Convexity, Classification, and Risk Bounds

Peter L. Bartlett*

Computer Science Division and Department of Statistics

University of California, Berkeley

bartlett@stat.berkeley.edu

Michael I. Jordan

Computer Science Division and Department of Statistics

University of California, Berkeley

jordan@stat.berkeley.edu

Jon D. McAuliffe

Department of Statistics

University of California, Berkeley

jon@stat.berkeley.edu

June 16, 2005

*Peter L. Bartlett and Michael I. Jordan are faculty in the Department of Statistics and the Division of Computer Science, and Jon D. McAuliffe is a graduate student in the Department of Statistics, at the University of California, Berkeley, 94720. We wish to acknowledge the support of grant number 0412995 from the National Science Foundation and MURI grant DAAD 190210383 from the ARO. The authors would also like to thank Gilles Blanchard, Olivier Bousquet, Pascal Massart, Ron Meir, Shahar Mendelson, Martin Wainwright and Bin Yu for helpful discussions.

Abstract

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 satisfy 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 we 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.

Keywords: machine learning, convex optimization, boosting, support vector machine, Rademacher complexity, empirical process theory

1. INTRODUCTION

Convexity has become an increasingly important theme in applied mathematics and engineering, having acquired a prominent role akin to the one 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). It is, however, only in recent years that the systematic exploitation of the algorithmic consequences of convexity has 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 et al., 1992, Cortes and Vapnik, 1995, Cristianini and Shawe-Taylor, 2000, Schölkopf and Smola, 2002), boosting (Freund and Schapire, 1997, Collins et al., 2002, Lebanon and Lafferty, 2002), and variational inference for graphical models (Jordan et al., 1999) are all based directly on ideas from convex optimization. These methods have had significant practical successes in applied areas such as bioinformatics, information management and signal processing (Feder et al., 2004, Joachims, 2002, Schölkopf et al., 2003).

If algorithms from convex optimization are to continue to make inroads into statistical theory and practice, it is important that we understand these algorithms not only from a computational point of view 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 the current paper 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 one if $\alpha \leq 0$ and zero otherwise, we wish to choose $f \in \mathcal{F}$ that minimizes the risk $R(f) = \mathbf{E}\ell(Yf(X)) = P(Y \neq \mathrm{sign}(f(X)))$.

Given a sample $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$, 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 et al., 1997). Indeed, the loss function $\ell(Yf(X))$ is non-convex in its (scalar) argument, and, while 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 θ, then the expectation of $\phi(Yf(X))$ is convex in θ (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 et al., 2000). In the machine learning literature, these convexity-based methods have largely displaced earlier non-convex methods such as neural networks.

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) have provided such a result under the assumption that the surrogate ϕ is differentiable, monotone, strictly convex, and satisfies $\phi(0) = 1$. This handles all of the cases shown in Figure 1 except the support vector machine. Steinwart (2005) has 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 continuous. Other results on Bayes-risk consistency have been presented by Breiman (2000), Jiang (2004), Mannor and Meir (2001), and Mannor et al. (2002).

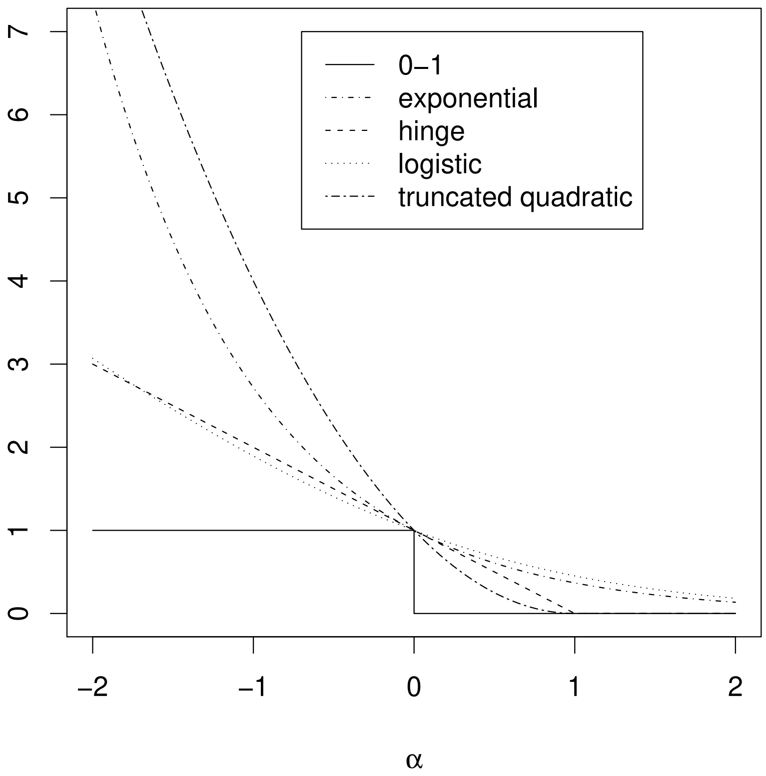

Figure 1: A plot of the 0-1 loss function and surrogates corresponding to various practical classifiers. These functions are plotted as a function of the margin $\alpha = yf(x)$. Note that a classification error is made if and only if the margin is negative; thus the 0-1 loss is a step function that is equal to one 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 8.

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 VC-dimension (Bartlett, 1998, Shawe-Taylor et al., 1998), such as the function classes used by AdaBoost (see, e.g., Schapire et al., 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, it is necessary to find general quantitative relationships between the approximation and estimation errors associated with ϕ and those associated with 0-1 loss. This point has been emphasized by Zhang (2004), who has presented several examples of such relationships. 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 us refer to $R_\phi(f) = \mathbf{E}\phi(Yf(X))$ as the "ϕ-risk," and let $R_\phi^* = \inf_f R_\phi(f)$ denote the "optimal ϕ-risk." We show that, for all measurable f,

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

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, while conditions such 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, $\text{sign}(2\eta(x) - 1)$. 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 (2001) considered the rate of convergence 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, we might expect a convergence rate no faster than $1/\sqrt{n}$. However, under Tsybakov's assumption, it can be as fast as $1/n$. We show that minimizing empirical ϕ-risk also leads to surprisingly fast convergence rates under this assumption. In particular, if ϕ is uniformly convex, 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:

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

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 zero—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 zero for the minimizer of the empirical risk. On the other hand, we can also split the upper bound on excess risk into an estimation term and an approximation term:

$$\psi(R(f) - R^*) \leq R_\phi(f) - R_\phi^* = \left(R_\phi(f) - \inf_{g \in \mathcal{F}} R_\phi(g)\right) + \left(\inf_{g \in \mathcal{F}} R_\phi(g) - R_\phi^*\right).$$

Often, it is possible to minimize ϕ-risk efficiently. Thus, while 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 paper 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. We present our conclusions in Section 5.

