

# ON SURROGATE LOSS FUNCTIONS AND $f$ -DIVERGENCES<sup>1</sup>

BY XUANLONG NGUYEN, MARTIN J. WAINWRIGHT  
 AND MICHAEL I. JORDAN

*Duke University, Statistical and Applied Mathematical Sciences  
 Institute(SAMSI), University of California, Berkeley*

The goal of binary classification is to estimate a discriminant function  $\gamma$  from observations of covariate vectors and corresponding binary labels. We consider an elaboration of this problem in which the covariates are not available directly but are transformed by a dimensionality-reducing quantizer  $Q$ . We present conditions on loss functions such that empirical risk minimization yields Bayes consistency when both the discriminant function and the quantizer are estimated. These conditions are stated in terms of a general correspondence between loss functions and a class of functionals known as Ali-Silvey or  $f$ -divergence functionals. Whereas this correspondence was established by Blackwell [*Proc. 2nd Berkeley Symp. Probab. Statist.* **1** (1951) 93–102. Univ. California Press, Berkeley] for the 0–1 loss, we extend the correspondence to the broader class of surrogate loss functions that play a key role in the general theory of Bayes consistency for binary classification. Our result makes it possible to pick out the (strict) subset of surrogate loss functions that yield Bayes consistency for joint estimation of the discriminant function and the quantizer.

**1. Introduction.** Consider the classical problem of binary classification: given a pair of random variables  $(X, Y) \in (\mathcal{X}, \mathcal{Y})$ , where  $\mathcal{X}$  is a Borel subset of  $\mathbb{R}^d$  and  $\mathcal{Y} = \{-1, +1\}$ , and given of a set of samples  $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ , the goal is to estimate a discriminant function that predicts the binary label  $Y$  given the covariate vector  $X$ . The accuracy of any discriminant function is generally assessed in terms of 0–1 loss as follows. Letting  $\mathbb{P}$  denote the distribution of  $(X, Y)$ , and letting  $\gamma : \mathcal{X} \rightarrow \mathbb{R}$  denote a given discriminant function, we seek to minimize the expectation of the 0–1 loss; that is, the error probability  $\mathbb{P}(Y \neq \text{sign}(\gamma(X)))$ .<sup>2</sup> Unfortunately, the 0–1 loss is a nonconvex function, and practical classification algorithms, such as boosting and the support vector machine, are based on relaxing the 0–1 loss to a convex upper bound or approximation, yielding a *surrogate loss*

Received September 2007; revised September 2007.

<sup>1</sup>Supported in part by grants from Intel Corporation, Microsoft Research, Grant 0412995 from the National Science Foundation, and an Alfred P. Sloan Foundation Fellowship (MJW).

*AMS 2000 subject classifications.* 62G10, 68Q32, 62K05.

*Key words and phrases.* Binary classification, discriminant analysis, surrogate losses,  $f$ -divergences, Ali-Silvey divergences, quantizer design, nonparametric decentralized detection, statistical machine learning, Bayes consistency.

<sup>2</sup>We use the convention that  $\text{sign}(\alpha) = 1$  if  $\alpha > 0$  and  $-1$  otherwise.

1 *function* to which empirical risk minimization procedures can be applied. A sig- 1  
 2 nificant achievement of the recent literature on binary classification has been the 2  
 3 delineation of necessary and sufficient conditions, under which such relaxations 3  
 4 yield Bayes consistency [2, 9, 12, 13, 19, 22]. 4

5 In many practical applications, this classical formulation of binary classification 5  
 6 is elaborated to include an additional stage of “feature selection” or “dimension re- 6  
 7 duction,” in which the covariate vector  $X$  is transformed into a vector  $Z$  according 7  
 8 to a data-dependent mapping  $Q$ . An interesting example of this more elaborate 8  
 9 formulation is a “distributed detection” problem, in which individual components of 9  
 10 the  $d$ -dimensional covariate vector are measured at spatially separated locations, 10  
 11 and there are communication constraints that limit the rate at which the measure- 11  
 12 ments can be forwarded to a central location where the classification decision is 12  
 13 made [21]. This communication-constrained setting imposes severe constraints on 13  
 14 the choice of  $Q$ : any mapping  $Q$  must be a separable function, specified by a col- 14  
 15 lection of  $d$  univariate, discrete-valued functions that are applied component-wise 15  
 16 to  $X$ . The goal of decentralized detection is to specify and analyze data-dependent 16  
 17 procedures for choosing such functions, which are typically referred to as “quan- 17  
 18 tizers.” More generally, we may abstract the essential ingredients of this problem 18  
 19 and consider a problem of experimental design, in which  $Q$  is taken to be a possi- 19  
 20 bly stochastic mapping  $\mathcal{X} \rightarrow \mathcal{Z}$ , chosen from some constrained class  $\mathcal{Q}$  of possible 20  
 21 quantizers. In this setting, the discriminant function is a mapping  $\gamma : \mathcal{Z} \rightarrow \mathbb{R}$ , 21  
 22 chosen from the class  $\Gamma$  of all measurable functions on  $\mathcal{Z}$ . Overall, the problem is to 22  
 23 simultaneously determine both the mapping  $Q$  and the discriminant function  $\gamma$ , 23  
 24 using the data  $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ , so as to jointly minimize the Bayes error 24  
 25  $R_{\text{Bayes}}(\gamma, Q) := \mathbb{P}(Y \neq \text{sign}(\gamma(Z)))$ . 25

26 As alluded to above, when  $Q$  is fixed, it is possible to give general conditions 26  
 27 under which relaxations of 0–1 loss yield Bayes consistency. As we will show in 27  
 28 the current paper, however, these conditions *no longer suffice* to yield consistency 28  
 29 in the more general setting, in which the choice of  $Q$  is also optimized. Rather, 29  
 30 in the setting of jointly estimating the discriminant function  $\gamma$  and optimizing the 30  
 31 quantizer  $Q$ , new conditions need to be imposed. It is the goal of the current paper 31  
 32 to present such conditions and, moreover, to provide a general theoretical under- 32  
 33 standing of their origin. Such an understanding turns out to repose not only on 33  
 34 analytic properties of surrogate loss functions (as in the  $Q$ -fixed case), but on a 34  
 35 relationship between the family of surrogate loss functions and another class of 35  
 36 functions known as *f-divergences* [1, 7]. In rough terms, an *f*-divergence 36  
 37 between two distributions is defined by the expectation of a convex function of their 37  
 38 likelihood ratio. Examples include the Hellinger distance, the total variational 38  
 39 distance, Kullback–Leibler divergence and Chernoff distance, as well as various other 39  
 40 divergences popular in the information theory literature [20]. In our setting, these 40  
 41 *f*-divergences are applied to the class-conditional distributions induced by apply- 41  
 42 ing a fixed quantizer  $Q$ . 42  
 43 43

1 An early hint of the relationship between surrogate losses and  $f$ -divergences 1  
 2 can be found in a seminal paper of Blackwell [3]. In our language, Blackwell's re- 2  
 3 sult can be stated in the following way: if a quantizer  $Q_A$  induces class-conditional 3  
 4 distributions whose  $f$ -divergence is smaller than the  $f$ -divergence induced by a 4  
 5 quantizer  $Q_B$ , then there exists some set of prior probabilities for the class labels 5  
 6 such that  $Q_A$  results in a smaller probability of error than  $Q_B$ . This result suggests 6  
 7 that any analysis of quantization procedures based on 0–1 and surrogate loss func- 7  
 8 tions might usefully attempt to relate surrogate loss functions to  $f$ -divergences. 8  
 9 Our analysis shows that this is indeed a fruitful suggestion, and that Blackwell's 9  
 10 idea takes its most powerful form when we move beyond 0–1 loss to consider 10  
 11 the full set of surrogate loss functions studied in the recent binary classification 11  
 12 literature. 12

13 Blackwell's result [3] has had significant historical impact on the signal process- 13  
 14 ing literature (and thence on the distributed detection literature). Consider, in a 14  
 15 manner complementary to the standard binary classification setting in which the 15  
 16 quantizer  $Q$  is assumed known, the setting in which the discriminant function  $\gamma$  16  
 17 is assumed known and only the quantizer  $Q$  is to be estimated. This is a standard 17  
 18 problem in the signal processing literature (see, e.g., the papers [10, 11, 17]), and 18  
 19 solution strategies typically involve the selection of a specific  $f$ -divergence to be 19  
 20 optimized. Typically, the choice of an  $f$ -divergence is made somewhat heuristi- 20  
 21 cally, based on the grounds of analytic convenience, computational convenience 21  
 22 or asymptotic arguments. 22

23 Our results in effect provide a broader and more rigorous framework for justi- 23  
 24 fying the use of various  $f$ -divergences in solving quantizer design problems. We 24  
 25 broaden the problem to consider the joint estimation of the discriminant function 25  
 26 and the quantizer. We adopt a decision-theoretic perspective in which we aim to 26  
 27 minimize the expectation of 0–1 loss, but we relax to surrogate loss functions that 27  
 28 are convex approximations of 0–1 loss, with the goal of obtaining computationally 28  
 29 tractable minimization procedures. By relating the family of surrogate loss func- 29  
 30 tions to the family of  $f$ -divergences, we are able to specify equivalence classes of 30  
 31 surrogate loss functions. The conditions that we present for Bayes consistency are 31  
 32 expressed in terms of these equivalence classes. 32  
 33

34 1.1. *Our contributions.* In order to state our contributions more precisely, let 34  
 35 us introduce some notation and definitions. Given the distribution  $\mathbb{P}$  of the pair 35  
 36  $(X, Y)$ , consider a discrete space  $\mathcal{Z}$ , and let  $Q(z|x)$  denote a *quantizer*—a condi- 36  
 37 tional probability distribution on  $\mathcal{Z}$  for almost all  $x$ . Let  $\mu$  and  $\pi$  denote measures 37  
 38 over  $\mathcal{Z}$  that are induced by  $Q$  as follows: 38  
 39

$$40 \quad (1a) \quad \mu(z) := \mathbb{P}(Y = 1, Z = z) = p \int_x Q(z|x) d\mathbb{P}(x|Y = 1), \quad 40$$

$$41 \quad (1b) \quad \pi(z) := \mathbb{P}(Y = -1, Z = z) = q \int_x Q(z|x) d\mathbb{P}(x|Y = -1), \quad 42$$

43

43

1 where  $p$  and  $q$  denote the prior probabilities  $p = \mathbb{P}(Y = 1)$  and  $q = \mathbb{P}(Y = -1)$ . 1  
 2 We assume that  $\mathcal{Q}$  is restricted to some constrained class  $\mathcal{Q}$ , such that both  $\mu$  and 2  
 3  $\pi$  are strictly positive measures. 3

4 An  $f$ -divergence is defined as 4

$$5 \quad (2) \quad I_f(\mu, \pi) := \sum_z \pi(z) f\left(\frac{\mu(z)}{\pi(z)}\right), \quad 5$$

6 where  $f: [0, +\infty) \rightarrow \mathbb{R} \cup \{+\infty\}$  is a continuous convex function. Different 6  
 7 choices of convex  $f$  lead to different divergence functionals [1, 7]. 7

8 The loss functions that we consider are known as *margin-based* loss functions. 8  
 9 Specifically, we study convex loss functions  $\phi(y, \gamma(z))$  that are of the form 9  
 10  $\phi(y\gamma(z))$ , where the product  $y\gamma(z)$  is known as the *margin*. Note in particular that 10  
 11 0–1 loss can be written in this form, since  $\phi_{0-1}(y, \gamma(z)) = \mathbb{I}(y\gamma(z) \leq 0)$ . Given 11  
 12 such a margin-based loss function, we define the  $\phi$ -risk  $R_\phi(\gamma, \mathcal{Q}) = \mathbb{E}\phi(Y\gamma(Z))$ . 12  
 13 Statistical procedures will be defined in terms of minimizers of  $R_\phi$  with respect to 13  
 14 the arguments  $\gamma$  and  $\mathcal{Q}$ , with the expectation replaced by an empirical expectation 14  
 15 defined by samples  $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ . 15

16 With these definitions, we now summarize our main results, which are stated 16  
 17 technically in Theorems 1–3. The first result (Theorem 1) establishes a general 17  
 18 correspondence between the family of  $f$ -divergences and the family of optimized 18  
 19  $\phi$ -risks. In particular, let  $R_\phi(\mathcal{Q})$  denote the *optimal*  $\phi$ -risk, meaning the  $\phi$ -risk 19  
 20 obtained by optimizing over the discriminant  $\gamma$  as follows: 20  
 21

$$22 \quad R_\phi(\mathcal{Q}) := \inf_{\gamma \in \Gamma} R_\phi(\mathcal{Q}, \gamma). \quad 22$$

