Bernoulli **17**(4), 2011, 1368–1385
DOI: 10.3150/10-BEJ320

Support vector machines with a reject option

MARTEN WEGKAMP[1] and MING YUAN[2]

[1]*Department of Mathematics and Department of Statistical Science, Cornell University, Ithaca, NY 14853, USA. E-mail: marten.wegkamp@cornell.edu*
[2]*Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA. E-mail: myuan@isye.gatech.edu*

This paper studies ℓ_1 regularization with high-dimensional features for support vector machines with a built-in reject option (meaning that the decision of classifying an observation can be withheld at a cost lower than that of misclassification). The procedure can be conveniently implemented as a linear program and computed using standard software. We prove that the minimizer of the penalized population risk favors sparse solutions and show that the behavior of the empirical risk minimizer mimics that of the population risk minimizer. We also introduce a notion of classification complexity and prove that our minimizers adapt to the unknown complexity. Using a novel oracle inequality for the excess risk, we identify situations where fast rates of convergence occur.

Keywords: adaptive prediction; classification with a reject option; lasso; oracle inequalities; sparsity; support vector machines; statistical learning

1. Introduction

In this paper we further investigate the new classification rules introduced in [1, 11] with a built-in reject option in the standard binary classification setting, where we observe independent realizations (X_i, Y_i), $i = 1, \ldots, n$, of a random pair (X, Y) in $\mathcal{X} \times \{-1, +1\}$ (here, $\mathcal{X}$ is an arbitrary space). A discriminant function $f : \mathcal{X} \to \mathbb{R}$ classifies an observation $x \in \mathcal{X}$ into one of two classes, labeled -1 or $+1$. Viewing $f(x)$ as a proxy value of the conditional probability $\eta(x) = \mathbb{P}\{Y = 1 | X = x\}$, we are less confident for small values of $|f(x)|$, corresponding to $\eta(x)$ near $1/2$. Our strategy is to report $\operatorname{sgn}(f(x)) \in \{-1, 1\}$ if $|f(x)|$ exceeds some prescribed threshold τ and withhold decision otherwise. Assuming that the cost of making a wrong decision is 1 and that of withholding a decision is d, the appropriate risk function is

$$R_\ell(f) = \mathbb{E}[\ell(Yf(X))] = \mathbb{P}\{Yf(X) < -\tau\} + d\mathbb{P}\{|Yf(X)| \le \tau\}$$

with the discontinuous loss function

$$\ell(z) = \begin{cases} 1, & \text{if } z < -\tau, \\ d, & \text{if } |z| \le \tau, \\ 0, & \text{otherwise.} \end{cases}$$

Since we always reject if $d = 0$ and never reject if $d \ge 1/2$ (see [5]), we take $0 < d \le 1/2$ in what follows without loss of generality. Although the minimizer of this risk is not unique, all such minimizers correspond to the unique classification rule that assigns $-1, +1$ or withhold decision, depending on which of $1-\eta$, η or d is smallest. The smallest risk is $\mathbb{E}[\min\{\eta(X), 1-\eta(X), d\}]$ and we may interpret the cost d as the largest conditional probability of misclassification that is considered tolerable.

In practice, minimization of the empirical counterpart $\hat{R}_\ell(f) = (1/n)\sum_{i=1}^n \ell(Y_i f(X_i))$ of $R_\ell(f)$ over a large class of functions f is computationally not feasible. For this reason, we could replace the loss function ℓ by a convex surrogate loss function and consider discriminant functions f of the form $\mathsf{f}_\lambda(x) = \sum_{j=1}^M \lambda_j f_j(x)$ based on a set of known functions $f_j : \mathcal{X} \to \mathbb{R}$ and coefficients $\lambda_j \in \mathbb{R}$, $1 \le j \le M$. Following [1], we will consider the generalized hinge loss

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

with slope $a = (1-d)/d > 1$. Observe that $\phi(z)$ is piecewise linear, so that minimization of the empirical risk

$$\hat{R}_\phi(\mathsf{f}_\lambda) = \frac{1}{n}\sum_{i=1}^n \phi(Y_i \mathsf{f}_\lambda(X_i)) \tag{1.1}$$

can be solved by a tractable linear program. Crucial for the choice of $\phi(z)$ is that it is classification calibrated: the unique minimizer

$$f_0(x) = \begin{cases} -1, & \text{if } \eta(x) < d, \\ 0, & \text{if } d \le \eta(x) \le 1-d, \\ +1, & \text{if } \eta(x) > 1-d \end{cases}$$

of $R_\phi(f) = \mathbb{E}[\phi(Yf(X))]$ also minimizes the risk $R_\ell(f) = \mathbb{E}[\ell(Yf(X))]$ over all measurable $f : \mathcal{X} \to \mathbb{R}$ for all $\tau < 1$; see, for example, [1, 12].

At this point it is important to note that truncating the minimizer $\operatorname{sgn}(2\eta - 1)$ of the hinge-loss-based risk $\mathbb{E}(1 - Yf(X))_+$ does not yield the optimal rule for any positive threshold τ. This is the reason why we generalize the hinge loss instead. In addition to the generalized hinge loss, there are also other choices of the surrogate loss function and corresponding truncation value τ that are classification calibrated. The treatment for the generalized hinge loss differs considerably from that for other losses, such as the logistic, exponential and quadratic loss, which are smoother. We refer to [12] for a detailed discussion.

Observe that $\phi(z) \geq \ell(z)$ for all $\tau \leq 1-d$ and, subsequently, $\mathbb{E}[\ell(Yf(X))] \leq \mathbb{E}[\phi(Yf(X))]$. It is shown in [1] that a similar relationship remains true for the excess risks, that is, the inequality

$$\mathbb{E}[\ell(Yf(X))] - \mathbb{E}[\ell(Yf_0(X))] \leq \mathbb{E}[\phi(Yf(X))] - \mathbb{E}[\phi(Yf_0(X))]$$

holds for all $d \leq \tau \leq 1-d$. This property is useful for deriving oracle inequalities in terms of the ℓ-risk since minimization of (1.1) produces oracle inequalities in terms of the ϕ-risk rather than the ℓ-risk directly.

Of particular interest here is the case where the number of basis functions, M, is large when compared with the sample size n. Usually, the minimization of the empirical risk $\hat{R}_\phi(\mathsf{f}_\lambda)$ is computed under a restriction on the quadratic term $\sum_{j=1}^M \lambda_j^2$. Here, we opt instead for an ℓ_1-type restriction $\|\lambda\|_{\ell_1} := \sum_{j=1}^M |\lambda_j|$ and estimate f_λ by $\mathsf{f}_{\hat{\lambda}(r)}$, where

$$\hat{\lambda}(r) := \arg\min_{\lambda \in \mathbb{R}^M} (\hat{R}_\phi(\mathsf{f}_\lambda) + r\|\lambda\|_{\ell_1}) \tag{1.2}$$

and $r > 0$ is a tuning parameter. The choice of an ℓ_1 penalty reflects our preference for sparse solutions, which is desirable when M is large.

In the remainder of this paper, we study the properties of $\hat{\lambda}(r)$ and its population counterpart,

$$\lambda(r) := \arg\min_{\lambda \in \mathbb{R}^M} (R_\phi(\mathsf{f}_\lambda) + r\|\lambda\|_{\ell_1}). \tag{1.3}$$

We establish oracle inequalities for $\lambda(r)$ and $\hat{\lambda}(r)$ in Sections 2 and 3, respectively. The results that we obtain are similar in spirit to those from [6, 8, 11]. However, [8, 11] do not discuss properties of $\lambda(r)$, and our results in Section 2 obtained here extend those proved by [6] in the context of twice differentiable loss functions. Furthermore, the oracle inequalities for the penalized empirical risk minimizer $\hat{\lambda}(r)$ in Section 3 are much sharper than earlier results from [11] for $0 \leq d \leq 1/2$ and [8] for $d = 1/2$. In particular, the new inequality reveals that the rate of convergence of the excess risk of $\mathsf{f}_{\hat{\lambda}}$ can be even faster than $1/n$ if the optimal discriminant function f_0 can be written as a linear combination of the f_j's in the dictionary. Moreover, we relax the condition on the dictionary and do not require that the parameter λ is bounded. We emphasize that our results hold, in particular, for $d = 1/2$, the case of support vector machines without a reject option, and generalize and extend the results obtained in [8]. In addition, novel empirical bounds on the error and reject rate are given. To demonstrate the feasibility of the ℓ_1-regularized support vector machine with a reject option, in Section 4 we formulate $\hat{\lambda}(r)$ as a solution of a linear program and report some numerical experiments. Some technical lemmas and a maximal inequality for a weighted empirical process are collected in the Appendix.

