The Annals of Statistics
2004, Vol. 32, No. 1, 56–134

STATISTICAL BEHAVIOR AND CONSISTENCY OF CLASSIFICATION METHODS BASED ON CONVEX RISK MINIMIZATION

BY TONG ZHANG

IBM T. J. Watson Research Center

We study how closely the optimal Bayes error rate can be approximately reached using a classification algorithm that computes a classifier by minimizing a convex upper bound of the classification error function. The measurement of closeness is characterized by the loss function used in the estimation. We show that such a classification scheme can be generally regarded as a (nonmaximum-likelihood) conditional in-class probability estimate, and we use this analysis to compare various convex loss functions that have appeared in the literature. Furthermore, the theoretical insight allows us to design good loss functions with desirable properties. Another aspect of our analysis is to demonstrate the consistency of certain classification methods using convex risk minimization. This study sheds light on the good performance of some recently proposed linear classification methods including boosting and support vector machines. It also shows their limitations and suggests possible improvements.

1. Motivation. In statistical machine learning, the goal is often to predict an unobserved output value y based on an observed input vector x. This requires us to estimate a functional relationship $y \approx f(x)$ from a set of example pairs of (x, y). Usually the quality of the predictor $f(x)$ can be measured by a problem dependent loss function $\ell(f(x), y)$. In machine learning analysis, one assumes that the training data are drawn from an underlying distribution D which is not known. Our goal is to find a predictor $f(x)$ so that the expected loss of f given below is as small as possible:

$$L(f(\cdot)) = \mathbf{E}_{X,Y}\ell\big(f(X), Y\big),$$

where we use $\mathbf{E}_{X,Y}$ to denote the expectation with respect to the true underlying distribution D.

In this paper we are mainly interested in binary-classification problems, where $y \in \{\pm 1\}$. We also consider the following prediction rule: predict $y = 1$ if $f(x) \geq 0$, and predict $y = -1$ otherwise. Note that the decision rule at $f(x) = 0$ is not important in our analysis. The classification error of $f(\cdot)$ at a point (x, y) is

Received September 2001; revised August 2002.
AMS 2000 subject classifications. 62G05, 62H30, 68T05.
Key words and phrases. Classification, consistency, boosting, large margin methods, kernel methods.

given by

$$I(f(x), y) = \begin{cases} 1, & \text{if } f(x)y < 0, \\ 1, & \text{if } f(x) = 0 \text{ and } y = -1, \\ 0, & \text{otherwise.} \end{cases}$$

Given a set of training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, independently drawn from D, we may consider finding $f(x)$ in a function class C that minimizes the empirical classification error

$$\frac{1}{n}\sum_{i=1}^{n} I\big(f(X_i), Y_i\big). \tag{1}$$

This method can be regarded as a stochastic approximation to the minimization of the true classification error

$$L(f(\cdot)) = \mathbf{E}_{X,Y} I\big(f(X), Y\big). \tag{2}$$

However, due to the nonconvexity of the classification error function I, the minimization of (1) is typically NP-hard. Recently a number of methods have been proposed to alleviate this computational problem. The basic idea is to minimize a convex upper bound of the classification error function $I(p, y)$. For example, AdaBoost [7] employs the exponential loss function $\exp(-py)$ [2, 3, 14, 7], and support vector machines (SVMs) employ a loss function of the form $\max(1 - py, 0)$ [16]. In general, let ϕ be a one variable convex function. We may consider the (approximate) minimization in a function class C with respect to the following empirical risk [In this paper, we only consider an estimation scheme that is invariant with respect to the transform $(p, y) \to (-p, -y)$ for clarity. A more general formulation without this symmetry can be useful for problems that have different penalties for errors made on in-class data and errors made on out-of-class data.]:

$$\frac{1}{n}\sum_{i=1}^{n} \phi\big(f(X_i)Y_i\big), \tag{3}$$

which can be regarded as a stochastic approximation to the true risk

$$Q(f(\cdot)) = \mathbf{E}_{X,Y}\phi\big(f(X)Y\big). \tag{4}$$

Note that computationally we shall also require that the function class C be a convex function class. The resulting estimation formulation then becomes a convex optimization problem which often can be efficiently solved. However, in this paper, we focus on the analysis of the loss function ϕ. Consequently, we do not need to assume that C is a convex function class unless indicated otherwise. Furthermore, we do not assume that the minimum of the optimization problem can be achieved by a function in C. As a result, in our analysis we will only consider approximate minimization over the function class C.

In the literature, the use of a convex upper bound of the classification error function has mainly been regarded as a computational heuristic to avoid the NP-hardness of minimizing the true classification error. The empirical success of the newly proposed learning methodologies such as boosting and support vector classification implies that these methods can yield good classifiers although they do not attempt to minimize the true classification error.

So far, the most influential explanation of their success is the so-called "margin" analysis. This concept has been used to explain both SVM [16] and boosting [13]. The basic idea is that using convex risk minimization one attempts to separate the value of $f(x)$ for in-class data and out-of-class as much as possible. However, in a statistical estimation procedure, one typically encounters two types of error: one introduced by the bias of the formulation, which we call *approximation error*, and the other introduced by the variance of using finite sample size, which we call *estimation error*. The margin idea mixes the two aspects together (although the analysis itself emphasizes estimation error) so that it is not clear which aspect is the main contribution to the success of these so-called margin maximization methods. Moreover, from the margin concept, we are unable to characterize the impact of different loss functions, and we are unable to analyze the closeness of a classifier obtained from convex risk minimization to the optimal Bayes classifier.

The approach presented in this paper is motivated from a different point of view given in [7], where the authors observed that one could replace the exponential loss function $\phi(v) = \exp(-v)$ in AdaBoost by the logistic regression loss $\phi(v) = \ln(1 + \exp(-v))$ to obtain a similar procedure. One justification of using logistic regression is that logistic regression can be regarded as a maximum likelihood estimate if the conditional in-class probability $\eta(x) = P(Y = 1|X = x)$ can be expressed as $1/(1 + \exp(-f(x)))$ for some $f(x) \in C$. By the well-known consistency result for the maximum-likelihood estimate for parametric function classes (under some regularity conditions), we know that it is possible to achieve the optimal Bayes error rate using logistic regression even though we do not directly minimize the classification error. However, this point of view treats logistic regression as a very special loss function which happens to be a maximum likelihood estimate.

On the other hand, there is no practical evidence that logistic regression shows any classification performance advantage over some other convex risk minimization formulations such as support vector machines. It is thus natural to ask the question whether consistency can also be achieved using other loss functions. More importantly, we would like to analyze different convex loss functions and characterize their classification behavior. So far these issues have not been fully addressed in the literature.

We would like to mention that recently a number of authors have started to investigate issues related to what we are interested in here. To our knowledge, Breiman is the first person to consider the consistency issue for boosting type algorithms. He showed in [4] that in the infinite sample case an arcing-style greedy

approximation procedure using the AdaBoost exponential loss function converges to the Bayes classifier. Although the approximation error analysis given in that paper is closely related to the ideas presented in this paper, it is quite specialized and does not contain more quantitative results such as Theorem 2.1 of this paper. Breiman also conjectured the possibility of obtaining similar consistency results for more general loss functions. This question can be answered using our analysis. In another related work, Bühlmann and Yu investigated various theoretical issues for the least squares formulation of boosting [5] and argued that the procedure can be as effective as other methods. However, unlike this paper, they did not focus on the approximation error aspect. In fact, we will see later that using other loss functions, it is possible to improve the approximation property of the least squares method. At about the same time of this work, Lugosi and Vayatis studied the consistency issue for certain forms of boosting methods using ideas similar to what we employ here [9]. However, the framework developed in their paper is quite different from this work. In particular, they did not study certain issues discussed here, such as the interpretation of convex risk minimization as conditional probability estimate and the associated analysis of loss functions. It is also possible to demonstrate the consistency of boosting-like procedures using results of this paper. For example, see [10]. For the support vector machine formulation, Ingo Steinwart independently obtained universal consistency in [15] using a different approach but without convergence rate results such as those in Section 4.

The goal of this paper is to study the impact of a convex loss function ϕ in an estimation scheme that approximately minimizes (4). We show that similar to logistic regression, schemes with other convex loss functions may also be regarded as methods to estimate the true conditional probability $\eta(x) = P(Y = 1|X = x)$ although they use different loss-function induced distance metrics [in this respect, the distance metric for logistic regression is the KL-divergence (relative entropy)] to measure the closeness. Such loss-function induced distances also characterize the closeness of the underlying function class C to a Bayes classifier. Although an estimation scheme with a general convex function ϕ does not correspond to a maximum likelihood estimate, consistency results can still be obtained in our analysis. As concrete examples of our analysis, we are specifically interested in the following loss functions:

- Least squares: $\phi(v) = (1 - v)^2$.
- Modified least squares: $\phi(v) = \max(1 - v, 0)^2$.
- SVM: $\phi(v) = \max(1 - v, 0)$.
- Exponential: $\phi(v) = \exp(-v)$.
- Logistic regression: $\phi(v) = \ln(1 + \exp(-v))$.

The above loss functions are of general interest. All of them have been used in practical applications.

In statistical estimation, one often encounters a trade-off between the approximation error (bias) and the estimation error (variance). The approximation error is deterministic and can roughly be characterized as the error of an approximately best predictor produced by the estimation scheme using an infinite number of samples. The error can be introduced through the following two factors: (1) the function class used in the estimation is not powerful enough to represent the best predictor; (2) the bias inherent to the estimation method (e.g., it may be numerically unstable, or it may produce a suboptimal predictor even if an optimal predictor is in the function class). The estimation error is introduced from the use of a finite number of samples that do not accurately represent the underlying distribution.

In our formulation, approximation error is determined by (4). It can be informally described as follows: If $f(x) \in C$ (approximately) minimizes (4), then the approximation error is the difference between the classification error of $f(x)$ and the optimal Bayes error. Clearly, the error is determined by the underlying distribution D, the function class C and the loss function ϕ. In this paper we do not restrict the underlying distribution D or the function class C. Therefore we shall only focus on the impact of different choices of ϕ. This analysis allows us to compare various loss functions under the same conditions.

The general framework for analyzing the approximation error of (4) with a generic convex loss function ϕ is presented in Section 2. Specific consequences of this analysis on loss functions that we are interested in are given in Section 3.

In Section 4, we consider the estimation error resulting from use of a finite sample size. Specifically, we show that some recently popularized kernel classification methods are universal in the sense that they can produce predictors with classification error that approaches the optimal Bayes error in probability for any underlying distribution when the sample size n goes to infinity. In order to achieve this, we need to choose a function class C so that the resulting approximation error using a convex risk minimization scheme with risk defined in (4) is zero for any conditional probability density $\eta(x) = P(Y = 1|X = x)$. Due to the tremendous expressive power of such a function class, we cannot use C directly in a finite-sample estimation method that minimizes the empirical risk in (3). Instead, we consider a complexity regularization approach in a kernel method. The underlying idea is to decompose the function class C as $C = \cup C_n$ ($C_1 \subset C_2 \subset \cdots \subset C$) such that the estimation error of minimizing (4) within C_n approaches zero as $n \to \infty$. Universality of kernel methods can then be established by appropriately choosing the complexity regularization parameter for each sample size.

