Convexity, Classification, and Risk Bounds

Peter L. BARTLETT, Michael I. JORDAN, and Jon D. MCAULIFFE

Many of the classification algorithms developed in the machine learning literature, including the support vector machine and boosting, can be viewed as minimum contrast methods that minimize a convex surrogate of the 0–1 loss function. The convexity makes these algorithms computationally efficient. The use of a surrogate, however, has statistical consequences that must be balanced against the computational virtues of convexity. To study these issues, we provide a general quantitative relationship between the risk as assessed using the 0–1 loss and the risk as assessed using any nonnegative surrogate loss function. We show that this relationship gives nontrivial upper bounds on excess risk under the weakest possible condition on the loss function—that it satisfies a pointwise form of Fisher consistency for classification. The relationship is based on a simple variational transformation of the loss function that is easy to compute in many applications. We also present a refined version of this result in the case of low noise, and show that in this case, strictly convex loss functions lead to faster rates of convergence of the risk than would be implied by standard uniform convergence arguments. Finally, we present applications of our results to the estimation of convergence rates in function classes that are scaled convex hulls of a finite-dimensional base class, with a variety of commonly used loss functions.

KEY WORDS: Boosting; Convex optimization; Empirical process theory; Machine learning; Rademacher complexity; Support vector machine.

1. INTRODUCTION

Convexity has become an increasingly important theme in applied mathematics and engineering, having taken on a prominent role akin to that played by linearity for many decades. Building on the discovery of efficient algorithms for linear programs, researchers in convex optimization theory have developed computationally tractable methods for large classes of convex programs (Nesterov and Nemirovskii 1994). Many fields in which optimality principles form the core conceptual structure have been changed significantly by the introduction of these new techniques (Boyd and Vandenberghe 2004).

Convexity arises in many guises in statistics as well, notably in properties associated with the exponential family of distributions (Brown 1986). But only recently has the systematic exploitation of the algorithmic consequences of convexity begun in statistics. One applied area in which this trend has been most salient is machine learning, where the focus has been on large-scale statistical models, for which computational efficiency is an imperative. Many of the most prominent methods studied in machine learning make significant use of convexity; in particular, support vector machines (Boser, Guyon, and Vapnik 1992; Cortes and Vapnik 1995; Cristianini and Shawe-Taylor 2000; Schölkopf and Smola 2002), boosting (Freund and Schapire 1997; Collins, Schapire, and Singer 2002; Lebanon and Lafferty 2002), and variational inference for graphical models (Jordan, Ghahramani, Jaakkola, and Saul 1999) are all based directly on ideas from convex optimization. These methods have had significant practical successes in such applied areas as bioinformatics, information management, and signal processing (Feder et al. 2004; Joachims 2002; Schölkopf, Tsuda, and Vert 2003).

If algorithms from convex optimization are to continue to make inroads into statistical theory and practice, we need to understand these algorithms not only from a computational standpoint, but also in terms of their statistical properties. What are the statistical consequences of choosing models and estimation procedures so as to exploit the computational advantages of convexity?

In this article we study this question in the context of discriminant analysis, a topic referred to as *classification* in the machine learning field. We consider the setting in which a covariate vector $X \in \mathcal{X}$ is to be classified according to a binary response $Y \in \{-1, 1\}$. The goal is to choose a discriminant function $f : \mathcal{X} \to \mathbb{R}$, from a class of functions $\mathcal{F}$, such that the sign of $f(X)$ is an accurate prediction of Y under an unknown joint measure P on (X, Y). We focus on 0–1 loss; thus, letting $\ell(\alpha)$ denote an indicator function that is 1 if $\alpha \leq 0$ and 0 otherwise, we wish to choose $f \in \mathcal{F}$ that minimizes the risk $R(f) = \mathbf{E}\ell(Yf(X)) = P(Y \neq \operatorname{sign}(f(X)))$.

Given a sample $\mathbf{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$, it is natural to consider estimation procedures based on minimizing the sample average of the loss, $\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i f(X_i))$. As is well known, however, such a procedure is computationally intractable for many nontrivial classes of functions (see, e.g., Arora, Babai, Stern, and Sweedyk 1997). Indeed, the loss function $\ell(Yf(X))$ is nonconvex in its (scalar) argument, and although not a proof, this suggests a source of the difficulty. Moreover, it suggests that we might base a tractable estimation procedure on minimization of a convex surrogate $\phi(\alpha)$ for the loss. In particular, if $\mathcal{F}$ consists of functions that are linear in a parameter vector $\boldsymbol{\theta}$, then the expectation of $\phi(Yf(X))$ is convex in $\boldsymbol{\theta}$ (by convexity of ϕ and linearity of expectation). Given a convex parameter space, we obtain a convex program and can exploit the methods of convex optimization. A wide variety of classification methods are based on this tactic; in particular, Figure 1 shows the (upper-bounding) convex surrogates associated with the support vector machine (Cortes and Vapnik 1995), AdaBoost (Freund and Schapire 1997), and logistic regression (Friedman, Hastie, and Tibshirani 2000). In the machine learning literature, these convexity-based methods have largely displaced earlier nonconvex methods, such as neural networks.

Peter L. Bartlett is Professor (E-mail: *bartlett@stat.berkeley.edu*) and Michael I. Jordan is Professor (E-mail: *jordan@stat.berkeley.edu*), Department of Statistics and the Computer Science Division, and Jon D. McAuliffe was a Graduate Student (E-mail: *jon@stat.berkeley.edu*), Department of Statistics, University of California, Berkeley, CA 94720, when this work was performed. He is now Assistant Professor, Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104. This work was supported by National Science Foundation grants 0412995 and 0434383 and MURI grant DAAD 190210383 from the ARO. The authors thank Gilles Blanchard, Olivier Bousquet, Pascal Massart, Ron Meir, Shahar Mendelson, Martin Wainwright, and Bin Yu for helpful discussions.

Journal of the American Statistical Association
March 2006, Vol. 101, No. 473, Theory and Methods
DOI 10.1198/016214505000000907

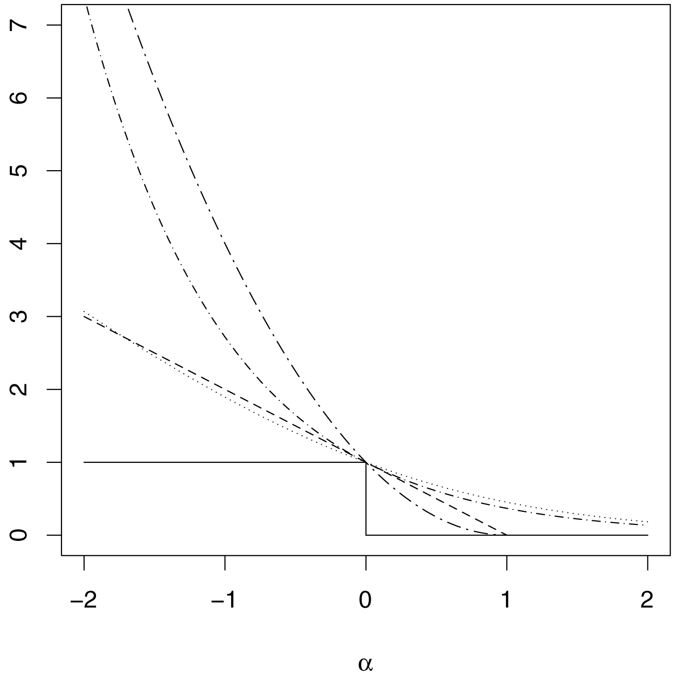

Figure 1. A Plot of the 0–1 Loss Function and Surrogates Corresponding to Various Practical Classifiers (—— 0–1; -·-· exponential; - - - hinge; ······ logistic; ·—· truncated quadratic). These functions are plotted as a function of the margin $\alpha = yf(x)$. Note that a classification error is made if and only if the margin is negative; thus the 0–1 loss is a step function that is equal to 1 for negative values of the abscissa. The curve labeled "logistic" is the negative log-likelihood, or scaled deviance, under a logistic regression model, "hinge" is the piecewise-linear loss used in the support vector machine, and "exponential" is the exponential loss used by the AdaBoost algorithm. The deviance is scaled so as to majorize the 0–1 loss; see Lemma 4.

A basic statistical understanding of the convexity-based setting has begun to emerge. In particular, when appropriate regularization conditions are imposed, it is possible to demonstrate the Bayes-risk consistency of methods based on minimizing convex surrogates for 0–1 loss. Lugosi and Vayatis (2004) provided such a result under the assumption that the surrogate ϕ is differentiable, monotone and strictly convex and satisfies $\phi(0) = 1$. This handles all of the cases shown in Figure 1 except the support vector machine. Steinwart (2005) demonstrated consistency for the support vector machine as well, in a general setting where $\mathcal{F}$ is taken to be a reproducing kernel Hilbert space and ϕ is assumed to be continuous. Other results on Bayes-risk consistency have been presented by Breiman (2004), Jiang (2004), Mannor and Meir (2001), and Mannor, Meir, and Zhang (2002).

Consistency results provide reassurance that optimizing a surrogate does not ultimately hinder the search for a function that achieves the Bayes risk, and thus allow such a search to proceed within the scope of computationally efficient algorithms. There is, however, an additional motivation for working with surrogates of 0–1 loss beyond the computational imperative. Minimizing the sample average of an appropriately behaved loss function has a regularizing effect; it is possible to obtain uniform upper bounds on the risk of a function that minimizes the empirical average of the loss ϕ, even for classes that are so rich that no such upper bounds are possible for the minimizer of the empirical average of the 0–1 loss. Indeed, a number of such results have been obtained for function classes with infinite Vapnik–Chervonenkis (VC) dimension (Bartlett 1998, Shawe-Taylor, Bartlett, Williamson, and Anthony 1998), such as the function classes used by AdaBoost (see, e.g., Schapire, Freund, Bartlett, and Lee 1998; Koltchinskii and Panchenko 2002). These upper bounds provide guidance for model selection and in particular help guide data-dependent choices of regularization parameters.

To carry this agenda further, we need to find general quantitative relationships between the approximation and estimation errors associated with ϕ and those associated with 0–1 loss. This point was emphasized by Zhang (2004), who presented several examples of such relationships. Here we simplify and extend Zhang's results, developing a general methodology for finding quantitative relationships between the risk associated with ϕ and the risk associated with 0–1 loss. In particular, let $R(f)$ denote the risk based on 0–1 loss, and let $R^* = \inf_f R(f)$ denote the Bayes risk. Similarly, let $R_\phi(f) = \mathbf{E}\phi(Yf(X))$ be called the "ϕ-risk," and let $R^*_\phi = \inf_f R_\phi(f)$ denote the "optimal ϕ-risk." We show that for all measurable f,

$$\psi\big(R(f) - R^*\big) \le R_\phi(f) - R^*_\phi, \tag{1}$$

for a nondecreasing function $\psi : [0, 1] \to [0, \infty)$. Moreover, we present a general variational representation of ψ in terms of ϕ, and show that this function is the optimal upper bound of the form (1), in the sense that any other function that satisfies (1) for all measurable f is everywhere no larger than ψ.

This result suggests that if ψ is well-behaved, then minimization of $R_\phi(f)$ may provide a reasonable surrogate for minimization of $R(f)$. Moreover, the result provides a quantitative way to transfer assessments of statistical error in terms of "excess ϕ-risk," $R_\phi(f) - R^*_\phi$, into assessments of error in terms of "excess risk," $R(f) - R^*$.

Although our principal goal is to understand the implications of convexity in classification, we do not impose a convexity assumption on ϕ at the outset. Indeed, whereas such conditions as convexity, continuity, and differentiability of ϕ are easy to verify and have natural relationships to optimization procedures, it is not immediately obvious how to relate such conditions to their statistical consequences. Thus we consider the weakest possible condition on ϕ: that it is "classification-calibrated," which is essentially a pointwise form of Fisher consistency for classification (Lin 2004). In particular, if we define $\eta(x) = P(Y = 1|X = x)$, then ϕ is classification-calibrated if, for x such that $\eta(x) \neq 1/2$, every minimizer f^* of the conditional expectation $\mathbf{E}[\phi(Yf^*(X))|X = x]$ has the same sign as the Bayes decision rule, $\operatorname{sign}(2\eta(x) - 1)$. We show that the upper bound (1) on excess risk in terms of excess ϕ-risk is nontrivial precisely when ϕ is classification-calibrated. Obviously, no such bound is possible when ϕ is not classification-calibrated.

The difficulty of a pattern classification problem is closely related to the behavior of the posterior probability $\eta(X)$. In many practical problems, it is reasonable to assume that for most X, $\eta(X)$ is not too close to 1/2. Mammen and Tsybakov (1999) introduced an elegant formulation of such an assumption, and Tsybakov (2004) considered the convergence rate of the risk of a function that minimizes empirical risk over some fixed class $\mathcal{F}$. He showed that under the assumption of low noise, the risk converges surprisingly quickly to the minimum over the class. If the minimum risk is nonzero, then we might expect a convergence rate no faster than $1/\sqrt{n}$. However, under Tsybakov's assumption, the convergence rate can be as fast as $1/n$. We show that minimizing the empirical ϕ-risk also leads to surprisingly fast convergence rates under this assumption. In particular, if ϕ is uniformly convex, then the minimizer of the empirical ϕ-risk has ϕ-risk that converges quickly to its optimal value, and the noise assumption allows an improvement in the relationship between excess ϕ-risk and excess risk.

These results suggest a general interpretation of pattern classification methods involving a convex contrast function. It is common to view the excess risk as a combination of an estimation term and an approximation term,

$$R(f) - R^* = \Big(R(f) - \inf_{g\in\mathcal{F}} R(g)\Big) + \Big(\inf_{g\in\mathcal{F}} R(g) - R^*\Big).$$

However, choosing a function with risk near-minimal over a class $\mathcal{F}$—that is, finding an f for which the estimation term above is close to 0—is, in a minimax setting, equivalent to the problem of minimizing empirical risk, and hence is computationally infeasible for typical classes $\mathcal{F}$ of interest. Indeed, for classes typically used by boosting and kernel methods, the estimation term in this expression does not converge to 0 for the minimizer of the empirical risk. In contrast, we can also split the upper bound on excess risk into an estimation term and an approximation term,

$$\begin{aligned}
&\psi\big(R(f) - R^*\big)\\
&\quad \le R_\phi(f) - R_\phi^*\\
&\quad = \Big(R_\phi(f) - \inf_{g\in\mathcal{F}} R_\phi(g)\Big) + \Big(\inf_{g\in\mathcal{F}} R_\phi(g) - R_\phi^*\Big).
\end{aligned}$$

Often, it is possible to minimize ϕ-risk efficiently. Thus, although finding an f with near-minimal risk might be computationally infeasible, finding an f for which this upper bound on risk is near-minimal can be feasible.

The article is organized as follows. Section 2 presents basic definitions and a statement and proof of (1). It also introduces the convexity assumption, and shows how it simplifies the computation of ψ. Section 3 presents a refined version of our main result in the setting of low noise. Section 4 presents bounds on the rate of convergence of the ϕ-risk of the empirical minimizer for strictly convex ϕ, and describes applications of these results to convex function classes, such as those used by AdaBoost. It also describes simulations that illustrate the theoretical results. Section 5 presents our conclusions. Proofs of all of our results are presented either in the main text or in Appendix A.

2. RELATING EXCESS RISK TO EXCESS ϕ-RISK

There are three sources of error to consider in a statistical analysis of classification problems: the classical estimation error due to finite sample size, the classical approximation error due to the size of the function space $\mathcal{F}$, and an additional source of approximation error due to the use of a surrogate in place of the 0–1 loss function. It is this last source of error that is our focus in this section. Thus, throughout this section, we work with population expectations and assume that $\mathcal{F}$ is the set of all measurable functions. This allows us to ignore errors due to the size of the sample and the size of the function space, and focus on the error due to the use of a surrogate for the 0–1 loss function.

We follow the tradition in the classification literature and refer to the function ϕ as a loss function, because it is a function that is to be minimized to obtain a discriminant. More precisely, $\phi(Yf(X))$ is generally referred to as a "margin-based loss function," where the quantity $Yf(X)$ is known as the "margin." (It is worth noting that margin-based loss functions are rather different from distance metrics, a point that we explore in App. B.)

This ambiguity in the use of "loss" will not confuse; in particular, we will be careful to distinguish the risk, which is an expectation over 0–1 loss, from the "ϕ-risk," which is an expectation over ϕ. Our goal in this section is to relate these two quantities.

2.1 Setup

Let $(\mathcal{X} \times \{-1, 1\}, \mathcal{G} \otimes 2^{\{-1,1\}}, P)$ be a probability space. Let X be the identity function on $\mathcal{X}$, and let Y be the identity function on $\{-1, 1\}$, so that P is the distribution of (X, Y); that is, for $A \in \mathcal{G} \otimes 2^{\{-1,1\}}$, $P((X, Y) \in A) = P(A)$. Let P_X on $(\mathcal{X}, \mathcal{G})$ be the marginal distribution of X, and let $\eta : \mathcal{X} \to [0, 1]$ be a measurable function such that $\eta(X)$ is a version of $P(Y = 1|X)$. Throughout this section, f is understood as a measurable mapping from $\mathcal{X}$ into $\mathbb{R}$.

Define the $\{0, 1\}$-*risk*, or just *risk*, of f as

$$R(f) = P\big(\operatorname{sign}(f(X)) \ne Y\big),$$

where $\operatorname{sign}(\alpha) = 1$ for $\alpha > 0$ and -1 otherwise. [The particular choice of the value of $\operatorname{sign}(0)$ is not important, but we need to fix some value in $\{\pm 1\}$ for the definitions that follow.] Based on an iid sample $\mathbf{D}_n = ((X_1, Y_1), \dots, (X_n, Y_n))$, we want to choose a function f_n with small risk.