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 be considered 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 the section we (a) work with population expectations and (b) 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, since 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 Appendix 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 Y the identity function on $\{-1, 1\}$, so that P is the distribution of (X, Y), i.e., 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

$$R(f) = P(\text{sign}(f(X)) \neq Y),$$

where $\text{sign}(\alpha) = 1$ for $\alpha > 0$ and -1 otherwise. (The particular choice of the value of sign(0) is not important, but we need to fix some value in $\{\pm 1\}$ for the definitions that follow.) Based on an i.i.d. sample $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$, 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 $\text{sign}(f(X)) = \text{sign}(\eta(X) - 1/2)$ a.s. on $\{\eta(X) \neq 1/2\}$ has $R(f) = R^*$.

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

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

Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$. Let $f_n = \hat{f}_\phi$ be a function in $\mathcal{F}$ which 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*

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

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 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*

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

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

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

and then

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

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

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

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 $\eta \neq 1/2$,

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

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.

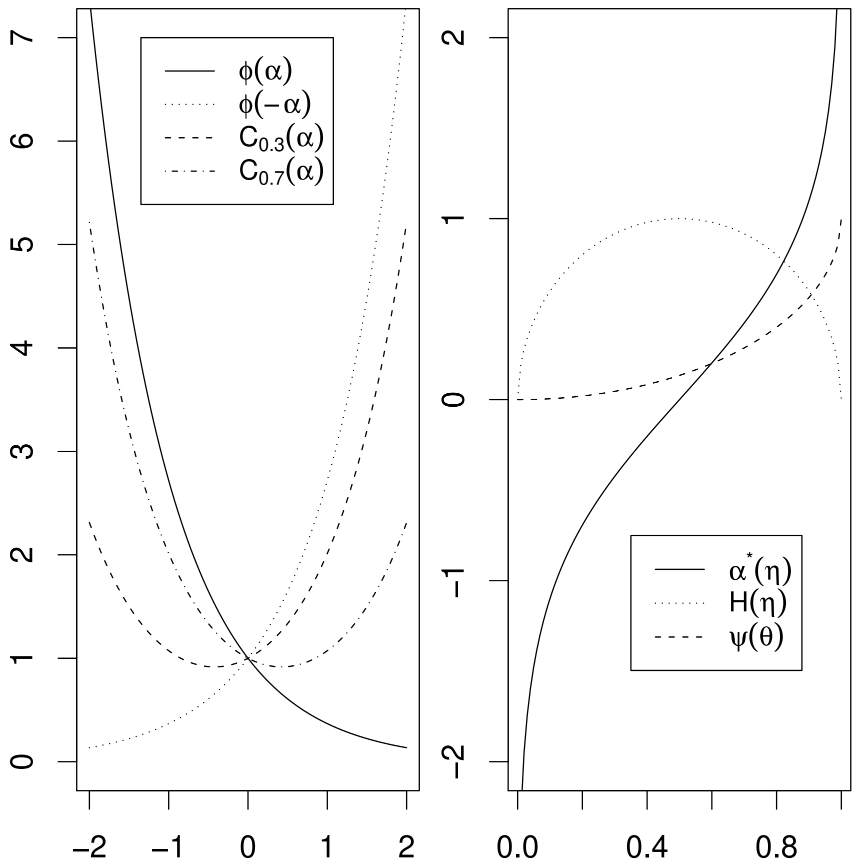

Figure 2: Exponential loss. The left panel shows $\phi(\alpha)$, its reflection $\phi(-\alpha)$, and two different convex combinations of these functions, for $\eta = 0.3$ and $\eta = 0.7$. Note that the minima of these combinations are the values $H(\eta)$, and the minimizing arguments are the values $\alpha^*(\eta)$. The right panel shows $H(\eta)$ and $\alpha^*(\eta)$ plotted as a function of η, and the ψ-transform $\psi(\theta)$.

Figure 2, left panel, shows $\phi(\alpha)$, $\phi(-\alpha)$, and the generic conditional ϕ-risk $C_\eta(\alpha)$ for $\eta = 0.3$ and $\eta = 0.7$. In this case, ϕ is strictly convex on $\mathbb{R}$, 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 it 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

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

Notice that this expression is correct also for η equal to 0 or 1. The right panel of Figure 2 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 $\eta \neq 1/2$, so the exponential loss is classification-calibrated.

2.3 The ψ-transform and the relationship between excess risks

We begin by defining a functional transform of the loss function. Theorem 3 below 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

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

Here $\overline{\text{co}}\, S$ is the closure of the convex hull of the set S, and 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 below in Lemma 6, part 7.

Recall that g is convex if and only if epi g is a convex set, and g is closed (epi g is a closed set) if and only if g is lower semicontinuous (Rockafellar, 1997). By Lemma 6, part 5, $\tilde{\psi}$ is continuous, so in fact the closure operation in Definition 2 is vacuous. We therefore 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 3. 1. *For any nonnegative loss function ϕ, any measurable $f : \mathcal{X} \to \mathbb{R}$ and any probability distribution on $\mathcal{X} \times \{\pm 1\}$,*

$$\psi(R(f) - R^*) \leq R_\phi(f) - R^*_\phi.$$

2. *Suppose $|\mathcal{X}| \geq 2$. 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 $[0, 1]$,*

$$\psi(\theta_i) \to 0 \quad \textit{if and only if} \quad \theta_i \to 0.$$

   (c) *For every sequence of measurable functions $f_i : \mathcal{X} \to \mathbb{R}$ and every probability distribution* *on* $\mathcal{X} \times \{\pm 1\}$,

$$R_\phi(f_i) \to R_\phi^* \quad \textit{implies} \quad R(f_i) \to R^*.$$

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 3(1) as $\psi^{-1}(R_\phi(f) - R_\phi^*)$. Invertibility follows from convexity of ψ together with Lemma 6, parts 6, 8, and 9 below.

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) show 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 zero. Clearly, the conclusions of Theorem 3 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 $R(f_i) \to R^*$. Theorem 3 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 zero, and in that case the ψ-transform takes a simplified form.

Theorem 4. *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 we present a proof of Theorem 3. 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 paper.

Lemma 5. *Suppose $g : \mathbb{R} \to \mathbb{R}$ 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 1, $g(\lambda x) = g(\lambda x + (1-\lambda)0) \leq \lambda g(x) + (1-\lambda) g(0) = \lambda g(x)$. To see 2, put $\lambda = y/x$ in 1. For 3, rewrite 2 as $g(y)/y \leq g(x)/x$. □

Lemma 6. *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* $\eta \in [0,1]$,

$$H(\eta) = H(1-\eta), \qquad H^-(\eta) = H^-(1-\eta), \qquad \psi(\eta) = \psi(-\eta).$$

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* $\theta \in (0, 1]$.

Proof. See Appendix A. $\square$

Proof. (Of Theorem 3). For Part 1, it is straightforward to show that

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

where $\mathbf{1}[\Phi]$ is 1 if the predicate Φ is true and 0 otherwise (see, for example, Devroye et al., 1996). We can apply Jensen's inequality, since ψ is convex by definition, and the fact that $\psi(0) = 0$ (Lemma 6, part 8) to show that

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

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

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

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 $C_{\eta(x)}(f(x)) \geq H(\eta(x))$.