2. Properties of the theoretical solution

We begin by studying $\lambda(r)$, the population version of $\hat{\lambda}(r)$. Recall that $\lambda(r)$ is defined by

$$\lambda(r) = \arg\min_{\lambda \in \mathbb{R}^M} \{R_\phi(\mathsf{f}_\lambda) + r\|\lambda\|_{\ell_1}\}. \tag{2.1}$$

In particular, $\lambda(0)$ minimizes the risk $R_\phi(\mathsf{f}_\lambda)$ over $\lambda \in \mathbb{R}^M$. By definition, we find that

$$R_\phi(\mathsf{f}_{\lambda(r)}) + r\|\lambda(r)\|_{\ell_1} \le R_\phi(\mathsf{f}_\lambda) + r\|\lambda\|_{\ell_1} \tag{2.2}$$

holds for all $\lambda \in \mathbb{R}^M$. This inequality applied to $\lambda = \lambda(0)$ has the following consequences.

Proposition 2.1. *Let* $I_0 = \{i : \lambda_i(0) \neq 0\}$ *be the support of* $\lambda(0)$.

(a) *If* $\|\lambda(0)\|_{\ell_1} = o(1/r)$ *as* $r \to 0$, *then* $R_\phi(\lambda(r)) \to R_\phi(\lambda(0))$ *as* $r \to 0$.
(b) $\|\lambda(r)\|_{\ell_1} \le \|\lambda(0)\|_{\ell_1}$ *for all* $r > 0$.
(c) $\sum_{j \notin I_0} |\lambda_j(r) - \lambda_j(0)| \le \sum_{j \in I_0} |\lambda_j(r) - \lambda_j(0)|$.

Proof. After applying inequality (2.2) to $\lambda = \lambda(0)$ and using the fact that $R_\phi(\mathsf{f}_{\lambda(0)}) \le R_\phi(\mathsf{f}_{\lambda(r)})$, we get

$$0 \le R_\phi(\mathsf{f}_{\lambda(r)}) - R_\phi(\mathsf{f}_{\lambda(0)}) \le r\|\lambda(0)\|_{\ell_1} - r\|\lambda(r)\|_{\ell_1} \le r\|\lambda(0)\|_{\ell_1},$$

which implies (a). The second claim follows from

$$R_\phi(\mathsf{f}_{\lambda(r)}) + r\|\lambda(r)\|_{\ell_1} \le R_\phi(\mathsf{f}_{\lambda(0)}) + r\|\lambda(0)\|_{\ell_1} \le R_\phi(\mathsf{f}_{\lambda(r)}) + r\|\lambda(0)\|_{\ell_1}.$$

For the proof of part (c), we first observe that $\|\lambda(r)\|_{\ell_1} \le \|\lambda(0)\|_{\ell_1}$ is equivalent to

$$\sum_{j \notin I_0} |\lambda_j(r)| \le \sum_{j \in I_0} |\lambda_j(0)| - \sum_{j \in I_0} |\lambda_j(r)|.$$

Next, we note that the term on the left equals $\sum_{j \notin I_0} |\lambda_j(0) - \lambda_j(r)|$ and we bound the term on the right by $\sum_{j \in I_0} |\lambda_j(0) - \lambda_j(r)|$ using the triangle inequality. This proves part (c). □

This result gives a simple condition for $R_\phi(\mathsf{f}_{\lambda(r)}) \to R_\phi(\mathsf{f}_{\lambda(0)})$ and shows that the ℓ_1 norm of the solution $\lambda(r)$ is always smaller than the ℓ_1 norm of $\lambda(0)$. Similar properties are established by [6] for minimizers of twice differentiable loss functions ϕ and ℓ_p norms for $p > 1$. In contrast, we consider here a non-differentiable loss function ϕ and $p = 1$.

Our target is a sparse vector $\theta \in \mathbb{R}^M$ with risk $R_\phi(\mathsf{f}_\theta)$ close to $R_\phi(\mathsf{f}_{\lambda(0)})$. Before we make this precise, we need to introduce a few concepts depending on the behavior of $\eta(X)$ near d and $1 - d$, and the set of functions f_j.

Definition 2.2 (Classification complexity). *The classification complexity is defined as the largest number $\alpha \geq 0$ such that, for some $A \geq 1$ and all $t > 0$,*

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

This notion of complexity is a generalization of Tsybakov's margin condition [9] for $d = 1/2$. The behavior of $\eta(X)$ is obviously not relevant in the interval $(d, 1-d)$, only at the endpoints d and $1-d$. The inequality always holds for $\alpha = 0$ and $A = 1$. In contrast, $\alpha = +\infty$ describes the easiest classification situation where we essentially require that $\eta(X)$ stays away from d and $1-d$ with probability one. If $\eta(X)$ has a density in the neighborhood of d and $1-d$, then we have that $\alpha = 1$.

Definition 2.3 (Restricted eigenvalue condition). *Let $\theta \in \mathbb{R}^M$, $c \geq 1$ and Ψ be the $M \times M$ matrix with entries $\Psi_{i,j} = 4\mathbb{E}[f_i(X)f_j(X)\omega(X)]$ with $\omega(X) = \eta(X)\{1-\eta(X)\}$. For $I = \{i : \theta_i \neq 0\}$, the support of θ, we define*

$$\kappa^2(\theta, c) = \inf_{\lambda \neq \theta \in \mathbb{R}^M \,:\, \|(\theta-\lambda)_{I^C}\|_{\ell_1} \leq c\|(\theta-\lambda)_I\|_{\ell_1}} \frac{(\theta-\lambda)'\Psi(\theta-\lambda)}{4\|(\theta-\lambda)_I\|_{\ell_2}^2}.$$

The condition $\kappa(\theta, c) > 0$ is a restrictive eigenvalue condition on the Gram matrix Ψ of the type introduced in [2] in the context of linear regression. Using similar reasoning as in [2], page 1714, it is implied by the local mutual coherence condition used in [11]. We are now in position to state an oracle inequality for the excess risk,

$$\Delta R_{\phi}(\mathsf{f}_{\lambda(r)}) := R_{\phi}(\mathsf{f}_{\lambda(r)}) - R_{\phi}(f_0), \tag{2.3}$$

of the regularized minimizer $\lambda(r)$ and the ℓ_1-distance between the vectors $\lambda(r)$ and θ.

Theorem 2.4. *Let α be the classification complexity, and θ be such that $R(\mathsf{f}_{\theta}) \leq R(\mathsf{f}_{\lambda(r)})$ and $\kappa = \kappa(\theta, 1) > 0$. Then, for any*

$$r \leq (2C_F)^{-(2+\alpha)/\alpha}\{4A(2d)^{\alpha}\}^{-1/\alpha}(\kappa^{-2}\|\theta\|_{\ell_0})^{-(1+\alpha)/\alpha} \tag{2.4}$$

with $C_F = \max_j \|f_j\|_{\infty} = \max_j \sup_x |f_j(x)|$ and $\|\theta\|_{\ell_0} = \sum_{j=1}^{M} I\{\theta_j \neq 0\}$, we have

$$\begin{aligned} &\Delta R_{\phi}(\mathsf{f}_{\lambda(r)}) + r\|\lambda(r) - \theta\|_{\ell_1} \\ &\quad \leq 3\Delta R_{\phi}(\mathsf{f}_{\theta}) + 6\{4A(2d)^{\alpha}\}^{1/(2+\alpha)}\|\mathsf{f}_{\theta} - f_0\|_{\infty}(\kappa^{-2}r^2\|\theta\|_{\ell_0})^{(1+\alpha)/(2+\alpha)}. \end{aligned} \tag{2.5}$$