23 In Theorem 1, we establish a precise correspondence between these optimal 23  
 24  $\phi$ -risks and the family of  $f$ -divergences. Theorem 1(a) addresses the forward 24  
 25 direction of this correspondence (from  $\phi$  to  $f$ ); in particular, we show that any 25  
 26 optimal  $\phi$ -risk can be written as  $R_\phi(\mathcal{Q}) = -I_f(\mu, \pi)$ , where  $I_f$  is the divergence 26  
 27 induced by a suitably chosen convex function  $f$ . We also specify a set of prop- 27  
 28 erties that any such function  $f$  inherits from the surrogate loss  $\phi$ . Theorem 1(b) 28  
 29 addresses the converse question: given an  $f$ -divergence, when can it be realized as 29  
 30 an optimal  $\phi$ -risk? We provide a set of necessary and sufficient conditions on any 30  
 31 such  $f$ -divergence and, moreover, specify a constructive procedure for determin- 31  
 32 ing *all* surrogate loss functions  $\phi$  that induce the specified  $f$ -divergence. 32  
 33

34 The relationship is illustrated in Figure 1; whereas each surrogate loss  $\phi$  induces 34  
 35 only one  $f$ -divergence, note that in general there are many surrogate loss functions 35  
 36 that correspond to the same  $f$ -divergence. As particular examples of the general 36  
 37 correspondence established in this paper, we show that the hinge loss corresponds 37  
 38 to the variational distance, the exponential loss corresponds to the Hellinger dis- 38  
 39 tance, and the logistic loss corresponds to the capacitory discrimination distance. 39  
 40

41 This correspondence, in addition to its intrinsic interest as an extension of 41  
 42 Blackwell’s work, has a number of consequences. In Section 3, we show that it 42  
 43

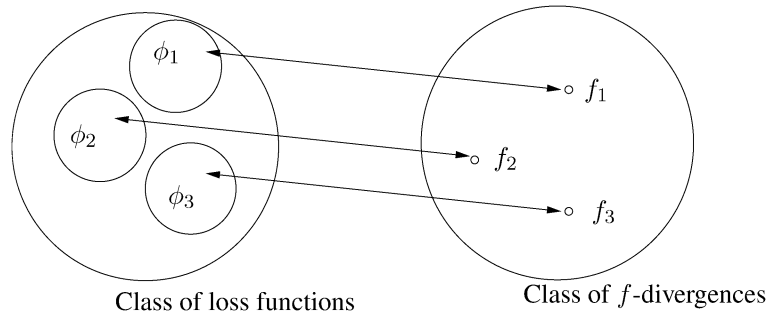


FIG. 1. Illustration of the correspondence between  $f$ -divergences and loss functions. For each loss function  $\phi$ , there exists exactly one corresponding  $f$ -divergence such that the optimized  $\phi$ -risk is equal to the negative  $f$ -divergence. The reverse mapping is, in general, many-to-one.

allows us to isolate a class of  $\phi$ -divergences for which empirical risk minimization is consistent in the joint (quantizer and discriminant) estimation setting. Note in particular (e.g., from Blackwell's work) that the  $f$ -divergence associated with the 0–1 loss is the total variational distance. In Theorem 2, we specify a broader class of  $\phi$ -losses that induce the total variational distance and prove that, under standard technical conditions, an empirical risk minimization procedure based on any such  $\phi$ -risk is Bayes consistent. This broader class includes not only the non-convex 0–1 loss, but also other convex and computationally tractable  $\phi$ -losses, including the hinge loss function that is well known in the context of support vector machines [6]. The key novelty in this result is that it applies to procedures that optimize simultaneously over the discriminant function  $\gamma$  and the quantizer  $Q$ .

One interpretation of Theorem 2 is as specifying a set of surrogate loss functions  $\phi$  that are universally equivalent to the 0–1 loss, in that empirical risk minimization procedures based on such  $\phi$  yield classifier-quantizer pairs  $(\gamma^*, Q^*)$  that achieve the Bayes risk. In Section 4, we explore this notion of universal equivalence between loss functions in more depth. In particular, we say that two loss functions  $\phi_1$  and  $\phi_2$  are *universally equivalent* if the optimal risks  $R_{\phi_1}(Q)$  and  $R_{\phi_2}(Q)$  induce the same ordering on quantizers, meaning the ordering  $R_{\phi_1}(Q_a) \leq R_{\phi_1}(Q_b)$  holds if and only if  $R_{\phi_2}(Q_a) \leq R_{\phi_2}(Q_b)$  for all quantizer pairs  $Q_a$  and  $Q_b$ . Thus, the set of surrogate loss functions can be categorized into subclasses by this equivalence, where of particular interest are all surrogate loss functions that are equivalent (in the sense just defined) to the 0–1 loss. In Theorem 3, we provide an explicit and easily tested set of conditions for a  $\phi$ -risk to be equivalent to the 0–1 loss. One consequence is that procedures based on a  $\phi$ -risk outside of this family cannot be Bayes consistent for joint optimization of the discriminant  $\gamma$  and quantizer  $Q$ . Thus, coupled with our earlier result in Theorem 2, we obtain a set of necessary and sufficient conditions on  $\phi$ -losses to be Bayes consistent in this joint estimation setting.

**2. Correspondence between  $\phi$ -loss and  $f$ -divergence.** Recall that in the setting of binary classification with  $Q$  fixed, it is possible to give conditions on the class of surrogate loss functions (i.e., upper bounds on or approximations of the 0–1 loss) that yield Bayes consistency. In particular, Bartlett, Jordan and McAuliffe [2] have provided the following definition of a *classification-calibrated loss*.

**DEFINITION 1.** Define  $\Phi_{a,b}(\alpha) = \phi(\alpha)a + \phi(-\alpha)b$ . A loss function  $\phi$  is *classification-calibrated* if for any  $a, b \geq 0$  and  $a \neq b$ :

$$(3) \quad \inf_{\{\alpha \in \mathbb{R} | \alpha(a-b) < 0\}} \Phi_{a,b}(\alpha) > \inf_{\{\alpha \in \mathbb{R} | \alpha(a-b) \geq 0\}} \Phi_{a,b}(\alpha).$$

The definition is essentially a pointwise form of a Fisher consistency condition that is appropriate for the binary classification setting. When  $Q$  is fixed, this definition ensures that under, fairly general conditions, the decision rule  $\gamma$  obtained by an empirical risk minimization procedure behaves equivalently to the Bayes optimal decision rule. Bartlett, Jordan and McAuliffe [2] also derived a simple lemma that characterizes classification-calibration for convex functions.

**LEMMA 1.** *Let  $\phi$  be a convex function. Then  $\phi$  is classification-calibrated if and only if it is differentiable at 0 and  $\phi'(0) < 0$ .*

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  $\phi: \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ :

A1:  $\phi$  is classification-calibrated;

A2:  $\phi$  is continuous;

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

$$(4) \quad \phi(\alpha^* - \varepsilon) \geq \phi(\alpha^* + \varepsilon).$$

The interpretation of assumption A3 is that one should penalize deviations away from  $\alpha^*$  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  $\alpha$ . Moreover, this assumption is satisfied by all of the loss functions commonly considered in the literature; in particular, any decreasing function  $\phi$  (e.g., hinge loss, logistic loss, exponential loss) satisfies this condition, as does the least squares loss (which is not decreasing). When  $\phi$  is convex, assumption A1 is equivalent to requiring that  $\phi$  be differentiable at 0 and  $\phi'(0) < 0$ . These facts also imply that the quantity  $\alpha^*$  defined in assumption A3 is strictly positive. Finally, although  $\phi$  is not defined for  $-\infty$ , we shall use the convention that  $\phi(-\infty) = +\infty$ .

1 In the following, we present the general relationship between optimal  $\phi$ -risks 1  
 2 and  $f$ -divergences. The easier direction is to show that any  $\phi$ -risk induces a cor- 2  
 3 responding  $f$ -divergence. The  $\phi$ -risk can be written in the following way: 3

$$4 \quad (5a) \quad R_\phi(\gamma, Q) = \mathbb{E}\phi(Y\gamma(Z)) \quad 4$$

$$5 \quad (5b) \quad = \sum_z \phi(\gamma(z))\mu(z) + \phi(-\gamma(z))\pi(z). \quad 5$$

6 For a fixed mapping  $Q$ , the optimal  $\phi$ -risk has the form 6  
 7

$$8 \quad R_\phi(Q) = \sum_{z \in \mathcal{Z}} \inf_\alpha (\phi(\alpha)\mu(z) + \phi(-\alpha)\pi(z)) \quad 8$$

$$9 \quad = \sum_z \pi(z) \inf_\alpha \left( \phi(-\alpha) + \phi(\alpha) \frac{\mu(z)}{\pi(z)} \right). \quad 9$$

10 For each  $z$ , define  $u(z) := \frac{\mu(z)}{\pi(z)}$ . With this notation, the function  $\inf_\alpha (\phi(-\alpha) +$  10  
 11  $\phi(\alpha)u)$  is concave as a function of  $u$  (since the minimum of a collection of linear 11  
 12 functions is concave). Thus, if we define 12

$$13 \quad (6) \quad f(u) := -\inf_\alpha (\phi(-\alpha) + \phi(\alpha)u), \quad 13$$

14 we obtain the relation 14

$$15 \quad (7) \quad R_\phi(Q) = -I_f(\mu, \pi). \quad 15$$

16 We have thus established the easy direction of the correspondence: given a loss 16  
 17 function  $\phi$ , there exists an  $f$ -divergence for which the relation (7) holds. Further- 17  
 18 more, the convex function  $f$  is given by the expression (6). Note that our argument 18  
 19 does not require convexity of  $\phi$ . 19

20 We now consider the converse. Given a divergence  $I_f(\mu, \pi)$  for some convex 20  
 21 function  $f$ , does there exist a loss function  $\phi$  for which  $R_\phi(Q) = -I_f(\mu, \pi)$ ? In 21  
 22 the theorem presented below, we answer this question in the affirmative. Moreover, 22  
 23 we present a constructive result: we specify necessary and sufficient conditions 23  
 24 under which there exist decreasing and convex surrogate loss functions for a given 24  
 25  $f$ -divergence, and we specify the form of all such loss functions. 25  
 26

27 Recall the notion of convex duality [18]: For a lower semicontinuous convex 27  
 28 function  $f: \mathbb{R} \rightarrow \mathbb{R} \cup \{\infty\}$ , the conjugate dual  $f^*: \mathbb{R} \rightarrow \mathbb{R} \cup \{\infty\}$  is defined as 28  
 29  $f^*(u) = \sup_{v \in \mathbb{R}} uv - f(v)$ . Consider an intermediate function: 29  
 30

$$31 \quad (8) \quad \Psi(\beta) = f^*(-\beta). \quad 31$$

32 Define  $\beta_1 := \inf\{\beta: \Psi(\beta) < +\infty\}$  and  $\beta_2 := \inf\{\beta: \Psi(\beta) \leq \inf \Psi\}$ . We are ready 32  
 33 to state our first main result. 33  
 34

35 THEOREM 1. (a) For any margin-based surrogate loss function  $\phi$ , there is 35  
 36 an  $f$ -divergence such that  $R_\phi(Q) = -I_f(\mu, \pi)$  for some lower semicontinuous 36  
 37 convex function  $f$ . 37  
 38  
 39  
 40  
 41  
 42  
 43

1 *In addition, if  $\phi$  is a decreasing convex loss function that satisfies conditions* 1  
 2 *A1, A2 and A3, then the following properties hold:* 2

- 3 (i)  $\Psi$  is a decreasing and convex function; 3  
 4 (ii)  $\Psi(\Psi(\beta)) = \beta$  for all  $\beta \in (\beta_1, \beta_2)$ ; 4  
 5 (iii) There exists a point  $u^* \in (\beta_1, \beta_2)$  such that  $\Psi(u^*) = u^*$ . 5  
 6 6

7 (b) *Conversely, if  $f$  is a lower semicontinuous convex function satisfying all* 7  
 8 *conditions (i)–(iii), there exists a decreasing convex surrogate loss  $\phi$  that induces* 8  
 9 *the  $f$ -divergence in the sense of equations (6) and (7).* 9

10 For proof of this theorem and additional properties, see Section 5.1. 10  
 11 11

12 REMARKS. (a) The existential statement in Theorem 1 can be strengthened 12  
 13 to a constructive procedure, through which we specify how to obtain any  $\phi$  loss 13  
 14 function that induces a given  $f$ -divergence. Indeed, in the proof of Theorem 1(b) 14  
 15 presented in Section 5.1, we prove that any decreasing surrogate loss function  $\phi$  15  
 16 satisfying conditions A1–A3 that induces an  $f$ -divergence must be of the form 16  
 17 17

$$(9) \quad \phi(\alpha) = \begin{cases} u^*, & \text{if } \alpha = 0, \\ \Psi^2(g(\alpha + u^*)), & \text{if } \alpha > 0, \\ g(-\alpha + u^*), & \text{if } \alpha < 0, \end{cases}$$

