The Annals of Statistics
2018, Vol. 46, No. 6B, 3246–3275
https://doi.org/10.1214/17-AOS1657

MULTICLASS CLASSIFICATION, INFORMATION, DIVERGENCE AND SURROGATE RISK

BY JOHN DUCHI[1], KHASHAYAR KHOSRAVI AND FENG RUAN[1]

Stanford University

We provide a unifying view of statistical information measures, multiway Bayesian hypothesis testing, loss functions for multiclass classification problems and multidistribution f-divergences, elaborating equivalence results between all of these objects, and extending existing results for binary outcome spaces to more general ones. We consider a generalization of f-divergences to multiple distributions, and we provide a constructive equivalence between divergences, statistical information (in the sense of DeGroot) and losses for multiclass classification. A major application of our results is in multiclass classification problems in which we must both infer a discriminant function γ—for making predictions on a label Y from datum X—and a data representation (or, in the setting of a hypothesis testing problem, an experimental design), represented as a quantizer q from a family of possible quantizers Q. In this setting, we characterize the equivalence between loss functions, meaning that optimizing either of two losses yields an optimal discriminant and quantizer q, complementing and extending earlier results of Nguyen et al. [*Ann. Statist.* **37** (2009) 876–904] to the multiclass case. Our results provide a more substantial basis than standard classification calibration results for comparing different losses: we describe the convex losses that are consistent for jointly choosing a data representation and minimizing the (weighted) probability of error in multiclass classification problems.

1. Introduction. Consider the multiclass classification problem: a decision maker receives a pair of random variables $(X, Y) \in \mathcal{X} \times \{1, \dots, k\}$, where Y is unobserved, and wishes to assign the variable X to one of the k classes $\{1, 2, \dots, k\}$ to minimize the probability of a misclassification. We represent the decision maker via a discriminant function $\gamma : \mathcal{X} \to \mathbb{R}^k$, where each component $\gamma_y(x)$, $y = 1, \dots, k$, represents the *margin* (or a score or perceived likelihood) the decision maker assigns to class y for datum x. The goal is then to minimize the expected loss, or L-*risk*,

$$R_L(\gamma) := \mathbb{E}\big[L\big(\gamma(X), Y\big)\big], \tag{1}$$

where $L(\gamma(x), y)$ measures the loss of margins $\gamma(x) \in \mathbb{R}^k$ when the true label of x is y and the expectation (1) is taken jointly over (X, Y). When L is the 0–1 loss,

Received September 2016; revised October 2017.

[1] Supported in part by NSF-CAREER Award 1553086 and the SAIL-Toyota Center for AI Research. Feng Ruan additionally supported by the Stanford Graduate Fellowship.

MSC2010 subject classifications. 62G10, 62K05, 62C05, 94A17, 68Q32.

Key words and phrases. f-divergence, risk, surrogate loss function, hypothesis test.

$L(\gamma(x), y) = 1\{\gamma_y(x) \le \gamma_i(x) \text{ for some } i \ne y\}$, the formulation (1) is the misclassification probability $\mathbb{P}(\operatorname{argmax}_y \gamma_y(X) \ne Y)$. We may also consider the classical k-category Bayesian experiment: given a random variable $X \in \mathcal{X}$ drawn according to one of the k hypotheses

$$H_1 : X \sim P_1, \qquad H_2 : X \sim P_2, \qquad \dots, \qquad H_k : X \sim P_k$$

with prior probabilities $\pi_1, \dots, \pi_k$, we wish to choose γ minimizing the expected $\mathbb{E}[L(\gamma(X), Y)]$ or posterior $\sum_y \mathbb{P}(Y = y \mid X = x) L(\gamma(x), y)$ loss.

We briefly note that often in decision theory, one has $L : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$, where $\mathcal{Y} = [k]$ and the action space $\mathcal{A} = \mathcal{Y}$ or is the probability simplex over $\mathcal{Y}$ (for prediction of distributions on $\mathcal{Y}$ [12, 13]); as we restrict ourselves to *margin-based* loss functions the action space $\mathbb{R}^k$ is most natural. For some margin-based losses, such as hinge-type losses, there is no straightforward probabilistic interpretation or link function mapping predictions to probability vectors. It is, of course, often the case that $\gamma_y(x)$ is a transformation of some hypothesized class-conditional probabilities, in which case if the action space $\mathcal{A} = \mathcal{Y}$ the natural decision is to map $\gamma(x)$ to $\delta(x) = \operatorname{argmax}_y \gamma_y(x)$.

In many applications, making decisions based on the raw X is undesirable—the vector X may be high-dimensional, carry useless information impinging on statistical efficiency or we may need to store or communicate the sample using limited memory or bandwidth. If all we wish to do is to classify a person as being taller or shorter than 160 centimeters, it makes little sense to track his or her blood type and eye color. With the increase in the number and variety of measurements we collect, such careful design choices are important for maintaining statistical power, interpretability, efficient downstream use and mitigating false discovery [3]. This desire to give "better" representations of data X has led to a rich body of work in statistics, machine learning and engineering, highlighting the importance of careful measurement, experimental design and data representation strategies [9, 23, 25, 29].

As Nguyen et al. [20] note in the binary case, one thus frequently extends the classical formulation (1) to include a stage in which a (data-dependent) $\mathsf{q} : \mathcal{X} \to \mathcal{Z}$ maps the vector X into a vector Z. A number of situations suggest such an approach. In most practical classification scenarios [26], an equivalent feature selection reduces the dimension of X or increases its interpretability. As a second motivation, consider the decentralized detection problem [20, 30] in communication applications in engineering, where remote sensors communicate data $X \sim P_i$ through limited bandwidth or memory. In this case, the central processor can infer the distribution P_i only after communication of the transformed vector $Z = \mathsf{q}(X)$, and one chooses a *quantizer* q from a family Q of (low complexity) quantizers. In fuller abstraction, we may treat the problem as a Bayesian experimental design problem, where the mapping $\mathsf{q} : \mathcal{X} \to \mathcal{Z}$ may be stochastic and is chosen from a family Q of possible experiments (observation channels). In each of the preceding examples, the incorporation of a quantizer q into the classification procedure poses a more complex problem, as one must simultaneously find a data representation q and discriminant γ. The goal, paralleling that for the risk (1), thus becomes joint minimization of the *quantized L-risk*

$$R_L(\gamma \mid \mathsf{q}) := \mathbb{E}\big[L\big(\gamma\big(\mathsf{q}(X)\big), Y\big)\big] \tag{2}$$

over a prespecified family Q of quantizers $\mathsf{q} : \mathcal{X} \to \mathcal{Z}$, where $\gamma : \mathcal{Z} \to \mathbb{R}^k$.

Often—for example, in the zero-one error case—the loss functions $L(\cdot, y)$ are nonconvex (even discontinuous), so population or empirical minimization is intractable. It is thus common to replace the loss with a convex *surrogate* and minimize this surrogate instead. A surrogate is *Fisher consistent* if its minimization yields a Bayes optimal discriminant γ for the original loss L [for any distribution $\mathbb{P}$ on (X, Y)]; researchers have characterized the Fisher consistency of convex surrogates for binary and multiclass classification [2, 19, 28, 33]. A weakness of such results is that they rely strongly on using the unrestricted class of all measurable discriminants $\gamma : \mathcal{X} \to \mathbb{R}^k$, and thus most "natural" convex losses are consistent [2, 33]. In this context, a major difficulty is to understand the consequences of using various surrogate losses, and requiring a restricted quantizer class Q is one approach to discovering nuanced properties of the relationship between surrogate and Bayes risk. Nguyen et al. [20] tackle this in the binary case, considering the problem of joint selection of the discriminant function $\gamma : \mathcal{Z} \to \mathbb{R}$ and quantizer $\mathsf{q} : \mathcal{X} \to \mathcal{Z}$. They exhibit a precise correspondence between binary margin-based loss functions and f-divergences—measures of the similarity between two probability distributions developed in information theory and statistics [1, 8, 31]—to give a general characterization of loss equivalence through classes of divergences. An interesting consequence of their results is that, in spite of positive results for Fisher consistency in binary classification problems [2, 19, 33], essentially only hinge-like losses are consistent for the 0–1 loss. We provide the extension of these results to the multiclass case.

Outline and discussion of our contributions. We build on this prior work to provide a unifying framework that relates statistical information measures, loss functions and generalized notions of entropy in the context of multiclass classification. To begin, we recall a generalization of f-divergences that applies to multiple distributions [11, 14], enumerating analogues of the positivity properties, data-processing inequalities and discrete approximation available in the binary case, as multiway f-divergences may be unfamiliar and they motivate our approach (Section 2). We begin our main contributions in Section 3, where we establish a correspondence between loss functions L, generalized entropy on discrete distributions [13] and multi-way f-divergences. To make this precise, define the probability simplex $\Delta_k := \{p \in \mathbb{R}^k_+ \mid \mathbf{1}^T p = 1\}$. Let $\pi \in \Delta_k$ be a prior distribution on the class label Y and $\widetilde{\pi}(x) \in \Delta_k$ be the posterior probabilities for Y conditional on the observation $X = x$. For concave $H : \Delta_k \to \mathbb{R}$, DeGroot [9] defines the *information* associated with the experiment (X, Y) as

$$I_H(X; \pi) := H(\pi) - \mathbb{E}\big[H\big(\widetilde{\pi}(X)\big)\big]. \tag{3}$$

The notion of H as a generalized entropy is clear here, as I is the gap between prior and posterior entropy and is always nonnegative. In this context, the value $H(\pi)$ measures the *uncertainty* of the experimenter (in some appropriate units) about the unknown class label y when his or her prior belief over y is π, so I is the gap between prior and posterior uncertainty [9].

To relate this type of entropy to loss functions, recall the well-known result [9, 13] that any loss $L : \mathbb{R}^k \times [k] \to \mathbb{R} \cup \{+\infty\}$ induces an entropy $H_L : \Delta_k \to \mathbb{R}$, also called the *pointwise Bayes risk* [11, 13, 24, 32], via

$$H_L(\pi) = \inf_{\alpha \in \mathbb{R}^k} \left\{ \sum_{i=1}^k \pi_i L(\alpha, i) \right\} - \mathbf{I}\{\pi \in \Delta_k\}, \tag{4}$$

where $\mathbf{I}\{\cdot\}$ is $+\infty$ if its argument is false (we drop this indicator in the future, defining H_L implicitly on Δ_k). We show an inverse construction, providing an explicit and constructable mapping from any concave function H to a loss L inducing H as the pointwise Bayes risk (4), where for each y the loss $\alpha \mapsto L(\alpha, y)$ is convex. In Section 3.2, we also develop the natural connections between these generalized entropies H and classification calibration [2, 19, 28, 33], in that our explicit L is generally calibrated.

In Section 4, we address the comparison of loss functions—building off of Nguyen et al.'s approach in the binary case [20]—and present our main results on consistency of joint selection of quantizer (data representation) q and discriminant γ. Using our correspondence between concave H, losses L, and f-divergences, we characterize the pairs of losses $L^{(1)}$ and $L^{(2)}$ for which equivalent quantizers and discriminants minimize the quantized risk (2) in the sense that there is a continuous concave h with $h(0) = 0$ such that

$$R_{L^{(2)}}(\gamma \mid \mathsf{q}) - \inf_{\gamma, \mathsf{q} \in \mathsf{Q}} R_{L^{(2)}}(\gamma \mid \mathsf{q}) \le h\Big(R_{L^{(1)}}(\gamma \mid \mathsf{q}) - \inf_{\gamma, \mathsf{q} \in \mathsf{Q}} R_{L^{(1)}}(\gamma \mid \mathsf{q})\Big)$$

for any γ and $\mathsf{q} \in \mathsf{Q}$. Another way to understand our results is as providing insight into classification calibration when the Bayes act (i.e., optimal discriminant γ) does not belong to the class of functions the statistician may choose in a classification problem. A substantial challenge for and criticism of the line of work on classification calibration and surrogate risk consistency [2, 19, 28, 33] is that the results say little for restricted families of classifiers. In this context, a corollary of our main contribution is as follows. The loss $L^{(1)}$ is *calibrated* [2, 28, 33] for $L^{(2)}$ if for any distribution P on $X \times Y$ and sequence $\gamma_n : \mathcal{X} \to \mathbb{R}^k$, $R_{L^{(1)}}(\gamma_n) \to \inf_\gamma R_{L^{(2)}}(\gamma)$ implies $R_{L^{(1)}}(\gamma_n) \to \inf_\gamma R_{L^{(2)}}(\gamma)$. Now, consider a collection Q of functions $\mathsf{q} : \mathcal{X} \to \mathcal{Z}$ for some set $\mathcal{Z}$, and then define the class of functions

$$\mathcal{G}(\mathsf{Q}) := \{\gamma \circ \mathsf{q} \mid \mathsf{q} \in \mathsf{Q} \text{ and } \gamma : \mathcal{Z} \to \mathbb{R}^k \text{ is measurable}\}.$$

Translated to this scenario, our main results—Theorems 1 and 2—imply the following.

