Journal of Machine Learning Research 11 (2010) 111-130 Submitted 2/09; Revised 11/09; Published 1/10

Classification Methods with Reject Option Based on Convex Risk Minimization

Ming Yuan MYUAN@ISYE.GATECH.EDU
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205, USA

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

Editor: Bin Yu

Abstract

In this paper, we investigate the problem of binary classification with a reject option in which one can withhold the decision of classifying an observation at a cost lower than that of misclassification. Since the natural loss function is non-convex so that empirical risk minimization easily becomes infeasible, the paper proposes minimizing convex risks based on surrogate convex loss functions. A necessary and sufficient condition for infinite sample consistency (both risks share the same minimizer) is provided. Moreover, we show that the excess risk can be bounded through the excess surrogate risk under appropriate conditions. These bounds can be tightened by a generalized margin condition. The impact of the results is illustrated on several commonly used surrogate loss functions.

Keywords: classification, convex surrogate loss, empirical risk minimization, generalized margin condition, reject option

1. Introduction

In binary classification, one observes independent realizations $(X_1,Y_1),\ldots,(X_n,Y_n)$ of the random pair (X,Y) where $X \in \mathcal{X}$ and $Y \in \mathcal{Y} = \{-1,1\}$. The goal is to learn from these training data a classification rule $g : \mathcal{X} \mapsto \mathcal{Y}$ that classifies an observation X into the two classes. It is recognized that in many applications the consequences of misclassification can be substantial. In such situations, a less specific response that reserves the right of not making a decision, sometimes referred to as a reject option (see, e.g., Herbei and Wegkamp, 2006), may even be more preferable than risking misclassification. This, for example, is typical in medical studies where screening of a certain disease can be done based on relatively inexpensive clinical measures. If the classification based on these measurements are satisfactory, nothing further needs to be done. But in the event that there are ambiguities, it would be more desirable to take a rejection option and seek more expensive studies to identify a subject's disease status. Similar approaches are often adopted in DNA sequencing or genotyping applications, where the rejection option is commonly referred to as a "no-call". Similar problems have attracted much attention in various application fields and also received increasing

amount of interest more recently in machine learning literature. See Ripley (1996) and Bartlett and Wegkamp (2008) and references therein.

To accommodate the reject option, we now seek a classification rule $g : \mathcal{X} \to \tilde{\mathcal{Y}}$ where $\tilde{\mathcal{Y}} = \{-1, 1, 0\}$ is an augmented response space and $g(X) = 0$ indicates that no definitive classification will be made for X or a reject option is taken. To measure the performance of a classification rule, we employ the following loss function that generalizes the usual 0-1 loss to account for reject option:

$$\ell[g(X), Y] = \begin{cases} 1 & \text{if } g(X) \neq Y \text{ and } g(X) \neq 0 \\ d & \text{if } g(X) = 0 \\ 0 & \text{if } g(X) = Y \end{cases} .$$

In other words, an ambiguous response ($g(X) = 0$) incurs a loss of d whereas misclassification incurs a loss of 1. Note that d is necessarily smaller than $1/2$. Otherwise, rather than taking a rejection option with a loss d, we can always flip a fair coin to randomly assign ± 1 as the value of $g(X)$, which incurs an average loss of $1/2 \leq d$. For this reason, we shall assume that $d < 1/2$ in what follows.

For any classification rule $g : \mathcal{X} \to \tilde{\mathcal{Y}}$, the risk function is then given by $R(g) = \mathbb{E}(\ell[g(X), Y])$ where the expectation is taken over the joint distribution of X and Y. It is not hard to show that the optimal classification rule $g^* := \arg\min R(g)$ is given by (see, e.g., Bartlett and Wegkamp, 2008)

$$g^*(X) = \begin{cases} 1 & \text{if } \eta(X) > 1 - d \\ 0 & \text{if } d \leq \eta(X) \leq 1 - d \\ -1 & \text{if } \eta(X) < d \end{cases}$$

where $\eta(X) = \mathbb{P}(Y = 1|X)$. The corresponding risk is

$$R^* := \inf R(g) = R(g^*) = \mathbb{E}\left(\min\{\eta(X), 1 - \eta(X), d\}\right).$$

Thus, the performance of any classification rule $g : \mathcal{X} \to \tilde{\mathcal{Y}}$ can be measured by the excess risk $\Delta R(g) := R(g) - R^*$.

Appealing to the general empirical risk minimization strategy, one could attempt to derive a classification rule from the training data by minimizing the empirical risk

$$R_n(g) = \frac{1}{n} \sum_{i=1}^{n} \ell[g(X_i), Y_i].$$

Similar to the usual 0-1 loss, however, ℓ is not convex in g; and direct minimization of R_n is typically an NP-hard problem. A common remedy is to consider a surrogate convex loss function. To this end, let $\phi : \mathbb{R} \mapsto \mathbb{R}$ be a convex function. Denote by

$$Q(f) = \mathbb{E}[\phi(Y f(X))]$$

the corresponding risk for a discriminant function $f : \mathcal{X} \mapsto \mathbb{R}$. Let $\hat{f}_n$ be the minimizer of

$$Q_n(f) = \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i f(X_i))$$

over a certain functional space $\mathcal{F}$ consisting of functions that map from $\mathcal{X}$ to $\mathbb{R}$. $\hat{f}_n$ can be conveniently converted to a classification rule $C(\hat{f}_n;\delta)$ as follows:

$$C(f(X);\delta) = \begin{cases} 1 & \text{if } f(X) > \delta \\ 0 & \text{if } |f(X)| \le \delta \\ -1 & \text{if } f(X) < -\delta \end{cases}$$

where $\delta > 0$ is a parameter that as we shall see plays a critical role in determining the performance of $C(\hat{f}_n,\delta)$.

In this paper, we investigate the statistical properties of this general convex risk minimization technique. To what extent $C(\hat{f}_n,\delta)$ mimics the optimal classification rule g^* plays a critical role in the success of this technique. Let f_ϕ^* be the minimizer of $Q(f)$. We shall assume throughout the paper that f_ϕ^* is uniquely defined. Typically f_ϕ^* reflects the limiting behavior of $\hat{f}_n$ when $\mathcal{F}$ is rich enough and there are infinitely many training data. Therefore the first question is whether or not f_ϕ^* can be used to recover the optimal rule g^*. A surrogate loss function ϕ that satisfies this property is often called infinite sample consistent (see, e.g., Zhang, 2004) or classification calibrated (see, e.g., Bartlett, Jordan and McAuliffe, 2006). A second question further concerns the relationship between the excess risk $\Delta R[C(f,\delta)]$ and the excess ϕ risk $\Delta Q(f) = Q(f) - \inf Q(f)$: Can we find an increasing function $\rho : \mathbb{R} \mapsto \mathbb{R}$ such that for all f,

$$\Delta R[C(f,\delta)] \le \rho\left(\Delta Q(f)\right) ? \tag{1}$$

Clearly the infinite sample consistency of ϕ implies that $\rho(0) = 0$. Such a bound on the excess risk provides useful tools in bounding the excess risk of $\hat{f}_n$. In particular, (1) indicates that

$$\Delta R[C(\hat{f}_n,\delta)] \le \rho\left(\Delta Q(\hat{f}_n)\right) = \rho\left[\Delta Q(\bar{f}) + \left(Q(\hat{f}_n) - Q(\bar{f})\right)\right],$$

where $\bar{f} = \arg\min_{f\in\mathcal{F}} Q(f)$. The first term $\Delta Q(\bar{f})$ on the right-hand side exhibits the approximation error of functional class $\mathcal{F}$ whereas the second term $Q(\hat{f}_n) - Q(\bar{f})$ is the estimation error.

