Journal of Machine Learning Research 9 (2008) 1823-1840 Submitted 11/06; Revised 8/07; Published 8/08

Classification with a Reject Option using a Hinge Loss

Peter L. Bartlett BARTLETT@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720-1776, USA

Marten H. Wegkamp WEGKAMP@STAT.FSU.EDU
Department of Statistics
Florida State University
Tallahassee, FL 32306-4330, USA

Editor: John Shawe-Taylor

Abstract

We consider the problem of binary classification where the classifier can, for a particular cost, choose not to classify an observation. Just as in the conventional classification problem, minimization of the sample average of the cost is a difficult optimization problem. As an alternative, we propose the optimization of a certain convex loss function ϕ, analogous to the hinge loss used in support vector machines (SVMs). Its convexity ensures that the sample average of this surrogate loss can be efficiently minimized. We study its statistical properties. We show that minimizing the expected surrogate loss—the ϕ-risk—also minimizes the risk. We also study the rate at which the ϕ-risk approaches its minimum value. We show that fast rates are possible when the conditional probability $\mathbb{P}(Y=1|X)$ is unlikely to be close to certain critical values.

Keywords: Bayes classifiers, classification, convex surrogate loss, empirical risk minimization, hinge loss, large margin classifiers, margin condition, reject option, support vector machines

1. Introduction

The aim of binary classification is to classify observations that take values in an arbitrary feature space $\mathcal{X}$ into one of two classes, labeled -1 or $+1$. A *discriminant function* $f:\mathcal{X}\to\mathbb{R}$ yields a classifier $\operatorname{sgn}(f(x))\in\{-1,+1\}$ that represents our guess of the label Y of a future observation X and we err if the margin $y\cdot f(x)<0$. The Bayes discriminant function

$$\mathbb{P}\{Y=1|X=x\}-\mathbb{P}\{Y=-1|X=x\}$$

minimizes the probability of misclassification $\mathbb{P}\{Yf(X)<0\}$. Observations x for which the conditional probability

$$\eta(x)=\mathbb{P}\{Y=+1|X=x\}$$

is close to $1/2$, are the most difficult to classify. In the extreme case where $\eta(x)=1/2$, we may just as well toss a coin to make a decision. While it is our aim to classify the majority of future observations in an automatic way, it is often appropriate to instead report a warning for those observations that are hard to classify (the ones having conditional probability $\eta(x)$ near the value $1/2$). This motivates the introduction of a *reject option* for classifiers, by allowing for a third decision, Ⓡ (*reject*), expressing doubt. For instance, in clinical trials it is important to be able to reject a tumor diagnostic classification since the consequences of misdiagnosis are severe and scientific expertise is required to make reliable determination. Although such classifiers are valuable in practice, few theoretical results are available in the statistical literature (Herbei and Wegkamp, 2006; Ripley, 1996). In the engineering community on the other hand this option is more common and empirically shown to effectively reduce the misclassification rate (Chow, 1970; Fumera and Roli, 2002, 2004; Fumera et al., 2000; Golfarelli et al., 1997; Györfi et al., 1978; Hansen et al., 1997; Landgrebe et al., 2006).

We propose to incorporate the reject option into our classification scheme by using a threshold value $0 \leq \delta < 1$ as follows. Given a discriminant function $f : \mathcal{X} \to \mathbb{R}$, we report $\mathrm{sgn}(f(x)) \in \{-1, 1\}$ if $|f(x)| > \delta$, but we withhold decision if $|f(x)| \leq \delta$ and report Ⓡ. In this note, we assume that the cost of making a wrong decision is 1 and the cost of using the reject option is $d > 0$. The appropriate risk function is then

$$L_{d,\delta}(f) = \mathbb{E}\ell_d(Yf(X)) = \mathbb{P}\{Yf(X) < -\delta\} + d\mathbb{P}\{|Yf(X)| \leq \delta\} \tag{1}$$

for the discontinuous loss

$$\ell_{d,\delta}(z) = \begin{cases} 1 & \text{if } z < -\delta, \\ d & \text{if } |z| \leq \delta, \\ 0 & \text{otherwise.} \end{cases}$$

The classifier associated with the discriminant function $f_d^*(x)$ that minimizes the risk $L_{d,\delta}(f)$ assigns $-1, 1$ or Ⓡ depending on which of $\eta(x)$, $1 - \eta(x)$ or d is smallest. Since we never reject if $d > 1/2$, we restrict ourselves to the cases $0 \leq d \leq 1/2$. The generalized Bayes discriminant function $f_d^*(x)$ is then

$$f_d^*(x) = \begin{cases} -1 & \text{if } \eta(x) < d \\ 0 & \text{if } d \leq \eta(x) \leq 1 - d \\ +1 & \text{if } \eta(x) > 1 - d \end{cases} \tag{2}$$

with risk

$$L_d^* = L_{d,\delta}(f_d^*) = \mathbb{E}\min\{\eta(X), 1 - \eta(X), d\}.$$

The case $(\delta, d) = (0, 1/2)$ reduces to the classical situation without the reject option. We emphasize that the rejection cost d should be known a priori. In a medical setting when determining whether a disease is present or absent, the reject option often leads to quantifiable costs for additional tests and perhaps in delays of treatment. The exact value of d will be dictated by such considerations. From the above we can also view d as an upper bound on the conditional probability of misclassification (given X) that is considered tolerable.

We postpone the discussion on the choice of the threshold δ until after Theorem 2.

Plug-in classification rules replace the regression function $\eta(x)$ by an estimate $\widehat{\eta}(x)$ in the formula for $f_d^*(x)$ above. It is shown by Herbei and Wegkamp (2006) that the rate of convergence of the risk (1) to the Bayes risk L_d^* of a general plug-in rule with reject option depends on how well $\widehat{\eta}(X)$ estimates $\eta(X)$ and on the behavior of $\eta(X)$ near the values d and $1 - d$. This condition on $\eta(X)$ nicely generalizes the margin condition of Tsybakov (2004) from the classical setting ($d = 1/2$) to our more general framework ($0 \leq d \leq 1/2$). The same paper derives oracle inequalities for the excess risk $L_{d,\delta}(\widehat{f}) - L_d^*$ of the (naive) empirical risk minimizer $\widehat{f}$ of $\sum_{i=1}^n \ell_{d,\delta}(Y_i f(X_i))$ based on n

independent observations (X_i,Y_i), over a class of discriminant functions $\mathcal{F}$. The results are in line with recent theoretical developments (Boucheron et al., 2006, 2005; Massart, 2007) of standard binary classification ($d = 1/2$). Despite its attractive theoretical properties, the naive empirical risk minimization method is often hard to implement. This paper addresses this pitfall by considering a convex surrogate for the loss function akin to the hinge loss that is used in SVMs. In the engineering literature, there are recently encouraging empirical results on SVMs with a reject option (Bounsiar et al., 2006; Fumera et al., 2003; Fumera and Roli, 2002; Tortorella, 2004).