COROLLARY 1. *Assume that the loss $L^{(1)}$ is calibrated for $L^{(2)}$ and let $H_i = H_{L^{(i)}}$ denote the associated pointwise Bayes risk* (4). *Then*

$$R_{L^{(1)}}(g_n) \to \inf_{g \in \mathcal{G}(\mathsf{Q})} R_{L^{(1)}}(g) \quad \textit{implies} \quad R_{L^{(2)}}(g_n) \to \inf_{g \in \mathcal{G}(\mathsf{Q})} R_{L^{(2)}}(g) \tag{5}$$

for any collection Q of mappings $\mathcal{X} \to \mathcal{Z}$, any set $\mathcal{Z}$, any distribution P on $\mathcal{X} \times \{1, \dots, k\}$, and sequence $g_n \in \mathcal{G}(\mathsf{Q})$ if and only if there exist $a > 0$, $b \in \mathbb{R}^k$, and $c \in \mathbb{R}$ such that $H_1(\pi) = aH_2(\pi) + b^T\pi + c$ for all $\pi \in \Delta_k$.

This corollary reposes on the connections we develop between losses, uncertainty measures and generalized f-divergences. Such measures of statistical information and divergence have been central to the design of communication and quantization schemes in signal processing [16, 18, 22, 30]; we characterize the divergence measures that, when optimized, yield optimal quantizers and detectors. We also provide a result showing when empirical minimization of a surrogate risk is consistent for the desired (original) risk.

A number of researchers have studied the connections between divergence measures and risk for binary and multicategory experiments; these point to the results we present. Indeed, [6] shows that if a quantizer q_1 induces class-conditional distributions with larger divergence than those induced by q_2, then there are prior probabilities such that q_1 allows tests with lower probability of error than q_2. Liese and Vajda [17] give a broad treatment of f-divergences, using their representation as the difference between prior and posterior risk in a binary experiment [21] to derive a number of their properties; see also the paper [24]. García-García and Williamson [11] show how multidistribution f-divergences [14] arise naturally in the context of multiclass classification problems as the gap between prior and posterior risk in classification, as in the work [17]. In the binary case, these results elucidate Nguyen et al.'s characterization of Fisher consistency for quantization and binary classification [20]. We pursue this line of research to draw the connections between Fisher consistency, information measures, multiclass classification, surrogate losses and divergences.

NOTATION. We let $\mathbf{0}$ and $\mathbf{1}$ denote the all-zeros and all-ones vectors, respectively. For a vector or collection of objects, we define $t_{1:m} = \{t_1, \dots, t_m\}$. The indicator function $\mathbf{I}\{\cdot\}$ is $+\infty$ if its argument is false, 0 otherwise, while $1\{\cdot\}$

is 1 if its argument is true, 0 otherwise. We let $\Delta_k = \{v \in \mathbb{R}_+^k : \mathbf{1}^T v = 1\}$ denote the probability simplex in $\mathbb{R}^k$. For $m \in \mathbb{N}$, we set $[m] = \{1, \dots, m\}$. We let $\operatorname{aff} A = \{\sum_{i=1}^m \lambda_i x_i \mid \lambda^T \mathbf{1} = 1, x_i \in A, m \in \mathbb{N}\}$ denote the affine hull of a set A, and $\operatorname{rel\,int} A$ denotes the interior of A relative to $\operatorname{aff} A$. We let $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ and $\underline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$. For $f : \mathbb{R}^k \to \overline{\mathbb{R}}$, we let $\operatorname{epi} f = \{(x, t) : f(x) \le t\}$ denote the epigraph of f. We say a convex function f is closed if $\operatorname{epi} f$ is a closed set, though we abuse notation and say that a concave f is closed if $\operatorname{epi}(-f)$ is closed. For a convex function $f : \mathbb{R}^k \to \overline{\mathbb{R}}$, we say that f is *strictly convex at a point* $t \in \mathbb{R}^k$ if for all $\lambda \in (0, 1)$ and $t_1, t_2 \ne t$ such that $t = \lambda t_1 + (1 - \lambda) t_2$ we have $f(t) < \lambda f(t_1) + (1 - \lambda) f(t_2)$. The (Fenchel) conjugate of a function $f : \mathbb{R}^k \to \overline{\mathbb{R}}$ is

$$f^*(s) = \sup_{t \in \mathbb{R}^k} \{s^T t - f(t)\}. \tag{6}$$

For any f, the conjugate f^* is closed convex (see [15], Chapter X). For measures ν and μ, we let $d\nu/d\mu$ denote the Radon–Nikodym derivative of ν with respect to μ. For random variables X_n, we say $X_n \overset{L_p}{\to} X_\infty$ if $\mathbb{E}[|X_n - X_\infty|^p] \to 0$.

2. Multidistribution *f*-divergences.

Divergence measures for probability distributions have significant statistical, decision- and information-theoretic applications, including in optimal testing, minimax rates of convergence and the design of communication schemes [1, 8, 16, 22]. Central to this work is the f-*divergence*, introduced by Ali and Silvey [1] and Csiszár [8] (see [17] for an overview). Given distributions P, Q defined on a common set $\mathcal{X}$, a closed convex function $f : [0, \infty) \to \overline{\mathbb{R}}$ satisfying $f(1) = 0$, and any measure μ dominating P and Q, the f-divergence between P and Q is

$$D_f(P \| Q) := \int_{\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right) q(x)\, d\mu(x) = \int f\left(\frac{dP}{dQ}\right) dQ. \tag{7}$$

Here, $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$ denote the densities of P and Q, respectively, and the value $u f(t/u)$ is defined appropriately for $t = 0$ and $u = 0$ (e.g., [17]). A number of classical divergence measures arise out of the f-divergence; taking $f(t) = t \log t$, $f(t) = \frac{1}{2}(\sqrt{t} - 1)^2$ or $f(t) = |t - 1|$ yields (resp.) the KL-divergence, squared Hellinger distance or total variation distance.

Central to our study of multiway hypothesis testing and classification is an understanding of relationships between multiple distributions. We use the following generalization [11, 14] of the f-divergence to multiple distributions.

DEFINITION 2.1. Let $P_1, \dots, P_k$ be probability distributions on a common σ-algebra $\mathcal{F}$ over a set $\mathcal{X}$. Let $f : \mathbb{R}_+^{k-1} \to \overline{\mathbb{R}}$ be a closed convex function satisfying $f(\mathbf{1}) = 0$. Let μ be any σ-finite measure such that $P_i \ll \mu$ for all i, and let $p_i =$

$dP_i/d\mu$. The *f-divergence* between $P_1, \dots, P_{k-1}$ and P_k is

$$D_f(P_1, \dots, P_{k-1} \| P_k) := \int f\left(\frac{p_1(x)}{p_k(x)}, \dots, \frac{p_{k-1}(x)}{p_k(x)}\right) p_k(x)\, d\mu(x). \tag{8}$$

We must specify the value of the integrand (8) when $p_k(x) = 0$. In this case, the function $\widetilde{f} : \mathbb{R}_+^k \to \overline{\mathbb{R}}$ defined, for an arbitrary $t' \in \operatorname{rel\,int\,dom} f$, by

$$\widetilde{f}(t, u) = \begin{cases} uf(t/u) & \text{if } u > 0, \\ \lim_{s \to 0} sf(t' - t + t/s) & \text{if } u = 0, \\ +\infty & \text{otherwise,} \end{cases} \tag{9}$$

is a closed convex function with value independent of t'; $\widetilde{f}$ is the closure of the *perspective* function $\mathbb{R}_+ \times \mathbb{R}^k \ni (u, t) \mapsto uf(t/u)$ of f (see [15], Proposition IV.2.2.2). Any time we consider the perspective we implicitly treat it as its closure (9).

We now enumerate a few properties of multiway f-divergences, showing how they naturally generalize classical binary f-divergences. We focus on basic properties that are useful for our further results on Bayes risk, classification and hypothesis testing and that parallel results in the binary case (7): they are well defined, have continuity properties with respect to discrete approximations and satisfy data-processing inequalities. While Györfi and Nemetz's original work [14] essentially contains the results, we carefully address infinite values [the closure (9)] and strict convexity, and we use them as definitional building blocks; we provide all proofs in the supplement [10], Section 10.

As our first step, we note that Definition 2.1 is independent of the base measure μ. (See [10], Section 10.1, for a proof generalizing [14], Corollary 1.)

LEMMA 2.1. *In expression* (8), *the value of the divergence does not depend on the choice of the dominating measure μ. Moreover,*

$$D_f(P_1, \dots, P_{k-1} \| P_k) \geq 0.$$

The inequality is strict if f is strictly convex at $\mathbf{1}$ *and the P_i are not identical.*

Given the importance of quantization to come, we now consider discrete approximations to the divergence. For an at most countable partition $\mathcal{P}$ of $\mathcal{X}$ into measurable sets A, we define the partitioned f-divergence

$$D_f(P_1, \dots, P_{k-1} \| P_k \mid \mathcal{P}) = \sum_{A \in \mathcal{P}} f\left(\frac{P_1(A)}{P_k(A)}, \dots, \frac{P_{k-1}(A)}{P_k(A)}\right) P_k(A).$$

As in the binary case [17, 31], we have the following approximability result generalizing [14], Theorem 6, to possibly infinite integrands: quantizers give arbitrarily good approximations to f-divergences (see [10], Section 10.2, for a proof).

PROPOSITION 1. *If f is a closed convex function with $f(\mathbf{1}) = 0$, then*

$$D_f(P_1, \dots, P_{k-1} \| P_k) = \sup_{\mathcal{P}} D_f(P_1, \dots, P_{k-1} \| P_k \mid \mathcal{P})$$

where the supremum is over finite partitions of $\mathcal{X}$.

In the binary case, f-divergences satisfy *data processing inequalities* [7, 8, 17], which state that processing or transforming an observation X drawn from the distributions P_1, P_2, decreases the divergence between them. To formalize this, recall that Q is a *Markov kernel* from a set $\mathcal{X}$ to $\mathcal{Z}$ if $Q(\cdot \mid x)$ is a probability distribution on $\mathcal{Z}$ for each $x \in \mathcal{X}$, and for each measurable $A \subset \mathcal{Z}$, the mapping $x \mapsto Q(A \mid x)$ is measurable. The following general data processing inequality shows that this holds in the multidistribution case as well, generalizing [14], Theorem 4, to possibly infinite f and the closure (9); we provide a proof in [10], Section 10.3.

PROPOSITION 2. *Let f be closed convex with $f(\mathbf{1}) = 0$, Q be a Markov kernel from $\mathcal{X}$ to $\mathcal{Z}$, and define the marginals $Q_P(A) = \int_{\mathcal{X}} Q(A \mid x)\, dP(x)$. Then*

$$D_f(Q_{P_1}, \dots, Q_{P_{k-1}} \| Q_{P_k}) \le D_f(P_1, \dots, P_{k-1} \| P_k).$$

This proposition is related to the relationships between risk, information and quantization we develop in Sections 3 and 4. Defining a *quantizer* q to be a measurable mapping $\mathsf{q} : \mathcal{X} \to \mathcal{Z}$ between measurable spaces $\mathcal{X}$ and $\mathcal{Z}$, the *quantized f divergence* is

$$D_f(P_1, \dots, P_{k-1} \| P_k \mid \mathsf{q}) := \sup_{\mathcal{P}} \sum_{A \in \mathsf{q}^{-1}(\mathcal{P})} f\left(\frac{P_1(A)}{P_k(A)}, \dots, \frac{P_{k-1}(A)}{P_k(A)}\right) P_k(A),$$

where $\mathcal{P}$ ranges over finite partitions of $\mathcal{Z}$ and $\mathsf{q}^{-1}(\mathcal{P}) = \{\mathsf{q}^{-1}(B) \mid B \in \mathcal{P}\}$. Proposition 2 immediately yields that quantization reduces information: the indicator $Q(A \mid x) = 1\{\mathsf{q}(x) \in A\}$ defines a Markov kernel, yielding the following.