Define the *Bayes risk*, $R^* = \inf_f R(f)$, where the infimum is over all measurable f. Then any f satisfying $\operatorname{sign}(f(X)) = \operatorname{sign}(\eta(X) - 1/2)$ a.s. on $\{\eta(X) \ne 1/2\}$ has $R(f) = R^*$.

Fix a function $\phi : \mathbb{R} \to [0, \infty)$. Define the *$\phi$-risk* of f as

$$R_\phi(f) = \mathbf{E}\phi(Yf(X)).$$

Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$. Let $f_n = \hat{f}_\phi$ be a function in $\mathcal{F}$ that minimizes the empirical expectation of $\phi(Yf(X))$,

$$\hat{R}_\phi(f) = \hat{\mathbf{E}}\phi(Yf(X)) = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i f(X_i)).$$

Thus we treat ϕ as specifying a contrast function that is to be minimized in determining the discriminant function f_n.

2.2 Basic Conditions on the Loss Function

Define the *conditional ϕ-risk*,

$$\mathbf{E}\big(\phi(Yf(X))|X = x\big) = \eta(x)\phi(f(x)) + (1 - \eta(x))\phi(-f(x)) \qquad \text{a.e. } (x).$$

It is useful to think of the conditional ϕ-risk in terms of a generic conditional probability $\eta \in [0, 1]$ and a generic classifier value $\alpha \in \mathbb{R}$. To express this viewpoint, we introduce the *generic conditional ϕ-risk*,

$$C_\eta(\alpha) = \eta\phi(\alpha) + (1 - \eta)\phi(-\alpha).$$

The notation suppresses the dependence on ϕ. The generic conditional ϕ-risk coincides with the conditional ϕ-risk of f at

$x \in \mathcal{X}$ if we take $\eta = \eta(x)$ and $\alpha = f(x)$. Here varying α in the generic formulation corresponds to varying f in the original formulation, for fixed x. As a useful illustration for the definitions that follow, consider a singleton domain $\mathcal{X} = \{x_0\}$. Minimizing the ϕ-risk corresponds to choosing $f(x_0)$ to minimize $C_{\eta(x_0)}(f(x_0))$.

For $\eta \in [0, 1]$, define the *optimal conditional ϕ-risk*,

$$H(\eta) = \inf_{\alpha \in \mathbb{R}} C_\eta(\alpha) = \inf_{\alpha \in \mathbb{R}} \big(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\big).$$

Then the *optimal ϕ-risk* satisfies

$$R_\phi^* := \inf_f R_\phi(f) = \mathbf{E}H(\eta(X)),$$

where the infimum is over measurable functions.

If the infimum in the definition of $H(\eta)$ is uniquely attained for some α, we can define $\alpha^* : [0, 1] \to \mathbb{R}$ by

$$\alpha^*(\eta) = \arg\min_{\alpha \in \mathbb{R}} C_\eta(\alpha).$$

In that case, we define $f_\phi^* : \mathcal{X} \to \mathbb{R}$, up to P_X-null sets, by

$$\begin{aligned} f_\phi^*(x) &= \arg\min_{\alpha \in \mathbb{R}} \mathbf{E}\big(\phi(Y\alpha) | X = x\big) \\ &= \alpha^*(\eta(x)) \end{aligned}$$

and then

$$R_\phi(f_\phi^*) = \mathbf{E}H(\eta(X)) = R_\phi^*.$$

For $\eta \in [0, 1]$, define

$$H^-(\eta) = \inf_{\alpha : \alpha(2\eta - 1) \le 0} C_\eta(\alpha).$$

This is the optimal value of the conditional ϕ-risk, under the constraint that the sign of the argument α disagrees with that of $2\eta - 1$.

We now turn to the basic condition that we impose on ϕ. This condition generalizes the requirement that the minimizer of $C_\eta(\alpha)$ (if it exists) has the correct sign. This is a minimal condition that can be viewed as a pointwise form of Fisher consistency for classification.

Definition 1. We say that ϕ is *classification-calibrated* if, for any $\eta \ne 1/2$,

$$H^-(\eta) > H(\eta).$$

Consider again a singleton domain $\mathcal{X} = \{x_0\}$. Minimizing ϕ-risk corresponds to choosing $f(x_0)$ to minimize $C_{\eta(x_0)}(f(x_0))$. The classification-calibrated condition requires that adding the constraint that $f(x_0)$ has the incorrect sign always leads to a strictly larger ϕ-risk.

Example 1 (Exponential loss). Consider the loss function $\phi(\alpha) = \exp(-\alpha)$ used by AdaBoost. Figure 2(a) shows $\phi(\alpha)$, $\phi(-\alpha)$, and the generic conditional ϕ-risk $C_\eta(\alpha)$ for $\eta = .3$ and $\eta = .7$. In this case ϕ is strictly convex on $\mathbb{R}$, and hence $C_\eta(\alpha)$ is also strictly convex on $\mathbb{R}$, for every η. So C_η is either minimal at a unique stationary point or attains no minimum. Indeed, if $\eta = 0$, then $C_\eta(\alpha) \to 0$ as $\alpha \to -\infty$; if $\eta = 1$, then $C_\eta(\alpha) \to 0$ as $\alpha \to \infty$. Thus we have $H(0) = H(1) = 0$ for exponential loss. For $\eta \in (0, 1)$, solving for the stationary point yields the unique minimizer

$$\alpha^*(\eta) = \frac{1}{2}\log\left(\frac{\eta}{1-\eta}\right).$$

We may then simplify the identity $H(\eta) = C_\eta(\alpha^*(\eta))$ to obtain

$$H(\eta) = 2\sqrt{\eta(1-\eta)}.$$

Note that this expression is also correct for η equal to 0 or 1. Figure 2(b) shows the graphs of α^* and H over the interval [0, 1]. It is easy to check that

$$H^-(\eta) \equiv \exp(0) = 1,$$

and this is strictly greater than $2\sqrt{\eta(1-\eta)}$ when $\eta \ne 1/2$, so the exponential loss is classification-calibrated.

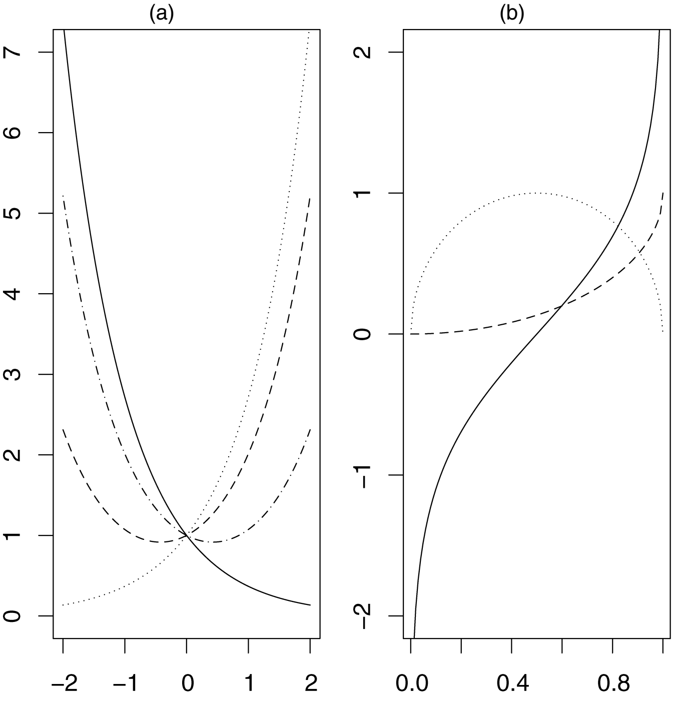

Figure 2. Exponential Loss. (a) $\phi(\alpha)$ (——), its reflection $\phi(-\alpha)$ (⋯⋯), and two different convex combinations of these functions [– – – $C_{.3}(\alpha)$; –·–· $C_{.7}(\alpha)$], for $\eta = .3$ and $\eta = .7$. Note that the minima of these combinations are the values $H(\eta)$, and the minimizing arguments are the values $\alpha^*(\eta)$. (b) $H(\eta)$ (⋯⋯) and $\alpha^*(\eta)$ (——) plotted as a function of η, and the ψ-transform $\psi(\theta)$ (– – –).

2.3 The ψ-Transform and the Relationship Between Excess Risks

We begin by defining a functional transform of the loss function. Then Theorem 1 shows that this transform gives optimal bounds on excess risk in terms of excess ϕ-risk.

Definition 2. We define the *ψ-transform* of a loss function as follows. Given $\phi : \mathbb{R} \to [0, \infty)$, define the function $\psi : [-1, 1] \to [0, \infty)$ by $\psi = \tilde{\psi}^{**}$, where

$$\tilde{\psi}(\theta) = H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right),$$

and $g^{**} : [-1, 1] \to \mathbb{R}$ is the Fenchel–Legendre biconjugate of $g : [-1, 1] \to \mathbb{R}$, which is characterized by

$$\operatorname{epi} g^{**} = \overline{\operatorname{co}} \operatorname{epi} g.$$

Here $\overline{\text{co}}\,S$ is the closure of the convex hull of the set S, and $\text{epi}\,g$ is the epigraph of the function g, that is, the set $\{(x,t) : x \in [0,1], g(x) \leq t\}$. The nonnegativity of ψ is established in Lemma 2, part 7.

Recall that g is convex if and only if $\text{epi}\,g$ is a convex set, and g is closed ($\text{epi}\,g$ is a closed set) if and only if g is lower semicontinuous (Rockafellar 1997). By Lemma 2, part 5, $\tilde{\psi}$ is continuous, so, in fact, the closure operation in Definition 2 is vacuous. We thus have that ψ is simply the functional convex hull of $\tilde{\psi}$ (also known as the greatest convex minorant of $\tilde{\psi}$),

$$\psi = \text{co}\,\tilde{\psi}.$$

This is equivalent to the epigraph convex hull condition of the definition; that is, ψ is the largest convex lower bound on $\tilde{\psi}$. This implies that $\psi = \tilde{\psi}$ if and only if $\tilde{\psi}$ is convex; see Example 9 for a loss function where $\tilde{\psi}$ is not convex.

The importance of the ψ-transform is shown by the following theorem.

Theorem 1.

1. For any nonnegative loss function ϕ, any measurable $f : \mathcal{X} \to \mathbb{R}$, and any probability distribution on $\mathcal{X} \times \{\pm 1\}$,
$$\psi\big(R(f) - R^*\big) \leq R_\phi(f) - R_\phi^*.$$
2. Suppose that $|\mathcal{X}| \geq 2$. For any nonnegative loss function ϕ, any $\epsilon > 0$, and any $\theta \in [0,1]$, there is a probability distribution on $\mathcal{X} \times \{\pm 1\}$ and a function $f : \mathcal{X} \to \mathbb{R}$ such that
$$R(f) - R^* = \theta$$
and
$$\psi(\theta) \leq R_\phi(f) - R_\phi^* \leq \psi(\theta) + \epsilon.$$
3. The following conditions are equivalent:
   a. ϕ is classification-calibrated.
   b. For any sequence (θ_i) in $[0,1]$,
   $$\psi(\theta_i) \to 0 \qquad \text{if and only if} \qquad \theta_i \to 0.$$
   c. For every sequence of measurable functions $f_i : \mathcal{X} \to \mathbb{R}$ and every probability distribution on $\mathcal{X} \times \{\pm 1\}$,
   $$R_\phi(f_i) \to R_\phi^* \quad \text{implies that} \quad R(f_i) \to R^*.$$

Here we mention that classification-calibration implies ψ is invertible on $[0,1]$, so in that case it is meaningful to write the upper bound on excess risk in Theorem 1, part 1 as $\psi^{-1}(R_\phi(f) - R_\phi^*)$. Invertibility follows from convexity of ψ together with Lemma 2, parts 6, 8, and 9.

Zhang (2004) has given a comparison theorem like parts 1 and 3b of this theorem for convex ϕ that satisfy certain conditions. These conditions imply an assumption on the rate of growth (and convexity) of $\tilde{\psi}$. Lugosi and Vayatis (2004) showed that a limiting result like part 3c holds for strictly convex, differentiable, monotonic ϕ. The following theorem shows that if ϕ is convex, classification-calibration is equivalent to a simple derivative condition on ϕ at 0. Clearly, the conclusions of Theorem 1 hold under weaker conditions than those assumed by Zhang (2004) or Lugosi and Vayatis (2004). Steinwart (2005) has shown that if ϕ is continuous and classification-calibrated, then $R_\phi(f_i) \to R_\phi^*$ implies that $R(f_i) \to R^*$. Theorem 1 shows that we may obtain a more quantitative statement of the relationship between these excess risks, under weaker conditions.

It is useful to note that when ϕ is convex, classification-calibration is equivalent to a condition on the derivative of ϕ at 0, and in that case the ψ-transform takes a simplified form.

Theorem 2.

1. Let ϕ be convex. Then ϕ is classification-calibrated if and only if it is differentiable at 0 and $\phi'(0) < 0$.
2. If ϕ is convex and classification-calibrated, then
$$\psi(\theta) = \phi(0) - H\left(\frac{1+\theta}{2}\right).$$

In the remainder of this section we present two preliminary lemmas and then present a proof of Theorem 1. Note that Section 3 presents several examples of calculations of the ψ-transform; some readers may want to visit that section first before proceeding to the proof.

The following elementary lemma will be useful throughout the article.

Lemma 1. Suppose that $g : \mathbb{R} \to \mathbb{R}$ is convex and $g(0) = 0$. Then

1. For all $\lambda \in [0,1]$ and $x \in \mathbb{R}$,
$$g(\lambda x) \leq \lambda g(x).$$
2. For all $x > 0$, $0 \leq y \leq x$,
$$g(y) \leq \frac{y}{x} g(x).$$
3. $g(x)/x$ is increasing on $(0, \infty)$.

Proof. For part 1, $g(\lambda x) = g(\lambda x + (1-\lambda)0) \leq \lambda g(x) + (1-\lambda)g(0) = \lambda g(x)$. To see part 2, put $\lambda = y/x$ in 1. For part 3, rewrite part 2 as $g(y)/y \leq g(x)/x$.

Lemma 2. For any nonnegative loss function ϕ, the functions H, H^-, and ψ have the following properties:

1. H and H^- are symmetric about $1/2$ and ψ is symmetric about 0. For all $\eta \in [0,1]$,
$$H(\eta) = H(1-\eta), \qquad H^-(\eta) = H^-(1-\eta),$$
and
$$\psi(\eta) = \psi(-\eta).$$
2. H is concave and, for $0 \leq \eta \leq 1$, it satisfies
$$H(\eta) \leq H\left(\frac{1}{2}\right) = H^-\left(\frac{1}{2}\right).$$
3. If ϕ is classification-calibrated, then $H(\eta) < H(1/2)$ for all $\eta \neq 1/2$.
4. H^- is concave on $[0, 1/2]$ and on $[1/2, 1]$, and, for $0 \leq \eta \leq 1$, it satisfies
$$H^-(\eta) \geq H(\eta).$$
5. H and H^- are continuous on $[0,1]$.
6. ψ and $\tilde{\psi}$ are continuous on $[-1,1]$.
7. ψ is nonnegative and minimal at 0.
8. $\psi(0) = 0$.
9. The following statements are equivalent:
   a. ϕ is classification-calibrated.
   b. $\psi(\theta) > 0$ for all $\theta \in (0,1]$.

For the proof see Appendix A.

Proof of Theorem 1. For part 1, it is straightforward to show that

$$\begin{aligned}R(f) - R^* &\\ &= R(f) - R(\eta - 1/2)\\ &= \mathbf{E}\big(\mathbb{1}\big[\operatorname{sign}(f(X)) \neq \operatorname{sign}(\eta(X) - 1/2)\big]|2\eta(X) - 1|\big),\end{aligned}$$

where $\mathbb{1}[\Phi]$ is 1 if the predicate Φ is true and 0 otherwise (see, e.g., Devroye, Györfi, and Lugosi 1996). We can apply Jensen's inequality, because ψ is convex by definition, and the fact that $\psi(0) = 0$ (Lemma 2, part 8) to show that

$$\begin{aligned}\psi\big(R(f) - R^*\big) &\\ &\le \mathbf{E}\psi\big(\mathbb{1}\big[\operatorname{sign}(f(X)) \neq \operatorname{sign}(\eta(X) - 1/2)\big]|2\eta(X) - 1|\big)\\ &= \mathbf{E}\big(\mathbb{1}\big[\operatorname{sign}(f(X)) \neq \operatorname{sign}(\eta(X) - 1/2)\big]\psi\big(|2\eta(X) - 1|\big)\big).\end{aligned}$$

Now, from the definition of ψ, we know that $\psi(\theta) \le \tilde{\psi}(\theta)$, so we have

$$\begin{aligned}\psi\big(R(f) - R^*\big) &\\ &\le \mathbf{E}\big(\mathbb{1}\big[\operatorname{sign}(f(X)) \neq \operatorname{sign}(\eta(X) - 1/2)\big]\tilde{\psi}\big(|2\eta(X) - 1|\big)\big)\\ &= \mathbf{E}\big(\mathbb{1}\big[\operatorname{sign}(f(X)) \neq \operatorname{sign}(\eta(X) - 1/2)\big]\\ &\qquad \times \big(H^-(\eta(X)) - H(\eta(X))\big)\big)\\ &= \mathbf{E}\Big(\mathbb{1}\big[\operatorname{sign}(f(X)) \neq \operatorname{sign}(\eta(X) - 1/2)\big]\\ &\qquad \times \Big(\inf_{\alpha:\alpha(2\eta(X)-1)\le 0} C_{\eta(X)}(\alpha) - H(\eta(X))\Big)\Big)\\ &\le \mathbf{E}\big(C_{\eta(X)}(f(X)) - H(\eta(X))\big)\\ &= R_\phi(f) - R_\phi^*,\end{aligned}$$

where we have used the fact that for any x, and in particular when $\operatorname{sign}(f(x)) = \operatorname{sign}(\eta(x) - 1/2)$, we have $C_{\eta(x)}(f(x)) \ge H(\eta(x))$.