The next section introduces a piecewise linear loss function $\phi_d(x)$ that generalizes the hinge loss function $\max\{0,1-x\}$ in that it allows for the reject option and $\phi_d(x) = \max\{0,1-x\}$ for $d = 1/2$. We prove that f_d^* in (2) also minimizes the risk associated with this new loss and that the excess risk $L_{d,\delta} - L_d^*$ can be bounded by $2d$ times the excess risk based on the piecewise linear loss ϕ_d if $\delta = 1/2$. Thus classifiers with small excess ϕ_d-risk automatically have small excess classification risk, providing theoretical justification of the more computationally appealing method.

In Section 3, we illustrate the computational convenience of the new loss, showing that the SVM classifier with reject option can be obtained by solving a standard convex optimization problem.

Finally, in Section 4, we show that fast rates (for instance, faster than $n^{-1/2}$) of the SVM classifier with reject option are possible under the same noise conditions on $\eta(X)$ used by Herbei and Wegkamp (2006). As a side effect, for the standard SVM (the special case of $d = 1/2$), our results imply fast rates without an assumption that $\eta(X)$ is unlikely to be near 0 and 1, a technical condition that has been imposed in the literature for that case (Blanchard et al., 2008; Tarigan and van de Geer, 2006).

2. Generalized Hinge Loss

Instead of the discontinuous loss $\ell_{d,\delta}$, we consider the convex surrogate loss

$$\phi_d(z) = \begin{cases} 1 - az & \text{if } z < 0, \\ 1 - z & \text{if } 0 \le z < 1, \\ 0 & \text{otherwise} \end{cases}$$

where $a = (1-d)/d \ge 1$. The next result states that the minimizer of the expectation of the discrete loss $\ell_{d,\delta}(z)$ and the convex loss $\phi_d(z)$ remains the same.

Proposition 1 *The Bayes discriminant function (2) minimizes the risk*

$$L_{\phi_d}(f) = \mathbb{E}\phi_d(Yf(X))$$

over all measurable $f : \mathcal{X} \to \mathbb{R}$. Furthermore,

$$dL_{\phi_d}(f_d^*) = L_{d,\delta}(f_d^*).$$

Proof Observe that

$$L_{\phi_d}(f) = \mathbb{E}\eta(X)\phi_d(f(X)) + \mathbb{E}(1-\eta(X))\phi_d(-f(X)).$$

Hence, for

$$r_{\eta,\phi_d}(z) = \eta\phi_d(z) + (1-\eta)\phi_d(-z) \tag{3}$$

it suffices to show that

$$z^* = \begin{cases} -1 & \text{if } \eta < 1/(1+a), \\ 0 & \text{if } 1/(1+a) \le \eta \le a/(1+a), \\ 1 & \text{if } \eta > a/(1+a) \end{cases}$$

minimizes $r_{\eta,\phi_d}(z)$. The function $r_{\eta,\phi_d}(z)$ can be written as

$$r_{\eta,\phi_d}(z) = \begin{cases} \eta - a\eta z & \text{if } z \le -1, \\ 1 + z(1-(1+a)\eta) & \text{if } -1 \le z \le 0, \\ 1 + z(-\eta + a(1-\eta)) & \text{if } 0 \le z \le 1, \\ z(a(1-\eta)) + (1-\eta) & \text{if } z \ge 1 \end{cases}$$

and it is now a simple exercise to verify that z^* indeed minimizes $r_{\eta,\phi_d}(z)$. Finally, since $L_{\phi_d}(f) = \mathbb{E} r_{\eta,\phi_d}(f(X))$ and

$$\begin{aligned} & \inf_z \eta\phi_d(z) + (1-\eta)\phi_d(-z) \\ = \ & \eta\phi_d(z^*) + (1-\eta)\phi_d(z^*) \\ = \ & \frac{\eta}{d}\mathbf{1}[\eta < d] + \mathbf{1}[d \le \eta \le 1-d] + \frac{1-\eta}{d}\mathbf{1}[\eta > 1-d], \end{aligned}$$

where $\mathbf{1}[A]$ denotes the indicator function of a set A, we find that

$$dL_{\phi_d}(f_d^*) = \mathbb{E}[\min(\eta(X), 1-\eta(X), d)] = L_d^*.$$

and the second claim follows as well. ■

We see that $\phi_d(z) \ge \ell_{d,\delta}(z)$ for all $z \in \mathbb{R}$ as long as $0 \le \delta \le 1-d$. Since this pointwise relation remains preserved under taking expected values, we immediately obtain $L_{d,\delta}(f) \le L_{\phi_d}(f)$. The following comparison theorem shows that a relation like this holds not only for the risks, but for the excess risks as well.

Theorem 2 *Let $0 \le d < 1/2$ and a measurable function f be fixed. For all $0 < \delta \le 1/2$, we have*

$$L_{d,\delta}(f) - L_d^* \le \frac{d}{\delta}\left(L_{\phi_d}(f) - L_{\phi_d}^*\right),$$

where $L_{\phi_d}^* = L_{\phi_d}(f_d^*)$. For $1/2 \le \delta \le 1-d$, we have

$$L_{d,\delta}(f) - L_d^* \le L_{\phi_d}(f) - L_{\phi_d}^*.$$

Finally, for $(\delta, d) = (0, 1/2)$, we have

$$L(f) - L^* \le L_\phi(f) - L_\phi^*, \tag{4}$$

where $L(f) := \mathbb{P}\{Yf(X) < 0\}$, $L^* := \mathbb{E}\min(\eta(X), 1-\eta(X))$ and $\phi(x) = \max\{0, 1-x\}$.

Remark 3 *The optimal multiplicative constant (d/δ or 1 depending on the value of δ) in front of the ϕ_d-excess risk is achieved at $\delta = 1/2$. For this choice, Theorem 2 states that*

$$L_{d,1/2}(f) - L_d^* \leq 2d\left(L_{\phi_d}(f) - L_{\phi_d}^*\right).$$

For all $d \leq \delta \leq 1-d$, the multiplicative constant in front of the ϕ_d-excess risk does not exceed 1. The choice $\delta = 1/2$ with the smallest constant $2d < 1$ is right in the middle of the interval $[d, 1-d]$. The choice $\delta = 1-d$ corresponds to the largest value of δ for which the piecewise constant function $\ell_{d,\delta}(z)$ is still majorized by the convex surrogate $\phi_d(z)$. For $\delta = d$ we will reject less frequently than for $\delta = 1-d$ and $\delta = 1/2$ can be seen as a compromise among these two extreme cases.

Inequality (4) is due to Zhang (2004).

