqrt{u}$ for $u \geq 0$. Let us augment the definition

of f with $f(u) = +\infty$ for $u < 0$; doing so does not alter the Hellinger (or Bhattacharyya) distances. Following the constructive procedure of Theorem 8, we begin by recovering Ψ from f:

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

Thus, we see that $u^* = 1$. If we let $g(u) = u$, then a possible surrogate loss function that realizes the Hellinger distance takes the form:

$$\phi(\alpha) \;\; = \;\; \begin{cases} 1 & \text{if } \alpha = 0 \\ \frac{1}{\alpha+1} & \text{if } \alpha > 0 \\ -\alpha + 1 & \text{if } \alpha < 0. \end{cases}$$

On the other hand, if we set $g(u) = \exp(u-1)$, then we obtain the exponential loss $\phi(\alpha) = \exp(-\alpha)$, agreeing with what was shown in Section 2.3. See Figure 4 for illustrations of these loss functions using difference choices of g.

Example 2 (Variational distance). In Section 2.3, we established that the hinge loss as well as the 0-1 loss both generate the variational distance. This f-divergence is based on the function $f(u) = -2\min(u, 1)$ for $u \geq 0$. As before, we can augment the definition by setting $f(u) = +\infty$ for $u < 0$, and then proceed to recover Ψ from f:

$$\Psi(\beta) = f^*(-\beta) = \sup_{u \in \mathbb{R}}(-\beta u - f(u)) = \begin{cases} 0 & \text{if } \beta > 2 \\ 2 - \beta & \text{if } 0 \leq \beta \leq 2 \\ +\infty & \text{if } \beta < 0. \end{cases}$$

By inspection, we see that $u^* = 1$. If we set $g(u) = u$, then we recover the hinge loss $\phi(\alpha) = (1-\alpha)_+$. On the other hand, choosing $g(u) = e^{u-1}$ leads to the following loss:

$$\phi(\alpha) \;\; = \;\; \begin{cases} (2 - e^{\alpha})_+ & \text{for } \alpha \leq 0 \\ e^{-\alpha} & \text{for } \alpha > 0. \end{cases} \tag{28}$$

Note that this choice of g does not lead to a convex loss ϕ, although this non-convex loss still induces f in the sense of Proposition 4. To ensure that ϕ is convex, g is any increasing convex function in $[1, +\infty)$ such that $g(u) = u$ for $u \in [1, 2]$. See Figure 4 for illustrations.

Example 3 (Capacitory discrimination distance). The capacitory discrimination distance is equivalent to an f-divergence with $f(u) = -u \log \frac{u+1}{u} - \log(u+1)$, for $u \geq 0$. Augmenting this function with $f(u) = +\infty$ for $u < 0$, we have

$$\Psi(\beta) = \sup_{u \in \mathbb{R}} -\beta u - f(u) = \begin{cases} \beta - \log(e^{\beta} - 1) & \text{for } \beta \geq 0 \\ +\infty & \text{otherwise.} \end{cases}$$

This representation shows that $u^* = \log 2$. If we choose $g(u) = \log(1 + \frac{e^u}{2})$, then we obtain the logistic loss $\phi(\alpha) = \log(1 + e^{-\alpha})$.

Example 4 (Triangular discrimination distance). Triangular discriminatory distance is equivalent to the negative of the harmonic distance; it is an f-divergence with $f(u) = -\frac{4u}{u+1}$ for $u \geq 0$. Let us augment f

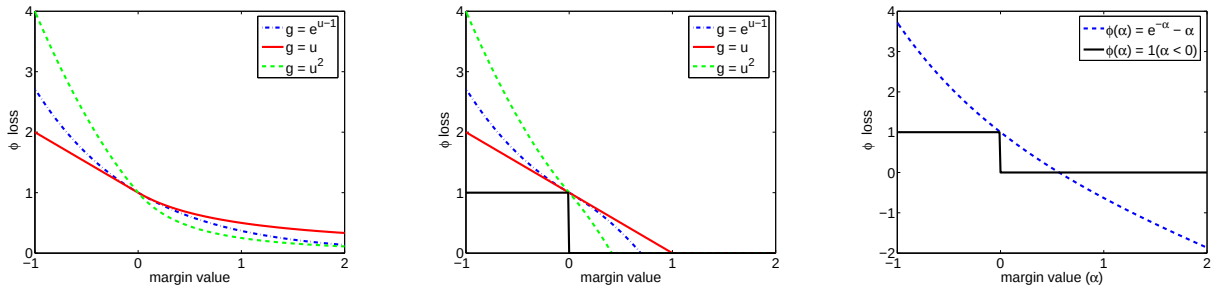

Figure 3. Panels (a) and (b) show examples of ϕ losses that induce the Hellinger distance and variational distance, respectively, based on different choices of the function g. Panel (c) shows a loss function that induces the symmetric KL divergence; for the purposes of comparison, the 0-1 loss is also plotted.

with $f(u) = +\infty$ for $u < 0$. Then we can write

$$\Psi(\beta) = \sup_{u \in \mathbb{R}}(-\beta u - f(u)) = \begin{cases} (2 - \sqrt{\beta})^2 & \text{for } \beta \geq 0 \\ +\infty & \text{otherwise.} \end{cases}$$

Clearly $u^* = 1$. In this case, setting $g(u) = u^2$ gives the least square loss $\phi(\alpha) = (1 - \alpha)^2$.

Example 5 (Another Kullback-Leibler based divergence). Recall that both the KL divergences (i.e., $KL(\mu||\pi)$ and $KL(\pi||\pi)$) are asymmetric; therefore, Corollary 9(b) implies that they are *not* realizable by any margin-based surrogate loss. However, a closely related functional is the *symmetric Kullback-Leibler* divergence (Bradt and Karlin, 1956):

$$KL_s(\mu, \pi) := KL(\mu||\pi) + KL(\pi||\mu). \tag{29}$$

It can be verified that this symmetrized KL divergence is an f-divergence, generated by the function $f(u) = -\log u + u \log u$ for $u \geq 0$, and $+\infty$ otherwise. Therefore, Corollary 9(a) implies that it can be generated by some surrogate loss function, but the form of this loss function is not at all obvious. Therefore, in order to recover an explicit form for some ϕ, we follow the constructive procedure of Theorem 8, first defining

$$\Psi(\beta) \quad = \quad \sup_{u \geq 0} \big\{ -\beta u + \log u - u \log u \big\}.$$

In order to compute the value of this supremum, we take the derivative with respect to u and set it to zero; doing so yields the zero-gradient condition $-\beta + 1/u - \log u - 1 = 0$. To capture this condition, we define a function $r : [0, +\infty) \to [-\infty, +\infty]$ via $r(u) = 1/u - \log u$. It is easy to see that $r(u)$ is a strictly decreasing function whose range covers the whole real line; moreover, the zero-gradient condition is equivalent to $r(u) = \beta + 1$. We can thus write $\Psi(\beta) = u + \log u - 1$ where $u = r^{-1}(\beta + 1)$, or equivalently

$$\Psi(\beta) \;=\; r(1/u) - 1 \;=\; r\left(\frac{1}{r^{-1}(\beta + 1)}\right) - 1.$$

It is straightforward to verify that the function Ψ thus specified is strictly decreasing and convex with $\Psi(0) = 0$, and that $\Psi(\Psi(\beta)) = \beta$ for any $\beta \in \mathbb{R}$. Therefore, Proposition 7 and Theorem 8 allow us to specify the

form of any convex surrogate loss function that generate the symmetric KL divergence; in particular, any such functions must be of the form (21):

$$\phi(\alpha) = \begin{cases} g(-\alpha) & \text{for } \alpha \leq 0 \\ \Psi(g(\alpha)) & \text{otherwise,} \end{cases}$$

where $g : [0, +\infty) \to [0, +\infty)$ is some increasing convex function satisfying $g(0) = 0$. As a particular example (and one that leads to a closed form expression for ϕ), let us choose $g(u) = e^u + u - 1$. Doing so leads to the surrogate loss function

$$\phi(\alpha) = e^{-\alpha} - \alpha - 1.$$

It can be verified by some calculations that the optimized ϕ-risk is indeed the symmetrized KL divergence. See Figure 4(c) for an illustration of this loss function.

5 On comparison of surrogate loss functions and quantizer designs

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. Our ultimate goal, however, is that of choosing an optimal Q, which can be viewed as a problem of experimental design (Blackwell, 1953). Accordingly, the remainder of this paper is devoted to the joint optimization of ϕ-risk (or more precisely, its empirical version) over both the discriminant function γ as well as the choice of experiment Q (hereafter referred to as a quantizer). In particular, we address the fundamental question associated with such an estimation procedure: for what loss functions ϕ does such joint optimization lead to minimum Bayes risk? Note that this question is not covered by standard consistency results (Zhang, 2004, Steinwart, 2005, Bartlett et al., 2005) on classifiers obtained from surrogate loss functions, because the optimization procedure involves both the discriminant function γ and the choice of quantizer Q.

5.1 Inequalities relating surrogate loss functions and f-divergences

The correspondence between surrogate loss functions and f-divergence allows one to compare surrogate ϕ-risks by comparing the corresponding f-divergences, and vice versa. For instance, since the optimal ϕ-risk for hinge loss is equivalent to the optimal ϕ-risk for 0-1 loss, we can say affirmatively that minimizing risk for hinge loss is equivalent to minimizing the Bayes risk. Moreover, it is well-known that the f-divergences are connected via various inequalities, some of which are summarized in the following lemma, proved in Appendix C:

Lemma 11. *The following inequalities among f-divergences hold:*

(a) $V^2 \leq \Delta \leq V$.