For part 2, the first inequality is from part 1. For the second inequality, fix $\epsilon > 0$ and $\theta \in [0, 1]$. From the definition of ψ, we can choose $\gamma, \alpha_1, \alpha_2 \in [0, 1]$, for which $\theta = \gamma\alpha_1 + (1 - \gamma)\alpha_2$ and $\psi(\theta) \ge \gamma\tilde{\psi}(\alpha_1) + (1 - \gamma)\tilde{\psi}(\alpha_2) - \epsilon/2$. Choose distinct $x_1, x_2 \in \mathcal{X}$, and choose P_X such that $P_X\{x_1\} = \gamma$, $P_X\{x_2\} = 1 - \gamma$, $\eta(x_1) = (1 + \alpha_1)/2$, and $\eta(x_2) = (1 + \alpha_2)/2$. From the definition of H^-, we can choose $f : \mathcal{X} \to \mathbb{R}$ such that $f(x_1) \le 0$, $f(x_2) \le 0$, $C_{\eta(x_1)}(f(x_1)) \le H^-(\eta(x_1)) + \epsilon/2$, and $C_{\eta(x_2)}(f(x_2)) \le H^-(\eta(x_2)) + \epsilon/2$. Then we have

$$\begin{aligned}R_\phi(f) - R_\phi^* &\\ &= \mathbf{E}\phi(Yf(X)) - \inf_g \mathbf{E}\phi(Yg(X))\\ &= \gamma\big(C_{\eta(x_1)}(f(x_1)) - H(\eta(x_1))\big)\\ &\qquad + (1 - \gamma)\big(C_{\eta(x_2)}(f(x_2)) - H(\eta(x_2))\big)\\ &\le \gamma\big(H^-(\eta(x_1)) - H(\eta(x_1))\big)\\ &\qquad + (1 - \gamma)\big(H^-(\eta(x_2)) - H(\eta(x_2))\big) + \epsilon/2\\ &= \gamma\tilde{\psi}(\alpha_1) + (1 - \gamma)\tilde{\psi}(\alpha_2) + \epsilon/2\\ &\le \psi(\theta) + \epsilon.\end{aligned}$$

Furthermore, because $\operatorname{sign}(f(x_1)) = \operatorname{sign}(f(x_2)) = -1$ but $\eta(x_1), \eta(x_2) \ge 1/2$,

$$\begin{aligned}R(f) - R^* &= \mathbf{E}|2\eta(X) - 1|\\ &= \gamma(2\eta(x_1) - 1) + (1 - \gamma)(2\eta(x_2) - 1)\\ &= \theta.\end{aligned}$$

For part 3, first note that for any ϕ, ψ is continuous on [0, 1] and $\psi(0) = 0$ by Lemma 2, parts 6 and 8, and hence $\theta_i \to 0$ implies that $\psi(\theta_i) \to 0$. Thus we can replace condition 3b by

3b′. For any sequence (θ_i) in [0, 1],

$$\psi(\theta_i) \to 0 \quad \text{implies that} \quad \theta_i \to 0.$$

To see that part (3a) implies 3b′, let ϕ be classification-calibrated, and let (θ_i) be a sequence that does not converge to 0. Define $c = \limsup \theta_i > 0$, and pass to a subsequence with $\lim \theta_i = c$. Then $\lim \psi(\theta_i) = \psi(c)$ by continuity, and $\psi(c) > 0$ by classification-calibration (Lemma 2, part 9). Thus, for the original sequence (θ_i), we see $\limsup \psi(\theta_i) > 0$, so we cannot have $\psi(\theta_i) \to 0$.

To see that part 3b′ implies 3c, suppose that $R_\phi(f_i) \to R_\phi^*$. By part 1, $\psi(R(f_i) - R^*) \to 0$, and part 3b′ implies that $R(f_i) \to R^*$.

Finally, to see that part 3c implies part 3a, suppose that ϕ is not classification-calibrated. By definition, we can choose $\eta \neq 1/2$ and a sequence $\alpha_1, \alpha_2, \ldots$ such that $\operatorname{sign}(\alpha_i(\eta - 1/2)) = -1$ but $C_\eta(\alpha_i) \to H(\eta)$. Fix $x \in \mathcal{X}$ and choose the probability distribution P so that $P_X\{x\} = 1$ and $P(Y = 1|X = x) = \eta$. Define a sequence of functions $f_i : \mathcal{X} \to \mathbb{R}$ for which $f_i(x) = \alpha_i$. Then $\lim R(f_i) > R^*$, and this is true for any infinite subsequence. But $C_\eta(\alpha_i) \to H(\eta)$ implies that $R_\phi(f_i) \to R_\phi^*$.

2.4 Examples

In this section we present several examples of the computation of the ψ-transform.

Example 2 (Exponential loss). Because $\phi(\alpha) = \exp(-\alpha)$ is convex, differentiable, and decreasing, Theorem 2, part 1 implies that it is classification-calibrated, as we have seen. We also noted that $H(\eta) = 2\sqrt{\eta(1 - \eta)}$. From Theorem 2, part 2,

$$\psi(\theta) = 1 - \sqrt{1 - \theta^2}.$$

Figure 2(b) shows the graph of ψ over the interval [0, 1]. (From Lemma 2, part 1, $\psi(\theta) = \psi(-\theta)$ for any ψ and any $\theta \in [-1, 1]$.)

Example 3 (Truncated quadratic loss). Now consider $\phi(\alpha) = (\max\{1 - \alpha, 0\})^2$, as depicted together with $\phi(-\alpha)$, $C_{.3}(\alpha)$, and $C_{.7}(\alpha)$ in Figure 3(a). This function is convex, differentiable, and decreasing at zero, and thus is classification-calibrated. If $\eta = 0$, then it is clear that any $\alpha \in (-\infty, -1]$ makes $C_\eta(\alpha)$ vanish. Similarly, any $\alpha \in [1, \infty)$ makes the conditional ϕ-risk vanish when $\eta = 1$. But when $0 < \eta < 1$, C_η is strictly convex with a (unique) stationary point, and solving for it yields

$$\alpha^*(\eta) = 2\eta - 1. \tag{2}$$

Note that, although α^* is in principle undefined at 0 and 1, we could choose to fix $\alpha^*(0) = -1$ and $\alpha^*(1) = 1$, which are valid settings. This would extend (2) to all of [0, 1].

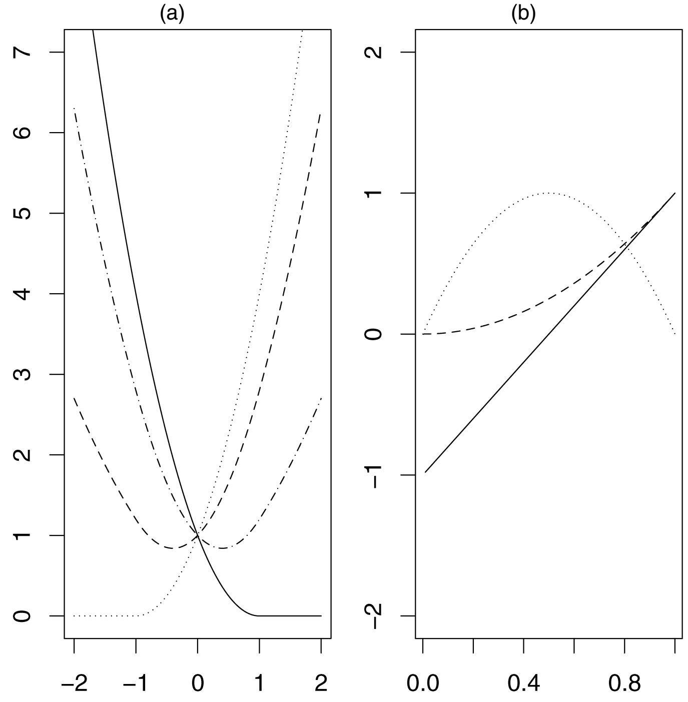

Figure 3. Truncated Quadratic Loss. (a) —— $\phi(\alpha)$; ······· $\phi(-\alpha)$; - - - $C_{.3}(\alpha)$; -·-· $C_{.7}(\alpha)$. (b) —— $\alpha^*(\eta)$; ······· $H(\eta)$; - - - $\Psi(\theta)$.

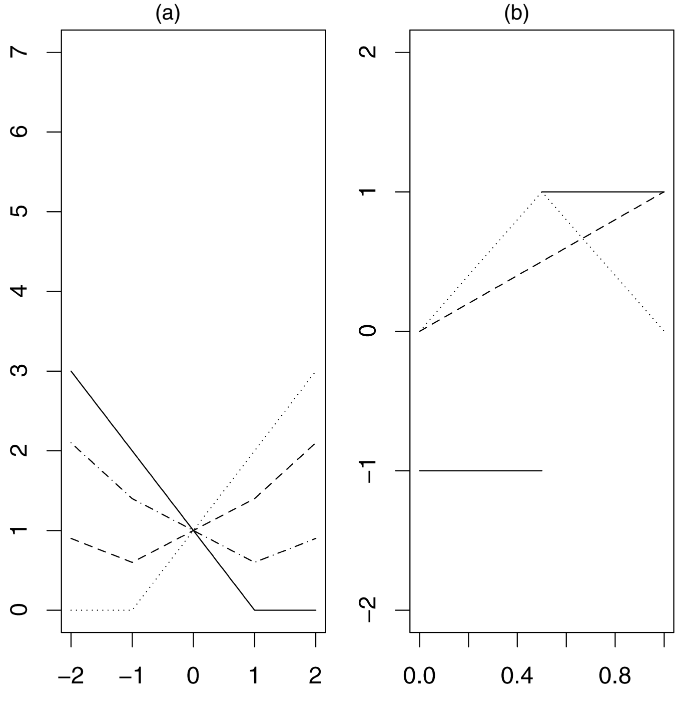

Figure 4. Hinge Loss. (a) —— $\phi(\alpha)$; ······· $\phi(-\alpha)$; - - - $C_{.3}(\alpha)$; -·-· $C_{.7}(\alpha)$. (b) —— $\alpha^*(\eta)$; ······· $H(\eta)$; - - - $\Psi(\theta)$.

As in Example 1, we may simplify the identity $H(\eta) = C_\eta(\alpha^*(\eta))$ for $0 < \eta < 1$ to obtain

$$H(\eta) = 4\eta(1-\eta),$$

which is also correct for $\eta = 0$ and 1, as noted. Thus,

$$\psi(\theta) = \theta^2.$$

Figure 3(b) shows α^*, H, and ψ.

Example 4 (Hinge loss). Here we take $\phi(\alpha) = \max\{1-\alpha, 0\}$, which is shown in Figure 4(a) along with $\phi(-\alpha)$, $C_{.3}(\alpha)$, and $C_{.7}(\alpha)$. Again, ϕ is convex and differentiable at 0 and has negative derivative at 0, so it is classification-calibrated. By direct consideration of the piecewise-linear form of $C_\eta(\alpha)$, it is easy to see that for $\eta = 0$, each $\alpha \le -1$ makes $C_\eta(\alpha)$ vanish, just as in Example 3. The same holds for $\alpha \ge 1$ when $\eta = 1$. Now for $\eta \in (0,1)$, we see that C_η decreases strictly on $(-\infty, -1]$ and increases strictly on $[1, \infty)$. Thus any minima must lie in $[-1, 1]$. But C_η is linear on $[-1, 1]$, so the minimum must be attained at 1 for $\eta > 1/2$, -1 for $\eta < 1/2$, and anywhere in $[-1, 1]$ for $\eta = 1/2$. We have argued that

$$\alpha^*(\eta) = \operatorname{sign}(\eta - 1/2) \tag{3}$$

for all $\eta \in (0,1)$ other than 1/2. Because (3) yields valid minima at 0, 1/2, and 1 as well, we could choose to extend it to the entire unit interval. Regardless, a simple direct verification as in the previous examples shows

$$H(\eta) = 2\min\{\eta, 1-\eta\}$$

for $0 \le \eta \le 1$, and so

$$\psi(\theta) = |\theta|.$$

We present α^*, H, and ψ in Figure 4(b).

Example 5 (Distance-weighted discrimination). Marron and Todd (2002) introduced the distance-weighted discrimination method for high-dimensional, small-sample-size problems. This method chooses an element of the unit ball in a reproducing kernel Hilbert space to minimize a certain criterion. It is straightforward to show that this criterion is an empirical ϕ-risk, for the loss function

$$\phi(\alpha) = \begin{cases} \dfrac{1}{\alpha} & \text{if } \alpha \ge \gamma \\ \dfrac{1}{\alpha}\left(2 - \dfrac{\alpha}{\gamma}\right) & \text{otherwise,} \end{cases}$$

where γ is a positive constant. Note that ϕ is convex, differentiable, decreasing, and hence classification-calibrated. It is easy to verify that

$$H(\eta) = \frac{1}{\eta}(1 + 2\min\{\eta, 1-\eta\}),$$

and hence

$$\psi(\theta) = \frac{|\theta|}{\gamma}.$$

Example 6 (ARC–X4). Breiman (1999) proposed ARC–X4, a boosting algorithm based on the convex cost function

$$\phi(\alpha) = |1-\alpha|^5.$$

More generally, consider the function $\phi(\alpha) = |1-\alpha|^p$ for $p > 1$. This is convex and has $\phi'(0) < 0$, so it is classification-calibrated. Furthermore, it is easy to verify that for $\eta \in (0,1)$,

$$\alpha^*(\eta) = \frac{\eta^{1/(p-1)} - (1-\eta)^{1/(p-1)}}{\eta^{1/(p-1)} + (1-\eta)^{1/(p-1)}},$$

and so

$$H(\eta) = \frac{2^p \eta(1-\eta)}{((1-\eta)^{1/(p-1)} + \eta^{1/(p-1)})^{p-1}}$$

and

$$\psi(\theta) = \phi(0) - H\left(\frac{1-\theta}{2}\right)$$
$$= 1 - \frac{2^{p-1}(1-\theta^2)}{((1-\theta)^{1/(p-1)} + (1+\theta)^{1/(p-1)})^{p-1}}.$$

Example 7 (Sigmoid loss). We conclude by examining a nonconvex loss function. Let $\phi(\alpha) = 1 - \tanh(k\alpha)$ for some fixed $k > 0$. Figure 5(a) depicts $\phi(\alpha)$ with $k = 1$, as well as $\phi(-\alpha)$, $C_{.3}(\alpha)$, and $C_{.7}(\alpha)$. Using the fact that tanh is an odd function, we can rewrite the conditional ϕ-risk as

$$C_\eta(\alpha) = 1 + (1 - 2\eta)\tanh(k\alpha). \tag{4}$$

From this expression, two facts are clear. First, when $\eta = 1/2$, every α minimizes $C_\eta(\alpha)$, because it is identically 1. Second, when $\eta \neq 1/2$, $C_\eta(\alpha)$ attains no minimum, because tanh has no maximal or minimal value on $\mathbb{R}$. Hence α^* is not defined for any η.

Inspecting (4), for $0 \le \eta < 1/2$ we obtain $H(\eta) = 2\eta$ by letting $\alpha \to -\infty$. Analogously, when $\alpha \to \infty$, we get $H(\eta) = 2(1 - \eta)$ for $1/2 < \eta \le 1$. Thus we have

$$H(\eta) = 2\min\{\eta, 1 - \eta\}, \qquad 0 \le \eta \le 1.$$

Because $H^-((1+\theta)/2) \equiv \phi(0) = 1$, we have

$$\tilde{\psi}(\theta) = |\theta|,$$

and convexity gives $\psi = \tilde{\psi}$. We present H and ψ in Figure 5(b). Finally, the foregoing considerations imply that sigmoid loss is classification-calibrated, provided that we note that the definition of classification-calibration requires nothing when $\eta = 1/2$.

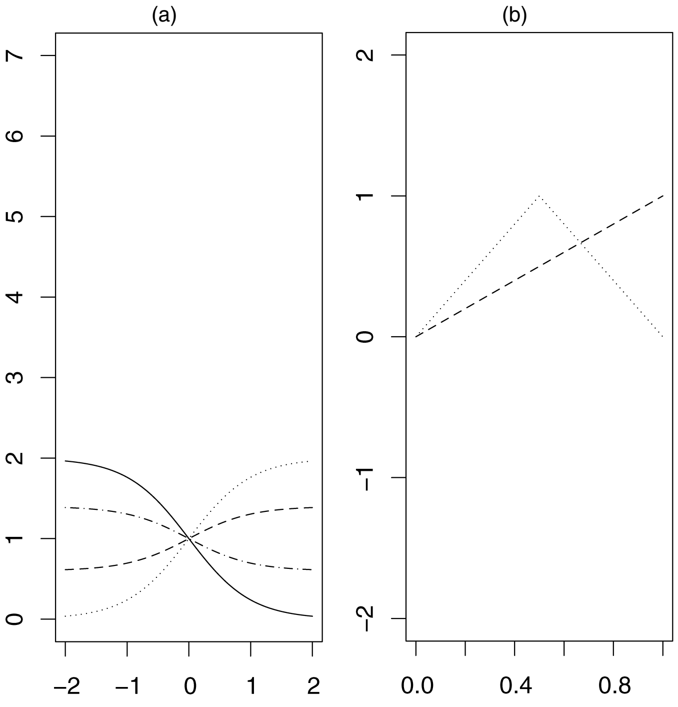

Figure 5. Sigmoid Loss. (a) —— $\phi(\alpha)$; ······· $\phi(-\alpha)$; - - - $C_{.3}(\alpha)$; -·-· $C_{.7}(\alpha)$. (b) ······· $H(\eta)$; - - - $\Psi(\theta)$.

