Adaptivity of Support Vector Machines with ℓ_1 Penalty

B. Tarigan and S.A. van de Geer
Mathematical Institute, University of Leiden

Abstract

We consider the problem of adaptation to model complexity and noise level, when using support vector machine loss in binary classification. We show that with ℓ_1 complexity regularization, adaptive rates can be obtained. [1]

1 Introduction

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

We regard (X, Y) as random variables with distribution P, and denote the distribution of X by Q. Moreover, we write the regression of Y on X as

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

Our aim is to find a classifier which has small probability of misclassification, or *prediction error*. The prediction error of a classifier f is

$$P(Yf(X) < 0).$$

Bayes rule is the classifier

$$f^* = \begin{cases} +1 & \text{if } \eta \geq 1/2 \\ -1 & \text{if } \eta < 1/2 \end{cases} \ .$$

It is easy to see that the prediction error is the smallest when using Bayes rule. The function η is not assumed to be known. To estimate Bayes rule we take a sample from P. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be observed i.i.d. copies of (X, Y). These observations are called the training set. We will assume $n \geq 8$ to avoid nonsense expressions later on. (In fact, we have the large n situation in mind.)

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 classification errors in the sample (see Vapnik (1995, 1998)). However, if $\mathcal{F}$ is a rich set, this classifier will be hard to compute. We will indeed consider a very high-dimensional class $\mathcal{F}$ in this paper. We use *support vector machine* loss instead of the number of misclassifications, to overcome computational problems. Moreover, we add an ℓ_1 penalty to the loss function, as computationally feasible complexity regularization method. We show that this procedure yields adaptive estimators.

Support vector machines (SVM's) have been introduced by Boser, Guyon and Vapnik (1992). An important book on SVM's is Schölkopf and Smola (2002). The empirical SVM loss function is

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ \ ,$$

[1] *AMS 1991 subject classifications.* Primary 62G07, secondary 62G08, 62H30, 68T10.
Key words and phrases. Binary classification, adaptation, margin, penalized classification rule, sparsity.
Running head: Adaptive support vector machines.

where z_+ denotes the positive part of $z \in \mathbf{R}$. This, we consider the hinge loss function. Define the theoretical SVM loss

$$R(f) = \mathbb{E}R_n(f) = \int (1 - yf(x))_+ dP(x, y) \ .$$

One may verify that SVM loss is consistent, in the sense that Bayes rule f^* satisfies (see Lin (2002))

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

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 $\mathcal{X}$, with

$$\int \psi_k^2 dQ \leq 1, \ k = 1, \ldots, m \ .$$

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

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

We will assume that the dimension m satisfies $m \leq n$, and in fact (see Subsection 3.2) essentially, $m \leq n/\log n$. This means however that m is allowed to be very large. To avoid overfitting, we propose to add an ℓ_1 complexity penalty to the empirical SVM loss.

Let for $f : \mathcal{X} \to \mathbf{R}$,

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

We consider a class of functions $\mathcal{F} \subset \{f_\alpha : \ \alpha \in \mathbf{R}^m\}$, satisfying for some given finite constant $K \geq 2$,

$$\|f - \tilde{f}\|_\infty \leq K, \ \forall \ f, \tilde{f} \in \mathcal{F} \ .$$

The assumption of a uniformly bounded class of functions will be crucial in our calculations. In some cases, the estimator $\hat{f}_n$ defined below in (1) does not change if we allow $\mathcal{F}$ to be the class of **all** functions f_α, $\alpha \in \mathbf{R}^m$, but this will very much depend on the situation. The dependency on K of our results will be given explicitly.

The ℓ_1 penalized SVM estimator $\hat{f}_n$ is defined as

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

where λ_n is a regularization parameter, which we choose as

$$\lambda_n = cK^2 \sqrt{\frac{\log n}{n}} \ .$$

Here c is a large enough universal constant. (Our rough calculations in Section 5 show that $c = 104$ will do in Theorem 2.1 below.) If $\{\psi_k\} \subset L_2(Q)$ contains no linear dependencies, the ℓ_1 penalty generally leads to sparse representations, i.e., 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 $\hat{\alpha}_k$. It is related to soft thresholding (see Donoho (1995)), and is referred to as the LASSO in Tibshirani (1996) and Hastie, Tibshirani and Friedman (2001).

We will study the theoretical SVM loss $R(\hat{f}_n)$ of the estimator $\hat{f}_n$. Note that $R(\hat{f}_n)$ is a random variable. The difference $R(f) - R(f^*)$ is called the SVM *excess risk* at f. We show in Theorem 2.1 that the SVM excess risk at $\hat{f}_n$ tends to zero in probability, with speed depending on the smoothness of the boundary of f^*, as well as on the *margin parameter*, or *noise level*, κ. The definition of the parameter κ is presented in Condition A below.