In the case when there is no reject option, these problems have been well investigated in recent years (Lin, 2002; Zhang, 2004; Bartlett, Jordan and McAuliffe, 2006). In this paper, we establish similar results when there is a reject option. The most significant difference between the two situations, with or without the reject option, is the role of δ. As we shall see, for some loss functions such as least squares, exponential or logistic, a good choice of δ yields classifiers that are infinite sample consistent. For other loss functions, however, such as the hinge loss, no matter how δ is chosen, the classification rule $C(f,\delta)$ cannot be infinite sample consistent.

The remainder of the paper is organized as follows. We first examine in Section 2 the infinite sample consistency for classification with reject option. After establishing a general result, we consider its implication on several commonly used loss functions. In Section 3, we establish bounds on the excess risk in the form of (1), followed by applications to the popular loss functions. We also show that under an additional assumption on the behavior of $\eta(X)$ near d and $1-d$ as in Herbei and Wegkamp (2006), generalizing the condition in the case of $d = 1/2$ of Mammen and Tsybakov (1999) and Tsybakov (2004), the bound (1) can be tightened considerably. Section 4 discusses rates of convergence of the empirical risk minimizer $\widehat{f}_n$ that minimizes the empirical risk $Q_n(f)$ over a bounded class $\mathcal{F}$. Section 5 considers extension to asymmetric loss where one type of misclassification may be more costly than the other. All proofs are relegated to Section 6.

2. Infinite Sample Consistency

We first give a general result on the infinite sample consistency of the classification rule $C(f_\phi^*, \delta)$.

Theorem 1 *Assume that ϕ is convex. Then the classification rule $C(f_\phi^*, \delta)$ for some $\delta > 0$ is infinite sample consistent, that is, $C(f_\phi^*, \delta) = g^*$ if and only if $\phi'(\delta)$ and $\phi'(-\delta)$ both exist, $\phi'(\delta) < 0$, and*

$$\frac{\phi'(\delta)}{\phi'(\delta) + \phi'(-\delta)} = d. \tag{2}$$

When there is no reject option, it is known that the necessary and sufficient condition for the infinite sample consistency is that ϕ is differentiable at 0 and $\phi'(0) < 0$ (see, e.g., Bartlett, Jordan and McAuliffe, 2006). As indicated by Theorem 1, the differentiability of ϕ at $\pm\delta$ plays a more prominent role in the general case when there is a reject option.

From Theorem 1 it is also evident that the infinite sample consistency depends on both ϕ and the choice of thresholding parameter δ. Observe that for any $\delta_1 < \delta_2$,

$$\phi'(-\delta_2) \le \phi'(-\delta_1) \le \phi'(\delta_1) \le \phi'(\delta_2),$$

which implies that the left-hand side of (2) is a decreasing function of δ. If ϕ is strictly convex, then it is strictly decreasing; and therefore there is at most one value of δ that satisfies (2). In other words, for strictly convex ϕ, there is at most one thresholding parameter δ such that $C(f_\phi^*, \delta) = g^*$. On the other hand, if ϕ is twice differentiable such that $\phi'(0) < 0$ and $\phi'(z) \ge 0$ as $z \to +\infty$, then for any $d < 1/2$, there always exists a $\delta > 0$ such that (2) holds. This is because the left-hand side of (2) is a decreasing function of δ, which approaches its supremum $1/2$ when $\delta \downarrow 0$ and 0 when δ increases. Moreover, the twice differentiability of ϕ ensures that the left-hand side of (2) is also a continuous function of δ. The following is therefore a direct consequence of Theorem 1:

Corollary 2 *If ϕ is strictly convex, then either there is a unique $\delta > 0$ such that $C(f_\phi^*, \delta)$ is infinite sample consistent; or $C(f_\phi^*, \delta)$ is not infinite sample consistent for any $\delta > 0$. In addition to convexity, if ϕ is twice differentiable such that $\phi'(0) < 0$ and $\phi'(z) \ge 0$ as $z \to +\infty$, then there always exists a $\delta > 0$ such that $C(f_\phi^*, \delta)$ is infinite sample consistent.*

Theorem 1 provides a general guideline on how to choose δ for common choices of convex losses. Below we look at several concrete examples.

2.1 Least Squares Loss

We first examine the least squares loss $\phi(z) = (1-z)^2$. Observe that

$$\frac{\phi'(\delta)}{\phi'(-\delta) + \phi'(\delta)} = \frac{1-\delta}{2}.$$

All conditions of Theorem 1 are met if and only if $\delta = 1 - 2d$.

Corollary 3 *For the least squares loss,*

$$C(f_\phi^*, 1-2d) = g^*.$$

2.2 Exponential Loss

Exponential loss, $\phi(z) = \exp(-z)$, is connected with boosting (Friedman, Hastie and Tibshirani, 2000). Because

$$\frac{\phi'(\delta)}{\phi'(-\delta)+\phi'(\delta)} = \frac{1}{1+\exp(2\delta)},$$

Therefore all conditions of Theorem 1 are met if and only if

$$\delta = \frac{1}{2}\log\left(\frac{1}{d}-1\right).$$

Corollary 4 *For the exponential loss,*

$$C\left(f_\phi^*, \frac{1}{2}\log\left(\frac{1}{d}-1\right)\right) = g^*.$$

2.3 Logistic Loss

Logisitic regression employs loss $\phi(z) = \ln(1+\exp(-z))$. Similar to before,

$$\frac{\phi'(\delta)}{\phi'(-\delta)+\phi'(\delta)} = \frac{1}{1+\exp(\delta)},$$

which suggests that all conditions of Theorem 1 are met if

$$\delta = \log\left(\frac{1}{d}-1\right).$$

Corollary 5 *For the logistic loss,*

$$C\left(f_\phi^*, \log\left(\frac{1}{d}-1\right)\right) = g^*.$$

2.4 Squared Hinge Loss

Squared hinge loss, $\phi(z) = (1-z)_+^2$, is another popular choice for which

$$\frac{\phi'(\delta)}{\phi'(-\delta)+\phi'(\delta)} = \frac{1-\delta}{2}.$$

Similar to the least squares loss, we have the following corollary.

Corollary 6 *For the squared hinge loss,*

$$C\left(f_\phi^*, 1-2d\right) = g^*.$$

2.5 Distance Weighted Discrimination

Marron, Todd and Ahn (2007) recently introduced the so-called distance weighted discrimination method where the following loss function (see, e.g., Bartlett, Jordan and McAuliffe, 2006) is used

$$\phi(z) = \begin{cases} \frac{1}{z} & \text{if } z \geq \gamma \\ \frac{1}{\gamma}\left(2 - \frac{z}{\gamma}\right) & \text{if } z < \gamma \end{cases}, \tag{3}$$

where $\gamma > 0$ is a constant. It is not hard to see that ϕ is convex. Moreover,

$$\phi'(z) = \begin{cases} -1/z^2 & \text{if } z \geq \gamma \\ -1/\gamma^2 & \text{if } z < \gamma \end{cases}.$$

Thus,

$$\frac{\phi'(\delta)}{\phi'(-\delta)+\phi'(\delta)} = \begin{cases} 1/2 & \text{if } \delta < \gamma \\ \frac{1/\delta^2}{1/\delta^2+1/\gamma^2} & \text{if } \delta > \gamma \end{cases}.$$

In other words, we have the following result for the distance weighted discrimination loss.