The following example illustrates the difficulties with nondifferentiability at 0, even if ϕ is decreasing and strictly convex.

Example 8. Consider

$$\phi(\alpha) = \begin{cases} e^{-2\alpha} & \text{if } \alpha \le 0 \\ e^{-\alpha} & \text{otherwise.} \end{cases}$$

Then ϕ is strictly convex and decreasing, but not classification-calibrated.

To see this, note that

$$\eta\phi(\alpha) + (1-\eta)\phi(-\alpha) = \begin{cases} \eta e^{-2\alpha} + (1-\eta)e^{\alpha} & \text{if } \alpha \le 0 \\ \eta e^{-\alpha} + (1-\eta)e^{2\alpha} & \text{otherwise.} \end{cases} \tag{5}$$

Taking derivatives and setting to 0 shows that (5) is minimized on the set $\{\alpha \le 0\}$ at

$$\alpha = \min\left(0, \frac{1}{3}\ln\frac{2\eta}{1-\eta}\right).$$

Thus, if $\eta < 1/2$ and $2\eta \ge 1 - \eta$ (i.e., $1/3 \le \eta < 1/2$), then the optimal α is at least 0.

2.5 Further Analysis of ψ

It is interesting to consider what properties of convex cost functions determine the optimal bound ψ on excess risk in terms of excess ϕ-risk. The following lemma shows that a flatter function ϕ leads to a better bound ψ. The measure of curvature that we consider involves the Bregman divergence of ϕ at 0. If ϕ is convex and classification-calibrated, then it is differentiable at 0, and we can define the Bregman divergence of ϕ at 0,

$$d_\phi(0, \alpha) = \phi(\alpha) - \big(\phi(0) + \alpha\phi'(0)\big).$$

We consider a symmetrized, normalized version of the Bregman divergence at 0, for $\alpha > 0$,

$$\xi(\alpha) = \frac{d_\phi(0, \alpha) + d_\phi(0, -\alpha)}{-\phi'(0)\alpha}.$$

Because ϕ is convex on $\mathbb{R}$, both ϕ and ξ are continuous, so we can define

$$\xi^{-1}(\theta) = \inf\{\alpha : \xi(\alpha) = \theta\}.$$

Lemma 3. For convex, classification-calibrated ϕ,

$$\psi(\theta) \ge -\phi'(0)\frac{\theta}{2}\xi^{-1}\left(\frac{\theta}{2}\right).$$

Notice that a slower increase of ξ (i.e., a less curved ϕ) gives better bounds on $R(f) - R^*$ in terms of $R_\phi(f) - R_\phi^*$.

2.6 General Loss Functions

All of the classification procedures mentioned in earlier sections use surrogate loss functions that either are upper bounds on 0–1 loss or can be transformed into upper bounds through a positive scaling factor. This is not a coincidence; as the next lemma establishes, it must be possible to scale any classification-calibrated ϕ into such a majorant.

Lemma 4. If $\phi : \mathbb{R} \to [0, \infty)$ is classification-calibrated, then there is a $\gamma > 0$ such that $\gamma\phi(\alpha) \ge \mathbb{1}[\alpha \le 0]$ for all $\alpha \in \mathbb{R}$.

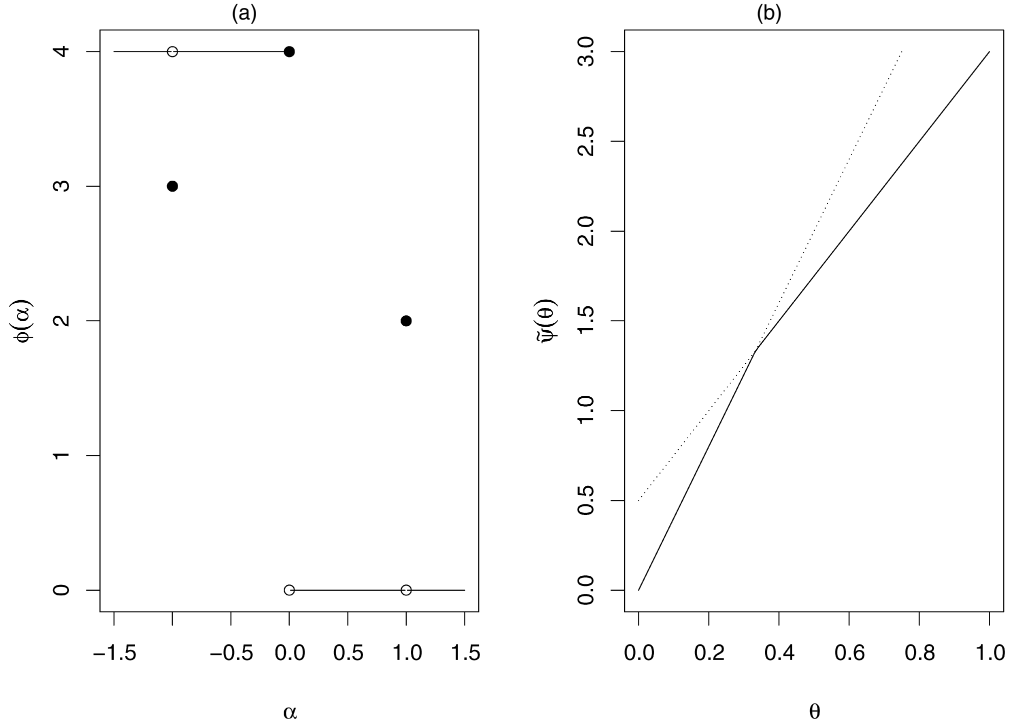

Figure 6. The Loss Function of Example 9 (a) and the Corresponding (nonconvex) $\tilde{\psi}$ (b). The dotted lines depict the graphs for the two linear functions of which $\tilde{\psi}$ is a pointwise minimum.

We have seen that for convex ϕ, the function $\tilde{\psi}$ is convex, and so $\psi = \tilde{\psi}$. The following example shows that, in general, we cannot avoid computing the convex lower bound ψ.

Example 9. Consider the following (classification-calibrated) loss function; see Figure 6(a):

$$\phi(\alpha) = \begin{cases} 4 & \text{if } \alpha \le 0, \alpha \ne -1 \\ 3 & \text{if } \alpha = -1 \\ 2 & \text{if } \alpha = 1 \\ 0 & \text{if } \alpha > 0, \alpha \ne 1. \end{cases}$$

It is easy to check that

$$H^-(\eta) = \begin{cases} \min\{4\eta, 2+\eta\} & \text{if } \eta \ge 1/2 \\ \min\{4(1-\eta), 3-\eta\} & \text{if } \eta < 1/2, \end{cases}$$

and that $H(\eta) = 4\min\{\eta, 1-\eta\}$. Thus

$$H^-(\eta) - H(\eta) = \begin{cases} \min\{8\eta - 4, 5\eta - 2\} & \text{if } \eta \ge 1/2 \\ \min\{4 - 8\eta, 3 - 5\eta\} & \text{if } \eta < 1/2, \end{cases}$$

so

$$\tilde{\psi}(\theta) = \min\left\{4\theta, \frac{1}{2}(5\theta + 1)\right\}.$$

This function, illustrated in Figure 6(b), is not convex; in fact, it is concave. Thus $\psi \ne \tilde{\psi}$.

3. TIGHTER BOUNDS UNDER LOW–NOISE CONDITIONS

Predicting the optimal class label is difficult in regions where $\eta(X)$ is close to $1/2$, because the information provided by the labels is most noisy there. In many practical pattern classification problems, it is reasonable to assume that the posterior probability $\eta(X)$ is unlikely to be very close to $1/2$. Hence it is important to understand how pattern classification methods perform under these "low-noise" conditions. To quantify the notion of low noise, consider the following two properties of a probability distribution on $\mathcal{X} \times \{\pm 1\}$, introduced by Mammen and Tsybakov (1999) and Tsybakov (2004):

M_β. For some c and all $\epsilon > 0$,

$$\Pr\left(0 < \left|\eta(X) - \frac{1}{2}\right| \le \epsilon\right) \le c\epsilon^\beta.$$

N_α. For some c and all measurable $f : \mathcal{X} \to \{\pm 1\}$,

$$\Pr\bigl(f(X)(\eta(X) - 1/2) < 0\bigr) \le c\bigl(R(f) - R^*\bigr)^\alpha. \tag{6}$$

These conditions are equivalent.

Lemma 5. For $0 \le \beta < \infty$, a probability distribution satisfies M_β iff it satisfies $N_{\beta/(1+\beta)}$. Furthermore, M_∞ is equivalent to N_1, because

$$\Pr\left(0 < \left|\eta(X) - \frac{1}{2}\right| < \frac{1}{2c}\right) = 0 \tag{7}$$

iff, for all measurable $f : \mathcal{X} \to \{\pm 1\}$,

$$\Pr\bigl(f(X)(\eta(X) - 1/2) < 0\bigr) \le c\bigl(R(f) - R^*\bigr). \tag{8}$$

In what follows, we say that *P has noise exponent* $\alpha \ge 0$ if it satisfies N_α. Recall that

$$\begin{aligned} R(f) - R^* &= \mathbf{E}\bigl(\mathbb{1}\bigl[f(X) \ne \operatorname{sign}(\eta(X) - 1/2)\bigr]|2\eta(X) - 1|\bigr) \\ &= \mathbf{E}\bigl(\mathbb{1}\bigl[f(X)(\eta(X) - 1/2) < 0\bigr]|2\eta(X) - 1|\bigr) \\ &\le P_X\bigl(f(X)(\eta(X) - 1/2) < 0\bigr), \end{aligned} \tag{9}$$

which implies that $\alpha \le 1$. If $\alpha = 0$, then this imposes no constraint on the noise. Take $c = 1$ to see that every probability measure satisfies N_1.

The following theorem shows that if the probability distribution is such that $\eta(X)$ is unlikely to be close to $1/2$, then the bound on the excess risk in terms of the excess ϕ-risk is improved. In cases where ψ is strictly convex, such as the exponential, quadratic, and logistic loss functions, this implies that performance improves in the presence of a favorable noise exponent, without knowledge of the noise exponent.

Theorem 3. Suppose that P has noise exponent $0 < \alpha \le 1$, and that ϕ is classification-calibrated. Then there is a $c > 0$ such that for any $f : \mathcal{X} \to \mathbb{R}$,

$$c\big(R(f) - R^*\big)^\alpha \psi\left(\frac{(R(f) - R^*)^{1-\alpha}}{2c}\right) \le R_\phi(f) - R_\phi^*.$$

Furthermore, this never gives a worse rate than the result of Theorem 1, because

$$\big(R(f) - R^*\big)^\alpha \psi\left(\frac{(R(f) - R^*)^{1-\alpha}}{2c}\right) \ge \psi\left(\frac{R(f) - R^*}{2c}\right).$$

Proof. Recalling the definition of low noise in (6), fix $c > 0$ such that for every $f : \mathcal{X} \to \mathbb{R}$,

$$P_X\big(\mathrm{sign}(f(X)) \ne \mathrm{sign}(\eta(X) - 1/2)\big) \le c\big(R(f) - R^*\big)^\alpha.$$

We approximate the error integral separately over a region with high noise and over the remainder of the input space. Toward this end, fix $\epsilon > 0$ (the noise threshold), and note that

$$\begin{aligned}
&R(f) - R^* \\
&\quad = \mathbf{E}\big(\mathbb{1}\big[\mathrm{sign}(f(X)) \ne \mathrm{sign}(\eta(X) - 1/2)\big]|2\eta(X) - 1|\big) \\
&\quad = \mathbf{E}\big(\mathbb{1}\big[|2\eta(X) - 1| < \epsilon\big] \\
&\qquad\quad \times \mathbb{1}\big[\mathrm{sign}(f(X)) \ne \mathrm{sign}(\eta(X) - 1/2)\big]|2\eta(X) - 1|\big) \\
&\qquad + \mathbf{E}\big(\mathbb{1}\big[|2\eta(X) - 1| \ge \epsilon\big] \\
&\qquad\quad \times \mathbb{1}\big[\mathrm{sign}(f(X)) \ne \mathrm{sign}(\eta(X) - 1/2)\big]|2\eta(X) - 1|\big) \\
&\quad \le c\epsilon(R(f) - R^*)^\alpha \\
&\qquad + \mathbf{E}\big(\mathbb{1}\big[|2\eta(X) - 1| \ge \epsilon\big] \\
&\qquad\quad \times \mathbb{1}\big[\mathrm{sign}(f(X)) \ne \mathrm{sign}(\eta(X) - 1/2)\big]|2\eta(X) - 1|\big).
\end{aligned}$$

Now, for any x,

$$\mathbb{1}\big[|2\eta(x) - 1| \ge \epsilon\big]|2\eta(x) - 1| \le \frac{\epsilon}{\psi(\epsilon)}\psi\big(|2\eta(x) - 1|\big). \quad (10)$$

Indeed, when $|2\eta(x) - 1| < \epsilon$, (10) follows from the fact that ψ is nonnegative (Lemma 2, parts 2, 8, and 9), and when $|2\eta(x) - 1| \ge \epsilon$, (10) follows from Lemma 1, part 2.

Thus, using the same argument as in the proof of Theorem 1,

$$\begin{aligned}
&R(f) - R^* \\
&\quad \le c\epsilon\big(R(f) - R^*\big)^\alpha \\
&\qquad + \frac{\epsilon}{\psi(\epsilon)}\mathbf{E}\big(\mathbb{1}\big[\mathrm{sign}(f(X)) \ne \mathrm{sign}(\eta(X) - 1/2)\big] \\
&\qquad\quad \times \psi\big(|2\eta(X) - 1|\big)\big) \\
&\quad \le c\epsilon\big(R(f) - R^*\big)^\alpha + \frac{\epsilon}{\psi(\epsilon)}\big(R_\phi(f) - R_\phi^*\big),
\end{aligned}$$

and hence

$$\left(\frac{R(f) - R^*}{\epsilon} - c\big(R(f) - R^*\big)^\alpha\right)\psi(\epsilon) \le R_\phi(f) - R_\phi^*.$$

Choosing

$$\epsilon = \frac{1}{2c}\big(R(f) - R^*\big)^{1-\alpha}$$

and substituting gives the first inequality. [We can assume that $R(f) - R^* > 0$, because the inequality is trivial otherwise.]

The second inequality follows from the fact that $\psi(\theta)/\theta$ is nondecreasing, which we know from Lemma 1, part 3.

4. ESTIMATION RATES

In previous sections we showed that the excess risk, $R(f) - R^*$, can be bounded in terms of the excess ϕ-risk, $R_\phi(f) - R_\phi^*$. In this section we give bounds on the excess ϕ-risk. Combined with our earlier results, these lead to bounds on the excess risk. We focus on methods that choose a function from a class $\mathcal{F}$ to minimize the *empirical ϕ-risk*,

$$\hat{R}_\phi(f) = \hat{\mathbf{E}}\phi(Yf(X)) = \frac{1}{n}\sum_{i=1}^n \phi(Y_i f(X_i)).$$

Let $\hat{f}$ denote the minimizer of the empirical ϕ-risk. We are interested in the convergence of $\hat{f}$'s excess ϕ-risk, $R_\phi(\hat{f}) - R_\phi^*$. We can split this excess ϕ-risk into an estimation error term and an approximation error term,

$$R_\phi(\hat{f}) - R_\phi^* = \Big(R_\phi(\hat{f}) - \inf_{f\in\mathcal{F}} R_\phi(f)\Big) + \Big(\inf_{f\in\mathcal{F}} R_\phi(f) - R_\phi^*\Big).$$

We focus on the first term, the estimation error term. We assume throughout that some $f^* \in \mathcal{F}$ achieves the infimum,

$$R_\phi(f^*) = \inf_{f\in\mathcal{F}} R_\phi(f).$$

The simplest way to bound $R_\phi(\hat{f}) - R_\phi(f^*)$ is to use a uniform convergence argument; if

$$\sup_{f\in\mathcal{F}} |\hat{R}_\phi(f) - R_\phi(f)| \le \epsilon_n, \quad (11)$$

then

$$\begin{aligned}
&R_\phi(\hat{f}) - R_\phi(f^*) \\
&\quad = \big(R_\phi(\hat{f}) - \hat{R}_\phi(\hat{f})\big) + \big(\hat{R}_\phi(\hat{f}) - \hat{R}_\phi(f^*)\big) \\
&\qquad + \big(\hat{R}_\phi(f^*) - R_\phi(f^*)\big) \\
&\quad \le 2\epsilon_n + \big(\hat{R}_\phi(\hat{f}) - \hat{R}_\phi(f^*)\big) \\
&\quad \le 2\epsilon_n,
\end{aligned}$$

because $\hat{f}$ minimizes $\hat{R}_\phi$. But this approach can give the wrong rate. For example, for a nontrivial class $\mathcal{F}$, the expectation of the supremum of the empirical process in (11) can decrease no faster than $1/\sqrt{n}$. But if $\mathcal{F}$ is a small class (e.g., if it is a subset of a finite-dimensional linear class) and $R_\phi(f^*) = 0$, then $R_\phi(\hat{f})$ should decrease as $\log n/n$.

Lee, Bartlett, and Williamson (1996) showed that better rates than those that follow from the uniform convergence argument can be obtained for the quadratic loss $\phi(\alpha) = (1 - \alpha)^2$ if $\mathcal{F}$ is convex, even if $R_\phi(f^*) > 0$. In particular, because the quadratic loss function is strictly convex, it is possible to bound the variance of the excess loss (i.e., the difference between the loss of a function f and that of the optimal f^*) in terms of its expectation. Because the variance decreases as we approach the optimal f^*, the risk of the empirical minimizer converges more quickly to the optimal risk than the simple uniform convergence results would suggest. Mendelson (2002) improved this result and extended it from prediction in $L_2(P_X)$ to prediction in $L_p(P_X)$ for other values of p. The proof used the idea of the modulus of convexity of a norm. In this section we use this idea to give a simpler proof of a more general bound when the loss function satisfies a strict convexity condition, and we obtain risk bounds.