Proof. Set $\delta = \lambda(r) - \theta$. Let $I = \{i : \theta_i \neq 0\}$ be the support of θ. It is straightforward to derive from Proposition 2.1 that

$$R_{\phi}(\mathsf{f}_{\lambda(r)}) + r\|\delta\|_{\ell_1} \leq R_{\phi}(\mathsf{f}_{\theta}) + 2r\|\delta_I\|_{\ell_1}$$

and, subsequently, that

$$r\|\delta_{I^C}\|_{\ell_1} \leq R_{\phi}(\mathsf{f}_{\theta}) - R_{\phi}(\mathsf{f}_{\lambda(r)}) + r\|\delta_I\|_{\ell_1} \leq r\|\delta_I\|_{\ell_1}.$$

The first inequality, combined with the assumption $\kappa = \kappa(\theta, 1) > 0$, yields

$$\begin{aligned}\Delta R_\phi(\mathsf{f}_{\lambda(r)}) + r\|\delta\|_{\ell_1} &\le \Delta R_\phi(\mathsf{f}_\theta) + \kappa^{-1}\|\mathsf{f}_\delta\|(r^2|I|)^{1/2} \\ &\le \Delta R_\phi(\mathsf{f}_\theta) + \kappa^{-1}\|\mathsf{f}_\lambda - f_0\|(r^2|I|)^{1/2} + \kappa^{-1}\|\mathsf{f}_\theta - f_0\|(r^2|I|)^{1/2},\end{aligned}$$

using the notation $\|\mathsf{f}\| = \mathbb{E}^{1/2}[\mathsf{f}^2(X)\omega(X)]$ and $\omega(X) = \eta(X)(1-\eta)(X)$. By Lemma A.1 in Appendix A, we find that

$$\|\mathsf{f}_\lambda - f_0\|^{2+2\alpha} \le 4A(2d)^\alpha \|\mathsf{f}_\lambda - f_0\|_\infty^{2+\alpha}\{\Delta R_\phi(\mathsf{f}_\lambda)\}^\alpha$$

for $\lambda = \theta$ and $\lambda = \lambda(r)$. After we plug this bound into the right-hand side of the previous display, we find that

$$\begin{aligned}&\Delta R_\phi(\mathsf{f}_{\lambda(r)}) + r\|\delta\|_{\ell_1} \\ &\quad\le \Delta R_\phi(\mathsf{f}_\theta) + \kappa^{-1}(r^2|I|)^{1/2}\{4A(2d)^\alpha\}^{1/(2+2\alpha)}\|\mathsf{f}_{\lambda(r)} - f_0\|_\infty^{(2+\alpha)/(2+2\alpha)}\{\Delta R_\phi(\mathsf{f}_{\lambda(r)})\}^{\alpha/(2+2\alpha)} \\ &\qquad + \kappa^{-1}(r^2|I|)^{1/2}\{4A(2d)^\alpha\}^{1/(2+2\alpha)}\|\mathsf{f}_\theta - f_0\|_\infty^{(2+\alpha)/(2+2\alpha)}\{\Delta R_\phi(\mathsf{f}_\theta)\}^{\alpha/(2+2\alpha)}.\end{aligned}$$

Next, we apply Young's algebraic inequality,

$$ab \le \frac{a^p}{p} + \frac{b^q}{q} \qquad \text{with } p > 1 \quad \text{and} \quad q = \frac{p}{p-1} \qquad \text{for all } a, b > 0,$$

to the last two terms on the right-hand side, with $p = (2+2\alpha)/\alpha$ and $q = (2+2\alpha)/(2+\alpha)$, to get

$$\begin{aligned}&\Delta R_\phi(\mathsf{f}_{\lambda(r)}) + r\|\delta\|_{\ell_1} \\ &\quad\le \Delta R_\phi(\mathsf{f}_\theta) + \frac{\alpha}{2+2\alpha}\{\Delta R_\phi(\mathsf{f}_{\lambda(r)}) + \Delta R_\phi(\mathsf{f}_\theta)\} \\ &\qquad + \frac{2+\alpha}{2+2\alpha}\{4A(2d)^\alpha\}^{1/(2+\alpha)}(\kappa^{-2}r^2|I|)^{(1+\alpha)/(2+\alpha)}(\|\mathsf{f}_{\lambda(r)} - f_0\|_\infty + \|\mathsf{f}_\theta - f_0\|_\infty).\end{aligned}$$

Since $\|\mathsf{f}_{\lambda(r)} - f_0\|_\infty \le \|\mathsf{f}_\theta - f_0\|_\infty + C_F\|\delta\|_{\ell_1}$, we deduce, after invoking (2.4), that

$$\begin{aligned}&(2+\alpha)\Delta R_\phi(\mathsf{f}_{\lambda(r)}) + (1+3\alpha/2)r\|\delta\|_{\ell_1} \\ &\quad\le (2+3\alpha)\Delta R_\phi(\mathsf{f}_\theta) + 2(2+\alpha)\{4A(2d)^\alpha\}^{1/(2+\alpha)}(\kappa^{-2}r^2|I|)^{(1+\alpha)/(2+\alpha)}\|\mathsf{f}_\theta - f_0\|_\infty,\end{aligned}$$

and the conclusion follows. □

It is interesting to see that the bound (2.5) crucially depends on the classification complexity parameter α and $\|\mathsf{f}_\theta - f_0\|_\infty$. In particular, if f_0 can itself be represented as a linear combination of the basis functions, then $f_0 = \mathsf{f}_{\lambda(0)}$. In this case, provided that $\kappa(\lambda(0), 1) > 0$, Theorem 2.4 implies that $\Delta R_\phi(\mathsf{f}_{\lambda(r)}) + r\|\lambda(r) - \lambda(0)\|_{\ell_1} \le 0$. In other words, we have the following corollary.

Corollary 2.5. *If* $f_0 = \mathsf{f}_{\lambda(0)}$ *and* $\kappa(\lambda(0), 1) > 0$, *then* $\lambda(r) = \lambda(0)$ *for any*

$$r \le (2C_F)^{-(2+\alpha)/\alpha}\{4A(2d)^\alpha\}^{-1/\alpha}(\kappa^{-2}\|\lambda(0)\|_{\ell_0})^{-(1+\alpha)/\alpha}.$$

3. ℓ_1-regularized empirical generalized hinge risk minimizers

In this section we study the estimate $\hat{\lambda}(2r)$. In what follows, we will simplify notation so as not to show dependence of $\hat{\lambda}$ on r whenever no confusion occurs. Again, we emphasize that our results hold, in particular, for $d = 1/2$, the case of a support vector machine without a reject option.

Note that the inequality

$$\hat{R}_\phi(\hat{\lambda}) + 2r\|\hat{\lambda}\|_{\ell_1} \le \hat{R}_\phi(\lambda) + 2r\|\lambda\|_{\ell_1} \tag{3.1}$$

applied to the vector of zeros $\lambda = (0, \ldots, 0)'$ implies that $\|\hat{\lambda}\|_{\ell_1} \le \phi(0)/(2r) = 1/(2r)$. This means that we can restrict our analysis to the set

$$\Lambda = \{\lambda \in \mathbb{R}^M : \|\lambda\|_{\ell_1} \le 1/(2r)\}.$$

The aim of this section is to show that $\hat{\lambda}$ is close to $\lambda(r)$ for a judiciously chosen tuning parameter r.

