Bernoulli **12**(6), 2006, 1045–1076

Classifiers of support vector machine type with ℓ_1 complexity regularization

BERNADETTA TARIGAN* and SARA A. VAN DE GEER**

Seminar für Statistik, ETH Zürich, LEO D11, 8092 Zürich, Switzerland.
*E-mail: *tarigan@stat.math.ethz.ch; **geer@stat.math.ethz.ch*

We study the binary classification problem with hinge loss. We consider classifiers that are linear combinations of base functions. Instead of an ℓ_2 penalty, which is used by the support vector machine, we put an ℓ_1 penalty on the coefficients. Under certain conditions on the base functions, hinge loss with this complexity penalty is shown to lead to an oracle inequality involving both model complexity and margin.

Keywords: binary classification; hinge loss; margin; oracle inequality; penalized classification rule; sparsity

1. Introduction

Let (X, Y) be random variables, with $X \in \mathcal{X}$ a *feature* and $Y \in \{-1, +1\}$ a binary *label*. The problem is to predict Y given X. A classifier is a function $f : \mathcal{X} \to \mathbb{R}$. Using the classifier f, we predict the label $+1$ when $f(X) \geqslant 0$, and the label -1 when $f(X) < 0$. Thus, a classification error occurs when $Yf(X) \leqslant 0$.

Let P be the distribution of the pair (X, Y), and denote the marginal distribution of X by Q. Moreover, write the regression of Y on X as

$$\eta(x) := P(Y = 1 | X = x), \qquad x \in \mathcal{X}.$$

Our aim is to find a classifier which makes the correct classification with high probability. The probability of misclassification by the classifier f or *prediction error* of f, is

$$P(Yf(X) \leqslant 0).$$

Bayes' (decision) rule is

$$f^* := \begin{cases} +1, & \text{if } \eta \geqslant \frac{1}{2}, \\ -1, & \text{if } \eta < \frac{1}{2}. \end{cases}$$

It is easy to see that the prediction error is smallest when using Bayes' rule. The function η is, however, not known. To estimate Bayes' rule, we take a sample from P. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be observed independent and identically distributed copies of (X, Y). These observations are called the *training set*. The sample size is assumed to be large enough to permit a nonparametric approach to the estimation problem. We assume $n \geqslant 8$ to avoid nonsense expressions later on.

1350–7265

Let $\mathcal{F}$ be a collection of classifiers. In empirical risk minimization, one chooses the classifier in $\mathcal{F}$ that has the smallest number of misclassifications in the sample (see Vapnik 1995, 1998). However, if $\mathcal{F}$ is a rich set, this classifier will generally be hard to compute. We will indeed consider a very high-dimensional class $\mathcal{F}$ in this paper. By replacing the number of misclassifications (i.e., 0/1 loss) by *hinge* loss one can overcome computational problems. The *support vector machine* (SVM) adds an ℓ_2 penalty (or quadratic penalty) to the hinge loss function. We propose instead to employ an ℓ_1 penalty. This yields a computationally feasible complexity regularization method and we show that the procedure can yield estimators that adjust to favourable distributions P.

The empirical hinge loss function is

$$R_n(f) := \frac{1}{n}\sum_{i=1}^{n} l(Y_i f(X_i)),$$

where $l(z) = (1-z)_+$, with z_+ denoting the positive part of $z \in \mathbb{R}$. The function l is called the *hinge* function. Define the theoretical hinge loss

$$R(f) := \mathbb{E}_n(f) = E(l(Yf(X))).$$

Hinge loss is consistent in the sense that Bayes' decision rule f^* minimizes the theoretical hinge loss

$$f^* = \arg\min_{\text{all } f} R(f)$$

(see Lin 2002).

For the collection of classifiers $\mathcal{F}$, we choose a subset of a high-dimensional linear space. Consider a given system $\{\psi_k : k = 1, \ldots, m\}$ of functions on $\mathcal{X}$. We call $\{\psi_k\}$ the collection of base functions. We assume throughout that $C_Q^2 < \infty$, where

$$C_Q^2 := \max_{1 \le k \le m} \int \psi_k^2 \, \mathrm{d}Q \tag{1}$$

is the largest squared $L_2(Q)$ norm of the base functions ψ_k. However, we do not require C_Q to be known.

For $\alpha \in \mathbb{R}^m$ define

$$f_\alpha(x) := \sum_{k=1}^{m} \alpha_k \psi_k(x), \qquad x \in \mathcal{X}.$$

We then take $\mathcal{F} \subset \{f_\alpha : \alpha \in \mathbb{R}^m\}$. The number of base functions ψ_k is allowed to be very large, up to

$$m \leqslant n^D, \tag{2}$$

for some $D \geqslant 1$. The SVM minimizes the empirical hinge loss with, to avoid overfitting, an ℓ_2 penalty on the coefficients α, that is, a quadratic penalty on the classifier f_α proportional to $\sum \alpha_k^2$ (or a weighted version thereof). In fact, classical SVMs take $\mathcal{F}$ not exactly as a (subset of a) finite-dimensional space, but rather as a reproducing kernel Hilbert space (see Section 2.3 for more details). SVMs were introduced by Boser *et al.* (1992), and have been applied extensively since then. The book by Schölkopf and Smola (2002) contains a good overview of SVMs and related learning theory.

As a variant of the SVM procedure, we propose to add an ℓ_1 complexity penalty, instead of an ℓ_2 complexity penalty, to the empirical hinge loss. The ℓ_1 penalty is proportional to the ℓ_1 norm $\sum_{k=1}^m |\alpha_k|$ of the coefficients. The ℓ_1 penalized minimum hinge loss estimator $\hat{f}_n$ is then defined as

$$\hat{f}_n := \arg\min_{f_\alpha \in \mathcal{F}} \left\{ R_n(f_\alpha) + \hat{\lambda}_n \sum_{k=1}^m |\alpha_k| \right\}, \tag{3}$$

where $\hat{\lambda}_n$ is a regularization parameter.

Under Conditions A, B and C below, an appropriate choice for the regularization parameter is

$$\hat{\lambda}_n := c \max(\hat{C}_n, 4) D \mathbf{K}^2 \sqrt{\frac{\log n}{n}}.$$

Here, c is required to be larger than some given universal constant, but is otherwise arbitrary. The quantity $\hat{C}_n$ is the largest empirical L_2 norm of the base functions $\{\psi_k\}$ (see (8)), that is, an estimate of C_Q defined in (1). The constant D is from (2). Finally, $\mathbf{K}$ is either an assumed given bound K_0 on the supremum norm of the functions in $\mathcal{F}$, or, under some other assumptions, $\mathbf{K} = 1$. More precisely, in Theorem 2.1, we require, for technical reasons, that

$$\|f\|_\infty \leqslant K_0 \qquad \forall f \in \mathcal{F}, \tag{4}$$

for some constant $K_0 \geqslant 1$, where

$$\|f\|_\infty := \sup_{x \in \mathcal{X}} |f(x)|$$

denotes the supremum norm of the function f. The dependency on K_0 of our results will be given explicitly. On the other hand, in Theorem 2.2, we let $\mathbf{K} = 1$. We assume there that for some constant K_n,

$$\|f - \tilde{f}\|_\infty \leqslant \max(K_n \|f - \tilde{f}\|_{1,\nu}, 2), \qquad \forall f, \tilde{f} \in \mathcal{F}. \tag{5}$$

Here, $\|\cdot\|_{1,\nu}$ denotes the $L_1(\nu)$ norm, with ν some measure on $\mathcal{X}$, depending on the other conditions (Conditions A and B). The constant K_n is allowed to be large, depending in fact on the rate an 'oracle' would have (see Theorem 2.2). Assumption (5) is more attractive than (4). Nevertheless, we first present the result under condition (4) because, as it turns out, our proof of Theorem 2.2 relies heavily on Theorem 2.1.

The ℓ_1 penalty generally leads to sparse representations, that is, it tends to result in an estimator $\hat{f}_n = \sum_{k=1}^{m} \hat{\alpha}_k \psi_k$ with only a few non-zero coefficients $\hat{\alpha}_k$ (see Zhu *et al.* 2003). It is related to soft thresholding (see Donoho 1995), and is referred to as the Lasso in Tibshirani (1996) and Hastie *et al.* (2001: Section 10.12).

The difference $R(f) - R(f^*)$ is called the hinge *excess risk* at f. We show in Theorems 2.1 and 2.2 that the hinge excess risk at $\hat{f}_n$ depends on the smoothness of the boundary of $\{f^* = 1\}$, as well as on the *margin* behaviour. The latter quantifies the identifiability of Bayes' decision rule. For definiteness, we assume it can be summarized in a *margin parameter* (or *noise level*) κ, defined in Condition A below.

Whether or not Theorems 2.1 and 2.2 produce good rates for the prediction error depends very much on the choice of the system of base functions $\{\psi_k\}$. A more detailed discussion of the problem can be found below, after the statement of the conditions and theorems.

In the literature, the term adaptivity generally refers to attaining (up to log terms) the minimax rate. Adaptivity and minimax rates have been established in Tsybakov (2004) and Tsybakov and van de Geer (2005). These papers use empirical risk minimization, which make the methods proposed there difficult to implement. In other work, for example, Koltchinskii (2001, 2006), Koltchinskii and Panchenko (2002) and Lugosi and Wegkamp (2004), Rademacher complexities are applied. Audibert (2004) establishes adaptivity to the margin for a Gibbs classifier. Scott and Nowak (2006) develop a computationally attractive tree method that adaptively attains minimax rates no faster than $n^{-1/2}$. However, their method only applies to low-dimensional input spaces. In Section 4.2, we establish rates that depend on margin and complexity, but we will not show that this is in fact (near) adaptation to minimax rates. To show the latter, one has to carefully define the class of probability measures over which one studies the minimax bounds (see also Remark 4.1).

This paper is organized as follows. In the next section, we present the conditions (Conditions A, B and C) and main theorems. The results are followed by a discussion of their impact and their relation to averaging classifiers and to kernel SVMs. Here we also address the problem that the hinge excess risk may not be a good approximation of the prediction error.

Section 3 takes a closer look at the conditions. The margin condition (Condition A) is shown to follow from assumptions on the amount of mass located near $\eta = \frac{1}{2}$, and possibly also on the behaviour near the boundaries $\eta = 0$ and $\eta = 1$. Also an extension is considered, as well as an extension of Condition B. It can be observed that the choice of the smoothing parameter $\hat{\lambda}_n$ does not depend on the constants appearing in Conditions A and B, and that the procedure adjusts to favourable values of these constants. Condition C is a technical condition on the base functions.

In Section 4 we consider an example. The rates obtained there depend on the roughness of the boundary of Bayes' decision rule and on the margin. We consider the case where Bayes' decision rule is a boundary fragment. We apply Haar wavelets, which provide piecewise constant approximations of the boundary and are closely related to the binary decision trees studied in Scott and Nowak (2006).

The proofs of the main theorems are given in Section 5. Here, we use the tools provided by empirical process theory, such as concentration and contraction inequalities. The proofs of the results in Section 4 can be found in Section 6.

2. A probability inequality for the ℓ_1 penalized minimum hinge loss estimator

2.1. Conditions and main theorems

Let ν be some measure on $\mathcal{X}$, and let $\|\cdot\|_{p,\nu}$ be the $L_p(\nu)$ norm ($1 \leqslant p < \infty$). We assume Conditions A and B below to hold for the same (unknown) measure ν.

Condition A is an identifiability condition, which we refer to as the *margin* condition.

Condition A. *There exist constants $\sigma > 0$ and $\kappa \geqslant 1$ such that, for all $f \in \mathcal{F}$,*

$$R(f) - R(f^*) \geqslant \|f - f^*\|_{1,\nu}^{\kappa}/\sigma^{\kappa}. \tag{6}$$

The parameter κ is called the *margin* parameter. Its value is generally not known.

Next we impose conditions on the system $\{\psi_k\}$. We use the notation $\psi = (\psi_1, \ldots, \psi_m)^{\mathrm{T}}$ and define $\Sigma_\nu := \int \psi\psi^{\mathrm{T}}\,\mathrm{d}\nu$ (assumed to exist). The smallest eigenvalue of Σ_ν is denoted by ρ_ν^2.

Condition B. *The smallest eigenvalue ρ_ν^2 of Σ_ν is non-zero.*

The value of ρ_ν^2 is generally also unknown.