For Part 2, the first inequality is from Part 1. For the second, 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 $\psi(\theta) \geq \gamma\tilde{\psi}(\alpha_1) + (1-\gamma)\tilde{\psi}(\alpha_2) - \epsilon/2$. 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 $f(x_1) \leq 0$, $f(x_2) \leq 0$, $C_{\eta(x_1)}(f(x_1)) \leq H^-(\eta(x_1)) + \epsilon/2$ and $C_{\eta(x_2)}(f(x_2)) \leq H^-(\eta(x_2)) + \epsilon/2$. Then we have

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

Furthermore, since $\mathrm{sign}(f(x_1)) = \mathrm{sign}(f(x_2)) = -1$ but $\eta(x_1), \eta(x_2) \geq 1/2$,

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

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

(3b') For any sequence (θ_i) in $[0,1]$,

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

To see that (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 6, 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 (3b') implies (3c), suppose that $R_\phi(f_i) \to R_\phi^*$. By Part 1, $\psi(R(f_i) - R^*) \to 0$, and (3b') implies $R(f_i) \to R^*$.

Finally, to see that (3c) implies (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 $\mathrm{sign}(\alpha_i(\eta - 1/2)) = -1$ 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 $R_\phi(f_i) \to R_\phi^*$.

$\square$

2.4 Examples

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

Example 2 (Exponential loss). Since $\phi(\alpha) = \exp(-\alpha)$ is convex, differentiable and decreasing, Theorem 4 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 4 Part 2,

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

The right panel of Figure 2 shows the graph of ψ over the interval $[0, 1]$. (We shall see below that $\psi(\theta) = \psi(-\theta)$ 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 $\phi(-\alpha)$, $C_{0.3}(\alpha)$, and $C_{0.7}(\alpha)$ in the left panel of Figure 3. This function is convex, differentiable, and decreasing at zero, and thus classification-calibrated. If $\eta = 0$, 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 $\eta = 1$. On the other hand, when $0 < \eta < 1$, C_η is strictly convex with a (unique)

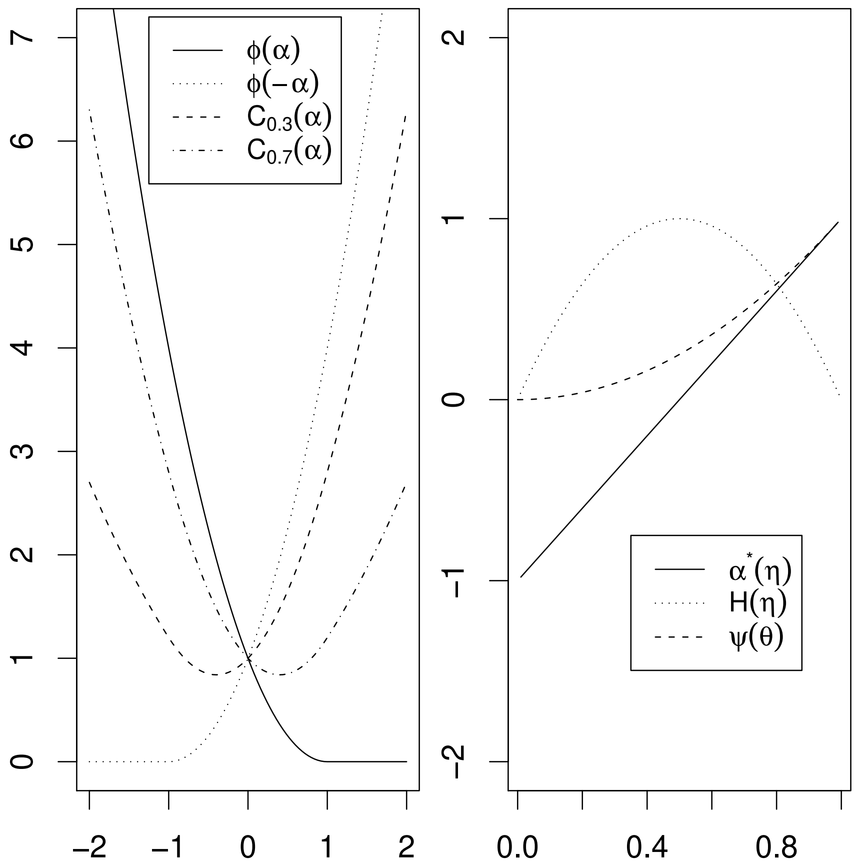

Figure 3: Truncated quadratic loss.

stationary point, and solving for it yields

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

Notice 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 $[0, 1]$.

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

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

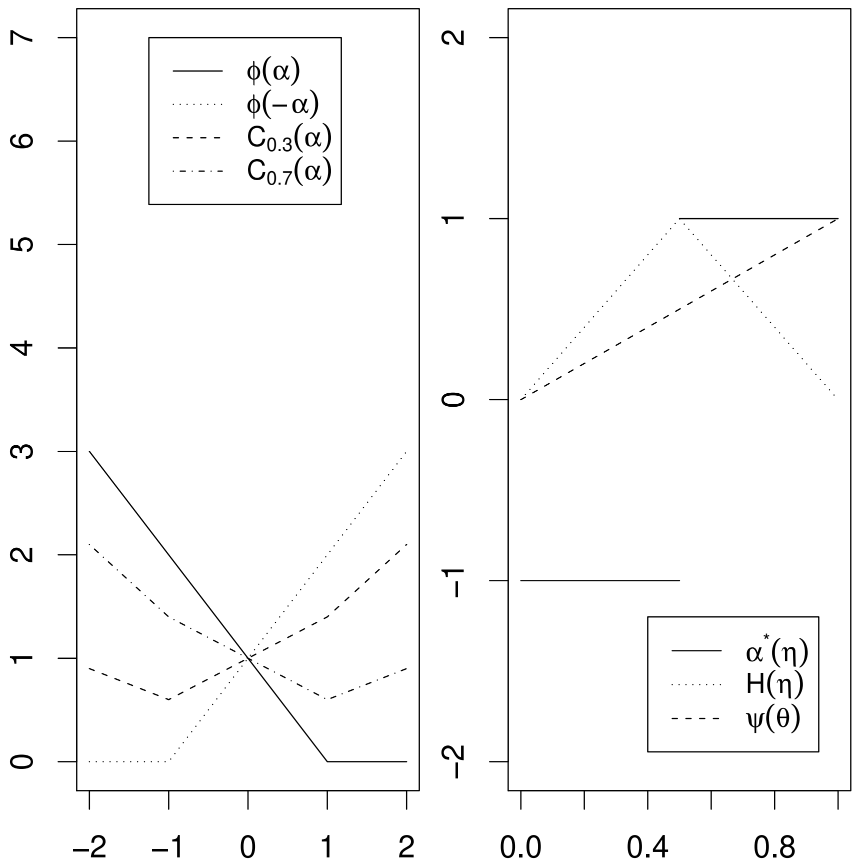

Figure 4: Hinge loss.