For convenience, throughout the paper we assume that all quantities appearing in the discussion are measurable whenever necessary.

2. Approximation error under convex risk minimization. In this section, we study the relationship between the classification error $L(f(\cdot))$ of a predictor f and the quantity $Q(f(\cdot))$ defined in (4).

We shall rewrite (4) as

$$Q(f(\cdot)) = \mathbf{E}_X\big[\eta(X)\phi(f(X)) + \big(1-\eta(X)\big)\phi\big(-f(X)\big)\big], \tag{5}$$

where we use $\eta(x)$ to denote the conditional probability $P(Y=1|X=x)$. $\mathbf{E}_X$ denotes the expectation over input data X. In the following, we will show that by minimizing $Q(\cdot)$, we also indirectly minimize the classification error.

The classification error of a predictor $f(x)$ can be written as

$$L(f(\cdot)) = \mathbf{E}_{f(X)\geq 0}\big(1-\eta(X)\big) + \mathbf{E}_{f(X)<0}\eta(X). \tag{6}$$

Note that we have used $\mathbf{E}_{f(X)\geq 0}$ to denote the expectation in the region $f(X)\geq 0$. In other words, $\mathbf{E}_{f(X)\geq 0}\, s(X) = \mathbf{E}_X\,[s(X)\mathbf{1}(f(X)\geq 0)]$, where $\mathbf{1}(\cdot)$ denotes the set indicator function. $\mathbf{E}_{f(x)<0}$ is similarly defined.

For convenience, we also introduce the notation:

$$Q(\eta, f) = \eta\phi(f) + (1-\eta)\phi(-f), \tag{7}$$

where we assume that $\eta\in[0,1]$.

Let $\mathbb{R}^*$ denote the extended real line ($\mathbb{R}^* = \mathbb{R}\cup\{-\infty,+\infty\}$). We extend a convex function $g:\mathbb{R}\to\mathbb{R}$ to a function $g:\mathbb{R}^*\to\mathbb{R}^*$ by defining $g(\infty) = \lim_{x\to\infty} g(x)$ and $g(-\infty) = \lim_{x\to-\infty} g(x)$. The extension is only for notational convenience. It ensures that the optimal minimizer $f_\phi^*(\eta)$ given below is well defined at $\eta = 0$ or 1 for certain loss functions.

DEFINITION 2.1. We define the function $f_\phi^*(\eta):[0,1]\to\mathbb{R}^*$ as

$$f_\phi^*(\eta) = \arg\min_{f\in\mathbb{R}^*} Q(\eta, f)$$

and

$$Q^*(\eta) = \inf_{f\in\mathbb{R}} Q(\eta, f) = Q\big(\eta, f_\phi^*(\eta)\big).$$

Note that $f_\phi^*(\eta)$ may not be uniquely determined. In such case, we may choose any $f_\phi^*(\eta)$ that minimizes the right-hand side. By symmetry, we have $Q(\eta,p) = Q(1-\eta,-p)$. This implies that if $f_\phi^*(\eta)$ minimizes the right-hand side, then $-f_\phi^*(1-\eta)$ also minimizes the right-hand side. Therefore we may choose f_ϕ^* such that $f_\phi^*(1-\eta) = -f_\phi^*(\eta)$. In the following, we always assume that this condition holds. In particular, it implies that $f_\phi^*(0.5) = 0$.

Clearly $f_\phi^*(\eta(x))$ minimizes $Q(f(x))$ in (5) among all measurable functions $f(x)$ by the construction of f_ϕ^*. In our analysis, it is convenient to introduce the following two quantities which are always nonnegative:

$$\Delta Q(\eta, f) = Q(\eta, f) - Q\big(\eta, f_\phi^*(\eta)\big) = Q(\eta, f) - Q^*(\eta),$$

$$\Delta Q(f(\cdot)) = Q(f(\cdot)) - Q\big(f_\phi^*(\eta(\cdot))\big) = \mathbf{E}_X \Delta Q\big(\eta(X), f(X)\big).$$

The second quantity measures the closeness of the risk of a function $f(\cdot)$ defined in (5) to the optimal risk $Q(f_\phi^*(\eta(\cdot)))$.

Both f_ϕ^* and Q^* are easy to compute given a convex loss function ϕ. We list the following results for formulations that we are specially interested in:

- Least squares: $f_\phi^*(\eta) = 2\eta - 1$; $Q^*(\eta) = 4\eta(1-\eta)$.
- Modified least squares: $f_\phi^*(\eta) = 2\eta - 1$; $Q^*(\eta) = 4\eta(1-\eta)$.
- SVM: $f_\phi^*(\eta) = \operatorname{sign}(2\eta - 1)$; $Q^*(\eta) = 1 - |2\eta - 1|$.
- Exponential: $f_\phi^*(\eta) = \frac{1}{2}\ln\frac{\eta}{1-\eta}$; $Q^*(\eta) = 2\sqrt{\eta(1-\eta)}$.
- Logistic regression: $f_\phi^*(\eta) = \ln\frac{\eta}{1-\eta}$; $Q^*(\eta) = -\eta\ln\eta - (1-\eta)\ln(1-\eta)$.

We are now ready to provide the intuition on why it is possible to obtain a Bayes classifier through the minimization of (4) which by itself only leads to an upper bound of the true classification error. Observe that for all of the above examples of ϕ, $f_\phi^*(\eta) > 0$ when $\eta > 0.5$. This implies if we let $f(x) = f_\phi^*(\eta(x))$ which minimizes (4), then the corresponding decision rule leads to the Bayes classifier. Since $f_\phi^*(\eta(x))$ minimizes (4) among all measurable functions $f(x)$, if the function class C contains $f_\phi^*(\eta(x))$, we are able to find $f(x) \in C$ that achieves the Bayes classification error.

Although the above intuition is quite simple and clear, there are still a number of technical issues that need to be resolved. First, $f_\phi^*(\eta(x))$ may not be the unique minimizer of (5). This means that even if $f_\phi^*(\eta(x)) \in C$, it is still unclear whether any minimizer of (5) in C leads to the Bayes classifier.

Additionally, we do not require $f_\phi^*(\eta(x))$ to be in C; and in general, we do not require that the minimum of (5) in C can be achieved by an element in C. In this case, it is necessary to bound the classification error of $f(x)$ in terms of $\Delta Q(f(\cdot))$, which is given by the following theorem.

THEOREM 2.1. *Assume $f_\phi^*(\eta) > 0$ when $\eta > 0.5$. Assume there exist $c > 0$ and $s \geq 1$ such that for all $\eta \in [0,1]$,*

$$|0.5 - \eta|^s \leq c^s \Delta Q(\eta, 0).$$

Then for any measurable function $f(x)$

$$L(f(\cdot)) \leq L^* + 2c\Delta Q(f(\cdot))^{1/s},$$

where L^* is the optimal Bayes error $L^* = L(2\eta(\cdot) - 1)$.

PROOF. Using (6), it is easy to verify that

$$\begin{aligned} &L(f(\cdot)) - L\big(2\eta(\cdot) - 1\big) \\ &\quad = \mathbf{E}_{\eta(X)\geq 0.5, f(X)<0}\big(2\eta(X) - 1\big) + \mathbf{E}_{\eta(X)<0.5, f(X)\geq 0}\big(1 - 2\eta(X)\big) \\ &\quad \leq \mathbf{E}_{(2\eta(X)-1)f(X)\leq 0}|2\eta(X) - 1| \\ &\quad \leq 2\big(\mathbf{E}_{(2\eta(X)-1)f(X)\leq 0}|\eta(X) - 0.5|^s\big)^{1/s}. \end{aligned} \tag{8}$$

The last inequality follows from the Jensen's inequality. Using the assumption of the theorem, we obtain

$$L(f(\cdot)) - L\big(2\eta(\cdot) - 1\big) \le 2c\big(\mathbf{E}_{(2\eta(X)-1)f(X)\le 0}\Delta Q\big(\eta(X), 0\big)\big)^{1/s}. \tag{9}$$

If we can further show that $(2\eta(x) - 1)f(x) \le 0$ implies $Q(\eta(x), 0) \le Q(\eta(x), f(x))$, then

$$\mathbf{E}_{(2\eta(X)-1)f(X)\le 0}\Delta Q\big(\eta(X), 0\big) \le \mathbf{E}_X \Delta Q\big(\eta(X), f(X)\big) = \Delta Q(f(\cdot)).$$

Combining this inequality with (9), we obtain the theorem. Therefore, in the following we only need to prove the fact that $(2\eta - 1)p \le 0$ implies $Q(\eta, 0) \le Q(\eta, p)$. To see this, we consider the following three cases:

- $\eta > 0.5$: By assumption, we have $f_\phi^*(\eta) > 0$. Now $(2\eta - 1)p \le 0$ implies $p \le 0$. Using $0 \in [p, f_\phi^*(\eta)]$ and the convexity of $Q(\eta, p)$ with respect to p, we obtain $Q(\eta, 0) \le \max(Q(\eta, p), Q(\eta, f_\phi^*(\eta))) = Q(\eta, p)$.
- $\eta < 0.5$: Note that we require $f_\phi^*(\eta) = -f_\phi^*(1 - \eta)$. Therefore $f_\phi^*(\eta) < 0$. Since $(2\eta - 1)p \le 0$ implies that $p \ge 0$, we have $0 \in [f_\phi^*(\eta), p]$, which implies that $Q(\eta, 0) \le \max(Q(\eta, p), Q(\eta, f_\phi^*(\eta))) = Q(\eta, p)$.
- $\eta = 0.5$: Note that $f_\phi^*(\eta) = 0$ which implies $Q(\eta, 0) \le Q(\eta, p)$ for all values of p.

This completes the proof of the theorem. □

We have made no special effort to ensure that the bound in the above theorem is as tight as possible. In many practical applications, the conditional probability $\eta(x)$ is close to 0 or 1. In this case, if the assumption of the theorem holds with $s > 1$, then the assumption will also hold with $s = 1$ but with a larger constant c. Using this idea, refined bounds can be easily obtained.

COROLLARY 2.1. *Under the assumptions of Theorem* 2.1, *we have*

$$L(f(\cdot)) \le L^* + 2c \inf_{\delta>0}\bigg[\big(\mathbf{E}_{|\eta(X)-0.5|<\delta}\Delta Q\big(\eta(X), f(X)\big)\big)^{1/s} + \left(\frac{c}{\delta}\right)^{s-1}\Delta Q(f(\cdot))\bigg].$$

PROOF. If $|\eta(x) - 0.5| \ge \delta$, we have

$$|\eta(x) - 0.5| \le \delta^{1-s}c^s \Delta Q(\eta, 0). \tag{10}$$

Now we can decompose (8) as

$$\mathbf{E}_{(2\eta(X)-1)f(X)\le 0, |\eta(X)-0.5|<\delta}|2\eta(X) - 1| + \mathbf{E}_{(2\eta(X)-1)f(X)\le 0, |\eta(X)-0.5|\ge\delta}|2\eta(X) - 1|.$$

We can bound the first term by Jensen's inequality as in the proof of Theorem 2.1, and the second term using (10). The rest of the proof follows the same argument as that of Theorem 2.1. □

Theorem 2.1 implies that if we obtain ϕ by approximately minimizing (5) so that $\Delta Q(f(\cdot))$ is small, then the classification error rate of $f(\cdot)$ is close to that of the Bayes error rate $L^* = L(f_\phi^*(\eta(\cdot)))$. In particular, if we can find a sequence of predictors $f_k(\cdot) \in C$ such that $\Delta Q(f_k(\cdot)) \to 0$, then we are able to achieve a classification error rate arbitrarily close to that of the Bayes error rate in C by approximately minimizing (5).