Corollary 7 *For the loss (3),*

$$C\left(f_\phi^*, [(1-d)/d]^{1/2}\gamma\right) = g^*.$$

2.6 Hinge Loss

The popular support vector machine employs the hinge loss, $\phi(z) = (1-z)_+$. The hinge loss is differentiable everywhere except 1. Therefore

$$\frac{\phi'(\delta)}{\phi'(-\delta)+\phi'(\delta)} = \begin{cases} \frac{1}{2} & \text{if } 0 < \delta < 1 \\ 0 & \text{if } \delta > 1 \end{cases}.$$

Because $0 < d < 1/2$, there does not exist a δ such that all conditions of Theorem 1 are met. As a matter of fact, for any $\delta > 0$, $C(f_\phi^*, \delta) \neq g^*$. Motivated by this observation, Bartlett and Wegkamp (2008) introduce the following modification to the hinge loss:

$$\phi(z) = \begin{cases} 1 - az & \text{if } z \leq 0 \\ 1 - z & \text{if } 0 < z \leq 1 \\ 0 & \text{if } z > 1 \end{cases}, \tag{4}$$

where $a > 1$. Note that with this modification,

$$\frac{\phi'(\delta)}{\phi'(-\delta)+\phi'(\delta)} = \begin{cases} 1/(a+1) & \text{if } 0 < \delta < 1 \\ 0 & \text{if } \delta > 1 \end{cases}.$$

Therefore, we have the following corollary.

Corollary 8 *For the modified hinge loss (4) and any $\delta < 1$, if $a = (1-d)/d$, then*

$$C\left(f_\phi^*, \delta\right) = g^*.$$

It is interesting to note that for the examples we considered previously, a specific choice of δ is needed to ensure the infinite sample consistent. Whereas for the modified hinge loss, a range of choice of δ can serve the same purpose. However, as we shall see in the next section, different choices of δ for the modified hinge loss may result in slightly different bound on the excess risk with $\delta = 1/2$ appearing to be more preferable in that it yields the smallest upper bound of the excess risk.

3. Excess Risk

We now turn to the excess risk $\Delta R[C(f,\delta)]$ and show how it can be bounded through the excess ϕ risk

$$\Delta Q(f) := Q(f) - Q(f_\phi^*).$$

Recall that the infinite sample consistency established in the previous section means that $\Delta Q(f) = 0$ implies throughout this section that $\Delta R(C(f,\delta)) = 0$. For brevity, we shall assume implicitly that δ is chosen in accordance with Theorem 1 to ensure infinite sample consistency. Write

$$Q_{\eta(X)}(z) = \eta(X)\phi(z) + (1-\eta(X))\phi(-z).$$

By definition,

$$Q_{\eta(X)}(f_\phi^*(X)) = \inf_z Q_{\eta(X)}(z).$$

Denote

$$\Delta Q_\eta(f) = Q_\eta(f) - Q_\eta(f_\phi^*)$$

where we suppress the dependence of η, f and f_ϕ^* on X for brevity.

Theorem 9 *Assume that ϕ is convex, $\phi'(\delta)$ and $\phi'(-\delta)$ both exist, $\phi'(\delta) < 0$, and (2) holds. In addition, suppose that there exist constants $C > 0$ and $s \geq 1$ such that*

$$\begin{aligned} |\eta - d|^s &\leq C^s \Delta Q_\eta(-\delta); \\ |(1-\eta) - d|^s &\leq C^s \Delta Q_\eta(\delta). \end{aligned}$$

Then

$$\Delta R[C(f,\delta)] \leq 2C\left[\Delta Q(f)\right]^{1/s}. \tag{5}$$

It is immediate from Theorem 9 that $\Delta Q(\hat{f}_n) \to_p 0$ implies $\Delta R(\hat{f}_n) \to_p 0$. In other words, consistency in terms of ϕ risk implies the consistency in terms of loss ℓ. It is worth noting that the constant in the upper bound can be tightened under stronger conditions.

Theorem 10 *In addition to the assumptions of Theorem 9, assume that*

$$\begin{aligned} (2\eta - 1)_+^s &\leq C^s \Delta Q_\eta(-\delta); \\ (1 - 2\eta)_+^s &\leq C^s \Delta Q_\eta(\delta). \end{aligned}$$

Then

$$\Delta R[C(f,\delta)] \leq C\left[\Delta Q(f)\right]^{1/s}.$$

We can improve the bounds even further by the following margin condition. Assume that for some $\alpha \geq 0$ and $A \geq 1$

$$\mathbb{P}\{|\eta(X) - z| \leq t\} \leq At^\alpha \tag{6}$$

for all $0 \leq t < d$ at $z = d$ and $z = 1 - d$. This assumption was introduced in Herbei and Wegkamp (2006) and generalizes the margin condition of Mammen and Tsybakov (1999) and Tsybakov (2004). It is always met for $\alpha = 0$ and $A = 1$. The other extreme is for $\alpha \to +\infty$ - the case where $\eta(X)$ stays away from d and $1 - d$ with probability one.

Theorem 11 *In addition to the assumptions of Theorem 9, assume that (6) holds for some $\alpha \geq 0$ and $A \geq 1$. Then, for some K depending on A and α,*

$$\Delta R\left[C(f,\delta)\right] \leq K\left[\Delta Q(f)\right]^{1/(s+\beta-\beta s)}, \tag{7}$$

where $\beta = \alpha/(1+\alpha)$.

In case $\alpha = 0$, the exponent $1/(s+\beta-\beta s)$ is $1/s$ on the right hand side in (7) above, and the situation is as in Theorem 9. For $\alpha \to +\infty$, the bound (7) improves upon the one in Theorem 9 as the exponent $1/(s+\beta-\beta s)$ converges to 1.

We now examine the consequences of Theorems 9, 10 and 11 on several common loss functions.

3.1 Least Squares

Note that for the least squares loss

$$\Delta Q_\eta(f) = (2\eta - 1 - f)^2.$$

Simple algebraic manipulations show that

$$\begin{aligned} \Delta Q_\eta(-\delta) &= 4|\eta - d|^2; \\ \Delta Q_\eta(\delta) &= 4|(1-\eta) - d|^2. \end{aligned}$$

Therefore, by Theorems 9 and 11,

Corollary 12 *For the least squares loss,*

$$\Delta R[C(f,1-2d)] \leq \left[\Delta Q(f)\right]^{1/2}.$$

Furthermore, if the margin condition (6) holds, then

$$\Delta R[C(f,1-2d)] \leq K\left[\Delta Q(f)\right]^{\frac{1+\alpha}{2+\alpha}}$$

for some constant $K > 0$.

3.2 Exponential Loss

An application of Taylor expansion yields (see, e.g., Zhang, 2004)

$$\Delta Q_\eta(f) \geq 2\left(\eta - \frac{1}{1+\exp(-2f)}\right)^2.$$

Therefore,

$$\begin{aligned} \Delta Q_\eta(-\delta) &\geq 2|\eta - d|^2; \\ \Delta Q_\eta(\delta) &\geq 2|(1-\eta) - d|^2. \end{aligned}$$

Therefore, by Theorems 9 and 11,

Corollary 13 *For the exponential loss,*

$$\Delta R\left[C\left(f,\frac{1}{2}\log\left(\frac{1}{d}-1\right)\right)\right] \leq \sqrt{2}\left[\Delta Q(f)\right]^{1/2}.$$

Furthermore, if the margin condition (6) holds, then

$$\Delta R\left[C\left(f,\frac{1}{2}\log\left(\frac{1}{d}-1\right)\right)\right] \leq K\left[\Delta Q(f)\right]^{\frac{1+\alpha}{2+\alpha}}$$

for some constant $K > 0$.