COROLLARY 2. *Let f be closed convex, satisfy $f(\mathbf{1}) = 0$ and q be a quantizer of $\mathcal{X}$. Then*

$$D_f(P_1, \dots, P_{k-1} \| P_k \mid \mathsf{q}) \le D_f(P_1, \dots, P_{k-1} \| P_k).$$

We also see that if q_1 and q_2 are quantizers of $\mathcal{X}$, and q_1 induces a finer partition of $\mathcal{X}$ than q_2, meaning that for $x, x' \in \mathcal{X}$ the equality $\mathsf{q}_1(x) = \mathsf{q}_1(x')$ implies $\mathsf{q}_2(x) = \mathsf{q}_2(x')$, we have

$$D_f(P_1, \dots, P_{k-1} \| P_k \mid \mathsf{q}_2) \le D_f(P_1, \dots, P_{k-1} \| P_k \mid \mathsf{q}_1).$$

This type of ordering is central to this work: any multiclass loss L induces a unique f-divergence, and consistency of discriminants $\gamma : \mathcal{X} \to \mathbb{R}^k$ for a loss L after quantization is intimately tied to the preservation (and relative ordering) of information as related to the quantized risk (2).

3. Risks, information measures and f-divergences. Having reviewed the basic properties of f-divergences, we turn to a more detailed look at their relationships with multiway hypothesis tests, multiclass classification, generalized entropies and statistical information relating multiple distributions. We build a correspondence between these that parallels that for binary experiments and classification problems [17, 20, 24].

We first recapitulate the probabilistic model for classification and Bayesian hypothesis testing problems from the Introduction. We have a prior $\pi \in \Delta_k$ and probability distributions $P_1, \dots, P_k$ defined on a set $\mathcal{X}$. The coordinate $Y \in [k]$ is drawn according to a multinomial with probabilities π, and conditional on $Y = y$, we draw $X \sim P_y$. Following DeGroot [9], we refer to this as an *experiment*. Associated with any experiment is a family of information as follows. Let $\widetilde{\pi}$ be the posterior distribution on Y given observation $X = x$, $\widetilde{\pi}_i(x) = \pi_i \, dP_i(x)/(\sum_{j=1}^k \pi_j \, dP_j(x))$. Given any closed concave $H : \mathbb{R}_+^k \to \mathbb{R}$, which we refer to as *generalized entropy* (see [13], Section 3.3; DeGroot [9] calls H an uncertainty function), the *information* associated with the experiment is the reduction of entropy (uncertainty) from prior to posterior (3),

$$I_H(X; \pi) = H(\pi) - \mathbb{E}\big[H\big(\widetilde{\pi}(X)\big)\big].$$

The expectation is taken over $X \sim \sum_{i=1}^k \pi_i P_i$. That $I_H(X; \pi) \geq 0$ is immediate by concavity; DeGroot [9], Theorem 2.1, shows that $I_H(X; \pi) \geq 0$ for all distributions $P_1, \dots, P_k$ and priors π if and only if H is concave on Δ_k.

In this section, we develop equivalence results between multiclass classification losses and risk, multiway f-divergences and entropy measures. Concretely, consider $L : \mathbb{R}^k \times [k] \to \mathbb{R}$, and recall the risk (1), defined as $R_L(\gamma) = \mathbb{E}[L(\gamma(X), Y)]$, where $\gamma \in \Gamma$, the set of measurable functions $\gamma : \mathcal{X} \to \mathbb{R}^k$. As in equation (4) in the Introduction, each loss L induces the entropy H_L on Δ_k via $H_L(\pi) = \inf_{\alpha \in \mathbb{R}^k} \sum_{i=1}^k \pi_i L(\alpha, i)$, also called the pointwise Bayes risk [11, 13, 24, 32]. In Section 3.1, we give an explicit inverse mapping showing how each generalized entropy H is induced by (at least one) *convex* loss function L, that is, $L(\cdot, i)$ is convex for each i. In Section 3.2, we illustrate consistency properties the entropy H implies about the convex loss L inducing it. We connect these results in Section 3.3 with multiway f-divergences. For any loss L and associated entropy/Bayes risk H_L, for all $\pi \in \Delta_k$ there exists a convex function $f_{L,\pi} : \mathbb{R}_+^{k-1} \to \mathbb{R}$ with $f_{L,\pi}(\mathbf{1}) = 0$ such that the gap between the prior Bayes L-risk—the best expected loss attainable without observing X—and the posterior Bayes risk $\inf_\gamma R_L(\gamma)$ is

$$H_L(\pi) - \inf_{\gamma \in \Gamma} R_L(\gamma) = H_L(\pi) - \mathbb{E}\big[H_L\big(\widetilde{\pi}(X)\big)\big] = D_{f_{L,\pi}}(P_1, \dots, P_{k-1} \| P_k)$$

(see [11, 13]). The inverse direction is new, and given any closed convex function $f : \mathbb{R}_+^{k-1} \to \mathbb{R}$ with $f(\mathbf{1}) = 0$, we construct convex losses $L(\cdot, i)$, an associated generalized entropy H_L, and prior $\pi = \mathbf{1}/k \in \Delta_k$ satisfying

$$D_f(P_1, \dots, P_{k-1} \| P_k) = \inf_{\alpha \in \mathbb{R}^k} \sum_{i=1}^k \pi_i L(\alpha, i) - \inf_{\gamma \in \Gamma} R_L(\gamma).$$

3.1. *Generalized entropies and losses.* We construct a natural bidirectional mapping between losses and generalized entropies, giving a few examples to illustrate. For any loss $L : \mathbb{R}^k \times [k] \to \overline{\mathbb{R}}$, the construction (4) of H_L yields a closed concave function, as H_L is the infimum of linear functionals of π. It is thus a generalized entropy [13] (or uncertainty function [9]), and the gap $H_L(\pi) - \mathbb{E}[H_L(\widetilde{\pi}(X))]$ between prior and posterior entropy is nonnegative. The following two examples with zero-one loss are illustrative.

EXAMPLE 1 (Zero-one loss). Consider the zero one loss

$$L^{\mathrm{zo}}(\alpha, y) = 1\{\alpha_y \le \alpha_j \text{ for some } j \ne y\},$$

where $y \in [k]$. Then we have

$$H_{L^{\mathrm{zo}}}(\pi) = \inf_\alpha \left\{ \sum_{i=1}^k \pi_i 1\{\alpha_i \le \alpha_j \text{ for some } j \ne i\} \right\} = 1 - \max_j \pi_j.$$

This generalized entropy is concave, nonnegative and satisfies $H_{L^{\mathrm{zo}}}(\pi) = 0$ if and only if $\pi = e_i$ for a standard basis vector e_i.

EXAMPLE 2 (Cost-weighted classification). In some scenarios, we allow different costs for classifying certain classes y as others; for example, it may be less costly to misclassify a benign tumor as cancerous than the opposite. In this case, we use a matrix $C = [c_{yi}]_{y,i=1}^k \in \mathbb{R}_+^{k \times k}$, where $c_{yi} \ge 0$ is the cost for classifying an observation of class y as class i (i.e., assigning $X \sim P_i$ instead of P_y in the experiment). We assume $c_{yy} = 0$ for each y and define

$$L^{\mathrm{cw}}(\alpha, y) = \max_i \left\{ c_{yi} \mid \alpha_i = \max_j \alpha_j \right\}, \qquad \alpha \in \mathbb{R}^k, \tag{10}$$

the maximal loss for those indices of α attaining $\max_j \alpha_j$. Let $C = [c_1 \ \cdots \ c_k]$ be the column representation of C. If $c_y^T \pi = \min_l c_l^T \pi$, then by choosing any α such that $\alpha_y > \alpha_j$ for all $j \ne y$, we have

$$H_{L^{\mathrm{cw}}}(\pi) = \inf_\alpha \left\{ \sum_{y=1}^k \pi_y \max_i \left\{ c_{yi} \mid \alpha_i = \max_j \alpha_j \right\} \right\} = \min_l \pi^T c_l.$$

The entropy $H_{L^{\mathrm{cw}}}$ is concave, nonnegative and satisfies $H_{L^{\mathrm{cw}}}(e_i) = 0$ for standard basis vectors e_i; Example 1 corresponds to $C = \mathbf{1}\mathbf{1}^T - I_{k \times k}$.

The forward mapping (4) from losses L to entropy H_L is straightforward, though it is many-to-one. Using convex duality and conjugacy arguments, we can show an inverse mapping. This construction is new, though precursors for proper scoring rules and predictions in the probability simplex exist (cf. [13] or [12], Theorem 2); these *characterize* proper scoring rules, but it is not always clear how to generate *convex* losses from these. Before stating the proposition, we recall the definition (6) of the Fenchel conjugate $f^*(s) = \sup_t \{s^T t - f(t)\}$.

PROPOSITION 3. *For any closed concave* $H : \Delta_k \to \underline{\mathbb{R}}$, *the losses*

$$L(\cdot, i) : \mathbb{R}^k \to \overline{\mathbb{R}}, \qquad L(\alpha, i) = -\alpha_i + (-H)^*(\alpha), \tag{11}$$

$i \in \{1, \dots, k\}$, *are closed*, *convex and satisfy the equality* (4) *that* $H \equiv H_L$.

PROOF. Standard Fenchel conjugacy relationships [15] imply

$$H(\pi) = \inf_{\alpha \in \mathbb{R}^k} \{-\pi^T \alpha + (-H)^*(\alpha)\} \qquad \text{where } (-H)^*(\alpha) = \sup_{\pi \in \Delta_k} \{\alpha^T \pi + H(\pi)\}.$$

Defining $L(\alpha, i) = -\alpha_i + (-H)^*(\alpha)$ for $i = 1, \dots, k$, we can write

$$\begin{aligned} H(\pi) &= \inf_{\alpha \in \mathbb{R}^k} \{-\pi^T \alpha + (-H)^*(\alpha)\} \\ &= \inf_{\alpha \in \mathbb{R}^k} \{-\pi^T \alpha + \pi^T \mathbf{1} \cdot (-H)^*(\alpha)\} \\ &= \inf_{\alpha \in \mathbb{R}^k} \left\{\sum_{i=1}^k \pi_i L(\alpha, i)\right\}. \end{aligned}$$

□

Proposition 3 shows that associated with every concave entropy defined on the simplex, there is at least one set of *convex* loss functions $L(\cdot, i)$ generating the entropy via the infimal representation (4), and there is thus a mapping from loss functions to entropies and from entropies to *convex* losses: given any loss L, we may construct a convex loss L^{cvx} with $H_L = H_{L^{\mathrm{cvx}}}$. The mapping from entropies H to loss functions generating H as in (4) is one-to-many, as any losses $L^{(1)}$ and $L^{(2)}$ with the same range satisfy $H_{L^{(1)}} = H_{L^{(2)}}$.

3.2. *Surrogate risk consistency and generalized entropies.* Our construction (11) of loss functions is a somewhat privileged construction, as it often yields desirable properties of the convex loss function itself, especially as related to the nonconvex zero-one loss. Indeed, it is often the case that the convex loss L so generated is Fisher consistent; to make this explicit, we recall the following definition [28, 33].

DEFINITION 3.1. Let $L : \mathbb{R}^k \times [k] \to \overline{\mathbb{R}}$. Then L is *classification calibrated for the zero-one loss* if for any $\pi \in \Delta_k$ and i^* such that $\pi_{i^*} < \max_j \pi_j$,

$$\inf_{\alpha \in \mathbb{R}^k} \left\{ \sum_{i=1}^k \pi_i L(\alpha, i) \right\} < \inf_{\alpha \in \mathbb{R}^k} \left\{ \sum_{i=1}^k \pi_i L(\alpha, i) : \alpha_{i^*} \geq \max_j \alpha_j \right\}. \tag{12}$$

Given a matrix $C \in \mathbb{R}_+^{k \times k}$ as in Example 2, L is *classification calibrated for the cost matrix* C if for any $\pi \in \Delta_k$ and any i^* with $c_{i^*}^T \pi > \min_j c_j^T \pi$,

$$\inf_{\alpha \in \mathbb{R}^k} \left\{ \sum_{i=1}^k \pi_i L(\alpha, i) \right\} < \inf_{\alpha \in \mathbb{R}^k} \left\{ \sum_{i=1}^k \pi_i L(\alpha, i) : \alpha_{i^*} \geq \max_j \alpha_j \right\}. \tag{13}$$

Tewari and Bartlett [28], Theorem 2, and Zhang [33], Theorem 3, show the importance of Definition 3.1: let $R(\gamma)$ be the zero-one or cost-weighted risk (Examples 1–2). If L is lower-bounded, then it is classification calibrated (with respect to zero-one or the cost-weighted loss) if and only if for any sequence $\gamma_n : \mathcal{X} \to \mathbb{R}^k$ and distribution $\mathbb{P}$ on $X \times Y$ we have *Fisher consistency*, that is,

$$R_L(\gamma_n) \to \inf_{\gamma \in \Gamma} R_L(\gamma) \quad \text{implies} \quad R(\gamma_n) \to \inf_{\gamma \in \Gamma} R(\gamma).$$

That is, classification calibration (with respect to zero-one-risk or the cost-weighted risk) is equivalent to surrogate risk consistency of the loss L. Because of the predominance of the zero-one loss in the literature, we use "classification calibration" without any qualification to mean "classification calibration with respect to zero-one loss."