The modulus of convexity of an arbitrary strictly convex function (rather than a norm) is a key notion in formulating our results. Recall that a *pseudometric* d on a set S satisfies all of the axioms of a metric, except that there can be $a \neq b$ with $d(a,b) = 0$.

Definition 3 (Modulus of convexity). Given a pseudometric d defined on a convex subset S of a vector space, and a convex function $f : S \to \mathbb{R}$, the *modulus of convexity* of f with respect to d is the function $\delta : [0,\infty) \to [0,\infty]$ satisfying

$$\delta(\epsilon) = \inf\left\{\frac{f(x_1)+f(x_2)}{2} - f\left(\frac{x_1+x_2}{2}\right) : x_1, x_2 \in S, d(x_1,x_2) \geq \epsilon\right\}.$$

If $\delta(\epsilon) > 0$ for all $\epsilon > 0$, then we say that f is *strictly convex* with respect to d.

For example, for $S = \mathbb{R}$, d denoting the Euclidean distance, and $f(\alpha) = \alpha^2$, the modulus of convexity is $\delta(\epsilon) = \epsilon^2/4$. For $S = [-a, a]$ and the same metric, $f(\alpha) = e^\alpha$ has modulus of convexity $e^{-a}((1+e^\epsilon)/2 - e^{\epsilon/2}) = e^{-a}\epsilon^2/8 + o(\epsilon^2)$.

We consider loss functions ϕ that also satisfy a Lipschitz condition with respect to a pseudometric d on $\mathbb{R}$; we say that $\phi : \mathbb{R} \to \mathbb{R}$ is Lipschitz with respect to d, with constant L, if

$$\text{for all } a, b \in \mathbb{R}, \qquad |\phi(a) - \phi(b)| \leq L \cdot d(a,b).$$

Note that if d is a *metric* and ϕ is convex, then ϕ necessarily satisfies a Lipschitz condition on any compact subset of $\mathbb{R}$ (Rockafellar 1997).

Assumption A. The loss function $\phi : \mathbb{R} \to \mathbb{R}$ and the class $\mathcal{F}$ of real functions on $\mathcal{X}$ satisfy the following conditions. For some pseudometric d on $\mathbb{R}$, there are constants L, c, r, and B, such that the following conditions obtain:

A.1. ϕ is classification-calibrated.
A.2. ϕ is Lipschitz with constant L, with respect to d.
A.3. ϕ is convex with modulus of convexity $\delta(\epsilon) \geq c\epsilon^r$ with respect to d.
A.4. $\mathcal{F}$ is convex.
A.5. For all $f \in \mathcal{F}$, $x_1, x_2 \in \mathcal{X}$, and $y_1, y_2 \in \mathcal{Y}$, $d(y_1 f(x_1), y_2 f(x_2)) \leq B$.

Define the *excess loss class* $g_\mathcal{F}$ as

$$g_\mathcal{F} = \{g_f : f \in \mathcal{F}\} = \big\{(x,y) \mapsto \phi(yf(x)) - \phi(yf^*(x)) : f \in \mathcal{F}\big\},$$

where $f^* = \arg\min_{f \in \mathcal{F}} \mathbf{E}\phi(Yf(X))$. Notice that functions in $g_\mathcal{F}$ can take negative values, but they all have nonnegative expectation. We are interested in bounds on the excess ϕ-risk, $R_\phi(\hat{f}) - R_\phi^*$, where $\hat{f}$ is the minimizer of the empirical ϕ-risk. This is equivalent to the expectation of $g_{\hat{f}}$, where $g_{\hat{f}}$ is the element of the loss class with minimal sample average.

In the following theorem, we exploit the concentration of measure phenomenon to give a bound on the excess ϕ-risk. A standard uniform convergence argument, described at the beginning of this section, could proceed by considering the supremum of the empirical process indexed by the loss class,

$$\mathbf{E}\sup\{\mathbf{E}g - \hat{\mathbf{E}}g : g \in g_\mathcal{F}\}.$$

This corresponds to considering the maximal deviation between expectations and sample averages over the loss class. Instead, we use an approach introduced by Bartlett and Mendelson (2005) (see also Massart 2000b; Koltchinskii and Panchenko 2000; Mendelson 2002; Lugosi and Wegkamp 2004; Bartlett, Bousquet, and Mendelson 2005). We divide the excess loss class into subsets of different expectation, $\{g \in g_\mathcal{F} : \mathbf{E}g = \epsilon\}$, and consider the suprema of the empirical processes indexed by such subsets,

$$\xi_{g_\mathcal{F}}(\epsilon) = \mathbf{E}\sup\{\mathbf{E}g - \hat{\mathbf{E}}g : g \in g_\mathcal{F}, \mathbf{E}g = \epsilon\}.$$

(Note that the function $\xi_{g_\mathcal{F}}$ depends on the sample size n, but we simplify the notation by omitting this dependence.) For strictly convex Lipschitz ϕ and convex $\mathcal{F}$, the variance of each excess loss function is bounded in terms of its expectation, which allows us to replace the maximal deviation over the whole class by the maximal deviation over a small subset of the class: those functions with expectation ϵ_n^*, where ϵ_n^* is the fixed point of the map $\epsilon \mapsto \xi_{g_\mathcal{F}}(\epsilon)$.

Theorem 4. Suppose that the loss function ϕ and the function class $\mathcal{F}$ satisfy Assumption A. Then there is a constant K such that, with probability at least $1 - \delta$, the minimizer $\hat{f} \in \mathcal{F}$ of the empirical ϕ-risk satisfies

$$R_\phi(\hat{f}) \leq \inf_{f \in \mathcal{F}} R_\phi(f) + \epsilon_n,$$

where

$$\epsilon_n = K\max\left\{\epsilon_n^*, \left(\frac{c_r L^2 \ln(1/\delta)}{n}\right)^{1/(2-\beta)}, \frac{BL\ln(1/\delta)}{n}\right\},$$

$$\epsilon_n^* \geq \xi_{g_\mathcal{F}}(\epsilon_n^*),$$

$$c_r = \begin{cases} (2c)^{-2/r} & \text{if } r \geq 2 \\ (2c)^{-1}B^{2-r} & \text{otherwise,} \end{cases}$$

and

$$\beta = \min\left(1, \frac{2}{r}\right).$$

Thus there is a constant c' such that for any probability distribution P on $\mathcal{X} \times \mathcal{Y}$ with noise exponent α, with probability at least $1 - \delta$,

$$c'\big(R(\hat{f}) - R^*\big)^\alpha \psi\left(\frac{(R(\hat{f}) - R^*)^{1-\alpha}}{2c'}\right) \leq \epsilon_n + \inf_{f \in \mathcal{F}} R_\phi(f) - R_\phi^*.$$

It is instructive to consider the various components of the classification risk in this bound. The estimation error, ϵ_n, increases as the complexity of the class $\mathcal{F}$ increases and decreases as the sample size increases. The approximation error, $\inf_{f\in\mathcal{F}} R_\phi(f) - R_\phi^*$, is expressed in terms of the ϕ-risk. It decreases as the class $\mathcal{F}$ increases. Finally, using the convex surrogate ϕ in place of the 0–1 loss affects the bound through the rate of growth of the function of $R(\hat{f}) - R^*$ that appears on the left side. The rate of decrease of classification risk improves as the noise exponent increases.

Consider the impact on the bound of the modulus of convexity of the loss function. For flatter loss functions, where the exponent of the modulus of convexity $r > 2$, the rate can be no better than $n^{-1/(2-2/r)} = n^{-r/(2(r-1))}$, which approaches $n^{-1/2}$ as r gets large. For more curved ϕ, with $r \le 2$, the rate can be as good as n^{-1}. In contrast, we have seen that a more curved ϕ leads to a worse ψ. But, if the noise exponent is $\alpha = 1$, then the bound is optimized by a more curved ϕ, with $r \ge 2$.

Shen, Tseng, Zhang, and Wong (2003) showed that fast rates are also possible under the low-noise assumption for a particular nonconvex ϕ. In that case, however, minimization of empirical ϕ-risk requires the use of heuristics, because the optimization problem cannot be solved efficiently.

In the remainder of this section, we present a proof of Theorem 4. This proof has two key ingredients, which we capture in a pair of lemmas. The first lemma shows that if the variance of an excess loss function is bounded in terms of its expectation, then we can obtain faster rates than would be implied by the uniform convergence bounds. The second lemma presents simple conditions on the loss function that ensure that this variance bound is satisfied for convex function classes.

Lemma 6. Consider a class $\mathcal{F}$ of functions $f:\mathcal{X}\to\mathbb{R}$ with $\sup_{f\in\mathcal{F}}\|f\|_\infty \le B$. Let P be a probability distribution on $\mathcal{X}$, and suppose that there are $c \ge 1$ and $0 < \beta \le 1$ such that, for all $f \in \mathcal{F}$,

$$\mathbf{E}f^2(X) \le c(\mathbf{E}f)^\beta. \tag{12}$$

Fix $0 < \alpha, \epsilon < 1$. Suppose that if some $f \in \mathcal{F}$ has $\hat{\mathbf{E}}f \le \alpha\epsilon$ and $\mathbf{E}f \ge \epsilon$, then some $f' \in \mathcal{F}$ has $\hat{\mathbf{E}}f' \le \alpha\epsilon$ and $\mathbf{E}f = \epsilon$. Then with probability at least $1 - e^{-x}$, any $f \in \mathcal{F}$ satisfies

$$\hat{\mathbf{E}}f \le \alpha\epsilon \qquad \Longrightarrow \qquad \mathbf{E}f \le \epsilon,$$

provided that

$$\epsilon \ge \max\left\{\epsilon^*, \left(\frac{9cKx}{(1-\alpha)^2 n}\right)^{1/(2-\beta)}, \frac{4KBx}{(1-\alpha)n}\right\},$$

where K is an absolute constant and

$$\epsilon^* \ge \frac{6}{1-\alpha}\xi_{\mathcal{F}}(\epsilon^*).$$

As an aside, notice that assuming that the distribution has noise exponent α can lead to a condition of the form (12). To see this, let f^* be the Bayes decision rule and consider the class of functions $\{\alpha g_f : f \in \mathcal{F}, \alpha \in [0,1]\}$, where

$$g_f(x,y) = \ell(f(x),y) - \ell(f^*(x),y)$$

and ℓ is the 0–1 loss. Then the condition

$$P_X\big(f(X) \ne f^*(X)\big) \le c\big(\mathbf{E}\ell(f(X),Y) - \mathbf{E}\ell(f^*(X),Y)\big)^\alpha$$

can be rewritten as

$$\mathbf{E}g_f^2(X,Y) \le c(\mathbf{E}g_f(X,Y))^\alpha.$$

Thus we can obtain a version of Tsybakov's result for small function classes from Lemma 6: If the Bayes decision rule f^* is in $\mathcal{F}$, then the function $\hat{f}$ that minimizes empirical risk has

$$\hat{\mathbf{E}}g_{\hat{f}} = \hat{R}(f) - \hat{R}(f^*) \le 0,$$

and so with high probability has $\mathbf{E}g_{\hat{f}} = R(f) - R^* \le \epsilon$ under the conditions of the theorem. If $\mathcal{F}$ is a VC class, then we have $\epsilon \le c\log n/n$ for some constant c, which is surprisingly fast when $R^* > 0$.

The second ingredient in the proof of Theorem 4 is the following lemma, which gives conditions that ensure a variance bound of the kind required for the previous lemma [condition (12)]. For a pseudometric d on $\mathbb{R}$ and a probability distribution on $\mathcal{X}$, we can define a pseudometric $\tilde{d}$ on the set of uniformly bounded real functions on $\mathcal{X}$,

$$\tilde{d}(f,g) = \big(\mathbf{E}d\big(f(X),g(X)\big)^2\big)^{1/2}.$$

If d is the usual metric on $\mathbb{R}$, then $\tilde{d}$ is the $L_2(P)$ pseudometric.

Lemma 7. Consider a convex class $\mathcal{F}$ of real-valued functions defined on $\mathcal{X}$, a convex loss function $\ell:\mathbb{R}\to\mathbb{R}$, and a pseudometric d on $\mathbb{R}$. Suppose that ℓ satisfies the following conditions:

1. ℓ is Lipschitz with respect to d, with constant L,

$$\text{for all } a,b\in\mathbb{R}, \qquad |\ell(a) - \ell(b)| \le Ld(a,b).$$

2. $R(f) = \mathbf{E}\ell(f)$ is a strictly convex functional with respect to the pseudometric $\tilde{d}$, with modulus of convexity $\tilde{\delta}$,

$$\tilde{\delta}(\epsilon) = \inf\left\{\frac{R(f)+R(g)}{2} - R\left(\frac{f+g}{2}\right) : \tilde{d}(f,g) \ge \epsilon\right\}.$$

Suppose that f^* satisfies $R(f^*) = \inf_{f\in\mathcal{F}} R(f)$, and define

$$g_f(x) = \ell(f(x)) - \ell(f^*(x)).$$

Then

$$\mathbf{E}g_f \ge 2\tilde{\delta}(\tilde{d}(f,f^*)) \ge 2\tilde{\delta}\left(\frac{\sqrt{\mathbf{E}g_f^2}}{L}\right).$$

We apply the lemma to a class of functions of the form $(x,y)\mapsto yf(x)$, with the loss function $\ell = \phi$. (The lemma can be trivially extended to a loss function $\ell:\mathbb{R}\times\mathcal{Y}\to\mathbb{R}$ that satisfies a Lipschitz constraint uniformly over $\mathcal{Y}$.)

In our application, the following result implies that we can estimate the modulus of convexity of R_ϕ with respect to the pseudometric $\tilde{d}$ if we have some information about the modulus of convexity of ϕ with respect to the pseudometric d.

Lemma 8. Suppose that a convex function $\ell:\mathbb{R}\to\mathbb{R}$ has modulus of convexity δ with respect to a pseudometric d on $\mathbb{R}$, and that for some fixed $c, r > 0$, every $\epsilon > 0$ satisfies

$$\delta(\epsilon) \ge c\epsilon^r.$$

Then for functions $f:\mathcal{X}\to\mathbb{R}$ satisfying $\sup_{x_1,x_2} d(f(x_1), f(x_2)) = B$, the modulus of convexity $\tilde{\delta}$ of $R(f) = \mathbf{E}\ell(f)$ with respect to the pseudometric $\tilde{d}$ satisfies

$$\tilde{\delta}(\epsilon) \ge c_r\epsilon^{\max\{2,r\}},$$

where $c_r = c$ if $r \ge 2$ and $c_r = cB^{r-2}$ otherwise.

It is also possible to prove a converse result, that the modulus of convexity of ϕ is at least the infimum over probability distributions of the modulus of convexity of R. [To see this, we choose a probability distribution concentrated on the $x \in \mathcal{X}$ where $f_1(x)$ and $f_2(x)$ achieve the infimum in the definition of the modulus of convexity.]

Proof of Theorem 4. Consider the class $\{g_f : f \in \mathcal{F}\}$ where, for each $f \in \mathcal{F}$,

$$g_f(x, y) = \phi(yf(x)) - \phi(yf^*(x)),$$

and where $f^* \in \mathcal{F}$ minimizes $R_\phi(f) = \mathbf{E}\phi(Yf(X))$. Applying Lemma 8, we see that the functional $R(f) = \mathbf{E}\phi(f)$, defined for functions $(x, y) \mapsto yf(x)$, has modulus of convexity

$$\tilde{\delta}(\epsilon) \geq c_r \epsilon^{\max\{2,r\}},$$

where $c_r = c$ if $r \geq 2$ and $c_r = cB^{r-2}$ otherwise. From Lemma 7,

$$\mathbf{E}g_f \geq 2c_r \left(\frac{\sqrt{\mathbf{E}g_f^2}}{L} \right)^{\max\{2,r\}},$$

which is equivalent to

$$\mathbf{E}g_f^2 \leq c_r' L^2 (\mathbf{E}g_f)^{\min\{1,2/r\}}$$

with

$$c_r' = \begin{cases} (2c)^{-2/r} & \text{if } r \geq 2 \\ (2c)^{-1}B^{2-r} & \text{otherwise.} \end{cases}$$

To apply Lemma 6 to the class $\{g_f : f \in \mathcal{F}\}$, we need to check the condition. Suppose that g_f has $\hat{\mathbf{E}}g_f \leq \alpha\epsilon$ and $\mathbf{E}g_f \geq \epsilon$. Then, by the convexity of $\mathcal{F}$ and the continuity of ϕ, some $f' = \gamma f + (1 - \gamma)f^* \in \mathcal{F}$, for $0 \leq \gamma \leq 1$ has $\mathbf{E}g_f = \epsilon$. Jensen's inequality shows that

$$\begin{aligned} \hat{\mathbf{E}}g_f &= \hat{\mathbf{E}}\phi\big(Y\big(\gamma f(X) + (1 - \gamma)f^*(X)\big)\big) - \hat{\mathbf{E}}\phi(Yf^*(X)) \\ &\leq \gamma\big(\hat{\mathbf{E}}\phi(Yf(x)) - \hat{\mathbf{E}}\phi(Yf^*(X))\big) \leq \alpha\epsilon. \end{aligned}$$

Applying Lemma 6 we have, with probability at least $1 - e^{-x}$, that any g_f with $\hat{\mathbf{E}}g_f \leq \epsilon/2$ also has $\mathbf{E}g_f \leq \epsilon$, provided that

$$\epsilon \geq \max\left\{\epsilon^*, \left(\frac{36c_r' L^2 Kx}{n}\right)^{1/(2-\min\{1,2/r\})}, \frac{16KBLx}{n}\right\},$$

where $\epsilon^* \geq 12\xi_{g_\mathcal{F}}(\epsilon^*)$. In particular, if $\hat{f} \in \mathcal{F}$ minimizes empirical risk, then

$$\hat{\mathbf{E}}g_{\hat{f}} = \hat{R}_\phi(\hat{f}) - \hat{R}_\phi(f^*) \leq 0 < \frac{\epsilon}{2},$$

and hence $\mathbf{E}g_{\hat{f}} \leq \epsilon$.