18 where  $g : [u^*, +\infty) \rightarrow \overline{\mathbb{R}}$  is some increasing continuous and convex function such 18  
 19 that  $g(u^*) = u^*$ , and  $g$  is right-differentiable at  $u^*$  with  $g'(u^*) > 0$ . 19  
 20 20

21 (b) Another consequence of Theorem 1 is that any  $f$ -divergence can be 21  
 22 obtained from a rather large set of surrogate loss functions; indeed, different such 22  
 23 losses are obtained by varying the function  $g$  in our constructive specification (9). 23  
 24 In Section 2.1, we provide concrete examples of this constructive procedure and 24  
 25 the resulting correspondences. For instance, we show that the variational distance 25  
 26 corresponds to the 0–1 loss and the hinge loss, while the Hellinger distance corre- 26  
 27 sponds to the exponential loss. Both divergences are also obtained from many less 27  
 28 familiar loss functions. 28  
 29 29

30 (c) Although the correspondence has been formulated in the population set- 30  
 31 ting, it is the basis of a constructive method for specifying a class of surrogate loss 31  
 32 functions that yield a Bayes consistent estimation procedure. Indeed, in Section 3, 32  
 33 we exploit this result to isolate a subclass of surrogate convex loss functions that 33  
 34 yield Bayes-consistent procedures for joint  $(\gamma, Q)$  minimization procedures. In- 34  
 35 terestingly, this class is a strict subset of the class of classification-calibrated loss 35  
 36 functions, all of which yield Bayes-consistent estimation procedure in the standard 36  
 37 classification setting (e.g., [2]). For instance, the class that we isolate contains the 37  
 38 hinge loss, but *not* the exponential loss or the logistic loss functions. Finally, in 38  
 39 Section 4, we show that, in a suitable sense, the specified subclass of surrogate 39  
 40 loss functions is the only one that yields consistency for the joint  $(\gamma, Q)$  estima- 40  
 41 tion problem. 41  
 42 42  
 43 43



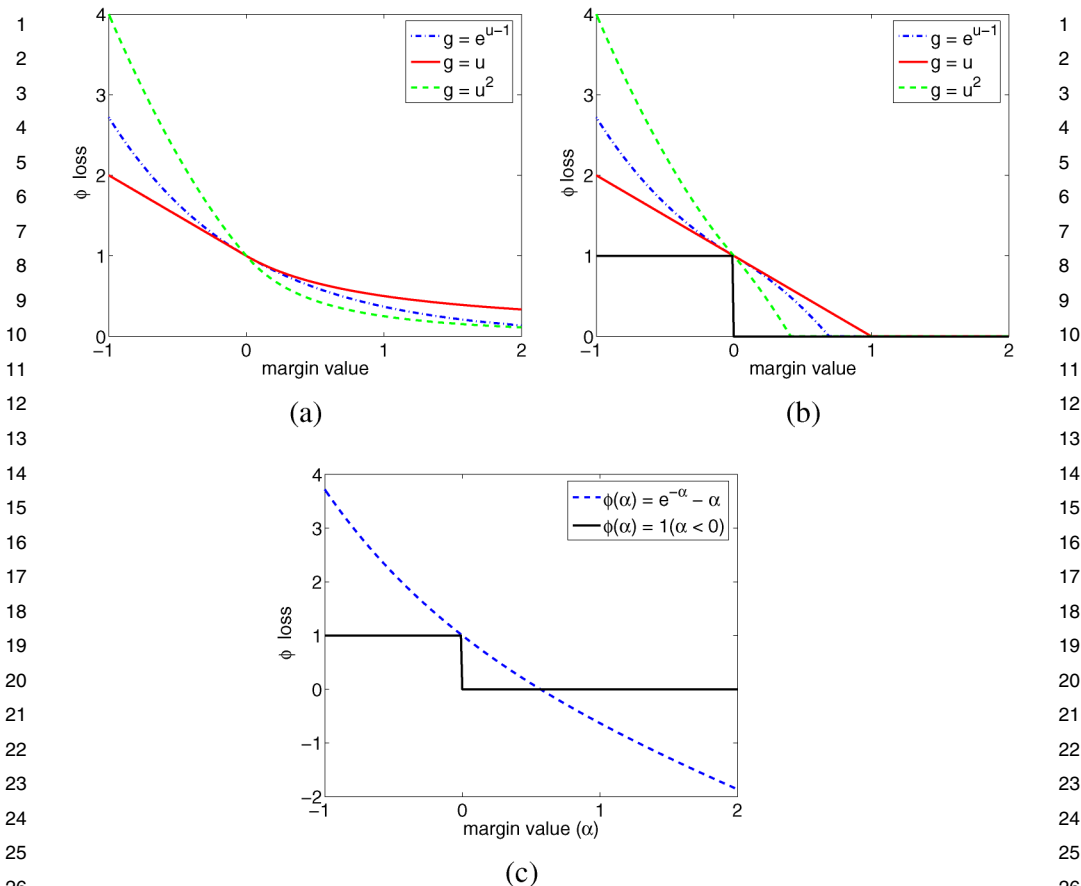


FIG. 2. Panels (a) and (b) show examples of  $\phi$  losses that induce the Hellinger distance and variational distance, respectively, based on different choices of the function  $g$ . Panel (c) shows a loss function that induces the symmetric KL divergence; for the purposes of comparison, the 0–1 loss is also plotted.

2.1.2. Exponential loss and Hellinger distance. Now, consider the exponential loss  $\phi(\alpha) = \exp(-\alpha)$ . In this case, a little calculation shows that the optimal discriminant is  $\gamma(z) = \frac{1}{2} \log \frac{\mu(z)}{\pi(z)}$ . The optimal risk for exponential loss is given by

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

where  $h(\mu, \pi) := \frac{1}{2} \sum_{z \in \mathcal{Z}} (\sqrt{\mu(z)} - \sqrt{\pi(z)})^2$  denotes the Hellinger distance between measures  $\mu$  and  $\pi$ . Conversely, 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  $f$  by setting  $f(u) = +\infty$  for  $u < 0$ ; doing so does not alter the Hellinger (or Bhattacharyya) distances. As be-

1 fore,

$$2 \quad \Psi(\beta) = f^*(-\beta) = \sup_{u \in \mathbb{R}}(-\beta u - f(u)) = \begin{cases} 1/\beta, & \text{when } \beta > 0, \\ +\infty, & \text{otherwise.} \end{cases} \quad 3$$

4 Thus, we see that  $u^* = 1$ . If we let  $g(u) = u$ , then a possible surrogate loss function  
5 that realizes the Hellinger distance takes the form: 6

$$7 \quad \phi(\alpha) = \begin{cases} 1, & \text{if } \alpha = 0, \\ \frac{1}{\alpha + 1}, & \text{if } \alpha > 0, \\ -\alpha + 1, & \text{if } \alpha < 0. \end{cases} \quad 8$$

9 On the other hand, if we set  $g(u) = \exp(u - 1)$ , then we obtain the exponential  
10 loss  $\phi(\alpha) = \exp(-\alpha)$ . See Figure 2 for illustrations of these loss functions. 11

12 *2.1.3. Least squares loss and triangular discrimination distance.* Letting  
13  $\phi(\alpha) = (1 - \alpha)^2$  be the least squares loss, the optimal discriminant is given by  
14  $\gamma(z) = \frac{\mu(z) - \pi(z)}{\mu(z) + \pi(z)}$ . Thus, the optimal risk for least squares loss takes the form 15

$$16 \quad R_{\text{sqr}}(Q) = \sum_{z \in \mathcal{Z}} \frac{4\mu(z)\pi(z)}{\mu(z) + \pi(z)} = 1 - \sum_{z \in \mathcal{Z}} \frac{(\mu(z) - \pi(z))^2}{\mu(z) + \pi(z)} = 1 - \Delta(\mu, \pi), \quad 17$$

18 where  $\Delta(\mu, \pi)$  denotes the *triangular discrimination distance* [20]. Conversely,  
19 the triangular discriminatory distance is equivalent to the negative of the harmonic  
20 distance; it is an  $f$ -divergence with  $f(u) = -\frac{4u}{u+1}$  for  $u \geq 0$ . Let us augment  $f$   
21 with  $f(u) = +\infty$  for  $u < 0$ . We have 22

$$23 \quad \Psi(\beta) = \sup_{u \in \mathbb{R}}(-\beta u - f(u)) = \begin{cases} (2 - \sqrt{\beta})^2, & \text{for } \beta \geq 0, \\ +\infty, & \text{otherwise.} \end{cases} \quad 24$$

25 Clearly,  $u^* = 1$ . In this case, setting  $g(u) = u^2$  gives the least square loss  $\phi(\alpha) =$   
26  $(1 - \alpha)^2$ . 27

28 *2.1.4. Logistic loss and capacitory discrimination distance.* Let  $\phi(\alpha) =$   
29  $\log(1 + \exp(-\alpha))$  be the logistic loss. Then,  $\gamma(z) = \log \frac{\mu(z)}{\pi(z)}$ . As a result, the  
30 optimal risk for logistic loss is given by 31

$$32 \quad R_{\log}(Q) = \sum_{z \in \mathcal{Z}} \mu(z) \log \frac{\mu(z) + \pi(z)}{\mu(z)} + \pi(z) \log \frac{\mu(z) + \pi(z)}{\pi(z)} \quad 33$$

$$34 \quad = \log 2 - KL\left(\mu \parallel \frac{\mu + \pi}{2}\right) - KL\left(\pi \parallel \frac{\mu + \pi}{2}\right) = \log 2 - C(\mu, \pi), \quad 35$$

36 where  $KL(U, V)$  denotes the Kullback–Leibler divergence between two measures  
37  $U$  and  $V$ , and  $C(U, V)$  denotes the *capacitory discrimination distance* [20]. Con-  
38 versely, the capacitory discrimination distance is equivalent to an  $f$ -divergence 39  
40 41  
42  
43

1 with  $f(u) = -u \log \frac{u+1}{u} - \log(u+1)$ , for  $u \geq 0$ . As before, augmenting this func- 1  
 2 tion with  $f(u) = +\infty$  for  $u < 0$ , we have 2

$$3 \quad \Psi(\beta) = \sup_{u \in \mathbb{R}} (-\beta u - f(u)) = \begin{cases} \beta - \log(e^\beta - 1), & \text{for } \beta \geq 0, \\ +\infty, & \text{otherwise.} \end{cases} 4 \quad 5$$

6 This representation shows that  $u^* = \log 2$ . If we choose  $g(u) = \log(1 + \frac{e^u}{2})$ , then 6  
 7 we recover the logistic loss  $\phi(\alpha) = \log[1 + \exp(-\alpha)]$ . 7  
 8

9 **2.1.5. Another symmetrized Kullback–Leibler divergence.** Recall that both the 9  
 10 KL divergences [i.e.,  $KL(\mu\|\pi)$  and  $KL(\pi\|\mu)$ ] are asymmetric; therefore, Corol- 10  
 11 lary 3 (see Section 5.1) implies that they are *not* realizable by any margin-based 11  
 12 surrogate loss. However, a closely related functional is the *symmetric Kullback–* 12  
 13 *Leibler* divergence [5]: 13  
 14

$$15 \quad (11) \quad KL_s(\mu, \pi) := KL(\mu\|\pi) + KL(\pi\|\mu). 15$$

16 It can be verified that this symmetrized KL divergence is an  $f$ -divergence, 16  
 17 generated by the function  $f(u) = -\log u + u \log u$  for  $u \geq 0$ , and  $+\infty$  otherwise. 17  
 18 Theorem 1 implies that it can be generated by surrogate loss functions of form (9), 18  
 19 but the form of this loss function is not at all obvious. Therefore, in order to re- 19  
 20 cover an explicit form for some  $\phi$ , we follow the constructive procedure outlined 20  
 21 in the remarks following Theorem 1, first defining 21  
 22

$$23 \quad \Psi(\beta) = \sup_{u \geq 0} \{-\beta u + \log u - u \log u\}. 23$$

24 In order to compute the value of this supremum, we take the derivative with 24  
 25 respect to  $u$  and set it to zero; doing so yields the zero-gradient condition 25  
 26  $-\beta + 1/u - \log u - 1 = 0$ . To capture this condition, we define a function 26  
 27  $r: [0, +\infty) \rightarrow [-\infty, +\infty]$  via  $r(u) = 1/u - \log u$ . It is easy to see that  $r$  is 27  
 28 strictly decreasing function whose range covers the whole real line; moreover, 28  
 29 the zero-gradient condition is equivalent to  $r(u) = \beta + 1$ . We can thus write 29  
 30  $\Psi(\beta) = u + \log u - 1$  where  $u = r^{-1}(\beta + 1)$ , or, equivalently, 30  
 31

$$32 \quad \Psi(\beta) = r(1/u) - 1 = r\left(\frac{1}{r^{-1}(\beta + 1)}\right) - 1. 32$$