Before we prove the theorem, we need an intermediate result. We define the functions

$$\xi(\eta) = \eta\mathbf{1}[\eta < d] + d\mathbf{1}[d \leq \eta \leq 1-d] + (1-\eta)\mathbf{1}[\eta > 1-d]$$

and

$$\begin{aligned} H(\eta) &= \inf_z \eta\phi_d(z) + (1-\eta)\phi_d(-z) \\ &= \frac{\eta}{d}\mathbf{1}[\eta < d] + \mathbf{1}[d \leq \eta \leq 1-d] + \frac{1-\eta}{d}\mathbf{1}[\eta > 1-d]. \end{aligned}$$

(We suppress their dependence on d in our notation.) Their expectations are $L_d^* = \mathbb{E}\xi(\eta(X))$ and $L_{\phi_d}^* = \mathbb{E}H(\eta(X))$, respectively. Furthermore, we define

$$\begin{aligned} H_{-1}(\eta) &= \inf_{z<-\delta}\left(\eta\phi_d(z) + (1-\eta)\phi_d(-z)\right), \\ H_{\circledR}(\eta) &= \inf_{|z|\leq\delta}\left(\eta\phi_d(z) + (1-\eta)\phi_d(-z)\right), \\ H_1(\eta) &= \inf_{z>\delta}\left(\eta\phi_d(z) + (1-\eta)\phi_d(-z)\right); \\ \xi_{-1}(\eta) &= \eta - \xi(\eta), \\ \xi_{\circledR}(\eta) &= d - \xi(\eta), \\ \xi_1(\eta) &= 1 - \eta - \xi(\eta). \end{aligned}$$

Proposition 4 *Let $0 \leq d < 1/2$.*

If $0 < \delta \leq 1/2$, then, for $b \in \{-1, 1, \circledR\}$,

$$\xi_b(\eta) \leq \frac{\delta}{d}\{H_b(\eta) - H(\eta)\}.$$

If $d \leq \delta \leq 1-d$, then, for $b \in \{-1, 1, \circledR\}$,

$$\xi_b(\eta) \leq H_b(\eta) - H(\eta).$$

If $(\delta, d) = (0, 1/2)$, then, for $b \in \{-1, 1, \circledR\}$,

$$\xi_b(\eta) \leq H_b(\eta) - H(\eta).$$

The proof is in the appendix.

Proof [Proof of Theorem 2] Recall that $L_{d,\delta}(f) = \mathrm{P}(\eta\mathbf{1}[f<-\delta]+d\mathbf{1}[-\delta\le f\le\delta]+(1-\eta)\mathbf{1}[f>\delta])$ and $L_{\phi_d}(f) = \mathrm{P}r_{\eta,\phi_d}(f)$ with r_{η,ϕ_d} defined in the proof of Proposition 1. Here P is the probability measure of X and $\mathrm{P}g = \int g d\mathrm{P}$ for any P-integrable g. Assume $0<\delta\le 1/2$ and $0\le d<1/2$. Define $\psi(x) = x\delta/d$. By linearity of ψ, we have for any measurable function f,

$$\begin{aligned}\psi(L_{d,\delta}(f)-L_d^*) &= \mathrm{P}\left(\mathbf{1}[f<-\delta]\,\psi(\xi_{-1}(\eta))+\mathbf{1}[-\delta\le f\le\delta]\,\psi(\xi_{®}(\eta))\right.\\ &\qquad\left.+\mathbf{1}[f>\delta]\,\psi(\xi_1(\eta))\right).\end{aligned}$$

Invoke now Proposition 4 to deduce

$$\begin{aligned}\psi(L_{d,\delta}(f)-L_d^*) &\le \mathrm{P}\left(\mathbf{1}[f<-\delta]\,[H_{-1}(\eta)-H(\eta)]+\mathbf{1}[-\delta\le f\le\delta]\,[H_{®}(\eta)-H(\eta)]\right.\\ &\qquad\left.+\mathbf{1}[f>\delta]\,[H_1(\eta)-H(\eta)]\right)\\ &\le \mathrm{P}\left\{r_{\eta,\phi_d}(f)-H(\eta)\right\}\end{aligned}$$

and conclude the proof by observing that the term on the right of the previous inequality equals $L_{\phi_d}(f)-L_{\phi_d}^*$.

For the case $(\delta,d)=(0,1/2)$ and the case (δ,d) with $d\le\delta\le 1-d$ and $0\le d<1/2$, take $\psi(x)=x$. ■

3. SVM Classifiers with Reject Option

In this section, we consider an SVM-like classifier for classification with a reject option, and show that it can be obtained by solving a quadratically constrained quadratic program (QCQP).

Let $K:\mathcal{X}^2\to\mathbb{R}$ be the kernel of a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, and let $\|f\|$ be the norm of f in $\mathcal{H}$. The SVM classifier with reject option is the minimizer of the empirical ϕ_d-risk subject to a constraint on the RKHS norm.[1] The following theorem shows that this classifier is the solution to a QCQP, that is, it is the minimizer of a convex quadratic criterion on a convex subset of Euclidean space defined by quadratic inequalities. Thus, the classifier can be found efficiently using general-purpose algorithms.

Theorem 5 *For any $x_1,\ldots,x_n\in\mathcal{X}$ and $y_1,\ldots,y_n\in\{-1,1\}$, let $\widehat{f}\in\mathcal{H}$ be the solution to*

$$\begin{aligned}&\textit{minimize} && f\mapsto\sum_{i=1}^n\phi_d\left(y_i f(x_i)\right)\\ &\textit{such that} && \|f\|^2\le r^2,\end{aligned}$$

where $r>0$. Then we can represent $\widehat{f}$ as the finite sum

$$\widehat{f}(x)=\sum_{i=1}^n\widehat{\alpha}_i K(x_i,x),$$

[1] Notice that we parameterize the optimization problem in terms of the constraint on the RKHS norm, rather than in terms of its Lagrange multiplier, which is more standard. The regularization path—the set of solutions to these problems as the parameter of the optimization problem varies—is identical.

where $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_n$ is the solution to the following QCQP.

$$\begin{array}{ll}
\displaystyle\min_{\alpha_i,\xi_i,\gamma_i} & \displaystyle\frac{1}{n}\sum_{i=1}^{n}\left(\xi_i + \frac{1-2d}{d}\gamma_i\right) \\
\textit{such that} & \displaystyle\sum_{i,j}\alpha_i\alpha_j K(x_i,x_j) \le r^2 \\
& \xi_i \ge 0, \qquad \gamma_i \ge 0, \\
& \displaystyle\xi_i \ge 1 - y_i\sum_{j=1}^{n}\alpha_j K(x_i,x_j), \\
& \displaystyle\gamma_i \ge -y_i\sum_{j=1}^{n}\alpha_j K(x_i,x_j) \qquad\qquad \textit{for } i = 1,\ldots,n.
\end{array}$$