Observe that in Theorem 2.1, both $\Delta Q(\eta, 0)$ and $\Delta Q(f(\cdot)) = \mathbf{E}_X \Delta Q(\eta(X), f(X))$ rely on the quantity $\Delta Q(\eta, f) = Q(\eta, f) - Q^*(\eta)$. It is thus worthwhile to further analyze it. In order to do so, we need the following definition of Bregman divergence which originally appeared in [1]:

DEFINITION 2.2. For a convex function ϕ, we define its Bregman divergence as

$$d_\phi(f_1, f_2) = \phi(f_2) - \phi(f_1) - \phi'(f_1)(f_2 - f_1).$$

For a concave function g, the Bregman divergence is defined as

$$d_g(\eta_1, \eta_2) = d_{-g}(\eta_1, \eta_2).$$

Note that in the above definition, $\phi'(f)$ in general denotes a subgradient of a convex function $\phi(f)$ at f. A subgradient p^* of a convex function $\phi(f)$ at p is a value such that $\phi(q) \geq \phi(p) + p^*(q - p)$ for all q (see [11], Section 23). Clearly, by definition, the Bregman divergence is always nonnegative. However, in general, a subgradient of a convex function at a point may not always exist, and even when it exists it may not be unique. To avoid such difficulties, in this paper we only use Bregman divergence for differentiable convex functions except in Section 4.3. In this case, subgradient becomes derivative which is uniquely defined.

The following lemma shows that Q^* is concave, which is useful in our later discussion.

LEMMA 2.1. *$Q^*(\eta)$ is a concave function of η ($\eta \in [0, 1]$).*

PROOF. Consider $0 \leq \eta_1 \leq \eta_2 \leq 1$ and $t \in [0, 1]$. Let $\eta = t\eta_1 + (1 - t)\eta_2$.

$$\begin{aligned} Q^*(\eta) &= \eta\phi\big(f_\phi^*(\eta)\big) + (1 - \eta)\phi\big(-f_\phi^*(\eta)\big) \\ &= t\big(\eta_1\phi\big(f_\phi^*(\eta)\big) + (1 - \eta_1)\phi\big(-f_\phi^*(\eta)\big)\big) \\ &\quad + (1 - t)\big(\eta_2\phi\big(f_\phi^*(\eta)\big) + (1 - \eta_2)\phi\big(-f_\phi^*(\eta)\big)\big) \\ &\geq t\big(\eta_1\phi\big(f_\phi^*(\eta_1)\big) + (1 - \eta_1)\phi\big(-f_\phi^*(\eta_1)\big)\big) \end{aligned}$$

$$+ (1-t)\big(\eta_2\phi\big(f_\phi^*(\eta_2)\big) + (1-\eta_2)\phi\big(-f_\phi^*(\eta_2)\big)\big)$$
$$= tQ^*(\eta_1) + (1-t)Q^*(\eta_2),$$

where the inequality above follows from the definition of f_ϕ^*. $\square$

For any convex function ϕ, by the definition of $f_\phi^*(\cdot)$, we have the first-order condition

$$\eta\phi'\big(f_\phi^*(\eta)\big) - (1-\eta)\phi'\big(-f_\phi^*(\eta)\big) = 0, \tag{11}$$

where $\phi'(p)$ denotes a subgradient of ϕ at p. This implies

$$\begin{aligned}\Delta Q(\eta, p) &= \eta\big[\phi(p) - \phi\big(f_\phi^*(\eta)\big)\big] + (1-\eta)\big[\phi(-p) - \phi\big(-f_\phi^*(\eta)\big)\big] \\ &= \eta\big[\phi(p) - \phi\big(f_\phi^*(\eta)\big) - \phi'\big(f_\phi^*(\eta)\big)\big(p - f_\phi^*(\eta)\big)\big] \\ &\quad + (1-\eta)\big[\phi(-p) - \phi\big(-f_\phi^*(\eta)\big) - \phi'\big(-f_\phi^*(\eta)\big)\big(f_\phi^*(\eta) - p\big)\big] \\ &= \eta d_\phi\big(f_\phi^*(\eta), p\big) + (1-\eta)d_\phi\big(-f_\phi^*(\eta), -p\big).\end{aligned}$$

In the above, $\Delta Q(\eta, p)$ is expressed using the Bregman divergence of ϕ. We may also express $\Delta Q(\eta, p)$ using the Bregman divergence of Q^*. Assume that ϕ and $f_\phi^*(\eta)$ are differentiable. We have from (11) that for all $p = f_\phi^*(\bar\eta)$, where $\bar\eta \in [0, 1]$,

$$\begin{aligned}Q^{*\prime}(\bar\eta) &= \frac{d}{d\bar\eta}\big[\bar\eta\phi\big(f_\phi^*(\bar\eta)\big) + (1-\bar\eta)\phi\big(-f_\phi^*(\bar\eta)\big)\big] \\ &= \phi\big(f_\phi^*(\bar\eta)\big) - \phi\big(-f_\phi^*(\bar\eta)\big) + \big[\bar\eta\phi'\big(f_\phi^*(\bar\eta)\big) - (1-\bar\eta)\phi'\big(-f_\phi^*(\bar\eta)\big)\big]f_\phi^{*\prime}(\bar\eta) \\ &= \phi(p) - \phi(-p).\end{aligned}$$

We thus have

$$\begin{aligned}\Delta Q(\eta, p) &= \eta\phi(p) + (1-\eta)\phi(-p) - Q^*(\eta) \\ &= (\eta - \bar\eta)\big(\phi(p) - \phi(-p)\big) + \bar\eta\phi(p) + (1-\bar\eta)\phi(-p) - Q^*(\eta) \\ &= (\eta - \bar\eta)Q^{*\prime}(\bar\eta) + Q^*(\bar\eta) - Q^*(\eta) \\ &= d_{Q^*}(\bar\eta, \eta).\end{aligned}$$

We summarize the above derivations in the following theorem:

THEOREM 2.2. *If ϕ is differentiable, then Bregman divergence d_ϕ is uniquely defined. We have the equality*:

$$\Delta Q(\eta, p) = \eta d_\phi\big(f_\phi^*(\eta), p\big) + (1-\eta)d_\phi\big(-f_\phi^*(\eta), -p\big).$$

If furthermore f_ϕ^* is differentiable, then Q^* is also differentiable. Assume that $p = f_\phi^*(\bar\eta)$. Then

$$\Delta Q(\eta, p) = d_{Q^*}(\bar\eta, \eta).$$

The above results are useful for calculating quantities in Theorem 2.1. If we assume that f_ϕ^* is invertible, then its inverse function $f_\phi^{*-1}(f(x))$ can be regarded as a conditional probability estimate. From $\Delta Q(\eta, p) = d_{Q^*}(\bar{\eta}, \eta)$, we obtain

$$\Delta Q(f(\cdot)) = \mathbf{E}_X \Delta Q\big(\eta(X), f(X)\big) = \mathbf{E}_X d_{Q^*}\big(f_\phi^{*-1}(f(X)), \eta(X)\big).$$

It is clear that by minimizing $Q(f(\cdot))$ as in (5), we are effectively minimizing the expected distance of the conditional probability $f_\phi^{*-1}(f(x))$ associated with $f(x)$ and the true conditional in-class probability $\eta(x)$. The distance is specified by the Bregman divergence of Q^*. Intuitively, we try to estimate the true in-class conditional probability even though the underlying method may not correspond to a maximum likelihood estimate. The conditional probability information resulting from a convex risk minimization method is very useful in applications. It can provide much richer information than the simple binary classification error bound given in Theorem 2.1.

We would like to mention that different formulations can share the same form of f_ϕ^* which determines the corresponding probability model. However, the function f_ϕ^* does not fully characterize the behavior of the formulation since the loss function induced distance $\Delta Q(\eta, p)$ in Theorem 2.2 can still behave differently.

3. Examples of approximation error analysis. In this section, we apply the general framework of approximation error analysis outlined in Section 2 to those loss functions that we are specially interested in. From Theorem 2.1, we can obtain the following generic classification error bound:

COROLLARY 3.1. *Under the assumptions of Theorem* 2.1, *let*

$$\varepsilon_1 = \inf_{f(\cdot)\in C} \mathbf{E}_X \Delta Q\big(\eta(X), f(X)\big).$$

Assume we find $\hat{f}(\cdot)$ *that approximately minimizes* (4) *as*

$$Q\big(\hat{f}(\cdot)\big) \le \inf_{f\in C} Q(f(\cdot)) + \varepsilon_2.$$

Then

$$\mathbf{E}_X \Delta Q\big(\eta(X), f(X)\big) \le \varepsilon_1 + \varepsilon_2$$

and

$$L\big(\hat{f}(\cdot)\big) \le L^* + 2c(\varepsilon_1 + \varepsilon_2)^{1/s}.$$

In the following, we are interested in estimating c, s and $\Delta Q(\eta, f)$ for each formulation.

3.1. *Least squares.* We consider the case $\phi(v) = (1 - v)^2$. $f_\phi^*(\eta) = 2\eta - 1$. The subgradient of ϕ is uniquely determined as $\phi'(v) = 2(v - 1)$. The Bregman divergence of ϕ is $d_\phi(p_1, p_2) = (p_2 - p_1)^2$. We thus obtain from Theorem 2.2 that

$$\Delta Q(\eta, p) = (2\eta - 1 - p)^2.$$

Specifically, we have

$$|\eta - 0.5|^2 = 0.5^2 \Delta Q(\eta, 0).$$

This implies that we can choose $c = 0.5$ and $s = 2$ in Corollary 3.1.

Therefore by using the least squares method for the classification problem, we attempt to minimize the expected squared difference of the predictor and the conditional in-class probability $\eta(x)$. This means that $(f(x) + 1)/2$ (truncated to $[0, 1]$) can be regarded as an approximation to the conditional in-class probability. Therefore least squares classification can be regarded as a nonmaximum likelihood conditional density estimation method.

3.2. *Modified least squares.* We consider the case $\phi(v) = \max(0, 1 - v)^2$. $f_\phi^*(\eta) = 2\eta - 1$. The subgradient of ϕ is uniquely determined as $\phi'(v) = -2\max(0, 1 - v)$. For $p_1 \in [-1, 1]$, the Bregman divergence of ϕ is

$$d_\phi(p_1, p_2) = (p_2 - p_1)^2 - \max(0, p_2 - 1)^2.$$

We thus obtain from Theorem 2.2 that

$$\Delta Q(\eta, p) = (2\eta - 1 - p)^2 - \eta \max(0, p - 1)^2 - (1 - \eta)\min(0, p + 1)^2.$$

Specifically, we have

$$|\eta - 0.5|^2 \le 0.5^2 \Delta Q(\eta, 0).$$

This implies that we can choose $c = 0.5$ and $s = 2$ in Corollary 3.1.