33 It is straightforward to verify that the function  $\Psi$  thus specified is strictly decreas- 33  
 34 ing and convex with  $\Psi(0) = 0$ , and that  $\Psi(\Psi(\beta)) = \beta$  for any  $\beta \in \mathbb{R}$ . Therefore, 34  
 35 Theorem 1 allow us to specify the form of any convex surrogate loss function that 35  
 36 generates the symmetric KL divergence; in particular, any such functions must be 36  
 37 of the form (9): 37  
 38

$$39 \quad \phi(\alpha) = \begin{cases} g(-\alpha), & \text{for } \alpha \leq 0, \\ \Psi(g(\alpha)), & \text{otherwise,} \end{cases} 39 \quad 40$$

41  
 42  
 43

1 where  $g : [0, +\infty) \rightarrow [0, +\infty)$  is some increasing convex function satisfying 1  
 2  $g(0) = 0$ . As a particular example (and one that leads to a closed form expres- 2  
 3 sion for  $\phi$ ), let us choose  $g(u) = e^u + u - 1$ . Doing so leads to the surrogate loss 3  
 4 function 4

$$5 \quad \phi(\alpha) = e^{-\alpha} - \alpha - 1, \quad 5$$

6 as illustrated in Figure 2(c). 6  
 7 7

8 **3. Bayes consistency via surrogate losses.** As shown in Section 2.1.1, if we 8  
 9 substitute the (nonconvex) 0–1 loss function into the linking equation (6), then 9  
 10 we obtain the variational distance  $V(\mu, \pi)$  as the  $f$ -divergence associated with 10  
 11 the function  $f(u) = \min\{u, 1\}$ . A bit more broadly, let us consider the subclass of 11  
 12  $f$ -divergences defined by functions of the form 12

$$13 \quad (12) \quad f(u) = -c \min\{u, 1\} + au + b, \quad 13$$

14 where  $a, b$  and  $c$  are scalars with  $c > 0$ . (For further examples of such losses, in 15  
 16 addition to the 0–1 loss, see Section 2.1.) The main result of this section is that 16  
 17 there exists a subset of surrogate losses  $\phi$  associated with an  $f$ -divergence of the 17  
 18 form (12) that, when used in the context of a risk minimization procedure for 18  
 19 jointly optimizing  $(\gamma, Q)$  pairs, yields a Bayes consistent method. 19

20 We begin by specifying some standard technical conditions under which our 20  
 21 Bayes consistency result holds. Consider sequences of increasing compact func- 21  
 22 tion classes  $\mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq \dots \subseteq \Gamma$  and  $\mathcal{D}_1 \subseteq \mathcal{D}_2 \subseteq \dots \subseteq \mathcal{Q}$ . Recall that  $\Gamma$  denotes the 22  
 23 class of all measurable functions from  $\mathcal{Z} \rightarrow \mathbb{R}$ , whereas  $\mathcal{Q}$  is a constrained class 23  
 24 of quantizer functions  $Q$ , with the restriction that  $\mu$  and  $\pi$  are strictly positive 24  
 25 measures. Our analysis supposes that there exists an oracle that outputs an optimal 25  
 26 solution to the minimization problem 26

$$27 \quad (13) \quad \min_{(\gamma, Q) \in (\mathcal{C}_n, \mathcal{D}_n)} \hat{R}_\phi(\gamma, Q) = \min_{(\gamma, Q) \in (\mathcal{C}_n, \mathcal{D}_n)} \frac{1}{n} \sum_{i=1}^n \sum_{z \in \mathcal{Z}} \phi(Y_i \gamma(z)) Q(z | X_i), \quad 27$$

28 and let  $(\gamma_n^*, Q_n^*)$  denote one such solution. Let  $R_{\text{Bayes}}^*$  denote the minimum Bayes 28  
 29 risk achieved over the space of decision rules  $(\gamma, Q) \in (\Gamma, \mathcal{Q})$ : 29

$$30 \quad (14) \quad R_{\text{Bayes}}^* := \inf_{(\gamma, Q) \in (\Gamma, \mathcal{Q})} R_{\text{Bayes}}(\gamma, Q). \quad 30$$

31 We refer to the nonnegative quantity  $R_{\text{Bayes}}(\gamma_n^*, Q_n^*) - R_{\text{Bayes}}^*$  as the *excess Bayes* 31  
 32 *risk* of our estimation procedure. We say that such an estimation procedure is *uni-* 32  
 33 *versally consistent* if the excess Bayes risk converges to zero, that is, if under the 33  
 34 (unknown) Borel probability measure  $\mathbb{P}$  on  $\mathcal{X} \times \mathcal{Y}$ , we have 34

$$35 \quad (15) \quad \lim_{n \rightarrow \infty} R_{\text{Bayes}}(\gamma_n^*, Q_n^*) = R_{\text{Bayes}}^* \quad \text{in probability.} \quad 35$$

36 In order to analyze the statistical behavior of this algorithm and to establish 36  
 37 universal consistency for appropriate sequences  $(\mathcal{C}_n, \mathcal{D}_n)$  of function classes, we 37  
 38 follow a standard strategy of decomposing the Bayes error in terms of two types 38  
 39 of errors: 39  
 40 40  
 41 41  
 42 42  
 43 43

- 1 • the *approximation error* associated with function classes  $\mathcal{C}_n \subseteq \Gamma$ , and  $\mathcal{D}_n \subseteq \mathcal{Q}$ : 1

2 (16) 
$$\mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) = \inf_{(\gamma, Q) \in (\mathcal{C}_n, \mathcal{D}_n)} \{R_\phi(\gamma, Q)\} - R_\phi^*,$$
 3

4 where  $R_\phi^* := \inf_{(\gamma, Q) \in (\Gamma, \mathcal{Q})} R_\phi(\gamma, Q)$ ; 4

- 5 • the *estimation error* introduced by the finite sample size  $n$ : 5

6 (17) 
$$\mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) = \mathbb{E} \sup_{(\gamma, Q) \in (\mathcal{C}_n, \mathcal{D}_n)} |\hat{R}_\phi(\gamma, Q) - R_\phi(\gamma, Q)|,$$
 7

8 where the expectation is taken with respect to the (unknown) measure  $\mathbb{P}^n(X, Y)$ . 8

9 For asserting universal consistency, we impose the standard conditions: 9

10 (18) **Approximation condition:** 
$$\lim_{n \rightarrow \infty} \mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) = 0.$$
 10

11 (19) **Estimation condition:** 
$$\lim_{n \rightarrow \infty} \mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) = 0 \quad \text{in probability.}$$
 11

12 **Conditions on loss function  $\phi$ :** Our consistency result applies to the class of 12  
13 surrogate losses that satisfy the following: 13

14 B1:  $\phi$  is continuous, convex, and classification-calibrated; 14

15 B2: For each  $n = 1, 2, \dots$ , we assume that 15

16 (20) 
$$M_n := \max_{y \in \{-1, +1\}} \sup_{(\gamma, Q) \in (\mathcal{C}_n, \mathcal{D}_n)} \sup_{z \in \mathcal{Z}} |\phi(y\gamma(z))| < +\infty.$$
 16

17 With this set-up, the following theorem ties together the Bayes error with the 17  
18 approximation error and estimation error and provides sufficient conditions for 18  
19 universal consistency for a *suitable subclass* of surrogate loss functions. 19

20 **THEOREM 2.** *Consider an estimation procedure of the form (13), using a sur-* 20  
21 *rogate loss  $\phi$ . Recall the prior probabilities  $p = \mathbb{P}(Y = 1)$  and  $q = \mathbb{P}(Y = -1)$ .* 21  
22 *For any surrogate loss  $\phi$  satisfying conditions B1 and B2 and inducing an* 22  
23  *$f$ -divergence of the form (12) for any  $c > 0$ , and for  $a, b$  such that  $(a - b)(p - q) \geq$*  23  
24 *0, we have:* 24

- 25 (a) *For any Borel probability measure  $\mathbb{P}$ , there holds, with probability at least* 25  
26  *$1 - \delta$ :* 26

27 
$$R_{\text{Bayes}}(\gamma_n^*, Q_n^*) - R_{\text{Bayes}}^*$$
 27  
28 
$$\leq \frac{2}{c} \left\{ 2\mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) + \mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) + 2M_n \sqrt{2 \frac{\ln(2/\delta)}{n}} \right\};$$
 28

- 29 (b) *Universal Consistency:* *For function classes satisfying the approxima-* 29  
30 *tion (18) and estimation conditions (19), the estimation procedure (13) is univer-* 30  
31 *sally consistent:* 31

32 (21) 
$$\lim_{n \rightarrow \infty} R_{\text{Bayes}}(\gamma_n^*, Q_n^*) = R_{\text{Bayes}}^* \quad \text{in probability.}$$
 32

1     REMARKS. (i) Note that both the approximation and the estimation errors are 1  
 2 with respect to the  $\phi$ -loss, but the theorem statement refers to the excess Bayes 2  
 3 risk. Since the analysis of approximation and estimation conditions such as those 3  
 4 in equation (18) and (19) is a standard topic in statistical learning, we will not 4  
 5 discuss it further here. We note that our previous work analyzed the estimation 5  
 6 error for certain kernel classes [15]. 6

7     (ii) It is worth pointing out that in order for our result to be applicable to an 7  
 8 *arbitrary* constrained class of  $\mathcal{Q}$  for which  $\mu$  and  $\pi$  are strictly positive measures, 8  
 9 we need the additional constraint that  $(a - b)(p - q) \geq 0$ , where  $a, b$  are scalars 9  
 10 in the  $f$ -divergence (12) and  $p, q$  are the unknown prior probabilities. Intuitively, 10  
 11 this requirement is needed to ensure that the approximation error due to varying 11  
 12  $Q$  within  $\mathcal{Q}$  dominates the approximation error due to varying  $\gamma$  (because the 12  
 13 optimal  $\gamma$  is determined only after  $Q$ ) for arbitrary  $\mathcal{Q}$ . Since  $p$  and  $q$  are generally 13  
 14 unknown, the only  $f$ -divergences that are practically useful are the ones for which 14  
 15  $a = b$ . One such  $\phi$  is the hinge loss, which underlies the support vector machine. 15

16     Finally, we note that the proof of Theorem 2 relies on an auxiliary result that is 16  
 17 of independent interest. In particular, we prove that for any function classes  $\mathcal{C}$  and 17  
 18  $\mathcal{D}$ , for certain choice of surrogate loss  $\phi$ , the excess  $\phi$ -risk is related to the excess 18  
 19 Bayes risk as follows. 19  
 20

21     LEMMA 2. *Let  $\phi$  be a surrogate loss function satisfying all conditions spec-* 21  
 22 *ified in Theorem 2. Then, for any classifier-quantizer pair  $(\gamma, Q) \in (\mathcal{C}, \mathcal{D})$ , we* 22  
 23 *have* 23

$$24 \quad (22) \quad \frac{c}{2}[R_{\text{Bayes}}(\gamma, Q) - R_{\text{Bayes}}^*] \leq R_{\phi}(\gamma, Q) - R_{\phi}^*. \quad 24$$

25     This result (22) demonstrates that in order to achieve joint Bayes consistency— 25  
 26 that is, in order to drive the excess Bayes risk to zero, while optimizing over the 26  
 27 pair  $(\gamma, Q)$ —it suffices to drive the excess  $\phi$ -risk to zero. 27  
 28  
 29  
 30

31     **4. Comparison between loss functions.** We have studied a broad class of 31  
 32 loss functions corresponding to  $f$ -divergences of the form (12) in Theorem 1. 32  
 33 A subset of this class in turn yields Bayes consistency for the estimation pro- 33  
 34 cedure (13) as shown in Theorem 2. A natural question is, are there any other 34  
 35 surrogate loss functions that also yield Bayes consistency? 35

36     A necessary condition for achieving Bayes consistency using estimation proce- 36  
 37 dure (13) is that the constrained minimization over surrogate  $\phi$ -risks should yield a 37  
 38  $(Q, \gamma)$  pair that minimizes the expected 0–1 loss subject to the same constraints. In 38  
 39 this section, we show that only surrogate loss functions that induce  $f$ -divergence 39  
 40 of the form (12) can actually satisfy this property. We establish this result by de- 40  
 41 veloping a general way of comparing different loss functions. In particular, by ex- 41  
 42 ploiting the correspondence between surrogate losses and  $f$ -divergences, we are 42  
 43 able to compare surrogate losses in terms of their corresponding  $f$ -divergences. 43

1 4.1. *Connection between 0–1 loss and  $f$ -divergences.* The connection be- 1  
 2 tween  $f$ -divergences and 0–1 loss that we develop has its origins in seminal work 2  
 3 on comparison of experiments by Blackwell and others [3–5]. In particular, we 3  
 4 give the following definition. 4