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

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

The right panel of Figure 3 shows α^*, H, and ψ.

Example 4 (Hinge loss). Here we take $\phi(\alpha) = \max\{1 - \alpha, 0\}$, which is shown in the left panel of Figure 4 along with $\phi(-\alpha)$, $C_{0.3}(\alpha)$, and $C_{0.7}(\alpha)$. Again, ϕ is convex, 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 $\alpha \leq -1$ makes $C_\eta(\alpha)$ vanish, just as in Example 3. The same holds for $\alpha \geq 1$ 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. Since (3) yields valid minima at 0, 1/2, and 1 also, we could choose to extend it to the whole unit interval. Regardless, a simple direct verification as in the previous examples shows

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

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

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

We present α^*, H, and ψ in the right panel of Figure 4.

Example 5 (Distance weighted discrimination). Marron and Todd (2002) introduced the distance weighted discrimination method for high dimension, low 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

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

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

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

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 $p > 1$. It 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

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

and

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

Example 7 (Sigmoid loss). We conclude by examining a non-convex loss function. Let $\phi(\alpha) = 1 - \tanh(k\alpha)$ for some fixed $k > 0$. Figure 5, left panel, depicts $\phi(\alpha)$ with $k = 1$, as well as $\phi(-\alpha)$, $C_{0.3}(\alpha)$, and $C_{0.7}(\alpha)$. 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

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

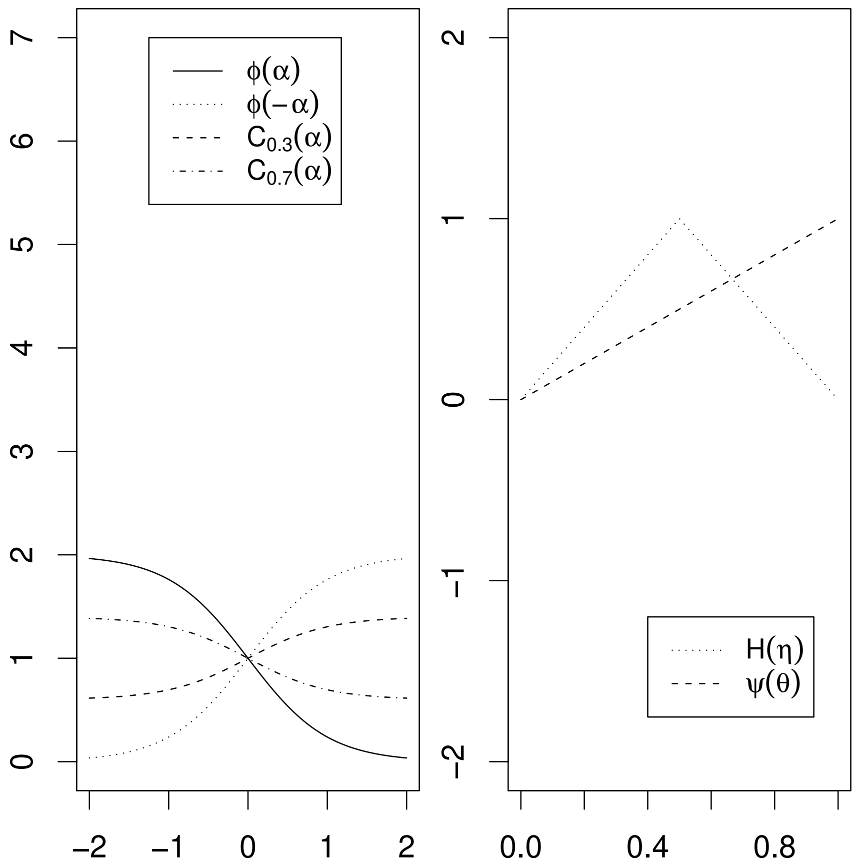

Figure 5: Sigmoid loss.

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

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

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

The following example illustrates the difficulties with nondifferentiability at zero, 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, notice 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 zero 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 $2\eta \ge 1-\eta$ (that is, $1/3 \le \eta < 1/2$), 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 we consider involves the Bregman divergence of ϕ at 0. If ϕ is convex and classification-calibrated, then it is differentiable at zero, and we can define the Bregman divergence of ϕ at 0:

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

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

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

Since 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 7. *For convex, classification-calibrated ϕ,*

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

Notice that a slower increase of ξ (that is, 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 utilize surrogate loss functions which are either upper bounds on 0-1 loss or can be transformed into upper bounds via 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 8. *If $\phi : \mathbb{R} \to [0, \infty)$ is classification-calibrated, then there is a $\gamma > 0$ such that $\gamma\phi(\alpha) \geq \mathbf{1}[\alpha \leq 0]$ for all $\alpha \in \mathbb{R}$.*

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

Example 9. Consider the following (classification-calibrated) loss function; see the left panel of Figure 6.

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

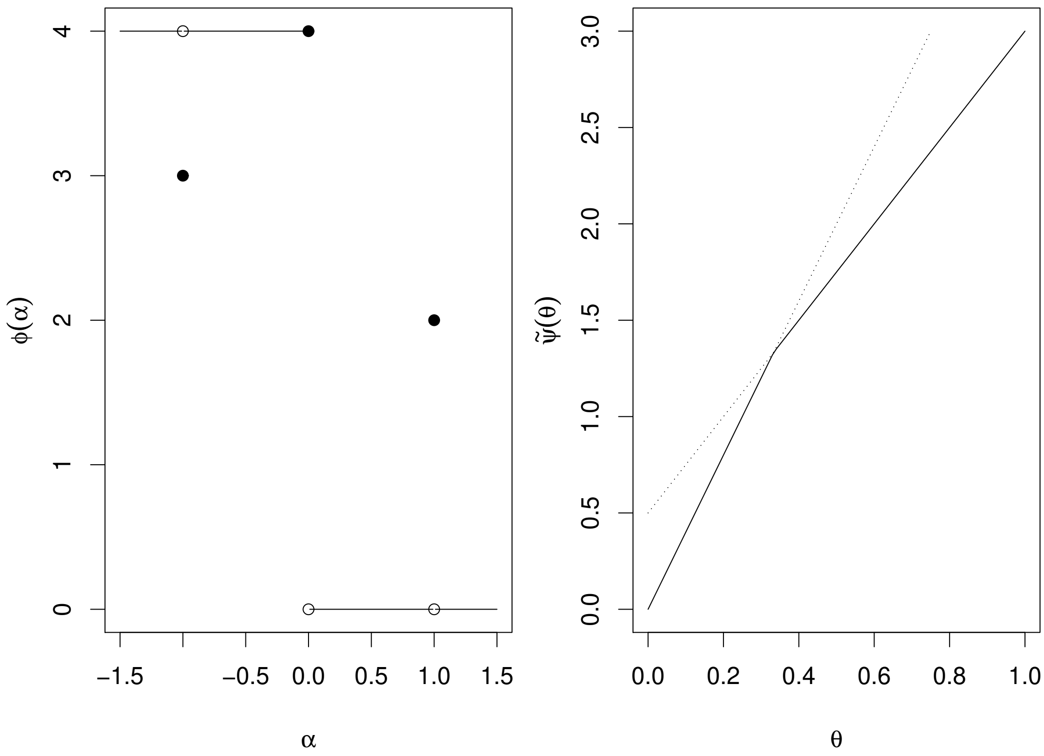

Figure 6: Left panel, the loss function of Example 9. Right panel, the corresponding (nonconvex) $\tilde{\psi}$. The dotted lines depict the graphs for the two linear functions of which $\tilde{\psi}$ is a pointwise minimum.

