CLASSIFICATION WITH A REJECT OPTION USING A HINGE LOSS

PETER L. BARTLETT AND MARTEN H. WEGKAMP

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.

Key words and phrases: Bayes classifiers; classification; convex surrogate loss; empirical risk minimization; hinge loss; large margin classifiers; margin condition; reject option; support vector machines.
MSC 2000: Primary 62C05; secondary 62G05, 62G08

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, labelled -1 or $+1$. Since an observation X does not fully determine its label y, we construct a classifier $t : \mathcal{X} \to \{-1, +1\}$ that represents our guess $t(X)$ of the label Y of a future observation X. The rule with the smallest probability of error $\mathbb{P}\{t(X) \neq Y\}$ is the Bayes rule

$$t^*(x) \;:=\; \begin{cases} -1 & \text{if } \mathbb{P}\{Y = -1|X = x\} \geq \mathbb{P}\{Y = +1|X = x\} \\ +1 & \text{otherwise} \end{cases} \tag{1}$$

Observations x for which the conditional probability

$$\eta(x) := \mathbb{P}\{Y = +1|X = x\} \tag{2}$$

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. In case the reject option is invoked, no decision is made. Although such classifiers are valuable in practice, few theoretical results are available—see [10] and [7] for references.

In this note, we assume that the cost of making a wrong decision is 1 and the cost of utilizing the reject option is $d > 0$. Given a classifier with reject option $t : \mathcal{X} \to \{-1, 1, \circledR\}$, the appropriate risk function is

$$L_d(t) := d\mathbb{P}\{t(X) = \circledR\} + \mathbb{P}\{t(X) \neq Y, t(X) \neq \circledR\}. \tag{3}$$

It is easy to see that the rule minimizing this risk assigns $-1, 1$ or ® depending on which of $\eta(x)$, $1 - \eta(x)$ or d is smallest. According to this rule, which we refer to as the Bayes rule, we need never reject if $d \geq 1/2$. For this reason we restrict ourselves to the cases $0 \leq d \leq 1/2$ and the Bayes rule with reject option is then

$$t_d^*(x) \quad := \quad \begin{cases} -1 & \text{if } \eta(x) < d \\ +1 & \text{if } \eta(x) > 1 - d \\ \circledR & \text{otherwise} \end{cases} \tag{4}$$

with risk

$$L_d^* := L_d(t_d^*) = \mathbb{E} \min\{\eta(X), 1 - \eta(X), d\}. \tag{5}$$

The case $d = 1/2$ reduces to the classical situation without the reject option and (4) coincides with (1).

Herbei and Wegkamp [7] study "plug-in" rules that replace the regression function $\eta(x)$ by an estimate $\widehat{\eta}(x)$ in the formula for $t_d^*(x)$ above. They show that the rate of convergence of the risk (3) to the Bayes risk L_d^* 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 Tsybakov's "noise" condition (cf. [13]) from the classical setting ($d = 1/2$) to our more general framework ($0 \leq d \leq 1/2$). The same paper considers classifiers $\widehat{t}$ that minimize the empirical counterpart

$$\widehat{L}_d(t) := \frac{d}{n}\sum_{i=1}^{n} \mathbb{I}\{t(X_i) = \circledR\} + \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{t(X_i) \neq Y_i, t(X_i) \neq \circledR\} \tag{6}$$

of the risk $L_d(t)$ over a class $\mathcal{T}$ of classifiers $t : \mathcal{X} \to \{-1, +1, \circledR\}$ with reject option, based on a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of independent copies of the pair (X, Y). For such rules $\widehat{t}$,

[7] establish oracle inequalities for the excess risk of the form

$$L_d(\widehat{t}) - L_d^* \le C_0 \inf_{t\in\mathcal{T}} [L_d(t) - L_d^*] + \Delta_n$$

for some constant $C_0 > 1$ and the remainder Δ_n depends on the sample size n, the "complexity" of the class $\mathcal{T}$ and the behavior of $\eta(X)$ near the values d and $1-d$. Their findings are in line with the recent theoretical developments of standard binary classification without the reject option ($d = 1/2$), see e.g. [3], [4], [9]. Despite the attractive theoretical properties of this method, it 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.