Theorem 3.1. *If, for some* $p \ge 1$,

$$r \ge \frac{1-d}{d} C_F \left\{ 9\sqrt{\frac{2\log 2(M \vee n)}{n}} + 2\frac{p\log_2 n}{\sqrt{2M \vee 2n}} + \sqrt{\frac{2\log 1/\delta}{n}} \right\}, \tag{3.2}$$

then for all $\theta \in \Lambda$, *with probability larger than* $1 - \delta$,

$$\Delta R_\phi(\mathsf{f}_{\hat{\lambda}}) + r\|\hat{\lambda}\|_{\ell_1} \le \Delta R_\phi(\mathsf{f}_\theta) + 3r\|\theta\|_{\ell_1} + n^{-p}$$

and, moreover,

$$\Delta R_\phi(\mathsf{f}_{\hat{\lambda}}) + r\|\hat{\lambda} - \theta\|_{\ell_1} \le \Delta R_\phi(\mathsf{f}_\theta) + 4r\|\theta\|_{\ell_1} + n^{-p}.$$

Proof. Write $\hat{\delta} = \hat{\lambda} - \theta$. Let $\varepsilon = r^{-1}n^{-p}$ and define

$$\hat{r} = \sup_{\lambda \in \Lambda} \frac{\{\hat{R}_\phi(\mathsf{f}_\lambda) - R_\phi(\mathsf{f}_\lambda)\} - \{\hat{R}_\phi(\mathsf{f}_\theta) - R_\phi(\mathsf{f}_\theta)\}}{\|\lambda - \theta\|_{\ell_1} + \varepsilon}. \tag{3.3}$$

By Propositions B.1 and B.2 in Appendix B,

$$\mathbb{P}\{\hat{r} \le r\} \ge 1 - \delta$$

for the choice r given in (3.2). Rewriting the inequality (3.1), we find that

$$\begin{aligned} R_\phi(\mathsf{f}_{\hat\lambda}) &\le R_\phi(\mathsf{f}_\theta) + \{\hat R_\phi(\mathsf{f}_\theta) - R_\phi(\mathsf{f}_\theta)\} - \{\hat R_\phi(\mathsf{f}_{\hat\lambda}) - R_\phi(\mathsf{f}_{\hat\lambda})\} \\ &\quad + 2r\|\theta\|_{\ell_1} - 2r\|\hat\lambda\|_{\ell_1} \\ &\le R_\phi(\mathsf{f}_\theta) + \hat r(\|\hat\delta\|_{\ell_1} + \varepsilon) + 2r\|\theta\|_{\ell_1} - 2r\|\hat\lambda\|_{\ell_1}. \end{aligned} \tag{3.4}$$

Thus, on the event $\hat r \le r$, after adding $r\|\hat\lambda\|_{\ell_1}$ to both sides, we obtain

$$R_\phi(\mathsf{f}_{\hat\lambda}) + r\|\hat\lambda\|_{\ell_1} \le R_\phi(\mathsf{f}_\theta) + 3r\|\theta\|_{\ell_1} + r\varepsilon,$$

which proves the first claim. Adding $r\|\hat\delta\|_{\ell_1}$ to both sides easily yields the second claim. □

A direct consequence of Theorem 3.1 is the following corollary which states that in the sparse setting where $r\|\lambda(r)\|_{\ell_1} \to 0$, the estimator $\hat\lambda(2r)$ behaves like the penalized minimizer $\lambda(r)$ in terms of their risk.

Corollary 3.2. *Suppose that $r\|\lambda(r)\|_{\ell_1} \to 0$ as $n \to \infty$ for r satisfying (3.2). Then, with probability at least $1-\delta$,*

$$|\{R_\phi(\hat\lambda) + r\|\hat\lambda\|_{\ell_1}\} - \{R_\phi(\lambda(r)) + r\|\lambda(r)\|_{\ell_1}\}| \to 0$$

as $n \to \infty$. In particular, when taking $\theta = \lambda(0)$, we have $|R_\phi(\hat\lambda) - R_\phi(\lambda(0))| \to 0$ and $\|\hat\lambda(2r) - \lambda(0)\|_{\ell_1} = o(1/r)$.

Proof. We combine the basic property (3.1) applied to $\theta = \lambda(r)$ and Theorem 3.1, and we find that on the event $\hat r \le r$,

$$R_\phi(\lambda(r)) + r\|\lambda(r)\|_{\ell_1} \le R_\phi(\hat\lambda) + r\|\hat\lambda\|_{\ell_1} \le R_\phi(\lambda(r)) + r\|\lambda(r)\|_{\ell_1} + \{2r\|\lambda(r)\|_{\ell_1} + r\varepsilon\}.$$

The result then follows from $\{2r\|\lambda(r)\|_{\ell_1} + r\varepsilon\} \to 0$. □

We emphasize that the above results do not impose any restrictions on the dictionary $\{f_j\}$. If we are willing to make assumptions on the Gram matrix Ψ, then we obtain a more refined result.

Theorem 3.3. *For all r satisfying (3.2) and $\theta \in \Lambda$ such that $\kappa = \kappa(\theta, 7) > 0$ and*

$$(\kappa^2 r^{\alpha/(1+\alpha)}\|\theta\|_{\ell_0})^{(1+\alpha)/(2+\alpha)} < c \tag{3.5}$$

for some (small) c depending on C_F, α, A and d, we have, for some C depending on c, that

$$\Delta R_\phi(\mathsf{f}_{\hat\lambda}) + \tfrac{1}{2}r\|\hat\lambda - \theta\|_{\ell_1} \le 3\Delta R_\phi(\theta) + C\|\mathsf{f}_\theta - f_0\|_\infty(\kappa^{-2}r^2\|\theta\|_{\ell_0})^{(1+\alpha)/(2+\alpha)} + n^{-p}$$

holds with probability at least $1-\delta$.

Proof. Recall that $\varepsilon = r^{-1}n^{-p}$. We may assume without loss of generality that

$$R_\phi(\mathsf{f}_\theta) + \varepsilon r \le R_\phi(\mathsf{f}_{\hat\lambda}) + \tfrac{1}{2}r\|\hat\delta\|_{\ell_1} \tag{3.6}$$

holds, since otherwise the statement holds trivially. Consequently, on the event $\hat r \le r$, using (3.4) and (3.6), we get

$$\begin{aligned}
R_\phi(\mathsf{f}_{\hat\lambda}) &\le R_\phi(\mathsf{f}_\theta) + \varepsilon r + r\|\hat\delta\|_{\ell_1} + 2r\|\theta\|_{\ell_1} - 2r\|\hat\lambda\|_{\ell_1} \\
&\le R_\phi(\mathsf{f}_{\hat\lambda}) + \tfrac{3}{2}r\|\hat\delta\|_{\ell_1} + 2r\|\theta\|_{\ell_1} - 2r\|\hat\lambda\|_{\ell_1} \\
&= R_\phi(\mathsf{f}_{\hat\lambda}) + \tfrac{3}{2}r\|\hat\delta_I\|_{\ell_1} + \tfrac{3}{2}r\|\hat\lambda_{I^c}\|_{\ell_1} + 2r\|\theta\|_{\ell_1} - 2r\|\hat\lambda\|_{\ell_1} \\
&= R_\phi(\mathsf{f}_{\hat\lambda}) + \tfrac{3}{2}r\|\hat\delta_I\|_{\ell_1} + 2r\|\theta\|_{\ell_1} - 2r\|\hat\lambda_I\|_{\ell_1} - \tfrac{1}{2}r\|\hat\delta_{I^c}\|_{\ell_1} \\
&\le R_\phi(\mathsf{f}_{\hat\lambda}) + \tfrac{7}{2}r\|\hat\delta_I\|_{\ell_1} - \tfrac{1}{2}r\|\hat\lambda_{I^c}\|_{\ell_1}
\end{aligned}$$

so that $\|\hat\delta_{I^c}\|_{\ell_1} \le 7\|\hat\delta_I\|_{\ell_1}$, where I is the support of θ. On the other hand,

$$\begin{aligned}
R_\phi(\mathsf{f}_{\hat\lambda}) + \tfrac{1}{2}r\|\hat\delta\|_{\ell_1} &\le R_\phi(\mathsf{f}_\theta) + \varepsilon r + \tfrac{3}{2}r\|\hat\delta\|_{\ell_1} + 2r\|\theta\|_{\ell_1} - 2r\|\hat\lambda\|_{\ell_1} \\
&\le R_\phi(\mathsf{f}_\theta) + \varepsilon r + \tfrac{3}{2}r\|\hat\delta_I\|_{\ell_1} + 2r\|\hat\delta_I\|_{\ell_1} - \tfrac{1}{2}r\|\hat\lambda_{I^c}\|_{\ell_1} \\
&\le R_\phi(\mathsf{f}_\theta) + \tfrac{7}{2}r\|\hat\delta_I\|_{\ell_1} + r\varepsilon.
\end{aligned}$$

The remainder of the proof follows that of Theorem 2.4, with $\kappa = \kappa(\theta, 7)$. □

This result differs from [11] (and [8] for the case $d = 1/2$) in the appearance of the norm $\|f_0 - \mathsf{f}_\theta\|_\infty$ on the right-hand side of the (oracle) inequality. This implies that for $f_0 = \mathsf{f}_\theta$ and for some sparse $\theta = \lambda(0)$ satisfying the conditions of Theorem 3.3, we can expect fast rates, regardless of the classification complexity! Another important difference with both papers is that no restriction is imposed on the sup-norm of f_λ. Such a condition is unnatural as $|\mathsf{f}_\lambda| \le C$ may overrule the restriction that the penalty term $r\|\lambda\|_{\ell_1}$ imposes.