5  
 6 DEFINITION 2. The quantizer  $Q_1$  dominates  $Q_2$  if  $R_{\text{Bayes}}(Q_1) \leq R_{\text{Bayes}}(Q_2)$  6  
 7 for any choice of prior probability  $q = \mathbb{P}(Y = -1) \in (0, 1)$ . 7  
 8

9 Recall that a choice of quantizer design  $Q$  induces two conditional distribu- 9  
 10 tions, say  $P(Z|Y = 1) \sim P_1$  and  $P(Z|Y = -1) \sim P_{-1}$ . From here onward, we use 10  
 11  $P_{-1}^Q$  and  $P_1^Q$  to denote the fact that both  $P_{-1}$  and  $P_1$  are determined by the spe- 11  
 12 cific choice of  $Q$ . By “parameterizing” the decision-theoretic criterion in terms 12  
 13 of loss function  $\phi$  and establishing a precise correspondence between  $\phi$  and the 13  
 14  $f$ -divergence, we obtain an arguably simpler proof of the classical theorem [3, 4] 14  
 15 that relates 0–1 loss to  $f$ -divergences. 15  
 16

17 PROPOSITION 1 [3, 4]. For any two quantizer designs  $Q_1$  and  $Q_2$ , the follow- 17  
 18 ing statements are equivalent: 18  
 19

- 20 (a)  $Q_1$  dominates  $Q_2$  [i.e.,  $R_{\text{Bayes}}(Q_1) \leq R_{\text{Bayes}}(Q_2)$  for any prior probability 20  
 21  $q \in (0, 1)$ ]; 21  
 22 (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  $f(u) =$  22  
 23  $-\min(u, c)$  for some  $c > 0$ ; 23  
 24 (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  $f$ . 24  
 25

26 PROOF. We first establish the equivalence (a)  $\Leftrightarrow$  (b). By the correspon- 26  
 27 dence between 0–1 loss and an  $f$ -divergence with  $f(u) = -\min(u, 1)$ , we have 27  
 28  $R_{\text{Bayes}}(Q) = -I_f(\mu, \pi) = -I_{f_q}(P_1, P_{-1})$ , where  $f_q(u) := qf(\frac{1-q}{q}u) = -(1 -$  28  
 29  $q) \min(u, \frac{q}{1-q})$ . Hence, (a)  $\Leftrightarrow$  (b). 29  
 30

31 Next, we prove the equivalence (b)  $\Leftrightarrow$  (c). The implication (c)  $\Rightarrow$  (b) is im- 31  
 32 mediate. Considering the reverse implication (b)  $\Rightarrow$  (c), we note that any convex 32  
 33 function  $f(u)$  can be uniformly approximated over a bounded interval as a sum 33  
 34 of a linear function and  $-\sum_k \alpha_k \min(u, c_k)$ , where  $\alpha_k > 0, c_k > 0$  for all  $k$ . For 34  
 35 a linear function  $f$ ,  $I_f(P_{-1}, P_1)$  does not depend on  $P_{-1}, P_1$ . Using these facts, 35  
 36 (c) follows easily from (b).  $\square$  36  
 37

38 COROLLARY 1. The quantizer  $Q_1$  dominates  $Q_2$  if and only if  $R_\phi(Q_1) \leq$  38  
 39  $R_\phi(Q_2)$  for any loss function  $\phi$ . 39  
 40

41 PROOF. By Theorem 1(a), we have  $R_\phi(Q) = -I_f(\mu, \pi) = -I_{f_q}(P_1, P_{-1})$ , 41  
 42 from which the corollary follows, using Proposition 1.  $\square$  42  
 43

1 Corollary 1 implies that if  $R_\phi(Q_1) \leq R_\phi(Q_2)$  for some loss function  $\phi$ , then  
 2  $R_{\text{Bayes}}(Q_1) \leq R_{\text{Bayes}}(Q_2)$  for some set of prior probabilities on the hypothesis  
 3 space. This implication justifies the use of a given surrogate loss function  $\phi$  in  
 4 place of the 0–1 loss for *some* prior probability; however, for a given prior proba-  
 5 bility, it gives no guidance on how to choose  $\phi$ . Moreover, the prior probabilities on  
 6 the label  $Y$  are typically unknown in many applications. In such a setting, Black-  
 7 well’s notion of  $Q_1$  dominating  $Q_2$  has limited usefulness. With this motivation in  
 8 mind, the following section is devoted to development of a more stringent method  
 9 for assessing equivalence between loss functions.

10  
 11 4.2. *Universal equivalence.* Suppose that the loss functions  $\phi_1$  and  $\phi_2$  realize  
 12 the  $f$ -divergences associated with the convex function  $f_1$  and  $f_2$ , respectively. We  
 13 then have the following definition.

14  
 15 DEFINITION 3. The surrogate loss functions  $\phi_1$  and  $\phi_2$  are *universally equiv-*  
 16 *alent*, denoted by  $\phi_1 \stackrel{u}{\approx} \phi_2$ , if for any  $\mathbb{P}(X, Y)$  and quantization rules  $Q_1, Q_2$ , there  
 17 holds:

$$18 \quad R_{\phi_1}(Q_1) \leq R_{\phi_1}(Q_2) \Leftrightarrow R_{\phi_2}(Q_1) \leq R_{\phi_2}(Q_2). \quad 18$$

19  
 20 In terms of the corresponding  $f$ -divergences, this relation is denoted by  $f_1 \stackrel{u}{\approx} f_2$ .  
 21

22 Observe that this definition is very stringent, in that it requires that the order-  
 23 ing between optimal  $\phi_1$  and  $\phi_2$  risks holds for all probability distributions  $\mathbb{P}$  on  
 24  $\mathcal{X} \times \mathcal{Y}$ . However, this stronger notion of equivalence is needed for nonparametric  
 25 approaches to classification, in which the underlying distribution  $\mathbb{P}$  is only weakly  
 26 constrained.

27 The following result provides necessary and sufficient conditions for two  
 28  $f$ -divergences to be universally equivalent.

29  
 30 THEOREM 3. Let  $f_1$  and  $f_2$  be continuous, nonlinear and convex functions  
 31 on  $[0, +\infty) \rightarrow \mathbb{R}$ . Then,  $f_1 \stackrel{u}{\approx} f_2$  if and only if  $f_1(u) = cf_2(u) + au + b$  for some  
 32 constants  $c > 0$  and  $a, b$ .

33  
 34 An important special case is when one of the  $f$ -divergences is the variational  
 35 distance. In this case, we have the following.

36  
 37 COROLLARY 2. (a) All  $f$ -divergences based on continuous convex  
 38  $f : [0, +\infty) \rightarrow \infty$  that are universally equivalent to the variational distance have  
 39 the form

$$40 \quad (23) \quad f(u) = -c \min(u, 1) + au + b \quad \text{for some } c > 0. \quad 40$$

41  
 42 (b) The 0–1 loss is universally equivalent only to those loss functions whose corre-  
 43 sponding  $f$ -divergence is based on a function of the form (23).

The above result establishes that only those surrogate loss functions corresponding to the variational distance yield universal consistency in a strong sense, meaning for any underlying  $\mathbb{P}$  and a constrained class of quantization rules.

**5. Proofs.** In this section, we provide detailed proofs of our main results, as well as some auxiliary results.

*5.1. Proofs of Theorem 1 and auxiliary properties.* Our proof proceeds via connecting some intermediate functions. First, let us define, for each  $\beta$ , the inverse mapping

$$(24) \quad \phi^{-1}(\beta) := \inf\{\alpha : \phi(\alpha) \leq \beta\},$$

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

LEMMA 3. *Suppose that  $\phi$  is a convex loss satisfying assumptions A1, A2 and A3.*

(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  $\phi$  is continuous at  $\phi^{-1}(\beta)$ .*

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

Using the function  $\phi^{-1}$ , we define a new function  $\tilde{\Psi} : \mathbb{R} \rightarrow \overline{\mathbb{R}}$  by

$$(25) \quad \tilde{\Psi}(\beta) := \begin{cases} \phi(-\phi^{-1}(\beta)), & \text{if } \phi^{-1}(\beta) \in \mathbb{R}, \\ +\infty, & \text{otherwise.} \end{cases}$$

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

$$(26) \quad \tilde{\beta}_1 := \inf\{\beta : \tilde{\Psi}(\beta) < +\infty\} \quad \text{and} \quad \tilde{\beta}_2 := \inf\{\beta : \tilde{\Psi}(\beta) = \inf \tilde{\Psi}\}.$$

It is simple to check that  $\inf \phi = \inf \tilde{\Psi} = \phi(\alpha^*)$ , and  $\tilde{\beta}_1 = \phi(\alpha^*)$ ,  $\tilde{\beta}_2 = \phi(-\alpha^*)$ . Furthermore, by construction, we have  $\tilde{\Psi}(\tilde{\beta}_2) = \phi(\alpha^*) = \tilde{\beta}_1$ , as well as  $\tilde{\Psi}(\tilde{\beta}_1) = \phi(-\alpha^*) = \tilde{\beta}_2$ . The following properties of  $\tilde{\Psi}$  are particularly useful for our main results.

LEMMA 4. *Suppose that  $\phi$  is a convex loss satisfying assumptions A1, A2 and A3. We have:*

(a)  *$\tilde{\Psi}$  is strictly decreasing in the interval  $(\tilde{\beta}_1, \tilde{\beta}_2)$ . If  $\phi$  is decreasing, then  $\tilde{\Psi}$  is also decreasing in  $(-\infty, +\infty)$ . In addition,  $\tilde{\Psi}(\beta) = +\infty$  for  $\beta < \tilde{\beta}_1$ .*

(b)  *$\tilde{\Psi}$  is convex in  $(-\infty, \tilde{\beta}_2]$ . If  $\phi$  is a decreasing function, then  $\tilde{\Psi}$  is convex in  $(-\infty, +\infty)$ .*

(c)  *$\tilde{\Psi}$  is lower semi-continuous, and continuous in its domain.*

(d) *For any  $\alpha \geq 0$ ,  $\phi(\alpha) = \tilde{\Psi}(\phi(-\alpha))$ . In particular, there exists  $u^* \in (\tilde{\beta}_1, \tilde{\beta}_2)$  such that  $\tilde{\Psi}(u^*) = u^*$ .*

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

Let us proceed to part (a) of the theorem. The statement for general  $\phi$  has already proved in the derivation preceding the theorem statement. Now, supposing that a decreasing convex surrogate loss  $\phi$  satisfies assumptions A1, A2 and A3, then

$$\begin{aligned} f(u) &= - \inf_{\alpha \in \mathbb{R}} (\phi(-\alpha) + \phi(\alpha)u) \\ &= - \inf_{\{\alpha, \beta \mid \phi^{-1}(\beta) \in \mathbb{R}, \phi(\alpha) = \beta\}} (\phi(-\alpha) + \beta u). \end{aligned}$$

For  $\beta$  such that  $\phi^{-1}(\beta) \in \mathbb{R}$ , there might be more than one  $\alpha$  such that  $\phi(\alpha) = \beta$ . However, our assumption (4) ensures that  $\alpha = \phi^{-1}(\beta)$  results in minimum  $\phi(-\alpha)$ . Hence,

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

By Lemma 4(b), the fact that  $\phi$  is decreasing implies that  $\tilde{\Psi}$  is convex. By convex duality and the lower semicontinuity of  $\tilde{\Psi}$  (from Lemma 4(c)), we can also write

$$(27) \quad \tilde{\Psi}(\beta) = \tilde{\Psi}^{**}(\beta) = f^*(-\beta).$$

Thus,  $\tilde{\Psi}$  is identical to the function  $\Psi$  defined in equation (8). The proof of part (a) is complete, thanks to Lemma 4. Furthermore, it can be shown that  $\phi$  must have the form (9). Indeed, from Lemma 4(d), 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  $\phi$  is decreasing and convex on the interval  $(-\infty, 0]$ , for any  $\alpha \geq 0$ , we can write

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

where  $g$  is some increasing continuous and convex function. From Lemma 4(d), 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  $\phi$  is classification-calibrated, we require that  $\phi$  be differentiable at 0 and  $\phi'(0) < 0$ . These conditions in turn imply that  $g$  must be right-differentiable at  $u^*$ , with  $g'(u^*) > 0$ .