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 the Bayes classifier (4) is a transformed version of the minimizer of the risk associated with this new loss and that the excess risk $L_d - L_d^*$ can be bounded by $2d$ times the excess risk based on the piecewise linear loss ϕ_d. 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 quadratic program.

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 in [7]. 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 ([5], [12]).

2. Generalized hinge loss

We can associate any real-valued function f with a classifier with reject option using the transformation

$$t_{f,\delta}(x) = \begin{cases} -1 & \text{if } f(x) < -\delta, \\ \text{Ⓡ} & \text{if } |f(x)| \le \delta, \\ +1 & \text{if } f(x) > \delta, \end{cases} \tag{7}$$

where $0 \le \delta \le 1$ is an arbitrary positive number. (We recommend the value $\delta = 1/2$ but we postpone the discussion on the choice of δ until after Theorem 2.) According to (3), the risk of $t_{f,\delta}(x)$ equals then

$$\begin{aligned} L_d(t_{f,\delta}) &= \mathbb{P}\{t_{f,\delta}(X) \neq Y, t_{f,\delta}(X) \neq \circledR\} + d\mathbb{P}\{t_{f,\delta}(X) = \circledR\} \\ &= \mathbb{P}\{f(X) < -\delta, Y = 1\} + \mathbb{P}\{f(X) > \delta, Y = -1\} + d\mathbb{P}\{-\delta \leq f(X) \leq \delta\} \\ &= \mathbb{P}\{Yf(X) < -\delta\} + d\mathbb{P}\{|Yf(X)| \leq \delta\}. \end{aligned}$$

For example, the Bayes classifier $t_d^*(x)$ defined in (4) corresponds to the function

$$f_d^*(x) = \begin{cases} -1 & \text{if } \eta(x) < d, \\ 0 & \text{if } d \leq \eta(x) \leq 1-d, \\ 1 & \text{if } \eta(x) > 1-d. \end{cases} \tag{8}$$

Although the function $f_d^*(x)$ is not unique, the classifier $t_d^*(x)$ is the unique minimizer of $L_d(t_{f,\delta})$ over all measurable $f : \mathcal{X} \to \mathbb{R}$. We see that

$$L_d(t_{f,\delta}) = L_{d,\delta}(f) := \mathbb{E}\ell_{d,\delta}(Yf(X)) \tag{9}$$

for the discontinuous loss

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

The choice $(\delta, d) = (0, 1/2)$ corresponds to the classical case $L_d(t_{f,\delta}) = \mathbb{P}\{Yf(X) < 0\}$. All other choices $\delta > 0$ lead to classification that allows for the reject option with minimal risk

$$L_{d,\delta}(f_d^*) = \mathbb{E}\ell_{d,\delta}(Yf_d^*(X)) = L_d^*.$$

Instead of the discontinuous loss $\ell_{d,\delta}$, we consider a convex surrogate loss. Define the piecewise linear function

$$\phi_d(\alpha) = \begin{cases} 1 - a\alpha & \text{if } \alpha < 0, \\ 1 - \alpha & \text{if } 0 \leq \alpha < 1, \\ 0 & \text{otherwise} \end{cases}$$

where $a = (1-d)/d \geq 1$.

The next result states that the minimizer of the expectation of the discrete loss $\ell_{d,\delta}$ and the convex loss $\phi_d(\alpha)$ remains the same: f_d^* defined in (8) minimizes both $\mathbb{E}\ell_{d,\delta}(Yf(X))$ and $\mathbb{E}\phi_d(Yf(X))$ and the minimal risks are related via the equality $L_{d,\delta}(f_d^*) = dL_{\phi_d}(f_d^*)$.

Proposition 1. *The minimizer of the risk*

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

over all measurable $f : \mathcal{X} \to \mathbb{R}$ is the Bayes rule (8). Furthermore,

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

Proof. Note that

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

Hence, for

$$r_{\eta,\phi_d}(\alpha) = \eta\phi_d(\alpha) + (1-\eta)\phi_d(-\alpha) \tag{10}$$

it suffices to show that

$$\alpha^* = \begin{cases} -1 & \text{if } \eta < 1/(1+a), \\ 0 & \text{if } 1/(1+a) \leq \eta \leq a/(1+a), \\ 1 & \text{if } \eta(x) > a/(1+a) \end{cases} \tag{11}$$