Since ΔQ is rather complicated, it is useful to give simplified upper and lower bounds. Such bounds are given in the lemma below. Using the lower bound, we can see that

$$\mathbf{E}_X\big(2\eta(X) - 1 - T(f(X))\big)^2 \le \varepsilon_1 + \varepsilon_2,$$

where $T(p)$ truncates p to the interval $[-1, 1]$. Since ε_1 in the modified least squares case is always smaller than the corresponding ε_1 for least squares, the approximation bound for modified least squares is always better than that of least squares. This implies that in general the modified least squares method can perform better than least squares for classification problems.

LEMMA 3.1. *Let* $T(p) = \min(\max(p, -1), 1)$. *Then for all* p *and* $\eta \in [0, 1]$ *we have*

$$\Delta Q(\eta, p) = \big(2\eta - 1 - T(p)\big)^2 + |2\eta - 1 - T(p)||p - T(p)|\frac{|p| + 3}{2}.$$

This implies the bounds:

$$\big(2\eta - 1 - T(p)\big)^2 \le \Delta Q(\eta, p) \le \big(2\eta - 1 - T(p)\big)^2 + |2\eta - 1 - T(p)|\frac{(|p|+1)^2}{2}.$$

PROOF. Note that

$$\begin{aligned}\Delta Q(\eta, p) &= \Delta Q\big(\eta, T(p)\big) + Q(\eta, p) - Q\big(\eta, T(p)\big)\\ &= \big(2\eta - 1 - T(p)\big)^2 + \eta\big(\max(0, 1-p)^2 - \max\big(0, 1 - T(p)\big)^2\big)\\ &\quad + (1-\eta)\big(\max(0, 1+p)^2 - \max\big(0, 1+T(p)\big)^2\big).\end{aligned}$$

We consider the following three cases:

- $p \in [-1, 1]$: Since $p = T(p)$, the lemma holds.
- $p > 1$: Note that $T(p) = 1$. We have

$$\begin{aligned}\Delta Q(\eta, p) &= \big(2\eta - 1 - T(p)\big)^2 + (1-\eta)\big((1+p)^2 - \big(1+T(p)\big)^2\big)\\ &= \big(2\eta - 1 - T(p)\big)^2 + |2\eta - 1 - T(p)|(p-1)\frac{p+3}{2}.\end{aligned}$$

- $p < -1$: Note that $T(p) = -1$. We have

$$\begin{aligned}\Delta Q(\eta, p) &= \big(2\eta - 1 - T(p)\big)^2 + \eta\big((1-p)^2 - \big(1-T(p)\big)^2\big)\\ &= \big(2\eta - 1 - T(p)\big)^2 + |2\eta - 1 - T(p)|(-p-1)\frac{-p+3}{2}.\end{aligned}$$

Combining the above three cases, we obtain the lemma. □

Using the lemma, we obtain the following result:

COROLLARY 3.2. *Consider* $\phi(v) = \max(0, 1-v)^2$. *Let* $T(p) = \min(\max(p, -1), 1)$ *and*

$$D\big(\eta(\cdot), f(\cdot)\big) = \mathbf{E}_X\big(2\eta(X) - 1 - T(f(X))\big)^2.$$

Let

$$\varepsilon_1 = \inf_{f \in C}\big[D\big(\eta(\cdot), f(\cdot)\big) + 0.5D^{1/2}\big(\eta(\cdot), f(\cdot)\big)\mathbf{E}_X^{1/2}\big(f(X)+1\big)^4\big].$$

Assume we find $\hat{f}(\cdot)$ *that approximately minimizes* (4) *as*

$$Q\big(\hat{f}(\cdot)\big) \le \inf_{f \in C} Q(f(\cdot)) + \varepsilon_2.$$

Then

$$\mathbf{E}_X\big(2\eta(X) - 1 - T(f(X))\big)^2 \le \varepsilon_1 + \varepsilon_2$$

and

$$L\big(\hat{f}(\cdot)\big) \le L^* + (\varepsilon_1 + \varepsilon_2)^{1/2}.$$

PROOF. Using the Schwarz inequality, we obtain

$$D^{1/2}(\eta(\cdot), f(\cdot))\mathbf{E}_X^{1/2}\big(f(X)+1\big)^4 \geq \mathbf{E}_X\big[|2\eta(X)-1-T(f(X))|\big(|f(X)|+1\big)^2\big].$$

Now apply Lemma 3.1. We obtain

$$\inf_{f\in C}\mathbf{E}_X\Delta Q\big(\eta(X), f(X)\big) \leq \varepsilon_1.$$

This proves the corollary. □

If $\inf_{f\in C}\mathbf{E}_X(2\eta(X)-1-T(f(X)))^2\mathbf{E}_X(f(X)+1)^4 = 0$, then we can find $f\in C$ by approximately minimizing (4) to obtain classifiers that have classification error arbitrarily close to that of the Bayes error. Roughly speaking, if $\eta(x)$ can be approximated as $(T(f(x))+1)/2$ for $f(x)\in C$, then the modified least squares method gives a good estimate of the conditional in-class probability $\eta(x)$. Note that in the case of the least squares method, we require that $\eta(x)$ be well approximated by $(f(x)+1)/2$ for $f(x)\in C$ in order to obtain a good conditional in-class probability estimate. This shows that in general, the modified least squares method is superior to the standard least squares method for classification problems.

3.3. *SVM.* We consider the case $\phi(v)=\max(0,1-v)$. $f_\phi^*(\eta)=\operatorname{sign}(2\eta-1)$. Since the subgradient of ϕ at 1 is not uniquely defined, we will directly compute $\Delta Q(f(\cdot))$ rather than use Theorem 2.2. It is easy to obtain

$$\begin{aligned}\Delta Q(\eta,p) &= \eta\big(\phi(p)-\phi\big(f_\phi^*(\eta)\big)\big)+(1-\eta)\big(\phi(-p)-\phi\big(-f_\phi^*(\eta)\big)\big)\\ &= \eta\max(0,1-p)+(1-\eta)\max(0,1+p)-1+|2\eta-1|.\end{aligned}$$

This implies that

$$\Delta Q(\eta,0)=\eta+(1-\eta)-1+|2\eta-1|=|2\eta-1|.$$

We may set $s=1$ and $c=0.5$ in Corollary 3.1.

In order to compare the SVM loss with other losses, we need to rewrite $\Delta Q(\eta,p)$ in a more intuitive form. In fact, it is not hard to check that

$$\Delta Q(\eta,p)=\begin{cases}(p-1)(1-\eta)+\big(1-\operatorname{sign}(2\eta-1)\big)|2\eta-1|, & p\geq 1,\\ |p-\operatorname{sign}(2\eta-1)||2\eta-1|, & p\in[-1,1],\\ |p+1|\eta+\big(1+\operatorname{sign}(2\eta-1)\big)|2\eta-1|, & p\leq -1.\end{cases}$$

From this formulation, we observe that in order for $\varepsilon_1=\mathbf{E}_X\Delta Q(\eta(X),f(X))$ to be small, the following has to be satisfied on average:

- if $\eta(x)$ is close to 0.5: $f(x)-T(f(x))$ is small (where $T(p)$ denotes the truncation of p to $[-1,1]$, which has been defined earlier);

- otherwise:
  - if $|f(x)| \leq 1$: $|f(x) - \operatorname{sign}(2\eta(x) - 1)|$ is small,
  - otherwise: $|f(x) - \operatorname{sign}(2\eta(x) - 1)||2\eta(x) - 1 - \operatorname{sign}(f(x))|$ is small.

Roughly speaking, for $\eta(x)$ not close to 0.5, we require $f(x) \approx \operatorname{sign}(2\eta(x) - 1)$ but allow $f(x) > 1$ when $\eta(x) \approx 1$ and allow $f(x) < -1$ when $\eta(x) \approx 0$. The latter two conditions correspond to the margin argument which motivated SVM [16] and has also been used to explain the effectiveness of boosting [13]. Although we can see that it naturally comes out of the approximation error analysis, the original concept was used to bound the estimation error with an emphasis on separable problems. In fact, one can observe from our analysis that the margin idea is mostly useful for problems that are nearly separable [$\eta(x)$ is close to 0 or 1]. It is not very useful if $\eta(x)(1-\eta(x))$ is not small since if $f(x) \in C$ is close to the optimal Bayes classifier, then $f(x)$ must satisfy the condition $f(x) \approx \operatorname{sign}(2\eta(x) - 1)$. Therefore our analysis not only provides a statistical justification for the margin concept for nearly separable problems, but also shows its limitation in the general case.

Our analysis also shows a significant disadvantage of the SVM formulation for problems that are not nearly separable. Note that if an SVM performs well, then it computes a predictor $\hat{f}(x)$ that has a small value $\Delta Q(\hat{f}(\cdot))$. For a point x such that $\eta(x)(1 - \eta(x))$ is not close to zero, we have $\hat{f}(x) \approx \operatorname{sign}(2\eta(x) - 1)$. That is, $\hat{f}(x)$ is clustered at ± 1. It gives similar values even though the corresponding conditional in-class probability $\eta(x)$ can be very different. This implies that the predictor computed by SVM does not carry any reliable probability information. By looking at the output $\hat{f}(x)$ at any given point x, it is difficult to tell how confident the prediction is. However, such confidence information is often extremely valuable in practical applications. In this respect, the modified least squares loss has a significant advantage over the standard SVM loss.

3.4. *Exponential loss.* We consider the case $\phi(v) = \exp(-v)$. Note that $f_\phi^*(\eta) = \frac{1}{2}\ln\frac{\eta}{1-\eta}$ and $Q^*(\eta) = 2\sqrt{\eta(1-\eta)}$. $f_\phi^*(\eta)$ is invertible,

$$f_\phi^{*-1}(p) = \frac{1}{1 + e^{-2p}}.$$

From Theorem 2.2 and remarks thereafter, we know that $1/(1 + e^{-2f(x)})$ can be regarded as an approximation to the true conditional probability function $\eta(x)$, and its closeness to $\eta(x)$ is measured by the expected Bregman divergence of $2\sqrt{\eta(1-\eta)}$,

$$\Delta Q(\eta, p) = d_{Q^*}(\bar{\eta}, \eta) = (\eta - \bar{\eta})(e^{-p} - e^{p}) + 2\sqrt{\bar{\eta}(1-\bar{\eta})} - 2\sqrt{\eta(1-\eta)},$$

where $\bar{\eta} = 1/(1 + e^{-2p})$. This implies that

$$\Delta Q(\eta, 0) = 1 - 2\sqrt{\eta(1-\eta)} \geq 2(\eta - 0.5)^2.$$

Therefore we may let $s = 2$ and $c = 2^{-1/2}$ in Corollary 3.1.