The prediction error excess risk is $P(Yf(X) < 0) - P(Yf^*(X) < 0)$. Rates of convergence for the SVM excess risk imply the same rates for the prediction error excess risk, as Bartlett, Jordan and McAuliffe (2003) have shown that

$$P(Yf(X) < 0) - P(Yf^*(X) < 0) \leq R(f) - R(f^*) \ .$$

We show that the rate of convergence following from Theorem 2.1 is, in the example considered in Section 4, up to log-terms equal to the minimax rates for the prediction error excess risk, the latter being established in Mammen and Tsybakov (1999).

Adaptivity and minimax rates have also been obtained in Tsybakov (2004), Tsybakov and van de Geer (2003) and Koltchinskii (2003). These papers use empirical risk minimization, which make the methods proposed there computationally infeasible. In other work, e.g., Koltchinskii (2001), Koltchinskii and Panchenko (2002), Lugosi and Wegkamp (2004), a different concept of adaptivity is studied, and for example Rademacher complexities.

This paper is organized as follows. In the next section, we present the conditions and main theorem. Section 3 discusses the conditions. Section 4 shows in an example that minimax rates can be obtained. The proof of the main theorem is given in Section 5.

2 Conditions and main theorem

Let $\|\cdot\|_p$ be the $L_p(Q)$ norm ($1 \leq p < \infty$). We start out with an identifiability condition, which we refer to as *margin* condition.

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

$$R(f) - R(f^*) \geq \|f - f^*\|_1^\kappa / \sigma^\kappa \ . \tag{2}$$

Next we impose conditions on the system $\{\psi_k\}$. Let $\mathcal{A}^*$ be some subset of $\mathbf{R}^m$, and $N(\cdot) : \mathcal{A}^* \to \{1, \ldots, m\}$ be some map.

Condition B. *For some $0 \leq \beta \leq 1$ and for all $\alpha \in \mathbf{R}^m$ and $\alpha^* \in \mathcal{A}^*$, with f_α and f_{α^*} in $\mathcal{F}$, it holds that*

$$\sum_{\alpha_k^* \neq 0} |\alpha_k - \alpha_k^*| \leq c_\psi \sqrt{N(\alpha^*)} \|f_\alpha - f_{\alpha^*}\|_1^\beta \ . \tag{3}$$

Here c_ψ is a constant, possibly depending on the system $\{\psi_k\}$ and possibly also depending on K.

Remark 2.1 In Subsection 3.2.1, we will give an example where Condition B holds with $N(\alpha^*) = \#\{\alpha_k \neq 0\}$, $\mathcal{A}^* = \mathbf{R}^m$, and $\beta = 1/2$. The example of Subsection 3.2.2 has somewhat similar $N(\alpha^*)$, but $\beta = 1$.

Remark 2.2 It is clear that in Condition A, the constants κ and σ depend on P. But also in Condition B, the set $\mathcal{A}^*$, the map $N(\cdot)$, and the constants β and c_ψ may depend on P, in particular on sparse approximations f_{α^*} of Bayes rule f^*. We also note that we restricted ourselves to maps $N(\cdot)$ taking values in $\{1, \ldots, m\}$, and assumed the bound (3) with the square root $\sqrt{N(\cdot)}$ and the power $\|f_\alpha - f_{\alpha^*}\|_1^\beta$. This can be extended, with on the right hand side of (3) more general strictly increasing functions of $\|f_\alpha - f_{\alpha^*}\|_1$. However, it is not clear to us whether such generalizations are important, and we restricted ourselves to facilitate the exposition.

Remark 2.3 Unless we take $\beta = 0$ in (3), Condition B is about identifiability of α given f_α. For systems that are linearly dependent in $L_2(Q)$, identifiability is clearly an issue. In that case, an adjustment of Condition B may be needed, which depends on whether sparse approximations resemble ℓ_1 restricted approximations. We refer to Donoho (2004a,b) for results in the latter direction.

The last two conditions are technical conditions that we needed to make the proof work.

Condition C. *We require that*

$$\sum_{k=1}^{m} |\alpha_k - \alpha_k^*| \leq K\sqrt{\frac{n}{\log n}} \ \forall \ \alpha, \alpha^* \ \textit{with} \ f_\alpha, f_{\alpha^*} \in \mathcal{F} \ . \tag{4}$$

Condition D. *The system of basis functions satisfies*

$$\max_{k=1,\ldots,m} \|\psi_k\|_\infty \leq \sqrt{\frac{n}{\log n}} \ . \tag{5}$$

We are now ready to formulate our main theorem. This theorem says that the ℓ_1 penalized SVM estimator balances approximation error and estimation error. Here, given $\bar{N} \in \{1, \ldots, m\}$, the approximation error is

$$\inf\left\{R(f_{\alpha^*}) - R(f^*) : \ f_{\alpha^*} \in \mathcal{F}, \ \alpha^* \in \mathcal{A}^*, \ N(\alpha^*) = \bar{N}\right\}$$

and the estimation error is

$$V_n(\bar{N}) = \begin{cases} 2\delta^{-\frac{\beta}{\kappa-\beta}}[4\sigma^2 c_\psi^2 \lambda_n^2 \bar{N}]^{\frac{\kappa}{2(\kappa-\beta)}} & \kappa > \beta \\ 0 & \kappa = \beta = 1 \end{cases} \ , \tag{6}$$

where $0 < \delta \leq 1/2$ is fixed, but otherwise arbitrary.