Let us turn to part (b) of the theorem. Since  $f$  is lower semicontinuous by assumption, convex duality allows us to write

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



1 which is equal to  $-I_f(\pi, \mu)$ , thereby showing that the  $f$ -divergence is symmetric. 1

2 (b)  $\Rightarrow$  (c): By assumption, the following relation holds for any measures  $\mu$  2  
3 and  $\pi$ :

$$4 \quad (29) \quad \sum_z \pi(z) f(\mu(z)/\pi(z)) = \sum_z \mu(z) f(\pi(z)/\mu(z)). \quad 4$$

5  
6  
7 Take any instance of  $z = l \in \mathcal{Z}$ , and consider measures  $\mu'$  and  $\pi'$ , which are 7  
8 defined on the space  $\mathcal{Z} - \{l\}$  such that  $\mu'(z) = \mu(z)$  and  $\pi'(z) = \pi(z)$  for all 8  
9  $z \in \mathcal{Z} - \{l\}$ . Since condition (29) also holds for  $\mu'$  and  $\pi'$ , it follows that 9

$$10 \quad \pi(z) f(\mu(z)/\pi(z)) = \mu(z) f(\pi(z)/\mu(z)) \quad 10$$

11  
12 for all  $z \in \mathcal{Z}$  and any  $\mu$  and  $\pi$ . Hence,  $f(u) = uf(1/u)$  for any  $u > 0$ . 12

13 (c)  $\Rightarrow$  (a): It suffices to show that all sufficient conditions specified by Theo- 13  
14 rem 1 are satisfied. 14

15 Since any  $f$ -divergence is defined by applying  $f$  to a likelihood ratio [see defi- 15  
16 nition (2)], we can assume  $f(u) = +\infty$  for  $u < 0$  without loss of generality. Since 16  
17  $f(u) = uf(1/u)$  for any  $u > 0$ , it can be verified using subdifferential calculus [8] 17  
18 that for any  $u > 0$ , there holds 18

$$19 \quad (30) \quad \partial f(u) = f(1/u) + \partial f(1/u) \frac{-1}{u}. \quad 19$$

20  
21 Given some  $u > 0$ , consider any  $v_1 \in \partial f(u)$ . Combined with equation (30) and the 21  
22 equality  $f(u) = uf(1/u)$ , we have 22  
23

$$24 \quad (31) \quad f(u) - v_1 u \in \partial f(1/u). \quad 24$$

25 By definition of conjugate duality,  $f^*(v_1) = v_1 u - f(u)$ . 25

26 Letting  $\Psi(\beta) = f^*(-\beta)$  as in Theorem 1, we have 26  
27

$$28 \quad \begin{aligned} \Psi(\Psi(-v_1)) &= \Psi(f^*(v_1)) = \Psi(v_1 u - f(u)) \\ &= f^*(f(u) - v_1 u) = \sup_{\beta \in \mathbb{R}} (\beta f(u) - \beta v_1 u - f(\beta)). \end{aligned} \quad 28$$

29  
30  
31 Note that from equation (31), the supremum is achieved at  $\beta = 1/u$ , so that we 31  
32 have  $\Psi(\Psi(-v_1)) = -v_1$  for any  $v_1 \in \partial f(u)$  for  $u > 0$ . In other words,  $\Psi(\Psi(\beta)) =$  32  
33  $\beta$  for any  $\beta \in \{-\partial f(u), u > 0\}$ . Convex duality and the definition  $\Psi(\beta) = f^*(-\beta)$  33  
34 imply that  $\beta \in -\partial f(u)$  for some  $u > 0$  if and only if  $-u \in \partial \Psi(\beta)$  for some  $u > 0$ . 34  
35 This condition on  $\beta$  is equivalent to the subdifferential  $\partial \Psi(\beta)$  containing some 35  
36 negative value, which is satisfied by any  $\beta \in (\beta_1, \beta_2)$ , so that  $\Psi(\Psi(\beta)) = \beta$  for 36  
37  $\beta \in (\beta_1, \beta_2)$ . In addition, since  $f(u) = +\infty$  for  $u < 0$ ,  $\Psi$  is a decreasing function. 37  
38 Now, as an application of Theorem 1, we conclude that  $I_f$  is realizable by some 38  
39 (decreasing) surrogate loss function.  $\square$  39  
40

41 The following result establishes a link between (un)boundedness and the prop- 41  
42 erties of the associated  $f$ . 42  
43

COROLLARY 4. Assume that  $\phi$  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,  $\phi$  is unbounded from below if and only if  $f$  is 1-coercive, that is,  $f(x)/\|x\| \rightarrow +\infty$  as  $\|x\| \rightarrow \infty$ .

PROOF.  $\phi$  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. [8].  $\square$

Consequently, for any decreasing and lower-bounded  $\phi$  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 [5] are 1-coercive, meaning that any associated surrogate loss  $\phi$  cannot be bounded below.

5.2. *Proof of Theorem 2.* First let us prove Lemma 2:

PROOF. Since  $\phi$  has form (9), it is easy to check that  $\phi(0) = (c - a - b)/2$ . Now, note that

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

In addition,

$$R_{\phi}(\gamma, Q) - R_{\phi}^* = R_{\phi}(\gamma, Q) - R_{\phi}(Q) + R_{\phi}(Q) - R_{\phi}^*.$$

By Theorem 1(a),

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

Therefore, the lemma will be immediate once we can show that

$$\begin{aligned} \frac{c}{2} \sum_{z: (\mu(z) - \pi(z))\gamma(z) < 0} |\mu(z) - \pi(z)| &\leq R_{\phi}(\gamma, Q) - R_{\phi}(Q) \\ (32) \qquad \qquad \qquad &= \sum_{z \in \mathcal{Z}} \pi(z) \phi(-\gamma(z)) + \mu(z) \phi(\gamma(z)) \\ &\quad - c \min\{\mu(z), \pi(z)\} + ap + bq. \end{aligned}$$

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

$$2 \quad (33) \quad \pi(z)\phi(-\gamma(z)) + \mu(z)\phi(\gamma(z)) \geq \pi(z)\phi(0) + \mu(z)\phi(0). \quad 3$$

4 Indeed, without loss of generality, suppose  $\mu(z) > \pi(z)$ . Since  $\phi$  is classification-  
5 calibrated, the convex function (with respect to  $\alpha$ )  $\pi(z)\phi(-\alpha) + \mu(z)\phi(\alpha)$   
6 achieves its minimum at some  $\alpha \geq 0$ . Hence, for any  $\alpha \leq 0$ ,  $\pi(z)\phi(-\alpha) +$   
7  $\mu(z)\phi(\alpha) \geq \pi(z)\phi(0) + \mu(z)\phi(0)$ . Hence, the statement (33) is proven. The RHS  
8 of equation (32) is lower bounded by

$$9 \quad \sum_{z: (\mu(z) - \pi(z))\gamma(z) < 0} (\pi(z) + \mu(z))\phi(0) - c \min\{\mu(z), \pi(z)\} + ap + bq \quad 10$$

$$11 \quad = \sum_{z: (\mu(z) - \pi(z))\gamma(z) < 0} (\pi(z) + \mu(z)) \frac{c - a - b}{2} - c \min\{\mu(z), \pi(z)\} \quad 12$$

$$13 \quad + ap + bq \quad 14$$

$$15 \quad \geq \frac{c}{2} \sum_{z: (\mu(z) - \pi(z))\gamma(z) < 0} |\mu(z) - \pi(z)| - (a + b)(p + q)/2 + ap + bq \quad 16$$

$$17 \quad = \frac{c}{2} \sum_{z: (\mu(z) - \pi(z))\gamma(z) < 0} |\mu(z) - \pi(z)| + \frac{1}{2}(a - b)(p - q) \quad 18$$

$$19 \quad \geq \frac{c}{2} \sum_{z: (\mu(z) - \pi(z))\gamma(z) < 0} |\mu(z) - \pi(z)|. \quad 20$$

21 This completes the proof of the lemma.  $\square$  22

23 We are now equipped to prove Theorem 2. For part (a), first observe that the  
24 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  
25 changes the values of  $(X_i, Y_i)$  for some index  $i \in \{1, \dots, n\}$ . Hence, applying  
26 McDiarmid's inequality yields concentration around the expected value [14], or  
27 (alternatively stated) we have that, with probability at least  $1 - \delta$ , 28

$$29 \quad (34) \quad \left| \sup_{\gamma \in \mathcal{C}_n, Q \in \mathcal{D}_n} |\hat{R}_\phi(\gamma, Q) - R_\phi(\gamma, Q)| - \mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) \right| \leq M_n \sqrt{2 \ln(1/\delta)/n}. \quad 30$$

31 Suppose that  $R_\phi(\gamma, Q)$  attains its minimum over the compact subset  $(\mathcal{C}_n, \mathcal{D}_n)$   
32 at  $(\gamma_n^\dagger, Q_n^\dagger)$ . Then, using Lemma 2, we have 33

$$34 \quad \frac{c}{2} (R_{\text{Bayes}}(\gamma_n^*, Q_n^*) - R_{\text{Bayes}}^*) \leq R_\phi(\gamma_n^*, Q_n^*) - R_\phi^* \quad 35$$

$$36 \quad = R_\phi(\gamma_n^*, Q_n^*) - R_\phi(\gamma_n^\dagger, Q_n^\dagger) + R_\phi(\gamma_n^\dagger, Q_n^\dagger) - R_\phi^* \quad 36$$

$$37 \quad = R_\phi(\gamma_n^*, Q_n^*) - R_\phi(\gamma_n^\dagger, Q_n^\dagger) + \mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n). \quad 37$$

Hence, using the inequality (34), we have, with probability at least  $1 - \delta$ ,

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

from which Theorem 2(a) follows.

For part (b), this statement follows by applying (a) with  $\delta = 1/n$ .

**5.3. Proof of Theorem 3.** One direction of the theorem (“if”) is easy. We focus on the other direction. The proof relies on the following technical result.

**LEMMA 5.** *Given a continuous convex function  $f: \mathbb{R}^+ \rightarrow \mathbb{R}$ , for any  $u, v \in \mathbb{R}^+$ , define*

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

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

- (1)  $T_f(u, v)$  are nonempty 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 the interval  $(u, v)$ .

Now, let us proceed to prove Theorem 3. The convex function  $f: [0, \infty) \rightarrow \mathbb{R}$  is continuous on  $(0, \infty)$  and hence is almost everywhere differentiable on  $(0, \infty)$  (see [16]). 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  $v$  be an arbitrary 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, we may assume that  $f_1(v) = f_2(v) = 0$ ; if not, we simply consider the 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 5 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:

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

1 In either case, we have

$$2 \quad (uf'_1(u) - f_1(u) - vd_1)(f'_2(u) - d_2) = (uf'_2(u) - f_2(u) - vd_2)(f'_1(u) - d_1). \quad 3$$

4 Let  $g_1, g_2$  be defined by  $f_1(u) = g_1(u) + d_1u$ ,  $f_2(u) = g_2(u) + d_2u$ . Then,   
 5  $(ug'_1(u) - g_1(u) - vd_1)g'_2(u) = (ug'_2(u) - g_2(u) - vd_2)g'_1(u)$ , implying that   
 6  $(g_1(u) + vd_1)g'_2(u) = (g_2(u) + vd_2)g'_1(u)$  for any  $u$  where  $f_1$  and  $f_2$  are both   
 7 differentiable. Since  $u$  and  $v$  can be chosen almost everywhere,  $v$  is chosen so that   
 8 there does not exist any open interval for  $u$  such that  $g_2(u) + vd_2 = 0$ . It follows   
 9 that  $g_1(u) + vd_1 = c(g_2(u) + vd_2)$  for some constant  $c$  and this constant  $c$  has to   
 10 be the same for any  $u$  due to the continuity of  $f_1$  and  $f_2$ . Hence, we have  $f_1(u) =$    
 11  $g_1(u) + d_1u = cg_2(u) + d_1u + cvd_2 - vd_1 = cf_2(u) + (d_1 - cd_2)u + cvd_2 - vd_1$ .   
 12 It is now simple to check that  $c > 0$  is necessary and sufficient for  $I_{f_1}$  and  $I_{f_2}$  to   
 13 have the same monotonicity.   
 14

15 **A. Proof of Lemma 3.** (a) Since  $\phi^{-1}(\beta) < +\infty$ , we have  $\phi(\phi^{-1}(\beta)) =$    
 16  $\phi(\inf\{\alpha : \phi(\alpha) \leq \beta\}) \leq \beta$ , where the final inequality follows from the lower semi-   
 17 continuity of  $\phi$ . If  $\phi$  is continuous at  $\phi^{-1}(\beta)$ , then we have  $\phi^{-1}(\beta) = \min\{\alpha :$    
 18  $\phi(\alpha) = \beta\}$ , in which case we have  $\phi(\phi^{-1}(\beta)) = \beta$ .   
 19

(b) Due to convexity and the inequality  $\phi'(0) < 0$ , it follows that  $\phi$  is a strictly   
 20 decreasing function in  $(-\infty, \alpha^*]$ . Furthermore, for all  $\beta \in \mathbb{R}$  such that  $\phi^{-1}(\beta) <$    
 21  $+\infty$ , we must have  $\phi^{-1}(\beta) \leq \alpha^*$ . Therefore, definition 24 and the (decreasing)   
 22 monotonicity of  $\phi$  imply that for any  $a, b \in \mathbb{R}$ , if  $b \geq a \geq \inf \phi$ , then  $\phi^{-1}(a) \geq$    
 23  $\phi^{-1}(b)$ , which establishes that  $\phi^{-1}$  is a decreasing function. In addition, we have   
 24  $a \geq \phi^{-1}(b)$  if and only if  $\phi(a) \leq b$ .   
 25