Proof The fact that $\widehat{f}$ can be represented as a finite sum over the kernel basis functions is a standard argument (Kimeldorf and Wahba, 1971; Cox and O'Sullivan, 1990). It follows from Pythagoras' theorem in Hilbert space: the squared RKHS norm can be split into the squared norm of the component in the space spanned by the kernel basis functions $x \mapsto K(x_i,x)$ and that of the component in the orthogonal subspace. Since the cost function depends on f only at the points x_i, and the reproducing property $f(x_i) = \langle K(x_i,\cdot), f\rangle$ shows that these values depend only on the component of f in the space spanned by the kernel basis functions, the orthogonal subspace only makes the constraint harder to satisfy, but does not affect the cost function. Thus, a minimizing $\widehat{f}$ can be represented in terms of the solution $\widehat{\alpha}$ to the minimization

$$\begin{array}{ll}
\displaystyle\min_{\alpha_1,\ldots,\alpha_n} & \displaystyle\frac{1}{n}\sum_{i=1}^{n}\phi_d\left(y_i\sum_{j=1}^{n}\alpha_j K(x_i,x_j)\right) \\
\text{such that} & \displaystyle\sum\sum_{1\le i,j\le n}\alpha_i\alpha_j K(x_i,x_j) \le r^2.
\end{array}$$

But then it is easy to see that we can decompose ϕ_d as

$$\phi_d(\beta) = \max\{0, 1-\beta\} + \frac{1-2d}{d}\max\{0, -\beta\}.$$

Parameterizing ϕ_d using the slack variables

$$\xi_i = \max\{0, 1 - y_i f(x_i)\}, \qquad \gamma_i = \max\{0, -y_i f(x_i)\}$$

gives the QCQP. ■

4. Tsybakov's Margin Condition, Bernstein Classes, and Fast Rates

In this section, we consider methods that choose the function $\widehat{f}$ from some class $\mathcal{F}$ so as to minimize the empirical ϕ_d-risk

$$\widehat{L}_{\phi_d}(f) = \frac{1}{n}\sum_{i=1}^{n}\phi_d(Y_i f(X_i)).$$

For instance, to analyze the SVM classifier with reject option, we could consider classes $\mathcal{F}_n = \{f \in \mathcal{H} : \|f\| \le c_n\}$ for some sequence of constants c_n. We are interested in bounds on the excess ϕ_d-risk, that is, the difference between the ϕ_d-risk of $\widehat{f}$ and the minimal ϕ_d-risk over all measurable functions, of the form

$$\mathbb{E}L_{\phi_d}(\widehat{f}) - L^*_{\phi_d} \le 2 \inf_{f\in\mathcal{F}} \left(L_{\phi_d}(f) - L^*_{\phi_d}\right) + \varepsilon_n.$$

Such bounds can be combined with an assumption on the rate of decrease of the approximation error $\inf_{f\in\mathcal{F}_n}\left(L_{\phi_d}(f) - L^*_{\phi_d}\right)$ for a sequence of classes $\mathcal{F}_n$ used by a method of sieves, and thus provide bounds on the rate of convergence of risk $L_{d,\delta}(\widehat{f})$ to the optimal Bayes risk L^*_d.

For many binary classification methods (including empirical risk minimization, plug-in estimates, and minimization of the sample average of a suitable convex loss), the estimation error term ε_n approaches zero at a faster rate when the conditional probability $\eta(X)$ is unlikely to be close to the critical value of $1/2$ (Audibert and Tsybakov, 2007; Bartlett et al., 2006; Blanchard et al., 2008; Steinwart and Scovel, 2007; Tarigan and van de Geer, 2006; Tsybakov, 2004). For plug-in rules, Herbei and Wegkamp (2006) showed an analogous result for classification with a reject option, where the corresponding condition concerns the probability that $\eta(X)$ is close to the critical values of d and $1-d$. In this section, we prove a bound on the excess ϕ_d-risk of $\widehat{f}$ that converges rapidly when a condition of this kind applies. We begin with a precise statement of the condition. For $d = 1/2$, it is equivalent to the margin condition of Tsybakov (2004).

Definition 6 *We say that η satisfies the margin condition at d with exponent $\alpha > 0$ if there is a $c \ge 1$ such that for all $t > 0$,*

$$\mathbb{P}\{|\eta(X) - d| \le t\} \le ct^{\alpha} \text{ and } \mathbb{P}\{|\eta(X) - (1-d)| \le t\} \le ct^{\alpha}.$$

The reason that conditions of this kind allow fast rates is related to the variance of the excess ϕ_d-loss,

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

where f^*_d minimizes the ϕ_d-risk. Notice that the expectation of g_f is precisely the excess risk of f, $\mathbb{E}g_f(X,Y) = L_{\phi_d}(f) - L^*_{\phi_d}$. We will show that when η satisfies the margin condition at d with exponent α, the variance of each g_f is bounded in terms of its expectation, and thus approaches zero as the ϕ-risk of f approaches the minimal value. Classes for which this occurs are called Bernstein classes.

Definition 7 *We say that $\mathcal{G} \subset L_2(\mathrm{P})$ is a (β, B)-Bernstein class with respect to the probability measure P $(0 < \beta \le 1, B \ge 1)$ if every $g \in \mathcal{G}$ satisfies*

$$\mathrm{P}g^2 \le B(\mathrm{P}g)^{\beta}.$$

We say that $\mathcal{G}$ has a Bernstein exponent β with respect to P if there exists a constant B for which $\mathcal{G}$ is a (β, B)-Bernstein class.

Lemma 8 *If η satisfies the margin condition at d with exponent α, then for any class $\mathcal{F}$ of measurable uniformly bounded functions, the class $\mathcal{G} = \{g_f : f \in \mathcal{F}\}$ has a Bernstein exponent $\beta = \alpha/(1+\alpha)$.*

The result relies on the following two lemmas. The first shows that the excess ϕ_d-risk is at least linear in a certain pseudo-norm of the difference between f and f_d^*. It is similar to the $L_1(\mathrm{P})$ norm, but it penalizes f less for large excursions that have little impact on the ϕ_d-risk. For example, if $\eta(x)=1$, then the conditional ϕ_d-risk is zero even if $f(x)$ takes a large positive value. For $\eta\in[0,1]$, define

$$\rho_\eta(f,f_d^*)=\begin{cases}\eta|f-f_d^*| & \text{if } \eta<d \text{ and } f<-1,\\ (1-\eta)|f-f_d^*| & \text{if } \eta>1-d \text{ and } f>1,\\ |f-f_d^*| & \text{otherwise,}\end{cases}$$

and recall the definition of the conditional ϕ_d-risk in (3).