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

$$r_{\eta,\phi_d}(\alpha) = \begin{cases} \eta - a\eta\alpha & \text{if } \alpha \leq -1, \\ 1 + \alpha(1-(1+a)\eta) & \text{if } -1 \leq \alpha \leq 0, \\ 1 + \alpha(-\eta + a(1-\eta)) & \text{if } 0 \leq \alpha \leq 1, \\ \alpha(a(1-\eta)) + (1-\eta) & \text{if } \alpha \geq 1 \end{cases}$$

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

$$\begin{aligned} & \inf_\alpha \eta\phi_d(\alpha) + (1-\eta)\phi_d(-\alpha) \\ = \; & \eta\phi_d(\alpha^*) + (1-\eta)\phi_d(\alpha^*) \\ = \; & \frac{\eta}{d}\{\eta \leq d\} + 1\{d \leq \eta \leq 1-d\} + \frac{1-\eta}{d}\{\eta \geq 1-d\}, \end{aligned}$$

we find that

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

and the second claim follows as well. □

We see that $\phi_d(\alpha) \geq \ell_{d,\delta}(\alpha)$ for all $\alpha \in \mathbb{R}$ as long as $0 \leq \delta \leq 1-d$. Since this pointwise relation remains preserved under taking expected values, we immediately obtain $L_d(t_{f,\delta}) \leq L_{\phi_d}(f)$. 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 $0 \leq d < 1/2$ be fixed. For all $0 < \delta \leq 1/2$, we have*

$$L_d(t_{f,\delta}) - L_d^* \leq \frac{d}{\delta}\left(L_{\phi_d}(f) - L_{\phi_d}^*\right), \tag{12}$$

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

$$L_d(t_{f,\delta}) - L_d^* \leq L_{\phi_d}(f) - L_{\phi_d}^*. \tag{13}$$

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

$$L(t_f) - L^* \leq L_\phi(f) - L^*_\phi, \tag{14}$$

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

Remark. 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

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

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 $\ell_{d,\delta}(\alpha)$ is still majorized by the convex surrogate $\phi_d(\alpha)$. 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 (14) is due to Zhang [14].

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

$$\xi(\eta) = \eta\{\eta \leq d\} + d\{d \leq \eta \leq 1-d\} + (1-\eta)\{\eta \geq 1-d\}$$

and

$$\begin{aligned} H(\eta) &= \inf_\alpha \eta\phi_d(\alpha) + (1-\eta)\phi_d(-\alpha) \\ &= \frac{\eta}{d}\{\eta \leq d\} + 1\{d \leq \eta \leq 1-d\} + \frac{1-\eta}{d}\{\eta \geq 1-d\}. \end{aligned}$$

(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

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

Proposition 3. *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 $b \in \{-1, 1, \circledR\}$,

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

Proof. First we compute

$$\begin{aligned}
\inf_{\alpha \leq -1} r_\phi(\alpha) &= \frac{\eta}{d}, \\
\inf_{-1 \leq \alpha \leq -\delta} r_\phi(\alpha) &= \frac{\eta}{d}\{\eta \leq d\} + \left(\frac{\delta}{d}\eta + 1 - \delta\right)\{\eta \geq d\} \\
\inf_{-\delta \leq \alpha \leq 0} r_\phi(\alpha) &= \{\eta \geq d\} + \left(1 - \delta + \eta\frac{\delta}{d}\right)\{\eta \leq d\} \\
\inf_{0 \leq \alpha \leq \delta} r_\phi(\alpha) &= \{\eta \leq 1 - d\} + \left(1 + \frac{\delta}{d} - \delta - \frac{\delta}{d}\eta\right)\{\eta \geq 1 - d\} \\
\inf_{\delta \leq \alpha \leq 1} r_\phi(\alpha) &= \frac{1-\eta}{d}\{\eta \geq 1 - d\} + \left(1 + \frac{\delta}{d} - \delta - \frac{\delta}{d}\eta\right)\{\eta \leq 1 - d\} \\
\inf_{\alpha \geq 1} r_\phi(\alpha) &= \frac{1-\eta}{d}
\end{aligned}$$