Theorem 2.1 *Let $\hat{f}_n$ be the ℓ_1 penalized SVM estimator defined in (1), with regularization parameter $\lambda_n = cK^2\sqrt{\log n/n}$, and c an appropriate universal constant. Suppose that conditions A, B, C and D hold. When $\kappa = \beta = 1$, we assume*

$$4\sigma^2 c_\psi^2 \lambda_n^2 m \leq \delta^2. \tag{7}$$

Define

$$\begin{aligned} \epsilon_n \quad = \quad & (1+4\delta)\inf\left\{R(f_{\alpha^*}) - R(f^*) + V_n(N(\alpha^*)) : \ f_{\alpha^*} \in \mathcal{F}, \ \alpha^* \in \mathcal{A}^*\right\} \\ & + (1+4\delta)\lambda_n K\sqrt{\frac{\log n}{n}} \ . \end{aligned} \tag{8}$$

Then, for a universal constant c_0,

$$\mathbb{P}\left(R(\hat{f}_n) - R(f^*) > \epsilon_n\right) \leq c_0 \exp[-\frac{K^2 \log n}{2}] \ . \tag{9}$$

Remark 2.4 The case $\kappa = \beta = 1$ is special. Note that condition (7) is met for sufficiently large n if we choose m somewhat smaller than $n/\log n$, say $m = n/(\log n)^{1+\gamma}$, with $\gamma > 0$.

Remark 2.5 Since $\lambda_n^2 \asymp \log n/n$, the estimation error is for $\kappa > \beta$ of order

$$V_n(\bar{N}) \asymp [\frac{\bar{N}\log n}{n}]^{\frac{\kappa}{2(\kappa-\beta)}}.$$

The worst case corresponds to $\kappa = \infty$, giving

$$V_n(\bar{N}) \asymp \sqrt{\frac{\bar{N}\log n}{n}}.$$

Remark 2.6 When $\beta = 1/2$, the estimation error (6) becomes

$$V_n(\bar{N}) \asymp [\lambda_n^2 \bar{N}]^{\frac{\kappa}{2\kappa-1}} \ .$$

This is as in Tsybakov and van de Geer (2003). However, the latter paper deals with a different approximation error, so that rates following from Theorem 2.1 with $\beta = 1/2$ may be quite different from those in Tsybakov and van de Geer (2003).

We conclude that the ℓ_1 penalized SVM estimator adapts to certain properties of the unknown distribution P, by trading off the estimation error and approximation error. This results in fast rates for the excess risk $R(\hat{f}_n) - R(f^*)$.

3 A discussion of the conditions

Conditions A and B together take care that the result follows easily from a probability inequality for the empirical process (see Lemma 5.1). Conditions C and D make sure that indeed the probability inequality holds (see Lemmas 5.3, 5.4 and 5.5). Conditions B, C and D however do belong together in the sense that they are all on the properties of the system $\{\psi_k\}$. We study Condition A in Subsection 3.1, and Conditions B,C and D in Subsection 3.2.

3.1 On Condition A

Condition A appears in similar form (for prediction error instead of SVM loss) in Mammen and Tsybakov (1999), Tsybakov (2004) and Tsybakov and van de Geer (2003). It follows essentially from conditions on the behavior of η near $\{\eta = 1/2\}$, and is therefore often called the *margin* condition, or condition on the *noise level*. Here, we need in addition to control the behavior near the boundaries $\{\eta = 0\}$ and $\{\eta = 1\}$. Define

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

Condition AA. *There exist constants $C \geq 1$ and $0 < a \leq \infty$ such that for all $z > 0$*

$$Q(\{\tau \leq z\}) \leq Cz^a \ . \tag{10}$$

We now show that Condition A holds with $\kappa = (a+1)/a$. The case where $a = \infty$ corresponds to the situation where the function η stays away from 0, 1 and $\frac{1}{2}$. In that case $\kappa = 1$. At the other extreme is $a \downarrow 0$, giving $\kappa \to \infty$. This can occur if η only takes values very near to $\frac{1}{2}$, so that Bayes rule is not much better than flipping a fair coin.