3.3 Logistic Loss

Similar to exponential loss, an application of Taylor expansion yields

$$\Delta Q_\eta(f) \geq 2\left(\eta - \frac{1}{1+\exp(-f)}\right)^2.$$

Therefore,

$$\begin{aligned} \Delta Q_\eta(-\delta) &\geq 2|\eta - d|^2; \\ \Delta Q_\eta(\delta) &\geq 2|(1-\eta) - d|^2. \end{aligned}$$

Therefore, by Theorems 9 and 11,

Corollary 14 *For the logistic loss,*

$$\Delta R\left[C\left(f,\log\left(\frac{1}{d}-1\right)\right)\right] \leq \sqrt{2}\left[\Delta Q(f)\right]^{1/2}.$$

Furthermore, if the margin condition (6) holds, then

$$\Delta R\left[C\left(f,\log\left(\frac{1}{d}-1\right)\right)\right] \leq K\left[\Delta Q(f)\right]^{\frac{1+\alpha}{2+\alpha}}$$

for some constant $K > 0$.

3.4 Squared Hinge Loss

Simple algebraic derivation shows

$$\Delta Q_\eta(f) = (2\eta - 1 - f)^2 - \eta(f-1)_+^2 - (1-\eta)(f+1)_-^2.$$

Therefore,

$$\begin{aligned} \Delta Q_\eta(-\delta) &= 4|\eta - d|^2; \\ \Delta Q_\eta(\delta) &= 4|(1-\eta) - d|^2. \end{aligned}$$

By Theorems 9 and 11,

Corollary 15 *For the squared hinge loss,*

$$\Delta R\left[C(f,1-2d)\right] \leq \left[\Delta Q(f)\right]^{1/2}.$$

Furthermore, if the margin condition (6) holds, then

$$\Delta R[C(f,1-2d)] \leq K\left[\Delta Q(f)\right]^{\frac{1+\alpha}{2+\alpha}}$$

for some constant $K > 0$.

3.5 Distance Weighted Discrimination

Observe that

$$Q_\eta(z) = \begin{cases} \frac{\eta}{z} + \frac{(1-\eta)z}{\gamma^2} + \frac{2(1-\eta)}{\gamma} & \text{if } z \geq \gamma \\ \frac{2}{\gamma} + \frac{z}{\gamma^2}(1-2\eta) & \text{if } |z| < \gamma \\ \frac{2\eta}{\gamma} - \frac{\eta z}{\gamma^2} - \frac{1-\eta}{z} & \text{if } z \leq -\gamma \end{cases}.$$

Hence

$$\inf Q_\eta(z) = \frac{2}{\gamma}\left(\sqrt{\eta(1-\eta)} + \min\{\eta, 1-\eta\}\right)$$

and

$$f_\phi^* = \begin{cases} (\eta/(1-\eta))^{1/2}\gamma & \text{if } \eta > 1/2 \\ \text{any value in } [-\gamma,\gamma] & \text{if } \eta = 1/2 \\ ((1-\eta)/\eta)^{1/2}\gamma & \text{if } \eta < 1/2 \end{cases}.$$

Recall that $\delta = ((1-d)/d)^{1/2}\gamma$. Then

$$\begin{aligned} \Delta Q_\eta(\delta) &\geq \left(\frac{\eta}{\delta} + \frac{(1-\eta)\delta}{\gamma^2} - 2\sqrt{\eta(1-\eta)}/\gamma\right) \\ &= \left(\left(\frac{\eta}{\delta}\right)^{1/2} - \left(\frac{(1-\eta)\delta}{\gamma^2}\right)^{1/2}\right)^2 \\ &= \frac{\eta\delta}{\gamma^2}\left(\left(\frac{d}{1-d}\right)^{1/2} - \left(\frac{1-\eta}{\eta}\right)^{1/2}\right)^2 \\ &= \frac{\eta\delta}{\gamma^2}\left(\left(\frac{d}{1-d}\right)^{1/2} + \left(\frac{1-\eta}{\eta}\right)^{1/2}\right)^{-2}\left(\frac{d}{1-d} - \frac{1-\eta}{\eta}\right)^2 \\ &= \frac{\delta}{\gamma^2(1-d)^2}\left[\left(\frac{d}{1-d}\right)^{1/2}\eta^{1/2} + (1-\eta)^{1/2}\right]^{-2}(1-\eta-d)^2. \end{aligned}$$

Observe that

$$\left(\frac{d}{1-d}\right)^{1/2}\eta^{1/2} + (1-\eta)^{1/2} \leq (1-d)^{-1/2}.$$

Thus,

$$(1-\eta-d)^2 \leq \gamma(1-d)^{1/2}d^{1/2}\Delta Q_\eta(\delta).$$

Similarly,

$$(\eta-d)^2 \leq \gamma(1-d)^{1/2}d^{1/2}\Delta Q_\eta(-\delta).$$

From Theorems 9 and 11, we conclude that

Corollary 16 *For the distance weighted discrimination loss,*

$$\Delta R\left[C(f,((1-d)/d)^{1/2}\gamma)\right] \leq \gamma^{1/2}(1-d)^{1/4}d^{1/4}[\Delta Q(f)]^{1/2}.$$

Furthermore, if the margin condition (6) holds, then

$$\Delta R\left[C(f,((1-d)/d)^{1/2}\gamma)\right] \leq K\,[\Delta Q(f)]^{\frac{1+\alpha}{2+\alpha}}$$

for some constant $K > 0$.

3.6 Hinge Loss with Rejection Option

As shown by Bartlett and Wegkamp (2008), for the modified hinge loss (4),

$$\arg\min_z Q_\eta(z) = \begin{cases} -1 & \text{if } \eta \le d \\ 0 & \text{if } d < \eta < 1-d \\ 1 & \text{if } \eta > 1-d \end{cases} .$$

Simple algebraic manipulations lead to

$$\Delta Q_\eta(-\delta) = \begin{cases} (1-\delta)(d-\eta)/d & \text{if } \eta \le d \\ (\eta-d)\delta/d & \text{if } d < \eta < 1-d \\ 1-(1-\eta)/d+(\eta-d)\delta/d & \text{if } \eta > 1-d \end{cases} ,$$

and

$$\Delta Q_\eta(\delta) = \begin{cases} 1-\eta/d+(1-\eta-d)\delta/d & \text{if } \eta \le d \\ (1-\eta-d)\delta/d & \text{if } d < \eta < 1-d \\ (\delta-1)(1-\eta-d)/d & \text{if } \eta > 1-d \end{cases} .$$

Therefore,

$$\begin{aligned} \frac{\min\{\delta,1-\delta\}}{d}|\eta-d| &\le \Delta Q_\eta(-\delta); \\ \frac{\min\{\delta,1-\delta\}}{d}|(1-\eta)-d| &\le \Delta Q_\eta(\delta). \end{aligned}$$

Furthermore,

$$\begin{aligned} \frac{\min\{\delta,1-\delta\}}{d}(2\eta-1)_+ &\le \Delta Q_\eta(-\delta); \\ \frac{\min\{\delta,1-\delta\}}{d}(1-2\eta)_+ &\le \Delta Q_\eta(\delta). \end{aligned}$$

From Theorems 10 and 11, we conclude that

Corollary 17 *For the modified hinge loss and any $\delta < 1$,*

$$\Delta R[C(f,\delta)] \le \frac{d}{\min\{\delta,1-\delta\}}\Delta Q(f). \tag{8}$$

Furthermore, if the margin condition (6) holds, then

$$\Delta R[C(f,\delta)] \le K\Delta Q(f) \tag{9}$$

for some constant $K > 0$.