The last condition puts a normalization on the system of functions $\{\psi_k\}$.

Condition C. *We assume*

$$\max_{1\leqslant k\leqslant m} \|\psi_k\|_\infty \leqslant \sqrt{\frac{n}{\log n}}. \tag{7}$$

Recall also the requirement that $C_Q < \infty$, and $m \leqslant n^D$ for some $D \geqslant 1$.

We now introduce the concepts approximation error and estimation error. Let $N(\alpha)$ be the number of non-zero coefficients in the vector α:

$$N(\alpha) := \#\{\alpha_k \neq 0\}, \qquad \alpha \in \mathbb{R}^m.$$

Given $N \in \{1, \ldots, m\}$, the approximation error is

$$\inf\bigl\{R(f_\alpha) - R(f^*) : f_\alpha \in \mathcal{F},\, N(\alpha) = N\bigr\}.$$

Let

$$\hat{C}_n^2 := \max_{1\leqslant k\leqslant m} \frac{1}{n}\sum_{i=1}^{n} \psi_k^2(x_i) \tag{8}$$

be the empirical version of the constant C_Q^2 defined in (1), and let the smoothing parameter be

$$\hat{\lambda}_n := c(\hat{C}_n \vee 4)D\mathbf{K}^2\sqrt{\frac{\log n}{n}}. \tag{9}$$

Here, $c \geqslant c_0$, with c_0 a universal constant. (From Section 5, a suitable choice is $c_0 = 864$.) Moreover, here and throughout we use, for a, $b \in \mathbb{R}$, the notation $a \vee b := \max\{a, b\}$. Likewise, $a \wedge b := \min\{a, b\}$. The value of $\mathbf{K}$ will be specified in Theorems 2.1 and 2.2. We let λ_n be the theoretical version of $\hat{\lambda}_n$:

$$\lambda_n := c(C_Q \vee 4)D\mathbf{K}^2\sqrt{\frac{\log n}{n}}. \tag{10}$$

As function of n, the value of the theoretical smoothing parameter behaves as $\sqrt{\log n/n}$. This is as in hard and soft thresholding (see, for example, Donoho 1995).

Define (a bound for) the 'estimation error' as

$$V_n(N) := 2\delta^{-1/(2\kappa-1)}(18\sigma\lambda_n^2 ND\mathbf{K}/\rho_\nu^2)^{\kappa/(2\kappa-1)}, \tag{11}$$

where $0 < \delta \leqslant \frac{1}{2}$ is fixed but otherwise arbitrary. Theorem 2.1 tells us that the estimation error and approximation error are traded off over all $f_\alpha \in \mathcal{F}$. The trade-off is reflected in the quantity

$$\epsilon_n := (1 + 4\delta)\inf\left\{ R(f_\alpha) - R(f^*) + V_n(N(\alpha)) + 2\lambda_n\mathbf{K}\sqrt{\frac{\log n}{n}} : f_\alpha \in \mathcal{F} \right\}. \tag{12}$$

By the trade-off, the ℓ_1 penalized minimum hinge loss estimator adjusts to certain properties of the unknown distribution P. Thus, it has the potential to produce fast rates for the excess risk $R(\hat{f}_n) - R(f^*)$.

Theorem 2.1. *Let $\hat{f}_n$ be the ℓ_1 penalized minimum hinge loss estimator defined in* (3)*, with regularization parameter $\hat{\lambda}_n$ given in* (9)*, where $c \geqslant c_0$, with c_0 an appropriate universal constant. Suppose that Conditions A, B, and C hold. Assume also that $\mathcal{F} \subset \{f_\alpha : \alpha \in \mathbb{R}^m\}$ and*

$$\|f\|_\infty \leqslant K_0 \qquad \forall f \in \mathcal{F},$$

where $K_0 \geqslant 1$. Let $\mathbf{K} = K_0$ in the definition (9) ((10)) *of $\hat{\lambda}_n$ (λ_n). Then, for a universal constant c_1,*

$$\mathbb{P}\Big(R(\hat{f}_n) - R(f^*) > \epsilon_n\Big) \leqslant \frac{c_1}{n^2}. \tag{13}$$

The estimation error $V_n(N)$ would be a bound for the error due to sampling, if a priori the estimator were required to have only a given set, of cardinality N, of non-zero coefficients. An oracle would choose the optimal set of non-zero coefficients by balancing sparseness and approximation error. In this sense, the theorem shows that the estimator mimics an oracle. Note that the balance is based on the ℓ_0 penalty which counts the number of non-zero coefficients. The similarity of ℓ_0 and ℓ_1 penalties is well known, and in fact goes through for underdetermined systems (see Donoho 2004a, 2004b).

It is of interest to examine the behaviour of the estimator for large n. Suppose the

constants C_Q, D, K_0, σ, κ and ρ_v are fixed (i.e., not dependent on the sample size n). Let us call this the *standard situation*. In the standard situation, $\lambda_n^2 = O(\log n/n)$, and the estimation error is of order

$$V_n(N) = O(N \log n/n)^{\kappa/(2\kappa-1)}.$$

For example (for a given N), the worst case corresponds to $\kappa = \infty$, giving $V_n(N) = O(\sqrt{N \log n/n})$. The rates for the estimation error are as in Tsybakov and van de Geer (2005). However, the latter paper deals with empirical risk minimization and prediction error excess risk (see below for the definition of the latter), which means that the rates established there may be quite different from those following from Theorem 2.1.

The prediction error excess risk is $P(Yf(X) \leqslant 0) - P(Yf^*(X) \leqslant 0)$. Rates of convergence for the hinge excess risk imply the same rates for the prediction error excess risk, as Zhang (2004) has shown that

$$P(Yf(X) \leqslant 0) - P(Yf^*(X) \leqslant 0) \leqslant R(f) - R(f^*) \tag{14}$$

(see also Bartlett *et al.* 2006). It is easy to see, however, that this inequality cannot be reversed. In particular, for $0 < \epsilon < 1$, the classifier $f := \epsilon f^*$ has zero prediction error excess risk, but, in view of Remark 3.1 below, hinge excess risk equal to the constant $(1-\epsilon)\int |1-2\eta|\,\mathrm{d}Q$. In the trade-off given by Theorem 2.1, however, it is the hinge excess risk that enters as approximation error. In other words, this trade-off may not reflect the trade-off between non-sparseness and prediction error excess risk. For that, we need a margin condition of the form of Condition A, with Bayes' rule f^* replaced by the mimimizer of the hinge loss over all $f \in \mathcal{F}$, and in addition an extended version of (14). In Section 4 we discuss an example where the two types of excess risk will be of the same order of magnitude.

We now consider a variant of Theorem 2.1, with essentially weaker conditions.

Theorem 2.2. *Let $\hat{f}_n$ be the ℓ_1 penalized minimum hinge loss estimator defined in* (3)*, with regularization parameter $\hat{\boldsymbol{\lambda}}_n$ given in* (9)*, where $c \geqslant c_0$, with c_0 an appropriate universal constant. Suppose that Conditions A, B, and C hold. Assume also that $\mathcal{F} \subset \{f_\alpha : \alpha \in \mathbb{R}^m\}$ is a convex set, and that for some constant K_n,*

$$\|f - \tilde{f}\|_\infty \leqslant (K_n \|f - \tilde{f}\|_{1,v}) \vee 2, \qquad \forall f, \tilde{f} \in \mathcal{F}. \tag{15}$$

Then, take $\mathbf{K} = 1$ in (9) *and* (10)*. If*

$$2\sigma \epsilon_n^{1/\kappa} K_n \leqslant 1, \tag{16}$$

we have, for a universal constant c_1,

$$\mathbb{P}\Big(R(\hat{f}_n) - R(f^*) > \epsilon_n\Big) \leqslant \frac{c_1}{n^2}. \tag{17}$$

Theorem 2.2 illustrates that the condition on the supremum norm of Theorem 2.1 can be weakened. The constant K_n will generally grow with n. For certain systems $\{\psi_k\}_{k=1}^m$, condition (16) is met when the number of base functions is small enough, yet large enough to allow approximation error and estimation error to balance. In other cases, inequality (15)

is a restriction on the allowed linear combinations. When there are no known bounds on κ and on the complexity of the problem, it is actually not possible to verify (16). This is, however, in line with Conditions A and B, which also cannot be verified.

In practice, we recommend that the ℓ_1 penalized estimator is computed over *all* f_α, $\alpha \in \mathbb{R}^m$, and that the choice of the smoothing parameter $\hat{\lambda}_n$ is decided upon by applying cross validation.

Remark 2.1. Our proof of the two theorems relies on the fact that one has the Lipschitz property

$$|l(y, f(x)) - l(y, \tilde{f}(x))| \leqslant |f(x) - \tilde{f}(x)|, \qquad \forall f, \tilde{f} \in \mathcal{F},$$

where $l(y, f(x)) = (1 - yf(x))_+$ is the hinge loss function. Theorem 2.2, moreover, uses the convexity of this loss function. The results can be extended to hold for any convex loss function $l(y, f_\alpha(x))$ with this Lipschitz property. The extension can, for example, be used to derive similar results to Theorems 2.1 and 2.2 for robust regression. Loubes and van de Geer (2002) and van de Geer (2003) proceed essentially along these lines, but study fixed design instead of random design.

Results for averages of classifiers, and kernel estimators, using ℓ_1 penalties, call for a different mathematical theory. We will briefly explain why in the next two subsections.

2.2. Averaging classifiers

When averaging classifiers, one introduces a collection of base classifiers $\{\psi_k\}$ and forms weighted averages $f_\alpha := \sum_k \alpha_k \psi_k$, where the weights α_k are assumed to be positive and sum to one. More generally, one may consider arbitrary linear combinations. One often supposes that the base classifiers $\{\psi_k\}$ form a Vapnik–Chervonenkis class of fixed dimension V (e.g., stumps). This set-up is different from ours in several respects. Firstly, the class of base classifiers may be infinite. However, one may usually replace it by a finite set, virtually without changing the situation. A more severe problem is that Σ_ν generally will have very small eigenvalues, as the base classifiers are highly correlated. And finally, Bayes' decision rule is generally not well approximated by such averages (unless it is itself one of the base classifiers). This means that the hinge excess risk for such averages is generally large. It is not clear, however, whether the same will be true for the prediction error excess risk. We conclude that Theorem 2.1 or Theorem 2.2 is not intended for the situation of averaging.

The picture is clearer when one alternatively considers estimating the regression function η, for example using exponential, quadratic or logistic loss. For these loss functions, Blanchard *et al.* (2003) have obtained rates of convergence for averaged classifiers. They also consider ℓ_1 penalties but different loss functions, and their results are not in the framework of sparseness. Their rates of convergence follow from the Vapnik–Chervonenkis dimension of the set of base classifiers.

2.3. Kernel representations

It is customary to minimize the hinge loss over a reproducing kernel Hilbert space, with kernel $\mathcal{K}$ (say) on $\mathcal{X} \times \mathcal{X}$. Suppose that $\mathcal{K}$ has eigenexpansion

$$\mathcal{K}(x, \tilde{x}) = \sum_{k=1}^{\infty} \beta_k \phi_k(x) \phi_k(\tilde{x}), \qquad (x, \tilde{x}) \in \mathcal{X} \times \mathcal{X}.$$

Here $\{\beta_k\}$ are the (non-zero) eigenvalues of $\mathcal{K}$, and $\{\phi_k\}$ are the eigenfunctions. Suppose we use the representation $f_\alpha = \sum_{k=1}^{\infty} \alpha_k \psi_k$, with $\psi_k = \phi_k$. Then in our set-up we employ the penalty

$$\text{pen}(f_\alpha) = \hat{\lambda}_n \sum_k |\alpha_k|. \tag{18}$$

This penalty is meaningful if Bayes' rule f^* can be well approximated by a sparse representation in terms of the eigenfunctions of the kernel $\mathcal{K}$. The more usual penalty is

$$\text{pen}(f_\alpha) = \lambda \|f_\alpha\|_{\mathcal{K}}^2, \tag{19}$$

where λ is a regularization parameter, and where $\|f_\alpha\|_{\mathcal{K}}^2 := \sum_k |\alpha_k|^2 / \beta_k^2$ (see, for example, Schölkopf and Smola 2002: Section 1.5). The eigenvalues $\{\beta_k\}$ of the kernel typically decrease to zero as $k \to \infty$ (for example, for Gaussian kernels the decay is exponentially fast), so that the penalty in (18) is substantially different from the more standard choice (19).