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

$$R(f) - R(f^*) \geq \sigma_K^{-1} \|f - f^*\|_1^{\frac{a+1}{a}} \ , \tag{11}$$

with

$$\sigma_K = (CK(a+1))^{\frac{1}{a}} \ (a+1)/a \ . \tag{12}$$

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

Proof. By straightforward manipulation, we obtain

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

This implies the inequality

$$R(f) - R(f^*) \geq \int |f - f^*| \tau dQ \ .$$

Hence, for any $z > 0$,

$$\begin{aligned}
R(f) - R(f^*) &\geq \int_{\tau > z} |f - f^*| \tau dQ \geq z \int_{\tau > z} |f - f^*| dQ \\
&= z\|f - f^*\|_1 - z \int_{\tau \leq z} |f - f^*| dQ \ .
\end{aligned}$$

But, since $\|f - f^*\|_\infty \leq K$,

$$\int_{\tau \leq z} |f - f^*| dQ \leq KQ(\{\tau \leq z\}) \leq CKz^a \ ,$$

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

$$R(f) - R(f^*) \geq z\|f - f^*\|_1 - CKz^{a+1} \ .$$

Take

$$z = \left(\frac{\|f - f^*\|_1}{CK(a+1)} \right)^{\frac{1}{a}} \ ,$$

to arrive at the result of the lemma. □

3.2 On conditions B, C and D

3.2.1 Orthonormal systems

Let $\{\psi_k\}_{k=1}^m$ be an orthonormal system in $L_2(Q)$, i.e.,

$$\int \psi_k \psi_l dQ = \begin{cases} 1 & \text{if } k = l \\ 0 & \text{if } k \neq l \end{cases} \ .$$

Lemma 3.2 *Condition B holds for orthonormal systems with $N(\alpha^*) = \#\{\alpha_k^* \neq 0\}$, $\alpha^* \in \mathcal{A}^* = \mathbf{R}^m$, $c_\psi = \sqrt{K}$ and $\beta = 1/2$. If $m \leq n/\log n$, Condition C holds too.*

Proof. One readily sees that

$$\|f_\alpha\|_2^2 = \sum_{k=1}^m |\alpha_k|^2 \ .$$

Take an arbitrary subset of $\{1, \ldots, m\}$, say $\{1. \ldots, N\}$, $N \leq m$. Then

$$\sum_{k=1}^N |\alpha_k| \leq \sqrt{N} (\sum_{k=1}^N \alpha_k^2)^{1/2} \leq \sqrt{N} \|f_\alpha\|_2 \ .$$

Now, let α and α^* be in $\mathbf{R}^m$ and assume both f_α and f_{α^*} are in $\mathcal{F}$, so that $|f_\alpha - f_{\alpha^*}|$ is bounded by K. Then we find

$$\sum_{k=1}^N |\alpha_k - \alpha_k^*| \leq \sqrt{N} \|f_\alpha - f_{\alpha^*}\|_2 \leq \sqrt{N} K^{1/2} \|f_\alpha - f_{\alpha^*}\|_1^{1/2} \ . \tag{13}$$

Thus, Condition B is met with $c_\psi = \sqrt{K}$ and $\beta = 1/2$.
Finally, when $m \leq n/\log n$, Condition C follows from (13) with $N = m$. □

Remark 3.1 Using the same reasoning as in Lemma 3.2, we can extend the situation to possibly non-orthogonal, but linearly independent systems. Let us define the $m \times m$ matrix

$$\Sigma = \int \psi \psi^T dQ \ .$$

Suppose that the smallest eigenvalue of Σ, $\lambda_{\min}^2$ say, is positive. Then Condition B is met with $c_\psi = \sqrt{K}/\lambda_{\min}$ and $\beta = 1/2$. When we assume in addition $m \leq \lambda_{\min} n/\log n$. Condition C holds too.

3.2.2 Wavelet systems

Let $\mathcal{X}$ be a subset of r-dimensional Euclidean space. We introduce a double index for the system of functions $\{\psi_k\} = \{\psi_{j,l},\ j \in I_l,\ l = 1, \ldots, L_0\}$, with l the index of the resolution level, and j the index of a function at a given level. Assume that the system of functions $\{\psi_{j,l},\ j \in I_l,\ l = 1, \ldots, L_0\}$ is orthonormal in $L_2(Q)$, as in Subsection 3.2.1. Then it follows from Lemma 3.2 that Condition B holds with $\beta = 1/2$. We will show that this can be improved to $\beta = 1$.

Assume for some constant $C_0 \geq 1$,

$$2^{rl}/C_0 \leq |I_l| \leq C_0 2^{rl} \ , \tag{14}$$

and

$$\sum_{j \in I_l} \|\psi_{j,l}\|_\infty \leq C_0^{1/2} 2^{rl/2},\ l = 1, \ldots, L_0 \ . \tag{15}$$

Conditions (14) and (15) are standard for compactly supported wavelet systems (see e.g. Härdle, Kerkyacharian, Picard and Tsybakov (1998)). We also assume that

$$2^{rL_0} \leq \frac{n}{16 C_0 \log n} \ . \tag{16}$$

Lemma 3.3 *Under (14) and (15), Condition B holds with $c_\psi = 4C_0$ and $\beta = 1$, and with $N(\alpha^*) = |I_L|$, $\alpha^* \in \mathcal{A}^* = \mathbf{R}^m$, where*

$$L = \min\{l : \alpha^*_{j,l} = 0,\ \forall\ (j,l) \text{ with } l > L\} \ .$$

If in addition (16) is met, then also Conditions C and D hold.