Notice that the corollary also suggests that $\delta = 1/2$ yields the best constant $2d$ in the upper bound. A similar result has also been recently established by Bartlett and Wegkamp (2008). It is also interesting to see that (8) cannot be further improved by the generalized margin condition (6) as the bounds (8) and (9) only differ by a constant.

4. Rates of Convergence for Empirical Risk Minimizers

In this section we briefly review the possible rates of convergence for minimizers of the empirical risk $Q_n(f) = (1/n)\sum_{i=1}^n \phi(Y_i f(X_i))$ over a convex class of discriminant functions $\mathcal{F}$; and show the implications of the excess risk bounds obtained in the previous section. The analysis of the generalized hinge loss is complicated and is treated in detail in Wegkamp (2007) and Bartlett and Wegkamp (2008). The other loss functions ϕ considered in this paper have in common that the modulus of convexity of Q,

$$\delta(\varepsilon) = \inf\left\{ \frac{Q(f)+Q(g)}{2} - Q\left(\frac{f+g}{2}\right) : \ \mathbb{E}[(f-g)^2(X)] \geq \varepsilon^2 \right\}$$

satisfies $\delta(\varepsilon) \geq c\varepsilon^2$ for some $c > 0$ and that, for some $L < \infty$,

$$|\phi(x) - \phi(x')| \leq L|x - x'| \text{ for all } x, x' \in \mathbb{R}.$$

We have the following result that imposes a restriction on the $1/n$-covering number $N_n = N(1/n, L_\infty, \mathcal{F})$, the cardinality of the set of closed balls with radius $1/n$ in L_∞ needed to cover $\mathcal{F}$.

Theorem 18 *Assume that $|f| \leq B$ for all $f \in \mathcal{F}$ and let $0 < \gamma < 1$. With probability at least $1-\gamma$,*

$$Q(\hat{f}_n) \quad \leq \quad \inf_{f\in\mathcal{F}} Q(f) + \frac{3L}{n} + 8\left(\frac{L^2}{2c} + \frac{B}{6}\right)\frac{\log(N_n/\gamma)}{n}$$

Together with the excess risk bounds from Theorems 9 and 11, we have

Corollary 19 *Under the assumptions of Theorems 9 and 18, we have, with probability at least $1-\gamma$,*

$$\Delta R(C(\hat{f}_n, \delta)) \quad \leq \quad 2C\left\{ \inf_{f\in\mathcal{F}} \Delta Q(f) + \frac{3L}{n} + 8\left(\frac{L^2}{2c} + \frac{LB}{3}\right)\frac{\log(N_n/\gamma)}{n} \right\}^{1/s}.$$

Furthermore, if the generalized margin condition (6) holds, then with probability at least $1-\gamma$,

$$\Delta R(C(\hat{f}_n, \delta)) \quad \leq \quad K\left\{ \inf_{f\in\mathcal{F}} \Delta Q(f) + \frac{3L}{n} + 8\left(\frac{L^2}{2c} + \frac{LB}{3}\right)\frac{\log(N_n/\gamma)}{n} \right\}^{1/(s+\beta-\beta s)}$$

for some constant $K > 0$.

In the special case where $\mathcal{F}$ consists of linear combinations

$$\mathsf{f}_\lambda(x) = \sum_{j=1}^M \lambda_j f_j(x)$$

of simple discriminant functions (decision stumps) $f_1, \ldots, f_M$ with $\sum_{j=1}^M |\lambda_j| \leq B$ and $|f_j| \leq 1$, we obtain the rate $(M \log n/n)^{1/(s+\beta-\beta s)}$. We can view B as a tuning parameter here, and if the functions f_j are near orthogonal in the sense that

$$\max_{1\leq i\neq j\leq M} \frac{\mathbb{E}[f_i(X) f_j(X)]}{\sqrt{\mathbb{E}[f_i^2(X)]\mathbb{E}[f_j^2(X)]}} \leq \frac{c}{|\lambda_0|_0}$$

for some small $c > 0$, a small modification of Theorem 1 in Wegkamp (2007) shows that we also adapt to the unknown sparsity of the minimizer λ_0 of $Q(\mathsf{f}_\lambda)$ over λ in that the rate becomes $(|\lambda_0|_0 \log n/n)^{1/(s+\beta-\beta s)}$ for suitably chosen $B = B(n)$.

5. Asymmetric Loss

We have focused thus far on the case where misclassifying from one class to the other, either $g(X) = 1$ while $Y = -1$ or $g(X) = -1$ while $Y = 1$, is assigned the same loss. In many applications, however, one type of misclassification may incur a heavier loss than the other. Such situations naturally arise in risk management or medical diagonsis. To this end, the following loss function can be adopted in place of ℓ:

$$\ell_\theta[g(X),Y] = \begin{cases} 1 & \text{if } g(X) = -1 \text{ and } Y = 1 \\ \theta & \text{if } g(X) = 1 \text{ and } Y = -1 \\ d & \text{if } g(X) = 0 \\ 0 & \text{if } g(X) = Y \end{cases}.$$

We shall assume that $\theta < 1$ without loss of generality. It can be shown that the rejection option is only available if $d < \theta/(1+\theta)$ (see, e.g., Herbei and Wegkamp, 2006), which we shall assume throughout the section. When this holds, the corresponding Bayes rule is given by (see, e.g., Herbei and Wegkamp, 2006)

$$g^*_\theta(X) = \begin{cases} 1 & \text{if } \eta(X) > 1 - d/\theta \\ 0 & \text{if } d \le \eta(X) \le 1 - d/\theta \\ -1 & \text{if } \eta(X) < d \end{cases}.$$

Instead of $C(\hat{f}_n, \delta)$, an asymmetrically truncated classification rule, $\hat{f}_n$, $C(\hat{f}_n; \delta_1, \delta_2)$, can be used for our purpose here where

$$C(f(X); \delta_1, \delta_2) = \begin{cases} 1 & \text{if } f(X) > \delta_1 \\ 0 & \text{if } -\delta_2 \le f(X) \le \delta_1 \\ -1 & \text{if } f(X) < -\delta_2 \end{cases}.$$

The behavior of the asymmetically truncated classification rule $C(\hat{f}_n; \delta_1, \delta_2)$ can be studied in a similar fashion as before. In particular, we have the following results in parallel to Theorems 1 and 9.