It is now easy to verify that

$$\begin{aligned}
H_{-1}(\eta) &= \inf_{\alpha < -\delta} \eta\phi_d(\alpha) + (1 - \eta)\phi_d(-\alpha) \\
&= \frac{\eta}{d}\{\eta \leq d\} + \left(\frac{\delta}{d}\eta + 1 - \delta\right)\{\eta \geq d\}
\end{aligned}$$

so that

$$\begin{aligned}
&H_{-1}(\eta) - H(\eta) = \\
&0\{\eta \leq d\} + \left(\frac{\delta}{d}\eta - \delta\right)\{d \leq \eta \leq 1 - d\} + \left(\frac{1+\delta}{d}\eta + 1 - \delta - \frac{1}{d}\right)\{\eta \geq 1 - d\}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\xi_{-1}(\eta) &= \eta - \xi(\eta) \\
&= 0\{\eta \leq d\} + (\eta - d)\{d \leq \eta \leq 1 - d\} + (2\eta - 1)\{\eta \geq 1 - d\}
\end{aligned}$$

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

$$\begin{aligned}
H_{\circledR}(\eta) &= \inf_{|\alpha|\le\delta} \eta\phi_d(\alpha) + (1-\eta)\phi_d(-\alpha) \\
&= \left(1 - \delta + \frac{\delta}{d}\eta\right)\{\eta \le d\} + \{d \le \eta \le 1-d\} \\
&\qquad + \left(1 - \delta + \frac{\delta}{d} - \frac{\delta}{d}\eta\right)\{\eta \ge 1-d\}
\end{aligned}$$

and

$$\begin{aligned}
H_{\circledR}(\eta) - H(\eta) &= \left(1 - \delta - \frac{1-\delta}{d}\eta\right)\{\eta \le d\} \\
&\qquad + \left(1 - \delta - \frac{1-\delta}{d} + \frac{1-\delta}{d}\eta\right)\{\eta \ge 1-d\}.
\end{aligned}$$

Since

$$\begin{aligned}
\xi_{\circledR}(\eta) &= d - \xi(\eta) \\
&= (d-\eta)\{\eta \le d\} + 0\{d \le \eta \le 1-d\} + (d-1+\eta)\{\eta \ge 1-d\}
\end{aligned}$$

we find that

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

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

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

and consequently

$$\begin{aligned}
H_1(\eta) - H(\eta) &= \left(1 - \delta + \frac{\delta}{d} - \frac{\delta}{d}\eta - \frac{\eta}{d}\right)\{\eta \le d\} \\
&\qquad + \left(\frac{\delta}{d} - \delta - \frac{\delta}{d}\eta\right)\{d \le \eta \le 1-d\} + 0\{\eta \ge 1-d\}.
\end{aligned}$$

Now,

$$\begin{aligned}
\xi_1(\eta) &= 1 - \eta - \xi(\eta) \\
&= (1-2\eta)\{\eta \le d\} + (1-\eta-d)\{d \le \eta \le 1-d\} + 0\{\eta \ge 1-d\},
\end{aligned}$$

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 3. Assume that $d \le \delta \le 1 - d$.

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

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

$\xi_{\circledR}(\eta) \le H_{\circledR}(\eta) - H(\eta) \iff (1+\delta-2d)\eta \le \delta(1-d)$. As $\eta \le d$, we need that $(1+\delta-2d)d \le \delta(1-d)$, that is, $(\delta-d)(1-2d) \ge 0$.

$\xi_{1}(\eta) \le H_{1}(\eta) - H(\eta) \iff (1+\delta-2d)\eta \le \delta(1-d)$. 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

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

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

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

Finally, if $\eta \ge 1-d$, 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 $(1+\delta-2d)(1-d) \ge (1+d\delta-2d) \iff (\delta-d)(1-2d) \ge 0$.

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

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

Proof of Theorem 2. 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,

$$\begin{aligned}\psi(L_d(t_{f,\delta}) - L_d^*) &= \mathrm{P}\{f < -\delta\}\psi(\phi_{-1}(\eta)) + \mathrm{P}\{-\delta \le f \le \delta\}\psi(\phi_{\circledR}(\eta)) \\ &\qquad + \mathrm{P}\{f > \delta\}\psi(\phi_1(\eta)),\end{aligned}$$

where P is the probability measure of X and $\mathrm{P}f = \int f \, d\mathrm{P}$. Invoke now Proposition 3 to deduce

$$\begin{aligned}\psi(L_d(t_{f,\delta}) - L_d^*) &\le \mathrm{P}\{f < -\delta\}\,[H_{-1}(\eta) - H(\eta)] + \mathrm{P}\{-\delta \le f \le \delta\}([H_{\circledR}(\eta) - H(\eta)] \\ &\quad + \mathrm{P}\{f > \delta\}\,[H_1(\eta) - H(\eta)]\end{aligned}$$

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 $d \leq \delta \leq 1 - d$ and $0 \leq d < 1/2$, take $\psi(x) = x$. □

3. SVM classifiers with reject option

In this section, we consider the SVM classifier with reject option, and show that it can be obtained by solving a quadratic program.