Now, due to the convexity of  $\phi$ , applying Jensen's inequality for any  $0 < \lambda < 1$ ,   
 26 we have  $\phi(\lambda\phi^{-1}(\beta_1) + (1 - \lambda)\phi^{-1}(\beta_2)) \leq \lambda\phi(\phi^{-1}(\beta_1)) + (1 - \lambda)\phi(\phi^{-1}(\beta_2)) \leq$    
 27  $\lambda\beta_1 + (1 - \lambda)\beta_2$ . Therefore,   
 28

$$29 \quad \lambda\phi^{-1}(\beta_1) + (1 - \lambda)\phi^{-1}(\beta_2) \geq \phi^{-1}(\lambda\beta_1 + (1 - \lambda)\beta_2), \quad 29$$

30 implying the convexity of  $\phi^{-1}$ .   
 31

32 **B. Proof of Lemma 4.**   
 33

34 **PROOF.** (a) We first prove the statement for the case of a decreasing func-   
 35 tion  $\phi$ . First, if  $a \geq b$  and  $\phi^{-1}(a) \notin \mathbb{R}$ , then  $\phi^{-1}(b) \notin \mathbb{R}$ ; hence,  $\Psi(a) = \Psi(b) =$    
 36  $+\infty$ . If only  $\phi^{-1}(b) \notin \mathbb{R}$ , then clearly  $\Psi(b) \geq \Psi(a)$  [since  $\Psi(b) = +\infty$ ]. If  $a \geq b$ ,   
 37 and both  $\phi^{-1}(a), \phi^{-1}(b) \in \mathbb{R}$ , then, from the previous lemma,  $\phi^{-1}(a) \leq \phi^{-1}(b)$ ,   
 38 so that  $\phi(-\phi^{-1}(a)) \leq \phi(-\phi^{-1}(b))$ , implying that  $\Psi$  is a decreasing function.   
 39

40 We next consider the case of a general function  $\phi$ . For  $\beta \in (\beta_1, \beta_2)$ , we have   
 41  $\phi^{-1}(\beta) \in (-\alpha^*, \alpha^*)$ , and hence  $-\phi^{-1}(\beta) \in (-\alpha^*, \alpha^*)$ . Since  $\phi$  is strictly decreas-   
 42 ing in  $(-\infty, \alpha^*]$ , then  $\phi(-\phi^{-1}(\beta))$  is strictly decreasing in  $(\beta_1, \beta_2)$ . Finally, when   
 43  $\beta < \inf \Psi = \phi(\alpha^*)$ ,  $\phi^{-1}(\beta) \notin \mathbb{R}$ , so  $\Psi(\beta) = +\infty$  by definition.   
 43

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

$$\begin{aligned} & \lambda\Psi(\gamma_1) + (1 - \lambda)\Psi(\gamma_2) \\ &= \lambda\phi(-\phi^{-1}(\gamma_1)) + (1 - \lambda)\phi(-\phi^{-1}(\gamma_2)) \\ &\geq \phi(-\lambda\phi^{-1}(\gamma_1) - (1 - \lambda)\phi^{-1}(\gamma_2)) \quad \text{since } \phi \text{ is convex} \\ &\geq \phi(-\phi^{-1}(\lambda\gamma_1 + (1 - \lambda)\gamma_2)) \\ &= \Psi(\lambda\gamma_1 + (1 - \lambda)\gamma_2), \end{aligned}$$

where the last inequality is due to the convexity of  $\phi^{-1}$  and decreasing  $\phi$ . Hence,  $\Psi$  is a convex function.

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  $\Psi$  is convex in  $(-\infty, \beta_2]$ .

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

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

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

(d) Since  $\phi$  is assumed to be classification-calibrated, Lemma 1 implies that  $\phi$  is differentiable at 0 and  $\phi'(0) < 0$ . Since  $\phi$  is convex, this implies that  $\phi$  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 equation (24),  $\phi^{-1}(\alpha) \leq -\phi^{-1}(\beta)$ . Therefore,

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

We have proved that  $\Psi$  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  $\phi$  is not a decreasing function, then the function  $\Psi$  need not be convex on the entire real line. For instance, the following loss function generates a function  $\Psi$  that is not convex:  $\phi(\alpha) = (1 - \alpha)^2$  when  $\alpha \leq 1$ , 0 when  $\alpha \in [0, 2]$ , and  $\alpha - 2$  otherwise. Then, we have  $\Psi(9) = \phi(2) = 0$ ,  $\Psi(16) = \phi(3) = 1$ ,  $\Psi(25/2) = \phi(-1 + 5/\sqrt{2}) = -3 + 5/\sqrt{2} > (\Psi(9) + \Psi(16))/2$ .  $\square$

## 1 C. Proof of Lemma 5. 1

2  
3 PROOF. Consider a joint distribution  $\mathbb{P}(X, Y)$  defined by  $\mathbb{P}(Y = -1) = q =$  3  
4  $1 - \mathbb{P}(Y = 1)$  and 4

$$5 \quad \mathbb{P}(X|Y = -1) \sim \text{Uniform}[0, b] \quad \text{and} \quad \mathbb{P}(X|Y = 1) \sim \text{Uniform}[a, c], \quad 5$$

6  
7 where  $0 < a < b < c$ . Let  $\mathcal{Z} = \{1, 2\}$ . We assume  $Z$  is produced by a determin- 8  
9 istic quantizer design  $Q$  specified by a threshold  $t \in (a, b)$ ; in particular, we set 9  
10  $Q(z = 1|x) = 1$  when  $x \geq t$ , and  $Q(z = 2|x) = 1$  when  $x < t$ . Under this quan- 10  
11 tizer design, we have 11

$$12 \quad \mu(1) = (1 - q) \frac{t - a}{c - a}; \quad \mu(2) = (1 - q) \frac{c - t}{c - a}; \quad 12$$

$$13 \quad \pi(1) = q \frac{t}{b}; \quad \pi(2) = q \frac{b - t}{b}. \quad 13$$

14  
15 Therefore, the  $f$ -divergence between  $\mu$  and  $\pi$  takes the form 15  
16  
17

$$18 \quad I_f(\mu, \pi) = \frac{qt}{b} f\left(\frac{(t - a)b(1 - q)}{(c - a)tq}\right) + \frac{q(b - t)}{b} f\left(\frac{(c - t)b(1 - q)}{(c - a)(b - t)q}\right). \quad 18$$

19  
20 If  $f_1 \stackrel{u}{\approx} f_2$ , then  $I_{f_1}(\mu, \pi)$  and  $I_{f_2}(\mu, \pi)$  have the same monotonicity property for 21  
22 any  $q \in (0, 1)$ , as well, as for any choice of the parameters  $q$  and  $a < b < c$ . Let 22  
23  $\gamma = \frac{b(1 - q)}{(c - a)q}$ , which can be chosen arbitrarily positive, and then define the function 23  
24  
25

$$26 \quad F(f, t) = tf\left(\frac{(t - a)\gamma}{t}\right) + (b - t)f\left(\frac{(c - t)\gamma}{b - t}\right). \quad 26$$

27  
28 Note that the functions  $F(f_1, t)$  and  $F(f_2, t)$  have the same monotonicity property, 28  
29 for any positive parameters  $\gamma$  and  $a < b < c$ . 29

30 We now claim that  $F(f, t)$  is a convex function of  $t$ . Indeed, using convex du- 30  
31 ality [18],  $F(f, t)$  can be expressed as follows: 31  
32

$$33 \quad F(f, t) = t \sup_{r \in \mathbb{R}} \left\{ \frac{(t - a)\gamma}{t} r - f^*(r) \right\} + (b - t) \sup_{s \in \mathbb{R}} \left\{ \frac{(c - t)\gamma}{b - t} s - f^*(s) \right\} \quad 33$$

$$34 \quad = \sup_{r, s} \{ (t - a)r\gamma - tf^*(r) + (c - t)s\gamma - tf^*(s) \}, \quad 34$$

35  
36 which is a supremum over a linear function of  $t$ , thereby showing that  $F(f, t)$  is 36  
37 convex of  $t$ . 37  
38

39 It follows that both  $F(f_1, t)$  and  $F(f_2, t)$  are subdifferentiable everywhere in 39  
40 their domains; since they have the same monotonicity property, we must have 40  
41

$$42 \quad (35) \quad 0 \in \partial F(f_1, t) \Leftrightarrow 0 \in \partial F(f_2, t). \quad 42$$

1 It can be verified using subdifferential calculus [8] that 1

$$2 \quad \partial F(f, t) = \frac{a\gamma}{t} \partial f\left(\frac{(t-a)\gamma}{t}\right) + f\left(\frac{(t-a)\gamma}{t}\right) 2$$

$$3 \quad - f\left(\frac{(c-t)\gamma}{b-t}\right) + \frac{(c-b)\gamma}{b-t} \partial f\left(\frac{(c-t)\gamma}{b-t}\right). 3$$

$$4 \quad 4$$

$$5 \quad 5$$

$$6 \quad 6$$

7 Letting  $u = \frac{(t-a)\gamma}{t}$ ,  $v = \frac{(c-t)\gamma}{b-t}$ , we have 7

$$8 \quad (36a) \quad 0 \in \partial F(f, t) 8$$

$$9 \quad (36b) \quad \Leftrightarrow 0 \in (\gamma - u)\partial f(u) + f(u) - f(v) + (v - \gamma)\partial f(v) 9$$

$$10 \quad \Leftrightarrow \exists \alpha \in \partial f(u), \beta \in \partial f(v) \text{ s.t.} 10$$

$$11 \quad (36c) 11$$

$$12 \quad 0 = (\gamma - u)\alpha + f(u) - f(v) + (v - \gamma)\beta 12$$

$$13 \quad \Leftrightarrow \exists \alpha \in \partial f(u), \beta \in \partial f(v) \text{ s.t.} 13$$

$$14 \quad (36d) 14$$

$$15 \quad \gamma(\alpha - \beta) = u\alpha - f(u) + f(v) - v\beta 15$$

$$16 \quad (36e) \quad \Leftrightarrow \exists \alpha \in \partial f(u), \beta \in \partial f(v) \text{ s.t. } \gamma(\alpha - \beta) = f^*(\alpha) - f^*(\beta). 16$$

17 By varying our choice of  $q \in (0, 1)$ , the number  $\gamma$  can take any positive value. 17

18 Similarly, by choosing different positive values of  $a, b, c$  (such that  $a < b < c$ ), we 18

19 can ensure that  $u$  and  $v$  can take on any positive real values such that  $u < \gamma < v$ . 19

20 Since equation (35) holds for any  $t$ , it follows that for any triples  $u < \gamma < v$ , (36e) 20

21 holds for  $f_1$  if and only if it also holds for  $f_2$ . 21

22 Considering a fixed pair  $u < v$ , first suppose that the function  $f_1$  is linear on the 22

23 interval  $[u, v]$  with a slope  $s$ . In this case, condition (36e) holds for  $f_1$  and any  $\gamma$  23

24 by choosing  $\alpha = \beta = s$ , which implies that condition (36e) also holds for  $f_2$  for 24

25 any  $\gamma$ . Thus, we deduce that  $f_2$  is also a linear function on the interval  $[u, v]$ . 25

26 Suppose, on the other hand, that  $f_1$  and  $f_2$  are both nonlinear in  $[u, v]$ . Due to 26

27 the monotonicity of subdifferentials, we have  $\partial f_1(u) \cap \partial f_1(v) = \emptyset$  and  $\partial f_2(u) \cap$  27

28  $\partial f_2(v) = \emptyset$ . Consequently, it follows that both  $T_{f_1}(u, v)$  and  $T_{f_2}(u, v)$  are non- 28

29 empty. If  $\gamma \in T_{f_1}(u, v)$ , then condition (36e) holds for  $f_1$  for some  $\gamma$ . Thus, it 29

30 must also hold for  $f_2$  using the same  $\gamma$ , which implies that  $\gamma \in T_{f_2}(u, v)$ . The 30

31 same argument can also be applied with the roles of  $f_1$  and  $f_2$  reversed, so we 31

32 conclude that  $T_{f_1}(u, v) = T_{f_2}(u, v)$ .  $\square$  32

## 33 REFERENCES 33

- 34 [1] ALI, S. M. and SILVEY, S. D. (1966). A general class of coefficients of divergence of one 34
- 35 distribution from another. *J. Roy. Statist. Soc. Ser. B* **28** 131–142. [MR0196777](#) 35
- 36 [2] BARTLETT, P., JORDAN, M. I. and MCAULIFFE, J. D. (2006). Convexity, classification and 36
- 37 risk bounds. *Amer. Statist. Assoc.* **101** 138–156. [MR2268032](#) 37
- 38 [3] BLACKWELL, D. (1951). Comparison of experiments. *Proc. 2nd Berkeley Symp. Probab. Statist.* 38
- 39 **1** 93–102. Univ. California Press, Berkeley. [MR0046002](#) 39
- 40 40
- 41 41
- 42 42
- 43 43