We now show how—under minor restrictions on the generalized entropy function H—the construction (11) yields classification calibrated losses.

DEFINITION 3.2. A convex function $f : \mathbb{R}^k \to \overline{\mathbb{R}}$ is $(\lambda, \kappa, \|\cdot\|)$-*uniformly convex* over $C \subset \mathbb{R}^k$ if it is closed and for all $t \in [0, 1]$ and $x_1, x_2 \in C$,

$$\begin{aligned} &f\bigl(tx_1 + (1-t)x_2\bigr) \\ &\qquad \leq tf(x_1) + (1-t)f(x_2) - \frac{\lambda}{2} t(1-t)\|x_1 - x_2\|^\kappa \bigl[(1-t)^{\kappa-1} + t^{\kappa-1}\bigr]. \end{aligned}$$

We say, without qualification, that f is uniformly convex on C if $\operatorname{dom} f \supset C$ and there exist $\lambda > 0$, a norm $\|\cdot\|$, and constant $\kappa < \infty$ such that Definition 3.2 holds; we say f is uniformly concave if $-f$ is uniformly convex. Definition 3.2 is an extension of the usual notion of *strong convexity*, which holds when $\kappa = 2$, and is essentially a quantified notion of strict convexity.

With this definition, we have the following two propositions. These two propositions, whose proofs we provide in [10], Section 7, show that generalized entropies naturally give rise to classification calibrated loss functions; we provide examples of these results in Section 3.4 to come.

PROPOSITION 4. *Assume that H is closed concave, symmetric and has* $\operatorname{dom} H = \Delta_k$, *and let L have definition* (11). *Additionally, assume that* (a) *H is strictly concave, and $\inf_\alpha \sum_{i=1}^k \pi_i L(\alpha, i)$ is attained for all $\pi \in \Delta_k$, or* (b) *H is uniformly concave. Then L is classification calibrated.*

Even when H is not strictly concave, we can give classification calibration results. Indeed, recall Example 1, which showed that for the zero-one-loss, we have $H_L(\pi) = 1 - \max_j \pi_j$.

PROPOSITION 5. *Let $H(\pi) = 1 - \max_j \pi_j$. The loss* (11) *defined by $L(\alpha, i) = -\alpha_i + (-H)^*(\alpha)$ is classification calibrated. Moreover, we have for any $\pi \in \Delta_k$ and $\alpha \in \mathbb{R}^k$ that*

$$\sum_{i=1}^k \pi_i L(\alpha, i) - \inf_\alpha \sum_{i=1}^k \pi_i L(\alpha, i) \geq \frac{1}{k}\left(\sum_{i=1}^k \pi_i L^{\text{zo}}(\alpha, i) - \inf_\alpha \sum_{i=1}^k \pi_i L^{\text{zo}}(\alpha, i)\right).$$

3.3. *Divergences, risk and generalized entropies.* In this section, we show that f-divergences as in Definition 2.1 have a precise correspondence with generalized entropies and losses; [11] establish the correspondence between f-divergences and entropy/pointwise Bayes risk H; our results show the important link from f directly *to* the loss L. We begin as in equation (4) with a concave generalized entropy H and loss L satisfying $H(\pi) = \inf_{\alpha \in \mathbb{R}^k} \sum_{i=1}^k \pi_i L(\alpha, i)$; by Proposition 3 it is no loss of generality to assume this correspondence. Let Γ be the collection of measurable functions $\gamma : \mathcal{X} \to \mathbb{R}^k$. The posterior Bayes risk for L is

$$H_L(\pi, P_{1:k}) := \inf_{\gamma \in \Gamma} \int_{\mathcal{X}} \sum_{i=1}^k \pi_i L\big(\gamma(x), i\big)\, dP_i(x) = \mathbb{E}\big[H_L\big(\widetilde{\pi}(X)\big)\big], \tag{14}$$

where $\widetilde{\pi}(x)$ is the posterior distribution on Y conditional on $X = x$. The information measure (3) is thus the gap between the prior Bayes L-risk and posterior Bayes L-risk. We may then write

$$\begin{aligned}
&\inf_{\alpha \in \mathbb{R}^k} \sum_{i=1}^k \pi_i L(\alpha, i) - \inf_{\gamma \in \Gamma} R_L(\gamma) \\
&\quad = H_L(\pi) - H_L(\pi, P_{1:k}) \\
&\quad = I_{H_L}(X; \pi) \\
&\quad = \int_{\mathcal{X}} \sup_\alpha \left(H_L(\pi) - \sum_{i=1}^{k-1} \pi_i L(\alpha, i) \frac{dP_i}{dP_k} - \pi_k L(\alpha, k)\right) dP_k \\
&\quad = D_{f_{L,\pi}}(P_{1:k-1} \| P_k),
\end{aligned}$$

where the closed convex function $f_{L,\pi} : \mathbb{R}_+^{k-1} \to \overline{\mathbb{R}}$ has definition

$$f_{L,\pi}(t) := \sup_{\alpha \in \mathbb{R}^k} \left(H_L(\pi) - \sum_{i=1}^{k-1} \pi_i L(\alpha, i) t_i - \pi_k L(\alpha, k) \right). \tag{15}$$

As $f_{L,\pi}$ is the supremum of affine functions of its argument t, it is closed convex and $f_{L,\pi}(\mathbf{1}) = H_L(\pi) - H_L(\pi) = 0$. That is, equation (15) shows that given any loss L or generalized entropy H, the information measure $I_{H_L}(X; \pi)$, gap between prior and posterior L-risk and $f_{L,\pi}$-divergence between distributions $P_1, \dots, P_{k-1}$ and P_k are identical.

We can also give a converse result that shows that every f-divergence can be written as the gap between prior and posterior risks for a convex loss function. We first recall the result that $D_f(P_{1:k-1} \| P_k)$ is a statistical information (3) based on an generalized entropy H associated with f. Except for the closure operation, this result is known (see [11], Theorem 3).

PROPOSITION 6. *For closed convex* $f : \mathbb{R}^{k-1} \to \overline{\mathbb{R}}$ *with* $f(\mathbf{1}) = 0$, *let*

$$H(t_1, \dots, t_k) = -k t_k f\left(\frac{t_1}{t_k}, \dots, \frac{t_{k-1}}{t_k}\right),$$

where we implicitly use the closure of the perspective [*Definition* (9)]. *Then*

$$D_f(P_1, \dots, P_{k-1} \| P_k) = H(\mathbf{1}/k) - \mathbb{E}\big[H\big(\widetilde{\pi}(X)\big)\big],$$

where the prior $\pi = \mathbf{1}/k$ *and the expectation is taken according to* $\sum_i \pi_i P_i$.

By combining Propositions 3 and 6 with the infimal representation (4) of H_L, we immediately obtain the following corollary, which is our explicit construction of a closed convex loss from an f-divergence.

COROLLARY 3. *Let* $\pi^0 = \mathbf{1}/k$. *For any closed and convex function* $f : \mathbb{R}^{k-1} \to \overline{\mathbb{R}}$ *such that* $f(\mathbf{1}) = 0$, *the convex losses defined by*

$$L(\alpha, i) = -\alpha_i + \sup_{\pi \in \Delta_k} \left\{ \pi^T \alpha - k \pi_k f\left(\frac{\pi_1}{\pi_k}, \dots, \frac{\pi_{k-1}}{\pi_k}\right) \right\}$$

satisfy $f(t) = \sup_\alpha \{H_L(\pi^0) - \sum_{i=1}^{k-1} \pi_i^0 L(\alpha, i) t_i - \pi_k^0 L(\alpha, k)\}$, *equation* (15). *Additionally,*

$$D_f(P_{1:k-1} \| P_k) = \inf_{\alpha \in \mathbb{R}^k} \sum_{i=1}^k \pi_i^0 L(\alpha, i) - \inf_\gamma \mathbb{E}\big[L\big(\gamma(X), Y\big)\big],$$

where the expectation is over $Y \sim \pi^0$ *and* $X \sim P_y$ *conditional on* $Y = y$.

For binary classification problems, Nguyen et al. [20], Theorem 1, provide an explicit construction of a closed convex margin-based loss inducing the f-divergence as in equation (15); the binary case allows a complete characterization of all such convex functions, which appears difficult in the multiclass case.

Corollary 3, coupled with the information representation given by the f-divergence (15), shows the complete equivalence between f-divergences, loss functions L and entropies H. For any f-divergence, there exists a loss function L and prior $\pi = \mathbf{1}/k$ such that $D_f(P_{1:k-1} \| P_k) = H_L(\pi) - H_L(\pi, P_{1:k})$. Conversely, for any loss function L and prior π, there exists a multiway f-divergence such that the gap $H_L(\pi) - H_L(\pi, P_{1:k}) = D_f(P_{1:k-1} \| P_k)$.

3.4. *Examples of generalized entropies and loss correspondences.* To complement our general results, we illustrate the correspondence between (concave) generalized entropies and the loss construction (11) through several examples, using Propositions 4 and 5 to guarantee classification calibration.

EXAMPLE 3 (Zero-one loss, Example 1, continued). We use the generalized entropy $H(\pi) = 1 - \max_j \pi_j$ generated by the zero-one loss to derive a convex loss function L that gives the same entropy via the representation (4). The conjugate to $-H$ is

$$(16)\qquad (-H)^*(\alpha) = 1 + \max\left\{\alpha_{(1)} - 1, \frac{\alpha_{(1)} + \alpha_{(2)}}{2} - \frac{1}{2}, \dots, \frac{\sum_{i=1}^k \alpha_{(i)}}{k} - \frac{1}{k}\right\},$$

where $\alpha_{(1)} \ge \alpha_{(2)} \ge \cdots$ are the entries of $\alpha \in \mathbb{R}^k$ in sorted order (see [10], Section 7.5, for a proof). Then the convex "family-wise" loss, named for its similarity to family-wise error control in hypothesis tests,

$$L^{\mathrm{fw}}(\alpha, i) = 1 - \alpha_i + \max_{l \in \{1,\dots,k\}} \left\{\frac{1}{l}\sum_{j=1}^{l} \alpha_{(j)} - \frac{1}{l}\right\}$$

generates the same entropy $H_{L^{\mathrm{fw}}}$ and associated f-divergence as the zero-one loss. Moreover, Proposition 5 guarantees that L^{fw} is classification calibrated (Definition 3.1). It appears that the loss L^{fw} is a new convex classification-calibrated loss function.

Rather than reconsidering Example 2, which we do later in the context of showing that distinct convex losses can yield the same generalized entropy, we now consider the multiclass logistic loss. The loss does *not* correspond to the zero-one loss, but it generates Shannon entropy and information.

EXAMPLE 4 (Logistic loss and entropy). For $1 \le i \le k$, define $p_i(\alpha) = e^{\alpha_i} / \sum_{j=1}^k e^{\alpha_j}$. The multiclass logistic loss is then

$$L(\alpha, i) = -\log p_i(\alpha) = \log\left(\sum_{j=1}^{k} e^{\alpha_j - \alpha_i}\right) \qquad \text{for } 1 \le i \le k.$$

The entropy associated with the loss is the familiar Shannon entropy,

$$H_L(\pi) = \inf_{\alpha \in \mathbb{R}^k} \left\{ -\sum_{i=1}^{k} \pi_i \log p_i(\alpha) \right\} = -\sum_{i=1}^{k} \pi_i \log \pi_i. \tag{17}$$

The conjugacy calculation (11) (i.e., our inverse construction from H to loss L) to generate L also yields the multiclass logistic loss. That the multiclass logistic loss is calibrated for the zero-one loss [33], Section 4.4, is now immediate: the negative Shannon entropy is strongly convex over the simplex Δ_k (this is Pinsker's inequality [7], Chapter 17.3), so the fact that logistic loss and Shannon entropy are dual via equation (11) and Proposition 4 yield calibration. The information measure (3) associated with the logistic loss is the mutual information between the observation X and label Y. Indeed, we have

$$\begin{aligned} I_H(X; \pi) &= H(\pi) - \mathbb{E}\big[H\big(\widetilde{\pi}(X)\big)\big] \\ &= H(Y) - \int_{\mathcal{X}} H(Y \mid X = x)\, d\overline{P}(x) \\ &= H(Y) - H(Y \mid X) = I(X; Y), \end{aligned}$$

where H denotes the Shannon entropy, $\overline{P} = \sum_{i=1}^{k} \pi_i P_i$ and $I(X; Y)$ is the usual (Shannon) mutual information between X and Y.

We include one final example to show that in some instances, many different *convex* losses can yield the same generalized entropy H.