Theorem 20 *Assume that ϕ is convex. Then $C(f^*_\phi, \delta_1, \delta_2)$ for some $\delta_1, \delta_2 > 0$ is infinite sample consistent, that is, $C(f^*_\phi, \delta_1, \delta_2) = g^*_\theta$ if and only if $\phi'(\pm\delta_1)$ and $\phi'(\pm\delta_2)$ exist; $\phi'(\delta_1), \phi'(\delta_2) < 0$; and*

$$\begin{aligned} \frac{\phi'(\delta_1)}{\phi'(-\delta_1) + \phi'(\delta_1)} &= \frac{d}{\theta}; \\ \frac{\phi'(\delta_2)}{\phi'(-\delta_2) + \phi'(\delta_2)} &= d. \end{aligned}$$

Furthermore, if $C(f^*_\phi, \delta_1, \delta_2)$ is infinite sample consistent and

$$\begin{aligned} |\theta(1-\eta) - d|^s &\le C^s \Delta Q_\eta(\delta_1); \\ |\eta - d|^s &\le C^s \Delta Q_\eta(-\delta_2), \end{aligned}$$

then

$$\Delta R_\theta[C(f, \delta_1, \delta_2)] \le 2C[\Delta Q(f)]^{1/s},$$

where $\Delta R_\theta(g) = R_\theta(g) - R_\theta(g^*_\theta)$ and $R_\theta(g) = \mathbb{E}[\ell_\theta(g(X), Y)]$.

Theorem 20 can be proved in the same fashion as Theorems 1 and 9 and is therefore omitted for brevity.

6. Proofs

Proof of Theorem 1. We first show the "if" part. Recall that

$$\begin{aligned} Q(f) &= \mathbb{E}[\phi(Yf(X))] \\ &= \mathbb{E}\left(\mathbb{E}\left[\phi(Yf(X))|X\right]\right) \\ &= \mathbb{E}\left[\eta(X)\phi(f(X)) + (1-\eta(X))\,\phi(-f(X))\right]. \end{aligned}$$

With slight abuse of notation, write

$$Q_{\eta(X)}(f(X)) = \eta(X)\phi(f(X)) + (1-\eta(X))\,\phi(-f(X)).$$

Then $f_\phi^*(X)$ minimizes $Q_{\eta(X)}(\cdot)$.

We now proceed by separately considering three different scenarios: (a) $\eta(X) < d$; (b) $\eta(X) > 1-d$; and (c) $d < \eta(X) < 1-d$. For brevity, we shall abbreviate the dependence of η and f_ϕ^* on X in the reminder of the proof when no confusion occurs.

First consider the case when $\eta < d$. Recall that $\phi'(-\delta) < \phi'(\delta) < 0$, and

$$\frac{\phi'(\delta)}{\phi'(-\delta)+\phi'(\delta)} = d.$$

Therefore,

$$\eta\phi'(-\delta) - (1-\eta)\phi'(\delta) > 0.$$

By the convexity of ϕ, for any $z > 0$,

$$\begin{aligned} \phi(z-\delta) - \phi(-\delta) &\geq \phi'(-\delta)z; \\ \phi(-z+\delta) - \phi(\delta) &\geq -\phi'(\delta)z. \end{aligned}$$

Hence

$$Q_\eta(z-\delta) - Q_\eta(-\delta) \geq \left[\eta\phi'(-\delta) - (1-\eta)\phi'(\delta)\right] z > 0,$$

which implies that $f_\phi^* \leq -\delta$.

It now suffices to show that $f_\phi^* \neq -\delta$. By the definition of $\phi'(-\delta)$ and $\phi'(\delta)$, for any $\varepsilon > 0$, there exists a $\zeta > 0$ such that for any $0 < z < \zeta$,

$$\begin{aligned} \frac{\phi(-z-\delta) - \phi(-\delta)}{-z} &\geq \phi'(-\delta) - \varepsilon; \\ \frac{\phi(z+\delta) - \phi(\delta)}{z} &\leq \phi'(\delta) + \varepsilon. \end{aligned}$$

Therefore for any $0 < z < \zeta$,

$$\begin{aligned} Q_\eta(-z-\delta) - Q_\eta(-\delta) &= \eta\left[\phi(-z-\delta) - \phi(-\delta)\right] + (1-\eta)\left[\phi(z+\delta) - \phi(\delta)\right] \\ &\leq -\eta\left[\phi'(-\delta) - \varepsilon\right] z + (1-\eta)\left[\phi'(\delta) + \varepsilon\right] z \\ &= \left(\left[(1-\eta)\phi'(\delta) - \eta\phi'(-\delta)\right] + \varepsilon\right) z. \end{aligned}$$

Recall that $(1-\eta)\phi'(\delta) - \eta\phi'(-\delta) < 0$. By setting ε small enough, we can ensure that $(1-\eta)\phi'(\delta) - \eta\phi'(-\delta) + \varepsilon$ remains negative. Hence

$$Q_\eta(-z-\delta) < Q_\eta(-\delta),$$

which implies that $f_\phi^* \neq -\delta$.

Now consider the case when $\eta > 1-d$. Observe that $Q_\eta(z) = Q_{1-\eta}(-z)$. From the previous discussion,

$$f_\phi^* = \arg\min_z Q_\eta(z) = -\arg\min_z Q_{1-\eta}(-z) > \delta.$$

At last, consider the case when $d < \eta < 1-d$. Observe that in this case,

$$\eta\phi'(\delta) - (1-\eta)\phi'(-\delta) > 0;$$
$$\eta\phi'(-\delta) - (1-\eta)\phi'(\delta) < 0.$$

Hence for any $z > 0$,

$$Q_\eta(z+\delta) - Q_\eta(\delta) \geq \left[\eta\phi'(\delta) - (1-\eta)\phi'(-\delta)\right] z > 0,$$

which implies that $f_\phi^* \leq \delta$. Similarly,

$$Q_\eta(-z-\delta) - Q_\eta(-\delta) \geq \left[-\eta\phi'(-\delta) + (1-\eta)\phi'(\delta)\right] z > 0,$$

which implies that $f_\phi^* \geq -\delta$. In summary, $f_\phi^* \in [-\delta, \delta]$.

We now consider the "only if" part. Let $[a_-, b_-]$ and $[a_+, b_+]$ be the subdifferential of ϕ at $-\delta$ and δ respectively. We need to show that $a_- = b_-$, $a_+ = b_+$ and $a_+/(a_+ + a_-) = d$. We begin by showing that $b_+ \leq 0$. Assume the contrary. The infinite sample consistency implies that for any $\eta > 1-d$, $f_\phi^* > \delta$. Because $b_+ > 0$, we have $\phi(f_\phi^*) > \phi(\delta)$. Together with the fact that $Q_\eta(f_\phi^*) < Q_\eta(\delta)$, this implies that $\phi(-f_\phi^*) < \phi(-\delta)$. Subsequently, we have $a_- > 0$. The convexity of ϕ also suggests that $a_- \leq a_+ \leq b_- \leq b_+$. Because

$$\begin{aligned} \phi(f_\phi^*) - \phi(\delta) &\geq b_+(f_\phi^* - \delta); \\ \phi(-\delta) - \phi(-f_\phi^*) &\leq a_-(f_\phi^* - \delta), \end{aligned}$$

we have

$$Q_\eta(f_\phi^*) - Q_\eta(\delta) \geq (\eta b_+ - (1-\eta)a_-)(f_\phi^* - \delta) > 0.$$

This contradiction suggests that $b_+ \leq 0$.