- 1 [4] BLACKWELL, D. (1953). Equivalent comparisons of experiments. *Ann. Statist.* **24** 265–272. 1  
 2 [MR0056251](#) 2
- 3 [5] BRADT, R. and KARLIN, S. (1956). On the design and comparison of certain dichotomous 3  
 4 experiments. *Ann. Statist.* **27** 390–409. [MR0087287](#) 4
- 5 [6] CORTES, C. and VAPNIK, V. (1995). Support-vector networks. *Machine Learning* **20** 273–297. 5  
 6 [7] CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and 6  
 indirect observation. *Studia Sci. Math. Hungar* **2** 299–318. [MR0219345](#) 6
- 7 [8] HIRIART-URRUTY, J. and LEMARÉCHAL, C. (2001). *Fundamentals of Convex Analysis.* 7  
 8 Springer, New York. [MR1865628](#) 8
- 9 [9] JIANG, W. (2004). Process consistency for adaboost. *Ann. Statist.* **32** 13–29. [MR2050999](#) 9
- 10 [10] KAILATH, T. (1967). The divergence and Bhattacharyya distance measures in signal selection. 10  
*IEEE Trans. Comm. Technology* **15** 52–60. 10
- 11 [11] LONGO, M., LOOKABAUGH, T. and GRAY, R. (1990). Quantization for decentralized hypoth- 11  
 12 esis testing under communication constraints. *IEEE Trans. Inform. Theory* **36** 241–255. 12  
 13 [MR1052776](#) 13
- 14 [12] LUGOSI, G. and VAYATIS, N. (2004). On the Bayes-risk consistency of regularized boosting 14  
 15 methods. *Ann. Statist.* **32** 30–55. [MR2051000](#) 15
- 16 [13] MANNOR, S., MEIR, R. and ZHANG, T. (2003). Greedy algorithms for classification— 16  
 17 consistency, convergence rates and adaptivity. *J. Machine Learning Research* **4** 713–741. 17  
[MR2072266](#) 17
- 18 [14] MCDIARMID, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics* 18  
 19 (J. Simons, ed.). Cambridge Univ. Press, Cambridge. [MR1036755](#) 19
- 20 [15] NGUYEN, X., WAINWRIGHT, M. J. and JORDAN, M. I. (2005). Nonparametric decentral- 20  
 21 ized detection using kernel methods. *IEEE Trans. Signal Processing* **53** 4053–4066. 21  
[MR2241114](#) 21
- 22 [16] PHELPS, R. R. (1993). *Convex Functions, Monotone Operators and Differentiability.* Springer, 22  
 23 New York. [MR1238715](#) 23
- 24 [17] POOR, H. V. and THOMAS, J. B. (1977). Applications of Ali-Silvey distance measures in 24  
 25 the design of generalized quantizers for binary decision systems. *IEEE Trans. Comm.* **25**  
 26 893–900. 26
- 27 [18] ROCKAFELLAR, G. (1970). *Convex Analysis.* Princeton Univ. Press, Princeton. [MR0274683](#) 27
- 28 [19] STEINWART, I. (2005). Consistency of support vector machines and other regularized kernel 28  
 machines. *IEEE Trans. Inform. Theory* **51** 128–142. [MR2234577](#) 28
- 29 [20] TOPSOE, F. (2000). Some inequalities for information divergence and related measures of dis- 29  
 30 crimination. *IEEE Trans. Inform. Theory* **46** 1602–1609. [MR1768575](#) 30
- 31 [21] TSITSIKLIS, J. N. (1993). Decentralized detection. In *Advances in Statistical Signal Process-* 31  
 32 *ing* 297–344. JAI Press, Greenwich. 32
- 33 [22] ZHANG, T. (2004). Statistical behavior and consistency of classification methods based on 33  
 convex risk minimization. *Ann. Statist.* **53** 56–134. [MR2051001](#) 33
- 34 34
- |  |  |    |
|--|--|----|
| 35 X. NGUYEN   | M. J. WAINWRIGHT   | 35 |
| 36 DEPARTMENT OF STATISTICAL SCIENCE   | M. I. JORDAN   | 36 |
| 37 DUKE UNIVERSITY   | DEPARTMENT OF STATISTICS   | 36 |
| 38 DURHAM, NORTH CAROLINA 27708  | AND  | 37 |
| 39 AND   | DEPARTMENT OF ELECTRICAL ENGINEERING   | 38 |
| 40 STATISTICAL AND APPLIED MATHEMATICAL  | AND COMPUTER SCIENCES  | 39 |
| 41 SCIENCES INSTITUTE (SAMSI)  | UNIVERSITY OF CALIFORNIA, BERKELEY   | 40 |
| 42 RESEARCH TRIANGLE PARK  | BERKELEY, CALIFORNIA 94720   | 40 |
| 43 DURHAM, NORTH CAROLINA 27709  | USA  | 41 |
| USA  | E-MAIL: <a href="mailto:wainwrig@stat.berkeley.edu">wainwrig@stat.berkeley.edu</a> | 42 |
| E-MAIL: <a href="mailto:xuanlong-nguyen@gmail.com">xuanlong-nguyen@gmail.com</a> | <a href="mailto:jordan@stat.berkeley.edu">jordan@stat.berkeley.edu</a>             | 42 |

## 1 THE LIST OF SOURCE ENTRIES RETRIEVED FROM MATHSCINET 1

2 The list of entries below corresponds to the Reference section of your article and was retrieved  
3 from MathSciNet applying an automated procedure. Please check the list and cross out those  
4 entries which lead to mistaken sources. Please update your references entries with the data  
5 from the corresponding sources, when applicable. More information can be found in the sup-  
6 port page:

7 <http://www.e-publications.org/ims/support/mrhelp.html>.  
8

- 9 [1] ALI, S. M. AND SILVEY, S. D. (1966). A general class of coefficients of divergence of one  
10 distribution from another. *J. Roy. Statist. Soc. Ser. B* **28** 131–142. MR0196777 (33 #4963)  
11 [2] BARTLETT, P. L., JORDAN, M. I., AND MCAULIFFE, J. D. (2006). Convexity, classification,  
12 and risk bounds. *J. Amer. Statist. Assoc.* **101** 138–156. MR2268032 (2008d:62057)  
13 [3] BLACKWELL, D. (1951). Comparison of experiments. In *Proceedings of the Second Berkeley*  
14 *Symposium on Mathematical Statistics and Probability, 1950*. University of California  
15 Press, Berkeley and Los Angeles, 93–102. MR0046002 (13,667f)  
16 [4] BLACKWELL, D. (1953). Equivalent comparisons of experiments. *Ann. Math. Statistics* **24**  
17 265–272. MR0056251 (15,47b)  
18 [5] BRADT, R. N. AND KARLIN, S. (1956). On the design and comparison of certain dichotomous  
19 experiments. *Ann. Math. Statist.* **27** 390–409. MR0087287 (19,332e)  
20 [6] Not Found!  
21 [7] CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and  
22 indirect observations. *Studia Sci. Math. Hungar.* **2** 299–318. MR0219345 (36 #2428)  
23 [8] HIRIART-URRUTY, J.-B. AND LEMARÉCHAL, C. (2001). *Fundamentals of convex analysis*.  
24 Grundlehren Text Editions. Springer-Verlag, Berlin. Abridged version of it Convex analy-  
25 sis and minimization algorithms. I [Springer, Berlin, 1993; MR1261420 (95m:90001)]  
26 and it II [ibid.; MR1295240 (95m:90002)]. MR1865628 (2002i:90002)  
27 [9] JIANG, W. (2004). Process consistency for AdaBoost. *Ann. Statist.* **32** 13–29. MR2050999  
28 (2005d:62106)  
29 [10] Not Found!  
30 [11] LONGO, M., LOOKABAUGH, T. D., AND GRAY, R. M. (1990). Quantization for decentralized  
31 hypothesis testing under communication constraints. *IEEE Trans. Inform. Theory* **36** 241–  
32 255. MR1052776 (91b:94008)  
33 [12] LUGOSI, G. AND VAYATIS, N. (2004). On the Bayes-risk consistency of regularized boosting  
34 methods. *Ann. Statist.* **32** 30–55. MR2051000 (2005d:62107)  
35 [13] MANNOR, S., MEIR, R., AND ZHANG, T. (2004). Greedy algorithms for classification—  
36 consistency, convergence rates, and adaptivity. *J. Mach. Learn. Res.* **4** 713–742.  
37 MR2072266 (2005f:68133)  
38 [14] MCDIARMID, C. (1989). On the method of bounded differences. In *Surveys in combinatorics,*  
39 *1989 (Norwich, 1989)*. London Math. Soc. Lecture Note Ser., Vol. **141**. Cambridge Univ.  
40 Press, Cambridge, 148–188. MR1036755 (91e:05077)  
41 [15] NGUYEN, X., WAINWRIGHT, M. J., AND JORDAN, M. I. (2005). Nonparametric decen-  
42 tralized detection using kernel methods. *IEEE Trans. Signal Process.* **53** 4053–4066.  
43 MR2241114 (2007b:94101)  
44 [16] PHELPS, R. R. (1993). *Convex functions, monotone operators and differentiability*, Second  
45 ed. Lecture Notes in Mathematics, Vol. **1364**. Springer-Verlag, Berlin. MR1238715  
46 (94f:46055)  
47 [17] Not Found!  
48 [18] ROCKAFELLAR, R. T. (1970). *Convex analysis*. Princeton Mathematical Series, No. 28.  
49 Princeton University Press, Princeton, N.J. MR0274683 (43 #445)

1	[19] STEINWART, I. (2005). Consistency of support vector machines and other regularized kernel	1
2	classifiers. <i>IEEE Trans. Inform. Theory</i> <b>51</b> 128–142. MR2234577 (2008e:62107)	2
3	[20] TOPSØE, F. (2000). Some inequalities for information divergence and related measures of	3
4	discrimination. <i>IEEE Trans. Inform. Theory</i> <b>46</b> 1602–1609. MR1768575 (2001i:94032)	4
5	[21] Not Found!	5
6	[22] ZHANG, T. (2004). Statistical behavior and consistency of classification methods based on	6
7	convex risk minimization. <i>Ann. Statist.</i> <b>32</b> 56–85. MR2051001 (2005d:62108)	7
8		8
9		9
10		10
11		11
12		12
13		13
14		14
15		15
16		16
17		17
18		18
19		19
20		20
21		21
22		22
23		23
24		24
25		25
26		26
27		27
28		28
29		29
30		30
31		31
32		32
33		33
34		34
35		35
36		36
37		37
38		38
39		39
40		40
41		41
42		42
43		43

1 META DATA IN THE PDF FILE 1

2 Following information will be included as pdf file Document Properties: 2

3  
 4 **Title** : On surrogate loss functions and f-divergences 4  
 5 **Author** : XuanLong Nguyen, Martin J. Wainwright, Michael I. Jordan 5  
 6 **Subject** : The Annals of Statistics, 0, Vol.0, No.00, 1-32 6  
 7 **Keywords**: 62G10, 68Q32, 62K05, Binary classification, discriminant 7  
 8 analysis, surrogate losses,  $f$ -divergences, Ali-Silvey divergences, 8  
 9 quantizer design, nonparametric decentralized detection, statis- 9  
 10 tical machine learning, Bayes consistency 10

11 THE LIST OF URI ADDRESSES 11

12  
 13 Listed below are all uri addresses found in your paper. The non-active uri addresses, if any, are 13  
 14 indicated as ERROR. Please check and update the list where necessary. The e-mail addresses 14  
 15 are not checked – they are listed just for your information. More information can be found in 15  
 16 the support page: 16

17 <http://www.e-publications.org/ims/support/urihelp.html>. 17

18 200 <http://www.imstat.org/aos/> [2:pp.1,1] OK 18  
 19 200 <http://dx.doi.org/10.1214/08-AOS595> [2:pp.1,1] OK 19  
 20 200 <http://www.imstat.org> [2:pp.1,1] OK 20  
 21 200 <http://www.ams.org/msc/> [2:pp.1,1] OK 21  
 22 --- <mailto:xuanlong-nguyen@gmail.com> [2:pp.29,29] Check skip 22  
 23 --- <mailto:wainwrig@stat.berkeley.edu> [2:pp.29,29] Check skip 23  
 24 --- <mailto:jordan@stat.berkeley.edu> [2:pp.29,29] Check skip 24

25 25  
 26 26  
 27 27  
 28 28  
 29 29  
 30 30  
 31 31  
 32 32  
 33 33  
 34 34  
 35 35  
 36 36  
 37 37  
 38 38  
 39 39  
 40 40  
 41 41  
 42 42  
 43 43