Lemma 9 *For* $\eta\in[0,1]$,

$$d\left(r_{\eta,\phi_d}(f)-r_{\eta,\phi_d}(f_d^*)\right)\geq(|\eta-d|\wedge|\eta-(1-d)|)\,\rho_\eta(f,f_d^*).$$

Proof Since r_{η,ϕ_d} is convex,

$$r_{\eta,\phi_d}(f)\geq r_{\eta,\phi_d}(f_d^*)+g(f-f_d^*)$$

for any g in the subgradient of $r_{\eta,\phi_d}(f)$ at f_d^*. In our case, r_{η,ϕ_d} is piecewise linear, with four pieces, and the subgradients include

$$\begin{array}{ll}\eta\frac{1-d}{d} & \text{at } f_d^*=-1,\\ |\eta-d|\frac{1}{d} & \text{at } f_d^*=-1,0,\\ |1-\eta-d|\frac{1}{d} & \text{at } f_d^*=0,1,\\ (1-\eta)\frac{1-d}{d} & \text{at } f_d^*=1.\end{array}$$

Thus, we have

$$\begin{aligned} & d(r_{\eta,\phi_d}(f)-r_{\eta,\phi_d}(f_d^*))\\ & \geq\begin{cases}\eta(1-d)|f-f_d^*| & \text{if } \eta<d \text{ and } f<-1,\\ |\eta-d||f-f_d^*| & \text{if } \eta<d \text{ and } f>-1,\\ (|\eta-d|\wedge|1-\eta-d|)\,|f-f_d^*| & \text{if } d\leq\eta\leq 1-d,\\ |1-\eta-d||f-f_d^*| & \text{if } \eta>1-d \text{ and } f<1,\\ (1-\eta)(1-d)|f-f_d^*| & \text{if } \eta>1-d,\ f>1.\end{cases}\\ & =\begin{cases}(1-d)\rho_\eta(f,f_d^*) & \text{if } \eta<d \text{ and } f<-1,\\ |\eta-d|\rho_\eta(f,f_d^*) & \text{if } \eta<d \text{ and } f>-1,\\ (|\eta-d|\wedge|1-\eta-d|)\,\rho_\eta(f,f_d^*) & \text{if } d\leq\eta\leq 1-d,\\ |1-\eta-d|\rho_\eta(f,f_d^*) & \text{if } \eta>1-d \text{ and } f<1,\\ (1-d)\rho_\eta(f,f_d^*) & \text{if } \eta>1-d,\ f>1.\end{cases}\\ & \geq(|\eta-d|\wedge|1-\eta-d|)\,\rho_\eta(f,f_d^*).\end{aligned}$$

■

We shall also use the following inequalities.

Lemma 10 *If $\|f\|_\infty = B$, then for $\eta \in [0,1]$,*

$$\rho_\eta(f, f_d^*) \le |f - f_d^*|,$$

and

$$\eta\,|\phi_d(f) - \phi_d(f_d^*)|^2 + (1-\eta)\,|\phi_d(-f) - \phi_d(-f_d^*)|^2 \le \left(\frac{1-d}{d}\right)^2 (B+1)\rho_\eta(f, f_d^*).$$

Proof The first inequality is immediate from the definition of ρ_η. To see the second, use the fact that ϕ_d is flat to the right of 1 to notice that

$$\begin{aligned} &\eta\,|\phi_d(f) - \phi_d(f_d^*)|^2 + (1-\eta)\,|\phi_d(-f) - \phi_d(-f_d^*)|^2 \\ &= \begin{cases} \eta\left|\phi_d(f) - \phi_d(f_d^*)\right|^2 & \text{if } \eta < d \text{ and } f < -1, \\ (1-\eta)\left|\phi_d(-f) - \phi_d(-f_d^*)\right|^2 & \text{if } \eta > 1-d \text{ and } f > 1. \end{cases} \end{aligned}$$

Since ϕ_d has Lipschitz constant $a = (1-d)/d$, this implies

$$\begin{aligned} &\eta\,|\phi_d(f) - \phi_d(f_d^*)|^2 + (1-\eta)\,|\phi_d(-f) - \phi_d(-f_d^*)|^2 \\ &\le \begin{cases} \eta a^2 |f - f_d^*|^2 & \text{if } \eta < d \text{ and } f < -1, \\ (1-\eta) a^2 |f - f_d^*|^2 & \text{if } \eta > 1-d \text{ and } f > 1, \\ a^2 |f - f_d^*|^2 & \text{otherwise} \end{cases} \\ &\le a^2(1+B)\rho_\eta(f, f_d^*), \end{aligned}$$

where the last inequality uses the fact that $|f - f_d^*| \le B+1$. ■

Proof [Proof of Lemma 8] By Lemma 9, we have

$$L_{\phi_d}(f) - L_{\phi_d}^* \ge d^{-1}\mathbb{E}\rho_\eta(f, f_d^*)\left(|\eta - (1-d)|I_{E_-} + |\eta - d|I_{E_+}\right),$$

with

$$E_- = \{|\eta - (1-d)| \le |\eta - d|\}, \qquad E_+ = \{|\eta - (1-d)| > |\eta - d|\}.$$

Using the assumption on η, there is an $A \ge 1$ such that for all $t > 0$

$$\mathbb{P}\{|\eta(X) - d| \le t\} \le At^\alpha \text{ and } \mathbb{P}\{\eta(X) - (1-d)| \le t\} \le At^\alpha.$$

Thus, for any set E,

$$\begin{aligned} P\rho_\eta(f, f_d^*)|\eta - (1-d)|I_E &\ge tP\rho_\eta(f, f_d^*)I_{\{|\eta-(1-d)|\ge t\}}I_E \\ &= tP\rho_\eta(f, f_d^*)I_E - tP\rho_\eta(f, f_d^*)I_{\{|\eta-(1-d)|<t\}}I_E \\ &\ge t\{P\rho_\eta(f, f_d^*)I_E - (B+1)At^\alpha\}, \end{aligned}$$

where B is such that $|f| \le B$ and hence $\rho_\eta(f, f_d^*) \le |f - f_d^*| \le B+1$. Similarly,

$$P\rho_\eta(f, f_d^*)|\eta - d|I_E \ge t\{P\rho_\eta(f, f_d^*)I_E - (B+1)At^\alpha\},$$

and we obtain

$$\begin{aligned} L_{\phi_d}(f) - L^*_{\phi_d} &\geq d^{-1}t\left(\mathrm{P}\rho_\eta(f,f^*_d)I_{E_+\cup E_-} - 2(B+1)At^\alpha\right) \\ &= d^{-1}t\left(\mathrm{P}\rho_\eta(f,f^*_d) - 2(B+1)At^\alpha\right). \end{aligned}$$

Choose

$$t = \left(\frac{\mathrm{P}\rho_\eta(f,f^*_d)}{4(B+1)A}\right)^{1/\alpha},$$

in the expression above, and we obtain

$$\mathbb{E}g_f(X,Y) = L_{\phi_d}(f) - L^*_{\phi_d} \geq \frac{1}{2d(4(B+1)A)^{1/\alpha}}\left(\mathrm{P}\rho_\eta(f,f^*_d)\right)^{(1+\alpha)/\alpha},$$

and so

$$\mathrm{P}\rho_\eta(f,f^*_d) \leq \left\{2d(4(B+1)A)^{1/\alpha}\right\}^{\alpha/(\alpha+1)}\left\{\mathbb{E}g_f(X,Y)\right\}^{\alpha/(1+\alpha)}.$$

In addition, by Lemma 10,

$$\begin{aligned} \mathbb{E}\{g_f(X,Y)\}^2 &= \mathbb{E}\mathbb{E}[\{g_f(X,Y)\}^2|X] \\ &= \mathrm{P}\left(\eta|\phi_d(f) - \phi_d(f^*_d)|^2 + (1-\eta)|\phi_d(-f) - \phi_d(-f^*_d)|^2\right) \\ &\leq (B+1)\left(\frac{1-d}{d}\right)^2 \mathrm{P}\rho_\eta(f,f^*_d). \end{aligned}$$