EXAMPLE 5 (Hinge losses). Define the pairwise multiclass hinge loss

$$L^{\text{hin}}(\alpha, i) = \sum_{j \neq i} [1 + \alpha_j]_+ + \mathbf{I}\{\mathbf{1}^T \alpha = 0\}.$$

We also consider the slight extension to weighted loss functions to address asymmetric losses of the form (10) from Example 2. In this case, given the loss matrix $C \in \mathbb{R}_+^{k \times k}$, we set

$$L^{\text{hin}}(\alpha, i) = \sum_{j=1}^{k} c_{ij} [1 + \alpha_j]_+ + \mathbf{I}\{\mathbf{1}^T \alpha = 0\}.$$

The loss $L(\alpha, i) = \sum_{j \neq i} c_{ij} [1 + \alpha_j - \alpha_i]_+$ yields a completely identical set of calculations without the restriction $\mathbf{1}^T \alpha = 0$, as it is invariant to shifts. A calculation (see [10], Section 7.6, for completeness) shows the generalized entropy (4) associated with the hinge loss with loss matrix $C = [c_1 \ \cdots \ c_k]$ is

$$H_{L^{\text{hin}}}(\pi) = \inf_{\alpha \in \mathbb{R}^k} \left\{ \sum_{i=1}^{k} \pi_i L^{\text{hin}}(\alpha, i) \right\} = k \min_{l} \pi^T c_l. \tag{18}$$

Such losses satisfy a number of classification calibration guarantees; we note one, essentially due to Zhang [33], Theorem 8. For completeness, we provide a proof in [10], Section 7.4.

OBSERVATION 1. Let $\phi : \mathbb{R} \to \mathbb{R}$ be any bounded below convex function, differentiable on $(-\infty, 0]$, with $\phi'(0) < 0$. Then $L(\alpha, y) = \sum_{i=1}^{k} c_{yi}\phi(-\alpha_i)$ is classification calibrated for the cost matrix C [Definition 3.1, equation (13)].

Taking $C = \mathbf{1}\mathbf{1}^T - I_{k\times k}$, we see that the hinge loss is calibrated for the zero-one loss (Example 1); taking arbitrary $C \in \mathbb{R}_+^{k\times k}$, the weighted hinge loss is calibrated for the cost matrix C. Even more, we have the following quantitative calibration guarantee in analogy with Proposition 5:

$$\begin{aligned}\sum_{i=1}^{k} \pi_i L^{\mathrm{hin}}(\alpha, i) &- \inf_{\alpha'} \sum_{i=1}^{k} \pi_i L^{\mathrm{hin}}(\alpha', i) \\ &\geq \sum_{i=1}^{k} \pi_i L^{\mathrm{cw}}(\alpha, i) - \inf_{\alpha'} \sum_{i=1}^{k} \pi_i L^{\mathrm{cw}}(\alpha', i)\end{aligned}$$

for all $\pi \in \Delta_k$ and $\alpha \in \mathbb{R}^k$, strengthening Observation 1. (We prove this as Lemma 7.9 [10], Section 7.7.) In the binary case, similar quantitative guarantees hold for *any* margin-based classification calibrated loss L for which $H_L = H_{L^{\mathrm{zo}}}$ (cf. [20], Lemma 2); we do not know if this extends to the multiclass case.

4. Comparison of loss functions. In Section 3, we demonstrated the correspondence between loss functions, generalized entropies, statistical information, f-divergences and (in restricted cases) classification calibration. These correspondences assume that decision makers have access to the entire observation X, which is often not the case; as noted in the Introduction, it is often beneficial to preprocess data to make it lower dimensional, communicate or store it efficiently or to improve statistical behavior. Thus, we now explore the impact quantization has on these concepts.

To motivate this further, consider that each of the family-wise loss L^{fw} of Example 3, logistic loss (Example 4) and any loss of the form $L(\alpha, y) = \sum_{i\neq y} \phi(-\alpha_i)$ for ϕ convex and decreasing with $\phi'(0) < 0$ (Example 5, Observation 1) is classification calibrated. This relates to one of the major criticisms of classification calibration: if the Bayes classifier (minimizer of risk over all functions $\mathcal{X} \to \mathbb{R}^k$) does not belong to the class of functions considered, classification calibration says little. In this context, we shed light on this issue by identifying losses that are consistent (calibrated) even with the additional selection of quantizer or data representation—a restriction of the class of possible functions as in the implication (5) in the Introduction.

4.1. *A model of quantization and experimental design.* We extend Nguyen et al.'s approach in the binary case [20] to the multiclass case by treating the design of an experiment or choice of data representation as a quantization problem, where a quantizer q maps the space $\mathcal{X}$ to a measurable space $\mathcal{Z}$. Then, for a loss L, prior $\pi \in \Delta_k$ on the label Y, and discriminant $\gamma : \mathcal{Z} \to \mathbb{R}^k$, we consider the quantized risk (2), which we recall is

$$R_{L,\pi}(\gamma \mid \mathsf{q}) := \mathbb{E}\big[L\big(\gamma\big(\mathsf{q}(X)\big), Y\big)\big].$$

Given class-conditional distributions $P_{1:k}$ (equivalently, hypotheses $H_i : P_i$ in the Bayesian testing setting) and collection Q of quantizers, our criterion is to choose the quantizer q that allows the best attainable risk. That is, we consider the quantized Bayes L-risk, defined as the infimum of the risk (2) over discriminants $\Gamma = \{\gamma : \mathcal{Z} \to \mathbb{R}^k\}$,

$$\inf_{\gamma \in \Gamma} R_{L,\pi}(\gamma \mid \mathsf{q}) = \int_{\mathcal{Z}} \inf_{\alpha} \sum_{i=1}^{k} \pi_i L(\alpha, i)\, dP_i^{\mathsf{q}}(z), \tag{19}$$

where $P^{\mathsf{q}}(A) = P(\mathsf{q}^{-1}(A))$ denotes the push-forward measure. The risk (19) measures the best attainable risk for a fixed choice of $\mathsf{q} \in \mathsf{Q}$; one thus seeks the design q giving the lowest quantized Bayes L-risk.

Whether for computational or analytic reasons, minimizing the loss (19) is often intractable; the zero-one loss L^{zo} (Example 1), for example, is non-convex and discontinuous. It is thus of interest to understand the asymptotic consequences of using a surrogate loss L in place of the desired loss (say L^{zo}) [2, 19, 28, 33], including the setting in which one incorporates further dimension reduction via the choice $\mathsf{q} \in \mathsf{Q}$. [20] introduce and study this problem for binary classification, giving a correspondence between f-divergences, loss functions and surrogate consistency with quantization. The consequences of using a surrogate for consistency of the resulting quantization and classification procedure in the multiclass case are *a-priori* unclear: we do not know when using such a surrogate can be done without penalty. To that end, we now characterize when two loss functions $L^{(1)}$ and $L^{(2)}$ provide equivalent criteria for choosing quantizers (experimental designs or data representations) according to the Bayes L-risk (19).

4.2. *Universal equivalence of loss functions.* Recalling our arguments in Section 3.3 that statistical information (the gap between prior and posterior risks) is a multiway f-divergence between distributions $P_1, \ldots, P_{k-1}$ and P_k, we give a quantized version of this construction. In analogy with the results of Section 3.3, the quantized statistical information is

$$\begin{aligned} I_{H_L}(X; \pi \mid \mathsf{q}) &:= H_L(\pi) - \mathbb{E}\big[H_L\big(\widetilde{\pi}\big(\mathsf{q}(X)\big)\big)\big] \\ &= \inf_{\alpha \in \mathbb{R}^k} \sum_{i=1}^{k} \pi_i L(\alpha, i) - \inf_{\gamma} R_{L,\pi}(\gamma \mid \mathsf{q}) \\ &= D_{f_{L,\pi}}(P_1, \ldots, P_{k-1} \| P_k \mid \mathsf{q}), \end{aligned} \tag{20}$$

where $H_L(\pi) = \inf_{\alpha\in\mathbb{R}^k} \sum_{i=1}^k \pi_i L(\alpha, i)$ as in (4), the convex function $f_{L,\pi}$ is defined as in expression (15) and does not depend on the quantizer q, and $\widetilde{\pi}(\mathsf{q}(X))$ denotes the posterior distribution on $Y \in [k]$ conditional on observing $\mathsf{q}(X)$. We extend Nguyen et al.'s notion of universal equivalence from the binary case, defining losses as equivalent if they induce the same ordering of quantizers q under the information measure (20).

DEFINITION 4.1. Loss functions $L^{(1)}$ and $L^{(2)}$ are *universally equivalent* for the prior π, denoted $L^{(1)} \overset{u}{\equiv}_\pi L^{(2)}$, if for any distributions $P_1, \dots, P_k$ on X and quantizers q_1 and q_2:

$$\begin{aligned} &I(X, \pi; H_{L^{(1)}} \mid \mathsf{q}_1) \le I(X, \pi; H_{L^{(1)}} \mid \mathsf{q}_2) \quad \text{if and only if} \\ &I(X, \pi; H_{L^{(2)}} \mid \mathsf{q}_1) \le I(X, \pi; H_{L^{(2)}} \mid \mathsf{q}_2). \end{aligned}$$

Definition 4.1 evidently is equivalent to the ordering condition

$$\begin{aligned} &\inf_\gamma R_{L^{(1)},\pi}(\gamma \mid \mathsf{q}_1) \le \inf_\gamma R_{L^{(1)},\pi}(\gamma \mid \mathsf{q}_2) \quad \text{if and only if} \\ &\inf_\gamma R_{L^{(2)},\pi}(\gamma \mid \mathsf{q}_1) \le \inf_\gamma R_{L^{(2)},\pi}(\gamma \mid \mathsf{q}_2), \end{aligned} \tag{21}$$

for all distributions $P_1, \dots, P_k$, on the quantized Bayes L-risk (19). This definition is somewhat stringent: losses are universally equivalent only if they induce the same quantizer ordering for all population distributions. If a quantizer q_1 is finer than q_2, all losses yield $I(X, \pi; H_L \mid \mathsf{q}_2) \le I(X, \pi; H_L \mid \mathsf{q}_1)$ by the data processing inequality (Corollary 2 of Section 2). The stronger equivalence notion is important for nonparametric classification settings in which the underlying distribution on (X, Y) is only weakly constrained and neither of a pair of quantizers $\mathsf{q}_1, \mathsf{q}_2 \in \mathsf{Q}$ is finer than the other.

Definition 4.1 and the representation (20) suggest that the entropy function H_L associated with the loss L through the infimal representation (4) and the f-divergence associated with L via the construction (15) are important for the equivalence of two loss functions. This is indeed the case. First, we have the following result on universal equivalence of loss functions based on their associated entropies.

THEOREM 1. *Let $L^{(1)}$ and $L^{(2)}$ be bounded below losses and $H_{L^{(1)}}$ and $H_{L^{(2)}}$ be the associated generalized entropies as in the construction* (4). *Then $L^{(1)}$ and $L^{(2)}$ are universally equivalent with respect to all priors π if and only if there exist $a > 0, b \in \mathbb{R}^k$, and $c \in \mathbb{R}$ such that for all $\pi \in \Delta_k$,*

$$H_{L^{(1)}}(\pi) = a H_{L^{(2)}}(\pi) + b^T \pi + c.$$

We can also characterize universal equivalence for a prior π.

THEOREM 2. *Let $\pi \in \Delta_k$ and as in Theorem 1 and let $L^{(1)}$ and $L^{(2)}$ be bounded below loss functions, with $f_\pi^{(1)}$ and $f_\pi^{(2)}$ the associated f-divergences as in the construction (15). Then $L^{(1)}$ and $L^{(2)}$ are universally equivalent with respect to the prior π if and only if there exist $a > 0, b \in \mathbb{R}^{k-1}$, and $c \in \mathbb{R}$ such that*

$$f_\pi^{(1)}(t) = af_\pi^{(2)}(t) + b^T t + c \qquad \textit{for all } t \in \mathbb{R}_+^{k-1}. \tag{22}$$

Nguyen et al. [20] prove Theorem 2 for binary classification problems ($k = 2$), using convex-conjugacy arguments. We outline our proofs (which apply for arbitrary k and so require a different set of tools) in Section 5.