We conjecture that for a choice of λ depending only on $\{X_i\}$ (and not on $\{Y_i\}$), the penalty in (19) cannot be adaptive to κ in the sense we put forward in Theorem 2.1. The reason why we believe this to be true is that with the quadratic penalty (19) a good choice for λ will be such that the estimation error, which depends on κ, is overruled. We do expect that λ in (19) can be chosen (rate-)adaptively using cross validation.

We remark that the kernel usually is allowed to depend on a second regularization parameter, called the *width*. For example, for $\mathcal{X}$ a compact in $\mathbb{R}^d$, one may apply the Gaussian kernel

$$\mathcal{K}(x, \tilde{x}) := \exp(-|x - \tilde{x}|^2 / h^2),$$

with width (proportional to) h^d. Both λ and h are often chosen data-dependent. Rates for the general kernels and the penalty (19), but with the restriction $\kappa = 1$, are given in Blanchard *et al.* (2004). The situation with Gaussian kernels, penalty (19) and known κ is examined in Steinwart and Scovel (2005).

3. On Conditions A, B and C

Conditions A and B together ensure that the result follows easily from a probability inequality for the empirical process (see Lemmas 5.1–5.3). Condition C makes sure that the probability inequality does indeed hold (see Lemmas 5.4–5.7).

3.1. On Condition A

Condition A is a lower bound for the hinge excess risk in terms of the $L_1(\nu)$ norm $\|\cdot\|_{1,\nu}$. We have restricted ourselves to this particular form for ease of exposition. A more general assumption is

$$R(f) - R(f^*) \geqslant G\Big(\|f - f^*\|_{1,\nu}^{1/2}\Big), \qquad \forall f \in \mathcal{F},$$

with $G(\cdot) := \int_0^{\cdot} g(z)\mathrm{d}z$ and g a continuous, strictly increasing function on $[0, \infty)$ satisfying $g(0) = 0$. The estimation error will then be

$$V_n(N) = 2\delta H\left(\frac{3\lambda_n\sqrt{2ND\mathbf{K}}}{\delta\rho_\nu}\right),$$

where

$$H(\cdot) := \int_0^{\cdot} g^{-1}(z)\mathrm{d}z.$$

This follows from the proof of Lemma 5.1, with Lemma 5.3 replaced therein by Young's inequality (for the latter, see, for example, Hardy *et al.* 1988: Section 8.3).

We restricted ourselves in Condition A to $G(z) = z^{2\kappa}/\sigma^{\kappa}$. This appears in similar form (for prediction error instead of hinge loss) in, for example, Mammen and Tsybakov (1999), Audibert (2004), Tsybakov (2004), Bartlett *et al.* (2006) and Scott and Nowak (2006). It follows essentially from conditions on the behaviour of η near $\{\eta = \frac{1}{2}\}$ and is therefore often called the *margin* condition, or condition on the *noise level*; see Condition AA below, which was first formulated by Tsybakov (2004).

Condition AA. *There exist constants $C \geqslant 1$ and $\gamma \geqslant 0$ such that, for all $z > 0$,*

$$Q(\{|1 - 2\eta| \leqslant z\}) \leqslant (Cz)^{1/\gamma}, \tag{20}$$

where, by convention, $(Cz)^{1/\gamma} = 1\{z \geqslant 1/C\}$ for $\gamma = 0$.

The case where $\gamma = 0$ corresponds to the situation where the function η stays away from $\frac{1}{2}$. This is the situation studied in Blanchard *et al.* (2004). The larger the value of γ the weaker (20) becomes, and for $\gamma = \infty$ it is satisfied for all distributions. If η only takes values very near to $\frac{1}{2}$, Bayes' decision rule is not much better than flipping a fair coin and (20) can only hold for large values of γ.

We will see that Condition A is closely intertwined with assumptions on the behaviour of η near $\{\eta = 0\}$ and $\{\eta = 1\}$ as well. In principle, values of η near 0 or 1 are favourable as they make the learning problem easier. However, these values make it harder to identify Bayes' rule in (say) $\|\cdot\|_{Q,1}$ norm. We show that Condition A holds with $\mathrm{d}\nu = \eta(1 - \eta)\mathrm{d}Q$ and $\kappa = 1 + \gamma$. This is a slight modification of Tsybakov (2004).

Lemma 3.1. *Suppose Condition AA is met. Then, for all f with $\|f - f^*\|_\infty \leqslant K$,*

$$R(f) - R(f^*) \geqslant \sigma_K^{-1} \|f - f^*\|_{1,\nu}^{1+\gamma}, \tag{21}$$

with $\mathrm{d}\nu = \eta(1-\eta)\mathrm{d}Q$ *and*

$$\sigma_K = C\left(\frac{K}{4}\left(\frac{1}{\gamma}+1\right)\right)^{\gamma}(1+\gamma). \tag{22}$$

Thus, Condition A holds with $\sigma = \sigma_K^{1/\kappa}$ *and* $\kappa = 1+\gamma$.

Proof. By straightforward manipulation, we obtain

$$\begin{aligned} R(f) - R(f^*) = &\int_{-1\leqslant f\leqslant 1} |(f-f^*)(1-2\eta)|\,\mathrm{d}Q \\ &+ \int_{f<-1,\eta\leqslant 1/2} |f-f^*|\eta\,\mathrm{d}Q + \int_{f<-1,\eta>1/2} |(f-f^*)(1-2\eta)|\,\mathrm{d}Q \\ &+ \int_{f<-1,\eta>1/2} (|f|-1)(1-\eta)\mathrm{d}Q + \int_{f>1,\eta\leqslant 1/2} |(f-f^*)(1-2\eta)|\,\mathrm{d}Q \\ &+ \int_{f>1,\eta\leqslant 1/2} (|f|-1)\eta\,\mathrm{d}Q + \int_{f>1,\eta>1/2} |f-f^*|(1-\eta)\mathrm{d}Q. \end{aligned}$$

This implies the inequality

$$R(f) - R(f^*) \geqslant \int |f-f^*|\,|1-2\eta|\,\mathrm{d}\nu,$$

with $\mathrm{d}\nu = \eta(1-\eta)\mathrm{d}Q$. Hence, for any $z > 0$,

$$\begin{aligned} R(f) - R(f^*) &\geqslant \int_{|1-2\eta|>z} |f-f^*|\,|1-2\eta|\,\mathrm{d}\nu \geqslant z\int_{|1-2\eta|>z} |f-f^*|\,\mathrm{d}\nu \\ &= z\|f-f^*\|_{1,\nu} - z\int_{|1-2\eta|\leqslant z} |f-f^*|\,\mathrm{d}\nu. \end{aligned}$$

But, since $\|f-f^*\|_\infty \leqslant K$ and $\eta(1-\eta) \leqslant \frac{1}{4}$,

$$\int_{|1-2\eta|\leqslant z} |f-f^*|\,\mathrm{d}\nu \leqslant \frac{K}{4}\,Q(\{|1-2\eta|\leqslant z\}) \leqslant (K/4)(Cz)^{1/\gamma},$$

where we invoke Condition AA. Thus, for all $z > 0$,

$$R(f) - R(f^*) \geqslant z\|f-f^*\|_{1,\nu} - \frac{K}{4}(Cz)^{1/\gamma}z.$$

When $\gamma > 0$, we take

$$z = \left(\frac{4\|f-f^*\|_{1,\nu}}{C^{1/\gamma}K(1/\gamma+1)}\right)^{\gamma},$$

and for $\gamma = 0$, we take $z \uparrow 1/C$. We thus arrive at the result of the lemma. □

Remark 3.1. An intermediate result of the proof of Lemma 3.1 is that

$$R(f) - R(f^*) \geqslant \int_{-1 \leqslant f \leqslant 1} |(f - f^*)(1 - 2\eta)| \, \mathrm{d}Q,$$

with equality if $\|f\|_\infty \leqslant 1$. For an f taking only the values ± 1, the hinge excess risk is therefore equal to twice the prediction error excess risk. (We will use this in Section 4.) The proof of Lemma 3.1 thus leads also to Zhang's (2004) inequality (see (14)).

Remark 3.2. The choice $\mathrm{d}\nu = \eta(1 - \eta)\mathrm{d}Q$ is in our view quite natural, as the conditional variance of $Yf(X)$ is equal to, given X, satisfies

$$\mathrm{E}(\mathrm{var}(yf(X))|X = 4 \int f^2 \eta(1 - \eta)\mathrm{d}Q.$$

There are, however, also other reasonable candidates for ν. For example, let us define

$$\tau := \min\{\eta, 1 - \eta, |1 - 2\eta|\},$$

and suppose that instead of Condition AA, one has for some set S, some $C \geqslant 1$ and some $\gamma \geqslant 0$,

$$Q(\{\tau \leqslant z\} \cap S) \leqslant (Cz)^{1/\gamma} \qquad \forall z > 0.$$

Then, from the same arguments as used in the proof of Lemma 3.1, one sees that Condition A holds for all $\|f - f^*\|_\infty \leqslant K$, with $\mathrm{d}\nu = 1_S \, \mathrm{d}Q$, $\kappa = 1 + \gamma$ and $\sigma_K = C(K(1/\gamma + 1))^\gamma(1 + \gamma)$. For the set S one may want to take $S = \mathcal{X}$ or $S = \{\eta \notin \{0, 1\}\}$. Recall that ν also plays a role in Condition B, which means we would like to take the set S as large as possible.

Of course, if η stays away from 0 and 1, say $t \leqslant \eta \leqslant 1 - t$ for some $0 < t < \frac{1}{2}$, then the choices $\mathrm{d}\nu = \eta(1 - \eta)\mathrm{d}Q$ and $\upsilon = Q$ discussed above are, up to constants, the same. In Blanchard *et al.* (2003) it is noted that one may force oneself into such a situation by adding extra noise, namely, by replacing Y_i by $Y_i' = \omega_i Y_i$ ($Y' = \omega_0 Y$), where $\{\omega_i\}$ is a sequence of independent random variables, with $\mathbb{P}(\omega_i = 1) = 1 - \mathbb{P}(\omega_i = -1) = \frac{3}{4}$, independent of $\{(X, Y), (X_i, Y_i)\}$. For any classifier f, the prediction error excess risk for predicting Y is equal to twice the prediction error excess risk for predicting its noisy variant Y'. Such a simple relation is generally not true for the hinge excess risk.

In Blanchard *et al.* (2004) the condition that η stays away from 0 and 1 is imposed as well, in order to enable a precise formulation of a good penalty in that context.

Remark 3.3. From Lemma 3.1, one sees that the condition that, for some K, $\|f - f^*\|_\infty \leqslant K$ may be needed for Condition A to hold. This is not a priori assumed in Theorem 2.2. Therefore, let us mention the following weaker version of Condition A. Suppose for simplicity that the infimum in the definition of ϵ_n is attained, say in f_{α^*}. So

$$f_{\alpha^*} = \arg\min\{R(f_\alpha) - R(f^*) + V_n(N(\alpha)) : f_\alpha \in \mathcal{F}\}.$$

Then, in Theorem 2.2, it suffices to assume (6) for those $f \in \mathcal{F}$ with $\|f - f_{\alpha^*}\|_\infty \leqslant 2$. Thus,

if $\|f_{\alpha^*} - f^*\|_\infty \leqslant K_0$ for some K_0, it suffices to assume (6) for those $f \in \mathcal{F}$ with $\|f - f^*\|_\infty \leqslant K_0 + 2$.