It is useful to obtain a lower bound of $\Delta Q(\eta, p)$ which is more intuitive than the Bregman divergence of $2\sqrt{\eta(1-\eta)}$. By using a Taylor expansion, we know that $\exists\, \eta'$ between $\bar{\eta}$ and η so that

$$\begin{aligned}\Delta Q(\eta, p) &= d_{Q^*}(\bar{\eta}, \eta) \\ &= -\frac{1}{2} Q^{*\prime\prime}(\eta')(\eta - \bar{\eta})^2 \\ &= \frac{1}{4}\big(\eta'(1-\eta')\big)^{-3/2}(\eta - \bar{\eta})^2 \geq 2\left(\eta - \frac{1}{1+e^{-2p}}\right)^2.\end{aligned}$$

This lower bound implies that if we can find $\hat{f}(x) \in C$ such that $\Delta Q(\hat{f}(\cdot))$ is small, then the expected squared difference between the estimated conditional probability $1/(1+e^{-2f(x)})$ and the true conditional probability $\eta(x)$ is also small. Although this result is similar to that of (modified) least squares, for practical problems such that $\eta(x)$ is close to 0 or 1, the quantity $\Delta Q(\hat{f}(\cdot))$ with the exponential loss can be significantly larger. To see this, we consider the following scenario: $0.5 \leq \eta \leq \bar{\eta} \leq 1$ or $0.5 \geq \eta \geq \bar{\eta} \geq 0$. Under this assumption we can obtain a lower bound of $\Delta Q(\eta, p)$ as

$$\begin{aligned}&\Delta Q(\eta, p) \\ &\quad = (\eta - \bar{\eta})\big(\exp(-p) - \exp(p)\big) + 2\sqrt{\bar{\eta}(1-\bar{\eta})} - 2\sqrt{\eta(1-\eta)} \\ &\quad \geq |\eta - \bar{\eta}|\big(\exp(|p|) - 1\big) \\ &\qquad - 2\sqrt{|\sqrt{\bar{\eta}(1-\bar{\eta})} - \sqrt{\eta(1-\eta)}|\big(\sqrt{\bar{\eta}(1-\bar{\eta})} + \sqrt{\eta(1-\eta)}\big)} \\ &\quad = |\eta - \bar{\eta}|\big(\exp(|p|) - 1\big) - 2\sqrt{|\bar{\eta}(1-\bar{\eta}) - \eta(1-\eta)|} \\ &\quad \geq |\eta - \bar{\eta}|\big(\exp(|p|) - 1\big) - 2\sqrt{|\bar{\eta} - \eta|}.\end{aligned}$$

Moreover, a similar upper bound on $\Delta Q(\eta, p)$ can be obtained without the assumptions on η and $\bar{\eta}$:

$$\begin{aligned}&\Delta Q(\eta, p) \\ &\quad = (\eta - \bar{\eta})\big(\exp(-p) - \exp(p)\big) + 2\sqrt{\bar{\eta}(1-\bar{\eta})} - 2\sqrt{\eta(1-\eta)} \\ &\quad \leq |\eta - \bar{\eta}| \exp(|p|) \\ &\qquad + 2\sqrt{|\sqrt{\bar{\eta}(1-\bar{\eta})} - \sqrt{\eta(1-\eta)}|\big(\sqrt{\bar{\eta}(1-\bar{\eta})} + \sqrt{\eta(1-\eta)}\big)} \\ &\quad = |\eta - \bar{\eta}| \exp(|p|) + 2\sqrt{|\bar{\eta}(1-\bar{\eta}) - \eta(1-\eta)|} \\ &\quad \leq |\eta - \bar{\eta}| \exp(|p|) + 2\sqrt{|\bar{\eta} - \eta|}.\end{aligned}$$

Clearly, if $f(x) \in C$ has a small $\Delta Q(f(\cdot))$, then for points such that $\eta(x) \approx 1$ [or $\eta(x) \approx 0$], we require that $1/(1+e^{-2f(x)}) \approx 1$ [or $1/(1+e^{-2f(x)}) \approx 0$]. This

implies that $|f(x)|$ is large. In this case, since $\exp(|f(x)|)$ is also large, it becomes more difficult to achieve a small value of $\Delta Q(\eta(x), f(x))$. Additionally, using the exponential loss, we compute a predictor such that $|f(x)|$ is large when $\eta(x) \approx 0, 1$, and $|f(x)|$ is small elsewhere. In the limit of zero error, $f(x)$ has to achieve values of $\pm\infty$ if $\eta(x) = 0, 1$. Such a predictor is clearly not very well behaved. A similar problem also exists for the logistic regression loss, although to a lesser degree.

3.5. *Logistic regression.* We consider the case $\phi(v) = \ln(1 + \exp(-v))$. Note that $f_\phi^*(\eta) = \ln \frac{\eta}{1-\eta}$ and $Q^*(\eta) = -\eta \ln \eta - (1-\eta)\ln(1-\eta)$. $f_\phi^*(\eta)$ is invertible,

$$f_\phi^{*-1}(p) = \frac{1}{1+e^{-p}}.$$

The logistic transform $1/(1+e^{-f(x)})$ of $f(x)$ can be regarded as an approximation to the true conditional in-class probability $\eta(x)$. The corresponding estimation method can be regarded as a maximum likelihood estimate with this probability model.

In logistic regression, Theorem 2.2 implies that the closeness of $1/(1+e^{-f(x)})$ to $\eta(x)$ is measured by the expected Bregman divergence of $-\eta \ln \eta - (1-\eta)\ln(1-\eta)$, which is essentially the relative entropy between $\eta(x)$ and $1/(1+e^{-f(x)})$ (also called KL-divergence),

$$\Delta Q(\eta, p) = \eta \ln[\eta(1+e^{-p})] + (1-\eta)\ln[(1-\eta)(1+e^{p})] = \mathrm{KL}\left(\eta \middle|\middle| \frac{1}{1+e^{-p}}\right).$$

This distance measurement is not surprising since it holds for all maximum likelihood estimation methods.

Let $\bar{\eta} = f_\phi^{*-1}(p) = 1/(1+\exp(-p))$. One may obtain a lower bound for $\mathrm{KL}(\eta||\bar{\eta})$ using Taylor expansion: $\exists\, \eta'$ between $\bar{\eta}$ and η such that

$$\begin{aligned}\Delta Q(\eta, p) = \mathrm{KL}(\eta||\bar{\eta}) &= -\frac{1}{2}Q^{*\prime\prime}(\eta')(\eta-\bar{\eta})^2 \\ &= \frac{1}{2\eta'(1-\eta')}(\eta-\bar{\eta})^2 \geq 2(\eta-\bar{\eta})^2.\end{aligned} \tag{12}$$

In particular,

$$\Delta Q(\eta, 0) \geq 2(\eta - 0.5)^2.$$

Therefore we may let $s = 2$ and $c = 2^{-1/2}$ in Corollary 2.1.

The lower bound of $\mathrm{KL}(\eta, \bar{\eta})$ in (12) implies that if we can find $\hat{p}(x) \in C$ such that $\Delta Q(\hat{p}(\cdot))$ is small, then the expected squared difference between the estimated conditional probability $1/(1+e^{-f(x)})$ and the true conditional probability $\eta(x)$ is also small. This result is similar to that of (modified) least squares. However, logistic regression has a similar problem as the exponential loss

when $\eta(x)$ is close to 0 or close to 1: $|f(x)|$ has to be very large in order to approximate such a value.

In order to compare the logistic regression loss with the exponential loss used in AdaBoost, we shall derive an upper bound on the KL-divergence in a way that is similar to the corresponding derivation in the exponential loss case (we do not attempt to optimize the bound),

$$\begin{aligned}
&\Delta Q(\eta, p)\\
&\quad = (\eta - \bar{\eta})\big(-\ln \bar{\eta} + \ln(1-\bar{\eta})\big) - (\bar{\eta}\ln\bar{\eta} - \eta\ln\eta)\\
&\qquad - \big((1-\bar{\eta})\ln(1-\bar{\eta}) - (1-\eta)\ln(1-\eta)\big)\\
&\quad \le 2\big|\eta - \bar{\eta}\big|\ln(1+e^{|p|}) + \big|(\bar{\eta}\ln\bar{\eta} - \eta\ln\eta)(\bar{\eta}\ln\bar{\eta} + \eta\ln\eta)\big|^{1/2}\\
&\qquad + \big|\big((1-\bar{\eta})\ln(1-\bar{\eta}) - (1-\eta)\ln(1-\eta)\big)\\
&\qquad\quad \times \big((1-\bar{\eta})\ln(1-\bar{\eta}) + (1-\eta)\ln(1-\eta)\big)\big|^{1/2}\\
&\quad \le 2|\eta - \bar{\eta}|\ln(1+e^{|p|}) + 2\sqrt{k|\bar{\eta} - \eta|},
\end{aligned}$$

where k is a constant. In the last inequality, we used the following derivation on the second and the third terms: using Taylor's expansion, there exists z between x and y so that

$$(x\ln x)^2 - (y\ln y)^2 = 2z\ln z(1+\ln z)(x-y) \le k|x-y|,$$

where $k = \sup_{0<z\le 1}|2z\ln z(1+\ln z)|$.

By comparing the upper bound of $\Delta Q(\eta, p)$ for logistic regression and the similar upper (or lower) bound of $\Delta Q(\eta, p)$ for the exponential loss, we find that logistic regression changes the exponential sensitivity $\exp(|p|)$ to $\ln(1+\exp(|p|))$ which behaves linearly when $|p|$ is large. This means that if $|p|$ is large, then $\Delta Q(\eta, p)$ is likely to be much smaller with the logistic regression loss. Also note that a predictor $f(x)$ with the exponential loss induces the same conditional probability estimate as that of the scaled predictor $2f(x)$ with logistic regression loss. This implies that the two loss functions share the same probability model (up to a scaling factor). We conclude from our analysis that logistic regression loss behaves better than the exponential loss for large $|f(x)|$ which occurs when $\eta(x)(1-\eta(x)) \approx 0$.

3.6. *Remarks on different loss functions.* Our analysis indicates that all loss functions considered in this section measure the closeness of a transformation of $f(x)$ to the true conditional in-class probability $\eta(x)$. This generalizes the conditional maximum likelihood estimate where the closeness is measured by the KL-divergence.

Both logistic regression and exponential losses utilize the logistic transform to relate a predictor $f(x)$ and the approximate conditional probability it represents.

However, we show that logistic regression is likely to give a better estimate when $|p|$ is large. On the other hand, both methods have the same drawback that $|f(x)|$ has to be very large in order to approximate the true conditional probability $\eta(x)$ well when $\eta(x)(1-\eta(x)) \approx 0$. This problem occurs for a loss ϕ such that the derivative of Q^* at 0 or 1 is ill-defined since in this case, the Bregman divergence d_{Q^*} can be ill-behaved.

In the case of logistic regression, the above phenomenon is related to a fundamental shortcoming of the maximum likelihood estimate. That is, it is not a robust rare-event estimator. For example, consider a coin toss experiment with a head probability of $\eta = 0.999$. Consider two models: the first is $\eta = 1$, and the second is $\eta = 0.1$. Clearly for most practical purposes (such as classification), the first model is much more accurate than the second one. However, if we use the maximum likelihood estimate with a sufficiently large number of samples, then we will almost always choose the second model since the first model has a large probability of giving the zero likelihood.

This problem can be easily avoided by using a convex loss ϕ such that the derivative of Q^* is well-behaved. For example, this is true for least squares, modified least squares and SVM losses. A significant drawback of SVM is that it approximates the binary-classification decision rule $\operatorname{sign}(2\eta(x) - 1)$, rather than the conditional probability $\eta(x)$ itself. This implies that an SVM classifier tends to give unreliable information on the confidence of its prediction. However, such information can be crucial for many practical applications. For example, in principle we cannot directly apply SVMs to a multi-class classification problem by training separate SVMs to predict each class and choose the class that is most confidently predicted—this scheme fails at points where the conditional probabilities for all classes are below 0.5.