4.3. *Consistency of empirical risk minimization.* A major application of these theorems is to show that certain nonconvex loss functions (such as the zero-one loss) are universally equivalent to convex loss functions, including variants of the hinge loss, by showing that their associated entropies are scalar multiples. As a first application of Theorems 1 and 2, however, we consider the Bayes consistency of empirical risk minimization for selecting a discriminant γ and quantizer q (in analogy with [20], Theorem 2). In this case, we receive a sample $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ and define the empirical risk

$$\widehat{R}_{L,n}(\gamma \mid \mathsf{q}) := \frac{1}{n}\sum_{i=1}^{n} L(\gamma(\mathsf{q}(X_i)), Y_i).$$

Now, let $\mathsf{Q}_1 \subset \mathsf{Q}_2 \subset \cdots \subset \mathsf{Q}$ be a nondecreasing collection of quantizers, indexed by sample size n, and similarly let $\Gamma_1 \subset \Gamma_2 \subset \cdots \subset \Gamma$ be a nondecreasing collection of discriminant functions, where we assume the collections satisfy the estimation and approximation error conditions

$$\begin{aligned} &\mathbb{E}\Big[\sup_{\gamma\in\Gamma_n,\mathsf{q}\in\mathsf{Q}_n} \big|\widehat{R}_{L,n}(\gamma \mid \mathsf{q}) - R_L(\gamma \mid \mathsf{q})\big|\Big] \le \varepsilon_n^{\mathrm{est}}, \\ &\inf_{\gamma\in\Gamma_n,\mathsf{q}\in\mathsf{Q}_n} R_L(\gamma \mid \mathsf{q}) - \inf_{\gamma\in\Gamma,\mathsf{q}\in\mathsf{Q}} R_L(\gamma \mid \mathsf{q}) \le \varepsilon_n^{\mathrm{app}}, \end{aligned} \tag{23}$$

where $\varepsilon_n^{\mathrm{est}} \to 0$ and $\varepsilon_n^{\mathrm{app}} \to 0$ as $n \to \infty$. Additionally, let R be the risk functional for the cost-weighted misclassification loss L^{cw} (Example 2), where $L^{\mathrm{cw}}(\alpha, y) = \max_i\{c_{yi} \mid \alpha_i = \max_j \alpha_j\}$. Then we have the following result.

THEOREM 3. *Assume the conditions (23) and that γ_n and q_n are approximate empirical L-risk minimizers satisfying*

$$\varepsilon_n^{\mathrm{opt}} := \mathbb{E}\Big[\widehat{R}_{L,n}(\gamma_n \mid \mathsf{q}_n) - \inf_{\gamma\in\Gamma_n,\mathsf{q}\in\mathsf{Q}_n} \widehat{R}_{L,n}(\gamma \mid \mathsf{q})\Big] \to 0 \qquad \textit{as } n \to \infty.$$

Let $R^\star(\mathsf{Q}) = \inf_{\gamma\in\Gamma,\mathsf{q}\in\mathsf{Q}} R(\gamma \mid \mathsf{q})$. If the loss L is classification calibrated and universally equivalent to the cost-weighted loss L^{cw}, then

$$R(\gamma_n \mid \mathsf{q}_n) - R^\star(\mathsf{Q}) \overset{L_1}{\to} 0.$$

Theorem 3 guarantees that under the estimation and approximation conditions (23), empirical risk minimization is consistent for minimizing the quantized Bayes risk whenever the loss L is classification calibrated and equivalent to the desired loss. The proof of Theorem 3 reposes on the following risk inequality, which may be of independent interest. The lemma is a consequence of the results on surrogate risk consistency for classification calibration [27, 28, 33] and our universal equivalence guarantees that exhibits the power of calibration and universal equivalence.

LEMMA 4.1. *Assume L is classification-calibrated and universally equivalent to the weighted misclassification loss L^{cw} with cost matrix $C \in \mathbb{R}_+^{k\times k}$. Then there exists a continuous concave function h with $h(0) = 0$ such that*

$$R(\gamma \mid \mathsf{q}) - \inf_{\gamma\in\Gamma,\mathsf{q}\in\mathsf{Q}} R(\gamma \mid \mathsf{q}) \le h\Big(R_L(\gamma \mid \mathsf{q}) - \inf_{\mathsf{q}\in\mathsf{Q}} R_L^\star(\mathsf{q})\Big).$$

With the choice $L(\alpha, y) = \sum_{i=1}^k c_{yi}[1+\alpha_i]_+ + \mathbf{I}\{\mathbf{1}^T\alpha = 0\}$ or $L(\alpha, y) = \sum_{i=1}^k c_{yi}[1+\alpha_i - \alpha_y]_+$, we may take $h(\varepsilon) = (1+\frac{1}{k})\varepsilon$, that is,

$$R(\gamma \mid \mathsf{q}) - \inf_{\gamma\in\Gamma,\mathsf{q}\in\mathsf{Q}} R(\gamma \mid \mathsf{q}) \le \left(1+\frac{1}{k}\right)\Big[R_L(\gamma \mid \mathsf{q}) - \inf_{\gamma\in\Gamma,\mathsf{q}\in\mathsf{Q}} R_L(\gamma \mid \mathsf{q})\Big].$$

Lemma 4.1 shows that the gap in *surrogate* risk provides a guaranteed upper bound on the true *cost-weighted* risk; in the case of the modified hinge losses of Example 5, this gap is linear. In the binary case, even stronger results are possible [20]—one may take $h(\varepsilon) = a\varepsilon$ (for some $a < \infty$) in Lemma 4.1 for any margin-based classification-calibrated loss universally equivalent to the 0–1 loss. This relies on the specific form any such binary convex loss must take (see equation (9) of [20]); our Examples 3 (the family-wise loss) and 5 show that fairly different-looking losses can be classification calibrated and universally equivalent to zero-one loss. We provide the proof of Lemma 4.1 in [10], Section 9.1. Theorem 3, which we prove in [10], Section 9.2, is then a consequence of this lemma and Theorem 1.

4.4. *Examples of universal equivalence.* In this section, we give several examples that build off of Theorems 1 and 2, showing that there exist convex losses that allow optimal joint design of quantizers (or measurement strategies) and discriminant functions, opening the way for potentially efficient convex optimization strategies. To that end, we give two hinge-like loss functions that are universally equivalent to the zero-one loss for all prior distributions π. We also give examples of classification calibrated loss functions that are not universally equivalent to the zero-one loss, although minimizing them without quantization yields Bayes-optimal classifiers.

EXAMPLE 6 (Cost-weighted losses). We return to Example 5, where we have $L^{\mathrm{hin}}(\alpha, i) = \sum_{j\neq i} c_{ij}[1+\alpha_j]_+ + \mathbf{I}\{\mathbf{1}^T\alpha = 0\}$. In this case, we have $H_{L^{\mathrm{hin}}}(\pi) =$

$k \min_l \pi^T c_l = k H_{L^{\rm cw}}(\pi)$, where $L^{\rm cw}$ denotes the cost-weighted misclassification error as in Example 2. Theorem 1 immediately guarantees that the (weighted) hinge loss is universally equivalent to the (weighted) 0–1 loss. The weighted hinge loss $L^{\rm hin}$ is also, as in Example 5, calibrated for the cost-weighted misclassification error.

EXAMPLE 7 (Max-type losses and zero-one loss). We return to Example 3 and let $L^{\rm fw}(\alpha, i) = 1 - \alpha_i + \max\{\alpha_{(1)} - 1, \frac{\alpha_{(1)}+\alpha_{(2)}}{2} - \frac{1}{2}, \ldots, \frac{\mathbf{1}^T\alpha}{k} - \frac{1}{k}\}$, the convex family-wise loss. By Example 3, the associated entropy is $H_{L^{\rm fw}}(\pi) = 1 - \max_j \pi_j = H_{L^{\rm zo}}$ for $\pi \in \Delta_k$, and Proposition 5 shows that $L^{\rm fw}$ is classification calibrated. We thus have that $L^{\rm fw}$ and the zero-one loss $L^{\rm zo}$ are universally equivalent by Theorems 1 and 2.

For our final example, we consider the logistic loss, which is classification calibrated but not universally equivalent to the zero-one loss.

EXAMPLE 8 (Logistic loss). The loss $L^{\log}(\alpha, i) = \log(\sum_{j=1}^k e^{\alpha_j - \alpha_i})$ has (Shannon) entropy $H(\pi) = -\sum_{i=1}^k \pi_i \log \pi_i$, as in Example 4. There are no a, b, c such that $H_{L^{\rm zo}}(\pi) = 1 - \max_j \pi_j = a H_{L^{\log}}(\pi) + b^T\pi + c$ for all $\pi \in \Delta_k$. Theorem 1 shows that the logistic loss is not universally equivalent to the zero-one loss. That is, in spite of its classification calibration, there are distributions $P_1, \ldots, P_k$, a collection Q of quantizers $\mathcal{X} \to \mathcal{Z}$, and a sequence $\gamma_n : \mathcal{Z} \to \mathbb{R}^k$ such that $R_{L^{\log}}(\gamma_n \mid \mathsf{q}_n) \to \inf_{\gamma, \mathsf{q} \in \mathsf{Q}} R_{L^{\log}}(\gamma \mid \mathsf{q})$, but $R_{L^{\rm zo}}(\gamma_n \mid \mathsf{q}_n) \not\to \inf_{\gamma, \mathsf{q} \in \mathsf{Q}} R_{L^{\rm zo}}(\gamma \mid \mathsf{q})$.

5. Proof of the Theorems 1 and 2. The remainder of the main body of the paper consists of the major parts of our arguments for Theorems 1 and 2. We divide the proof of the theorems into two parts. The "if" part is straightforward; the "only if" is substantially more complex.

Proof (*if direction*). We give the proof for Theorem 2; that for Theorem 1 is identical. Assume that $\operatorname{dom} f_\pi^{(1)} = \operatorname{dom} f_\pi^{(2)}$ and there exist $a > 0, b \in \mathbb{R}^{k-1}$, and $c \in \mathbb{R}$ such that equation (22) holds. By Definition 2.1 of multiway f-divergences, for any quantizer q, we have

$$D_{f_\pi^{(1)}}(P_1, \ldots, P_{k-1} \| P_k \mid \mathsf{q}) = a D_{f_\pi^{(2)}}(P_1, \ldots, P_{k-1} \| P_k \mid \mathsf{q}) + b^T \mathbf{1} + c,$$

as $\int_{\mathcal{X}} dP_i = 1$. Applying the relationship (20), we obtain

$$I(X, \pi; H_{L^{(1)}} \mid \mathsf{q}) = a I(X, \pi; H_{L^{(2)}} \mid \mathsf{q}) + b^T \mathbf{1} + c.$$

As $a > 0$, the universal equivalence of $L^{(1)}$ and $L^{(2)}$ follows immediately.

We turn to the "only if" part of the proofs of Theorems 1 and 2. A roadmap is as follows: we first define what we call *order equivalence* of convex functions, which is related to the equivalence of f-divergences and generalized entropies (Definition 5.1). Then, for any two loss functions $L^{(1)}$ and $L^{(2)}$ that are universally equivalent, we show that the associated entropies $H_{L^{(1)}}$ and $H_{L^{(2)}}$, as constructed in the infimal representation (4), and the functions $f^{(1)}$ and $f^{(2)}$ generating the f-divergences via expression (15), are order equivalent (Lemmas 5.1 and 5.2). After this, we provide a characterization of order equivalent closed convex functions (Lemma 5.3), which is the linchpin of our analysis. The lemma shows that for any two order equivalent closed convex functions f_1 and f_2 with $\operatorname{dom} f_1 = \operatorname{dom} f_2$, there are parameters $a > 0$, $b \in \mathbb{R}^k$, and $c \in \mathbb{R}$ such that $f^{(1)}(t) = af^{(2)}(t) + b^T t + c$ for all $t \in \operatorname{dom} f_1 = \operatorname{dom} f_2$. This proves the "only if" part of the Theorems 1 and 2, yielding the desired result. We present the main parts of the proof in the body of the paper, deferring technical nuances to the supplement [10].

5.1. *Universal equivalence and order equivalence.* By Definition 4.1 [and its equivalent variant stated (21)], universally equivalent losses $L^{(1)}$ and $L^{(2)}$ induce the same ordering of quantized information measures and f-divergences. The next definition captures this ordering slightly differently.

DEFINITION 5.1. Let $f_1 : \Omega \to \overline{\mathbb{R}}$ and $f_2 : \Omega \to \overline{\mathbb{R}}$ be closed convex functions, where $\Omega \subset \mathbb{R}^k$ is closed convex. Let $m \in \mathbb{N}$ be arbitrary and the matrices $A, B \in \mathbb{R}^{k \times m}$ satisfy $A\mathbf{1} = B\mathbf{1}$, where A has columns $a_i \in \Omega$ and B has columns $b_i \in \Omega$. Then f_1 and f_2 are *order-equivalent* if for all $m \in \mathbb{N}$ and all such matrices A and B we have

$$\sum_{j=1}^m f_1(a_j) \le \sum_{j=1}^m f_1(b_j) \quad \text{if and only if} \quad \sum_{j=1}^m f_2(a_j) \le \sum_{j=1}^m f_2(b_j). \tag{24}$$

As the above context suggests, order equivalence has strong connections with universal equivalence of loss functions L and associated f-divergences and generalized entropies. The next two lemmas make this explicit.