Combining these two inequalities shows that

$$\mathbb{E}\{g_f(X,Y)\}^2 \leq (B+1)\left(\frac{1-d}{d}\right)^2\left(2d(4A(B+1))^{1/\alpha}\right)^{\alpha/(\alpha+1)}\left(\mathbb{E}g_f(X,Y)\right)^{\alpha/(1+\alpha)}.$$

■

Remark 11 *Specialized to the case $(\delta,d) = (0,1/2)$, we note that Lemma 8 removes unnecessary technical restrictions on $\eta(X)$ near 0 and 1, imposed by Blanchard et al. (2008) and Tarigan and van de Geer (2006). This is consistent with results of Steinwart and Scovel (2007) on SVMs with Gaussian kernels.*

Lemma 8 provides the main ingredient for establishing fast rates of minimizers $\widehat{f}_d$ of the empirical risk $\widehat{L}_{\phi_d}(f)$.

In the theorem, we use the notation $N(\varepsilon, L_\infty, \mathcal{F})$ to denote the ε-covering number of $\mathcal{F}$ in L_∞, that is, the smallest number of closed ε-balls in L_∞ needed to cover $\mathcal{F}$. The countability assumption means that measurability is not an issue. It can be replaced by other mild sufficient conditions.

Theorem 12 *If η satisfies the margin condition at d with exponent α, $\mathcal{F}$ is a countable class of functions $f : \mathcal{X} \to \mathbb{R}$ satisfying $\|f\|_\infty \leq B$, and $\mathcal{F}$ satisfies*

$$\log N(\varepsilon, L_\infty, \mathcal{F}) \leq C\varepsilon^{-p}$$

for all $\varepsilon > 0$ *and some* $0 \le p \le 2$*, then there exists a constant* C' *independent of* n*, such that*

$$\mathbb{E}L_{\phi_d}(\widehat{f}_d) - L^*_{\phi_d} \le 2 \inf_{f \in \mathcal{F}} \left(L_{\phi_d}(f) - L^*_{\phi_d} \right) + C' n^{-\frac{1+\alpha}{2+p+\alpha+p\alpha}},$$

where $\widehat{f}_d = \arg\min_{f \in \mathcal{F}} \widehat{L}_{\phi_d}(f)$.

Proof We use the notation $\mathrm{P}g_f = \mathbb{E}g_f(X,Y)$ and

$$\mathbb{P}_n g_f = \frac{1}{n} \sum_{i=1}^n g_f(X_i, Y_i).$$

By definition of $\widehat{f}_d$, we have

$$\begin{aligned} L_{\phi_d}(\widehat{f}_d) - L^*_{\phi_d} &= \mathrm{P}g_{\widehat{f}_d} \\ &= 2\mathbb{P}_n g_{\widehat{f}_d} + (\mathrm{P} - 2\mathbb{P}_n) g_{\widehat{f}_d} \\ &\le 2 \inf_{f \in \mathcal{F}} \mathbb{P}_n g_f + \sup_{f \in \mathcal{F}} (\mathrm{P} - 2\mathbb{P}_n) g_f. \end{aligned}$$

Taking expected values on both sides, yields,

$$\mathbb{E}L_{\phi_d}(\widehat{f}_d) - L^*_{\phi_d} \le 2 \inf_{f \in \mathcal{F}} \left(L_{\phi_d}(f) - L^*_{\phi_d} \right) + \mathbb{E}\left[\sup_{f \in \mathcal{F}} (\mathrm{P} - 2\mathbb{P}_n) g_f \right].$$

Since $|g_f - g_{f'}| \le |f - f'|(1-d)/d$, it follows that

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} (\mathrm{P} - 2\mathbb{P}_n) g_f \right] \le \frac{1-d}{d} \varepsilon_n + \frac{1-d}{d} B \mathbb{P} \left\{ \sup_{f \in \mathcal{F}_n} (\mathrm{P} - 2\mathbb{P}_n) g_f \ge \varepsilon_n \right\},$$

where $\mathcal{F}_n$ is a minimal ε_n-covering net of $\mathcal{F}$ with

$$\varepsilon_n = M n^{-(1+\alpha)/(2+p+\alpha+p\alpha)}$$

for some constant M to be selected later. The union bound and Bernstein's exponential inequality for the tail probability of sums of bounded random variables in conjunction with Lemma 8, yield

$$\begin{aligned} \mathbb{P}\left\{ \sup_{f \in \mathcal{F}_n} (\mathrm{P} - 2\mathbb{P}_n) g_f \ge \varepsilon_n \right\} &\le \sum_{f \in \mathcal{F}_n} \mathbb{P}\left\{ (\mathrm{P} - \mathbb{P}_n) g_f \ge \frac{1}{2} (\mathrm{P}g_f + \varepsilon_n) \right\} \\ &\le |\mathcal{F}_n| \max_{f \in \mathcal{F}_n} \exp\left(-\frac{n}{8} \frac{(\varepsilon_n + \mathrm{P}g_f)^2}{\mathrm{P}g_f^2 + B(\varepsilon_n + \mathrm{P}g_f)/6} \right) \\ &\le \exp(C\varepsilon_n^{-p} - cn\varepsilon_n^{2-\beta}) \end{aligned}$$

with $0 \le \beta = \alpha/(1+\alpha) \le 1$ and some $c > 0$ independent of n. Conclude the proof by noting that

$$\exp(C\varepsilon_n^{-p} - cn\varepsilon_n^{2-\beta}) = \exp\left(-\frac{c}{2} n \varepsilon_n^{2-\beta} \right),$$

and by choosing the constant M in ε_n such that $C\varepsilon_n^{-p} = cn\varepsilon_n^{2-\beta}/2$ and $\exp(-n\varepsilon_n^{2-\beta}) = o(\varepsilon_n)$. ■

Remark 13 *The constant 2 in front of the minimal excess risk on the right could be made closer to* 1*, at the expense of increasing* C'*.*

Theorem 12 discusses minimizers of the empirical risk $\widehat{L}_{\phi_d}$ *over classes* $\mathcal{F}$ *of uniformly bounded functions. The analysis of SVMs that minimize* $\widehat{L}_{\phi_d}$ *plus a regularization term requires more work.*

Remark 14 *Consider for simplicity the case* $\mathcal{F}$ *is finite (*$p=0$*). Then, if the margin condition holds for* $\alpha=+\infty$*, we obtain from the proof of Theorem 12 rates of convergence of order* $\log|\mathcal{F}|/n$*. If* $\alpha=0$*, we in fact impose no restriction on* $\eta(X)$ *at all, and the rate equals* $(\log|\mathcal{F}|/n)^{1/2}$*.*