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 sum of the empirical ϕ_d-risk and a regularization term that is proportional to the squared RKHS norm. The following theorem shows that this classifier is the solution to a quadratic program, that is, it is the minimizer of a quadratic criterion on a subset of Euclidean space defined by linear inequalities. Thus, the classifier can be found efficiently using general-purpose algorithms.

Theorem 4. *For any $x_1, \ldots, x_n \in \mathcal{X}$ and $y_1, \ldots, y_n \in \{-1, 1\}$, let $f^* \in \mathcal{H}$ be the minimizer of the regularized risk functional*

$$f \mapsto \sum_{i=1}^{n} \phi_d\left(y_i f(x_i)\right) + \lambda \|f\|^2,$$

where $\lambda > 0$. Then we can represent f^* as the finite sum

$$f^*(x) = \sum_{i=1}^{n} \alpha_i^* k(x_i, x),$$

where $\alpha_1^*, \ldots, \alpha_n^*$ is the solution to the following quadratic program.

$$\begin{aligned}
\min_{\alpha_i, \xi_i, \gamma_i} \quad & \frac{1}{n} \sum_{i=1}^{n} \left(\xi_i + \frac{1-2d}{d} \gamma_i \right) + \lambda \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) \\
\textit{s.t.} \quad & \xi_i \geq 0, \qquad \gamma_i \geq 0, \\
& \xi_i \geq 1 - y_i \sum_{j=1}^{n} \alpha_j k(x_i, x_j), \\
& \gamma_i \geq -y_i \sum_{j=1}^{n} \alpha_j k(x_i, x_j) \qquad\qquad \textit{for } i = 1, \ldots, n.
\end{aligned}$$

Proof. The fact that f^* can be represented as a finite sum over the kernel basis functions is a standard argument (see [8, 6]). It follows from Pythagoras' theorem: 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 contributes only to the squared norm term and not to the cost term. Thus, a minimizing f^* can be represented in terms of the solution α^* to the minimization

$$\min_{\alpha_1,\ldots,\alpha_n} \qquad \frac{1}{n}\sum_{i=1}^{n}\phi_d\left(y_i\sum_{j=1}^{n}\alpha_j k(x_i, x_j)\right) + \lambda \sum\sum_{1\le i,j\le n}\alpha_i\alpha_j k(x_i, x_j).$$

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

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

Defining $\xi_i = \max\{0, 1 - y_i f(x_i)\}$ and $\gamma_i = \max\{0, -y_i f(x_i)\}$ gives the QP. □

4. Tsybakov's noise 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

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

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 to Bayes risk.

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 $1/2$ (see [13, 2, 5, 12, 1]). For plug-in rules, [7] 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 Tsybakov's margin condition [13].

Definition 5. *We say that η satisfies the margin condition at d with exponent α if there is an $A \geq 1$ such that for all $t > 0$,*

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

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

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

where f^* 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 6. *We say that $\mathcal{G} \subset L_2(\mathrm{P})$ is a (β, B)-Bernstein class with respect to the probability measure* P $(0 < \beta \leq 1,\ B \geq 1)$ *if every $g \in \mathcal{G}$ satisfies*

$$\mathrm{Pg}^2 \leq \mathrm{B}\{\mathrm{Pg}\}^{\beta}.$$

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 7. *If η satisfies the margin condition at d with exponent α, then for any class 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^*. 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

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

and recall the definition of the conditional ϕ_d-risk in (10).

Lemma 8. *For* $\eta \in [0,1]$,

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

Proof. Since r_{η,ϕ_d} is convex,

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

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

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

Thus, we have

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

□

We shall also use the following inequalities.