Proof. We have

$$\sum_{\alpha_k \neq 0} |\alpha_k| \leq \sum_{l=1}^{L} \sum_{j \in I_l} |\alpha_{j,l}| \leq \sum_{l=1}^{L} \sum_{j \in I_l} \|f_\alpha \psi_{j,l}\|_1$$

$$\leq \sum_{l=1}^{L} \sum_{j \in I_l} \|\psi_{j,l}\|_\infty \|f_\alpha\|_1 \leq \sum_{l=1}^{L} C_0^{1/2} 2^{rl/2} \|f_\alpha\|_1 \leq 4 C_0^{1/2} 2^{rL/2} \|f_\alpha\|_1 \tag{17}$$

$$\leq 4 C_0 \sqrt{N(\alpha)} \|f_\alpha\|_1 \ .$$

Hence, Condition B holds with $c_\psi = 4C_0$.

Taking $L = L_0$, (17) together with (16) yields that

$$\sum_{k=1}^{m} |\alpha_k| \leq 4 C_0^{1/2} 2^{rL_0/2} \|f_\alpha\|_1 \leq \sqrt{\frac{n}{\log n}} \|f_\alpha\|_1 \ .$$

So also

$$\sum_{k=1}^{m} |\alpha_k - \alpha^*_k| \leq \sqrt{\frac{n}{\log n}} \|f_\alpha - f_{\alpha^*}\|_1 \leq K \sqrt{\frac{n}{\log n}} \ ,$$

whenever both f_α and f_{α^*} are in $\mathcal{F}$. Hence, Condition C is met.

Condition D follows from (15) and (16),

$$\max_{j,l} \|\psi_{j,l}\|_\infty \leq C_0^{1/2} 2^{rL_0/2} \leq \sqrt{\frac{n}{\log n}} \ .$$

□

Remark 3.2 The situation can be extended to "anisotropic" cases. For example, suppose Q is Lebesgue measure on $[0,1]^r$. Let $\{\psi_{j,l}\}$ be a wavelet system on $[0,1]$, satisfying (14) and (15), with $r=1$. We make this into an orthonormal system on $[0,1]^r$ by taking all r-fold products

$$\prod_{t=1}^{r} \psi_{j_t,l_t},\ l_t \leq L_0,\ j_t \in I_{l_t},\ t=1,\ldots,r\ .$$

We invoke the notation

$$f_\alpha = \sum \alpha_{j_1\ldots j_r,l_1\ldots l_r}\psi_{j_1,l_1}\ldots\psi_{j_r,l_r}\ .$$

One could then for instance choose

$$N(\alpha) = \prod_{t=1}^{r} |I_{L_t}|\ ,$$

where $L_t = \min\{l:\ \alpha_{j_1\ldots,j_r,l_1,\ldots,l_r}=0\ \forall\ (j_1\ldots j_r,l_1\ldots l_r) \text{ with } l_t > L_t\}$.
This will be our approach in the example of Section 4.

3.2.3 Linearly dependent systems

It is important to realize that the choice of the basis functions plays a role in Conditions B, C and D, but also in considerations on their approximating properties. Recall that Bayes rule f^* takes only the values ± 1. Approximating such a function by for example a wavelet series is not always very natural, as a good approximation might require very many non zero coefficients. We discuss briefly the overcomplete systems *wedgelets* and *curvelets* as alternatives to wavelets.

Let us assume that $\mathcal{X}=[0,1]^2$ and that Q is Lebesgue measure. Suppose moreover that f^* is a boundary fragment, i.e.,

$$f^*(x) = \begin{cases} +1 & \text{if } x \in \{(u,v)\in[0,1]^2:\ g^*(u)\geq v\} \\ -1 & \text{else} \end{cases}\ . \tag{18}$$

Assume that g^* is Hölder continuous with exponent 1, i.e.,

$$|g^*(u)-g^*(\tilde{u})| \leq |u-\tilde{u}|,\ u,\tilde{u}\in[0,1]\ .$$

Suppose moreover that the derivative $\dot{g}^*$ exists, and is Hölder continuous with exponent γ, $0<\gamma<1$. In that case, we call $s=1+\gamma$ the smoothness of the boundary g^*. Donoho (1999) proves that there is a N term wedgelet approximation f_N^{Wedge} such that

$$\|f_N^{\text{Wedge}} - f^*\|_2^2 \leq C_s N^{-s}\ .$$

Moreover, the approximation f_N^{Wedge} itself also only takes the values ± 1. so that also

$$\|f_N^{\text{Wedge}} - f^*\|_1 = \frac{1}{2}\|f_N^{\text{Wedge}} - f^*\|_2^2 \leq \frac{C_s}{2}N^{-s}\ . \tag{19}$$