It is easy to check that

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

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

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

so

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

This function, illustrated in the right panel of Figure 6, is not convex; in fact it is concave. Thus $\psi \neq \tilde{\psi}$.

3. TIGHTER BOUNDS UNDER LOW NOISE CONDITIONS

Predicting the optimal class label is difficult in regions where $\eta(X)$ is close to $1/2$, since 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. In order 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 (2001).

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

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

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

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

These conditions are equivalent.

Lemma 9. *For $0 \leq \beta < \infty$, 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 $f : \mathcal{X} \to \{\pm 1\}$,

$$\Pr\left(f(X)(\eta(X) - 1/2) < 0\right) \leq c(R(f) - R^*). \tag{8}$$

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

$$\begin{aligned}
R(f) - R^* &= \mathbf{E}\left(\mathbf{1}\left[f(X) \neq \operatorname{sign}(\eta(X) - 1/2)\right] |2\eta(X) - 1|\right) \\
&= \mathbf{E}\left(\mathbf{1}\left[f(X)(\eta(X) - 1/2) < 0\right] |2\eta(X) - 1|\right) \qquad (9) \\
&\leq P_X\left(f(X)(\eta(X) - 1/2)\right),
\end{aligned}$$

which implies $\alpha \leq 1$. If $\alpha = 0$, 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 10. *Suppose P has noise exponent $0 < \alpha \leq 1$, and ϕ is classification-calibrated. Then there is a $c > 0$ such that for any $f : \mathcal{X} \to \mathbb{R}$,*

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

Furthermore, this never gives a worse rate than the result of Theorem 3, since

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

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

$$P_X\left(\operatorname{sign}(f(X)) \neq \operatorname{sign}(\eta(X) - 1/2)\right) \leq c\left(R(f) - R^*\right)^\alpha.$$

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

$$
\begin{aligned}
R(f) - R^* &= \mathbf{E}\left(\mathbf{1}\left[\operatorname{sign}(f(X)) \neq \operatorname{sign}(\eta(X) - 1/2)\right] |2\eta(X) - 1|\right) \\
&= \mathbf{E}\left(\mathbf{1}\left[|2\eta(X) - 1| < \epsilon\right] \mathbf{1}\left[\operatorname{sign}(f(X)) \neq \operatorname{sign}(\eta(X) - 1/2)\right] |2\eta(X) - 1|\right) \\
&\quad + \mathbf{E}\left(\mathbf{1}\left[|2\eta(X) - 1| \geq \epsilon\right] \mathbf{1}\left[\operatorname{sign}(f(X)) \neq \operatorname{sign}(\eta(X) - 1/2)\right] |2\eta(X) - 1|\right) \\
&\leq c\epsilon \left(R(f) - R^*\right)^{\alpha} \\
&\quad + \mathbf{E}\left(\mathbf{1}\left[|2\eta(X) - 1| \geq \epsilon\right] \mathbf{1}\left[\operatorname{sign}(f(X)) \neq \operatorname{sign}(\eta(X) - 1/2)\right] |2\eta(X) - 1|\right).
\end{aligned}
$$

Now, for any x,

$$
\mathbf{1}\left[|2\eta(x) - 1| \geq \epsilon\right] |2\eta(x) - 1| \leq \frac{\epsilon}{\psi(\epsilon)} \psi(|2\eta(x) - 1|). \tag{10}
$$

Indeed, when $|2\eta(x) - 1| < \epsilon$, (10) follows from the fact that ψ is nonnegative (Lemma 6, parts 8, 9), and when $|2\eta(x) - 1| \geq \epsilon$ it follows from Lemma 5(2).

Thus, using the same argument as in the proof of Theorem 3,

$$
\begin{aligned}
R(f) - R^* &\leq c\epsilon \left(R(f) - R^*\right)^{\alpha} + \frac{\epsilon}{\psi(\epsilon)} \mathbf{E}\left(\mathbf{1}\left[\operatorname{sign}(f(X)) \neq \operatorname{sign}(\eta(X) - 1/2)\right] \psi\left(|2\eta(X) - 1|\right)\right) \\
&\leq c\epsilon \left(R(f) - R^*\right)^{\alpha} + \frac{\epsilon}{\psi(\epsilon)} \left(R_\phi(f) - R_\phi^*\right),
\end{aligned}
$$

and hence,

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

Choosing

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

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

The second inequality follows from the fact that $\psi(\theta)/\theta$ is non-decreasing, which we know from Lemma 5, part 3. $\square$

4. ESTIMATION RATES

In previous sections, we have seen 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:

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

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

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

then

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

since $\hat{f}$ minimizes $\hat{R}_\phi$. However, 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 $1/\sqrt{n}$. However, if $\mathcal{F}$ is a small class (for instance, 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 et al. (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 (the difference between the loss of a function f and that of the optimal f^*) in terms of its expectation. Since 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 11 (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 $\epsilon > 0$, we say that f is *strictly convex* with respect to d.

For example, for $S = \mathbb{R}$, d 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

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

(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

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

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,

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

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 (2003)—see also Massart (2000b), Koltchinskii and Panchenko (2000), Mendelson (2002), Lugosi and Wegkamp (2004) and Bartlett et al. (2002). 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,

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

(Notice 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, and this 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 12. *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*

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

where

$$\begin{aligned}
\epsilon_n &= K \max\left\{\epsilon_n^*, \left(\frac{c_r L^2 \ln(1/\delta)}{n}\right)^{1/(2-\beta)}, \frac{BL\ln(1/\delta)}{n}\right\},\\
\epsilon_n^* &\ge \xi_{g_{\mathcal{F}}}(\epsilon_n^*),\\
c_r &= \begin{cases} (2c)^{-2/r} & \text{if } r \ge 2,\\ (2c)^{-1}B^{2-r} & \text{otherwise,}\end{cases}\\
\beta &= \min\left(1, \frac{2}{r}\right).
\end{aligned}$$

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 $1-\delta$,

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

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, the use of 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 hand side. The rate of decrease of classification risk improves as the noise exponent increases.

Consider the impact on the bound of the modulus of complexity of the loss function: for flatter loss functions, where the exponent of the modulus of complexity is $r > 2$, 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 $r \leq 2$, the rate can be as good as n^{-1}. On the other hand, we have seen that a more curved ϕ leads to a worse ψ. However, if the noise exponent is $\alpha = 1$, the bound is optimized by a more curved ϕ, with $r \geq 2$.