We have also studied the least squares method for classification and its modification. Both methods can provide reliable confidence information since they directly estimate the conditional probability $\eta(x)$. In both cases, the conditional probability estimates are given by $(f(x) + 1)/2$ truncated to the interval $[0, 1]$. Such an estimate is well-behaved even when $\eta(x)(1 - \eta(x)) \approx 0$. This is a significant advantage over logistic regression. We also show that in general the modified least squares method gives a better approximation to the conditional probability $\eta(x)$ than the standard least squares method. Using a similar argument, we may in fact obtain an even better method with the loss function:

$$\phi(v) = \begin{cases} -4v, & v < -1, \\ (v-1)^2, & v \in [-1, 1], \\ 0, & v > 1. \end{cases}$$

This new loss function, which we call modified Huber's loss, changes the modified least squares loss so that it penalizes misclassified points with $v < -1$ only linearly. Using this formula, we can obtain $f_\phi^*(\eta) = 2\eta - 1$ and $Q^*(\eta) = 4\eta(1-\eta)$.

The subgradient of ϕ is uniquely determined as $\phi'(v) = \max(\min(2(v-1), 0), -4)$. For $p_1 \in [-1, 1]$, the Bregman divergence is

$$d_\phi(p_1, p_2) = (p_2 - p_1)^2 - \max(0, p_2 - 1)^2 - \min(0, p_2 + 1)^2.$$

It is not difficult to check that

$$\begin{aligned}\Delta Q(\eta, p) &= \Delta Q\big(\eta, T(p)\big) + Q(\eta, p) - Q\big(\eta, T(p)\big) \\ &= \big(2\eta - 1 - T(p)\big)^2 + 2|2\eta - 1 - T(p)||p - T(p)|,\end{aligned}$$

where $T(p) = \min(\max(p, -1), 1)$. Obviously this distance metric is better than that of modified least squares in Lemma 3.1 when $|p|$ is large. This distance function behaves like that of SVM and logistic regression losses in that all three are linear in $|p|$ when $|p|$ is large. The new formulation is more attractive than the support vector formulation since it directly approximates the conditional probability $\eta(x)$. In addition, it is better behaved than logistic regression since it does not require a large value of $|f(x)|$ when $\eta(x)$ is close to 0 or 1.

4. Universal approximation and consistency. Consider a function class C and an appropriate convex loss function ϕ. If $\inf_{f \in C} \Delta Q(f)$, the loss function induced distance of the conditional in-class probability $\eta(x)$ to C, is small, then any $f(x) \in C$ that approximately minimizes (4) achieves a classification error close to the optimal Bayes error. In particular, if the distance is zero, then one can find a classifier by approximately minimizing (4) with classification error rate arbitrarily close to that of the Bayes rate.

We call a function class C universal with respect to a convex loss function ϕ if any measurable conditional density function $\eta(x)$ has a distance of zero to C. Section 4.1 proves a general universal approximation theorem. Examples of universal function classes are given in Section 4.2.

Given a function class C that is universal, we can find an approximate Bayes classifier by approximately minimizing (4). Since statistical machine learning requires the use of finite sample size to approximate the true underlying distribution, we need to use an estimation method to obtain a sample dependent predictor $\hat{f}_n(\cdot) \in C$ such that $Q(\hat{f}_n(\cdot))$ converges to $\inf_{f \in C} Q(f(\cdot))$ in probability. Section 4.3 discusses a complexity regularization approach to achieve this convergence. The results demonstrate the universality of some kernel based classification methods such as support vector machines. However in practice, one may also use other schemes to achieve this convergence. For example, the framework developed in this work has been applied in [7] to demonstrate the consistency of boosting like procedures using a different regularization scheme.

4.1. *A universal approximation theorem.* We only consider classification problems in $\mathbb{R}^d$. Although the analysis can be generalized to other measure spaces, we do not consider them here for simplicity.

DEFINITION 4.1. Let $U \subset \mathbb{R}^d$. We denote by $C(U)$ the Banach space of continuous functions: $U \to \mathbb{R}$ under the uniform-norm topology.

DEFINITION 4.2. We call a probability measure μ in $\mathbb{R}^d$ regular if it is defined on the Borel sets of $\mathbb{R}^d$.

Note that in the above definition we have only used a special case of regular measures in real analysis for which Lusin's theorem holds.

DEFINITION 4.3. We say a convex function $\phi : \mathbb{R} \to \mathbb{R}$ has property A if:

- ϕ is continuous in $\mathbb{R}$ and Q^* is continuous on $[0, 1]$.
- $\phi(p) < \phi(-p)$ for all $p > 0$.
- $f^*_\phi(\eta) \in (-\infty, +\infty)$ and is piecewise continuous in $(0, 1)$.

LEMMA 4.1. *Assume* $0 \le \delta < 0.5$. *Let* $\eta \in [0, 1]$ *and* $\eta_\delta = \min(\max(\eta, \delta), 1 - \delta)$. *If* ϕ *has property* A, *then*

$$Q\big(\eta, f^*_\phi(\eta_\delta)\big) \le Q^*(\eta_\delta).$$

PROOF. Due to the symmetry with respect to the transformation $\eta \to 1 - \eta$, we only need to consider the case $\eta > \eta_\delta > 0.5$. The condition $\phi(p) < \phi(-p)$ for all $p > 0$ implies that $f^*_\phi(\eta_\delta) > 0$, and thus $\phi(f^*_\phi(\eta_\delta)) - \phi(-f^*_\phi(\eta_\delta)) < 0$. Let $p_\delta = f^*_\phi(\eta_\delta)$. It follows that

$$\begin{aligned} Q(\eta, p_\delta) &= \eta\phi(p_\delta) + (1 - \eta)\phi(-p_\delta) \\ &= \eta_\delta\phi(p_\delta) + (1 - \eta_\delta)\phi(-p_\delta) + (\eta - \eta_\delta)\big(\phi(p_\delta) - \phi(-p_\delta)\big) \\ &< Q^*(\eta_\delta). \end{aligned}$$

□

We can now prove the following universal approximation theorem:

THEOREM 4.1. *Let* ϕ *be a convex function which has property* A. *Consider a function class* $C \subset C(U)$ *defined on a Borel set* $U \subset \mathbb{R}^d$. *If* C *is dense in* $C(U)$, *then for any regular probability measure* μ *of* $x \in \mathbb{R}^d$ *such that* $\mu(U) = 1$, *and any conditional probability* $P(Y = 1|X = x) = \eta(x) : \mathbb{R}^d \to [0, 1]$ (*measurable with respect to* μ),

$$\inf_{p \in C} \Delta Q(f(\cdot)) = 0,$$

where (X, Y) *is distributed according to* (μ, η).

PROOF. Given any $\varepsilon > 0$, by the continuity of Q^* we can find $\delta \in (0, 0.5)$ such that

$$\sup_{|\eta-\eta'|\le\delta} |Q^*(\eta) - Q^*(\eta')| < \varepsilon. \tag{13}$$

Define

$$\eta_\delta(x) = \min\big(\max(\eta(x), \delta), 1-\delta\big),$$

which is clearly measurable. Also define

$$K_\delta = \sup_{\delta\le\eta\le1-\delta} |f_\phi^*(\eta)|, \qquad M_\delta = \sup_{|z|\le K_\delta} |\phi(z)| + 1.$$

Using the assumptions on f_ϕ^* and ϕ, we have $K_\delta, M_\delta < +\infty$.

Since μ is regular, using Lusin's theorem in measure theory (e.g., see [12], page 55), we know that $f_\phi^*(\eta_\delta(x))$ can be approximated by a continuous function $\alpha'(x) \in C(U)$ such that $|\alpha'(x)| \le K_\delta$ and $P(f_\phi^*(\eta_\delta(x)) \ne \alpha'(x)) \le \varepsilon/(2M_\delta)$. This implies that

$$\mathbf{E}_X Q\big(\eta(X), \alpha'(X)\big) \le \mathbf{E}_X Q\big(\eta(X), f_\phi^*\big(\eta_\delta(X)\big)\big) + \varepsilon.$$

Using Lemma 4.1, and (13), we have

$$\mathbf{E}_X Q\big(\eta(X), f_\phi^*\big(\eta_\delta(X)\big)\big) \le \mathbf{E}_X Q^*\big(\eta_\delta(X)\big) \le \mathbf{E}_X Q^*\big(\eta(X)\big) + \varepsilon.$$

This implies that

$$\mathbf{E}_X Q\big(\eta(X), \alpha'(X)\big) \le \mathbf{E}_X Q^*\big(\eta(X)\big) + 2\varepsilon. \tag{14}$$

Since $\alpha'(x)$ is continuous and bounded in U, and ϕ is continuous, using the assumption that C is dense in $C(U)$, we can find $p_c \in C$ such that

$$\sup_{x\in U}\big[|\phi\big(p_c(x)\big) - \phi\big(\alpha'(x)\big)| + |\phi\big(-p_c(x)\big) - \phi\big(-\alpha'(x)\big)|\big] \le \varepsilon.$$

This implies that

$$\mathbf{E}_X Q\big(\eta(X), p_c(X)\big) \le \mathbf{E}_X Q\big(\eta(X), \alpha'(X)\big) + \varepsilon.$$

Together with (14), we obtain the theorem. □

Note that all convex functions ϕ considered in this paper have property A. Therefore from Theorem 2.1, we know that as long as we choose a function class $C \in C(U)$ such that C is dense in $C(U)$, we are able to achieve classification error arbitrarily close to the optimal Bayes error by approximately minimizing (4) within C.

4.2. *Universality of some function classes.* We consider function classes $\mathbb{R}^d \to \mathbb{R}$ consisting of linear combinations of functions of the form $h(w^T x + b)$, where $w \in \mathbb{R}^d$, $b \in \mathbb{R}$, and $h : \mathbb{R} \to \mathbb{R}$ is a fixed continuous function,

$$C_h = \left\{ \sum_{i=1}^{k} \alpha_i h(w_i^T x + b_i) : \alpha_i \in \mathbb{R}, w_i \in \mathbb{R}^d, b_i \in \mathbb{R}, k \in N \right\}.$$

These types of function classes have been extensively studied in the neural networks literature in the last ten years. For example, it is well known that any continuous function on a compact subset of $\mathbb{R}^d$ can be uniformly well approximated by a function in C_h if h is a sigmoidal function. This result means that two-level neural networks (with sigmoidal activation function h) are universal approximators. In this paper, we use the following general version of a neural network universal approximation result which was proved in [8]:

THEOREM 4.2 ([8]). *If h is a nonpolynomial continuous function, then C_h is dense in $C(U)$ for all compact subsets U of $\mathbb{R}^d$.*

We should mention that in the original theorem, the density result is stated for the restriction of $C(\mathbb{R}^d)$ to a compact subset U of $\mathbb{R}^d$. However, since any continuous function in $C(U)$ can be extended to a continuous function in $\mathbb{R}^d$ (a special case of the Tietze extension theorem in topology), the restriction of $C(\mathbb{R}^d)$ to U is equivalent to $C(U)$.