Given that $a_- \leq a_+ \leq b_- \leq b_+ \leq 0$, we have $|a_-| \geq |a_+| \geq |b_-| \geq |b_+|$, which implies that

$$\frac{b_+}{a_- + b_+} \leq \frac{a_+}{b_- + a_+}.$$

It suffices to show that

$$\frac{b_+}{a_- + b_+} \geq d \qquad \text{and} \qquad \frac{a_+}{b_- + a_+} \leq d.$$

Assume the contrary. First consider the case when $b_+/(a_- + b_+) < d$. Let η be such that $b_+/(a_- + b_+) < \eta < d$. By definition, for any $f < -\delta$,

$$\begin{aligned} \phi(f) - \phi(-\delta) &\geq a_-(f + \delta); \\ \phi(-f) - \phi(\delta) &\geq b_+(-f - \delta). \end{aligned}$$

Hence

$$Q_\eta(f) - Q_\eta(-\delta) \geq [\eta a_- - (1-\eta)b_+](f + \delta) > 0,$$

which implies that $\arg\min Q_\eta(z) \geq -\delta$. This contradicts with the infinite sample consistency. Therefore, $b_+/(a_-+b_+) \geq d$. Next we deal with the case of $a_+/(b_-+a_+) > d$. Let η be such that $a_+/(b_-+a_+) > \eta > d$. Following a similar argument as before, one can show that $Q_{1-\eta}(f) - Q_{1-\eta}(\delta) > 0$ for any $f > \delta$, which implies that $\arg\min Q_\eta(z) \leq \delta$. This again contradicts infinite sample consistency because $1-\eta < 1-d$. Therefore, $a_+/(b_-+a_+) \leq d$.

The proof is now concluded. ∎

Proof of Theorem 9. Recall that

$$Q_\eta(f) = \eta\phi(f) + (1-\eta)\phi(-f).$$

Similarly, write

$$R_\eta[C(f,\delta)] = \eta\ell(C(f,\delta),1) + (1-\eta)\ell(C(f,\delta),-1).$$

Also write $\Delta Q_\eta(f) = Q_\eta(f) - \inf Q_\eta(f)$ and $\Delta R_\eta(f) = R_\eta(f) - \inf R_\eta(f)$. It suffices to show that

$$\Delta R_\eta[C(f,\delta)] \leq 2C\left[\Delta Q_\eta(f)\right]^{1/s}. \tag{10}$$

The theorem can be deduced from (10) by Jensen's inequality:

$$\begin{aligned}
\Delta R[C(f,\delta)] &= \mathbb{E}\left[\Delta R_{\eta(X)}[C(f(X),\delta)]\right] \\
&\leq 2C\mathbb{E}\left[\Delta Q_{\eta(X)}(f(X))\right]^{1/s} \\
&\leq 2C\left(\mathbb{E}\left[\Delta Q_{\eta(X)}(f(X))\right]\right)^{1/s} \\
&= 2C\left[\Delta Q(f)\right]^{1/s}.
\end{aligned}$$

To show (10), we consider separately the different combinations of values of η and f. For brevity, we shall abbreviate their dependence on X in what follows.

Case 1. $\eta < d$ and $f < -\delta$. As shown before, in this case $f^*_\phi(X) < -\delta$. Thus,

$$\Delta R_\eta[C(f,\delta)] = 0 \leq C\left[\Delta Q_\eta(f)\right]^{1/s}.$$

Case 2. $\eta < d$ and $|f| < \delta$. Observe that

$$Q_\eta(f) - Q_\eta(-\delta) \geq \left[\eta\phi'(-\delta) - (1-\eta)\phi'(\delta)\right](f+\delta) = \frac{-\phi'(\delta)}{d}(d-\eta)(f+\delta) \geq 0.$$

Together with the fact that $C^s\Delta Q_\eta(-\delta) \geq |\eta-d|^s$, we have

$$\Delta Q_\eta(f) \geq \Delta Q_\eta(-\delta) \geq C^{-s}|\eta-d|^s.$$

Note that

$$\Delta R_\eta[C(f,\delta)] = d-\eta.$$

We have

$$\Delta R_\eta[C(f,\delta)] \leq C\left[\Delta Q_\eta(f)\right]^{1/s}.$$

Case 3. $\eta < d$ and $f > \delta$. Observe that

$$Q_\eta(f) - Q_\eta(\delta) \geq \left[\eta\phi'(\delta) - (1-\eta)\phi'(-\delta)\right](f-\delta) = \frac{-\phi'(\delta)}{d}(1-\eta-d)(f-\delta) \geq 0.$$

Together with the facts that $d < 1/2$ and $C^s \Delta Q_\eta(\delta) \geq |1-\eta-d|^s$, we have

$$\Delta Q_\eta(f) \geq \Delta Q_\eta(\delta) \geq C^{-s}|1-\eta-d|^s \geq (2C)^{-s}|1-2\eta|^s.$$

Note that

$$\Delta R_\eta[C(f,\delta)] = 1-2\eta.$$

Therefore,

$$\Delta R_\eta[C(f,\delta)] \leq 2C\left[\Delta Q_\eta(f)\right]^{1/s}.$$

Case 4. $d < \eta < 1-d$ and $f < -\delta$. Following a similar argument as before,

$$Q_\eta(f) - Q_\eta(-\delta) \geq \frac{-\phi'(\delta)}{d}(d-\eta)(f+\delta) \geq 0.$$

Therefore,

$$\Delta Q_\eta(f) \geq \Delta Q_\eta(-\delta) \geq C^{-s}|\eta-d|^s,$$

which, together with the fact that $\Delta R_\eta[C(f,\delta)] = \eta - d$, implies that

$$\Delta R_\eta[C(f,\delta)] \leq C\left[\Delta Q_\eta(f)\right]^{1/s}.$$

Case 5. $d < \eta < 1-d$ and $|f| < \delta$. In this case,

$$\Delta R_\eta[C(f,\delta)] = 0 \leq C\left[\Delta Q_\eta(f)\right]^{1/s}.$$

Case 6. $d < \eta < 1-d$ and $f > \delta$. Observe that

$$Q_\eta(f) - Q_\eta(\delta) \geq \frac{-\phi'(\delta)}{d}(1-\eta-d)(f-\delta) \geq 0.$$

Hence

$$\Delta Q_\eta(f) \geq \Delta Q_\eta(\delta) \geq C^{-s}|1-\eta-d|^s,$$

which, together with the fact that $\Delta R_\eta[C(f,\delta)] = 1-\eta-d$, implies that

$$\Delta R_\eta[C(f,\delta)] \leq C\left[\Delta Q_\eta(f)\right]^{1/s}.$$

Case 7. $\eta > 1-d$. Observe that $R_\eta[C(f,\delta)] = R_{1-\eta}[C(-f,\delta)]$, and $Q_\eta(f) = Q_{1-\eta}(-f)$. Because $1-\eta < d$, from Cases 1, 2 and 3, we have

$$\begin{aligned}
\Delta R_\eta[C(f,\delta)] &= \Delta R_{1-\eta}[C(-f,\delta)] \\
&\leq 2C\left[\Delta Q_{1-\eta}(-f)\right]^{1/s} \\
&= 2C\left[\Delta Q_\eta(f)\right]^{1/s}.
\end{aligned}$$

The proof is therefore completed. ∎

Proof of Theorem 10. The proof follows from the same argument as that of Theorem 9. The only difference takes place in Case 3 where under the current assumptions

$$\Delta R_\eta(f) = 1 - 2\eta \le C\left[\Delta Q_\eta(\delta)\right]^{1/s}. \blacksquare$$