Shen et al. (2003) show that fast rates are also possible under the low noise assumption for a particular non-convex ϕ. In that case, however, minimization of empirical ϕ-risk requires the use of heuristics, since the optimization problem cannot be solved efficiently.

In the remainder of this section we present a proof of Theorem 12. There are two key ingredients in the proof, 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 condition is satisfied for convex function classes.

Lemma 13. *Consider a class $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathbb{R}$ with $\sup_{f\in\mathcal{F}} \|f\|_\infty \leq B$. Let P be a* *probability distribution on $\mathcal{X}$, and suppose that there are $c \geq 1$ and $0 < \beta \leq 1$ such that, for all $f \in \mathcal{F}$,*

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

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

$$\hat{\mathbf{E}}f \leq \alpha\epsilon \Rightarrow \mathbf{E}f \leq \epsilon.$$

provided that

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

where K is an absolute constant and

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

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

$$P_X\left(f(X) \neq f^*(X)\right) \leq c\left(\mathbf{E}\ell(f(X),Y) - \mathbf{E}\ell(f^*(X),Y)\right)^\alpha$$

can be rewritten

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

Thus, we can obtain a version of Tsybakov's result for small function classes from Lemma 13: if the Bayes decision rule f^* is in $\mathcal{F}$, then the function $\hat{f}$ that minimizes empirical risk has

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

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

The second ingredient in the proof of Theorem 12 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 $\mathcal{X}$,

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

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

Lemma 14. *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:*

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

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

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

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

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

We shall 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 will imply 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 15. *Suppose that a convex function $\ell : \mathbb{R} \to \mathbb{R}$ has modulus of convexity δ with respect to a pseudometric d on $\mathbb{R}$, for some fixed $c, r > 0$, every $\epsilon > 0$ satisfies*

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

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

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

where $c_r = c$ if $r \geq 2$ 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 12)* Consider the class $\{g_f : f \in \mathcal{F}\}$ with, for each $f \in \mathcal{F}$,

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

where $f^* \in \mathcal{F}$ minimizes $R_\phi(f) = \mathbf{E}\phi(Yf(X))$. Applying Lemma 15, 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 14,

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

which is equivalent to

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

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 13 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

$$\hat{\mathbf{E}}g_f = \hat{\mathbf{E}}\phi(Y(\gamma f(X) + (1-\gamma)f^*(X))) - \hat{\mathbf{E}}\phi(Yf^*(X)) \leq \gamma\left(\hat{\mathbf{E}}\phi(Yf(x)) - \hat{\mathbf{E}}\phi(Yf^*(X))\right) \leq \alpha\epsilon.$$

Applying Lemma 13 we have, with probability at least $1 - e^{-x}$, any g_f with $\hat{\mathbf{E}}g_f \leq \epsilon/2$ also has $\mathbf{E}g_f \leq \epsilon$, provided

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

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

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

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

Combining with Theorem 10 shows that, for some c',

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

□

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

$$
d_\phi(a,b)=\inf\left\{|a-\alpha|+|\beta-b| :\ \phi \text{ constant on } (\min\{\alpha,\beta\},\max\{\alpha,\beta\})\right\}.
$$

For all except the truncated quadratic loss function, this corresponds to the standard metric on $\mathbb{R}$, $d_\phi(a,b)=|a-b|$. In all cases, $d_\phi(a,b)\le|a-b|$, 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 Vapnik-Chervonenkis dimension, and we constrain the size of the class by restricting the ℓ_1 norm of the linear parameters. If $\mathcal{G}$ is the VC-class, we write $\mathcal{F}=B\,\mathrm{absconv}(\mathcal{G})$, for some constant B, where

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

Theorem 16. *Let $\phi:\mathbb{R}\to\mathbb{R}$ be a convex loss function. Suppose that, on the interval $[-B,B]$, ϕ*

	$\phi(\alpha)$	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	$(\max\{0,1-\alpha\})^2$	$2(B+1)$	$\epsilon^2/4$

Table 1: Four convex loss functions defined on $\mathbb{R}$. 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.

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 i.i.d. 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 $\mathcal{F}=B\,\mathrm{absconv}(\mathcal{G})$, where $\mathcal{G}\subseteq\{\pm1\}^{\mathcal{X}}$ has VC-dimension V, and

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

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

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

4.2 Simulations

This section describes a set of simulations which 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 12. 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:

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

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 9 implies that η_q has noise exponent $1/(q+1)$. The constant C was chosen 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 $p > 1$:

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

The leading constant ensures 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 4 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, since $(\phi(\epsilon) + \phi(-\epsilon))/2 - \phi(0) = \epsilon^p$ 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 below. For each choice of n, p, and q, 25 repetitions of the following procedure were performed. First, we generated a dataset according to $(P_{\mathcal{X}}, \eta_q(x))$ and found the empirical risk minimizer $\hat{f}_n$ over $\mathcal{F}$, via 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 3 to obtain the inequality

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

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

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

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 12. For the class $\mathcal{F}$ under consideration, the centered empirical process $\xi_{\mathcal{F}}(\epsilon)$ utilized in the theorem can be pointwise upper-bounded using a local Rademacher average symmetrization, which is in turn bounded by the Dudley entropy integral. The derivation closely follows the proof of Theorem 16 (see also Bartlett et al. (2002)). These calculations reveal that a suitable upper bound on ϵ^* in Theorem 12 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 $n^{-1/(2-2/p)}$ 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 -0.7 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. Though the match when $p = 3.5$ is less exact, the simulated results appear compatible with the theoretical rate to within the noise tolerance.

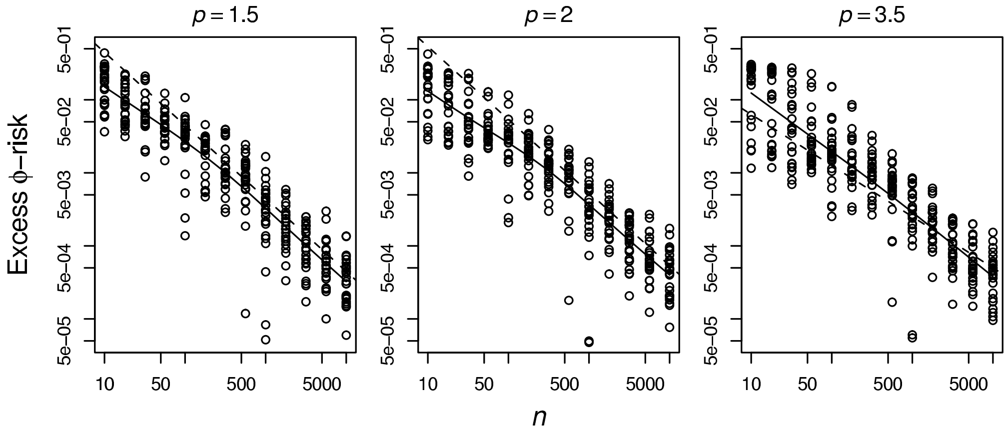

Figure 7: Rate plots. Each panel shows simulated excess ϕ-risk on the log scale versus simulated sample size on the log scale, for the choice of p given at top. We took $q = 1$ in each case. Natural cubic spline fits appear as solid lines. The dashed line depicts the slope corresponding to the theoretical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

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 which, based on an i.i.d. 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 which 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 readily verifiable in most practical examples. In addition, the convexification step 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 (2001), 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 are given by Breiman (2000), Mannor et al. (2002) and Lugosi and Vayatis (2004). Regarding the numerical optimization to determine $\hat{f}$, Zhang and Yu (2003) give novel bounds on the convergence rate for generic forward stagewise additive modeling (see also Zhang, 2002). These authors focus 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 6

Proof. 1 is immediate from the definitions.