We now consider bounds on the error and reject rates without an additional test sample. We write

$$\mathbb{P}_n\{Y\mathsf{f}_{\hat\lambda}(X) \le \beta\} = \frac{1}{n}\sum_{i=1}^n I\{Y_i\mathsf{f}_{\hat\lambda}(X_i) \le \beta\}$$

for any $\beta > 0$. The misclassification and rejection rate can be bounded above as follows.

Theorem 3.4. *If*

$$r(\gamma) \ge \frac{9C_F}{\gamma}\sqrt{\frac{2\log 2(M \vee n)}{n}} + \frac{2p\log_2(n)C_F}{\gamma\sqrt{2(M \vee n)}} + \frac{C_F}{\gamma}\sqrt{\frac{2\log(1/\delta)}{n}},$$

then, with probability at least $1-\delta$*, we have*

$$\mathbb{P}\{Y\mathsf{f}_{\hat\lambda}(X)\le-\tau\}\le\min_{\gamma>0}[\mathbb{P}_n\{Y\mathsf{f}_{\hat\lambda}(X)\le-\tau+\gamma\}+r(\gamma)\|\hat\lambda\|_{\ell_1}]+n^{-p},$$

$$\mathbb{P}\{|\mathsf{f}_{\hat\lambda}(X)|\le\tau\}\le\min_{\gamma>0}[\mathbb{P}_n\{|\mathsf{f}_{\hat\lambda}(X)|\le\tau+\gamma\}+r(\gamma)\|\hat\lambda\|_{\ell_1}]+n^{-p}.$$

Proof. Set

$$\varphi_\gamma(z)=\begin{cases}1, & \text{if } z<-\tau,\\ \dfrac{1}{\gamma}(\gamma-\tau-z), & \text{if } -\tau\le z\le-\tau+\gamma,\\ 0, & \text{if } z\ge-\tau+\gamma.\end{cases}$$

The following inequalities then hold uniformly in λ:

$$\begin{aligned}\mathbb{P}\{Y\mathsf{f}_\lambda(X)\le-\tau\}&\le\mathbb{P}_n\{Y\mathsf{f}_\lambda(X)\le-\tau+\gamma\}+R_{\varphi_\gamma}(\mathsf{f}_\lambda)-\hat R_{\varphi_\gamma}(\mathsf{f}_\lambda)\\&\le\mathbb{P}_n\{Y\mathsf{f}_\lambda(X)\le-\tau+\gamma\}+\hat r_0\{\|\lambda\|_{\ell_1}+\varepsilon\},\end{aligned}$$

where

$$\hat r_0=\sup_{\lambda\in\Lambda}\frac{|\hat R_{\varphi_\gamma}(\mathsf{f}_\lambda)-R_{\varphi_\gamma}(\mathsf{f}_\lambda)|}{\|\lambda\|_{\ell_1}+\varepsilon},$$

with ε given by $\varepsilon r(\gamma)=n^{-p}$. We can invoke Propositions B.1 and B.2 to complete the proof of the first claim. The proof of the second claim uses the reasoning above, with the only modification being that $\varphi_\gamma(z)$ is now given by

$$\varphi_\gamma(z)=\begin{cases}1, & \text{if } |z|<\tau,\\ \dfrac{1}{\gamma}(z+\gamma+\tau), & \text{if } -\tau-\gamma\le z\le-\tau,\\ -\dfrac{1}{\gamma}(z-\gamma-\tau), & \text{if } \tau\le z\le\tau+\gamma,\\ 0, & \text{if } |z|\ge\tau+\gamma;\end{cases}$$

the rest of the reasoning is unchanged. □

4. Numerical experiments

We now demonstrate the practical merits of $\hat\lambda(r)$ via a couple of numerical experiments. We begin by noting that the computation of $\hat\lambda(r)$ can be conveniently formulated as a linear program. Let $\xi_1,\dots,\xi_n$ be the slack variables such that

$$\xi_i\ge0,\qquad \xi_i\ge1-Y_if_\lambda(X_i),\qquad \xi_i\ge1-aY_if_\lambda(X_i).\tag{4.1}$$

Clearly the minimum ξ_i that satisfies these constraints is $\phi(Y_i f(X_i))$. We also introduce slack variables ξ_{n+i}, $i = 1, \ldots, M$, to represent $|\lambda_i|$, that is,

$$\xi_{n+i} \geq \lambda_i, \qquad \xi_{n+i} \geq -\lambda_i. \tag{4.2}$$

Using the slack variables, $\hat{\lambda}(r)$ can be given as the solution of the linear program

$$\min_{\lambda,\xi}[\xi_1 + \cdots + \xi_n + r(\xi_{n+1} + \cdots + \xi_{n+M})]$$

subject to

$$\begin{aligned} &\xi_i \geq 0, \qquad \xi_i \geq 1 - y_i h_i, \qquad \xi_i \geq 1 - aY_i h_i, \qquad i = 1, \ldots, n, \\ &\xi_{n+i} \geq \lambda_i, \qquad \xi_{n+i} \geq -\lambda_i, \qquad i = 1, \ldots, M, \\ &h_i = \sum_j \lambda_j f_j(X_i), \qquad i = 1, \ldots, n. \end{aligned}$$

To illustrate the merits of $\hat{\lambda}$, we implement the method described above and first apply it to a set of simulated examples. To fix ideas, we set $d = 0.25$ or, equivalently, $a = 3$. For each run, 50 positive instances ($Y = +1$) and 50 negative instances ($Y = -1$) were generated. Two hundred ($M = 200$) features (f_j's) were simulated from a multivariate normal distribution. For positive instances, the mean was set to $(1/\sqrt{2}, 1/\sqrt{2}, 0, \ldots, 0)'$, whereas for the negative instances, the mean was set to $(-1/\sqrt{2}, -1/\sqrt{2}, 0, \ldots, 0)'$. In both cases, the covariance matrix was the identity matrix. The operating characteristics of the method are demonstrated in Figure 1. On the left-hand side, the misclassification rate ($\mathbb{P}(Y\mathsf{f}_{\hat{\lambda}}(X) < -0.5)$), rejection rate ($\mathbb{P}(|Y\mathsf{f}_{\hat{\lambda}}(X)| < 0.5)$) and associated ℓ-risk of the ℓ_1-regularized generalized hinge loss ($R_\ell(\mathsf{f}_{\hat{\lambda}})$) are plotted as functions of the tuning parameter r for a typical simulation. The results are to be compared with the usual ℓ_1-regularized support vector machines where no rejection option is allowed. Since there is no rejection, the misclassification rate for the usual support vector machines coincides with its ℓ-risk. It is evident that by incorporating the rejection option, $\hat{\lambda}$ yields a smaller ℓ-risk, provided that both methods are optimally tuned. To further investigate the merits of allowing the rejection option, we repeated the experiment 200 times. The excess risk ΔR_ℓ of both the usual support vector machine and the proposed method are summarized in the plot on the right-hand side. It further confirms the advantage of $\hat{\lambda}$.