(b) $2h^2 \leq \Delta \leq 4h^2$. *As a result,* $\frac{1}{2}V^2 \leq 2h^2 \leq V$.

(c) $\frac{1}{2}\Delta \leq C \leq \log 2 \cdot \Delta$. *As a result,* $\frac{1}{2}V^2 \leq C \leq (\log 2)\, V$.

Using this lemma and our correspondence theorem, it is straightforward to derive the following connection between different risks.

Lemma 12. *The following inequalities among optimized ϕ-risks hold:*

(a) $R_{hinge}(Q) = 2R_{bayes}(Q)$.

(b) $2R_{bayes}(Q) \leq R_{sqr}(Q) \leq 1 - (1 - 2R_{bayes}(Q))^2$.

(c) $2 \cdot \log 2R_{bayes}(Q) \leq R_{log}(Q) \leq \log 2 - \frac{1}{2}(1 - 2R_{bayes}(Q))^2$.

(d) $2R_{bayes}(Q) \leq R_{exp}(Q) \leq 1 - \frac{1}{2}(1 - 2R_{bayes}(Q))^2$.

Note that Lemma 12 shows that all the ϕ-risks considered (i.e., hinge, square, logistic, and exponential) are bounded below by the variational distance (up to some constant multiplicative term). However, with the exception of hinge loss, these results do *not* tell us whether minimizing ϕ-risk leads to a classifier-quantizer pair (γ, Q) with minimal Bayes risk. We explore this issue in more detail in the sequel: more precisely, we specify all surrogate losses ϕ such that minimizing the associated ϕ-risk leads to the same optimal decision rule (Q, γ) as minimizing the Bayes risk.

5.2 Connection between 0-1 loss and f-divergences

The connection between f-divergences and 0-1 loss can be traced back to seminal work on comparison of experiments, pioneered by Blackwell and others (Blackwell, 1951, 1953, Bradt and Karlin, 1956).

Definition 13. *The quantizer Q_1 dominates Q_2 if $R_{Bayes}(Q_1) \leq R_{Bayes}(Q_2)$ for any choice of prior probabilities $q = \mathbb{P}(Y = -1) \in (0, 1)$.*

Recall that a choice of quantizer design Q induces two conditional distributions $P(Z|Y = 1) \sim P_1$ and $P(Z|Y = -1) \sim P_{-1}$. Hence, we shall use P_{-1}^Q and P_1^Q to denote the fact that both P_{-1} and P_1 are determined by the specific choice of Q. By "parameterizing" the decision-theoretic criterion in terms of loss function ϕ and establishing a precise correspondence between ϕ and the f-divergence, we can derive the following theorem that relates 0-1 loss and f-divergences:

Theorem 14. (Blackwell, 1951, 1953) *For any two quantizer designs Q_1 and Q_2, the following statement are equivalent:*

(a) *Q_1 dominates Q_2 (i.e., $R_{bayes}(Q_1) \leq R_{bayes}(Q_2)$ for any prior probabilities $q \in (0, 1)$).*

(b) *$I_f(P_1^{Q_1}, P_{-1}^{Q_1}) \geq I_f(P_1^{Q_2}, P_{-1}^{Q_2})$, for all functions f of the form $f(u) = -\min(u, c)$ for some $c > 0$.*

(c) *$I_f(P_1^{Q_1}, P_{-1}^{Q_1}) \geq I_f(P_1^{Q_2}, P_{-1}^{Q_2})$, for all convex functions f.*

We include a short proof of this result in Appendix D, using the tools developed in this paper. In conjunction with our correspondence between f-divergences and ϕ-risks, this theorem implies the following

Corollary 15. *The quantizer Q_1 dominates Q_2 if and only if $R_\phi(Q_1) \leq R_\phi(Q_2)$ for any loss function ϕ.*

Proof. By Proposition 4, we have $R_\phi(Q) = -I_f(\mu, \pi) = -I_{f_q}(P_1, P_{-1})$, from which the corollary follows using Theorem 14. □

Corollary 15 implies that if $R_\phi(Q_1) \leq R_\phi(Q_2)$ for some loss function ϕ, then $R_{bayes}(Q_1) \leq R_{bayes}(Q_2)$ for some set of prior probabilities on the hypothesis space. This implication justifies the use of a given surrogate loss function ϕ in place of the 0-1 loss for *some* prior probability; however, for a given prior probability, it gives no guidance on how to choose ϕ. Moreover, in many applications (e.g., decentralized detections), it is usually the case that the prior probabilities on the hypotheses are fixed, and the goal is to determine optimum quantizer design Q for this fixed set of priors. In such a setting, the Blackwell's notion of Q_1 dominating Q_2 has limited usefulness. With this motivation in mind, the following section is devoted to development of a more stringent method for assessing equivalence between loss functions.

5.3 Universal equivalence

In the following definition, the loss functions ϕ_1 and ϕ_2 realize the f-divergences associated with the convex function f_1 and f_2, respectively.

Definition 16. *The surrogate loss functions ϕ_1 and ϕ_2 are* universally equivalent, *denoted by $\phi_1 \overset{u}{\approx} \phi_2$, if for any $\mathbb{P}(X, Y)$ and quantization rules Q_1, Q_2, there holds:*

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

In terms of the corresponding f-divergences, this relation is denoted by $f_1 \overset{u}{\approx} f_2$.

Observe that this definition is very stringent, in that it requires that the ordering between optimized ϕ_1 and ϕ_2 risks holds for all probability distributions $\mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$. However, this notion of equivalence is needed for nonparametric approaches to classification, in which the underlying distribution $\mathbb{P}$ is not available in parametric form.

The following result provides necessary and sufficient conditions for two f-divergences to be universally equivalent:

Theorem 17. *Let f_1 and f_2 be convex functions on $[0, +\infty) \to \mathbb{R}$ and differentiable almost everywhere. Then $f_1 \overset{u}{\approx} f_2$ if and only if $f_1(u) = cf_2(u) + au + b$ for some constants $c > 0$ and a, b.*

Proof. The proof relies on the following technical result (see Appendix E for a proof):

Lemma 18. *Given a continuous convex function $f : \mathbb{R}^+ \to \mathbb{R}$, define, for any $u, v \in \mathbb{R}^+$, define:*

$$T_f(u, v) := \left\{ \frac{u\alpha - v\beta - f(u) + f(v)}{\alpha - \beta} = \frac{f^*(\alpha) - f^*(\beta)}{\alpha - \beta} \;\middle|\; \alpha \in \partial f(u), \beta \in \partial f(v), \alpha \neq \beta \right\}.$$

If $f_1 \overset{u}{\approx} f_2$, then for any $u, v > 0$, one of the following must be true:

1. *$T_f(u, v)$ are non-empty for both f_1 and f_2, and $T_{f_1}(u, v) = T_{f_2}(u, v)$.*
2. *Both f_1 and f_2 are linear in (u, v).*

Note that if function f is differentiable at u and v and $f'(u) \neq f'(v)$, then $T_f(u, v)$ is reduced to a number:

$$\frac{uf'(u) - vf'(v) - f(u) + f(v)}{f'(u) - f'(v)} = \frac{f^*(\alpha) - f^*(\beta)}{\alpha - \beta},$$

where $\alpha = f'(u)$, $\beta = f'(v)$, and f^* denotes the conjugate dual of f.
Let v is a point where both f_1 and f_2 are differentiable. Let $d_1 = f_1'(v)$, $d_2 = f_2'(v)$. Without loss of generality, assume $f_1(v) = f_2(v) = 0$ (if not, we can consider functions $f_1(u) - f_1(v)$ and $f_2(u) - f_2(v)$).

Now, for any u where both f_1 and f_2 are differentiable, applying Lemma 18 for v and u, then either f_1 and f_2 are both linear in $[v, u]$ (or $[u, v]$ if $u < v$), in which case $f_1(u) = cf_2(u)$ for some constant c, or the following is true:

$$\frac{uf_1'(u) - f_1(u) - vd_1}{f_1'(u) - d_1} = \frac{uf_2'(u) - f_2(u) - vd_2}{f_2'(u) - d_2}.$$

In either case, we have

$$(uf_1'(u) - f_1(u) - vd_1)(f_2'(u) - d_2) = (uf_2'(u) - f_2(u) - vd_2)(f_1'(u) - d_1).$$

Let $f_1(u) = g_1(u) + d_1 u$, $f_2(u) = g_2(u) + d_2 u$. Then, $(ug_1'(u) - g_1(u) - vd_1)g_2'(u) = (ug_2'(u) - g_2(u) - vd_2)g_1'(u)$, implying that $(g_1(u) + vd_1)g_2'(u) = (g_2(u) + vd_2)g_1'(u)$ for any u where f_1 and f_2 are both differentiable. It follows that $g_1(u) + vd_1 = c(g_2(u) + vd_2)$ for some constant c and this constant c has to be the same for any u due to the continuity of f_1 and f_2. Hence, we have $f_1(u) = g_1(u) + d_1 u = cg_2(u) + d_1 u + cvd_2 - vd_1 = cf_2(u) + (d_1 - cd_2)u + cvd_2 - vd_1$. It is now simple to check that $c > 0$ is necessary and sufficient for I_{f_1} and I_{f_2} to have the same monotonicity. □

An important special case is when one of the f-divergences is the variational distance. In this case, we have the following

Proposition 19. *(a) All f-divergences based on continuous convex $f : [0, +\infty) \to \infty$ that are universally equivalent to the variational distance have the form*

$$f(u) = -c \min(u, 1) + au + b \qquad \textit{for some } c > 0. \tag{30}$$

(b) The 0-1 loss is universally equivalent only to those loss functions whose corresponding f-divergence is based on a function of the form (30)*.*

Proof. Note that statement (b) follows immediately from statement (a). The proof in Theorem 17 does not exactly apply here, because it requires both f_1 and f_2 to be differentiable almost everywhere. We provide a modified argument in Appendix F. □

Theorem 17 shows that each class of equivalent f-divergences are restricted by a strong linear relationship. It is important to note, however, that this restrictiveness does *not* translate over to the classes of universally equivalent loss functions (by Theorem 8).