Proof of Theorem 11. The last part of the proof is based on the proof of Theorem 3 in Bartlett, Jordan and McAuliffe (2006). Let $g = C(f,\delta)$ be the classification rule with reject option based on $f : \mathcal{X} \to \mathbb{R}$ and set $g^* = C(f_\phi^*, \delta)$. We have shown above that under the assumptions of Theorem 9,

$$\begin{aligned}
|d - \eta| 1\{g \neq g^*\} \left(1\{g = -1\} + 1\{g^* = -1\}\right) &\le C\left[\Delta Q_\eta(f)\right]^{1/s} \\
|1 - d - \eta| 1\{g \neq g^*\} \left(1\{g = 1\} + 1\{g^* = 1\}\right) &\le C\left[\Delta Q_\eta(f)\right]^{1/s}.
\end{aligned}$$

Moreover, Lemma 1 in Herbei and Wegkamp (2006) states that

$$\begin{aligned}
\Delta R(g) \quad = \quad & \mathbb{E}\left[|d - \eta(X)| 1\{g(X) \neq g^*(X)\} \left(1\{g(X) = -1\} + 1\{g^*(X) = -1\}\right)\right] \\
& + \mathbb{E}\left[|1 - d - \eta(X)| 1\{g(X) \neq g^*(X)\} \left(1\{g(X) = 1\} + 1\{g^*(X) = 1\}\right)\right].
\end{aligned} \tag{11}$$

Hence, for any $\varepsilon > 0$,

$$\begin{aligned}
&\Delta R(g) \\
&= \mathbb{E}\left[|d - \eta(X)| 1\{|d - \eta(X)| \le \varepsilon\} 1\{g(X) \neq g^*(X)\} \left(1\{g(X) = -1\} + 1\{g^*(X) = -1\}\right)\right] \\
&+ \mathbb{E}\left[|d - \eta(X)| 1\{|d - \eta(X)| > \varepsilon\} 1\{g(X) \neq g^*(X)\} \left(1\{g(X) = -1\} + 1\{g^*(X) = -1\}\right)\right] \\
&+ \mathbb{E}\left[|1 - d - \eta(X)| 1\{|1 - d - \eta(X)| \le \varepsilon\} 1\{g(X) \neq g^*(X)\} \left(1\{g(X) = 1\} + 1\{g^*(X) = 1\}\right)\right] \\
&+ \mathbb{E}\left[|1 - d - \eta(X)| 1\{|1 - d - \eta(X)| > \varepsilon\} 1\{g(X) \neq g^*(X)\} \left(1\{g(X) = 1\} + 1\{g^*(X) = 1\}\right)\right] \\
&\le 2\varepsilon \mathbb{P}\{g^*(X) \neq g(X)\} + 2\varepsilon^{1-s} \Delta Q(f)
\end{aligned}$$

where we used (11) and the inequality $|x| 1\{|x| \ge \varepsilon\} \le |x|^r \varepsilon^{1-r}$ for $r \ge 1$. Using the bound

$$\mathbb{P}\{g(X) \neq g^*(X)\} \le \left[2(8A)^{1/\alpha} \Delta R(g)\right]^\beta$$

from the proof of Lemma 4 of Herbei and Wegkamp (2006), and choosing

$$\varepsilon = c\left[\Delta R(g)\right]^{1-\beta}$$

with $c = [2(8A)^{1/\alpha}]^\beta/4$ readily gives the desired claim with $K = 4Cc^{1-s}$. $\blacksquare$

Proof of Theorem 18. Recall that $\bar{f} \in \mathcal{F}$ minimizes $Q(f)$ over $f \in \mathcal{F}$. Let $h(yf(x)) = \phi(yf(x)) - \phi(y\bar{f}(x))$. Since

$$\begin{aligned}
\frac{Q(f) + Q(\bar{f})}{2} \quad &\ge \quad Q\left(\frac{f + \bar{f}}{2}\right) + c\mathbb{E}\left[(f - \bar{f})^2(X)\right] \\
&\ge \quad Q(\bar{f}) + c\mathbb{E}\left[(f - \bar{f})^2(X)\right],
\end{aligned}$$

we have

$$\begin{aligned}
\mathbb{E}\left[h^2(Yf(X))\right] \quad &\le \quad L^2 \mathbb{E}\left[(f - \bar{f})^2(X)\right] \\
&\le \quad \frac{L^2}{2c}\{Q(f) - Q(\bar{f})\} \\
&= \quad \frac{L^2}{2c}\mathbb{E}\left[h(Yf(X))\right],
\end{aligned}$$

see, for example, Bartlett, Jordan and McAuliffe (2006). Since $\widehat{f}_n$ minimizes $Q_n(f)$, we have

$$\begin{aligned} Q(\widehat{f}_n) - Q(\overline{f}) &= Ph(y\widehat{f}_n(x)) \\ &= 2\mathbb{P}_n h(Y\widehat{f}_n(X)) + (P-2\mathbb{P}_n)h(Y\widehat{f}_n(X)) \\ &\le 2\mathbb{P}_n h(Y\overline{f}(X)) + (P-2\mathbb{P}_n)h(Y\widehat{f}_n(X)) \\ &\le \sup_{f\in\mathcal{F}}(P-2\mathbb{P}_n)h(Yf(X)) \end{aligned}$$

where $\mathbb{P}h(Yf(X)) = \mathbb{E}[h(Yf(X))]$ and $\mathbb{P}_n h(Yf(X)) = (1/n)\sum_{i=1}^n h(Y_i f(X_i))$ for any $f \in \mathcal{F}$. Next we observe that

$$\sup_{f\in\mathcal{F}}(\mathbb{P}-2\mathbb{P}_n)h(Yf(X)) \le \frac{3L}{n} + \max_{f\in\mathcal{F}_n}(\mathbb{P}-2\mathbb{P}_n)h(Yf(X))$$

where $\mathcal{F}_n$ is the minimal $1/n$-net of $\mathcal{F}$. By Bernstein's inequality, we get

$$\begin{aligned} \mathbb{P}\left\{\sup_{f\in\mathcal{F}_n}(\mathbb{P}-2\mathbb{P}_n)h(Yf(X)) \ge t\right\} &\le N_n \exp\left[-\frac{n\{t+\mathbb{P}h(Yf(X))\}^2/8}{\mathbb{P}h^2(Yf(X)) + (2LB)\{t+\mathbb{P}h(Yf(X))\}/6}\right] \\ &\le N_n \exp\left[-\frac{nt}{8}\left(\frac{L^2}{2c}+\frac{LB}{3}\right)^{-1}\right] \end{aligned}$$

and the conclusion follows easily. ∎

Acknowledgments

The authors gratefully acknowledge the fact that part of the research was done while the authors were visiting the Isaac Newton Institute for Mathematical Sciences (Statistical Theory and Methods for Complex, High-Dimensional Data Programme) at Cambridge University during Spring 2008. The research of Ming Yuan was supported in part by NSF grants DMS-MPSA-0624841 and DMS-0846234. The research of Marten Wegkamp was supported in part by NSF Grant DMS-0706829.

References

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

P.L. Bartlett and M.H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9:1823-1840, 2008.

J. Friedman, T. Hastie and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28(2):337-407, 2000.

R. Herbei and M.H. Wegkamp. Classification with reject option. *Canadian Journal of Statistics*, 34(4):709-721, 2006.

Y. Lin. Support vector machines and the Bayes rule in classification. *Machine Learning and Knowledge Discovery*, 6:259-275, 2002.

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

J.S. Marron, M. Todd and J. Ahn. Distance weighted discrimination. *Journal of the American Statistical Association*, 102:1267-1271, 2007.

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

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

M.H. Wegkamp. Lasso type classifiers with a reject option. *Electronic Journal of Statistics*, 1:155-168, 2007.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32:56-134, 2004.