To further demonstrate the merits of the method, we apply it to the mixture data example considered in [4]. The training data consist of 200 data points generated from a pair of two-dimensional mixture densities. Similarly to [4], we consider a dictionary of Gaussian radial basis functions $f_j(\cdot) = \exp(-2\|\cdot - b_j\|^2), j = 1, \ldots, 100$, where the locations b_j are placed on a 10×10 equally spaced lattice. To fix ideas, we consider the case where $d = 0.25$. The optimal classification rule will classify an observation as $+1$ if the corresponding conditional probability $\mathbb{P}(Y = +1|X)$ is greater than 0.75 and as -1 if the conditional probability is less than 0.25. When the conditional probability is between 0.25 and 0.75, we withhold the decision. The corresponding decision boundaries

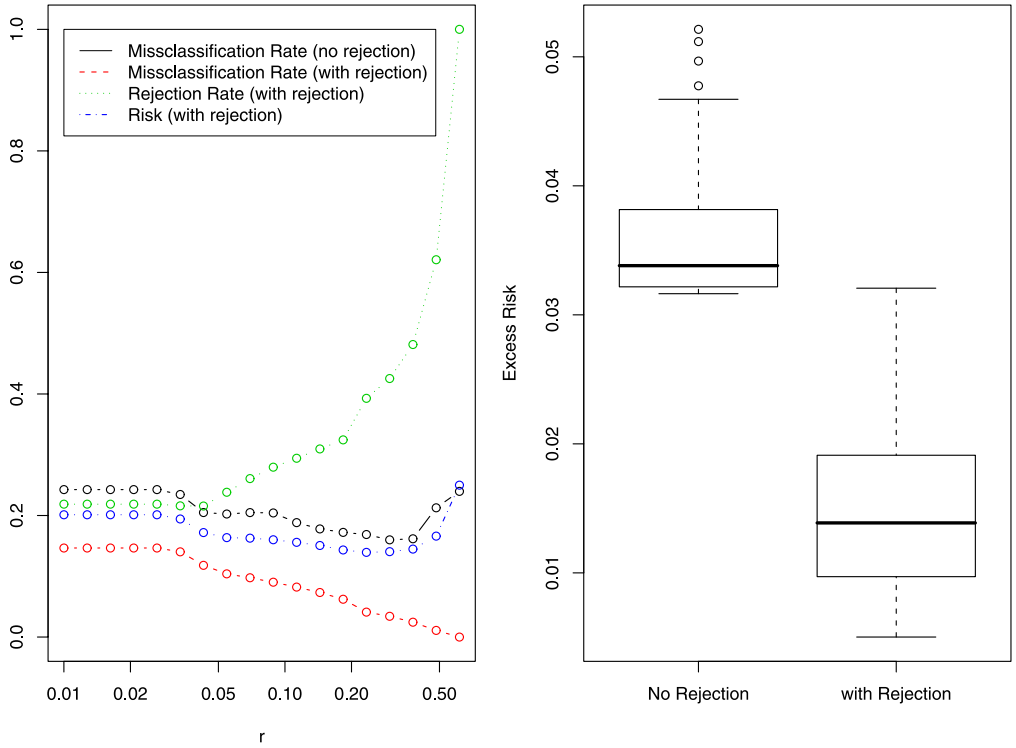

Figure 1. Simulation – the effect of rejection, misclassification rate and excess risk R_ℓ. The left-hand panel shows the three criteria as functions of the tuning parameter r for the support vector machine (SVM) with rejection option for a typical run. Also included is the misclassification rate for the usual SVM. It is evident that SVM with rejection option enjoys lower misclassification rate by withholding decision for "hard-to-classify" cases. The right-hand panel compares the excess ℓ-risk for SVM with or without rejection option. The box plots of the excess risk are produced based on 200 runs. This again confirms that SVM with rejection option leads to improved performance in terms of the ℓ loss.

are given in the right-hand panel of Figure 2. It is known that the usual SVM only targets the decision boundary identified with $\mathbb{P}(Y=+1|X)$ and cannot be used to recover the optimal decision boundaries given here; see, for instance, [12] for further discussion of this issue. In contrast, the SVM with rejection option is devised specifically for this purpose. To this end, we ran the SVM with rejection option with $a=3$ and $\tau=0.5$, as discussed earlier. The tuning parameter r was selected by tenfold cross-validation. The left-hand panel of Figure 2 gives the estimated decision boundaries. It is clear from the plot that SVM with rejection option successfully captured the main characteristics of the underlying probabilities. The main difference between the two sets of decision boundaries occurs in regions where no observations are available. As a result, the SVM with rejection option opted for withholding a decision.

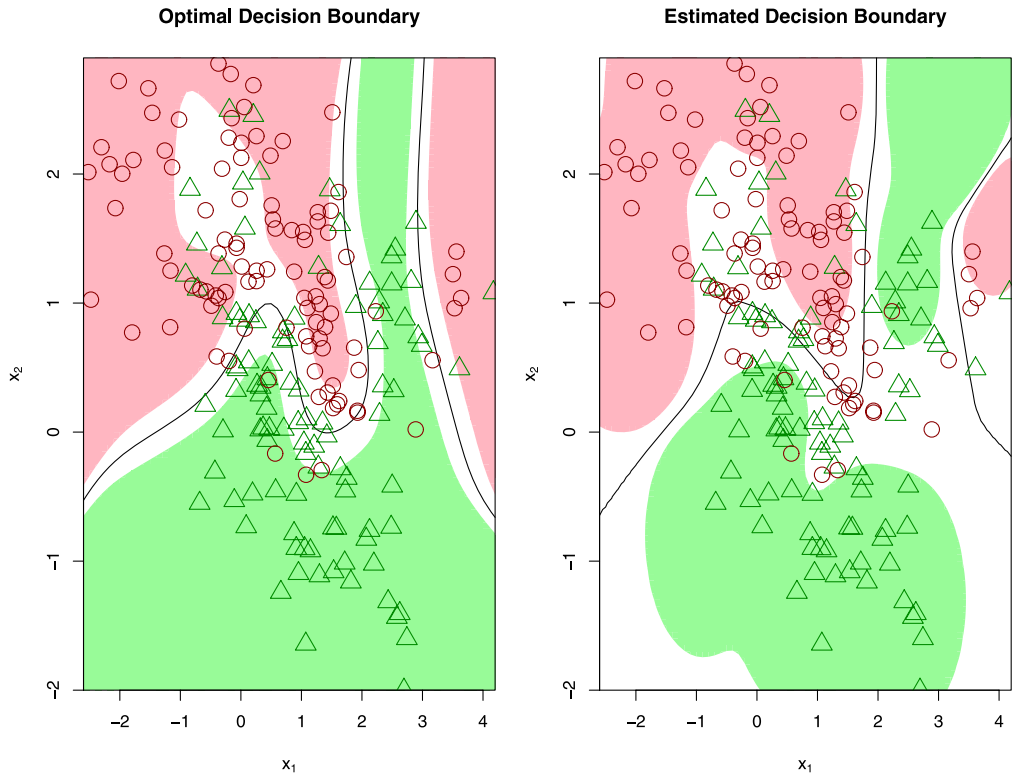

Figure 2. Mixture data – optimal and estimated decision boundaries. The left-hand panel gives the optimal decision boundary, whereas the right-hand panel corresponds to the SVM with rejection option. In both plots, positive cases are represented by red circles and negative cases by green triangles. The light red regions correspond to classification $Y = +1$ and light green regions to classification $Y = -1$. Areas where a decision is withheld are not shaded. The solid black line in the left-hand panel is the level set for $\mathbb{P}(Y = +1|X) = 0.5$. The solid black line in the right-hand panel is the level set for $\mathsf{f}_{\hat{\lambda}} = 0$.

Appendix A: Connection between excess risk and weighted L_2 norm

The next lemma is a technical result that links the excess risk $\Delta R_\phi(\lambda)$ to the L_2 norm:

$$\|\mathsf{f}_\lambda - f_0\| = \sqrt{\mathbb{E}[|\mathsf{f}_\lambda(X) - f_0(X)|^2 \omega(X)]}$$

with $\omega(X) = \eta(X)(1-\eta)(X)$. Its proof is rather technical and relies on results obtained in [1]. Essentially, $\|\mathsf{f}_\lambda - f_0\|_\infty$ replaces the suboptimal bound $1 + C_\Lambda C_F$ in [11].