5.4 Convex loss functions equivalent to 0-1 loss

This section is devoted to a more in-depth investigation of the class of surrogate loss functions ϕ that are universally equivalent to the 0-1 loss.

5.4.1 Explicit construction

We begin by presenting several examples of surrogate loss functions equivalent to 0-1 loss. From Proposition 19, any such loss must realize an f-divergence based on a function of the form (30). For simplicity, we let $a = b = 0$; these constants do not have any significant effect on the corresponding loss functions ϕ (only simple shifting and translation operations). Hence, we will be concerned only with loss functions whose corresponding f has the form $f(u) = -c \min(u, 1)$ for $u \geq 0$. Suppose that we augment the definition by setting $f(u) = +\infty$ for $u < 0$; with this modification, f remains a lower semicontinuous convex function. In Section 4, we considered this particular extension, and constructed all loss functions that were equivalent to the 0-1 loss (in particular, see equation (28)). As a special case, this class of loss functions includes the hinge loss function.

Choosing an alternative extension of f for $u < 0$ leads to a different set of loss functions, also equivalent to 0-1 loss. For example, if we set $f(u) = -k \min(u, 1)$ for $u < 0$ where $k \geq c$, then the resulting Ψ takes the form

$$\Psi(\beta) = \begin{cases} (c - \beta)_+ & \text{for } 0 \leq \beta \leq k \\ +\infty & \text{otherwise.} \end{cases}$$

In this case, the associated loss functions ϕ has the form:

$$\phi(\alpha) = \begin{cases} g(c/2 - \alpha) & \text{for } \alpha \leq 0 \\ (c - g(c/2 + \alpha))_+ & \text{when } g(c/2 + \alpha) \leq k \\ +\infty & \text{otherwise,} \end{cases} \tag{31}$$

where g is a increasing convex function such that $g(c/2) = c/2$. However, to ensure that ϕ is a convex function, it is simple to see that g has to be linear in the interval $[c/2, u]$ for some u such that $g(u) = k$.

5.4.2 A negative result

Thus, varying the extension of f for $u < 0$ (and subsequently the choice of g) leads to a large class of possible loss functions equivalent to the 0-1 loss. What are desirable properties of a surrogate loss function? Properties can be desirable either for computational reasons (e.g., convexity, differentiablity), or for statistical reasons (e.g., consistency). Unfortunately, in this regard, the main result of this section is a negative one: in particular, we prove that there is no differentiable surrogate loss that is universally equivalent to the 0-1 loss.

Proposition 20. *There does not exist a continuous and differentiable convex loss function ϕ that is universally equivalent to the 0-1 loss.*

Proof. From Proposition 19, any ϕ that is universally equivalent to 0-1 loss must generate an f-divergence of the form (30). Let $a = b = 0$ without loss of generality; the proof proceeds in the same way for the general case. First, we claim that regardless of how f is augmented for $u < 0$, the function Ψ always has the following form:

$$\Psi(\beta) = f^*(-\beta) \;=\; \sup_{u \in \mathbb{R}} \big\{ -\beta u - f(u) \big\} \;=\; \begin{cases} +\infty & \text{for } \beta < 0 \\ c - \beta & \text{for } 0 \leq \beta \leq c \\ \geq 0 & \text{otherwise.} \end{cases} \tag{32}$$

Indeed, for $\beta < 0$, we have

$$\Psi(\beta) \quad \geq \sup_{u \geq 0} \big\{ -\beta u + c \min(u, 1) \big\} \;=\; +\infty.$$

Turning to the case $\beta \in [0, c]$, we begin by observing that we must have $f(u) \geq -cu$ for $u \leq 0$ (since f is a convex function). Therefore,

$$\sup_{u<0} \big\{ -\beta u - f(u) \big\} \;\leq\; \sup_{u<0} \big\{ -\beta u + cu \big\} \;=\; 0.$$

On the other hand, we have $\sup_{u \geq 0} \big\{ -\beta u + c \min(u, 1) \big\} = c - \beta \geq 0$, so that we conclude that $\Psi(\beta) = c - \beta$ for $\beta \in [0, c]$. Finally, for $\beta \geq c$, we have $\Psi(\beta) \geq \sup_{u \geq 0} \big\{ -\beta u + c \min(u, 1) \big\} = 0$.

Given the form (32), Theorem 7 implies that the loss function ϕ must have the following form:

$$\phi(\alpha) = \begin{cases} g(c/2 - \alpha) & \text{when } \alpha \leq 0 \\ (c - g(c/2 + \alpha))_+ & \text{when } \alpha > 0 \text{ and } g(c/2 + \alpha) \leq c, \\ \geq 0 & \text{otherwise,} \end{cases} \tag{33}$$

where g is an increasing continuous convex function from $[c/2, +\infty)$ to $\mathbb{R}$ satisfying $g(c/2) = c/2$.

For ϕ to be differentiable, the function g has to be differentiable everywhere in its domain. Let $a > 0$ be the value such that $c = g(c/2 + a)$. Since ϕ achieves its minimum at a, $\phi'(a) = 0$. This implies that g has to satisfy $g'(c/2+a) = 0$. That would imply that g attains its minimum at $c/2+a$, but $g(c/2+a) = c > g(c/2)$, which leads to a contradiction.

□

6 Empirical risk minimization with surrogate convex loss functions

As discussed in Section 1, surrogate loss functions are widely used in statistical learning theory, where the goal is to learn a discriminant function given only indirect access to the distribution $\mathbb{P}(X, Y)$ via empirical samples. In this section, we demonstrate the utility of our correspondence between f-divergences and surrogate loss functions in the setting of the elaborated version of the classical discriminant problem, in which the goal is to choose both a discriminant function γ as well as a quantizer Q. As described in previous work (Nguyen et al., 2005), our strategy is the natural one given empirical data: in particular, we choose (Q, γ) by minimizing the empirical version of the ϕ-risk. It is worthwhile noting that without direct access to the distribution $\mathbb{P}(X, Y)$, it is impossible to compute or manipulate the associated f-divergences. In particular, without closed form knowledge of $\mu(z)$ and $\pi(z)$, it is impossible to obtain closed-form solution for the optimal discriminant γ, as required to compute the f-divergence (see Proposition 4). Nonetheless, the correspondence to f-divergences turns out to be useful, in that it allows us to establish Bayes consistency of the procedure based on ϕ-risks for choosing the quantizer and discriminant function.

6.1 Decentralized detection problem

We begin with further background and necessary notation for the decentralized detection problem; see Nguyen et al. (2005) for further details. Let S be an integer, representing some number of sensors that collect observations from the environment. More precisely, for each $t = 1, \ldots, S$, let $X^t \in \mathcal{X}^t$ represent the observation at sensor t, where $\mathcal{X}^t$ denotes the observation space. The covariate vector $X = (X^t,\ t = 1, \ldots, S)$ is obtained by concatenating all of these observations together. We assume that the global estimate $\widehat{Y}$ is to be formed by a *fusion center*. In the *centralized setting*, this fusion center is permitted access to the full vector X of observations. In this case, it is well-known (van Trees, 1990) that optimal decision rules, whether under Bayes error or Neyman-Pearson criteria, can be formulated in terms of the likelihood ratio $\mathbb{P}(X|Y = 1)/\mathbb{P}(X|Y = -1)$. In contrast, the defining feature of the *decentralized setting* is that the fusion center has access only to some form of summary of each observation X^t. More specifically, we suppose that each sensor $t = 1 \ldots, S$ is permitted to transmit a *message* Z^t, taking values in some space $\mathcal{Z}^t$. The fusion center, in turn, applies some decision rule γ to compute an estimate $\widehat{Y} = \gamma(Z^1, \ldots, Z^S)$ of Y based on its received messages.

For simplicity, let us assume that the input space $\mathcal{X}^t$ is identical for each $t = 1, \ldots, S$, and similarly, that the quantized space $\mathcal{Z}^t$ is the same for all t. The original observation space $\mathcal{X}^t$ can be either finite (e.g, having M possible values), or continuous (e.g., Gaussian measurements). The key constraint, giving rise to the decentralized nature of the problem, is that the corresponding message space $\mathcal{Z} = \{1, \ldots, L\}^S$ is discrete with finite number of values, and hence "smaller" than the observation space (i.e., $L \ll M$ in the case of discrete $\mathcal{X}$). The problem is to find, for each sensor $t = 1, \ldots, S$, a decision rule represented as a measurable function $Q^t : \mathcal{X}^t \to \mathcal{Z}^t$, as well as an overall decision rule represented by a measurable function $\gamma : \mathcal{Z} \to \{-1, +1\}$ at the fusion center so as to minimize the *Bayes risk* $\mathbb{P}(Y \neq \gamma(Z))$.