3.2. On Condition B

3.2.1. The role of the $\|\cdot\|_{1,\nu}$ norm

Recall first that f^* is a renormalized indicator function. In Condition A, the occurrence of the $\|\cdot\|_{1,\nu}$ norm is closely related to the fact that, for indicator functions, the $L_2(\nu)$ norm $\|\cdot\|_{2,\nu}$ is equal to $\|\cdot\|_{1,\nu}^{1/2}$. In our proof, the $L_2(\nu)$ norm appears as intermediate in the inequality

$$\left(\sum_k |\alpha_k|\right)^2 \leqslant N(\alpha)K\|f_\alpha\|_{1,\nu}/\rho_\nu^2,$$

where it is assumed that $\|f_\alpha\|_\infty \leqslant K$ (see Lemma 5.2). In Tarigan and van de Geer (2004) it is shown that when ν is the Lebesgue measure on $[0, 1]^d$, one has in fact the following improved inequality for standard compactly supported wavelet systems $\{\psi_k\}$ on $[0, 1]^d$,

$$\left(\sum_k |\alpha_k|\right)^2 \leqslant \text{const.}N(\alpha)\|f_\alpha\|_{1,\nu}^2,$$

provided that $\{k : \alpha_k \neq 0\}$ is the set of all wavelets up to a given resolution level. In this paper, we do not employ this improved variant to avoid digressions. Moreover, as pointed out in Section 4, wavelets may not lead to sparse approximations of Bayes' decision rule.

3.2.2. Improving the estimation error bound

The smallest eigenvalue ρ_ν^2 appears in our definition (11) of the estimation error. If ρ_ν tends to zero as n tends to infinity, this will slow down the reates. Therefore, it is desirable to have ρ_ν stay away from zero. However, it is as yet unclear to what extent one can find systems $\{\psi_k\}$ with this property and, at the same time, good approximating properties.

We now propose a possible improvement of the bound for the estimation error. We replace Condition B by the following condition:

Condition BB. *For each index set $\mathcal{J} \subset \{1, \ldots, m\}$ there exists a non-negative $\mathcal{N}_{\mathcal{J},\nu}$ such that, for all α with $\|f_\alpha\|_\infty \leqslant K$, one has*

$$\left(\sum_{k\in\mathcal{J}} |\alpha_k|\right)^2 \leqslant \mathcal{N}_{\mathcal{J},\nu}\|f_\alpha\|_{1,\nu}.$$

With this condition we also have an extension of Lemma 5.2.

Theorems 2.1 and 2.2 hold with Condition BB instead of B, if we make the following adjustments. Define, for $\mathcal{J}(\alpha) := \{k : \alpha_k \neq 0\}$,

$$\mathcal{N}_\nu(\alpha) := \mathcal{N}_{\mathcal{J}(\alpha),\nu}.$$

Next, let $V_n(\alpha)$ be definition (11) with N/ρ_ν^2 replaced by $\mathcal{N}_\nu(\alpha)$:

$$V_n(\alpha) = 2\delta^{-1/(2\kappa-1)}(18\sigma\lambda_n^2\mathcal{N}_\nu(\alpha)D\mathbf{K})^{\kappa/(2\kappa-1)}.$$

Then, Theorems 2.1 and 2.2 remain true if in (12) we replace $V_n(N(\alpha))$ by $V_n(\alpha)$.

We give an elementary lemma to verify Condition B. It shows that for systems orthogonal in $L_2(\nu)$, for example, only the eigenvalues of the system chosen by the oracle matter.

Lemma 3.2. *Suppose that for some strictly positive weights $\{w_k\}_{k=1}^m$, the smallest eigenvalue of $W\Sigma_\nu W$, with $W = \mathrm{diag}(w_1, \ldots, w_m)$, is equal to one. Then Condition BB holds with*

$$\mathcal{N}_{\mathcal{J},\nu} = \sum_{k\in\mathcal{J}} w_k^2.$$

Proof. Let $\|v\|^2 = v^{\mathrm{T}}v$ denote the squared length of a vector $v \in \mathbb{R}^m$. We know that, for all $v \in \mathbb{R}^m$,

$$\|v\|^2 \leqslant v^{\mathrm{T}}W\Sigma_\nu W v, \tag{23}$$

as $W\Sigma_\nu W$ has smallest eigenvalue equal to one. So

$$\left(\sum_{k\in\mathcal{J}} |\alpha_k|\right)^2 \leqslant \left(\sum_{k\in\mathcal{J}} w_k^2\right)\left(\sum_{k\in\mathcal{J}} \frac{\alpha_k^2}{w_k^2}\right),$$

and, by using (23) with $v = W^{-1}\alpha$,

$$\sum_{k\in\mathcal{J}} \frac{\alpha_k^2}{w_k^2} = \alpha^{\mathrm{T}}W^{-2}\alpha \leqslant \alpha^{\mathrm{T}}\Sigma_\nu\alpha = \|f_\alpha\|_{2,\nu}^2.$$

Finally, we invoke the fact that $\|f_\alpha\|_{2,\nu}^2 \leqslant K\|f_\alpha\|_{1,\nu}$ for $\|f_\alpha\|_\infty \leqslant K$. □

3.3. On Condition C

We need a bound on both the $L_2(Q)$ norm $\|\psi_k\|_{2,Q}$ and the supremum norm $\|\psi_k\|_\infty$, which holds for all k. This allows us to apply Bernstein's inequality (see Lemmas 5.5 and 5.7). The uniform bound $\|\psi_k\|_\infty \leqslant \sqrt{n/\log n}$ holds, for example, for most compactly supported wavelet systems on $[0, 1]^d$, with no more than about $(\log_2(n/\log n))/d$ resolution levels per dimension. This means that, up to constants, the number of functions (wavelets) m in $\{\psi_k\}$ is also no more than about $n/\log n$. This bound on the resolution can mean that the rate of approximation is limited beforehand (see the example of Section 4).

In general, we assume polynomial growth $m \leqslant n^D$. This is quite standard in model selection problems, and rates are generally logarithmic in the a priori number of parameters m.

4. An example: boundary fragments

The choice of the base functions $\{\psi_k\}$ plays a crucial role in considerations on their approximating properties. Recall that Bayes' decision rule f^* takes only the values ± 1. Approximating such a function by, for example, an orthogonal series is not always very natural, as a good approximation might require very many non-zero coefficients. *Wedgelets* (Donoho 1999) and *curvelets* (Candès and Donoho 2004) are good alternatives to wavelets. Because these are overcomplete systems, our Condition B does not hold, so that the results in Section 2 are not applicable. This is the reason why we have chosen to nevertheless study wavelet approximations, in particular Haar functions. As Haar functions consider successive splits of intervals, this approach is related to classification by dyadic trees. Scott and Nowak (2006) derive rates for dyadic decision trees in a context similar to our example.

The main purpose of this section is to illustrate that Theorem 2.1 (or Theorem 2.2) can produce rates that adjust to roughness of the boundary of Bayes' decision rule, as well as to the margin. We will make some simplifying assumptions (in particular, Assumptions 1–4 below) to avoid digressions.

We consider the case $\mathcal{X} = [0, 1]^2$. Moreover, we suppose that f^* is a boundary fragment, that is, for some function $g^* : [0, 1] \to [0, 1]$,

$$f^*(x) = \begin{cases} +1, & \text{if } x \in \{(u, v) \in [0, 1]^2 : g^*(u) \geqslant v\}, \\ -1, & \text{otherwise.} \end{cases} \tag{24}$$

We also suppose that the boundary g^* is exactly the set where the regression function η is equal to $\frac{1}{2}$:

$$\eta(u, v) \begin{cases} > \frac{1}{2}, & \text{if } g^*(u) > v, \\ = \frac{1}{2}, & \text{if } g^*(u) = v, \\ < \frac{1}{2} & \text{if } g^*(u) < v. \end{cases} \tag{25}$$

Let μ be Lebesgue measure on $[0, 1]^2$. For a function g on $[0, 1]$, we use the notation

$$\|g\|_{p,\mu} := \|\bar{g}\|_{p,\mu}, \qquad 1 \leqslant p < \infty,$$

where $\bar{g}(u, v) = g(u)$, $(u, v) \in [0, 1]^2$.

To bound the excess risk, we make the following assumptions.

Assumption 1. *The distribution Q of X has density $q = \mathrm{d}Q/\mathrm{d}\mu$, satisfying, for some constant $0 < c_q < \infty$,*

$$1/c_q \leqslant q \leqslant c_q.$$

Assumption 2. *For some constant $0 < s \leqslant \frac{1}{2}$, $s \leqslant g^*(u) \leqslant 1 - s$ for all u.*

Assumption 3. *For some constants $0 < c_\eta < \infty$ and $0 < \gamma < \infty$,*

$$|v - g^*(u)|^\gamma / c_\eta \leqslant |2\eta(u, v) - 1| \leqslant c_\eta |v - g^*(u)|^\gamma, \qquad \forall (u, v) \in [0, 1]^2.$$

Thus, we require in Assumption 3 that for each u, $2\eta(u, v)$ is Hölder continuous with exponent γ at $v = g^*(u)$, and also a Hölder type lower bound on its increments.

Lemma 4.1. *Let Assumptions 1–3 hold. Then Condition AA holds with* $C = c_\eta(2c_q)^\gamma \vee 1/s$. *Moreover, we have for each boundary fragment* f_g *with boundary* g, *that is,*

$$f_g(x) = \begin{cases} +1, & \text{if } x \in \{(u, v) \in [0, 1]^2 : g(u) \geqslant v\}, \\ -1, & \text{otherwise}, \end{cases} \tag{26}$$

the upper bound

$$R(f_g) - R(f^*) \leqslant 2c_\eta c_q \|g - g^*\|_{\kappa,\mu}^\kappa,$$

where $\kappa = 1 + \gamma$.

For $r \geqslant 1$, we define the class of Hölder continuous functions with exponent $1/r$,

$$\mathcal{G}_r(\text{Hölder}) := \{g : [0, 1] \to [0, 1] : |g(u) - g(\tilde{u})| \leqslant |u - \tilde{u}|^{1/r}, \forall u, \tilde{u}\} .$$

We call r the *roughness* parameter. We let $\mathcal{G}_0$ be the class of all constant functions on [0, 1].

The next lemma studies the approximation of functions in $\mathcal{G}_r$(Hölder). Later, we will see that Condition C results in a bound on the resolution level, and hence on the one-dimensional precision level of our measurements. This precision level, say δ, is defined as the smallest value such that our approximations are piecewise constant on the grid Δ^2, where $\Delta = \{k\delta : k = 0, 1, \ldots\}$. In our situation, we will have $\frac{1}{2}\sqrt{\log n/n} \leqslant \delta \leqslant \sqrt{\log n/n}$.

For $a > 0$ define $\lfloor a \rfloor$ as the largest integer less than or equal to $\lceil a \rceil$. Likewise a is the smallest integer greater than or equal to a.

Lemma 4.2. *Suppose* $g^* \in \mathcal{G}_r$(Hölder) *for some* $r \geqslant 1$. *Then for all* $\epsilon \geqslant \delta$, *there is a function* g_ϵ^* *which is constant on the intervals* $(u_{j-1}, u_j]$, $u_j = j\epsilon^r$, $j = 1, 2, \ldots, \lceil \epsilon^{-r} \rceil$, *and with values in* Δ *such that*

$$\|g^* - g_\epsilon^*\|_\infty \leqslant \epsilon + \delta.$$

Let $\{h_{j,l}\}$ be the orthonormal Haar basis of L_2([0, 1], Lebesguemeasure). So

$$h_{1,1} := 1,$$
$$h_{1,2} := 1_{[0,1/2)} - 1_{[1/2,1)},$$

and generally

$$2^{-(l-2)/2} h_{j,l} := 1_{[2(j-1)2^{-l+1},(2j-1)2^{-l+1})} - 1_{[(2j-1)2^{-l+1},(2j)2^{-l+1})}, \qquad j = 1, \ldots, 2^{l-2},\ l = 2, 3, \ldots . \tag{27}$$

We use the expansion

$$f_\alpha = \sum_{k=1}^{L} \sum_{l=1}^{L} \sum_i \sum_j \alpha_{i,j,k,l} h_{i,k} h_{j,l},$$

where $\{h_{j,l}\}$ is the one-dimensional Haar system. We take one-dimensional resolution levels L, with L the largest integer such that $2^{2(L-2)} \leqslant n/\log n$. This means we have one-dimensional measurement precision $\delta = 2^{-(L-1)} \leqslant \sqrt{\log n/n}$. As a consequence of Lemma 4.2, we thus obtain the following result:

Lemma 4.3. *Suppose Assumptions 1–3 are met. Let* $\mathcal{F} = \{f_\alpha : \|f_\alpha\|_\infty \leqslant K_0\}$, *where* $K_0 \geqslant 1$, *and let* $\delta_n = \sqrt{\log n/n}$. *Consider integers N, with* $N = JL^2$ *and where* $2 \leqslant J \leqslant \delta_n^{-r}$ *is an integer. If* $g^* \in \mathcal{G}_r$(Hölder), *we have*

$$\inf_{f_\alpha \in \mathcal{F}, N(\alpha) \leqslant N} R(f_\alpha) - R(f^*) \leqslant 2^{\kappa+1} c_q c_\eta \times \left(\left(\frac{2L^2}{N} \right)^{\kappa/r} + \delta_n^\kappa \right).$$

Furthermore, if $g^* \in \mathcal{G}_0$,

$$\inf_{f_\alpha \in \mathcal{F}, N(\alpha) = L} R(f_\alpha) - R(f^*) \leqslant 2 c_q c_\eta \delta_n^\kappa.$$

Finally, we assume the following:

Assumption 4. *For some constant* $0 < t < \frac{1}{2}$, *we have that* $t \leqslant \eta \leqslant 1 - t$.