Combining with Theorem 3 shows that for some c',

$$\begin{aligned} c'\big(R(\hat{f}) - R^*\big)^\alpha \psi\left(\frac{(R(\hat{f}) - R^*)^{1-\alpha}}{2c'}\right) & \\ \leq R_\phi(\hat{f}) - R_\phi^* & \\ = R_\phi(\hat{f}) - R_\phi(f^*) + R_\phi(f^*) - R_\phi^* & \\ \leq \epsilon + R_\phi(f^*) - R_\phi^*. & \end{aligned}$$

4.1 Examples

We consider four loss functions that satisfy the requirements for the fast convergence rates: the exponential loss function used in AdaBoost, the deviance function corresponding to logistic regression, the quadratic loss function, and the truncated quadratic loss function; see Table 1. These functions are illustrated in Figures 1 and 3. We use the pseudometric

$$\begin{aligned} d_\phi(a, b) = \inf\big\{&|a - \alpha| + |\beta - b| : \\ &\phi \text{ constant on } (\min\{\alpha, \beta\}, \max\{\alpha, \beta\})\big\}. \end{aligned}$$

For all functions except the truncated quadratic loss function, this corresponds to the standard metric on $\mathbb{R}$, $d_\phi(a, b) = |a - b|$. In all cases, $d_\phi(a, b) \leq |a - b|$, but for the truncated quadratic, d_ϕ ignores differences to the right of 1. It is easy to calculate the Lipschitz constant and modulus of convexity for each of these loss functions. These parameters are given in Table 1.

In the following result, we consider the function class used by algorithms such as AdaBoost: the class of linear combinations of classifiers from a fixed base class. We assume that this base class has finite VC dimension, and constrain the size of the class by restricting the ℓ_1 norm of the linear parameters. If $\mathcal{G}$ is the VC class, then we write $\mathcal{F} = B \operatorname{absconv}(\mathcal{G})$ for some constant B, where

$$B \operatorname{absconv}(\mathcal{G}) = \left\{ \sum_{i=1}^m \alpha_i g_i : m \in \mathbb{N}, \alpha_i \in \mathbb{R}, g_i \in \mathcal{G}, \|\alpha\|_1 = B \right\}.$$

Theorem 5. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex loss function. Suppose that on the interval $[-B, B]$, ϕ is Lipschitz with constant L_B and has modulus of convexity $\delta(\epsilon) = a_B\epsilon^2$ (both with respect to the pseudometric d).

For any probability distribution P on $\mathcal{X} \times \mathcal{Y}$ that has noise exponent α, there is a constant c' for which the following is true. For iid data $(X_1, Y_1), \ldots, (X_n, Y_n)$, let $\hat{f} \in \mathcal{F}$ be the minimizer of the empirical ϕ-risk, $R_\phi(f) = \hat{\mathbf{E}}\phi(Yf(X))$. Suppose that $\mathcal{F} = B \operatorname{absconv}(\mathcal{G})$, where $\mathcal{G} \subseteq \{\pm 1\}^\mathcal{X}$ has VC dimension V and

$$\epsilon_n^* \geq BL_B \max\left\{ \left(\frac{L_B a_B}{B}\right)^{1/(V+1)}, 1 \right\} n^{-(V+2)/(2(V+1))}.$$

Then, with probability at least $1 - \delta$,

$$R(\hat{f}) \leq R^* + c'\left(\epsilon_n^* + \frac{L_B(L_B/(2a_B) + B)\ln(1/\delta)}{n} + \inf_{f \in \mathcal{F}} R_\phi(f) - R_\phi^*\right)^{1/(2-\alpha)}.$$

Table 1. Four Convex Loss Functions Defined on $\mathbb{R}$

	$\phi(\alpha)$	L_B	$\delta(\epsilon)$
Exponential	$e^{-\alpha}$	e^B	$e^{-B}\epsilon^2/8$
Logistic	$\ln(1 + e^{-2\alpha})$	2	$e^{-2B}\epsilon^2/4$
Quadratic	$(1 - \alpha)^2$	$2(B + 1)$	$\epsilon^2/4$
Truncated quadratic	$(\max\{0, 1 - \alpha\})^2$	$2(B + 1)$	$\epsilon^2/4$

NOTE: On the interval $[-B, B]$, each has the indicated Lipschitz constant L_B and modulus of convexity $\delta(\epsilon)$ with respect to d_ϕ. All have a quadratic modulus of convexity.

4.2 Simulations

This section describes a set of simulations that illustrate the performance of the excess risk bound based on the ψ-transform, as well as the theoretical excess ϕ-risk rates obtained from Theorem 4. We took $\mathcal{X} = [-1, 1]^{10}$ as our covariate space, with $P_{\mathcal{X}}$ equal to the uniform distribution on $\mathcal{X}$. For the conditional distribution $\eta(x)$, we used members of a parameterized family based on the logistic function,

$$\eta_q(x) = P(Y = 1|x) = \sigma\big(C\,\mathrm{sign}(x_1)|x_1|^q\big), \qquad q > 0,$$

where $\sigma(u) = 1/(1 + \exp(-u))$. Varying q results in different noise exponents for the conditional distribution; it is straightforward to see that $P_{\mathcal{X}}(0 < |2\eta_q(X) - 1| < \epsilon) < (4/C)^{1/q}\epsilon^{1/q}$, so Lemma 5 implies that η_q has noise exponent $1/(q + 1)$. We chose the constant C so that $\eta_q(-1) = 1/4$ and $\eta_q(1) = 3/4$, for all q.

The margin-based loss functions in our simulations also came from a one-dimensional family, indexed by $p > 1$,

$$\phi_p(\alpha) = (p - 1)(2/p)^{p/(p-1)} - 2\alpha + |\alpha|^p.$$

The leading constant ensures that ϕ_p is nonnegative for all $\alpha \in [-1, 1]$. For $p > 1$, ϕ_p is convex with a negative first derivative at 0, so Theorem 2, part 1 tells us that it is classification-calibrated. Different choices of p lead to different values for the modulus of convexity exponent of ϕ_p, because $(\phi(\epsilon) + \phi(-\epsilon))/2 - \phi(0) = \epsilon^p$ for positive ϵ.

We took as a family of real-valued classifiers the convex hull of the coordinate functions, $\mathcal{F} = \mathrm{co}\{x \mapsto \beta_i(x) = x_i : i = 1, \ldots, 10\}$. Thus each $f \in \mathcal{F}$ has the form $f(u) = \lambda^\top u$ for some $\lambda \geq 0$, $\lambda^\top 1 \leq 1$. We simulated datasets of several sizes n between 10 and 10,000, using various values of p and q, as detailed later. For each choice of $n, p,$ and q, we performed 25 repetitions of the following procedure. First, we generated a dataset according to $(P_{\mathcal{X}}, \eta_q(x))$ and found the empirical risk minimizer $\hat{f}_n$ over $\mathcal{F}$, through a constrained convex optimization. Then we computed the 0–1 risk of $\hat{f}_n$, approximating the relevant integral with adaptive numerical quadrature. Subtracting the Bayes risk for the chosen distribution (depending on q), also approximated using quadrature, gave us the excess 0–1 risk of $\hat{f}_n$. Finally, we carried out a similar computation to determine the excess ϕ-risk of $\hat{f}_n$.

We illustrate the behavior of the upper bound on excess 0–1 risk obtained from the ψ-transform using these simulation results. A routine calculation along the lines of the examples in Section 2.3 shows that $\psi(\theta) = (p - 1)(2\theta/p)^{p/(p-1)}$. We appeal to Theorem 1 to obtain the inequality

$$(p - 1)\left(\frac{2(R(f) - R^*)}{p}\right)^{p/(p-1)} \leq R_\phi(f) - R_\phi^*.$$

Solving this inequality for excess 0–1 risk gives an upper bound as a function of excess ϕ-risk,

$$R(f) - R^* \leq \frac{p}{2}\left(\frac{R_\phi(f) - R_\phi^*}{p - 1}\right)^{(p-1)/p}.$$

We verified that, as expected, every excess 0–1 risk in our simulations obeyed the upper bound determined by its corresponding excess ϕ-risk, across all values of p and q.

We also used the simulations to illustrate the theoretical rates of convergence for excess ϕ-risk implied by Theorem 4. For the class $\mathcal{F}$ under consideration, the centered empirical process $\xi_{\mathcal{F}}(\epsilon)$ used in the theorem can be pointwise upper-bounded using a local Rademacher average symmetrization, which in turn is bounded by the Dudley entropy integral. The derivation closely follows the proof of Theorem 5 (see also Bartlett et al. 2005). These calculations reveal that a suitable upper bound on ϵ^* in Theorem 4 is $c(d/n)\log(nL/d)$, with $d = 10$ the dimension of $\mathcal{F}$ and c a universal constant. Thus, with probability at least $1 - e^{-x}$, we have the excess ϕ-risk bound

$$R_\phi(\hat{f}) - R_\phi(f^*) \leq c \max\left\{\left(\frac{c_r L^2 x}{n}\right)^{1/(2-\min\{1,2/p\})}, \frac{BLx}{n}, \frac{d}{n}\log\left(\frac{nL}{d}\right)\right\},$$

recalling that p is the modulus of convexity exponent for $\phi_p(\alpha)$. Treating the logarithmic factor as approximately constant, we therefore expect a rate of order n^{-1} for $1 < p < 2$ and of order $n^{-1/(2-2/p)}$ for $p \geq 2$.

Figure 7 presents the simulation results for $p \in \{1.5, 2, 3.5\}$, all with $q = 1$. Results with $q \in \{1.5, 2, 2.5, 3\}$ are similar, and indeed the theoretical rates do not vary with q. The solid lines are natural cubic spline fits, on the log–log scale, to the sample size and excess ϕ-risk from each simulation. The slope of each dashed line is the theoretical rate exponent implied by the bound: -1.0 for $p = 1.5$ and 2 and $-.7$ for $p = 3.5$. As the plots reveal, the agreement with theory when $p = 1.5$ and 2 is extremely good for large enough n. Although the match when $p = 3.5$ is less exact, the simulated results appear compatible with the theoretical rate to within the noise tolerance.

5. CONCLUSIONS

We have focused on the relationship between properties of a nonnegative margin-based loss function ϕ and the statistical performance of the classifier that, based on an iid training set, minimizes empirical ϕ-risk over a class of functions. We first derived a universal upper bound on the population misclassification risk of any thresholded measurable classifier in terms of its corresponding population ϕ-risk. The bound is governed by the ψ-transform, a convexified variational transform of ϕ. It is the tightest possible upper bound uniform over all probability distributions and measurable functions in this setting.

Using this upper bound, we characterized the class of loss functions that guarantee that every ϕ-risk consistent classifier sequence is also Bayes-risk consistent under any population distribution. Here ϕ-risk consistency denotes sequential convergence of population ϕ-risks to the smallest possible ϕ-risk of any measurable classifier. The characteristic property of such a ϕ, which we term classification-calibration, is a kind of pointwise Fisher consistency for the conditional ϕ-risk at each $x \in \mathcal{X}$. The necessity of classification-calibration is apparent; the sufficiency underscores its fundamental importance in elaborating the statistical behavior of large-margin classifiers.

For the special case of convex ϕ, which is widespread in practical applications, we demonstrated that classification-calibration is equivalent to the existence and strict negativity of the first derivative of ϕ at 0, a condition that is readily verifiable in most practical examples. In addition, the convexification step

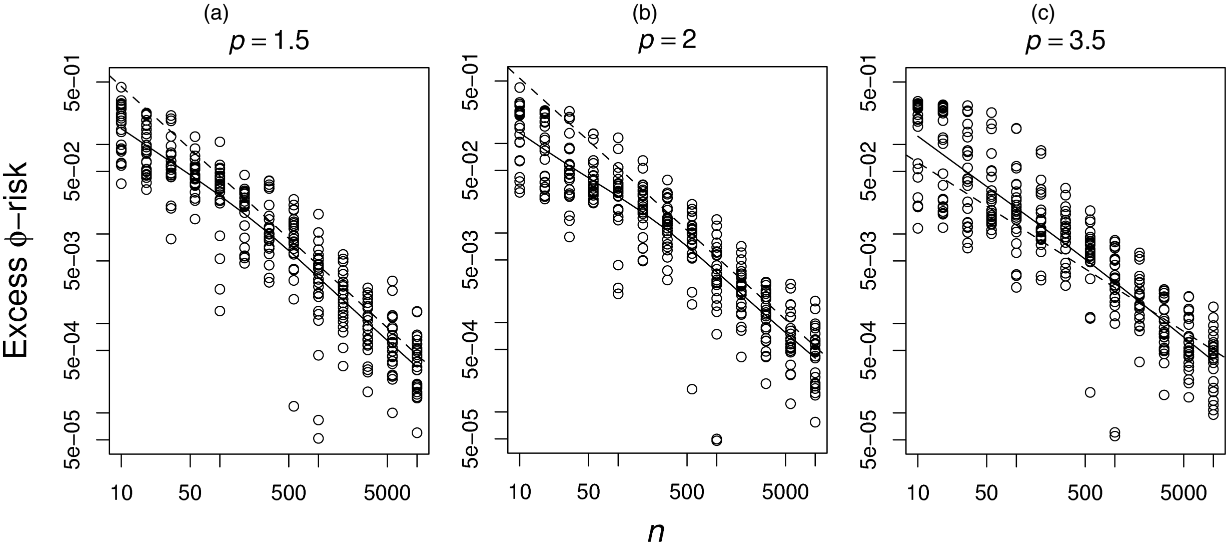

Figure 7. Rate Plots for (a) p = 1.5, (b) p = 2, and (c) p = 3.5. Each panel shows simulated excess ϕ-risk on the log scale versus simulated sample size on the log scale, for the choice of p given at top. We took q = 1 in each case. Natural cubic spline fits appear as solid lines. The dashed line depicts the slope corresponding to the theoretical rate for the chosen p. (The vertical position of the dashed line is not informative.)

in the ψ-transform is vacuous for convex ϕ, which simplifies the derivation of closed forms.

Under the low-noise assumption of Mammen and Tsybakov (1999) and Tsybakov (2004), we sharpened our original upper bound. We found that empirical ϕ-risk minimization yields convergence of ϕ-risk to that of the best-performing function in $\mathcal{F}$ as the sample size grows. For strictly convex ϕ, the convergence rate can be faster than that implied by standard uniform convergence arguments, depending on the strictness of convexity of ϕ and the complexity of $\mathcal{F}$. Combined with the low-noise condition, we saw that this implies fast rates of convergence of the misclassification risk to its optimal value. Simulations confirm the convergence rates of ϕ-risk predicted by the theory, for a linear class and a particular probability distribution. Simulations also show that the relationship between excess ϕ-risk and excess risk closely follows that predicted by the theory.

Two important issues that we have not treated are the approximation error for population ϕ-risk relative to $\mathcal{F}$, and algorithmic considerations in the minimization of empirical ϕ-risk. In the setting of scaled convex hulls of a base class, some approximation results have been given by Breiman (2004), Mannor et al. (2002), and Lugosi and Vayatis (2004). Regarding the numerical optimization to determine $\hat{f}$, Zhang and Yu (2005) gave novel bounds on the convergence rate for generic forward stagewise additive modeling (see also Zhang 2003). These authors focused on optimization of a convex risk functional over the entire linear hull of a base class, with regularization enforced by an early stopping rule.

APPENDIX A: PROOFS

Proof of Lemma 2

The proof of part 1 is immediate from the definitions. For part 2, concavity follows because H is an infimum of concave (affine) functions of η. Now, because H is concave and symmetric about 1/2, $H(1/2) = H((1/2)\eta + (1/2)(1-\eta)) \ge (1/2)H(\eta) + (1/2)H(1-\eta) = H(\eta)$. Thus H is maximal at 1/2. To see that $H(1/2) = H^-(1/2)$, note that $\alpha(2\eta - 1) \le 0$ for all α when $\eta = 1/2$.

To prove part 3, assume that there is an $\eta \ne 1/2$ with $H(\eta) = H(1/2)$. Fix a sequence $\alpha_1, \alpha_2, \ldots$ for which $\lim_{i\to\infty} C_{1/2}(\alpha_i) = H(1/2)$. By the assumption,

$$\liminf_{i\to\infty}\bigl(\eta\phi(\alpha_i) + (1-\eta)\phi(-\alpha_i)\bigr) \ge H(\eta)$$

$$= H(1/2) = \lim_{i\to\infty} \frac{\phi(\alpha_i) + \phi(-\alpha_i)}{2}. \quad \text{(A.1)}$$

Rearranging, we have

$$(\eta - 1/2)\liminf_{i\to\infty}\bigl(\phi(\alpha_i) - \phi(-\alpha_i)\bigr) \ge 0.$$

Because $H(1-\eta) = H(\eta)$, the same argument shows that

$$(\eta - 1/2)\liminf_{i\to\infty}\bigl(\phi(-\alpha_i) - \phi(\alpha_i)\bigr) \ge 0.$$

It follows that

$$\lim_{i\to\infty}\bigl(\phi(\alpha_i) - \phi(-\alpha_i)\bigr) = 0,$$

so that all of the expressions in (A.1) are equal. Hence, $H(\eta) = \lim_{i\to\infty} C_\eta(\alpha_i) = \lim_{i\to\infty} C_\eta(-\alpha_i)$, which implies that $H(\eta) = H^-(\eta)$. Thus if $H(\eta) = H(1/2)$, then ϕ is not classification-calibrated.