LEMMA 5.1. *If losses $L^{(1)}$ and $L^{(2)}$ are lower bounded and universally equivalent, then the associated entropies of the construction* (4) *are order equivalent over $\Delta_k \subset \mathbb{R}^k_+$.*

PROOF. Let H_i be the entropy (pointwise Bayes risk) associated with $L^{(i)}$, noting that $\operatorname{dom} H_1 = \operatorname{dom} H_2 = \Delta_k$ because $\inf_{\pi \in \Delta_k} H_i(\pi) > -\infty$. Let the matrices $A = [a_1 \cdots a_m] \in \mathbb{R}^{k \times m}_+$ and $B \in \mathbb{R}^{k \times m}_+$ satisfy $a_i, b_i \in \Delta_k$ for each $i = 1, \dots, m$, and let $v = \frac{1}{m} A\mathbf{1} = \frac{1}{m} B\mathbf{1} \in \Delta_k$. We show that $\sum_{j=1}^m H_1(a_j) \le \sum_{j=1}^m H_1(b_j)$ if and only if $\sum_{j=1}^m H_2(a_j) \le \sum_{j=1}^m H_2(b_j)$, that is, expression (24) holds, by constructing appropriate distributions $P_{1:k}$ and π, then applying the universal equivalence of $L^{(1)}$ and $L^{(2)}$.

Let M_0 be any integer large enough that $v_0 = \frac{1}{k}(1 + \frac{1}{M_0})\mathbf{1} - \frac{1}{M_0}v \in \mathbb{R}_+^k$, so that $v_0 \in \Delta_k$. Then define the vectors $\widetilde{a}_1 = v_0, \dots, \widetilde{a}_{mM_0} = v_0$, and let

$$A^{\text{ext}} = \begin{bmatrix} a_1 & \cdots & a_m & \widetilde{a}_1 & \cdots & \widetilde{a}_{mM_0} \end{bmatrix} \in \mathbb{R}_+^{k \times M} \quad \text{and}$$

$$B^{\text{ext}} = \begin{bmatrix} b_1 & \cdots & b_m & \widetilde{a}_1 & \cdots & \widetilde{a}_{mM_0} \end{bmatrix} \in \mathbb{R}_+^{k \times M},$$

where $M = (M_0 + 1)m$. These satisfy $A^{\text{ext}}\mathbf{1} = B^{\text{ext}}\mathbf{1} = \frac{M}{k}\mathbf{1}$. We let a^{ext} and b^{ext} denote the columns of these extended matrices.

Now, let the spaces $\mathcal{X} = [M] \times [M]$ and $\mathcal{Z} = [M]$. Define quantizers $\mathsf{q}_1, \mathsf{q}_2 : \mathcal{X} \to \mathcal{Z}$ by $\mathsf{q}_1(i, j) = i$ and $\mathsf{q}_2(i, j) = j$. For $l = 1, \dots, k$, define the distributions P_l on $\mathcal{X}$ by

$$P_l(i, j) = \frac{k^2}{M^2} \cdot a_{il}^{\text{ext}} b_{jl}^{\text{ext}}, \qquad \text{so} \sum_{j=1}^{M} P_l(i, j) = \frac{k}{M} a_{il}^{\text{ext}} \frac{k}{M} \sum_{j=1}^{M} b_{jl}^{\text{ext}} = \frac{k}{M} a_{il}^{\text{ext}}$$

and similarly $\sum_i P_l(i, j) = \frac{k}{M} b_{jl}^{\text{ext}}$. Let $\pi = \frac{1}{k}\mathbf{1}$ be the uniform prior distribution on the label $Y \in \{1, \dots, k\}$, and note that the posterior probability

$$\widetilde{\pi}(\mathsf{q}_1^{-1}(\{i\})) = \left[\frac{\pi_l \sum_j P_l(i, j)}{\sum_{l'} \pi_{l'} \sum_j P_l(i, j)} \right]_{l=1}^{k} = \left[\frac{a_{il}^{\text{ext}}}{\sum_{l'} a_{il'}^{\text{ext}}} \right]_{l=1}^{k} = a_i^{\text{ext}} \in \Delta_k,$$

because $P_l(\mathsf{q}_1^{-1}(i)) = \sum_j P_l(i, j) = \frac{k}{M} a_{il}^{\text{ext}}$, and similarly $\widetilde{\pi}(\mathsf{q}_2^{-1}(\{j\})) = b_j^{\text{ext}} \in \Delta_k$. Taking the expectation over $X \sim \sum_{l=1}^{k} \pi_l P_l$, we have

$$\begin{aligned} \mathbb{E}[H_L(\widetilde{\pi}(\mathsf{q}_1^{-1}(\mathsf{q}_1(X))))] &= \frac{1}{k} \sum_{i,l} P_l(\mathsf{q}_1^{-1}(i)) H_L(\widetilde{\pi}(\mathsf{q}_1^{-1}(i))) \\ &= \frac{1}{M} \sum_{i=1}^{M} H_L(a_i^{\text{ext}}), \end{aligned}$$

because $\sum_l a_{il}^{\text{ext}} = 1$. Similarly, $\mathbb{E}[H_L(\widetilde{\pi}(\mathsf{q}_2^{-1}(\mathsf{q}_2(X))))] = \frac{1}{M} \sum_{j=1}^{M} H_\pi(b_j^{\text{ext}})$. Recalling the definitions (3) and (20) of the (quantized) information associated with H, we have $I(X, \pi; H \mid \mathsf{q}_1) = H(\pi) - \frac{1}{M} \sum_{i=1}^{M} H(a_i^{\text{ext}})$ and $I(X, \pi; H \mid \mathsf{q}_2) = H(\pi) - \frac{1}{M} \sum_{i=1}^{M} H(b_i^{\text{ext}})$. Then the universal equivalence of losses $L^{(1)}$ and $L^{(2)}$ immediately implies for $\pi = \frac{1}{k}\mathbf{1}$ that

$$H_1(\pi) - \frac{1}{M} \sum_{i=1}^{M} H_1(a_i^{\text{ext}}) \le H_1(\pi) - \frac{1}{M} \sum_{i=1}^{M} H_1(b_i^{\text{ext}}) \quad \text{iff}$$

$$H_2(\pi) - \frac{1}{M} \sum_{i=1}^{M} H_2(a_i^{\text{ext}}) \le H_2(\pi) - \frac{1}{M} \sum_{i=1}^{M} H_2(b_i^{\text{ext}}).$$

Noting that $a_i^{\text{ext}} = b_i^{\text{ext}}$ for each $i \geq m+1$, we rearrange the preceding equivalent statements by adding $\frac{1}{M}\sum_{i \geq m+1} H(a_i^{\text{ext}})$ to each side to obtain that the H_i satisfy inequality (24). □

For f-divergences, a parallel result is possible; as the techniques are similar to those we use to prove Lemma 5.1 (by constructing an explicit discrete space $\mathcal{X}$ and quantizers q), we defer the proof to [10], Section 8.1.

LEMMA 5.2. *If losses $L^{(1)}$ and $L^{(2)}$ are universally equivalent for the prior π (Definition 4.1) and lower-bounded, the corresponding f-divergences $f_{L^{(1)},\pi}$ and $f_{L^{(2)},\pi}$ of construction (15) are order equivalent.*

5.2. *Characterization of the order equivalence of convex functions.* Lemmas 5.1 and 5.2 illustrate the intrinsic relationship between the universal equivalence (Definition 4.1) of losses and the order equivalence (Definition 5.1) of their associated generalized entropies and f-divergences. Therefore, it is natural to ask when convex functions are order equivalent. The lemma below characterizes this order equivalence, and coupled with Lemmas 5.1 and 5.2, it immediately implies Theorems 1 and 2.

LEMMA 5.3. *Let $f_1, f_2 : \Omega \to \mathbb{R}$ be closed convex functions, where $\Omega \subset \mathbb{R}^k$ is a convex set. Then f_1 and f_2 are order equivalent on Ω if and only if there exist $a > 0$, $b \in \mathbb{R}^k$, and $c \in \mathbb{R}$ such that for all $t \in \Omega$*

$$f_1(t) = af_2(t) + b^T t + c. \tag{25}$$

While the proof of Lemma 5.3 is complex, we provide a partial proof highlighting the most important parts of the argument, deferring technical details to the supplement. The essential idea is that Lemma 5.3 holds for simplices (and so it certainly holds for H_L); we can then cover any convex set Ω with a number of overlapping simplices to extend the result to all of Ω, which we do fully in [10], Section 8.2. To demonstrate Lemma 5.3 for simplices, we require the following.

DEFINITION 5.2. Vectors $u_0, u_1, \dots, u_m$ are *affinely independent* if

$$u_1 - u_0, \qquad u_2 - u_0, \qquad \dots, \qquad u_m - u_0,$$

are linearly independent. A set $E \subset \mathbb{R}^k$ is a *simplex* if $E = \text{Conv}\{u_0, u_1, \dots, u_k\}$ where $u_0, \dots, u_k$ are affinely independent.

Then the essential special case of Lemma 5.3 is the following result.

LEMMA 5.4. *Let $E = \text{Conv}\{u_0, \dots, u_k\} \subset \Omega$ where $u_0, \dots, u_k$ are affinely independent. If f_1 and f_2 are order equivalent, then there exist $a > 0$, $b \in \mathbb{R}^k$, and $c \in \mathbb{R}$ such that*

$$f_1(t) = af_2(t) + b^T t + c \qquad \textit{for all } t \in E.$$

The proof of Lemma 5.4 proceeds in a series of intermediate results, which we provide in turn, with proofs in [10], Section 8.3. Our first step is to argue that we need only prove equivalence results for convex functions on dense subsets of their domains.

LEMMA 5.5 ([15], Proposition IV.1.2.5). *Let $f_1, f_2 : \Omega \to \mathbb{R}$ be closed convex and satisfy $f_1(t) = f_2(t)$ for t in a dense subset of Ω. Then $f_1 = f_2$ on Ω.*

The first technical lemma we prove is essentially a direct consequence of the definition of order equivalence.

LEMMA 5.6. *Let $u_1, \dots, u_m \in \Omega$, $\alpha \in \mathbb{Q}^m$ satisfy $\mathbf{1}^T\alpha = 1$, and $v \in \Omega$ with $v = \sum_{i=1}^m \alpha_i u_i$. If $f_1, f_2 : \Omega \to \mathbb{R}$ are order equivalent, then*

$$\sum_{i=1}^m \alpha_i f_1(u_i) \le f_1(v) \quad \textit{if and only if} \quad \sum_{i=1}^m \alpha_i f_2(u_i) \le f_2(v).$$

Thus if $\alpha \in \mathbb{Q}^n$ satisfies $\mathbf{1}^T\alpha = 1$ and $u_1, \dots, u_n \in \Omega$, then

$$f_1\left(\sum_{i=1}^n \alpha_i u_i\right) = \sum_{i=1}^n \alpha_i f_1(u_i) \quad \text{iff} \quad f_2\left(\sum_{i=1}^n \alpha_i u_i\right) = \sum_{i=1}^n \alpha_i f_2(u_i). \tag{26}$$

The next lemma shows that we can force equality (25) to hold for the $k+1$ extreme points and centroid of any simplex in $\mathbb{R}^k$; it is intuitive because there are $k+2$ free parameters in the choices of $a > 0$, $b \in \mathbb{R}^k$, and $c \in \mathbb{R}$.

LEMMA 5.7. *Let $f_1, f_2 : \Omega \to \mathbb{R}$ be closed convex and let $u_0, \dots, u_k \in \Omega$ be affinely independent. There exist $a > 0$, $b \in \mathbb{R}^k$, and c such that $f_1(u) = af_2(u) + b^T u + c$ for $u \in \{u_0, \dots, u_k, u_{\text{cent}}\}$, where $u_{\text{cent}} = \frac{1}{k+1}\sum_{i=0}^k u_i$.*

Lastly, we have the following characterization of the linearity of convex functions over convex hulls.