Remark 15 *The entropy condition is satisfied for many classes. For instance, Kolmogorov and Tichomirov (1961) prove the following result for Sobolev spaces with parameter* β*. Let* $\mathcal{X}$ *be a bounded, convex subset of* $\mathbb{R}^d$ *and for every* $k=(k_1,\ldots,k_d)\in\mathbb{N}^d$*, define the differential operator* D^k *by*

$$D^k=\frac{\partial^{k_1+\ldots+k_d}}{\partial x_1^{k_1}\ldots\partial x_d^{k_d}}.$$

Let $\mathcal{F}=\mathcal{F}(\beta,c_1,c_2)$ *be the class of real valued, continuous functions* f *on* $\mathcal{X}$ *with uniformly bounded partial derivatives of order* $k\le\lfloor\beta\rfloor$ *(the greatest integer smaller than* β*),*

$$\max_{k_1+\ldots+k_n\le\lfloor\beta\rfloor}\max_{x\in\mathcal{X}}\left|D^k f(x)\right|\le c_1,$$

and which highest partial derivatives are Lipschitz of order $\beta-\lfloor\beta\rfloor$*,*

$$\max_{k_1+\ldots+k_n=\lfloor\beta\rfloor}\max_{x,y\in\mathcal{X},\ x\neq y}\frac{|D^k f(x)-D^k f(y)|}{\|x-y\|^{\beta-\lfloor\beta\rfloor}}\le c_2.$$

The constants c_1 *and* c_2 *are independent of* f*. Such classes have covering numbers (Kolmogorov and Tichomirov, 1961; van der Vaart and Wellner, 1996)*

$$\log N(\varepsilon,L_\infty,\mathcal{F})\le C_d\left(\frac{1}{\varepsilon}\right)^{d/\beta},$$

for every $\varepsilon>0$ *and some constant* C_d *depending on the dimension d and the constants* c_1 *and* c_2*, but not on* ε*. Applying the theorem with* $p=d/\beta$*, we obtain rates between* $n^{-\beta/(2\beta+d)}$ *(for* $\alpha=0$*) and* $n^{-\beta/(d+\beta)}$ *(for* $\alpha=+\infty$*).*

Another example is the case where $\mathcal{F}$ *is a subset of a RKHS. For instance, let* $\mathcal{H}$ *be the RKHS corresponding to the Gaussian kernel* $K(x,y)=\exp(-\|x-y\|^2/\sigma^2)$ *and let* $\|f\|$ *be the norm of f in* $\mathcal{H}$*. For* $\mathcal{F}=\mathcal{F}_R=\{f\in\mathcal{H}:\ \|f\|\le R\}$*, Zhou (2003) proves that, for* $\mathcal{X}=[0,1]^d$*, fixed R and fixed scale parameter* σ*, the entropy bound*

$$\log N(\varepsilon,L_\infty,\mathcal{F})\le C_d\log^{d+1}\left(\frac{R}{\varepsilon}\right)$$

for some $C_d<\infty$ *and the rates of convergence range between* $\sqrt{\log^{d+1}(n)/n}$ *(*$\alpha=0$*) and* $\log^{d+1}(n)/n$ *(*$\alpha=\infty$*). See also the results of Guo et al. (2002).*

Acknowledgments

The authors gratefully acknowledge the support of NSF, the first author through grant DMS-0434383 and the second author through grant DMS-0706829.

Appendix A. Proof of Proposition 4

First we compute

$$\begin{aligned}
\inf_{z\le -1} r_{\eta,\phi_d}(z) &= \frac{\eta}{d}, \\
\inf_{-1\le z\le -\delta} r_{\eta,\phi_d}(z) &= \frac{\eta}{d}\mathbf{1}[\eta \le d] + \left(\frac{\delta}{d}\eta + 1 - \delta\right)\mathbf{1}[\eta > d] \\
\inf_{-\delta\le z\le 0} r_{\eta,\phi_d}(z) &= \mathbf{1}[\eta \ge d] + \left(\frac{\delta}{d}\eta + 1 - \delta\right)\mathbf{1}[\eta < d] \\
\inf_{0\le z\le \delta} r_{\eta,\phi_d}(z) &= \mathbf{1}[\eta \le 1-d] + \left(1 + \frac{\delta}{d} - \delta - \frac{\delta}{d}\eta\right)\mathbf{1}[\eta > 1-d] \\
\inf_{\delta\le z\le 1} r_{\eta,\phi_d}(z) &= \frac{1-\eta}{d}\mathbf{1}[\eta > 1-d] + \left(1 + \frac{\delta}{d} - \delta - \frac{\delta}{d}\eta\right)\mathbf{1}[\eta \le 1-d] \\
\inf_{z\ge 1} r_{\eta,\phi_d}(z) &= \frac{1-\eta}{d}
\end{aligned}$$

It is now easy to verify that

$$\begin{aligned}
H_{-1}(\eta) &= \inf_{z<-\delta} \eta\phi_d(z) + (1-\eta)\phi_d(-z) \\
&= \frac{\eta}{d}\mathbf{1}[\eta < d] + \left(\frac{\delta}{d}\eta + 1 - \delta\right)\mathbf{1}[\eta \ge d]
\end{aligned}$$

so that

$$\begin{aligned}
&H_{-1}(\eta) - H(\eta) = \\
&\left(\frac{\delta}{d}\eta - \delta\right)\mathbf{1}[d \le \eta \le 1-d] + \left(\frac{1+\delta}{d}\eta + 1 - \delta - \frac{1}{d}\right)\mathbf{1}[\eta > 1-d]
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\xi_{-1}(\eta) &= \eta - \xi(\eta) \\
&= (\eta - d)\mathbf{1}[d \le \eta \le 1-d] + (2\eta - 1)\mathbf{1}[\eta > 1-d]
\end{aligned}$$

and we see that

$$\frac{\delta}{d}\xi_{-1}(\eta) \le H_{-1}(\eta) - H(\eta)$$

for all $0 < \delta \le 1$. Next, we compute

$$\begin{aligned}
H_{\circledR}(\eta) &= \inf_{|z|\le\delta} \eta\phi_d(z) + (1-\eta)\phi_d(-z) \\
&= \left(1 - \delta + \frac{\delta}{d}\eta\right)\mathbf{1}[\eta < d] + \mathbf{1}[d \le \eta \le 1-d] \\
&\qquad + \left(1 - \delta + \frac{\delta}{d} - \frac{\delta}{d}\eta\right)\mathbf{1}[\eta > 1-d]
\end{aligned}$$

and

$$\begin{aligned} H_{®}(\eta)-H(\eta) &= \left(1-\delta-\frac{1-\delta}{d}\eta\right)\mathbf{1}[\eta<d] \\ &\quad +\left(1-\delta-\frac{1-\delta}{d}+\frac{1-\delta}{d}\eta\right)\mathbf{1}[\eta>1-d]. \end{aligned}$$