For part 4, H^- is concave on $[0, 1/2]$ by the same argument as for the concavity of H. (Note that when $\eta < 1/2$, H^- is an infimum over a set of concave functions, but in this case when $\eta > 1/2$, it is an infimum over a different set of concave functions.) The inequality $H^- \ge H$ follows from the definitions.

For part 5, first note that the concavity of H implies that it is continuous on the relative interior of its domain, that is, $(0, 1)$. Thus, to show that H is continuous $[0, 1]$, it suffices (by symmetry) to show that it is left-continuous at 1. Because $[0, 1]$ is locally simplicial in the sense of Rockafellar (1997), his theorem 10.2 gives lower semicontinuity of H at 1 (equivalently, upper semicontinuity of the convex function $-H$ at 1). To see upper semicontinuity of H at 1, fix any $\epsilon > 0$ and choose α_ϵ such that $\phi(\alpha_\epsilon) \le H(1) + \epsilon/2$. Then for any η between $1 - \epsilon/(2\phi(-\alpha_\epsilon))$ and 1, we have

$$H(\eta) \le C_\eta(\alpha_\epsilon) \le H(1) + \epsilon.$$

Because this is true for any ϵ, $\limsup_{\eta\to 1} H(\eta) \le H(1)$, which is upper semicontinuity. Thus H is left-continuous at 1. The same argument shows that H^- is continuous on $(0, 1/2)$ and $(1/2, 1)$ and is left-continuous at $1/2$ and 1. Symmetry implies that H^- is continuous on the closed interval $[0, 1]$. The continuity of $\tilde{\psi}$ is now immediate.

To see part 6, observe that ψ is a closed convex function with locally simplicial domain $[-1, 1]$, so that its continuity follows by once again applying theorem 10.2 of Rockafellar (1997).

It follows immediately from parts 2 and 4 that $\tilde{\psi}$ is nonnegative and minimal at 0. Because epi ψ is the convex hull of epi $\tilde{\psi}$ (i.e., the set of all convex combinations of points in epi $\tilde{\psi}$), we see that ψ is also nonnegative and minimal at 0, which is part 7.

Part 8 follows immediately from 2.

To prove part 9, suppose first that ϕ is classification-calibrated. Then for all $\theta \in (0, 1]$, $\tilde{\psi}(\theta) > 0$. But every point in epi ψ is a convex combination of points in epi $\tilde{\psi}$, so if $(\theta, 0) \in$ epi ψ, then we can only have $\theta = 0$. Hence for $\theta \in (0, 1]$, points in epi ψ of the form (θ, c) must have $c > 0$, and closure of $\tilde{\psi}$ now implies $\psi(\theta) > 0$. For the converse, note that if ϕ is not classification-calibrated, then some $\theta > 0$ has $\tilde{\psi}(\theta) = 0$, and so $\psi(\theta) = 0$.

Proof of Theorem 2

Recall that a *subgradient* of ϕ at $\alpha \in \mathbb{R}$ is any value $m_\alpha \in \mathbb{R}$ such that $\phi(x) \ge \phi(\alpha) + m_\alpha(x - \alpha)$ for all x. To prove part 1, fix a convex function ϕ.

($\Rightarrow$) Because ϕ is convex, we can find subgradients $g_1 \ge g_2$ such that for all α,

$$\phi(\alpha) \ge g_1\alpha + \phi(0)$$

and

$$\phi(\alpha) \ge g_2\alpha + \phi(0).$$

Then we have

$$\begin{aligned}
&\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\\
&\quad \ge \eta\big(g_1\alpha + \phi(0)\big) + (1-\eta)\big(-g_2\alpha + \phi(0)\big)\\
&\quad = \big(\eta g_1 - (1-\eta)g_2\big)\alpha + \phi(0) \qquad \text{(A.2)}\\
&\quad = \left(\frac{1}{2}(g_1 - g_2) + (g_1 + g_2)\left(\eta - \frac{1}{2}\right)\right)\alpha + \phi(0). \qquad \text{(A.3)}
\end{aligned}$$

Because ϕ is classification-calibrated, for $\eta > 1/2$, we can express $H(\eta)$ as $\inf_{\alpha>0}[\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)]$. If (A.3) were greater than $\phi(0)$ for every $\alpha > 0$, then it would follow that for $\eta > 1/2$, $H(\eta) \ge \phi(0) \ge H(1/2)$, which, by Lemma 2, part 3, would be a contradiction. We now show that $g_1 > g_2$ implies this contradiction. Indeed, we can choose

$$\frac{1}{2} < \eta < \frac{1}{2} + \frac{g_1 - g_2}{2|g_1 + g_2|}$$

to show that $|(\eta - 1/2)(g_1 + g_2)| < (g_1 - g_2)/2$, so (A.3) is greater than $\phi(0)$ for all $\alpha > 0$. Thus if ϕ is classification-calibrated, then we must have $g_1 = g_2$, which implies that ϕ is differentiable at 0.

To see that we must also have $\phi'(0) < 0$, note that from (A.2), we have

$$\eta\phi(\alpha) + (1-\eta)\phi(-\alpha) \ge (2\eta - 1)\phi'(0)\alpha + \phi(0).$$

But for any $\eta > 1/2$ and $\alpha > 0$, if $\phi'(0) \ge 0$, then this expression is at least $\phi(0)$. Thus if ϕ is classification-calibrated, then we must have $\phi'(0) < 0$.

($\Leftarrow$) Suppose that ϕ is differentiable at 0 and has $\phi'(0) < 0$. Then the function $C_\eta(\alpha) = \eta\phi(\alpha) + (1-\eta)\phi(-\alpha)$ has $C'_\eta(0) = (2\eta - 1)\phi'(0)$. For $\eta > 1/2$, this is negative. It follows from the convexity of ϕ that $C_\eta(\alpha)$ is minimized by some $\alpha^* \in (0, \infty]$. To see this, note that for some $\alpha_0 > 0$, we have

$$C_\eta(\alpha_0) \le C_\eta(0) + \alpha_0 C'_\eta(0)/2.$$

But the convexity of ϕ, and hence of C_η, implies that for all α,

$$C_\eta(\alpha) \ge C_\eta(0) + \alpha C'_\eta(0).$$

In particular, if $\alpha \le \alpha_0/4$, then

$$C_\eta(\alpha) \ge C_\eta(0) + \frac{\alpha_0}{4}C'_\eta(0) > C_\eta(0) + \frac{\alpha_0}{2}C'_\eta(0) \ge C_\eta(\alpha_0).$$

Similarly, for $\eta < 1/2$, the optimal α is negative. This means that ϕ is classification-calibrated.

For the proof of part 2, note that part 1 implies that ϕ is differentiable at 0 and $\phi'(0) < 0$, and so

$$\begin{aligned}
\phi(0) &\ge H^-(\eta)\\
&= \inf_{\alpha:\alpha(\eta-1/2)\le 0}\big(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\big)\\
&\ge \inf_{\alpha:\alpha(\eta-1/2)\le 0}\big(\eta\big(\phi(0) + \phi'(0)\alpha\big) + (1-\eta)\big(\phi(0) - \phi'(0)\alpha\big)\big)\\
&= \phi(0) + \inf_{\alpha:\alpha(\eta-1/2)\le 0}\big((2\eta - 1)\phi'(0)\alpha\big)\\
&= \phi(0).
\end{aligned}$$

Thus $H^-(\eta) = \phi(0)$. The concavity of H (Lemma 2, part 2 implies that $\tilde{\psi} = H^-(\eta) - H(\eta) = \phi(0) - H(\eta)$ is convex, which implies that $\psi = \tilde{\psi}$.

Proof of Lemma 3

From the convexity of ϕ, we have

$$\begin{aligned}
\psi(\theta) &= H\left(\frac{1}{2}\right) - H\left(\frac{1+\theta}{2}\right)\\
&= \phi(0) - \inf_{\alpha>0}\left(\frac{1+\theta}{2}\phi(\alpha) + \frac{1-\theta}{2}\phi(-\alpha)\right)\\
&= \sup_{\alpha>0}\left(-\theta\phi'(0)\alpha + \frac{1+\theta}{2}\big(\phi(0) - \phi(\alpha) + \alpha\phi'(0)\big)\right.\\
&\qquad\qquad \left. + \frac{1-\theta}{2}\big(\phi(0) - \phi(-\alpha) - \alpha\phi'(0)\big)\right)\\
&= \sup_{\alpha>0}\left(-\theta\phi'(0)\alpha - \frac{1+\theta}{2}d_\phi(0,\alpha) - \frac{1-\theta}{2}d_\phi(0,-\alpha)\right)\\
&\ge \sup_{\alpha>0}\big(-\theta\phi'(0)\alpha - d_\phi(0,\alpha) - d_\phi(0,-\alpha)\big)\\
&= \sup_{\alpha>0}(\theta - \xi(\alpha))(-\phi'(0)\alpha)\\
&\ge \big(\theta - \xi(\xi^{-1}(\theta/2))\big)\big(-\phi'(0)\xi^{-1}(\theta/2)\big)\\
&= -\phi'(0)\frac{\theta}{2}\xi^{-1}\left(\frac{\theta}{2}\right),
\end{aligned}$$

where the first inequality used the fact that for all $\alpha \in [0, 1]$ and all $a, b > 0$, $\alpha a + (1-\alpha)b \le a + b$.

Proof of Lemma 4

Proceeding by contrapositive, suppose that no such γ exists. Because $\phi(\alpha) \ge \mathbb{1}[\alpha \le 0]$ on $(0, \infty)$, then we must have $\inf_{\alpha\le 0}\phi(\alpha) = 0$. But $\phi(\alpha) = C_1(\alpha)$, and hence

$$0 = \inf_{\alpha\le 0} C_1(\alpha) = H^-(1) \ge H(1) \ge 0.$$

Thus $H^-(1) = H(1)$, so ϕ is not classification-calibrated.

Proof of Lemma 5

We first show that N_α implies $M_{\alpha/(1-\alpha)}$. Consider the set $S = \{x : 0 < |\eta(x) - 1/2| \le \epsilon\}$, and let f be such that $S = \{x : f(x)(\eta(x) - 1/2) < 0\}$. Then N_α implies that

$$\begin{aligned}\epsilon \Pr(S) &\ge \int_S \left|\eta(x) - \frac{1}{2}\right| dP_X(x)\\ &= \frac{1}{2}\big(R(f) - R^*\big)\\ &\ge \frac{1}{2}\left(\frac{1}{c}\Pr(S)\right)^{1/\alpha}.\end{aligned}$$

Rearranging shows that

$$\Pr(S) \le (2\epsilon)^{\alpha/(1-\alpha)} c^{1/(1-\alpha)},$$

and hence the distribution satisfies $M_{\alpha/(1-\alpha)}$.

To see that M_β implies $N_{\beta/(1+\beta)}$, we fix $\epsilon > 0$ and $f : \mathcal{X} \to \{\pm 1\}$, define $S = \{x : f(x)(\eta(x) - 1/2) < 0\}$, and write

$$\begin{aligned}R(f) - R^* &= \mathbf{E}\big(\mathbb{1}[X \in S]|2\eta(X) - 1|\big)\\ &= 2\int_S \left|\eta(x) - \frac{1}{2}\right| dP_X(x)\\ &\ge 2\epsilon \int_S \mathbb{1}\left[\left|\eta(x) - \frac{1}{2}\right| > \epsilon\right] dP_X(x)\\ &= 2\epsilon\left(\Pr(S) - \int_S \mathbb{1}\left[0 < \left|\eta(x) - \frac{1}{2}\right| \le \epsilon\right] dP_X(x)\right)\\ &\ge 2\epsilon\big(\Pr(S) - c\epsilon^\beta\big).\end{aligned}$$

With $\epsilon = (\Pr(S)/(c(1+\beta)))^{1/\beta}$, this shows that

$$R(f) - R^* \ge \frac{2\beta}{c^{1/\beta}(1+\beta)^{(\beta+1)/\beta}}(\Pr(S))^{(\beta+1)/\beta},$$

and hence the distribution satisfies $N_{\beta/(\beta+1)}$.

Now consider the second part of the lemma. For any measurable $f : \mathcal{X} \to \{\pm 1\}$, (8) is equivalent to

$$\begin{aligned}&\Pr(A_f) \le c \int_{A_f} |2\eta(x) - 1|\, dP_X(x)\\ &\iff \qquad \int_{A_f} \frac{1}{c}\, dP_X(x) \le \int_{A_f} |2\eta(x) - 1|\, dP_X(x), \qquad (A.4)\end{aligned}$$

where $A_f = \{x : f(x)(\eta(x) - 1/2) < 0\}$. Note that A_f ranges over all measurable subsets of $\{x : |\eta(x) - 1/2| > 0\}$, so that (A.4) is true for all such A_f iff

$$\Pr\left(0 < |2\eta(X) - 1| < \frac{1}{c}\right) = 0,$$

which is (7).

Proof of Lemma 6

The proof of Lemma 6 uses techniques due to Bartlett and Mendelson (2005), which built on the work of Massart (2000b), Koltchinskii and Panchenko (2000), Mendelson (2002), Lugosi and Wegkamp (2004), and Bartlett et al. (2005). We use the following concentration inequality, which is a refinement, due to Rio (2001) and Klein (2002), of a result of Massart (2000a), following Talagrand (1994) and Ledoux (2001). The best estimates on the constants are due to Bousquet (2002).

Lemma A.1. There is an absolute constant K for which the following holds. Let $\mathcal{G}$ be a class of functions defined on $\mathcal{X}$ with $\sup_{g \in \mathcal{G}} \|g\|_\infty \le b$. Suppose that P is a probability distribution such that for every $g \in \mathcal{G}$, $\mathbf{E}g = 0$. Let $X_1, \ldots, X_n$ be independent random variables distributed according to P and set $\sigma^2 = \sup_{g \in \mathcal{G}} \operatorname{var} g$. Define

$$Z = \sup_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^n g(X_i).$$

Then, for every $x > 0$ and every $\rho > 0$,

$$\Pr\left\{Z \ge (1+\rho)\mathbf{E}Z + \sigma\sqrt{\frac{Kx}{n}} + \frac{K(1+\rho^{-1})bx}{n}\right\} \le e^{-x}.$$

To prove Lemma 6, from the condition on $\mathcal{F}$, we have

$$\begin{aligned}&\Pr\{\exists f \in \mathcal{F} : \hat{\mathbf{E}}f \le \alpha\epsilon, \mathbf{E}f \ge \epsilon\}\\ &\quad \le \Pr\{\exists f \in \mathcal{F} : \hat{\mathbf{E}}f \le \alpha\epsilon, \mathbf{E}f = \epsilon\}\\ &\quad = \Pr\big\{\sup\{\mathbf{E}f - \hat{\mathbf{E}}f : f \in \mathcal{F}, \mathbf{E}f = \epsilon\} \ge (1-\alpha)\epsilon\big\}.\end{aligned}$$

We bound this probability using Lemma A.1, with $\rho = 1$ and $\mathcal{G} = \{\mathbf{E}f - f : f \in \mathcal{F}, \mathbf{E}f = \epsilon\}$. This shows that

$$\Pr\{\exists f \in \mathcal{F} : \hat{\mathbf{E}}f \le \alpha\epsilon, \mathbf{E}f \ge \epsilon\} \le \Pr\{Z \ge (1-\alpha)\epsilon\} \le e^{-x},$$

provided that

$$2\mathbf{E}Z \le \frac{(1-\alpha)\epsilon}{3},$$

$$\sqrt{\frac{c\epsilon^\beta Kx}{n}} \le \frac{(1-\alpha)\epsilon}{3},$$

and

$$\frac{4KBx}{n} \le \frac{(1-\alpha)\epsilon}{3}.$$

(We have used the fact that $\sup_{f \in \mathcal{F}} \|f\|_\infty \le B$ implies that $\sup_{g \in \mathcal{G}} \|g\|_\infty \le 2B$.) Observing that

$$\mathbf{E}Z = \xi_{\mathcal{F}}(\epsilon),$$

and rearranging gives the result.

Proof of Lemma 7

The proof proceeds in two steps. The Lipschitz condition allows us to relate $\mathbf{E}g_f^2$ to $\tilde{d}(f, f^*)$, and the modulus of convexity condition, together with the convexity of $\mathcal{F}$, relates this to $\mathbf{E}g_f$.

We have

$$\begin{aligned}\mathbf{E}g_f^2 &= \mathbf{E}\big(\ell(f(X)) - \ell(f^*(X))\big)^2\\ &\le \mathbf{E}\big(Ld\big(f(X), f^*(X)\big)\big)^2\\ &= L^2(\tilde{d}(f, f^*))^2. \qquad (A.5)\end{aligned}$$

From the definition of the modulus of convexity,

$$\begin{aligned}\frac{R(f) + R(f^*)}{2} &\ge R\left(\frac{f + f^*}{2}\right) + \tilde{\delta}(\tilde{d}(f, f^*))\\ &\ge R(f^*) + \tilde{\delta}(\tilde{d}(f, f^*)),\end{aligned}$$

where the optimality of f^* in the convex set $\mathcal{F}$ implies the second inequality. Rearranging gives

$$\mathbf{E}g_f = R(f) - R(f^*) \ge 2\tilde{\delta}(\tilde{d}(f, f^*)).$$

Combining with (A.5) gives the result.