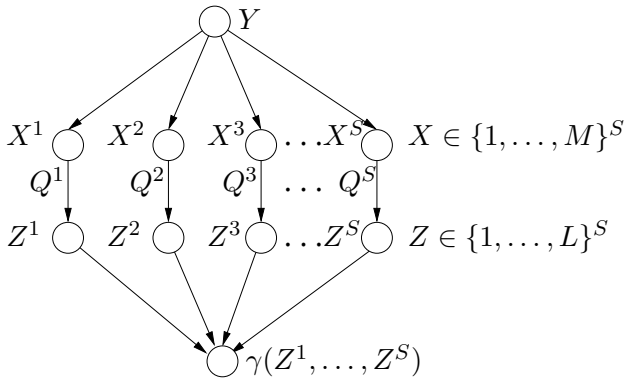

Figure 4. Decentralized detection system with S sensors, in which Y is the unknown hypothesis, $X = (X^1, \ldots, X^S)$ is the vector of sensor observations; and $Z = (Z^1, \ldots, Z^S)$ are the quantized messages transmitted from sensors to the fusion center.

Figure 4 provides a graphical representation of this decentralized detection problem. The single node at the top of the figure represents the hypothesis variable Y, and the outgoing arrows point to the collection of observations $X = (X^1, \ldots, X^S)$. The local decision rules Q^t lie on the edges between sensor observations X^t and messages Z^t. Finally, the node at the bottom is the fusion center, which collects all the messages.

Recall that the quantizer Q can be conveniently viewed as conditional probability distribution $Q(z|x)$, which implies that an aggregate observation x is mapped to an aggregate quantized message z with probability $Q(z|x)$. In particular, the decentralization constraints require that the conditional probability distributions $Q(z|x)$ factorize; i.e., for any realization z of Z, $Q(z|X) = \prod_{t=1}^{S} Q^t(z^t|X^t)$ with probability one. For the remainder of this section, however, we shall use $Q_z(x)$ to denote $Q(z|x)$, to highlight the formal view that the quantizer rule Q is a collection of measurable functions $Q_z : \mathcal{X} \to \mathbb{R}$ for $z \in \mathcal{Z}$.

In summary, our decentralized detection problem is a particular case of the elaborated discriminant problem—namely, a hypothesis testing problem with an additional component of experiment design, corresponding to the choice of the quantizer Q.

A learning algorithm for decentralized detection. Our previous work (Nguyen et al., 2005) introduced an algorithm for designing a decentralized detection system (i.e., both the quantizer and the classifier at the fusion center) based on surrogate loss functions. The algorithm operates on an i.i.d. set of data samples, and makes no assumptions about the underlying probability distribution $\mathbb{P}(X, Y)$. Such an approach is fundamentally different from the bulk of previous work on decentralized decentralization, which typically are based on restrictive parametric assumptions. This type of nonparametric approach is particularly useful in practical applications of decentralized detection (e.g., wireless sensor networks), where specifying an accurate parametric model for the probability distribution $\mathbb{P}(X, Y)$ may be difficult or infeasible.

Let $(x_i, y_i)_{i=1}^n$ be a set of i.i.d. samples from the (unknown) underlying distribution $\mathbb{P}(X, Y)$ over the covariate X and hypothesis $Y \in \{-1, +1\}$. Let $\mathcal{C}_n \subseteq \Gamma$ and $\mathcal{D}_n \subseteq \mathcal{Q}$ represent subsets of classifiers and quantizers, respectively. The algorithm chooses an optimum decision rule $(\gamma, Q) \in (\mathcal{C}_n, \mathcal{D}_n)$ by minimizing an empirical version of ϕ-risk:

$$\hat{R}_\phi(\gamma, Q) := \frac{1}{n} \sum_{i=1}^{n} \sum_{z} \phi(y_i \gamma(z)) Q_z(x_i). \tag{34}$$

It is worth noting that the perspective of surrogate ϕ-loss (as opposed to f-divergence) is the most natural in this nonparametric setting. Given that the minimization takes place over the subset $(\mathcal{C}_n, \mathcal{D}_n)$, there is no

closed-form solution for the minimizer $\gamma \in \mathcal{C}_n$ of problem (34) (even when the optimum Q is known). Hence, it is not even possible to formulate an equivalent closed-form problem in terms of f-divergences. Despite this fact, we demonstrate that the connection to f-divergences is nonetheless useful, in that it allows to address the consistency of the estimation procedure (34). In particular, we prove that for all ϕ that are universally equivalent to the 0-1 loss, this estimation procedure is indeed consistent (for suitable choices of the sequences of function classes $\mathcal{C}_n$ and $\mathcal{D}_n$). The analysis is inspired by frameworks recently developed by a number of authors (see, e.g., Zhang, 2004, Steinwart, 2005, Bartlett et al., 2005) for the standard case of classification (i.e., without any component of experiment design) in statistical machine learning.

6.2 A consistency theorem

For each $z \in \mathcal{Z}$, let us endow the space of functions $Q_z : \mathcal{X} \to \mathbb{R}$ with an appropriate topology, specifically that defined in the proof of Proposition 2.1 in Tsitsiklis (1993a), and endow the space of $\mathcal{Q}$ with the product topology, under which it is shown to be compact (Tsitsiklis, 1993a). In addition, the space of measurable functions $\gamma : \mathcal{Z} \to \{-1, 1\}$ is endowed with the uniform-norm topology.

Consider sequences of increasing compact function classes $\mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq \ldots \subseteq \Gamma$ and $\mathcal{D}_1 \subseteq \mathcal{D}_2 \subseteq \ldots \subseteq \mathcal{Q}$. This analysis supposes that there exists oracle that outputs an optimal solution to the minimization problem

$$\min_{(\gamma,Q)\in(\mathcal{C}_n,\mathcal{D}_n)} \hat{R}_\phi(\gamma, Q), \tag{35}$$

and let (γ_n^*, Q_n^*) denote one such solution. Let R_{bayes}^* denote the minimum Bayes risk achieved over the space of decision rules $(\gamma, Q) \in (\Gamma, \mathcal{Q})$. We refer to the non-negative quantity $R_{bayes}(\gamma_n^*, Q_n^*) - R_{bayes}^*$ the *excess Bayes risk* of our estimation procedure. We say that such an estimation procedure is *universally consistent* if the excess Bayes risk converges to zero (in probability) as $n \to \infty$. More precisely, we require that for any (unknown) Borel probability measure $\mathbb{P}(X, Y)$

$$\lim_{n\to\infty} R_{bayes}(\gamma_n^*, Q_n^*) = R_{bayes}^*. \tag{36}$$

In order to analyze statistical behavior of this algorithm and to establish universal consistency for appropriate sequences $(\mathcal{C}_n, \mathcal{D}_n)$ of function classes, we follow a standard strategy of decomposing the Bayes error in terms of two types of errors:

- the *approximation error* introduced by the bias of the function classes $\mathcal{C}_n \subseteq \Gamma$, and $\mathcal{D}_n \subseteq \mathcal{Q}$, and
- the *estimation error* introduced by the variance of using finite sample size n.

These quantities are defined as follows:

Definition 21. *The approximation error of the procedure is given by*

$$\mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) = \inf_{(\gamma,Q)\in(\mathcal{C}_n,\mathcal{D}_n)} \{R_\phi(\gamma, Q)\} - R_\phi^*, \tag{37}$$

where $R_\phi^* := \inf_{(\gamma,Q)\in(\Gamma,\mathcal{Q})} R_\phi(\gamma, Q)$.

Definition 22. *The estimation error is given by*

$$\mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) = \mathbb{E} \sup_{(\gamma,Q)\in(\mathcal{C}_n,\mathcal{D}_n)} \left| \hat{R}_\phi(\gamma, Q) - R_\phi(\gamma, Q) \right|, \tag{38}$$

where the expectation is taken with respect to the (unknown) measure $\mathbb{P}(X, Y)$.

Conditions on loss function ϕ. Our consistency result applies to the class of surrogate losses that are universally equivalent to the 0-1 loss. From Proposition 19, all such loss functions ϕ correspond to an f-divergence of the form

$$f(u) = -c\min(u, 1) + au + b, \tag{39}$$

for some constants $c > 0, a, b$. For any such ϕ, a straightforward calculation (see the proof of Proposition 4) shows that the optimum risk (for fixed quantizer Q) takes the form

$$R_\phi(Q) = -I_f(\mu, \pi) = c\sum_{z\in\mathcal{Z}} \min\{\mu(z), \pi(z)\} - ap - bq, \tag{40}$$

where $p = \mathbb{P}(Y = 1)$ and $q = \mathbb{P}(Y = -1) = 1 - p$.

Recall that any surrogate loss ϕ is assumed to be continuous, convex, and classification-calibrated (see Definition 1). For our proof, we require the additional technical conditions, expressed in terms of ϕ as well as its induced f-divergence (39):

$$(a - b)(p - q) \geq 0 \qquad \text{and} \qquad \phi(0) \geq 0. \tag{41}$$

Intuitively, these technical conditions are needed so that the approximation error due to varying Q dominates the approximation error due to varying γ (because the optimum γ is determined only after Q is). Simply letting, say, $a = b$ would suffice.