Now, it is clear that for the overcomplete system of wedgelets, Condition B is cumbersome (see also Remark 2.3). Nevertheless, let us speculate on the rate that could follow from Theorem 2.1. To this end, assume the following four conditions hold true. Assume that Condition A holds and that in fact

$$R(f_N^{\text{wedge}}) - R(f^*) \asymp N^{-s\kappa}$$

which is true if Condition A and (19) are tight. Now, recall first that Condition B is true for wavelets with $\beta=1$ and $N(\alpha^*)$ the number of terms in the highest resolution level used by α^*. Here, we assume, roughly speaking, that Condition B holds for wedgelets with $\beta=1/2$, and similar $N(\alpha^*)$, where α^* varies only over the set $\mathcal{A}^*$ of sparse approximations of f^*. Finally, we assume Conditions C and D for wedgelets. Theorem 2.1 would then yield the rate

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

with

$$\rho = \frac{\kappa s}{2\kappa s + 1 - s} \ .$$

Clearly, because of the many unverified assumptions, we do not claim to have established this rate for wedgelets. Nevertheless, it gives a benchmark for the result of the example of Section 4. There, we actually prove the same rate, but in a different context and using the linearly independent Haar basis instead of wedgelets (see Corollary 4.3).

Let us also briefly mention an example from Candès and Donoho (2004), where f^* is not necessarily a boundary fragment. Suppose that the set $\{f^* = 1\}$ has a C^2 boundary. Then the best N term wavelet approximation f_N^W would obey

$$\|f_N^W - f^*\|_2^2 \asymp \frac{1}{N}, \ N \to \infty \ ,$$

see Candès and Donoho (2004). They also that an N term curvelet approximation f_N^C satisfies

$$\|f_N^C - f^*\|_2^2 \asymp \frac{\log^3 N}{N^2}, \ N \to \infty \ .$$

We conclude that one on the one hand we need good (sparse) approximation properties to Bayes rule f^* of the system $\{\psi_k\}$ and on the other hand (some variant of) Condition B.

4 An example: boundary fragments

Suppose that space $\mathcal{X}$ is the 2-dimensional unit cube $[0,1]^2$, and that the distribution Q of X is Lebesgue measure.
Assume that f^* is a boundary fragment as in (18). To simplify, we consider a modification of the situation of Subsection 3.2.3. Recall that there, we assumed g^* Hölder continuous with exponent 1, and $\dot{g}^*$ Hölder continuous with exponent γ. Such a function can be well approximated by a piecewise linear function. In this section, we assume that g^* can be well approximated by a piecewise constant function. This allows us to avoid rather subtle approximation theory.
Let for all $N \in \{1, 2, \ldots\}$, $\mathcal{G}(N)$ be the class of piecewise linear functions

$$\mathcal{G}(N) = \{g(u) = \sum_{k=1}^{N} a_k 1\{(k-1)/N < u \leq k/N\} : \ (a_1, \ldots, a_N) \in \mathbf{R}^N\}.$$

We assume that

$$\min_{g \in \mathcal{G}(N)} \int |g^*(u) - g(u)| du \leq N^{-s}, \ \forall N \in \{1, 2 \ldots\} \ . \tag{20}$$

We now call s the *smoothness* parameter. Note that when $s \leq 1$, assumption (20) is true under the Hölder condition

$$|g^*(u) - g^*(\tilde{u})| \leq |u - \tilde{u}|^s, \ u, \tilde{u} \in [0,1] \ . \tag{21}$$

Now, let $\{\psi_{j,l}\}$ be the Haar basis of $L_2([0,1], \text{Lebesgue measure})$. So

$$\psi_{1,0} = 1$$

$$\psi_{1,1} = 1_{(0,1/2]} - 1_{(1/2,1]} \ ,$$

and generally,

$$2^{-(l+1)/2} \psi_{j,l} = 1_{(2(j-1)2^{-l}, (2j-1)2^{-l}]} - 1_{((2j-1)2^{-l}, (2j)2^{-l}]}, \ j = 1, \ldots, 2^{l-1} \ . \tag{22}$$

We use the expansion

$$f_\alpha = \sum_{i,j,k,l} \alpha_{i,j,k,l} \psi_{i,k} \psi_{j,l} \ .$$

The resolution levels k and l at taken most equal to L_0, where

$$2^{L_0} \leq \sqrt{\frac{n}{16C_0 \log n}} \ .$$

Lemma 4.1 *Assume f^* satisfies (20). Let $\epsilon \leq 1$ and*

$$\min\{\epsilon, \epsilon^{1/s}) \geq 2^{-L_0+2} \ .$$

Take $2^{-k_s+2} \leq \epsilon^{1/s}$ and $2^{-l_s+2} \leq \epsilon$. Then there is an $\alpha^* = (\alpha^*_{i,j,k,l})$ with $\alpha^*_{i,j,k,l} = 0$ whenever $k > k_s$ or $l > l_s$, such that

$$\|f_{\alpha^*} - f^*\|_1 \leq \epsilon \ .$$

Proof. Let for non-negative z, $\lfloor z \rfloor$ be the largest integer less than or equal to z. Define