Proof of Lemma 8

Fix functions $f_1, f_2 : \mathcal{X} \to \mathbb{R}$ with $\tilde{d}(f_1, f_2) = \sqrt{\mathbf{E}d^2(f_1(X), f_2(X))} \ge \epsilon$. We have

$$\begin{aligned}\frac{R(f_1)+R(f_2)}{2} &- R\left(\frac{f_1+f_2}{2}\right)\\ &= \mathbf{E}\left(\frac{\ell(f_1(X))+\ell(f_2(X))}{2} - \ell\left(\frac{f_1(X)+f_2(X)}{2}\right)\right)\\ &\ge \mathbf{E}\big(\delta\big(d\big(f_1(X), f_2(X)\big)\big)\big)\\ &\ge c\mathbf{E}d^r\big(f_1(X), f_2(X)\big)\\ &= c\mathbf{E}\big(d^2\big(f_1(X), f_2(X)\big)\big)^{r/2}.\end{aligned}$$

When the function $\xi(a) = a^{r/2}$ is convex (i.e., when $r \ge 2$), Jensen's inequality shows that

$$\frac{R(f_1)+R(f_2)}{2} - R\left(\frac{f_1+f_2}{2}\right) \ge c\epsilon^r.$$

Otherwise, we use the following convex lower bound on $\xi : [0, B^2] \to [0, B^r]$:

$$\xi(a) = a^{r/2} \ge B^r \frac{a}{B^2},$$

which follows from (the concave analog of) Lemma 1, part 2. This implies that

$$\frac{R(f_1)+R(f_2)}{2} - R\left(\frac{f_1+f_2}{2}\right) \ge cB^{r-2}\epsilon^2.$$

Proof of Theorem 5

It is clear that $\mathcal{F}$ is convex and satisfies the conditions of Theorem 4. That theorem implies that, with probability at least $1 - \delta$,

$$\big(R(\hat{f}) - R^*\big)^{2-\alpha} \le c'\Big(\epsilon_n + \inf_{f\in\mathcal{F}} R_\phi(f) - R_\phi^*\Big),$$

provided that

$$\epsilon_n \ge K \max\left\{\epsilon_n^*, \frac{L_B^2 \ln(1/\delta)}{2a_B n}, \frac{BL_B \ln(1/\delta)}{n}\right\},$$

where $\epsilon_n^* \ge \xi_{g_\mathcal{F}}(\epsilon_n^*)$. It remains to prove suitable upper bounds for ϵ_n^*.

By a classical symmetrization inequality (see, e.g., Van der Vaart and Wellner 1996), we can upper bound $\xi_{g_\mathcal{F}}$ in terms of local Rademacher averages,

$$\begin{aligned}\xi_{g_\mathcal{F}}(\epsilon) &= \mathbf{E}\sup\{\mathbf{E}g_f - \hat{\mathbf{E}}g_f : f \in \mathcal{F}, \mathbf{E}g_f = \epsilon\}\\ &\le 2\mathbf{E}\sup\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i g_f(X_i, Y_i) : f \in \mathcal{F}, \mathbf{E}g_f = \epsilon\right\},\end{aligned}$$

where the expectations are over the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ and the independent uniform (Rademacher) random variables $\sigma_i \in \{\pm 1\}$. The Ledoux and Talagrand (1991) contraction inequality and Lemma 7 imply that

$$\begin{aligned}\xi_{g_\mathcal{F}}(\epsilon) &\le 4L\mathbf{E}\sup\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i d_\phi\big(Y_i f(X_i), Y_i f^*(X_i)\big) : f \in \mathcal{F}, \mathbf{E}g_f = \epsilon\right\}\\ &\le 4L\mathbf{E}\sup\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i d_\phi\big(Y_i f(X_i), Y_i f^*(X_i)\big) : \right.\\ &\qquad\qquad \left. f \in \mathcal{F}, \tilde{d}_\phi(f, f^*)^2 \le 2a_B\epsilon\right\}\\ &= 4L\mathbf{E}\sup\left\{\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i, Y_i) : f \in \mathcal{F}_\phi, \mathbf{E}f^2 \le 2a_B\epsilon\right\},\end{aligned}$$

where

$$\mathcal{F}_\phi = \big\{(x, y) \mapsto d_\phi\big(yf(x), yf^*(x)\big) : f \in \mathcal{F}\big\}.$$

One approach to approximating these *local Rademacher averages* is through information about the rate of growth of covering numbers of the class. For some subset A of a pseudometric space (S, d), let $\mathcal{N}(\epsilon, A, d)$ denote the cardinality of the smallest ϵ-cover of A, that is, the smallest set $\hat{A} \subset S$ for which every $a \in A$ has some $\hat{a} \in \hat{A}$ with $d(a, \hat{a}) \le \epsilon$. Using Dudley's entropy integral (Dudley 1999), Mendelson (2002) showed the following result. Suppose that $\mathcal{F}$ is a set of $[-1, 1]$-valued functions on $\mathcal{X}$ and that there are $\gamma > 0$ and $0 < p < 2$ for which

$$\sup_P \mathcal{N}\big(\epsilon, \mathcal{F}, L_2(P)\big) \le \gamma\epsilon^{-p},$$

where the supremum is over all probability distributions P on $\mathcal{X}$. Then for some constant $C_{\gamma,p}$ (which depends only on γ and p),

$$\begin{aligned}\frac{1}{n}\mathbf{E}\sup&\left\{\sum_{i=1}^n \sigma_i f(X_i) : f \in \mathcal{F}, \mathbf{E}f^2 \le \epsilon\right\}\\ &\le C_{\gamma,p}\max\big\{n^{-2/(2+p)}, n^{-1/2}\epsilon^{(2-p)/4}\big\}.\end{aligned}$$

Because $d_\phi(a, b) \le |a - b|$, any ϵ-cover of $\{f - f^* : f \in \mathcal{F}\}$ is an ϵ-cover of $\mathcal{F}_\phi$, so that $\mathcal{N}(\epsilon, \mathcal{F}_\phi, L_2(P)) \le \mathcal{N}(\epsilon, \mathcal{F}, L_2(P))$.

Now for the class absconv$(\mathcal{G})$ with $d_{\mathrm{VC}}(\mathcal{G}) = d$, we have

$$\sup_P \mathcal{N}\big(\epsilon, \operatorname{absconv}(\mathcal{G}), L_2(P)\big) \le Cd\epsilon^{-2d/(d+2)};$$

(see, e.g., Van der Vaart and Wellner 1996). Applying Mendelson's result shows that

$$\begin{aligned}\frac{1}{n}\mathbf{E}\sup&\left\{\sum_{i=1}^n \sigma_i f(X_i) : f \in B\operatorname{absconv}(\mathcal{G}), \mathbf{E}f^2 \le \epsilon\right\}\\ &\le C_d\max\big\{Bn^{-(d+2)/(2d+2)}, B^{d/(d+2)}n^{-1/2}\epsilon^{1/(d+2)}\big\}.\end{aligned}$$

Solving for $\epsilon_n^* \ge \xi_{g_\mathcal{F}}(\epsilon_n^*)$ shows that it suffices to choose

$$\epsilon_n^* = C'_d BL_B \max\left\{\left(\frac{L_B a_B}{B}\right)^{1/(d+1)}, 1\right\} n^{-(d+2)/(2d+2)}$$

for some constant C'_d that depends only on d.

APPENDIX B: LOSS, RISK, AND DISTANCE

We could construe R_ϕ as the risk under a loss function $\ell_\phi : \mathbb{R} \times \{\pm 1\} \to [0, \infty)$ defined by $\ell_\phi(\hat{y}, y) = \phi(\hat{y}y)$. The following result establishes that loss functions of this form are fundamentally unlike distance metrics.

Lemma B.1. Suppose that $\ell_\phi : \mathbb{R}^2 \to [0, \infty)$ has the form $\ell_\phi(x, y) = \phi(xy)$ for some $\phi : \mathbb{R} \to [0, \infty)$. Then the following results hold:

1. ℓ_ϕ is not a distance metric on $\mathbb{R}$.
2. ℓ_ϕ is a pseudometric on $\mathbb{R}$ iff $\phi \equiv 0$, in which case ℓ_ϕ assigns distance 0 to every pair of reals.

Proof. By hypothesis, ℓ_ϕ is nonnegative and symmetric. Another requirement of a distance metric is definiteness; for all $x, y \in \mathbb{R}$,

$$x = y \qquad \Longleftrightarrow \qquad \ell_\phi(x, y) = 0. \tag{B.1}$$

But we may write any $z \in (0, \infty)$ in two different ways, as $\sqrt{z}\sqrt{z}$ and, for example, $\sqrt{2z}\sqrt{(1/2)z}$. Satisfying (B.1) requires $\phi(z) = 0$ in the former case and $\phi(z) > 0$ in the latter case, an impossibility. This proves part 1.

To prove part 2, recall that a pseudometric relaxes (B.1) to the requirement

$$x = y \qquad \Longrightarrow \qquad \ell_\phi(x, y) = 0. \tag{B.2}$$

Because each $z \geq 0$ has the form xy for $x = y = \sqrt{z}$, (B.2) amounts to the necessary condition that $\phi \equiv 0$ on $[0, \infty)$. The final requirement on ℓ_ϕ is the triangle inequality, which in terms of ϕ becomes

$$\phi(xz) \leq \phi(xy) + \phi(yz) \quad \text{for all } x, y, z \in \mathbb{R}. \tag{B.3}$$

Because ϕ must vanish on $[0, \infty)$, taking $y = 0$ in (B.3) shows that only the zero function can (and does) satisfy the constraint.

[Received April 2003. Revised June 2005.]

REFERENCES

Arora, S., Babai, L., Stern, J., and Sweedyk, Z. (1997), "The Hardness of Approximate Optima in Lattices, Codes, and Systems of Linear Equations," *Journal of Computer and System Sciences*, 54, 317–331.

Bartlett, P. L. (1998), "The Sample Complexity of Pattern Classification With Neural Networks: The Size of the Weights Is More Important Than the Size of the Network," *IEEE Transactions on Information Theory*, 44, 525–536.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005), "Local Rademacher Complexities," *The Annals of Statistics*, 33, 1497–1537.

Bartlett, P. L., and Mendelson, S. (2005), "Empirical Minimization," *Probability Theory and Related Fields*, to appear.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992), "A Training Algorithm for Optimal Margin Classifiers," in *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, New York: ACM Press, pp. 144–152.

Bousquet, O. (2002), "A Bennett Concentration Inequality and Its Application to Suprema of Empirical Processes," *Comptes Rendus de l'Académie des Sciences, Série I*, 334, 495–500.

Boyd, S., and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge, U.K.: Cambridge University Press.

Breiman, L. (1999), "Prediction Games and Arcing Algorithms," *Neural Computation*, 11, 1493–1517.

——— (2004), "Population Theory for Boosting Ensembles," *The Annals of Statistics*, 32, 1–11.

Brown, L. D. (1986), *Fundamentals of Statistical Exponential Families*, Hayward, CA: Institute of Mathematical Statistics.

Collins, M., Schapire, R. E., and Singer, Y. (2002), "Logistic Regression, Adaboost and Bregman Distances," *Machine Learning*, 48, 253–285.

Cortes, C., and Vapnik, V. (1995), "Support-Vector Networks," *Machine Learning*, 20, 273–297.

Cristianini, N., and Shawe-Taylor, J. (2000), *An Introduction to Support Vector Methods*, Cambridge, U.K.: Cambridge University Press.

Devroye, L., Györfi, L., and Lugosi, G. (1996), *A Probabilistic Theory of Pattern Recognition*, New York: Springer-Verlag.

Dudley, R. M. (1999), *Uniform Central Limit Theorems*, Cambridge, U.K.: Cambridge University Press.

Feder, M., Figueiredo, M. A. T., Hero, A. O., Lee, C.-H., Loeliger, H.-A., Nowak, R., Singer, A. C., and Yu, B. (2004), "Special Issue on Machine Learning Methods in Signal Processing," *IEEE Transactions on Signal Processing*, 52, 2152.

Freund, Y., and Schapire, R. E. (1997), "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, 55, 119–139.

Friedman, J., Hastie, T., and Tibshirani, R. (2000), "Additive Logistic Regression: A Statistical View of Boosting," *The Annals of Statistics*, 28, 337–374.

Jiang, W. (2004), "Process Consistency for Adaboost," *The Annals of Statistics*, 32, 13–29.

Joachims, T. (2002), *Learning to Classify Text Using Support Vector Machines*, Dordrecht: Kluwer Academic.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999), "Introduction to Variational Methods for Graphical Models," *Machine Learning*, 37, 183–233.

Klein, T. (2002), "Une Inégalité de Concentration à Gauche Pour les Processus Empiriques" [A Left Concentration Inequality for Empirical Processes], *Comptes Rendus de l'Académie des Sciences, Série I*, 334, 501–504.

Koltchinskii, V. I., and Panchenko, D. (2000), "Rademacher Processes and Bounding the Risk of Function Learning," in *High Dimensional Probability II*, Vol. 47, eds. E. Giné, D. M. Mason, and J. A. Wellner, Boston, MA: Birkhäuser, pp. 443–459.

——— (2002), "Empirical Margin Distributions and Bounding the Generalization Error of Combined Classifiers," *The Annals of Statistics*, 30, 1–50.

Lebanon, G., and Lafferty, J. (2002), "Boosting and Maximum Likelihood for Exponential Models," in *Advances in Neural Information Processing Systems 14*, eds. T. Dietterich, S. Becker, and Z. Ghahramani, Cambridge, MA: MIT Press, pp. 447–454.

Ledoux, M. (2001), *The Concentration of Measure Phenomenon*, Providence, RI: American Mathematical Society.

Ledoux, M., and Talagrand, M. (1991), *Probability in Banach Spaces: Isoperimetry and Processes*, New York: Springer-Verlag.

Lee, W. S., Bartlett, P. L., and Williamson, R. C. (1996), "Efficient Agnostic Learning of Neural Networks With Bounded Fan-in," *IEEE Transactions on Information Theory*, 42, 2118–2132.

Lin, Y. (2004), "A Note on Margin-Based Loss Functions in Classification," *Statistics and Probability Letters*, 68, 73–82.

Lugosi, G., and Vayatis, N. (2004), "On the Bayes Risk Consistency of Regularized Boosting Methods," *The Annals of Statistics*, 32, 30–55.

Lugosi, G., and Wegkamp, M. (2004), "Complexity Regularization via Localized Random Penalties," *The Annals of Statistics*, 32, 1679–1697.

Mammen, E., and Tsybakov, A. B. (1999), "Smooth Discrimination Analysis," *The Annals of Statistics*, 27, 1808–1829.

Mannor, S., and Meir, R. (2001), "Geometric Bounds for Generalization in Boosting," in *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory*, eds. D. Helmbold and R. Williamson, Springer-Verlag, pp. 461–472.

Mannor, S., Meir, R., and Zhang, T. (2002), "The Consistency of Greedy Algorithms for Classification," in *Proceedings of the Annual Conference on Computational Learning Theory*, eds. J. Kivinen and R. Sloan, Springer-Verlag, pp. 319–333.

Marron, J. S., and Todd, M. (2002), "Distance-Weighted Discrimination," Technical Report 1339, Cornell University, School of Operations Research and Industrial Engineering.

Massart, P. (2000a), "About the Constants in Talagrand's Concentration Inequality for Empirical Processes," *The Annals of Probability*, 28, 863–884.

——— (2000b), "Some Applications of Concentration Inequalities to Statistics," *Annales de la Faculté des Sciences de Toulouse*, IX, 245–303.

Mendelson, S. (2002), "Improving the Sample Complexity Using Global Data," *IEEE Transactions on Information Theory*, 48, 1977–1991.

Nesterov, Y., and Nemirovskii, A. (1994), *Interior-Point Polynomial Algorithms in Convex Programming*, Philadelphia: SIAM Publications.

Rio, E. (2001), "Inégalités de Concentration Pour les Processus Empiriques de Classes de Parties" [Concentration Inequalities for Set-Indexed Empirical Processes], *Probability Theory and Related Fields*, 119, 163–175.

Rockafellar, R. T. (1997), *Convex Analysis*, Princeton, NJ: Princeton University Press.

Schapire, R. E., Freund, Y., Bartlett, P., and Lee, W. S. (1998), "Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods," *The Annals of Statistics*, 26, 1651–1686.

Schölkopf, B., and Smola, A. (2002), *Learning With Kernels*, Cambridge, MA: MIT Press.

Schölkopf, B., Tsuda, K., and Vert, J.-P. (2003), *Kernel Methods in Computational Biology*, Cambridge, MA: MIT Press.

Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., and Anthony, M. (1998), "Structural Risk Minimization Over Data-Dependent Hierarchies," *IEEE Transactions on Information Theory*, 44, 1926–1940.

Shen, X., Tseng, G. C., Zhang, X., and Wong, W. H. (2003), "On Psi-Learning," *Journal of the American Statistical Association*, 98, 724–734.

Steinwart, I. (2005), "Consistency of Support Vector Machines and Other Regularized Kernel Classifiers," *IEEE Transactions on Information Theory*, 51, 128–142.

Talagrand, M. (1994), "Sharper Bounds for Gaussian and Empirical Processes," *The Annals of Probability*, 22, 28–76.

Tsybakov, A. (2004), "Optimal Aggregation of Classifiers in Statistical Learning," *The Annals of Statistics*, 32, 135–166.

Van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer-Verlag.

Zhang, T. (2003), "Sequential Greedy Approximation for Certain Convex Optimization Problems," *IEEE Transactions on Information Theory*, 49, 682–691.

——— (2004), "Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization," *The Annals of Statistics*, 32, 56–85.

Zhang, T., and Yu, B. (2005), "Boosting With Early Stopping: Convergence and Consistency," *The Annals of Statistics*, 33, 1538–1579.