See Remark 3.2 for a discussion of this assumption.

Theorem 4.1. *Suppose that Assumptions 1–4 hold. Let* $\mathcal{F} = \{f_\alpha : \|f_\alpha\|_\infty \leqslant K_0\}$. *Consider the standard situation, that is, the case where the constants* K_0, c_q, c_η, s, γ, r *and* t *do not depend on* n. *Let* $\kappa = 1 + \gamma$. *Then*

$$\mathbb{E} R(\hat{f}_n) - R(f^*) = O\left(\frac{\log^2 n}{n} \right)^\beta,$$

with $\beta = \kappa/(2\kappa - 1 + r)$. *If* $g^* \in \mathcal{G}_0$, *we have*

$$\mathbb{E} R(\hat{f}_n) - R(f^*) = O\left(\frac{\log^{3/2} n}{n} \right)^\beta,$$

with

$$\beta = \begin{cases} \dfrac{\kappa}{2\kappa - 1}, & \text{if } \kappa \geqslant \dfrac{3}{2}, \\ \dfrac{\kappa}{2}, & \text{if } \kappa \leqslant \dfrac{3}{2}. \end{cases}$$

Note that the coefficient β is increasing in r, so the rates are faster when r is smaller. Moreover, when $r \geqslant 1$, the coefficient β is increasing in κ, that is, the rates are then faster for larger values of κ. We conclude that the ℓ_1 penalized minimum hinge loss estimator adjusts to the values of both roughness r and margin parameter κ.

For the case $g^* \in \mathcal{G}_0$, we do not obtain the rate $(\log^{3/2} n/n)^{\kappa/(2\kappa-1)}$ for all values of κ

due to limited precision level. If it were known a priori that g^* is constant, one would only have to consider the one-dimensional problem, and a precision level of order $n/\log n$ could be taken. In that case, the rate would be of order $(\log^{3/2} n/n)^{\kappa/(2\kappa-1)}$ for all values of κ.

Remark 4.1. Note that for roughness $r \geqslant 1$, the rates in Theorem 4.1 become better as κ increases, which makes the definition of minimax rates a subtle matter. The situation is as in Scott and Nowak (2006). They present a formulation of minimax rates, but their concept of roughness is different from ours.

5. Proof of Theorems 2.1 and 2.2

5.1. Proof of Theorem 2.1

Let us write $\hat{f}_n = f_{\hat{\alpha}_n}$. Moreover, let

$$I(\alpha) := \sum_{k=1}^{m} |\alpha_k|, \qquad \alpha \in \mathbb{R}^m,$$

and

$$\nu_n(f) := \sqrt{n}(R_n(f) - R(f)), \qquad f \in \mathcal{F}.$$

Up to and including Lemma 5.6, we fix an arbitrary $\alpha^* \in \mathbb{R}^m$, with $f_{\alpha^*} \in \mathcal{F}$. The result of Theorem 2.1 then follows from taking the infimum, over all such f_{α^*}, of $\epsilon_n(f_{\alpha^*})$ defined below in (29). This is done at the very end of this section.

Set $K := 2K_0$. Let Ω^* be the set

$$\Omega^* := \left\{ \sup_{f \in \mathcal{F}} \frac{|\nu_n(f) - \nu_n(f_{\alpha^*})|}{I(\alpha - \alpha^*) + K\sqrt{\log n/n}} \leqslant \sqrt{n}\frac{\lambda_n}{2} \right\} \bigcap \left\{ \frac{\lambda_n^2}{4} \leqslant \hat{\lambda}_n^2 \leqslant 4D\lambda_n^2 \right\}. \tag{28}$$

Recall that

$$\lambda_n = c'(C_Q \vee 4)DK^2\sqrt{\frac{\log n}{n}}, \qquad \hat{\lambda}_n = c'(\hat{C}_n \vee 4)DK^2\sqrt{\frac{\log n}{n}},$$

where $4c' = c \geqslant 4c_0'$. Moreover, we take $c_0' = 216$ ($c_0 = 4c_0'$). We show in Lemmas 5.4–5.7 that, under Condition C, the set $\{\omega \notin \Omega^*\}$ has probability at most

$$\bar{c}_1 \exp(-K^2 \log n/2) + 2\exp(-2\log n).$$

Here, $\bar{c}_1$ is an appropriate universal constant. Lemma 5.1 below tells us that Conditions A and B yield, on Ω^*, the bound

$$\epsilon_n(f_{\alpha^*}) = (1 + 4\delta)\left\{ R(f_{\alpha^*}) - R(f^*) + V_n(N(\alpha^*)) + \lambda_n K\sqrt{\frac{\log n}{n}} \right\} \tag{29}$$

for the excess risk $R(\hat{f}_n) - R(f^*)$. Lemmas 5.2 and 5.3 are tools used in Lemma 5.1.

Lemma 5.1. *Assume Conditions A and B. Then, on* Ω^*,

$$R(\hat{f}_n) - R(f^*) \leqslant \epsilon_n(f_{\alpha^*}), \tag{30}$$

where $\epsilon_n(f_{\alpha^*})$ *is given in* (29).

Proof. We use similar arguments to those in Loubes and van de Geer (2002), van de Geer (2003) and Tsybakov and van de Geer (2005). Define $N^* = N(\alpha^*)$ and, for each $\alpha \in \mathbb{R}^m$,

$$I_1(\alpha) := \sum_{k:\alpha_k^* \neq 0} |\alpha_k|, \qquad I_2(\alpha) := I(\alpha) - I_1(\alpha) = \sum_{k:\alpha_k^*=0} |\alpha_k| \ .$$

Then

$$\begin{aligned} R(\hat{f}_n) - R(f_{\alpha^*}) = & -\Big(\nu_n(\hat{f}_n) - \nu_n(f_{\alpha^*})\Big)/\sqrt{n} + \hat{\lambda}_n(I(\alpha^*) - I(\hat{\alpha}_n)) \\ & + [R_n(\hat{f}_n) + \hat{\lambda}_n I(\hat{\alpha}_n)] - [R_n(f_{\alpha^*}) + \hat{\lambda}_n I(\alpha^*)] \\ \leqslant & -\Big(\nu_n(\hat{f}_n) - \nu_n(f_{\alpha^*})\Big)/\sqrt{n} + \hat{\lambda}_n\big(I(\alpha^*) - I(\hat{\alpha}_n)\big). \end{aligned}$$

The latter inequality is true because

$$R_n(\hat{f}_n) + \hat{\lambda}_n I(\hat{\alpha}_n) \leqslant R_n(f_{\alpha^*}) + \hat{\lambda}_n I(\alpha^*),$$

as $\hat{f}_n$ is the minimizer of the penalized empirical hinge loss over $\mathcal{F}$, and $f_{\alpha^*} \in \mathcal{F}$.

Thus, on Ω^*,

$$\begin{aligned} R(\hat{f}_n) - R(f_{\alpha^*}) &\leqslant \frac{\lambda_n}{2}\left(I(\hat{\alpha}_n - \alpha^*) + K\sqrt{\frac{\log n}{n}}\right) + \hat{\lambda}_n\big(I(\alpha^*) - I(\hat{\alpha}_n)\big) \\ &= \frac{\lambda_n}{2}\left(I_1(\hat{\alpha}_n - \alpha^*) + I_2(\hat{\alpha}_n) + K\sqrt{\frac{\log n}{n}}\right) + \hat{\lambda}_n\big(I_1(\alpha^*) - I_1(\hat{\alpha}_n) - I_2(\hat{\alpha}_n)\big), \end{aligned}$$

where we use the fact that $I_2(\hat{\alpha}_n - \alpha^*) = I_2(\hat{\alpha}_n)$ and $I_2(\alpha^*) = 0$. Since $\lambda_n/2 \leqslant \hat{\lambda}_n$ on Ω^*, we find on that set that

$$R(\hat{f}_n) - R(f_{\alpha^*}) \leqslant \frac{\lambda_n}{2}\left(I_1(\hat{\alpha}_n - \alpha^*) + K\sqrt{\frac{\log n}{n}}\right) + \hat{\lambda}_n\big(I_1(\alpha^*) - I_1(\hat{\alpha}_n)\big).$$

Now use the fact that $I_1(\alpha^*) - I_1(\hat{\alpha}_n) \leqslant I_1(\hat{\alpha}_n - \alpha^*)$, and that, on Ω^*, $\hat{\lambda}_n \leqslant 2\sqrt{D}\lambda_n$. Invoking the bounds $\frac{1}{2} \leqslant 1 \leqslant \sqrt{D}$, we obtain that, on Ω^*,

$$R(\hat{f}_n) - R(f_{\alpha^*}) \leqslant 3\lambda_n\sqrt{D}I_1(\hat{\alpha}\mathrm{a}_n - \alpha^*) + \lambda_n K\sqrt{\log n/n}.$$

We now use Lemma 5.2, and the triangle inequality, to arrive at

$$R(\hat{f}_n) - R(f_{\alpha^*}) \leqslant 3\lambda_n\sqrt{N^* DK}\|\hat{f}_n - f^*\|_{1,\nu}^{1/2}/\rho_\nu + 3\lambda_n\sqrt{N^* DK}\|f_{\alpha^*} - f^*\|_{1,\nu}^{1/2}/\rho_\nu + \lambda_n K\sqrt{\log n/n}. \tag{31}$$

Let us use the shorthand notation

$$\hat{d} := R(\hat{f}_n) - R(f^*), \qquad d^* := R(f_{\alpha^*}) - R(f^*).$$

Then the application of Lemma 5.3 below (with $\upsilon = 3\lambda_n\sqrt{N^* DK}/\rho_\nu$ and respectively $t = \|\hat{f}_n - f^*\|_{1,\nu}^{1/2}$ and $t = \|f_{\alpha^*} - f^*\|_{1,\nu}^{1/2}$), and Condition A, to the first two terms on the right-hand side of (31) yields

$$\hat{d} \leqslant \delta(\hat{d} + d^*) + 2\delta\left(\frac{3\lambda_n}{\rho_\nu}\frac{\sqrt{\sigma N^* DK}}{\delta}\right)^{2\kappa/(2\kappa-1)} + \lambda_n K\sqrt{\frac{\log n}{n}}.$$

Since, for $\delta \leqslant \frac{1}{2}$, the inequality $(1+\delta)/(1-\delta) \leqslant 1 + 4\delta$ holds, we have now shown that

$$\hat{d} \leqslant (1+4\delta)\left\{d^* + 2\delta\left(\frac{3\lambda_n}{\rho_\nu}\frac{\sqrt{\sigma N^* DK}}{\delta}\right)^{2\kappa/(2\kappa-1)} + \lambda_n K\sqrt{\frac{\log n}{n}}\right\}$$

$$= (1+4\delta)\left\{d^* + 2\delta^{-1/(2\kappa-1)}\left[\frac{9\sigma\lambda_n^2 N^* DK}{\rho_\nu^2}\right]^{\kappa/(2\kappa-1)} + \lambda_n K\sqrt{\frac{\log n}{n}}\right\}.$$

□

Lemma 5.2. *Assume Condition B. Let $\mathcal{J} \subset \{1, \ldots, m\}$ be some index set with cardinality $N = |\mathcal{J}|$. Then, for $\|f_\alpha\|_\infty \leqslant K$,*

$$\left(\sum_{k\in\mathcal{J}} |\alpha_k|\right)^2 \leqslant NK\|f_\alpha\|_{1,\nu}/\rho_\nu^2.$$

Proof. Clearly

$$\left(\sum_{k\in\mathcal{J}} |\alpha_k|\right)^2 \leqslant N\sum_k \alpha_k^2.$$

But

$$\sum_k \alpha_k^2 = \alpha^{\mathrm{T}}\alpha \leqslant \alpha^{\mathrm{T}}\Sigma_\nu\alpha/\rho_\nu^2 = \|f_\alpha\|_{2,\nu}^2/\rho_\nu^2 \leqslant K\|f_\alpha\|_{1,\nu}/\rho_\nu^2 .$$

□

Lemma 5.1 makes use of Lemma 5.3 below. Such inequalities are routinely used in the recent classification literature (see, for example, Tsybakov and van de Geer 2005). Lemma

5.3 is an immediate consequence of Young's inequality (see Hardy *et al.* 1988: Section 8.3), using some straightforward bounds to simplify the expressions.