Lemma A.1. *Let $\alpha > 0$ be as in Definition 2.2. Then, for all $\lambda \in \mathbb{R}^M$,*

$$\|\mathsf{f}_\lambda - f_0\|^{2+2\alpha} \le 4A(2d)^\alpha \|\mathsf{f}_\lambda - f_0\|_\infty^{2+\alpha} \{\Delta R_\phi(\lambda)\}^\alpha. \tag{A.1}$$

Proof. Let $f : \mathcal{X} \to \mathbb{R}$ be arbitrary and set

$$\rho_\eta(f, f_0) = \begin{cases} \eta|f - f_0|, & \text{if } \eta < d \text{ and } f < -1, \\ (1-\eta)|f - f_0|, & \text{if } \eta > 1 - d \text{ and } f > 1, \\ |f - f_0|, & \text{otherwise,} \end{cases}$$

then [1], Lemma 9, states that

$$\Delta R_\phi(\lambda) \ge d^{-1}\mathbb{E}[\rho_\eta(f, f_0)(X)(|\eta(X) - (1-d)|I_{\{X \in E_-\}} + |\eta(X) - d|I_{\{X \in E_+\}})]$$

with

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

Using (A.1), for any set E,

$$\begin{aligned} &\mathbb{E}[\rho_\eta(f, f_0)(X)|\eta(X) - (1-d)|I_{\{X \in E\}}] \\ &\quad \ge t\mathbb{E}[\rho_\eta(f, f_0)(X)I_{\{|\eta(X)-(1-d)| \ge t\}} I_{\{X \in E\}}] \\ &\quad = t\mathbb{E}[\rho_\eta(f, f_0)(X)I_{\{X \in E\}}] - t\mathbb{E}[\rho_\eta(f, f_0)(X)I_{\{|\eta(X)-(1-d)| < t, X \in E\}}] \\ &\quad \ge t\mathbb{E}[\rho_\eta(f, f_0)(X)I_{\{X \in E\}} - \|f - f_0\|_\infty A t^\alpha]. \end{aligned}$$

Similarly,

$$\mathbb{E}[\rho_\eta(f, f_0)(X)|\eta(X) - d|I_{\{X \in E\}}] \ge t\mathbb{E}[\rho_\eta(f, f_0)(X)I_{\{X \in E\}} - \|f - f_0\|_\infty A t^\alpha],$$

and we obtain

$$\begin{aligned} \Delta R_\phi(\lambda) &\ge d^{-1}t\mathbb{E}[\rho_\eta(\mathsf{f}_\lambda, f_0)(X)I_{\{X \in E_+ \cup E_-\}} - 2\|\mathsf{f}_\lambda - f_0\|_\infty A t^\alpha] \\ &= d^{-1}t\mathbb{E}[\rho_\eta(\mathsf{f}_\lambda, f_0)(X) - 2\|\mathsf{f}_\lambda - f_0\|_\infty A t^\alpha]. \end{aligned}$$

Plugging

$$t = \left(\frac{\mathbb{E}[\rho_\eta(\mathsf{f}_\lambda, f_0)(X)]}{4A\|\mathsf{f}_\lambda - f_0\|_\infty}\right)^{1/\alpha}$$

into the preceding expression, we obtain

$$\Delta R_\phi(\lambda) \ge \frac{(\mathbb{E}[\rho_\eta(\mathsf{f}_\lambda, f_0)(X)])^{(1+\alpha)/\alpha}}{2d(4A\|\mathsf{f}_\lambda - f_0\|_\infty)^{1/\alpha}}.$$

Since

$$\|\mathsf{f}_\lambda - f_0\|^2 = \mathbb{E}[\omega(X)(\mathsf{f}_\lambda - f_0)^2(X)] \le \|\mathsf{f}_\lambda - f_0\|_\infty \mathbb{E}[\omega(X)|\mathsf{f}_\lambda(X) - f_0(X)|],$$

we get, for all λ,

$$\Delta R_\phi(\lambda) \geq \frac{(\mathbb{E}[\omega(X)|\mathsf{f}_\lambda(X) - f_0(X)|])^{(1+\alpha)/\alpha}}{2d(4A\|\mathsf{f}_\lambda - f_0\|_\infty)^{1/\alpha}}$$
$$\geq \frac{(\|\mathsf{f}_\lambda - f_0\|^2)^{(1+\alpha)/\alpha}}{2d(4A)^{1/\alpha}\|\mathsf{f}_\lambda - f_0\|_\infty^{(2+\alpha)/\alpha}}.$$

The claim follows. □

Remark A.2. If $|\mathsf{f}_\lambda| \leq 1$, then $\rho_\eta(\mathsf{f}_\lambda, f_0) = |\mathsf{f}_\lambda - f_0|$. Hence, if we restrict the parameters λ such that f_λ are bounded by 1, then we can impose the restricted eigenvalue condition on the matrix with entries $\mathbb{E}[f_i(X)f_j(X)]$ instead of $\mathbb{E}[f_i(X)f_j(X)\omega(X)]$.

Appendix B: A maximal inequality for a weighted empirical process

Recall that $\Lambda = \{\lambda \in \mathbb{R}^M : \|\lambda\|_{\ell_1} \leq 1/(2r)\}$ and let $\theta \in \Lambda$ and $\varepsilon > 0$. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function with Lipschitz constant C_φ and define the risks

$$R_\varphi(\mathsf{f}_\lambda) = \mathbb{E}[\varphi(Y\mathsf{f}_\lambda(X))],$$
$$\hat{R}_\varphi(\mathsf{f}_\lambda) = \frac{1}{n}\sum_{i=1}^n \varphi(Y_i\mathsf{f}_\lambda(X_i)).$$

Finally, let $\varepsilon > 0$ and set

$$\hat{r}(\varphi, \theta, \varepsilon) = \sup_{\lambda \in \Lambda} \frac{|\{\hat{R}_\varphi(\mathsf{f}_\lambda) - R_\varphi(\mathsf{f}_\lambda)\} - \{\hat{R}_\varphi(\mathsf{f}_\theta) - R_\varphi(\mathsf{f}_\theta)\}|}{\|\theta - \lambda\|_{\ell_1} + \varepsilon}.$$

We prove a maximal inequality for $\hat{r}(\varphi, \theta, \varepsilon)$ which slightly generalizes the result obtained in [11].

Proposition B.1. *Let $0 < \delta < 1$ and set*

$$r(\varphi, \theta, \varepsilon) = \mathbb{E}[\hat{r}(\varphi, \theta, \varepsilon)] + C_\varphi C_F \sqrt{\frac{2\log(1/\delta)}{n}}.$$

Then,

$$\mathbb{P}\{r(\varphi, \theta, \varepsilon) \geq \hat{r}(\varphi, \theta, \varepsilon)\} \geq 1 - \delta.$$

Proof. First, observe that changing a pair (X_i, Y_i) in $\hat{r}$ changes it by at most $2C_\varphi C_F/n$. The result follows immediately after applying McDiarmid's exponential inequality [3], Theorem 2.2, page 8. □

We now control the expectation of $\hat{r}(\varphi, \theta, \varepsilon)$.

Proposition B.2. *Set* $J = \lceil \log_2(1/\{\varepsilon r\}) \rceil$. *Then,*

$$\mathbb{E}[\hat{r}(\varphi, \theta, \varepsilon)] \leq 9 C_\varphi C_F \sqrt{\frac{2 \log 2(M \vee n)}{n}} + \frac{2 J C_\varphi C_F}{\sqrt{2(M \vee n)}}.$$

Proof. Let $\sigma_1, \ldots, \sigma_n$ be independent Rademacher variables, taking the values ± 1, each with probability $1/2$, independent of the data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Set

$$\hat{R}^0_\varphi(\mathsf{f}_\lambda) = \frac{1}{n} \sum_{i=1}^n \sigma_i \varphi(Y_i \mathsf{f}_\lambda(X_i)).$$

A standard symmetrization trick [3], page 18, shows that

$$\begin{aligned}
\mathbb{E}[\hat{r}(\varphi, \theta, \varepsilon)] &\leq 2\mathbb{E}\left[\sup_{\lambda \in \Lambda} \frac{|\hat{R}^0_\varphi(\mathsf{f}_\lambda) - \hat{R}^0_\varphi(\mathsf{f}_\theta)|}{\|\lambda - \theta\|_{\ell_1} + \varepsilon} \right] \\
&\leq 2\mathbb{E}\left[\sup_{\|\lambda - \theta\|_{\ell_1} \leq \varepsilon} \frac{|\hat{R}^0_\varphi(\mathsf{f}_\lambda) - \hat{R}^0_\varphi(\mathsf{f}_\theta)|}{\|\lambda - \theta\|_{\ell_1} + \varepsilon} \right] + 2\mathbb{E}\left[\sup_{\varepsilon \leq \|\lambda - \theta\|_{\ell_1} \leq 1/r} \frac{|\hat{R}^0_\varphi(\mathsf{f}_\lambda) - \hat{R}^0_\varphi(\theta)|}{\|\lambda - \theta\|_{\ell_1} + \varepsilon} \right] \\
&= (I) + (II).
\end{aligned}$$