Since

$$\begin{aligned} \xi_{®}(\eta) &= d-\xi(\eta) \\ &= (d-\eta)\mathbf{1}[\eta<d]+(d-1+\eta)\mathbf{1}[\eta>1-d] \end{aligned}$$

we find that

$$\frac{\delta}{d}\xi_{®}(\eta)\le H_{®}(\eta)-H(\eta)$$

provided $0<\delta\le 1/2$. Finally, we find that

$$\begin{aligned} H_1(\eta) &= \inf_{z>\delta}\eta\phi_d(z)+(1-\eta)\phi_d(-z) \\ &= \frac{1-\eta}{d}\mathbf{1}[\eta>1-d]+\left(\frac{\delta}{d}+1-\delta-\frac{\delta}{d}\eta\right)\mathbf{1}[\eta\le 1-d] \end{aligned}$$

and consequently

$$\begin{aligned} H_1(\eta)-H(\eta) &= \left(1-\delta+\frac{\delta}{d}-\frac{\delta}{d}\eta-\frac{\eta}{d}\right)\mathbf{1}[\eta<d] \\ &\quad +\left(\frac{\delta}{d}-\delta-\frac{\delta}{d}\eta\right)\mathbf{1}[d\le\eta\le 1-d]. \end{aligned}$$

Now,

$$\begin{aligned} \xi_1(\eta) &= 1-\eta-\xi(\eta) \\ &= (1-2\eta)\mathbf{1}[\eta<d]+(1-\eta-d)\mathbf{1}[d\le\eta\le 1-d], \end{aligned}$$

and we find that

$$\frac{\delta}{d}\xi_1(\eta)\le H_1(\eta)-H(\eta)$$

provided $0<\delta\le 1$.

We now verify the second claim of Proposition 4. Assume that $d\le\delta\le 1-d$.

First we consider the case $\eta<d$. Then

$\xi_{-1}(\eta)\le H_{-1}(\eta)-H(\eta)$ holds trivially.

$\xi_{®}(\eta)\le H_{®}(\eta)-H(\eta)\iff(1-\delta-d)\eta\le(1-\delta-d)d$. As $\eta\le d$, we need that $\delta\le 1-d$.

$\xi_1(\eta)\le H_1(\eta)-H(\eta)\iff(1+\delta-2d)\eta\le\delta(1-d)$. As $\eta\le d$, we need that $(1+\delta-2d)d\le\delta(1-d)$, equivalently, $(\delta-d)(1-2d)\ge 0$.

Next, if $d\le\eta\le 1-d$, we see that

$\xi_{-1}(\eta) \le H_{-1}(\eta) - H(\eta) \iff (\delta - d)\eta \ge d(\delta - d).$

$\xi_{®}(\eta) \le H_{®}(\eta) - H(\eta)$ holds trivially.

$\xi_1(\eta) \le H_1(\eta) - H(\eta) \iff (\delta - d)\eta \le (1-d)(\delta - d).$

Finally, if $\eta > 1-d$, we find that

$\xi_{-1}(\eta) \le H_{-1}(\eta) - H(\eta) \iff (1+\delta-2d)\eta \ge (1+d\delta-2d)$. For $\eta \ge 1-d$ this holds provided $(1+\delta-2d)(1-d) \ge (1+d\delta-2d) \iff (\delta-d)(1-2d) \ge 0$.

$\xi_{®}(\eta) \le H_{®}(\eta) - H(\eta) \iff (1-\delta-d)\eta \ge (1-d)(1-\delta-d).$

$\xi_1(\eta) \le H_1(\eta) - H(\eta)$ holds trivially.

This concludes the proof of the second claim, since $d \le \delta \le 1-d$. The last claim for the case $(\delta, d) = (0, 1/2)$ follows as well from the preceding calculations.

References

J. Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers under margin conditions. *Annals of Statistics*, 35(2):608–633, 2007.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Annals of Statistics*, 36(2):489–531, 2008.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

S. Boucheron, O. Bousquet, and G. Lugosi. Introduction to statistical learning theory. In O. Bousquet, U. von Luxburg, and G. Rätsch, editors, *Advanced Lectures in Machine Learning*, pages 169–207. Springer, 2006.

A. Bounsiar, E. Grall, and P. Beauseroy. A kernel based rejection method for supervised classification. *International Journal of Computational Intelligence*, 3(4):312–321, 2006.

C.K. Chow. On optimum error and reject trade-off. *IEEE Transactions on Information Theory*, 16: 41–46, 1970.

D. Cox and F. O'Sullivan. Asymptotic analysis of penalized likelihood and related estimators. *Annals of Statistics*, 18:1676–1695, 1990.

G. Fumera and F. Roli. Suppport vector machines with embedded reject option. In S. Lee and A. Verri, editors, *Pattern Recognition with Support Vector Machines*, volume 2388, pages 68–82. Springer, 2002.

G. Fumera and F. Roli. Analysis of error-reject trade-off in linearly combined multiple classifiers. *Pattern Recognition*, 37:1245–1265, 2004.

G. Fumera, F. Roli, and G. Giacinto. Reject option with multiple thresholds. *Pattern Recognition*, 33:2099–2101, 2000.

G. Fumera, I. Pillai, and F. Roli. Classification with reject option in text categorisation systems. In *Proceedings of the 12th International Conference on Image Analysis and Processing*, pages 582–587. IEEE Computer Society, 2003.

M. Golfarelli, D. Maio, and D. Maltoni. On the error-reject trade-off in biometric verification systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:786 –796, 1997.

Y. Guo, P.L. Bartlett, J. Shawe-Taylor, and R.C Williamson. Covering numbers for support vector machines. *IEEE Transactions on Information Theory*, 48(1):239 – 250, 2002.

L. Györfi, Z. Györfi, and I. Vajda. Bayesian decision with rejection. *Problems of Control and Information Theory*, 8:445–452, 1978.

L. K. Hansen, C. Lissberg, and P. Salamon. The error-reject tradeoff. *Open Systems and Information Dynamics*, 4:159–184, 1997.

R. Herbei and M. H. Wegkamp. Classification with reject option. *Canadian Journal of Statistics*, 4 (4):709–721, 2006.

G.S̃. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971.

A.N. Kolmogorov and V.M. Tichomirov. ε-entropy and ε-capacity of sets in functional spaces. *American Mathematical Society Translations*, 17:277–364, 1961.

C.W. Landgrebe, D.M.J. Tax, P. Paclik, and R.P.W. Duin. The interaction between classification and reject performance for distance-based reject-option classifiers. *Pattern Recognition Letters*, 27(8):908–917, 2006.

P. Massart. *Concentration Inequalities and Model Selection*, volume 1896. Springer, 2007.

B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

I. Steinwart and C. Scovel. Fast rates for support vector machines using gaussian kernels. *Annals of Statistics*, 35(2):575–607, 2007.

B. Tarigan and S. A. van de Geer. Classifiers of support vector machine type with ℓ_1 complexity regularization. *Bernoulli*, 12(6):1045–1076, 2006.

F. Tortorella. Reducing the classification cost of support vector classifiers through an ROC-based rejection rule. *Pattern Analysis and Applications*, 7:128 – 143, 2004.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32: 135–166, 2004.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

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

D.X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49(7):1743–1752, 2003.