In the following we consider kernel induced function classes. Let K be a symmetric positive kernel. That is, $K(a, b) = K(b, a)$, and the $n \times n$ Gram matrix $G = [K(x_i, x_j)]_{i,j=1,\ldots,n}$ is always positive semi-definite. We have the following definition:

DEFINITION 4.4. Let $H_0 = \{\sum_{i=1}^{L} \alpha_i K(x_i, x) : L \in N, \alpha_i \in \mathbb{R}\}$. H_0 is an inner product space with norm defined as

$$\left\| \sum_i \alpha_i K(x_i, \cdot) \right\| = \left(\sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) \right)^{1/2}.$$

Let H be the closure of H_0 under the norm $\|\cdot\|$. Then H forms a Hilbert space, called the reproducing kernel Hilbert space of K.

Note that if $\|\sum_i \alpha_i K(x_i, \cdot)\| = 0$, then $\sum_{i=1} \alpha_i K(x_i, x) = 0$ for all x (otherwise, the Gram matrix with x_i and x will not be positive semi-definite). Therefore the inner product in the above definition is well defined. In fact, we have the following bound:

PROPOSITION 4.1. *Let K be a symmetric positive kernel. $\forall f(x) \in H$ and variable x*

$$|f(x)| \le \|f(\cdot)\| K(x,x)^{1/2}.$$

Given a datum x, we can associate it with the function $K(x,\cdot) \in H$. It is easy to check that for all $f(\cdot) \in H$: $f(x) = \langle f(\cdot), K(x,\cdot)\rangle$, where we use $\langle\cdot,\cdot\rangle$ to denote the inner product in H. This fact will be used in the next section. For further information on reproducing kernel Hilbert spaces, we refer interested readers to [17].

We now consider kernel functions of the form

$$K_h([x_1,b_1],[x_2,b_2]) = h(x_1^T x_2 + b_1 b_2),$$

where h can be expressed as a Taylor's expansion with nonnegative coefficients. It is well known that K_h is a positive definite kernel in this case. An easy way to see this is by writing $K_h(z_1,z_2) = \sum_i \psi_i(z_1)\psi_i(z_2)$ using Taylor's expansion: K_h now acts as an inner product in the so-called feature space $[\psi_i(z)]$.

We denote by H_h the corresponding reproducing kernel Hilbert space of K_h in $\mathbb{R}^{d+1}$. It induces a function class $\bar{H}_h$ in $\mathbb{R}^d$ with $f \in H_h \to \bar{f} \in \bar{H}_h$ defined by $\bar{f}(x) = f([x,1])$. Clearly, $C_h \subset \bar{H}_h$. Without causing confusion, in the following we also denote $\bar{f}$ by f for simplicity,

$$f(x) = \bar{f}(x) = f([x,1]).$$

4.3. *Estimation error and consistency of kernel formulations.* Consider the estimation problem

$$\hat{f}_n = \arg\inf_{f\in H_h}\left[\frac{1}{n}\sum_{i=1}^{n}\phi\big(\bar{f}(X_i)Y_i\big) + \frac{\lambda_n}{2}\|f\|^2\right], \tag{15}$$

where $\lambda_n > 0$ is a small regularization parameter. (X_i, Y_i) are training data drawn from the unknown underlying distribution D. (15) can be regarded as a stochastic approximation to the minimization of (4).

Although (15) is formulated as an infinite dimensional optimization problem, it is well known that the computation can be performed in a finite dimensional space (e.g., see [17, 18]). Let $f^\perp$ be the orthogonal projection of f onto the subspace V_X spanned by $g_i(x) = h(X_i^T x + 1) \in \bar{H}_h$ $(i = 1,\dots,n)$. Then by definition $f(X_i) - f^\perp(X_i) = \langle (f - f^\perp), g_i\rangle = 0$. Since $\forall f \notin V_X$ we have $\|f^\perp\| < \|f\|$, it is easy to observe that $f^\perp$ always has a smaller objective value in (15) than f if $f \notin V_X$. Therefore the optimal solution $\hat{f}_n$ must lie in the finite dimensional space V_X. Since the formulation is strictly convex and the right-hand side approaches infinity as $f \to \infty$, it has a unique finite solution which lies in V_X.

To obtain estimation error on kernel methods, we can use the leave-one-out analysis. Some relatively general results for kernel methods were obtained in [18]. The following bound is slightly weaker than a corresponding result in [18] but is easier to prove.

THEOREM 4.3. *Let* $\hat{f}^{[k]}$ *be the solution of* (15) *with the kth datum removed from the training set. Then*

$$\|\hat{f}_n(\cdot) - \hat{f}_n^{[k]}(\cdot)\| \le \frac{2}{\lambda_n n}|\phi'(\hat{f}_n(X_k)Y_k)|h(X_k^T X_k + 1)^{1/2},$$

where ϕ' *denotes a subgradient of* ϕ.

PROOF. The minimizer $\hat{f}_n$ of (15) lies in the finite dimensional space V_X spanned by $g_i(x) = h(X_i^T x + 1) \in \bar{H}_h$ $(i = 1, \ldots, n)$. Using the linear representation $f(X_i) = \langle f, g_i \rangle$, and Theorem 23.8 in [11], we know that there exist subgradients $\phi'(\hat{f}_n(X_i)Y_i)$ $(i = 1, \ldots, n)$ such that the following first-order condition holds:

$$\frac{1}{n}\sum_{i=1}^{n} \phi'(\langle \hat{f}_n, g_i \rangle Y_i) g_i Y_i + \lambda_n \hat{f}_n = 0. \tag{16}$$

We also have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n-1} \phi(\hat{f}_n(X_i)Y_i) + \frac{\lambda_n}{2}\|\hat{f}_n\|^2 \\
&\quad + \left[\frac{1}{n}\sum_{i=1}^{n-1} \phi'(\hat{f}_n(X_i)Y_i)(\hat{f}^{[n]}(X_i) - \hat{f}_n(X_i))Y_i + \lambda_n \langle \hat{f}_n, \hat{f}^{[n]} - \hat{f}_n \rangle\right] \\
&\quad + \frac{\lambda_n}{2}\|\hat{f}^{[n]} - \hat{f}_n\|^2 \\
&\quad = \frac{1}{n}\sum_{i=1}^{n-1} \left[\phi(\hat{f}^{[n]}(X_i)Y_i) - d_\phi(\hat{f}_n(X_i)Y_i, \hat{f}^{[n]}(X_i)Y_i)\right] + \frac{\lambda_n}{2}\|\hat{f}^{[n]}\|^2 \\
&\quad \le \frac{1}{n}\sum_{i=1}^{n-1} \phi(\hat{f}^{[n]}(X_i)Y_i) + \frac{\lambda_n}{2}\|\hat{f}^{[n]}\|^2,
\end{aligned}$$

where d_ϕ denotes the Bregman divergence of ϕ. Also note that by the definition of $\hat{f}^{[n]}$, we have

$$\frac{1}{n}\sum_{i=1}^{n-1} \phi(\hat{f}_n(X_i)Y_i) + \frac{\lambda_n}{2}\|\hat{f}_n\|^2 \ge \frac{1}{n}\sum_{i=1}^{n-1} \phi(\hat{f}^{[n]}(X_i)Y_i) + \frac{\lambda_n}{2}\|\hat{f}^{[n]}\|^2.$$

Therefore by comparing the above two inequalities, we obtain

$$\begin{aligned}
&\frac{\lambda_n}{2}\|\hat{f}_n - \hat{f}^{[n]}\|^2 \\
&\quad\le -\left[\frac{1}{n}\sum_{i=1}^{n-1}\phi'\big(\hat{f}_n(X_i)Y_i\big)Y_i\langle g_i, \hat{f}^{[n]} - \hat{f}_n\rangle + \lambda_n\langle \hat{f}_n, \hat{f}^{[n]} - \hat{f}_n\rangle\right] \\
&\quad\le \left\|\frac{1}{n}\sum_{i=1}^{n-1}\phi'\big(\hat{f}_n(X_i)Y_i\big)g_iY_i + \lambda_n\hat{f}_n\right\|\|\hat{f}^{[n]} - \hat{f}_n\| \\
&\quad= \frac{1}{n}|\phi'\big(\hat{f}_n(X_n)Y_n\big)|h(x_n^Tx_n+1)^{1/2}\|\hat{f}^{[n]} - \hat{f}_n\|.
\end{aligned}$$

The equality follows from (16) and the fact that $\|g_n\|^2 = h(x_n^Tx_n + 1)$. By canceling the factor $\|\hat{f}^{[n]} - \hat{f}_n\|$ from the above inequality, we obtain the desired bound with $k = n$. □

The above theorem can be used to derive leave-one-out estimates for different kernel formulations in (15). In fact, a straightforward application of Theorem 4.3 and Proposition 4.1 lead to the following leave-one-out cross-validation error bound:

$$\begin{aligned}
&\sum_{k=1}^{n}\phi\big(\hat{f}^{[k]}(X_k)Y_k\big) \\
&\quad\le \sum_{k=1}^{n}\sup_{|\beta_k|\le 1}\phi\left(\hat{f}_n(X_k)Y_k + \frac{2\beta_k}{\lambda_n n}|\phi'\big(\hat{f}_n(X_k)Y_k\big)|h\big(X_k^TX_k+1\big)\right).
\end{aligned} \tag{17}$$

Note that the expected leave-one-out error of any estimator $\hat{f}$ is equivalent to the expected generalization error of $\hat{f}$. By Markov's inequality, for all $\varepsilon > 0$, we have

$$P\left(Q\big(\hat{f}(\cdot)\big) > \inf_{f\in\bar{H}_h}Q(f(\cdot)) + \varepsilon\right) \le \frac{1}{\varepsilon}\left(\mathbf{E}Q\big(\hat{f}(\cdot)\big) - \inf_{f\in\bar{H}_h}Q(f(\cdot))\right).$$

This implies that if we can choose λ_n such that

$$\lim_{n\to\infty}\mathbf{E}Q\big(\hat{f}_n(\cdot)\big) = \inf_{f\in\bar{H}_h}Q(f(\cdot)),$$

then $Q(\hat{f}_n(\cdot))$ converges to $\inf_{f\in\bar{H}_h}Q(f(\cdot))$ in probability.

Therefore, to show consistency, we only need to estimate leave-one-out bounds using Theorem 4.3 for formulations we are interested in. For simplicity, from now on we assume that $P(h(X^TX + 1) \le M^2) = 1$.