$$\bar{g}_s = \lfloor 2^{l_s} g^*_s \rfloor 2^{-l_s} \ ,$$

where

$$g^*_s = \arg \min_{g \in \mathcal{G}(2^{k_s})} \int |g^*(u) - g(u)| du \ .$$

Take

$$\bar{f}(u,v) = 2 1\{v \leq \bar{g}_s(u)\} - 1 \ .$$

Then

$$\|\bar{f} - f^*\|_1 = 2 \int |\bar{g}_s(u) - g^*(u)| du$$

$$= 2(\int |\bar{g}_s(u) - g^*_s(u)| du + \int |g^*_s(u) - g^*(u)| du)$$

$$\leq 2(2^{-l_s} + 2^{-k_s s}) \leq \epsilon \ .$$

The function $\bar{f}$ is constant on the sets $((i-1)2^{-k_s}, i2^{-k_s}] \times ((j-1)2^{-l_s}, j2^{-l_s}]$, $i \in \{1, \ldots, k_s\}$, $j \in \{1, \ldots, l_s\}$. Therefore the coefficients $\alpha^*_{i,j,k,l}$ of $\bar{f}$ in the wavelet expansion

$$\bar{f} = \sum_{i,j,k,l} \alpha^*_{i,j,k,l} \psi_{i,k} \psi_{j,l}$$

are zero whenever $k > k_s$ or $l > l_s$. □

Define now

$$k(\alpha) = \min\{k_0 : \alpha_{i,j,k,l} = 0 \text{ whenever } k > k_0\} \ ,$$
$$l(\alpha) = \min\{l_0 : \alpha_{i,j,k,l} = 0 \text{ whenever } l > l_0\} \ ,$$

and

$$N(\alpha) = |I_{k(\alpha)}| \times |I_{l(\alpha)}| \ .$$

Lemma 4.2 *Suppose g^* satisfies (20). For $N \in \{2^k, \ k = 0, 1, \ldots\}$, and*

$$\max\{N^{\frac{s}{s+1}}, N^{\frac{1}{s+1}} \leq 2^{L_0 - 2}\},$$

we have the inequality

$$\min_{\alpha^*: \ N(\alpha^*) = N} \|f_{\alpha^*} - f^*\|_1 \leq 16 N^{-\frac{s}{s+1}}. \tag{23}$$

Proof. Suppose $N = 2^{k-2}$, $k \geq 2$. Let $\lfloor z \rfloor$ denote the integer part of $z \geq 0$ and define

$$t = \lfloor \frac{sk}{s+1} \rfloor + 1 \ .$$

Then obviously

$$N \leq 2^{t-1} 2^{\lfloor t/s \rfloor} \ .$$

Define $\epsilon = 2^{-t+2}$ and apply Lemma 4.1 with $k_s = \lfloor t/s \rfloor + 1$ and $l_s = t$. One then finds

$$\min_{\alpha^*: \ N(\alpha^*) \leq N} \|f_{\alpha^*} - f^*\|_1 \leq 42^{-t} \leq 16 N^{-\frac{s}{s+1}} \ ,$$

where in the last inequality, we used $t \geq \frac{sk}{s+1}$. □

Corollary 4.1 *Suppose that (20) is met, with unknown smoothness s. Assume Condition A holds with unknown margin parameter κ, and that in fact the lower bound for $R(f) - R(f^*)$ is tight, i.e.,*

$$R(f) - R(f^*) \asymp \|f - f^*\|_1^\kappa, \ f \in \mathcal{F} \ .$$

Take the the above Haar system (22), and use $\{\psi_{i,k}\psi_{j,l}\}$ as two-dimensional system. Assume resolution levels k and l at most $\sqrt{n/16C_0 \log n}$. Furthermore, suppose that $s > 1/(2\kappa - 1)$, i.e. that we do not need the finer resolution levels. By Lemma 4.2, the approximation error is now

$$\inf_{\alpha^*: \ N(\alpha^*) \leq N} R(f_{\alpha^*}) - R(f^*) \asymp N^{-\frac{s\kappa}{s+1}} \ .$$

so that the application of Theorem 2.1 gives the rate

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

with

$$\rho = \frac{\kappa s}{2\kappa s + 1 - s} \ .$$

Remark 4.1 For $s \leq 1$, this corresponds up to the log-term to the minimax rate over Hölder classes of boundaries (see Mammen and Tsybakov (1999)). It is down to log-terms the same rate as given in Tsybakov and van de Geer (2003). Tsybakov (2004) proves similar adaptive rates for nested models, using aggregation of classifiers.

5 Proof of Theorem 2.1

Let us write

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

and

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

Throughout most this section, we fix an arbitrary $\alpha^* \in \mathcal{A}^*$, with $f_{\alpha^*} \in \mathcal{F}$. The result of Theorem 2.1 then follows from taking the infimum over all such f_{α^*}. This is done at the very end of this section.