LEMMA 5.8. *Let $f : \Omega \to \mathbb{R}$ be convex with $u_1, \dots, u_m \in \Omega$ and $u_{\text{cent}} = \frac{1}{m}\sum_{i=1}^m u_i$. If $f(u_{\text{cent}}) = \frac{1}{m}\sum_{i=1}^m f(u_i)$, then*

$$f\left(\sum_{i=1}^m \lambda_i u_i\right) = \sum_{i=1}^m \lambda_i f(u_i) \qquad \textit{for all } \lambda \in \mathbb{R}_+^m \textit{ with } \mathbf{1}^T\lambda = 1.$$

With the four Lemmas 5.5–5.8, we can now prove Lemma 5.4. By rotating with $u_i - u_0$ and shifting by u_0, it is no loss of generality to assume that the functions f_i are defined on $V = \{v \in \mathbb{R}_+^k \mid \mathbf{1}^T v \le 1\}$, so that f_1 and f_2 are continuous, defined, convex and order equivalent on V. We make one further reduction. Let

$e_i \in \mathbb{R}^k$ for $1 \le i \le k$ be the standard basis for $\mathbb{R}^k$ and $e_0 = \mathbf{0}$ be shorthand for the all-zeros vector. Further, let $e_{\text{center}} = \frac{1}{k+1}\sum_{i=0}^k e_i$ be the centroid of V (so $V = \text{Conv}\{e_0, \dots, e_k\}$). Lemma 5.7 guarantees the existence of $a > 0, b \in \mathbb{R}^k$, and $c \in \mathbb{R}$ such that

$$f_1(v) = af_2(v) + b^T v + c \qquad \text{for } v \in \{e_0, e_1, \dots, e_k, e_{\text{center}}\}.$$

Now, let $h_1(v) = f_1(v)$ and $h_2(v) = af_2(v) + b^T v + c$, so h_1 and h_2 are convex, order equivalent on V, and satisfy $h_1(v) = h_2(v)$ for $v \in \{e_0, \dots, e_k, e_{\text{center}}\}$. Thus, Lemma 5.4 is equivalent to showing that if h_1, h_2 are convex, order equivalent, and equal on the extreme points and centroid of V, then

$$h_1(v) = h_2(v) \qquad \text{for } v \in V = \{v \in \mathbb{R}_+^k \mid \mathbf{1}^T v \le 1\}. \tag{27}$$

We divide our discussion into two cases.

Linear case. Suppose that $h_1(e_{\text{center}}) = \frac{1}{k+1}\sum_{i=0}^k h_1(e_i)$. Then by order equivalence of h_1 and h_2 [equation (26)], we have $h_2(e_{\text{center}}) = \frac{1}{k+1}\sum_{i=0}^k h_2(e_i)$. Lemma 5.8 thus implies that h_1 and h_2 are linear on $V = \text{Conv}\{e_0, \dots, e_k\}$, equal on the vertices of V, and hence equal on its interior.

Nonlinear case. By convexity, we have $h_1(e_{\text{center}}) < \frac{1}{k+1}\sum_{i=0}^k h_1(e_i)$, and order equivalence (Lemma 5.6) implies $h_2(e_{\text{center}}) < \frac{1}{k+1}\sum_{i=0}^k h_2(e_i)$. For $v \in V = \text{Conv}\{e_0, \dots, e_k\}$, we use $v_0 = 1 - \mathbf{1}^T v$ for shorthand, so we may write $v = \sum_{i=0}^k v_i e_i$ and have $[v_0\ v_1\ \cdots\ v_k]^T \in \Delta_{k+1}$. Now, fix an arbitrary $v \in V \cap \mathbb{Q}^k$. We wish to show that $h_1(v) = h_2(v)$. To that end, we consider consider the gaps due convexity of $h_j(e_{\text{center}})$ to the values of $h_j(e_i)$ *relative* to those from $h_j(v)$ to $h_j(e_i)$, defining the linear functions $\varphi_j : [0, 1] \to \mathbb{R}$ by

$$\varphi_j(r) := (1-r)\left[h_j(e_{\text{center}}) - \frac{1}{k+1}\sum_{i=0}^k h_j(e_i)\right] + r\left[\sum_{i=0}^k v_i h_j(e_i) - h_j(v)\right]$$

for $j = 1, 2$. Then

$$\varphi_j(0) = h_j(e_{\text{center}}) - \frac{1}{k+1}\sum_{i=0}^k h_j(e_i) < 0$$

by assumption, and by convexity,

$$\varphi_j(1) = \sum_{i=0}^k v_i h_j(e_i) - h_j(v) \ge 0.$$

The key is that the order equivalence of h_1 and h_2 on V implies that

$$\text{sign}(\varphi_1(r)) = \text{sign}(\varphi_2(r)) \qquad \text{for } r \in [0, 1], \tag{28}$$

so that φ_1 and φ_2 have the same zero crossing $r^\star > 0$, that is, there exists $0 < r^\star \le 1$ with $\varphi_1(r^\star) = \varphi_2(r^\star) = 0$. [We prove equality (28) presently.] At this $r^\star > 0$, we find

$$0 = \varphi_1(r^\star) - \varphi_2(r^\star) = -r^\star h_1(v) + r^\star h_2(v),$$

where we use that $h_1(e_i) = h_2(e_i)$ for $i = 0, \dots, k$ and $h_1(e_{\text{center}}) = h_2(e_{\text{center}})$. That is, $h_1(v) = h_2(v)$, and as $v \in V \cap \mathbb{Q}^k$ is arbitrary and $\mathbb{Q}^k$ is dense, Lemma 5.5 extends the equality $h_1 = h_2$ to all of V. Expression (27) holds.

Returning to the sign equivalence (28), for $r > 0$, we may divide $\varphi_j(r)$ by r, and we have $\varphi_j(r) \le 0$ if and only if

$$\frac{1-r}{r}\left[h_j(e_{\text{center}}) - \frac{1}{k+1}\sum_{i=0}^{k} h_j(e_i)\right] + \sum_{i=0}^{k} v_i h_j(e_i) \le h_j(v).$$

Defining $\alpha_i = v_i - \frac{1-r}{r(k+1)} \in \mathbb{Q}$ for $i = 0, \dots, k$ and $\alpha_{k+1} = \frac{1-r}{r}$, the inequality $\varphi_j(r) \le 0$ is equivalent to $\sum_{i=0}^{k} \alpha_i h_j(e_i) + \alpha_{k+1} h_j(e_{\text{center}}) \le h_j(v)$. A calculation yields $\mathbf{1}^T \alpha = 1$ and $\sum_{i=0}^{k} \alpha_i e_i + \alpha_{k+1} e_{\text{center}} = v$, and applying Lemma 5.6 immediately yields that $\varphi_1(r) \le 0$ if and only if $\varphi_2(r) \le 0$ for all $r \in (0, 1] \cap \mathbb{Q}$. Noting that $\varphi_1(0) < 0$ and $\varphi_2(0) < 0$, we obtain equality (28).

6. Discussion.

Rather than recapitulating our contributions, we point out a few directions we believe will prove interesting for further study. While Corollary 1 shows that some convex losses are surrogate-risk consistent even with restricted families of classifiers, it does not apply to the practical case in which the collection of discriminants γ is a (convex subset of a) finite-dimensional vector space. This longstanding problem certainly deserves further work. Another direction, a bit further afield, is to investigate the links between this work and objective Bayesian approaches and reference priors [4, 5]. In this line of work, one has a family $\{P_\theta\}_{\theta \in \Theta}$ of probability models on an observation space $\mathcal{X}$ and before performing inference chooses a prior π on θ to maximize $I_\pi(X; \theta)$, the (Shannon) information between $X \sim P_\theta$ and $\theta \sim \pi$. For tasks *other* than minimizing log loss, it may be sensible to use a notion of information and entropy corresponding to the desired loss. Our notions of loss equivalence, including construction of convex losses equivalent to nonconvex losses, could provide insight in such situations.

Acknowledgments. We thank Peter Grünwald for a very careful review of the paper, which helped us to elucidate our contributions, as well as a number of suggestions for improving clarity and communication of the results. We also thank the Associate Editor for helpful feedback and detailed comments.

SUPPLEMENTARY MATERIAL

Proofs of Results (DOI: 10.1214/17-AOS1657SUPP; .pdf). We provide proofs of all deferred results not included in the main text.

REFERENCES

[1] ALI, S. M. and SILVEY, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B* **28** 131–142. MR0196777

[2] BARTLETT, P. L., JORDAN, M. I. and MCAULIFFE, J. D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* **101** 138–156. MR2268032

[3] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300.

[4] BERGER, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Anal.* **1** 385–402.

[5] BERNARDO, J. M. (2005). Reference analysis. In *Bayesian Thinking*, *Modeling and Computation* (D. Day and C. R. Rao, eds.). *Handbook of Statistics* **25** 17–90. Elsevier, Amsterdam.

[6] BLACKWELL, D. (1951). Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* 93–102. Univ. California Press, Berkeley, CA. MR0046002

[7] COVER, T. M. and THOMAS, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley, Hoboken, NJ. MR2239987

[8] CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and indirect observation. *Stud. Sci. Math. Hung.* **2** 299–318.

[9] DEGROOT, M. H. (1962). Uncertainty, information, and sequential experiments. *Ann. Math. Stat.* **33** 404–419.

[10] DUCHI, J. C., KHOSRAVI, K. and RUAN, F. (201). Supplement to “Multiclass classification, information, divergence and surrogate risk.” DOI:10.1214/17-AOS1657SUPP.

[11] GARCÍA-GARCÍA, D. and WILLIAMSON, R. C. (2012). Divergences and risks for multiclass experiments. In *Proceedings of the Twenty Fifth Annual Conference on Computational Learning Theory*.

[12] GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

[13] GRÜNWALD, P. D. and DAWID, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Ann. Statist.* **32** 1367–1433. MR2089128

[14] GYÖRFI, L. and NEMETZ, T. (1978). f-dissimilarity: A generalization of the affinity of several distributions. *Ann. Inst. Statist. Math.* **30** 105–113.

[15] HIRIART-URRUTY, J.-B. and LEMARÉCHAL, C. (1993). *Convex Analysis and Minimization Algorithms. II. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **306**. Springer, Berlin. MR1295240

[16] KAILATH, T. (1967). The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. Commun. Technol.* **15** 52–60.

[17] LIESE, F. and VAJDA, I. (2006). On divergences and informations in statistics and information theory. *IEEE Trans. Inform. Theory* **52** 4394–4412.

[18] LONGO, M., LOOKABAUGH, T. D. and GRAY, R. M. (1990). Quantization for decentralized hypothesis testing under communication constraints. *IEEE Trans. Inform. Theory* **36** 241–255. MR1052776

[19] LUGOSI, G. and VAYATIS, N. (2004). On the Bayes-risk consistency of regularized boosting methods. *Ann. Statist.* **32** 30–55. MR2051000

[20] NGUYEN, X., WAINWRIGHT, M. J. and JORDAN, M. I. (2009). On surrogate loss functions and f-divergences. *Ann. Statist.* **37** 876–904. MR2502654

[21] ÖSTERREICHER, F. and VAJDA, I. (1993). Statistical information and discrimination. *IEEE Trans. Inform. Theory* **39** 1036–1039. MR1237725

[22] POOR, H. V. and THOMAS, J. B. (1977). Applications of Ali–Silvey distance measures in the design of generalized quantizers for binary decision systems. *IEEE Trans. Commun.* **25** 893–900.

[23] PUKELSHEIM, F. (2006). *Optimal Design of Experiments*. *Classics in Applied Mathematics* **50**. SIAM, Philadelphia, PA. MR2224698

[24] REID, M. and WILLIAMSON, R. (2011). Information, divergence, and risk for binary experiments. *J*. *Mach*. *Learn*. *Res*. **12** 731–817.

[25] ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull*. *Amer. Math*. *Soc*. **55** 527–535.

[26] SCHAPIRE, R. E. and FREUND, Y. (2012). *Boosting*: *Foundations and Algorithms*. MIT Press, Cambridge, MA. MR2920188

[27] STEINWART, I. (2007). How to compare different loss functions and their risks. *Constr. Approx*. **26** 225–287. MR2327600

[28] TEWARI, A. and BARTLETT, P. L. (2007). On the consistency of multiclass classification methods. *J*. *Mach*. *Learn*. *Res*. **8** 1007–1025.

[29] TISHBY, N., PEREIRA, F. and BIALEK, W. (1999). The information bottleneck method. In *The* 37*'th Allerton Conference on Communication*, *Control*, *and Computing*.

[30] TSITSIKLIS, J. N. (1993). Decentralized detection. In *Advances in Signal Processing* **2** 297–344. JAI Press, London.

[31] VAJDA, I. (1972). On the f-divergence and singularity of probability measures. *Period*. *Math*. *Hungar.* **2** 223–234.

[32] WILLIAMSON, R. C., VERNET, E. and REID, M. D. (2016). Composite multiclass losses. *J*. *Mach*. *Learn*. *Res*. **17**. MR3595157

[33] ZHANG, T. (2003/04). Statistical analysis of some multi-category large margin classification methods. *J*. *Mach*. *Learn*. *Res*. **5** 1225–1251. MR2248016

J. DUCHI
K. KHOSRAVI
F. RUAN
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: jduchi@stanford.edu
khosravi@stanford.edu
fengruan@stanford.edu