Lemma 5.3. *For all $\kappa \geqslant 1$, and all positive υ, t and δ,*

$$\upsilon t \leqslant \delta t^{2\kappa}/\sigma^{\kappa} + \delta^{-1/(2\kappa-1)}(\sigma\upsilon^2)^{\kappa/(2\kappa-1)}.$$

We now will show that the set Ω^* has probability close to one. To this end, a concentration inequality will be applied. Theorem 5.1 is from Massart (2000), who improves the constants from Ledoux (1997). These authors actually assume certain measurability conditions. To avoid digressions, we will skip all measurability issues.

Theorem 5.1. *Let $Z_1, \ldots, Z_n$ be independent and identically distributed copies of a random variable $Z \in \mathcal{Z}$. Let Γ be a class of real-valued functions on $\mathcal{Z}$ satisfying $\sup_z|\gamma(z)| \leqslant K$ for all $\gamma \in \Gamma$. Define*

$$\mathbf{Z} := \sup_{\gamma\in\Gamma}\left|\frac{1}{n}\sum_{i=1}^{n}\{\gamma(Z_i) - \mathbb{E}\gamma(Z_i)\}\right| \tag{32}$$

and

$$\tau^2 := \sup_{\gamma\in\Gamma}\operatorname{var}(\gamma(Z)). \tag{33}$$

Then, for any positive z,

$$\mathbb{P}\Big(\mathbf{Z} \geqslant 2\mathbb{E}\mathbf{Z} + \tau\sqrt{8z/n} + 69Kz/(2n)\Big) \leqslant \exp(-z). \tag{34}$$

Lemma 5.4. *Define $\mathcal{F}_M := \{f_\alpha \in \mathcal{F} : I(\alpha - \alpha^*) \leqslant M, \|f_\alpha - f_{\alpha^*}\|_\infty \leqslant K\}$, and*

$$\mathbf{Z}_M := \sup_{f_\alpha\in\mathcal{F}_M}|\nu_n(f_\alpha) - \nu_n(f_{\alpha^*})|/\sqrt{n}.$$

Then, for all M satisfying $C_Q M \geqslant K\sqrt{\log n/n}$ (where C_Q is given in (1)*), we have*

$$\mathbb{P}\Big(\mathbf{Z}_M \geqslant 2\mathbb{E}\mathbf{Z}_M + 36K^2C_QM\sqrt{\log n/n}\Big) \leqslant \exp(-(C_Q^2M^2 \vee K^2)\log n).$$

Proof. In Theorem 5.1 we take

$$\Gamma = \{\gamma_\alpha : f_\alpha \in \mathcal{F}_M\},$$

where

$$\gamma_\alpha(x, y) = l(yf_\alpha(x)) - l(yf_{\alpha^*}(x)),$$

and where $l(z) = (1 - z)_+$ is the hinge function. Since l is Lipschitz, we have

$$|\gamma_\alpha(x, y)| \leqslant |f_\alpha(x) - f_{\alpha^*}(x)|.$$

Note first that this implies $\mathbf{Z}_M \leqslant K$, so that it suffices to consider values M with $C_QM \leqslant K\sqrt{n/\log n}$. The Lipschitz property also implies that $\tau^2 \leqslant \sup_{f_\alpha\in\mathcal{F}_M}\|f_\alpha - f_{\alpha^*}\|_{2,Q}^2$. So $\tau \leqslant C_QM \wedge K := \tau_1$. We now take $z = (C_Q^2M^2 \vee K^2)\log n$.

Then, for $K \leqslant C_Q M \leqslant K\sqrt{n/\log n}$,

$$\begin{aligned}\tau_1\sqrt{8z/n} + 69Kz/(2n) &= KC_Q M\sqrt{8\log n/n} + 69KC_Q^2 M^2 \log n/(2n)\\ &\leqslant 3KC_Q M\sqrt{\log n/n} + \frac{69}{2}K^2 C_Q M\sqrt{\log n/n}\\ &\leqslant 36K^2 C_Q M\sqrt{\log n/n},\end{aligned}$$

where we use the fact that $K \geqslant 2$. Moreover, for $K\sqrt{\log n/n} \leqslant C_Q M \leqslant K$,

$$\begin{aligned}\tau_1\sqrt{8z/n} + 69Kz/(2n) &= KC_Q M\sqrt{8\log n/n} + 69KK^2 \log n/(2n)\\ &\leqslant 3KC_Q M\sqrt{\log n/n} + \frac{69}{2}K^2 C_Q M\sqrt{\log n/n}\\ &\leqslant 36K^2 C_Q M\sqrt{\log n/n}.\end{aligned}$$

The result thus follows from Theorem 5.1. □

Lemma 5.5. *Suppose Condition C is met. For* $\mathbf{Z}_M$ *defined in Lemma 5.4, it holds that*

$$\mathbb{E}\mathbf{Z}_M \leqslant 36(C_Q \vee 4)DM\sqrt{\log n/n}. \tag{35}$$

Proof. This follows from similar arguments to those in van de Geer (2003), using the fact that the function $z \mapsto l(z) = (1-z)_+$, $z \in \mathbb{R}$, is Lipschitz. Let us briefly summarize these arguments. Let $\epsilon_1, \ldots, \epsilon_n$ be a Rademacher sequence independent of $(X_1, Y_1), \ldots, (X_n, Y_n)$. By symmetrization and the contraction inequality (see Ledoux and Talagrand 1991), we find that

$$\begin{aligned}\mathbb{E}\mathbf{Z}_M &\leqslant 4\mathbb{E}\left(\sup_{f_\alpha \in \mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f_\alpha(X_i) - f_{\alpha^*}(X_i))\right|\right)\\ &\leqslant 4M\mathbb{E}\left(\max_{k=1,\ldots,m}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\psi_k(X_i)\right|\right).\end{aligned}$$

By Bernstein's inequality (see, for example, Shorack and Wellner 1986: 855), we know that, for any $z > 0$,

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n \epsilon_i\psi_k(X_i)\right| \geqslant z\right) \leqslant 2\exp\left(-\frac{nz^2}{2z\|\psi_k\|_\infty/3 + 2\|\psi_k\|_{2,Q}^2}\right).$$

Use Condition C, which says that for all k, $\|\psi_k\|_\infty \leqslant \sqrt{n/\log n}$. Moreover, $\|\psi_k\|_{2,Q} \leqslant C_Q$ for all k. We find for all $z \geqslant 1$, using $m \leqslant n^D$ and $D \geqslant 1$,

$$\begin{aligned}
&\mathbb{P}\left(\max_{k=1,\dots,m}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\psi_k(X_i)\right| \geqslant 3(C_Q\vee 4)Dz\sqrt{\frac{\log n}{n}}\right)\\
&\quad\leqslant 2m\exp\left(-\frac{9(C_Q\vee 4)^2D^2z\log n}{2(C_Q\vee 4)D+2C_Q^2}\right)\\
&\quad\leqslant 2m\exp(-9Dz\log n/4)\\
&\quad\leqslant 2\exp(-5Dz\log n/4).
\end{aligned}$$

Now, for any positive random variable U and any positive t,

$$\mathbb{E}(U)=\int_0^\infty \mathbb{P}(U\geqslant z)\mathrm{d}z \leqslant t\left(1+\int_1^\infty \mathbb{P}(U\geqslant tz)\mathrm{d}z\right).$$

Apply this with

$$U=\max_{k=1,\dots,m}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\psi_k(X_i)\right|.$$

It follows that

$$\mathbb{E}\mathbf{Z}_M \leqslant 3(1+2)4M(C_Q\vee 4)D\sqrt{\frac{\log n}{n}}=36M(C_Q\vee 4)D\sqrt{\frac{\log n}{n}},$$

where we use the bound

$$\frac{\exp(-5D\log n/4)}{5D\log n/4}\leqslant 1,$$

because $D\geqslant 1$ and $n\geqslant 8$. □

Next, we show that, for $\lambda_n\geqslant 216(C_Q\vee 4)DK^2\sqrt{\log n/n}$, the set

$$\left\{|\nu_n(\hat{f}_n)-\nu_n(f_{\alpha^*})|\leqslant\sqrt{n}\frac{\lambda_n}{2}\left(I(\hat{\alpha}_n-\alpha^*)+K\sqrt{\frac{\log n}{n}}\right)\right\}$$

has probability at least $1-\bar{c}_1\exp(-K^2\log n/2)$.

Lemma 5.6. *Suppose Condition C is met. We have, for a universal constant $\bar{c}_1$,*

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f_\alpha\in\mathcal{F},\|f_\alpha-f_{\alpha^*}\|_\infty\leqslant K}\left|\frac{\nu_n(f_\alpha)-\nu_n(f_{\alpha^*})}{I(\alpha-\alpha^*)+K\sqrt{\log n/n}}\right|>108(C_Q\vee 4)DK^2\sqrt{\log n}\right)\\
&\quad\leqslant\bar{c}_1\exp\left(-\frac{K^2\log n}{2}\right).
\end{aligned}\tag{36}$$

Proof. This follows from the peeling device, which is designed to establish bounds for

weighted empirical process, as discussed in van de Geer (2000: Section 5.3). We split $\mathbb{R}^m$ into the sets

$$S_1 = \{\alpha : C_Q I(\alpha - \alpha^*) \leqslant K\sqrt{\log n/n}\},$$

$$S_2 = \{\alpha : K\sqrt{\log n/n} < C_Q I(\alpha - \alpha^*) \leqslant K\}$$

$$\subseteq \bigcup_{j=0}^{j_0} \{\alpha : 2^{-(j+1)}K < C_Q I(\alpha - \alpha^*) \leqslant 2^{-j}K\},$$

with $2^{-j_0} < \sqrt{\log n/n}$, and

$$S_3 = \{\alpha : C_Q I(\alpha - \alpha^*) \geqslant K\}$$

$$= \bigcup_{j=1}^{\infty} \{\alpha : 2^{j-1}K < C_Q I(\alpha - \alpha^*) \leqslant 2^j K\}.$$

The combination of Lemma 5.4 and 5.5, and invoking $K \geqslant 2\mathrm{r}$, yields that for $C_Q M \geqslant K\sqrt{\log n/n}$,

$$\mathbb{P}\left(\mathbf{Z}_M \geqslant 54(C_Q \vee 4)DMK^2\sqrt{\frac{\log n}{n}}\right) \leqslant \exp(-(C_Q^2 M^2 \vee K^2)\log n) \ . \tag{37}$$

We find on the set S_1,

$$\mathbb{P}\left(\sup_{f_\alpha \in \mathcal{F}, \alpha \in S_1} \left| \frac{\nu_n(f_\alpha) - \nu_n(f_{\alpha^*})}{I(\alpha - \alpha^*) + K\sqrt{\log n/n}} \right| \geqslant 108(C_Q \vee 4)DK^2\sqrt{\log n}\right)$$

$$\leqslant \mathbb{P}\left(\sup_{f_\alpha \in \mathcal{F}, \alpha \in S_1} |\nu_n(f_\alpha) - \nu_n(f_{\alpha^*})| \geqslant 108(K\sqrt{\log n/n})(C_Q \vee 4)DK^2\sqrt{\log n}\right)$$

$$\leqslant \exp(-K^2 \log n).$$

Next, we consider the set S_2. Take j_0 as the smallest integer such that $2^{-j_0} < \sqrt{\log n/n}$. Then, from (37),

$$\mathbb{P}\left(\sup_{f_\alpha \in \mathcal{F}, \alpha \in S_2} \left| \frac{\nu_n(f_\alpha) - \nu_n(f_{\alpha^*})}{I(\alpha - \alpha^*) + K\sqrt{\log n/n}} \right| \geqslant 108(C_Q \vee 4)DK^2\sqrt{\log n}\right)$$

$$\leqslant \sum_{j=0}^{j_0} \mathbb{P}\left(\sup_{f_\alpha \in \mathcal{F}, C_Q I(\alpha - \alpha^*) \leqslant 2^{-j}K} \left|\nu_n(f_\alpha) - \nu_n(f_{\alpha^*})\right| \geqslant 54(2^{-j}K)(C_Q \vee 4)DK^2\sqrt{\log n}\right)$$

$$\leqslant \log n \exp(-K^2 \log n),$$

as, for $n \geqslant 8$, $j_0 + 1 \leqslant \log n$.