Any surrogate loss that satisfies all of these conditions (continuous, convex, classification-calibrated, universally equivalent to 0-1 loss, and condition (41)) is said to satisfy *property* $\mathcal{P}$. Throughout this section, we shall assume that the loss function ϕ has property $\mathcal{P}$. In addition, for each $n = 1, 2, \ldots$, we assume that

$$M_n \;:=\; \max_{y\in\{-1,+1\}} \sup_{(\gamma,Q)\in(\mathcal{C}_n,\mathcal{D}_n)} \sup_{z\in\mathcal{Z}} |\phi(y\gamma(z))| \;<\; +\infty. \tag{42}$$

The following theorem ties together the Bayes error with the approximation error and estimation error, and provides sufficient conditions for universal consistency:

Theorem 23. *Let $\mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq \ldots \subseteq \Gamma$ and $\mathcal{D}_1 \subseteq \mathcal{D}_2 \subseteq \ldots \subseteq \mathcal{Q}$ be nested sequences of compact function classes, and consider the estimation procedure* (35) *using a surrogate loss ϕ that satisfies property $\mathcal{P}$.*

(a) For any Borel probability measure $\mathbb{P}(X, Y)$, with probability at least $1 - \delta$, there holds:

$$R_{bayes}(\gamma_n^*, Q_n^*) - R_{bayes}^* \;\leq\; \frac{2}{c}\left\{2\mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) + \mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) + 2M_n\sqrt{2\frac{\ln(2/\delta)}{n}}\right\}.$$

(b) Universal Consistency: *Suppose that the function classes satisfy the following properties:*

Approximation condition: $\lim_{n\to\infty} \mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) = 0$.

Estimation condition: $\lim_{n\to\infty} \mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) = 0$ *and* $\lim_{n\to\infty} M_n\sqrt{\ln n/n} = 0$.

Then the estimation procedure (35) *is universally consistent:*

$$\lim_{n\to\infty} R_{bayes}(\gamma_n^*, Q_n^*) = R_{bayes}^* \qquad \textit{in probability}. \tag{43}$$

The proof of this theorem relies on an auxiliary result that is of independent interest. In particular, we prove that for any function classes $\mathcal{C}$ and $\mathcal{D}$, and surrogate loss satisfying property $\mathcal{P}$, the excess ϕ-risk is related to the excess Bayes risk as follows:

Proposition 24. *Let ϕ be a loss function that has property $\mathcal{P}$. Then any classifier-quantizer pair $(\gamma, Q) \in (\mathcal{C}, \mathcal{D})$, we have*

$$\frac{c}{2}\big[R_{bayes}(\gamma, Q) - R^*_{bayes}\big] \quad \leq \quad R_\phi(\gamma, Q) - R^*_\phi. \tag{44}$$

See Appendix G for a proof of this result. A consequence of equation (44) is that in order to achieve Bayes consistency (i.e., driving the excess Bayes risk to zero), it suffices to drive the excess ϕ-risk to zero.

With Proposition 24, we are now equipped to prove Theorem 23:

Proof. (a) First observe that the value of $\sup_{\gamma\in\mathcal{C}_n, Q\in\mathcal{D}_n} |\hat{R}_\phi(\gamma, Q) - R_\phi(\gamma, Q)|$ varies by at most $2M_n/n$ if one changes the values of (x_i, y_i) for some index $i \in \{1, \ldots, n\}$. Hence, applying McDiarmid's inequality yields concentration around the expected value, or (alternatively stated) that with probability at least $1 - \delta$,

$$\left| \sup_{\gamma\in\mathcal{C}_n, Q\in\mathcal{D}_n} |\hat{R}_\phi(\gamma, Q) - R_\phi(\gamma, Q)| - \mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) \right| \leq M_n\sqrt{2\ln(1/\delta)/n}. \tag{45}$$

Suppose that $R_\phi(\gamma, Q)$ attains its minimum over the compact subset $(\mathcal{C}_n, \mathcal{D}_n)$ at $(\gamma^\dagger_n, Q^\dagger_n)$. Then, using Proposition 24, we have

$$\begin{aligned}
\frac{c}{2}(R_{bayes}(\gamma^*_n, Q^*_n) - R^*_{bayes}) \quad &\leq \quad R_\phi(\gamma^*_n, Q^*_n) - R^*_\phi \\
&= \quad R_\phi(\gamma^*_n, Q^*_n) - R_\phi(\gamma^\dagger_n, Q^\dagger_n) + R_\phi(\gamma^\dagger_n, Q^\dagger_n) - R^*_\phi) \\
&= \quad R_\phi(\gamma^*_n, Q^*_n) - R_\phi(\gamma^\dagger_n, Q^\dagger_n) + \mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n)
\end{aligned}$$

Hence, using equation (45), we have with probability at least $1 - \delta$:

$$\begin{aligned}
\frac{c}{2}(R_{bayes}(\gamma^*_n, Q^*_n) - R^*_{bayes}) \quad &\leq \quad \hat{R}_\phi(\gamma^*_n, Q^*_n) - \hat{R}_\phi(\gamma^\dagger_n, Q^\dagger_n) + 2\mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) + 2M_n\sqrt{2\ln(2/\delta)/n} + \mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) \\
&\leq \quad 2\mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) + \mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) + 2M_n\sqrt{2\ln(2/\delta)/n},
\end{aligned}$$

from which Theorem 23(a) follows.

(b) This statement follows by applying (a) with $\delta = 1/n$, and noting that $R_{bayes}(\gamma^*_n, Q^*_n) - R^*_{bayes}$ is bounded. □

A natural question is under what conditions the approximation and estimation conditions of Theorem 23 hold. We conclude this section by stating some precise conditions on the function classes that ensure that the approximation condition holds. Let U be a Borel subset of $\mathcal{X}$ such that $\mathbb{P}_X(U) = 1$, and let $C(U)$ denote the Banach space of continuous functions $Q_z(x)$ mapping U to $\mathbb{R}$. If $\cup_{n=1}^\infty \mathcal{D}_n$ is dense in $\mathcal{Q} \cap C(U)$ and if $\cup_{n=1}^\infty \mathcal{C}_n$ is dense in Γ, then the approximation condition in Theorem 23 holds. In order to establish this fact, note that $R_\phi(\gamma, Q)$ is a continuous function with respect to (γ, Q) over the compact space $(\Gamma, \mathcal{Q})$. (Here compactness is defined with respect to the topology defined in the proof of Proposition 2.1 in Tsitsiklis (1993a).) The approximation condition then follows by applying Lusin's approximation theorem for regular measures, using an argument similar to the proof of Theorem 4.1 in Zhang (2004).

6.3 Estimation error for kernel classes

For the estimation condition in Theorem 23(b) to hold the sequence of function classes $(\mathcal{C}_n, \mathcal{D}_n)_{n=1}^{\infty}$ has to increase sufficiently slowly in "size" with respect to n. In this section, we analyze the behavior of this estimation error for a certain kernel-based function class. Throughout this section, in addition to the conditions imposed on ϕ in the preceding section, we assume that the loss function ϕ is Lipschitz with constant L_ϕ. We also assume without loss of generality that $\phi(0) = 0$ (otherwise, one could consider the modified loss function $\phi(\alpha) - \phi(0)$).

First of all, we require a technical definition of a particular measure of function class complexity:

Definition 25. *Let $\mathcal{F}$ be a class of measurable functions mapping from its domain to $\mathbb{R}$. The* Rademacher complexity *of $\mathcal{F}$ is given by*

$$\mathfrak{R}(\mathcal{F}) = \frac{2}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i f(X_i)\right|, \tag{46}$$

where $\sigma_i,\ i = 1,\ldots n$ are i.i.d. Bernoulli variables (taking values $\{-1,+1\}$ equiprobably), and the expectation is taken over both $\sigma_1,\ldots,\sigma_n$ and $X_1,\ldots,X_n$.

For analyzing the estimation error, the relevant class of functions takes the form

$$\mathcal{G} \;:=\; \Big\{g : \mathcal{X}\to\mathbb{R} \;\Big|\; g(x) = \gamma(\mathrm{argmax}_z Q_z(x)) \quad \text{for some } (\gamma, Q)\in(\mathcal{C},\mathcal{D}\cap\mathcal{Q}_0)\Big\}. \tag{47}$$

We now show that the Rademacher complexity of this class can be used to upper bound the estimation error:

Lemma 26. *For a Lipschitz ϕ (with constant L_ϕ), the estimation error is upper bounded by the Rademacher complexity of $\mathcal{G}$ as follows:*

$$\mathcal{E}_1(\mathcal{C},\mathcal{D}) \le 2L_\phi\mathfrak{R}(\mathcal{G}). \tag{48}$$

Proof. Using the standard symmetrization method (van der Vaart and Wellner, 1996), we have:

$$\begin{aligned}\mathcal{E}_1(\mathcal{C},\mathcal{D}) &\le \mathfrak{R}(\mathcal{H})\\ &= \frac{2}{n}\mathbb{E}\sup_{(\gamma,Q)\in(\mathcal{C},\mathcal{D})}\left|\sum_{i=1}^{n}\sigma_i\sum_{z\in\mathcal{Z}}\phi(y_i\gamma(z))Q_z(x_i)\right|\end{aligned}$$

where $\mathcal{H}$ is the function class given by

$$\mathcal{H} \;:=\; \Big\{h : \mathcal{X}\times\{\pm1\}\to\mathbb{R} \;\Big|\; h(x,y) = \sum_{z\in\mathcal{Z}}\phi(y\gamma(z))Q_z(x) \quad \text{for some } (\gamma,Q)\in(\mathcal{C},\mathcal{D})\Big\}$$

Let $\mathcal{H}_0$ be the subset of $\mathcal{H}$ defined by restricting to $Q \in \mathcal{Q}_0$. Since $\mathcal{Q} = \mathrm{co}\mathcal{Q}_0$ (where co denotes the convex hull), it follows that $\mathcal{H} = \mathrm{co}\mathcal{H}_0$, from which it follows from a result in Bartlett and Mendelson (2002) that $\mathfrak{R}(\mathcal{H}) = \mathfrak{R}(\mathcal{H}_0)$. For $h \in \mathcal{H}_0$, we have $h(x,y) = \phi(y\gamma(\mathrm{argmax}_z Q_z(x)))$. Using results from Bartlett and Mendelson (2002) again, we conclude that $\mathfrak{R}(\mathcal{H}_0) \le 2L_\phi\mathfrak{R}(\mathcal{G})$, □

Using Lemma 26, in order for the estimation condition to hold, it is sufficient to choose the function classes so that the Rademacher complexity converges to zero as n tends to infinity. The function classes used in practice often correspond to classes defined by reproducing kernel Hilbert spaces (RKHS). Accordingly, herein we focus our analysis on such a kernel class.