For 2, concavity follows because H is an infimum of concave (affine) functions of η. Now, since H is concave and symmetric about 1/2, $H(1/2) = H((1/2)\eta + (1/2)(1-\eta)) \geq (1/2)H(\eta) +$

$(1/2)H(1-\eta) = H(\eta)$. Thus H is maximal at $1/2$. To see that $H(1/2) = H^-(1/2)$, notice that $\alpha(2\eta - 1) \le 0$ for all α when $\eta = 1/2$.

To prove 3, assume that there is an $\eta \neq 1/2$ 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,

$$\liminf_{i\to\infty} \left(\eta\phi(\alpha_i) + (1-\eta)\phi(-\alpha_i)\right) \ge H(\eta) = H(1/2) = \lim_{i\to\infty} \frac{\phi(\alpha_i) + \phi(-\alpha_i)}{2}, \tag{13}$$

Rearranging, we have

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

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

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

It follows that

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

so all the expressions in (13) 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 $H(\eta) = H(1/2)$, ϕ is not classification-calibrated.

For 4, H^- is concave on $[0, 1/2]$ by the same argument as for the concavity of H. (Notice 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 5, first notice that the concavity of H implies that it is continuous on the relative interior of its domain, i.e., $(0, 1)$. 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, on the other hand, 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

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

Since this is true for any ϵ, $\limsup_{\eta \to 1} H(\eta) \leq H(1)$, 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 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 6, observe that ψ is a closed convex function with locally simplicial domain $[-1, 1]$, so its continuity follows by once again applying Theorem 10.2 of Rockafellar (1997).

It follows immediately from 2 and 4 that $\tilde{\psi}$ is nonnegative and minimal at 0. Since 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 7.

8 follows immediately from 2.

To prove 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, 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, notice that if ϕ is not classification-calibrated, then some $\theta > 0$ has $\tilde{\psi}(\theta) = 0$, and so $\psi(\theta) = 0$. □

Proof of Theorem 4

Recall that a *subgradient* of ϕ at $\alpha \in \mathbb{R}$ is any value $m_\alpha \in \mathbb{R}$ such that $\phi(x) \geq \phi(\alpha) + m_\alpha(x - \alpha)$ for all x.

Proof. **Part 1:** Fix a convex function ϕ.

($\Longrightarrow$) Since ϕ is convex, we can find subgradients $g_1 \geq g_2$ such that, for all α,

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

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

Then we have

$$
\begin{aligned}
\eta\phi(\alpha) + (1-\eta)\phi(-\alpha) &\geq \eta(g_1\alpha + \phi(0)) + (1-\eta)(-g_2\alpha + \phi(0)) \\
&= (\eta g_1 - (1-\eta)g_2)\alpha + \phi(0) \qquad (14) \\
&= \left(\frac{1}{2}(g_1 - g_2) + (g_1 + g_2)\left(\eta - \frac{1}{2}\right)\right)\alpha + \phi(0). \qquad (15)
\end{aligned}
$$

Since 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 (15) were greater than $\phi(0)$ for every $\alpha > 0$, it would then follow that for $\eta > 1/2$, $H(\eta) \geq \phi(0) \geq H(1/2)$, which, by Lemma 6, part 3, is 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 (15) is greater than $\phi(0)$ for all $\alpha > 0$. Thus, if ϕ is classification-calibrated, we must have $g_1 = g_2$, which implies ϕ is differentiable at 0.

To see that we must also have $\phi'(0) < 0$, notice that, from (14), we have

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

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

($\Longleftarrow$) 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, notice that for some $\alpha_0 > 0$, we have

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

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

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

In particular, if $\alpha \leq \alpha_0/4$,

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

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

Part 2: Part 1 implies that ϕ is differentiable at zero and $\phi'(0) < 0$. Hence we have

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

Thus, $H^-(\eta) = \phi(0)$. The concavity of H (Lemma 6, part 2) implies $\tilde{\psi} = H^-(\eta) - H(\eta) = \phi(0) - H(\eta)$ is convex, which implies that $\psi = \tilde{\psi}$. $\square$

Proof of Lemma 7

Proof. From convexity of ϕ, we have

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

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

Proof of Lemma 8

Proof. Proceeding by contrapositive, suppose no such γ exists. Since $\phi(\alpha) \geq \mathbf{1}\left[\alpha \leq 0\right]$ on $(0,\infty)$, we must then have $\inf_{\alpha\leq 0}\phi(\alpha) = 0$. But $\phi(\alpha) = C_1(\alpha)$, hence

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

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

Proof of Lemma 9

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

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

Rearranging shows that

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

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

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

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

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

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

$$\begin{aligned} & \Pr(A_f) \le c \int_{A_f} |2\eta(x) - 1| \, dP_X(x) \\ \iff \qquad & \int_{A_f} \frac{1}{c} \, dP_X(x) \le \int_{A_f} |2\eta(x) - 1| \, dP_X(x), \end{aligned} \tag{16}$$

where $A_f = \{x : f(x)(\eta(x) - 1/2) < 0\}$. Notice that A_f ranges over all measurable subsets of $\{x : |\eta(x) - 1/2| > 0\}$, so that (16) 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 13

The proof of Lemma 13 uses techniques due to Bartlett and Mendelson (2003), which built on the work of Massart (2000b), Koltchinskii and Panchenko (2000), Mendelson (2002), Lugosi and Wegkamp (2004), Bartlett et al. (2002). 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 17. *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 $X_1, ..., X_n$ 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 $\rho > 0$,

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

Proof. (of Lemma 13)

From the condition on $\mathcal{F}$, we have

$$\begin{aligned}\Pr\left\{\exists f \in \mathcal{F} : \hat{\mathbf{E}} f \leq \alpha\epsilon,\, \mathbf{E} f \geq \epsilon\right\} &\leq \Pr\left\{\exists f \in \mathcal{F} : \hat{\mathbf{E}} f \leq \alpha\epsilon,\, \mathbf{E} f = \epsilon\right\} \\ &= \Pr\left\{\sup\left\{\mathbf{E} f - \hat{\mathbf{E}} f : f \in \mathcal{F},\, \mathbf{E} f = \epsilon\right\} \geq (1-\alpha)\epsilon\right\}.\end{aligned}$$

We bound this probability using Lemma 17, with $\rho = 1$ and $\mathcal{G} = \{\mathbf{E} f - f : f \in \mathcal{F},\, \mathbf{E} f = \epsilon\}$. This shows that

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

provided that

$$\begin{aligned} 2\mathbf{E} Z &\leq \frac{(1-\alpha)\epsilon}{3}, \\ \sqrt{\frac{c\epsilon^{\beta} K x}{n}} &\leq \frac{(1-\alpha)\epsilon}{3}, \text{ and} \\ \frac{4KBx}{n} &\leq \frac{(1-\alpha)\epsilon}{3}. \end{aligned}$$

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

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

and rearranging gives the result.