Finally, we consider the set S_3. We find

$$\mathbb{P}\left(\sup_{f_\alpha\in\mathcal{F},\alpha\in S_3}\left|\frac{\nu_n(f_\alpha)-\nu_n(f_{\alpha^*})}{I(\alpha-\alpha^*)+K\sqrt{\log n/n}}\right|\geqslant 108(C_Q\vee 4)DK^2\sqrt{\log n}\right)$$

$$\leqslant\sum_{j=1}^{\infty}\mathbb{P}\left(\sup_{f_\alpha\in\mathcal{F},C_QI(\alpha-\alpha^*)\leqslant 2^jK}|\nu_n(f_\alpha)-\nu_n(f_{\alpha^*})|\geqslant 54(2^jK)(C_Q\vee 4)DK^2\sqrt{\log n}\right)$$

$$\leqslant\sum_{j=1}^{\infty}\exp(-2^{2j}K^2\log n).$$

We conclude that

$$\mathbb{P}\left(\sup_{f_\alpha\in\mathcal{F}}\left|\frac{\nu_n(f_\alpha)-\nu_n(f_{\alpha^*})}{I(\alpha-\alpha^*)+K\sqrt{\log n/n}}\right|\geqslant 108(C_Q\vee 4)DK^2\sqrt{\log n}\right)$$

$$\leqslant\exp(-K^2\log n)+\log n\exp(-K^2\log n)+\sum_{j=1}^{\infty}\exp(-2^{2j}K^2\log n)$$

$$\leqslant\bar{c}_1\exp\left(-\frac{K^2\log n}{2}\right),$$

for a universal constant $\bar{c}_1$. □

Lemma 5.7. *Suppose Condition C holds. Then*

$$\mathbb{P}\left(\frac{\lambda_n^2}{4}\leqslant\hat{\lambda}_n^2\leqslant 4D\lambda_n^2\right)\geqslant 1-2\exp(-2\log n). \tag{38}$$

Proof. Recall the definitions

$$\lambda_n=c'(C_Q\vee 4)DK^2\sqrt{\frac{\log n}{n}},\qquad \hat{\lambda}_n=c'(\hat{C}_n\vee 4)DK^2\sqrt{\frac{\log n}{n}},$$

and

$$C_Q^2=\max_{1\leqslant k\leqslant m}\|\psi_k\|_{2,Q}^2,\qquad \hat{C}_{\mathrm{n}}^2=\max_{1\leqslant k\leqslant m}\frac{1}{n}\sum_{i=1}^{n}\psi_k^2(X_i).$$

We first bound the probability of the set $\{\hat{\lambda}_n^2<\lambda_n^2/4\}$. We consider two cases: $C_Q\leqslant 4$ and $C_Q>4$. If $C_Q\leqslant 4$, we have

$$\hat{C}_n\vee 4=\begin{cases}\hat{C}_n\geqslant C_Q\vee 4, & \text{if } \hat{C}_n>4,\\ 4=C_Q\vee 4, & \text{if } \hat{C}_n\leqslant 4.\end{cases}$$

In other words, if $C_Q\leqslant 4$, we have $\hat{\lambda}_n\geqslant\lambda_n$, and so the set $\{\hat{\lambda}_{\mathrm{n}}^2<\lambda_{\mathrm{n}}^2/4\}$ has probability zero.

Now, let $\psi_{\max}$ a base function for which the maximum $L_2(Q)$ norm is attained. Then, clearly,

$$\hat{C}_n^2 \geqslant \|\psi_{\max}\|_{2,Q_n}^2.$$

From Bernstein's inequality, we now establish that

$$\begin{aligned}\mathbb{P}(\hat{C}_n^2 < C_Q^2/4) &\leqslant \mathbb{P}(\|\psi_{\max}\|_{2,Q_n}^2 - \|\psi_{\max}\|_{2,Q}^2 < -3C_Q^2/4)\\ &\leqslant \exp\left[-\frac{n9C_Q^4/16}{C_Q^2\|\psi_{\max}^2\|_\infty/2 + 2C_Q^2\|\psi_{\max}^2\|_{2,Q}^2}\right]\\ &\leqslant \exp[-9C_Q^2\log n/40] \leqslant \exp[-C_Q^2\log n/8],\end{aligned}$$

since $\|\psi_{\max}^2\|_{2,Q} = C_Q^2$ and, by Condition C, $\|\psi_{\max}^2\|_\infty \leqslant n/\log n$. With $C_Q > 4$, this gives

$$\mathbb{P}(\hat{C}_n^2 < C_Q^2/4) \leqslant \exp(-2\log n).$$

Next, we bound the probability of $\{\hat{\lambda}_n^2 > 4D\lambda_n^2\}$. We consider the cases $\hat{C}_n \leqslant 4$ and $\hat{C}_n > 4$. For $\hat{C}_n \leqslant 4$, one has

$$(\hat{C}_n \vee 4)^2 = 4^2 \leqslant 4D(C_Q \vee 4)^2,$$

so that $\{\hat{\lambda}_n^2 \leqslant 4D\lambda_n^2\}$ trivially holds. Turning to the case $\hat{C}_n > 4$, note first that, again by Bernstein's inequality,

$$\begin{aligned}\mathbb{P}\left(\max_{1\leqslant k\leqslant m}\|\psi_k\|_{2,Q_n}^2 - \|\psi_k\|_{2,Q}^2 > 3D(C_Q\vee 4)^2\right) &\leqslant m\exp\left[-\frac{9D^2(C_Q\vee 4)^4\log n}{2D(C_Q\vee 4)^2 + 2C_Q^2}\right]\\ &\leqslant m\exp[-9D(C_Q\vee 4)^2\log n/4]\\ &\leqslant \exp[-5D(C_Q\vee 4)^2\log n/4] \leqslant \exp[-2\log n].\end{aligned}$$

But clearly, if

$$\max_{1\leqslant k\leqslant m}\|\psi_k\|_{2,Q_n}^2 - \|\psi_k\|_{2,Q}^2 \leqslant 3D(C_Q\vee 4)^2,$$

one has, for $\hat{C}_n > 4$,

$$(\hat{C}_n\vee 4)^2 = \hat{C}_n^2 \leqslant 3D(C_Q\vee 4)^2 + C_Q^2 \leqslant 4D(C_Q\vee 4)^2.$$

□

To conclude the proof of Theorem 2.1, we observe that, for any $z \geqslant 0$,

$$\mathbb{P}\Big(R(\hat{f}_n) - R(f^*) > z\Big) = \lim_{z_t\downarrow z}\mathbb{P}\Big(R(\hat{f}_n) - R(f^*) > z_t\Big),$$

because a distribution function is right-continuous. Let ϵ_n be defined as in (12):

$$\epsilon_n = \inf\big\{\epsilon_n(f_{\alpha^*}) : \alpha^* \in \mathbb{R}^m, f_{\alpha^*} \in \mathcal{F}\big\}.$$

We may then write

$$\epsilon_n = \lim_{t\to\infty} \epsilon_{n,t},$$

for a sequence $\{\epsilon_{n,t}\}_{t=1}^{\infty}$ with

$$\epsilon_{n,t} = \epsilon_n(f_{\alpha_t^*}),$$

for some $\alpha_t^* \in \mathbb{R}^m$, $f_{\alpha_t^*} \in \mathcal{F}$, $t = 1, 2, \ldots$. Therefore, by Lemmas 5.1–5.7.

$$\begin{aligned}\mathbb{P}\Big(R(\hat{f}_n) - R(f^*) > \epsilon_n\Big) &= \lim_{t\to\infty} \mathbb{P}(R(\hat{f}_n) - R(f^*) > \epsilon_{n,t}) \\ &\leqslant c_1 \exp(-2\log n),\end{aligned}$$

where $c_1 = \bar{c}_1 + 2$.

5.2. Proof of Theorem 2.2

For simplicity, let us assume that the infimum in

$$\inf\{R(f_\alpha) - R(f^*) + V_n(N(\alpha)) : f_\alpha \in \mathcal{F}\}$$

is attained for some $f_{\alpha^*} \in \mathcal{F}$:

$$f_{\alpha^*} := \arg\min\{R(f_\alpha) - R(f^*) + V_n(N(\alpha)) : f_\alpha \in \mathcal{F}\}.$$

Define

$$t_n := \frac{K_n\|\hat{f}_n - f_{\alpha^*}\|_{1,\nu}}{2 + K_n\|\hat{f}_n - f_{\alpha^*}\|_{1,\nu}}.$$

Let

$$\tilde{f}_n := (1 - t_n)\hat{f}_n + t_n f_{\alpha^*}.$$

Then $\tilde{f}_n \in \mathcal{F}$ because $\mathcal{F}$ is convex. Moreover,

$$\|\tilde{f}_n - f_{\alpha^*}\|_\infty = \frac{2\|\hat{f}_n - f_{\alpha^*}\|_\infty}{2 + K_n\|\hat{f}_n - f_{\alpha^*}\|_{1,\nu}} \leqslant \frac{2\|\hat{f}_n - f_{\alpha^*}\|_\infty}{K_n\|\hat{f}_n - f_{\alpha^*}\|_{1,\nu}} \leqslant 2.$$

Observe also that by the convexity of the hinge loss and of the ℓ_1 norm, for $\tilde{\alpha}_n = (1 - t_n)\hat{\alpha}_n + t_n\alpha^*$,

$$R_n(\tilde{f}_n) - R_n(f_{\alpha^*}) + \hat{\lambda}_n(I(\tilde{\alpha}_n) - I(\alpha^*)) \leqslant (1 - t_n)[R_n(\hat{f}_n) - R_n(f_{\alpha^*}) + \hat{\lambda}_n(I(\hat{\alpha}_n) - I(\alpha^*))] \leqslant 0.$$

Let Ω^* be defined as in (28), but with $\mathcal{F}$ replaced by $\mathcal{F} := \mathcal{F} \cap \{\|f - f_{\alpha^*}\|_\infty \leqslant 2\}$. Then, by the same arguments as in the proof of Theorem 2.1, with $\hat{f}_n$ now replaced by $\tilde{f}_n$, and with $K = 2$, we see that, on Ω^*,

$$R(\tilde{f}_n) - R(f^*) \leqslant \epsilon_n.$$

Next, Condition A implies

$$\|\tilde{f}_n - f^*\|_{1,\nu}^\kappa \leqslant \sigma^\kappa(R(\tilde{f}_n) - R(f^*)).$$

On the other hand, by the triangle inequality and again Condition A,

$$\|\tilde{f}_n - f^*\|_{1,\nu} \geq \|\tilde{f}_n - f_{\alpha^*}\|_{1,\nu} - \|f^*_\alpha - f^*\|_{1,\nu} \geq (1 - t_n)\|\hat{f}_n - f_{\alpha^*}\|_{1,\nu} - \sigma\epsilon_n^{1/\kappa},$$

because $R(f_{\alpha^*}) - R(f^*) \leq (1 + 4\delta)(R(f_{\alpha^*}) - R(f^*)) \leq \epsilon_n$. So

$$R(\tilde{f}_n) - R(f^*) \leq \epsilon_n$$

implies

$$(1 - t_n)\|\hat{f}_n - f_{\alpha^*}\|_{1,\nu} \leq 2\sigma\epsilon_n^{1/\kappa},$$

or

$$\|\hat{f}_n - f_{\alpha^*}\|_{1,\nu} \leq 2\sigma\epsilon_n^{1/\kappa} + K_n\sigma\epsilon_n^{1/\kappa}\|\hat{f}_n - f_{\alpha^*}\|_{1,\nu},$$

or, since $2K_n\sigma\epsilon_n^{1/\kappa} \leq 1$,

$$\|\hat{f}_n - f_{\alpha^*}\|_{1,\nu} \leq 4\sigma\epsilon_n^{1/\kappa}.$$

But then, using once again the fact that $2K_n\sigma\epsilon_n^{1/\kappa} \leq 1$,

$$\|\hat{f}_n - f_{\alpha^*}\|_\infty \leq 2.$$

In other words, on Ω^*, we have that $\hat{f}_n \in \mathcal{F}$.