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

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

and

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

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

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

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

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

□

Proof of Lemma 7. By Lemma 8, we have

$$L_{\phi_d}(f)-L^*_{\phi_d}\ge d^{-1}\mathrm{P}\rho_\eta(f,f^*)\left(|\eta-(1-d)|I_{E_-}+|\eta-d|I_{E_+}\right),$$

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,

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

where B is such that $|f|\le B$. Similarly,

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

and we obtain

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

Choose

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

in the expression above, and we obtain

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

and so

$$\mathrm{P}\rho_\eta(f, f^*) \leq \left\{ 2d(4(B+1)A)^{1/\alpha} \right\}^{\alpha/(\alpha+1)} \left\{ \mathbb{E} g_f(X,Y) \right\}^{\alpha/(1+\alpha)}. \tag{15}$$

In addition, by Lemma 9,

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

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. Specialized to the case $(\delta, d) = (0, 1/2)$, we note that Lemma 7 removes unnecessary technical restrictions on $\eta(X)$ near 0 and 1, imposed in [5] and [12].

Lemma 7 provides the main ingredient for establishing fast rates of minimizers $\widehat{f}_d$ of the empirical risk

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

Theorem 10. *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 $0 \leq p \leq 2$, then there exists a constant C' independent of n, such that

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

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 $|g_f - g_{f'}| \le |f - f'|$, it follows that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}(\mathrm{P} - 2\mathbb{P}_n)g_f\right] \le 4\varepsilon_n + 2B\mathbb{P}\left\{\sup_{f\in\mathcal{F}_n}(\mathrm{P} - 2\mathbb{P}_n)g_f \ge \varepsilon_n\right\},$$

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

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

The union bound and Bernstein's exponential inequality for the tail probability of sums of bounded random variables in conjunction with Lemma 7, yield

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

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 A in ε_n such that $C\varepsilon_n^{-p} = -2cn\varepsilon_n^{2-\beta}$ and $\exp(-n\varepsilon_n^{2-\beta}) = o(\varepsilon_n)$. □

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

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'.

References

1. J. Y. Audibert and A. B. Tsybakov (2005). Fast convergence rates for plug-in classifiers under margin conditions. *Annals of Statistics.* (To appear)
2. P. L. Bartlett, M. I. Jordan and J. D. McAuliffe (2006). Convexity, classification, and risk bounds, *Journal of the American Statistical Association*, 101(473):138-156.
3. S. Boucheron, O. Bousquet and G. Lugosi (2004a). Introduction to statistical learning theory. In *Advanced Lectures in Machine Learning* (O. Bousquet, U. von Luxburg, and G. Rätsch, Editors), 169–207. Springer, New-York.
4. S. Boucheron, O. Bousquet and G. Lugosi (2005). Theory of Classification: a Survey of Recent Advances. *ESAIM: Probability and Statistics*, 9:323-375.
5. G. Blanchard, O. Bousquet and P. Massart (2004). Statistical Performance of Support Vector Machines. *Manuscript* (http://ida.first.fraunhofer.de/∼blanchard/publi).
6. D. Cox and F. O'Sullivan (1990). Asymptotic analysis of penalized likelihood and related estimators. *Annals of Statistics*, 18:1676–1695, 1990.
7. R. Herbei and M. H. Wegkamp (2006). Classification with reject option. *Canadian Journal of Statistics.* (To appear)
8. G. S. Kimeldorf and G. Wahba (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95.
9. P. Massart (2003). St Flour Lecture Notes. *Manuscript.* (http://www.math.u-psud.fr/∼massart/stf2003_massart.pdf)
10. B. D. Ripley (1996). *Pattern recognition and neural networks.* Cambridge University Press, Cambridge.
11. I. Steinwart and C. Scovel (2004). Fast Rates for Support Vector Machines using Gaussian Kernels. *Los Alamos National Laboratory Technical Report LA-UR-04-8796.*
12. B. Tarigan and S. A. van de Geer (2004). Adaptivity of support vector machines with ℓ_1 penalty. *Technical Report MI 2004-14, University of Leiden.*
13. A. B. Tsybakov (2004). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32, 135–166.
14. T. Zhang (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32: 56-85.

Computer Science Division and Department of Statistics, University of California, Berkeley
E-mail address: bartlett@cs.berkeley.edu

Department of Statistics, Florida State University, Tallahassee
E-mail address: wegkamp@stat.fsu.edu