Briefly, a kernel class of functions is defined as follows. Let $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a positive semidefinite kernel function with $\sup_{z,z'} K(z,z') < +\infty$. Given a kernel function K, we can associate a feature map $\Phi : \mathcal{Z} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with inner product $\langle ., . \rangle$ and for all $z, z' \in \mathcal{Z}$, $K(z,z') = \langle \Phi(z),\, \Phi(z') \rangle$. As a reproducing kernel Hilbert space, any function $\gamma \in \mathcal{H}$ can be expressed as an inner product $\gamma(z) = \langle w,\, \Phi(z) \rangle$, where w can be expressed as $w = \sum_{i=1}^m \alpha_i \Phi(z_i)$ for some $\alpha_1, \ldots, \alpha_m$ and $z_1, \ldots, z_m \in \mathcal{Z}$ for some m. See Aronszajn (1950) and Saitoh (1988) for general mathematical background on reproducing kernel Hilbert spaces, and Schölkopf and Smola (2002) for more details on learning approaches using kernel methods.

If we use this type of kernel class, then the classification rule γ can be written as $\gamma(z) = \sum_{i=1}^m \alpha_i K(z, z_i)$. Suppose that $\mathcal{C}$ is the subset of $\mathcal{H}$ given by

$$\mathcal{C} \;:=\; \left\{ \gamma \;\middle|\; \gamma(z) = \langle w,\, \Phi(z) \rangle,\;\; ||w|| \leq B \right\}, \tag{49}$$

where $B > 0$ is a constant that controls the "size" of the space. Assume further that the space $\mathcal{X}$ is discrete with M^S possible values, and that $\mathcal{Z}$ has L^S possible values. (Recall that S is the total number of covariates $(X_1, \ldots, X_S)$). In Nguyen et al. (2005) we proved that for the function class $\mathcal{G}$ defined in (47), the Rademacher complexity $\mathfrak{R}(\mathcal{G})$ is upper bounded by

$$\frac{2B}{n} \left[\mathbb{E} \sup_{Q \in \mathcal{D}_0} \sum_{i=1}^n K(\mathrm{argmax}_z Q_z(X_i), \mathrm{argmax}_z Q_z(X_i)) + 2(n-1)\sqrt{n/2} \sup_{z,z'} K(z,z') \sqrt{2MS \log L} \right]^{1/2}, \tag{50}$$

which decays with order $O(1/n^{1/4})$. (We note in passing that this $O(1/n^{1/4})$ rate is not tight, but the bound is nonetheless useful for its particularly simple form).

It follows from Lemma 26 and equation (50) that $\mathcal{E}_1(\mathcal{C}, \mathcal{D}) = O(B/n^{1/4})$, where B is the constant used to control the "size" of the function class $\mathcal{C}$ defined in equation (49). Let B_n denote the constant for the corresponding function class $\mathcal{C}_n$, and let $(B_n)_{n=1}^\infty$ be an increasing sequence such that $B_n \to +\infty$. Then, we see from the bound (50) that if B_n increases sufficiently slowly (i.e., slower than $n^{1/4}$), then the estimation error $\mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) \to 0$. Note also that

$$|\gamma(z)| \;\leq\; ||w||^{1/2} ||\Phi(z)||^{1/2} \;=\; O(B_n^{1/2}),$$

so that $M_n = O(B_n^{1/2})$ (where M_n is defined in equation (42)). As a consequence, we have $M_n \sqrt{\ln n / n} \to 0$, so that the estimation condition of condition of Theorem 23(b) holds.

7 Concluding remarks

The main contribution of this paper is a precise explication of the correspondence between loss functions that act as surrogates to the 0-1 loss (which are widely used in statistical machine learning), and the class of f-divergences (which are widely used in information theory and signal processing, and arise as error exponents in the large deviations setting). The correspondence helps explicate the use of various divergences in signal processing and quantization theory, as well as explain the behavior of surrogate loss functions often used in machine learning and statistics. Building on this foundation, we defined the notion of universal equivalence among divergences (and their associated loss functions). As an application of these ideas, we investigated the statistical behavior of a practical nonparametric kernel-based algorithm for designed decentralized hypothesis testing rules, and in particular proved that it is strongly consistent under appropriate conditions.

Acknowledgements

The authors were partially supported by grants from Intel Corporation; Microsoft Research; Grant 0412995 from the National Science Foundation; and an Alfred P. Sloan Foundation Fellowship (MJW).

A Proof of Lemma 5

(a) Since $\phi^{-1}(\beta) < +\infty$, we have $\phi(\phi^{-1}(\beta)) = \phi(\inf\{\alpha : \phi(\alpha) \leq \beta\}) \leq \beta$, where the final inequality follows from the lower semi-continuity of ϕ. If ϕ is continuous at $\phi^{-1}(\beta)$, then we have $\phi^{-1}(\beta) = \min\{\alpha : \phi(\alpha) = \beta\}$, in which case we have $\phi(\phi^{-1}(\beta)) = \beta$.
(b) Due to convexity and the inequality $\phi'(0) < 0$, it follows that ϕ is a strictly decreasing function in $(-\infty, \alpha^*]$. Furthermore, for all $\beta \in \mathbb{R}$ such that $\phi^{-1}(\beta) < +\infty$, we must have $\phi^{-1}(\beta) \leq \alpha^*$. Therefore, definition (17) and the (decreasing) monotonicity of ϕ imply that for any $a, b \in \mathbb{R}$, if $b \geq a \geq \inf \phi$, then $\phi^{-1}(a) \geq \phi^{-1}(b)$, which establishes that ϕ^{-1} is a decreasing function. In addition, we have $a \geq \phi^{-1}(b)$ if and only if $\phi(a) \leq b$.

Now, due to the convexity of ϕ, applying Jensen's inequality for any $0 < \lambda < 1$, we have $\phi(\lambda\phi^{-1}(\beta_1) + (1-\lambda)\phi^{-1}(\beta_2)) \leq \lambda\phi(\phi^{-1}(\beta_1) + (1-\lambda)\phi(\phi^{-1}(\beta_2)) \leq \lambda\beta_1 + (1-\lambda)\beta_2$. Therefore,

$$\lambda\phi^{-1}(\beta_1) + (1-\lambda)\phi^{-1}(\beta_2) \geq \phi^{-1}(\lambda\beta_1 + (1-\lambda)\beta_2),$$

implying the convexity of ϕ^{-1}.

B Proof of Lemma 6

(a) We first prove the statement for the case of a decreasing function ϕ. First, if $a \geq b$ and $\phi^{-1}(a) \notin \mathbb{R}$, then $\phi^{-1}(b) \notin \mathbb{R}$, hence $\Psi(a) = \Psi(b) = +\infty$. If only $\phi^{-1}(b) \notin \mathbb{R}$, then clearly $\Psi(b) \geq \Psi(a)$ (since $\Psi(b) = +\infty$). If $a \geq b$, and both $\phi^{-1}(\alpha), \phi^{-1}(\beta) \in \mathbb{R}$, then from the previous lemma, $\phi^{-1}(a) \leq \phi^{-1}(b)$, so that $\phi(-\phi^{-1}(a)) \leq \phi(-\phi^{-1}(b))$, implying that Ψ is a decreasing function.

We next consider the case of a general function ϕ. For $\beta \in (\beta_1, \beta_2)$, we have $\phi^{-1}(\beta) \in (-\alpha^*, \alpha^*)$, and hence $-\phi^{-1}(\beta) \in (-\alpha^*, \alpha^*)$. Since ϕ is strictly decreasing in $(-\infty, \alpha^*]$, then $\phi(-\phi^{-1}(\beta))$ is strictly decreasing in (β_1, β_2). Finally, when $\beta < \inf \Psi = \phi(\alpha^*)$, $\phi^{-1}(\beta) \notin \mathbb{R}$, so $\Psi(\beta) = +\infty$ by definition.

(b) First of all, assume that ϕ is decreasing. By applying Jensen's inequality, for any $0 < \lambda < 1$, and γ_1, γ_2, we have:

$$\begin{aligned}
\lambda\Psi(\gamma_1) + (1-\lambda)\Psi(\gamma_2)) &= \lambda\phi(-\phi^{-1}(\gamma_1)) + (1-\lambda)\phi(-\phi^{-1}(\gamma_2) \\
&\geq \phi(-\lambda\phi^{-1}(\gamma_1) - (1-\lambda)\phi^{-1}(\gamma_2)) \qquad \text{due to convexity of } \phi \\
&\geq \phi(-\phi^{-1}(\lambda\gamma_1 + (1-\lambda)\gamma_2)) \qquad \text{due to convexity of } \phi^{-1}, \quad \text{and decreasing } \phi \\
&= \Psi(\lambda\gamma_1 + (1-\lambda)\gamma_2),
\end{aligned}$$

implying the convexity of Ψ.