□

Proof of Lemma 14

Proof. 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

$$\begin{aligned}
\mathbf{E}g_f^2 &= \mathbf{E}\left(\ell(f(X)) - \ell(f^*(X))\right)^2 \\
&\le \mathbf{E}\left(Ld(f(X), f^*(X))\right)^2 \\
&= L^2\left(\tilde{d}(f, f^*)\right)^2. \qquad (17)
\end{aligned}$$

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 (17) gives the result. □

Proof of Lemma 15

Proof. 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

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

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 $\xi : [0, B^2] \to [0, B^r]$,

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

which follows from (the concave analog of) Lemma 5, part 2. This implies

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

□

Proof of Theorem 16

Proof. It is clear that $\mathcal{F}$ is convex and satisfies the conditions of Theorem 12. That theorem implies that, with probability at least $1 - \delta$,

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

provided that

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

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

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

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

where the expectations are over the sample $(X_1, Y_1) \dots, (X_n, Y_n)$ and the independent uniform (Rademacher) random variables $\sigma_i \in \{\pm 1\}$. The Ledoux and Talagrand (1991) contraction inequality and Lemma 14 imply

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

where

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

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 $d(a,\hat{a})\leq\epsilon$. Using Dudley's entropy integral (Dudley, 1999), Mendelson (2002) has shown the following result: Suppose that $\mathcal{F}$ is a set of $[-1,1]$-valued functions on $\mathcal{X}$, and there is a $\gamma>0$ and $0<p<2$ for which

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

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

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

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

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

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

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

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

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

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

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 18. *Suppose $\ell_\phi : \mathbb{R}^2 \to [0,\infty)$ has the form $\ell_\phi(x,y) = \phi(xy)$ for some $\phi : \mathbb{R} \to [0,\infty)$. Then*

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 zero to every pair of reals.*

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

$$x = y \iff \ell_\phi(x,y) = 0. \tag{18}$$

But we may write any $z \in (0,\infty)$ in two different ways, as $\sqrt{z}\sqrt{z}$ and, for example, $\sqrt{2z}\sqrt{(1/2)z}$. To satisfy (18) requires $\phi(z) = 0$ in the former case and $\phi(z) > 0$ in the latter, an impossibility.

This proves 1.

To prove 2, recall that a pseudometric relaxes (18) to the requirement

$$x = y \implies \ell_\phi(x, y) = 0. \tag{19}$$

Since each $z \geq 0$ has the form xy for $x = y = \sqrt{z}$, (19) 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

$$\phi(xz) \leq \phi(xy) + \phi(yz), \quad \text{for all } x, y, z \in \mathbb{R}. \tag{20}$$

Since ϕ must vanish on $[0, \infty)$, taking $y = 0$ in (20) shows that only the zero function can (and does) satisfy the constraint. $\square$

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(2):525–536.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2002), "Local Rademacher complexities," *Annals of Statistics*. To appear.

Bartlett, P. L. and Mendelson, S. (2003), "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*, pages 144–152, New York. ACM Press.

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 University Press, Cambridge, UK.

Breiman, L. (1999), "Prediction games and arcing algorithms," *Neural Computation*, 11:1493–1517.

Breiman, L. (2000), "Some infinity theory for predictor ensembles," Technical Report 577, Department of Statistics, University of California, Berkeley.

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

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 University Press, Cambridge.

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

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

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(8):2152–2152.

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(1):119–139.

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

Jiang, W. (2004), "Process consistency for Adaboost," *Annals of Statistics*, 32(1):13–29.

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

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(6):501–504.

Koltchinskii, V. and Panchenko, D. (2002), "Empirical margin distributions and bounding the generalization error of combined classifiers," *Annals of Statistics*, 30(1):1–50.

Koltchinskii, V. I. and Panchenko, D. (2000), "Rademacher processes and bounding the risk of function learning," In Giné, E., Mason, D. M., and Wellner, J. A., editors, *High Dimensional Probability II*, volume 47, pages 443–459. Birkhäuser.

Lebanon, G. and Lafferty, J. (2002), "Boosting and maximum likelihood for exponential models," In T. Dietterich, S. Becker & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14*, pages 447–454.

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

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

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(6):2118–2132.

Lin, Y. (2004), "A note on margin-based loss functions in classification," *Statistics and Probability Letters*, 68(1):73–82.

Lugosi, G. and Vayatis, N. (2004), "On the Bayes risk consistency of regularized boosting methods," *Annals of Statistics*, 32(1):30–55.

Lugosi, G. and Wegkamp, M. (2004), "Complexity regularization via localized random penalties," *Annals of Statistics*, 32(4):1679–1697.

Mammen, E. and Tsybakov, A. B. (1999), "Smooth discrimination analysis," *The Annals of Statistics*, 27(6):1808–1829.

Mannor, S. and Meir, R. (2001), "Geometric bounds for generalization in boosting," In D. Helmbold & R. Williamson (Eds.), *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory*, pages 461–472.

Mannor, S., Meir, R., and Zhang, T. (2002), "The consistency of greedy algorithms for classification," In J. Kivinen & R. Sloan (Eds.), *Proceedings of the Annual Conference on Computational Learning Theory*, pages 319–333.

Marron, J. S. and Todd, M. (2002), "Distance weighted discrimination," Technical Report No. 1339, School of Operations Research and Industrial Engineering, Cornell University.

Massart, P. (2000a), "About the constants in Talagrand's concentration inequality for empirical processes," *Annals of Probability*, 28(2):863–884.

Massart, P. (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(7):1977–1991.

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

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(2):163–175.

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

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(5):1651–1686.

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

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

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(5):1926–1940.

Shen, X., Tseng, G. C., Zhang, X., and Wong, W. H. (2003), "On psi-learning," *Journal of 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(1):128–142.

Talagrand, M. (1994), "Sharper bounds for Gaussian and empirical processes," *Annals of Probability*, 22(1):28–76.

Tsybakov, A. (2001), "Optimal aggregation of classifiers in statistical learning," Technical Report PMA-682, Université Paris VI.

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

Zhang, T. (2002), "Sequential greedy approximation for certain convex optimization problems," Technical Report RC22309, IBM T. J. Watson Research Center, Yorktown Heights.

Zhang, T. (2004), "Statistical behavior and consistency of classification methods based on convex risk minimization," *Annals of Statistics*, 32(1):56–85.

Zhang, T. and Yu, B. (2003), "Boosting with early stopping: Convergence and consistency," *Annals of Statistics*. To appear.