But this means that on Ω^*, we can apply the arguments of Theorem 2.1, to arrive at

$$R(\hat{f}_n) - R(f_{\alpha^*}) \leq \epsilon_n.$$

Since $\mathbf{P}(\Omega^*) \geq 1 - c_1/n^2$, this completes the proof. □

6. Proof of the results in Section 4

Proof of Lemma 4.1. Consider, for some $z > 0$, the set $|2\eta - 1| \leq z$. Since $|2\eta(u, v) - 1| \geq |v - g^*(u)|^\gamma/c_\eta$, we have

$$\{|2\eta - 1| \leq z\} \subset \{(u, v)\colon |v - g^*(u)| \leq (c_\eta z)^{1/\gamma}\}.$$

It follows that when $(c_\eta z)^{1/\gamma} \leq s$,

$$\begin{aligned} Q(\{|2\eta - 1| \leq z\}) &\leq Q(\{(u, v) : |v - g^*(u)| \leq (c_\eta z)^{1/\gamma}\}) \\ &\leq c_q\mu(\{(u, v) : |v - g^*(u)| \leq (c_\eta z)^{1/\gamma}\}) = 2c_q(c_\eta z)^{1/\gamma}. \end{aligned}$$

So Condition AA holds with $C = c_\eta(2c_q)^\gamma \vee 1/s$.

By Remark 3.1, since f_g takes values in $\{\pm 1\}$, we have

$$R(f_g) - R(f^*) = 2\int_{f_g \neq f^*} |2\eta - 1|\, \mathrm{d}Q.$$

Hence,

$$R(f_g) - R(f^*) = 2\int\int_{g(u)\wedge g^*(u)}^{g(u)\vee g^*(u)} |2\eta(u, v) - 1| q(u, v)\mathrm{d}v\,\mathrm{d}u$$

$$\leqslant 2c_\eta c_q \int\int_{g(u)\wedge g^*(u)}^{g(u)\vee g^*(u)} |v - g^*(u)|^\gamma\,\mathrm{d}u$$

$$= \frac{2c_\eta c_q}{\kappa}\|g - g^*\|_{\kappa,\mu}^\kappa \leqslant 2c_\eta c_q \|g - g^*\|_{\kappa,\mu}^\kappa.$$

□

Proof of Lemma 4.2. Let

$$g_\epsilon^*(u) := \lfloor g^*(u_j)/\delta \rfloor \delta, \qquad u_{j-1} < u \leqslant u_j.$$

Then

$$|g_\epsilon^*(u) - g^*(u)| \leqslant \delta + |u_j - u|^{1/r} \leqslant \delta + \epsilon.$$

□

Proof of Lemma 4.3. Let $g^* \in \mathcal{G}_r$(Hölder). Consider the integer l such that $2^{l-1} \leqslant J \leqslant 2^l$. Take $\epsilon = 2^{-(l-1)/r}$. Then $\lceil \epsilon^{-r} \rceil = \epsilon^{-r} \leqslant J$. Moreover, $\epsilon \leqslant (J/2)^{-1/r}$. The function g_ϵ^* is piecewise constant on at most J intervals. We note that for the one-dimensional expansion in the Haar basis, of the indicator function of a half-interval $\mathrm{l}_{[a,1]}$, with $a \in \Delta$, we need no more than L non-zero coefficients. So we need at most $L^2 J$ non-zero coefficents to expand f_ϵ^*. Here, f_ϵ^* is the boundary fragment with boundary g_ϵ^*. Hence, by Lemma 4.2,

$$\|g_\epsilon^* - g^*\|_\infty \leqslant (J/2)^{-1/r} + \delta_n.$$

Note also that f_ϵ^* takes only the values ± 1, so $f_\epsilon^* \in \mathcal{F}$.

Finally, if $g^* \in \mathcal{G}_0$, we consider the function $g_{\delta_n}^* := g^*/\delta_n\delta_n$. We clearly need no more than L coefficients to expand the boundary fragment $f_{\delta_n}^*$ corresponding to $g_{\delta_n}^*$.

The proof is completed by applying Lemma 4.1 and the inequality $(a + b)^\kappa \leqslant 2^\kappa(a^\kappa + b^\kappa)$, $a, b > 0$. □

Proof of Theorem 4.1. By Lemma 4.1, Condition A holds with $\kappa = 1 + \gamma$ and with $\mathrm{d}\nu = \eta(1 - \eta)\mathrm{d}Q$.

Now the base functions $\{\psi_k\}$ have $L_2(\mu)$ norm equal to one, and are orthogonal in $L_2(\mu)$. So, by Assumptions 1 and 4, we have that $\mathrm{d}\nu = \eta(1 - \eta)\mathrm{d}Q \geqslant (s^2/c_q)\mathrm{d}\mu$. Therefore, we know that the smallest eigenvalue ρ_ν^2 of Σ_ν satisfies $\rho_\nu^2 \geqslant s^2/c_q$. Thus, Condition B is met as well.

Condition C is met, since $2^{2(L-2)} \leqslant \log n/n$ implies

$$\|\psi_k\|_\infty \leqslant \sqrt{n/\log n}, \qquad \forall k.$$

The result now follows from Theorem 2.1. To see this, we invoke Lemma 4.3.

When $g^* \in \mathcal{G}_r$(Hölder), we let

$$N := \left\lfloor \left(\frac{n}{\log^2} \right)^{r/(2\kappa+r-1)} \right\rfloor L^2. \tag{39}$$

Then

$$J := \frac{N}{L^2} \leqslant \left(\frac{n}{\log^2 n} \right)^{r/(2\kappa+r-1)} \leqslant \left(\frac{n}{\log n} \right)^{r/(2\kappa+r-1)} \leqslant \left(\frac{n}{\log n} \right)^{r/2} = \delta_n^{-r}.$$

For the estimation error, we now have

$$V_n(N) = O\left(\frac{N \log n}{n} \right)^{\kappa/(2\kappa-1)} = O\left(\frac{\log^2 n}{n} \right)^{\kappa/(2\kappa+r-1)}.$$

In view of Lemma 4.3,

$$\inf_{N(\alpha)=N} R(f_\alpha) - R(f^*) = O\left(\frac{L^2}{N} \right)^{\kappa/r} = O\left(\frac{\log^2 n}{n} \right)^{\kappa/(2\kappa+r-1)}.$$

When $g^* \in \mathcal{G}_0$, the result immediately follows from Theorem 2.1 by taking $N := L$ and applying Lemma 4.3. □

Acknowledgements

The research for this paper was supported in part by Netherlands Organization for Scientific Research (NWO) grant 613.000.218.

References

Audibert, J.-Y. (2004) Classification under polynomial entropy and margin assumptions and randomized estimators. Preprint PMA-908, Laboratoire de Probabilités et Modèles Aléatoires. www.proba.jussieu.fr/mathdoc/textes/PMA-908.pdf

Bartlett, P.L., Jordan, M.I. and McAuliffe, J.D. (2006) Convexity, classification and risk bounds. *J. Amer. Statist. Assoc.*, **101**, 138–156.

Blanchard, G., Lugosi, G. and Vayatis, N. (2003) On the rate of convergence of regularized boosting classifiers. *J. Mach. Learn. Res.*, **4**, 861–894.

Blanchard, G., Bousquet, O. and Massart, P. (2004) Statistical performance of support vector machines. Manuscript. http://ida.first.fraunhofer.de/∼blanchard/publi/index.html

Boser, B., Guyon, I. and Vapnik, V.N. (1992) A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pp. 142–152. New York: Association for Computing Machine.

Candès, E.J. and Donoho, D.L. (2004) New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Comm. Pure Appl. Math.*, **57**, 219–266.

Donoho, D.L. (1995) Denoising via soft-thresholding. *IEEE Trans. Inform. Theory*, **41**, 613–627.

Donoho, D.L. (1999) Wedgelets: nearly minimax estimation of edges. *Ann. Statist.*, **27**, 859–897.

Donoho, D.L. (2004a) For most large underdetermined systems of equations, the minimal ℓ^1-norm

near-solution approximates the sparsest near-solution. Technical report, Stanford University. www-stat.stanford.edu/∼donoho/Reports/2004/l1l0approx.pdf

Donoho, D.L. (2004b) For most large underdetermined systems of linear equations, the minimal ℓ^1-norm solution is also the sparsest solution. Technical report, Stanford University. www-stat.stanford.edu/∼donoho/Reports/2004/l1l0EquivCorrected.pdf

Hardy, G.H., Littlewood, J.E. and Pólya, G. (1988) *Inequalities*, 2nd edn. Cambridge: Cambridge University Press.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. New York: Springer-Verlag.

Koltchinskii, V. (2001) Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory*, **47**, 1902–1914.

Koltchinskii, V. (2006) Local Rademacher complexities and oracle inequalities in risk minimization. To appear in *Ann. Statist.*, **34**(6).

Koltchinskii, V. and Panchenko, D. (2002) Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.*, **30**, 1–50.

Ledoux, M. (1997) On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Statist.*, **1**, 63–87.

Ledoux, M. and Talagrand, M. (1991) *Probability in Banach Spaces: Isoperimetry and Processes*. New York: Springer-Verlag.

Lin, Y. (2002) Support vector machines and the Bayes rule in classification. *Data Min. Knowledge Discovery*, **6**, 259–275.

Loubes, J.-M. and van de Geer, S. (2002) Adaptive estimation in regression, using soft thresholding type penalties. *Statist. Neerlandica*, **56**, 453–478.

Lugosi, G. and Wegkamp, M. (2004) Complexity regularization via localized random penalties. *Ann. Statist.*, **32**, 1679–1697.

Mammen, E. and Tsybakov, A.B. (1999) Smooth discrimination analysis. *Ann. Statist.*, **27**, 1808–1829.

Massart, P. (2000) About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, **28**, 863–884.

Schölkopf, B. and Smola, A. (2002) *Learning with Kernels*. Cambridge, MA: MIT Press.

Scott, C. and Nowak, R. (2006) Minimax-optimal classification with dyadic decision trees. *IEEE Trans. Inform. Theory*, **52**, 1335–1353.

Shorack, G.R. and Wellner, J.A. (1986) *Empirical Processes with Applications to Statistics*. New York: Wiley.

Steinwart, I. and Scovel, S. (2005) Fast rates for support vector machines using Gaussian kernels. Technical report LA-UR 04-8796, Los Alamos National Laboratory. www.c3.lanl.gov/ml/pubs/2004_fastratesa/paper.pdf

Tarigan, B. and van de Geer, S.A. (2004) Adaptivity of support vector machines with ℓ_1 penalty. Technical report MI 2004-14, University of Leiden. http://www.stat.math.ethz.ch/∼geer/reports.html

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B*, **58**, 267–288.

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

Tsybakov, A.B. and van de Geer, S.A. (2005) Square root penalty: adaptation to the margin in classification and in edge estimation. *Ann. Statist.*, **33**, 1203–1224.

van de Geer, S. (2000) *Empirical Processes in M-Estimation*. Cambridge: Cambridge University Press.

van de Geer, S. (2003) Adaptive quantile regression. In M.G. Akritas and D.N. Politis (eds), *Recent Advances and Trends in Nonparametric Statistics*, pp. 235–250. Amsterdam: Elsevier.

Vapnik, V.N. (1995) *The Nature of Statistical Learning Theory.* New York: Springer-Verlag.
Vapnik, V.N. (1998) *Statistical Learning Theory.* New York: Wiley.
Zhang, T. (2004) Statistical behaviour and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, **32**, 56–84.
Zhu, J., Rosset, S., Hastie, T. and Tibshirani, R. (2003) 1-norm support vector machines. *Neural Inform. Process. Syst.*, **16**.

Received February 2005 and revised April 2006