The first term

$$\begin{aligned}
I &= 2\mathbb{E}\left[\sup_{\|\lambda - \theta\|_{\ell_1} \leq \varepsilon} \frac{|\hat{R}^0_\varphi(\mathsf{f}_\lambda) - \hat{R}^0_\varphi(\mathsf{f}_\theta)|}{\|\lambda - \theta\|_{\ell_1} + \varepsilon} \right] \\
&= \mathbb{E}\left[\sup_{\|\lambda - \theta\|_{\ell_1} \leq \varepsilon} \frac{1}{\|\lambda - \theta\|_{\ell_1} + \varepsilon} \left| \frac{1}{n} \sum_{i=1}^n \{\varphi(Y_i \mathsf{f}_\lambda(X_i)) - \varphi(Y_i \mathsf{f}_\theta(X_i))\} \right| \right]
\end{aligned}$$

can be bounded using the contraction principle for Rademacher processes; see [7], pages 112–113. For this, we observe that the function $g(z) = \varphi(z_0 + z) - \varphi(z_0)$ is Lipschitz with Lipschitz constant C_φ and $g(0) = 0$. Consequently,

$$\begin{aligned}
(I) &\leq 2\frac{C_\varphi}{\varepsilon} \mathbb{E}\left[\sup_{\|\lambda - \theta\|_{\ell_1} \leq \varepsilon} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i Y_i \mathsf{f}_{\lambda - \theta}(X_i) \right| \right] \\
&\leq 2\frac{C_\varphi}{\varepsilon} \mathbb{E}\left[\sup_{\|\lambda - \theta\|_{\ell_1} \leq \varepsilon} \|\lambda - \theta\|_{\ell_1} \max_{1 \leq j \leq M} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i Y_i f_j(X_i) \right| \right] \\
&\leq 2C_\varphi \mathbb{E}\left[\max_{1 \leq j \leq M} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i Y_i f_j(X_i) \right| \right]
\end{aligned}$$

$$\leq 2C_\varphi C_F \frac{\sqrt{2\log(2M)}}{\sqrt{n}}.$$

The last maximal inequality can be found in [3], Lemma 2.2, page 7, which uses the fact that the variables $\sigma_i Y_i f_j(X_i)$ are sub-Gaussian,

$$\mathbb{E}\left[\exp\left\{s\sum_{i=1}^n \sigma_i Y_i f_j(X_i)\right\}\right] \leq \exp(ns^2 C_F^2/2)$$

for all s, which follows, in turn, from [3], Lemma 2.1, page 5.

The second term (II) requires a peeling argument [10], page 70. Since $0 \leq \hat{r} \leq 2C_\varphi C_F$ almost surely, we can use the bound

$$(II) \leq \zeta + 2C_\varphi C_F \mathbb{P}\left\{\sup_{\varepsilon \leq \|\lambda-\theta\|_{\ell_1} \leq 1/r} 2\frac{|\hat{R}^0_\varphi(\mathsf{f}_\lambda) - \hat{R}^0_\varphi(\mathsf{f}_\theta)|}{\|\lambda-\theta\|_{\ell_1}+\varepsilon} \geq \zeta\right\}. \tag{B.1}$$

Observe that for any $\zeta > 0$,

$$\begin{aligned}&\mathbb{P}\left\{\sup_{\varepsilon \leq \|\lambda-\theta\|_{\ell_1} \leq 1/r} 2\frac{|\hat{R}^0_\varphi(\mathsf{f}_\lambda) - \hat{R}^0_\varphi(\mathsf{f}_\theta)|}{\|\lambda-\theta\|_{\ell_1}+\varepsilon} \geq \zeta\right\}\\ &\quad\leq \sum_{j=1}^J \mathbb{P}\left\{\sup_{2^{j-1}\varepsilon \leq \|\lambda-\theta\|_{\ell_1} \leq 2^j\varepsilon} |\hat{R}^0_\varphi(\mathsf{f}_\lambda) - \hat{R}^0_\varphi(\mathsf{f}_\theta)| \geq 2^{j-2}\varepsilon\zeta\right\}.\end{aligned}$$

Now, set

$$Z_j = \sup_{\|\lambda-\theta\|_{\ell_1} \leq 2^j\varepsilon} |\hat{R}^0_\varphi(\mathsf{f}_\lambda) - \hat{R}^0_\varphi(\mathsf{f}_\theta)|$$

and the same considerations leading to the final bound of (I) above yield

$$\mathbb{E}[Z_j] \leq 2^j \varepsilon C_\phi C_F \frac{\sqrt{2\log(2M)}}{\sqrt{n}}$$

and for $t = 1/\sqrt{2}$, we obtain

$$(II) \leq \zeta + 2C_\varphi C_F \sum_{j=1}^J \mathbb{P}\{Z_j - \mathbb{E}[Z_j] \geq 2^{j-2}\varepsilon\zeta - \mathbb{E}[Z_j]\}.$$

A change of a single pair (X_i, Y_i) changes Z_j by at most $2C_\varphi C_F(2^j\varepsilon)/n$, so that another application of the bounded differences inequality [3], Theorem 2.2, page 8, gives, by taking

$$\zeta = 7C_\varphi C_F \frac{\sqrt{2\log 2(M \vee n)}}{\sqrt{n}},$$

the final bound

$$\begin{aligned}
&\sum_{j=1}^{J} \mathbb{P}\{Z_j - \mathbb{E}[Z_j] \geq 2^{j-2}\varepsilon\zeta - \mathbb{E}[Z_j]\} \\
&\quad \leq \sum_{j=1}^{J} \mathbb{P}\left\{ Z_j - \mathbb{E}[Z_j] \geq t \cdot 2^j C_\varphi C_F \varepsilon \frac{\sqrt{2\log(2M \vee 2n)}}{\sqrt{n}} \right\} \\
&\quad \leq J \exp\left\{ -2\frac{t^2 (C_\varphi C_F 2^j \varepsilon)^2 2\log(2M \vee 2n)}{(2C_\phi C_F 2^j \varepsilon)^2} \right\} \\
&\quad = J(2M \vee 2n)^{-t^2} < J/\sqrt{2M \vee 2n}.
\end{aligned}$$

Finally, we invoke (B.1) to complete the proof of Proposition B.1. □

Acknowledgements

The research of Marten Wegkamp was supported in part by NSF Grant DMS-0706829. The research of Ming Yuan was supported in part by NSF Grant DMS-08-46234 and NIH Grant R01GM076274-01.

References

[1] Bartlett, P.L. and Wegkamp, M.H. (2008). Classification with a reject option using a hinge loss. *J. Mach. Learn. Res.* **9** 1823–1840. MR2438825

[2] Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[3] Devroye, L. and Lugosi, G. (2000). *Combinatorial Methods in Density Estimation.* New York: Springer. MR1843146

[4] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning.* New York: Springer. MR1851606

[5] Herbei, R. and Wegkamp, M.H. (2006). Classification with reject option. *Canad. J. Statist.* **34** 709–721. MR2347054

[6] Koltchinskii, V. (2009). Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincaré Probab. Statist.* **45** 7–57. MR2500227

[7] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces.* New York: Springer. MR1102015

[8] Tarigan, B. and van de Geer, S.A. (2006). Classifiers of support vector machine type with ℓ_1 complexity regularization. *Bernoulli* **12** 1045–1076. MR2274857

[9] Tsybakov, A.B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002

[10] van de Geer, S.A. (2000). *Empirical Processes in M-estimation.* Cambridge: Cambridge Univ. Press.

[11] Wegkamp, M.H. (2007). Lasso type classifiers with a reject option. *Electron. J. Statist.* **1** 155–168. MR2312148

[12] Yuan, M. and Wegkamp, M.H. (2010). Classification methods with reject option based on convex risk minimization. *J. Mach. Learn. Res.* **11** 111–130. MR2591623

Received June 2009 and revised January 2010