We start with the following simple bound:

COROLLARY 4.1. *Under the assumptions of Theorem* 4.3, *and further that* $\phi(\cdot) \geq 0$ *and* $P(h(X^T X + 1) \leq M^2) = 1, \forall k$ *the expected leave-one-out error can be bounded as*

$$\mathbf{E}Q(\hat{f}^{[k]}) \leq \inf_{f \in H_h} \left[Q(\bar{f}(\cdot)) + \frac{\lambda_n}{2} \|f\|^2 \right] + \frac{2M^2 \Delta_\phi^2}{\lambda_n n},$$

where the expectation is with respect to the training samples $(X_1, Y_1), \ldots, (X_n, Y_n)$, *and*

$$\Delta_\phi = \sup \left\{ |\phi'(z)| : |z| \leq \sqrt{\frac{2\phi(0)}{\lambda_n}} M \right\}.$$

PROOF. Note that by the definition of $\hat{f}_n$, we obtain

$$\frac{1}{n} \sum_{i=1}^{n} \phi(\hat{f}_n(X_i) Y_i) + \frac{\lambda_n}{2} \|\hat{f}_n\|^2 \leq \phi(0).$$

Therefore $\|\hat{f}_n\| \leq \sqrt{2\phi(0)/\lambda_n}$. This implies that $|\hat{f}_n(X_k)| \leq \sqrt{2\phi(0)/\lambda_n} M$ for all k. Similarly $|\hat{f}^{[k]}(X_k)| \leq \sqrt{2\phi(0)/\lambda_n} M$ for all k. Therefore from Taylor expansion

$$\phi(\hat{f}^{[k]}(X_k) Y_k) - \phi(\hat{f}_n(X_k) Y_k) \leq \Delta_\phi |\hat{f}^{[k]}(X_k) - \hat{f}_n(X_k)| \leq \Delta_\phi^2 \frac{2M^2}{\lambda_n n}$$

for all k. Note that the second inequality follows from Theorem 4.3 and Proposition 4.1. Summing over k, we obtain

$$\frac{1}{n} \sum_{k=1}^{n} \phi(\hat{f}^{[k]}(X_k) Y_k) \leq \inf_{f \in H_h} \left[\frac{1}{n} \sum_{k=1}^{n} \phi(\bar{f}(X_k) Y_k) + \frac{\lambda_n}{2} \|f\|^2 \right] + \frac{2M^2 \Delta_\phi^2}{\lambda_n n}.$$

Taking expectation with respect to the training data, we obtain the corollary. □

We list bounds of Δ_ϕ for loss functions considered in this paper:

- SVM: $\Delta_\phi \leq 1$.
- Logistic regression: $\Delta_\phi \leq 1$.
- Modified Huber: $\Delta_\phi \leq 4$.
- Least squares: $\Delta_\phi \leq \sqrt{\frac{8}{\lambda_n}} M + 2$.
- Modified least squares: $\Delta_\phi \leq \sqrt{\frac{8}{\lambda_n}} M + 2$.
- Exponential: $\Delta_\phi \leq \exp(\sqrt{\frac{2}{\lambda_n}} M)$.

Although Corollary 4.1 is useful for certain loss functions, in many cases better bounds can be obtained from Theorem 4.3 using refined analysis (e.g., see [18]). We will only consider the least squares loss and the modified least squares loss here for simplicity.

COROLLARY 4.2. *Consider* $\phi(v) = (1-v)^2$ *or* $\phi(v) = \max(1-v, 0)^2$. *Then under the assumptions of Theorem* 4.3, *for any* k *the expected leave-one-out error can be bounded as*

$$\mathbf{E}Q\big(\hat{f}_n^{[k]}(\cdot)\big) \le \left(1 + \frac{4M^2}{\lambda_n n}\right)^2 \inf_{f(\cdot)\in H_h}\left[Q\big(\bar{f}(\cdot)\big) + \frac{\lambda_n}{2}\|f(\cdot)\|^2\right],$$

where $P(h(X^T X + 1) \le M^2) = 1$.

PROOF. In both cases we have $\phi(v+\delta)^{1/2} \le \phi(v)^{1/2} + (\delta^2)^{1/2}$. This implies that for all $\beta_k \in [-1, 1]$ $(= 1, \ldots, n)$,

$$\begin{aligned}
&\left[\sum_{k=1}^n \phi\left(\hat{f}_n(X_k)Y_k + \frac{2\beta_k}{\lambda_n n}|\phi'\big(\hat{f}_n(X_k)Y_k\big)|M^2\right)\right]^{1/2} \\
&\quad\le \left[\sum_{k=1}^n \phi\big(\hat{f}_n(X_k)Y_k\big)\right]^{1/2} + \left[\sum_{k=1}^n \left(\frac{2\beta_k}{\lambda_n n}|\phi'\big(\hat{f}_n(X_k)Y_k\big)|M^2\right)^2\right]^{1/2} \\
&\quad= \left[\sum_{k=1}^n \phi\big(\hat{f}_n(X_k)Y_k\big)\right]^{1/2} + \left[\sum_{k=1}^n \frac{16\beta_k^2 M^4}{(\lambda_n n)^2}\phi\big(\hat{f}_n(X_k)Y_k\big)\right]^{1/2} \\
&\quad\le \left(1 + \frac{4M^2}{\lambda_n n}\right) \inf_{f\in H_h}\left[\sum_{i=1}^n \phi\big(\bar{f}(X_i)Y_i\big) + \frac{\lambda_n n}{2}\|f\|^2\right]^{1/2},
\end{aligned}$$

where the equality follows from the fact that $\phi'(v)^2 = 4\phi(v)$. Now using (17) and taking expectation with respect to the training data, we obtain the corollary. □

Using Corollary 4.2 for least squares and modified least squares, and Corollary 4.1 for other formulations, we immediately obtain the following theorem:

THEOREM 4.4. *Let* h *be an entire function with nonnegative Taylor coefficients. Assume we choose* λ_n *in* (15) *such that* $\lambda_n \to 0$ *and* $\lambda_n n \to \infty$ *with the following loss functions* ϕ: *least squares*, *modified least squares*, *modified Huber's loss*, *SVM or logistic regression*; *or we choose* λ_n *such that* $\lambda_n \to 0$ *and* $\lambda_n \log^2 n \to \infty$ *with the exponential loss. Then for any distribution* D *with regular input probability measure which is bounded almost everywhere in* $\mathbb{R}^d$, *we have*

$$\lim_{n\to\infty} \mathbf{E}Q\big(\hat{f}_n(\cdot)\big) = \inf_{f\in\bar{H}_h} Q(f(\cdot)).$$

Moreover, if h *is not a polynomial, then*

$$\lim_{n\to\infty} \mathbf{E}\Delta Q\big(\hat{f}_n(\cdot)\big) = 0.$$

This implies that the classification error of $\hat{f}_n(\cdot)$ *converges to the optimal Bayes error in probability as* $n \to \infty$.

5. Conclusion. In this paper we have studied how close the optimal Bayes error rate can be approximately reached using a classification algorithm that computes a classifier by minimizing a convex upper bound of the classification error function.

We have separated approximation and estimation aspects of the problem. The former is introduced through the bias of the formulation, which we investigated in Section 2. In particular, we related the quantity $Q(f(\cdot))$ defined in (4) to the classification performance of $f(\cdot)$ measured by $L(f(\cdot)) - L^*$.

In our framework, the method of minimizing (4) can be regarded as computing an approximation of the conditional in-class probability $\eta(x)$. In this regard, the closeness of the true conditional probability $\eta(x)$ and an estimated approximation is defined by a loss-function induced distance.

We have analyzed this distance function for a number of convex loss functions. We have shown that exponential and logistic regression losses are not well-behaved when the conditional probability $\eta(x)$ is close to 0 or 1. Although a support vector machine does not suffer from this problem, it computes a predictor that approximates $\text{sign}(2\eta(x) - 1)$. This implies that a support vector machine does not provide reliable confidence information (for its prediction) which can be very useful in practice. In particular, they are not directly applicable to multi-class classification problems. We have also shown that both least squares and modified least squares methods can lead to reliable conditional probability estimates, and they are well-behaved when $\eta(x)$ is close to 0 or 1. In addition, the modified least squares method gives better approximation than the standard least squares. We also proposed a new convex loss, which we call modified Huber's loss, to further enhance the modified least squares formulation. In our analysis, this new loss function achieves the best overall approximation behavior.

The analysis of loss functions introduced in this paper can also be used to demonstrate the consistency of statistical classification methods using convex risk minimization. In particular, we proved a universal approximation theorem in Section 4. Using this result, we obtained universal consistency of certain kernel based classification methods such as support vector machines.

Acknowledgments. The author would like to thank anonymous referees for pointing out related work, and for constructive suggestions that helped to improve the presentation of the paper.

REFERENCES

[1] BREGMAN, L. M. (1967). The relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Computational Mathematics and Mathematical Physics* **7** 200–217.

[2] BREIMAN, L. (1998). Arcing classifiers (with discussion). *Ann. Statist.* **26** 801–849.

[3] BREIMAN, L. (1999). Prediction games and arcing algorithms. *Neural Computation* **11** 1493–1517.

[4] BREIMAN, L. (2000). Some infinity theory for predictor ensembles. Technical Report 577, Dept. Statistics, Univ. California, Berkeley.
[5] BÜHLMANN, P. and YU, B. (2003). Boosting with L_2-loss: Regression and classification. *J. Amer. Statist. Assoc.* **98** 324–339.
[6] FREUND, Y. and SCHAPIRE, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* **55** 119–139.
[7] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion). *Ann. Statist.* **28** 337–407.
[8] LESHNO, M., LIN, YA. V., PINKUS, A. and SCHOCKEN, S. (1993). Multilayer feedforward networks with a non-polynomial activation function can approximate any function. *Neural Networks* **6** 861–867.
[9] LUGOSI, G. and VAYATIS, N. (2004). On the Bayes-risk consistency of regularized boosting methods. *Ann. Statist.* **32** 30–55.
[10] MANNOR, S., MEIR, R. and ZHANG, T. (2002). The consistency of greedy algorithms for classification. In *Proc. 15th Annual Conference on Computational Learning Theory. Lecture Notes in Comput. Sci.* **2375** 319–333. Springer, New York.
[11] ROCKAFELLAR, R. T. (1970). *Convex Analysis*. Princeton Univ. Press.
[12] RUDIN, W. (1987). *Real and Complex Analysis*, 3rd ed. McGraw-Hill, New York.
[13] SCHAPIRE, R. E., FREUND, Y., BARTLETT, P. and LEE, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Statist.* **26** 1651–1686.
[14] SCHAPIRE, R. E. and SINGER, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning* **37** 297–336.
[15] STEINWART, I. (2002). Support vector machines are universally consistent. *J. Complexity* **18** 768–791.
[16] VAPNIK, V. N. (1998). *Statistical Learning Theory*. Wiley, New York.
[17] WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelpfia.
[18] ZHANG, T. (2001). A leave-one-out cross validation bound for kernel methods with applications in learning. In *Proc. 14th Annual Conference on Computational Learning Theory* 427–443. Springer, New York.

IBM T. J. WATSON RESEARCH CENTER
P.O. BOX 218
YORKTOWN HEIGHTS, NEW YORK 10598
USA
E-MAIL: tzhang@watson.ibm.com

DISCUSSION

BY PETER L. BARTLETT, MICHAEL I. JORDAN AND JON D. MCAULIFFE

University of California, Berkeley

The authors have contributed three significant papers that provide, among other insights, an understanding of the consistency of several "large margin" methods for pattern classification. In two-class classification, the aim is to find a function $f : \mathcal{X} \to \mathbb{R}$ that accurately predicts a binary response variable $Y \in \{\pm 1\}$ using the covariate $X \in \mathcal{X}$, in the sense that $R(f) = \mathbf{E}\ell(Yf(X))$, the risk of the