In general, the above arguments go through for any $\gamma_1, \gamma_2 \in [\beta_1, \beta_2]$. Since $\Psi(\beta) = +\infty$ for $\beta < \beta_1$, this implies that Ψ is convex in $(-\infty, \beta_2]$.

(c) For any $a \in \mathbb{R}$, from the definition of ϕ^{-1}, and due to the continuity of ϕ,

$$\begin{aligned}\{\beta \,\big|\, \Psi(\beta) = \phi(-\phi^{-1}(\beta)) \leq a\} &= \{\beta \,\big|\, -\phi^{-1}(\beta) \geq \phi^{-1}(a)\} \\ &= \{\beta \,\big|\, \phi^{-1}(\beta) \leq -\phi^{-1}(a)\} \\ &= \{\beta \,\big|\, \beta \geq \phi(-\phi^{-1}(a))\}\end{aligned}$$

is a closed set. Similarly, $\{\beta \in \mathbb{R} \,\big|\, \Psi(\beta) \geq a\}$ is a closed set. Hence Ψ is continuous in its domain.

(d) Since ϕ is assumed to be classification-calibrated, Lemma 2 implies that ϕ is differentiable at 0 and $\phi'(0) < 0$. Since ϕ is convex, this implies that ϕ is strictly decreasing for $\alpha \leq 0$. As a result, for any $\alpha \geq 0$, let $\beta = \phi(-\alpha)$, then we obtain $\alpha = -\phi^{-1}(\beta)$. Since $\Psi(\beta) = \phi(-\phi^{-1}(\beta))$, we have $\Psi(\beta) = \phi(\alpha)$. Hence, $\Psi(\phi(-\alpha)) = \phi(\alpha)$. Letting $u^* = \phi(0)$, then we have $\Psi(u^*) = u^*$, and $u^* \in (\beta_1, \beta_2)$.

(e) Let $\alpha = \Psi(\beta) = \phi(-\phi^{-1}(\beta)$. Then from (17), $\phi^{-1}(\alpha) \leq -\phi^{-1}(\beta)$. Therefore,

$$\Psi(\Psi(\beta)) = \Psi(\alpha) = \phi(-\phi^{-1}(\alpha)) \leq \phi(\phi^{-1}(\beta)) \leq \beta.$$

We have proved that Ψ is strictly decreasing for $\beta \in (\beta_1, \beta_2)$. As such, $\phi^{-1}(\alpha) = -\phi^{-1}(\beta)$. We also have $\phi(\phi^{-1}(\beta)) = \beta$. It follows that $\Psi(\Psi(\beta)) = \beta$ for all $\beta \in (\beta_1, \beta_2)$.

Remark: With reference to statement (b), if ϕ is not a decreasing function, then the function Ψ need not be convex on the entire real line. For instance, the following loss function generates a function Ψ that is not convex:

$$\phi(\alpha) = \begin{cases} (1-\alpha)^2 & \text{when } \alpha \leq 1 \\ 0 & \text{when } 1 \leq \alpha \leq 2 \\ \alpha - 2 & \text{otherwise.} \end{cases}$$

We have $\Psi(9) = \phi(2) = 0, \Psi(16) = \phi(3) = 1, \Psi(25/2) = \phi(-1 + 5/\sqrt{2}) = -3 + 5/\sqrt{2} > (\Psi(9) + \Psi(16))/2$.

C Proof of Lemma 11

(a) The inequality $\Delta \leq V$ is trivial. On the other hand, the inequality $V^2 \leq \Delta$ follows by applying the Cauchy-Schwarz inequality:

$$\Delta = \sum_z \left(\frac{|\mu(z) - \pi(z)|}{\sqrt{\mu(z) + \pi(z)}} \right)^2 \sum_z \left(\sqrt{\mu(z) + \pi(z)} \right)^2 \geq \left(\sum_z |\mu(z) - \pi(z)| \right)^2 = V^2(\mu, \pi).$$

(b) Note that for any $z \in \mathcal{Z}$, we have $1 \leq \frac{(\sqrt{\mu(z)} + \sqrt{\pi(z)})^2}{\mu(z) + \pi(z)} \leq 2$. Applying these inequalities in the following expression

$$\Delta(\mu, \pi) = \sum_{z \in \mathcal{Z}} (\sqrt{\mu(z)} - \sqrt{\pi(z)})^2 \frac{(\sqrt{\mu(z)} + \sqrt{\pi(z)})^2}{\mu(z) + \pi(z)}$$

yields $2h^2 \leq \Delta \leq 4h^2$.

(c) See Topsoe (2000) for a proof.

D Proof of Theorem 14

We first establish the equivalence (a) $\Leftrightarrow$ (b). By the correspondence between 0-1 loss and an f-divergence with $f(u) = -\min(u, 1)$, and the remark following Proposition 4, we have $R_{bayes}(Q) = -I_f(\mu, \pi) = -I_{f_q}(P_1, P_{-1})$, where $f_q(u) := qf(\frac{1-q}{q}u) = -(1-q)\min(u, \frac{q}{1-q})$. Hence, (a) $\Leftrightarrow$ (b).

Next, we prove the equivalence (b) $\Leftrightarrow$ (c). The implication (c) $\Rightarrow$ (b) is immediate. Considering the reverse implication (b) $\Rightarrow$ (c), we note that any convex function $f(u)$ can be uniformly approximated as a sum of a linear function and $-\sum_k \alpha_k \min(u, c_k)$ where $\alpha_k > 0, c_k > 0$ for all k. For a linear function f, $I_f(P_{-1}, P_1)$ does not depend on P_{-1}, P_1. Using these facts, Statement (c) follows from Statement (b).

E Proof of Lemma 18

Consider a joint distribution $\mathbb{P}(X, Y)$ defined by $\mathbb{P}(Y = -1) = q = 1 - \mathbb{P}(Y = 1)$ and

$$\mathbb{P}(X|Y = -1) \sim \text{Uniform}[0, b], \quad \text{and} \quad \mathbb{P}(X|Y = 1) \sim \text{Uniform}[a, c],$$

where $0 < a < b < c$. Let $Z \in \{1, 2\}$ be a quantized version of X. We assume Z is produced by a deterministic quantizer design Q specified by a threshold $t \in (a, b)$; in particular, we set $Q(z = 1|x) = 1$ when $x \geq t$, and $Q(z = 2|x) = 1$ when $x < t$. Under this quantizer design, we have

$$\begin{aligned} \mu(1) &= (1-q)\frac{t-a}{c-a}; \;\; \mu(2) = (1-q)\frac{c-t}{c-a} \\ \pi(1) &= q\frac{t}{b}; \;\; \pi(2) = q\frac{b-t}{b}. \end{aligned}$$

Therefore, the f-divergence between μ and π takes the form:

$$I_f(\mu, \pi) = \frac{qt}{b} f\left(\frac{(t-a)b(1-q)}{(c-a)tq}\right) + \frac{q(b-t)}{b} f\left(\frac{(c-t)b(1-q)}{(c-a)(b-t)q}\right).$$

If $f_1 \overset{u}{\approx} f_2$, then $I_{f_1}(\mu, \pi)$ and $I_{f_1}(\mu, \pi)$ have the same monotonicity property for any $q \in (0, 1)$ as well for for any choice of the parameters q and $a < b < c$. Let $\gamma = \frac{b(1-q)}{(c-a)q}$, which can be chosen arbitrarily positive, and then define the function

$$F(f, t) = tf\left(\frac{(t-a)\gamma}{t}\right) + (b-t)f\left(\frac{(c-t)\gamma}{b-t}\right).$$

Note that the functions $F(f_1, t)$ and $F(f_2, t)$ have the same monotonicity property, for any positive parameters γ and $a < b < c$.

We now claim that $F(f, t)$ is a convex function of t. Indeed, using convex duality (Rockafellar, 1970), $F(f, t)$ can be expressed as follows:

$$\begin{aligned} F(f, t) &= t \sup_{r \in \mathbb{R}} \left\{ \frac{(t-a)\gamma}{t} r - f^*(r) \right\} + (b-t) \sup_{s \in \mathbb{R}} \left\{ \frac{(c-t)\gamma}{b-t} s - f^*(s)) \right\} \\ &= \sup_{r,s} \left\{ \frac{(t-a)r}{\gamma} - tf^*(r) + \frac{(c-t)s}{\gamma} - tf^*(s) \right\}, \end{aligned}$$

which is a supremum over a linear function of t, thereby showing that $F(f, t)$ is convex of t.

It follows that both $F(f_1, t)$ and $F(f_2, t)$ are subdifferentiable everywhere in their domains; since they have the same monotonicity property, we must have

$$0 \in \partial F(f_1, t) \Leftrightarrow 0 \in \partial F(f_2, t). \tag{51}$$

It can be verified using subdifferential calculus (e.g. Hiriart-Urruty and Lemaréchal, 2001) that:

$$\partial F(f, t) = \frac{a\gamma}{t}\partial f\left(\frac{(t-a)\gamma}{t}\right) + f\left(\frac{(t-a)\gamma}{t}\right) - f\left(\frac{(c-t)\gamma}{b-t}\right) + \frac{(c-b)\gamma}{b-t}\partial f\left(\frac{(c-t)\gamma}{b-t}\right).$$

Letting $u = \frac{(t-a)\gamma}{t}$, $v = \frac{(c-t)\gamma}{b-t}$, we have

$$\begin{aligned}
0 \in \partial F(f, t) \quad &\Leftrightarrow \quad 0 \in (\gamma - u)\partial f(u) + f(u) - f(v) + (v - \gamma)\partial f(v) && \text{(52a)}\\
&\Leftrightarrow \quad \exists \alpha \in \partial f(u), \beta \in \partial f(v) \ \text{ s.t. } 0 = (\gamma - u)\alpha + f(u) - f(v) + (v - \gamma)\beta && \text{(52b)}\\
&\Leftrightarrow \quad \exists \alpha \in \partial f(u), \beta \in \partial f(v) \ \text{ s.t. } \gamma(\alpha - \beta) = u\alpha - f(u) + f(v) - v\beta && \text{(52c)}\\
&\Leftrightarrow \quad \exists \alpha \in \partial f(u), \beta \in \partial f(v) \ \text{ s.t. } \gamma(\alpha - \beta) = f^*(\alpha) - f^*(\beta). && \text{(52d)}
\end{aligned}$$

By varying our choice of $q \in (0, 1)$, the number γ can take any positive value. Similarly, by choosing different positive values of a, b, c (such that $a < b < c$), we can ensure that u and v can take on any positive real values such that $u < \gamma < v$. Since equation (51) holds for any t, it follows that for any triples $u < \gamma < v$, equation (52d) holds for f_1 if and only if it also holds for f_2.

Considering a fixed pair $u < v$, first suppose that the function f_1 is linear on the interval $[u, v]$ with a slope s. In this case, equation (52d) holds for f_1 and any γ by choosing $\alpha = \beta = s$, which implies that equation (52d) also holds for f_2 for any γ. Thus, we deduce that f_2 is also a linear function on the interval $[u, v]$.

Suppose, on the other hand, that f_1 and f_2 are both non-linear in $[u, v]$. Due to the monotonicity of subdifferentials, we have $\partial f_1(u) \cap \partial f_1(v) = \emptyset$ and $\partial f_2(u) \cap \partial f_2(v) = \emptyset$. Consequently, it follows that both $T_{f_1}(u, v)$ and $T_{f_2}(u, v)$ are non-empty. If $\gamma \in T_{f_1}(u, v)$, then (52d) holds for f_1 for some γ. Thus, it must also hold for f_2 using the same γ, which implies that $\gamma \in T_{f_2}(u, v)$. The same argument can also be applied with the roles of f_1 and f_2 reversed, so that we conclude that $T_{f_1}(u, v) = T_{f_2}(u, v)$.

F Proof of Proposition 19

Using Lemma 18, the proof of Proposition 19 follows relatively easily. Note that the variational distance corresponds to $f_1(u) = |u - 1| = u + 1 - 2\min\{u, 1\}$, which is linear above and below 1. Therefore, the same must be true for any continuous convex function f_2. All such functions can indeed be written as $-c\min(u, 1) + au + b$, for some constant c, a, b. In order for f_2 to have the same monotonicity as f_1, it is necessary and sufficient that $c > 0$.

G Proof of Proposition 24

Following a similar construction as in the proof of Proposition 20, all ϕ satisfying property $\mathcal{P}$ have $\phi(0) = (c-a-b)/2$. Now, note that

$$\begin{aligned} R_{bayes}(\gamma, Q) - R^*_{bayes} &= R_{bayes}(\gamma, Q) - R_{bayes}(Q) + R_{bayes}(Q) - R^*_{bayes} \\ &= \sum_{z\in\mathcal{Z}} \pi(z)\mathbb{I}(\gamma(z) > 0) + \mu(z)\mathbb{I}(\gamma(z) < 0) - \min\{\mu(z), \pi(z)\} + R_{bayes}(Q) - R^*_{bayes} \\ &= \sum_{z:(\mu(z)-\pi(z))\gamma(z)<0} |\mu(z) - \pi(z)| + R_{bayes}(Q) - R^*_{bayes}. \end{aligned}$$

In addition,

$$R_\phi(\gamma, Q) - R^*_\phi = R_\phi(\gamma, Q) - R_\phi(Q) + R_\phi(Q) - R^*_\phi.$$

By Proposition 4,

$$\begin{aligned} R_\phi(Q) - R^*_\phi &= -I_f(\mu, \pi) - \inf_{Q\in\mathcal{Q}}(-I_f(\mu, \pi)) \\ &= c\sum_{z\in\mathcal{Z}} \min\{\mu(z), \pi(z)\} - \inf_{Q\in\mathcal{Q}} c\sum_{z\in\mathcal{Z}} \min\{\mu(z), \pi(z)\} \\ &= c(R_{bayes}(Q) - R^*_{bayes}). \end{aligned}$$

Therefore, the lemma would be immediate once we could show that

$$\begin{aligned} \frac{c}{2} \sum_{z:(\mu(z)-\pi(z))\gamma(z)<0} |\mu(z) - \pi(z)| &\le R_\phi(\gamma, Q) - R_\phi(Q) \\ &= \sum_{z\in\mathcal{Z}} \pi(z)\phi(-\gamma(z)) + \mu(z)\phi(\gamma(z)) - c\min\{\mu(z), \pi(z)\} + ap + bq. \quad (53) \end{aligned}$$

It is simple to check that for any $z \in \mathcal{Z}$ such that $(\mu(z) - \pi(z))\gamma(z) < 0$, there holds:

$$\pi(z)\phi(-\gamma(z)) + \mu(z)\phi(\gamma(z)) \ge \pi(z)\phi(0) + \mu(z)\phi(0). \quad (54)$$

Indeed, w.o.l.g., suppose $\mu(z) > \pi(z)$. Since ϕ is classification-calibrated, the convex function (with respect to α) $\pi(z)\phi(-\alpha) + \mu(z)\phi(\alpha)$ achieves its minimum at some $\alpha \ge 0$. Hence, for any $\alpha \le 0$, $\pi(z)\phi(-\alpha) + \mu(z)\phi(\alpha) \ge \pi(z)\phi(0) + \mu(z)\phi(0)$. Hence, (54) is proven. The RHS of Eqn. (53) is lower bounded by:

$$\begin{aligned} &\sum_{z:(\mu(z)-\pi(z))\gamma(z)<0} (\pi(z) + \mu(z))\phi(0) - c\min\{\mu(z), \pi(z)\} + ap + bq \\ =& \sum_{z:(\mu(z)-\pi(z))\gamma(z)<0} (\pi(z) + \mu(z))\frac{c-a-b}{2} - c\min\{\mu(z), \pi(z)\} + ap + bq \\ =& \frac{c}{2} \sum_{z:(\mu(z)-\pi(z))\gamma(z)<0} |\mu(z) - \pi(z)| - (a+b)(p+q)/2 + ap + bq \\ =& \frac{c}{2} \sum_{z:(\mu(z)-\pi(z))\gamma(z)<0} |\mu(z) - \pi(z)| + \frac{1}{2}(a-b)(p-q) \\ \ge& \frac{c}{2} \sum_{z:(\mu(z)-\pi(z))\gamma(z)<0} |\mu(z) - \pi(z)|. \end{aligned}$$

This completes the proof.

References

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Royal Stat. Soc. Series B*, 28:131–142, 1966.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68: 337–404, 1950.

P. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification and risk bounds. *Journal of the American Statistical Association*, 2005. To appear.

P. Bartlett and S. Mendelson. Gaussian and Rademacher complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

D. Blackwell. Comparison of experiments. *Proceeding of 2nd Berkeley Symposium on Probability and Statistics*, 1:93–102, 1951.

D. Blackwell. Equivalent comparisons of experiments. *Annals of Statistics*, 24(2):265–272, 1953.

R. S. Blum, S. A. Kassam, and H. V. Poor. Distributed detection with multiple sensors: Part II — advanced topics. *Proceedings of the IEEE*, 85:64–79, 1997.

R. Bradt and S. Karlin. On the design and comparison of certain dichotomous experiments. *Annals of Statistics*, 27(2):390–409, 1956.

L. Breiman. Arcing classifiers. *Annals of Statistics*, 26:801–824, 1998.

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

C. Chong and S. P. Kumar. Sensor networks: Evolution, opportunities, and challenges. *Proceedings of the IEEE*, 91:1247–1256, 2003.

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

T. Cover and J. Thomas. *Elements of information theory*. Wiley, 1991.

I. Csiszaŕ. Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungar*, 2:299–318, 1967.

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.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28:337–374, 2000.

J. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2001.

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

M. Longo, T. Lookabaugh, and R. Gray. Quantization for decentralized hypothesis testing under communication contraints. *IEEE Trans. on Information Theory*, 36(2):241–255, 1990.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Nonparametric decentralized detection using kernel methods. *IEEE Transactions on Signal Processing*, In press, November 2005.

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 Communications*, 25:893–900, 1977.

G. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

S. Saitoh. *Theory of Reproducing Kernels and its Applications*. Longman Scientific & Technical, Harlow, UK, 1988.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *IEEE Trans. Info. Theory*, 51:128–142, 2005.

F. Topsoe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46:1602–1609, 2000.

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

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

A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY, 1996.

H. L. van Trees. *Detection, Estimation and Modulation Theory*. Krieger Publishing Co., Melbourne, FL, 1990.

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