Let Ω^* be the set

$$\Omega^* = \left\{ |\nu_n(\hat{f}_n) - \nu_n(f_{\alpha^*})| \leq \sqrt{n}\lambda_n \left(I(\hat{\alpha}_n - \alpha^*) + K\sqrt{\log n/n} \right) \right\} \ .$$

We show in Lemmas 5.3, 5.4 and 5.5, that under Conditions C and D, the set $\{\omega \notin \Omega^*\}$ has probability at most $c_0 \exp[-K^2 \log n/2]$. Lemma 5.1 below tells us that Conditions A and B yield, on Ω^*, the bound

$$\epsilon_n(f_{\alpha^*}) = (1 + 4\delta) \left\{ R(f_{\alpha^*}) - R(f^*) + V_n(N(\alpha^*)) + \lambda_n K \sqrt{\frac{\log n}{n}} \right\} \tag{24}$$

for the excess risk $R(\hat{f}_n) - R(f^*)$.

Lemma 5.1 *Assume Conditions A and B, and, if $\kappa = \beta = 1$,*

$$4\sigma^2 \lambda_n^2 c_\psi^2 m \leq \delta^2 \ . \tag{25}$$

Then on Ω^*,

$$R(\hat{f}_n) - R(f^*) \leq \epsilon_n(f_{\alpha^*}) \ , \tag{26}$$

where $\epsilon_n(f_{\alpha^*})$ is given in (24).

Proof. We use similar arguments as in Loubes and van de Geer (2002),van de Geer (2003) and Tsybakov and van de Geer (2003). Define $N^* = N(\alpha^*)$, and

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

Then

$$\begin{aligned} R(\hat{f}_n) - R(f_{\alpha^*}) &= -\left(\nu_n(\hat{f}_n) - \nu_n(f_{\alpha^*})\right)/\sqrt{n} + \lambda_n(I(\alpha^*) - I(\hat{\alpha}_n)) \\ &\quad + [R_n(\hat{f}_n) + \lambda_n I(\hat{\alpha}_n)] - [R_n(f_{\alpha^*}) + \lambda_n I(\alpha^*)] \\ &\leq -\left(\nu_n(\hat{f}_n)] - \nu_n(f_{\alpha^*})\right)/\sqrt{n} + \lambda_n \left(I(\alpha^*) - I(\hat{\alpha}_n)\right). \end{aligned}$$

Thus, on Ω^*,

$$\begin{aligned} R(\hat{f}_n) - R(f_{\alpha^*}) &\leq \lambda_n \left(I(\hat{\alpha}_n - \alpha^*) + K\sqrt{\log n/n}\right) + \lambda_n \left(I(\alpha^*) - I(\hat{\alpha}_n)\right) \\ &\leq 2\lambda_n I_1(\hat{\alpha}_n - \alpha^*) + \lambda_n K\sqrt{\log n/n} \ . \end{aligned}$$

By Condition B, we arrive at

$$R(\hat{f}_n) - R(f_{\alpha^*}) \leq 2\lambda_n c_\psi \sqrt{N^*} \|\hat{f}_n - f_{\alpha^*}\|_1^\beta + \lambda_n K\sqrt{\log n/n} \ . \tag{27}$$

Now, let us use the short hand notation

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

Then (27) can be rewritten as

$$\hat{d} \leq 2\lambda_n c_\psi \sqrt{N^*} \|\hat{f}_n - f_{\alpha^*}\|_1^\beta + \lambda_n K\sqrt{\log n/n} + d^* \ . \tag{28}$$

By the triangle inequality and Condition A,

$$\|\hat{f}_n - f_{\alpha^*}\|_1^\beta \leq \sigma(\hat{d})^{\frac{\beta}{\kappa}} + \sigma(d^*)^{\frac{\beta}{\kappa}} \ .$$

When $\kappa > \beta$, we may invoke Lemma 5.2 below to obtain

$$\hat{d} \leq \delta(\hat{d} + d^*) + 2\delta^{-\frac{\beta}{\kappa-\beta}} [2\sigma\lambda_n c_\psi \sqrt{N^*}]^{\frac{\kappa}{\kappa-\beta}} + \lambda_n K\sqrt{\log n/n} + d^* \ .$$

When $\kappa = \beta = 1$, we get from (28) that

$$\begin{aligned} \hat{d} &\leq 2\sigma\lambda_n c_\psi \sqrt{N^*}(\hat{d} + d^*) + \lambda_n K\sqrt{\log n/n} + d^* \\ &\leq \delta(\hat{d} + d^*) + \lambda_n K\sqrt{\log n/n} + d^* \ , \end{aligned}$$

where we applied assumption (25).
Since for $\delta \leq 1/2$,

$$\frac{1+\delta}{1-\delta} \leq 1 + 4\delta \ ,$$

we now have shown that for $\kappa > \beta$,

$$\hat{d} \leq (1+4\delta)\left\{d^* + 2\delta^{-\frac{\beta}{\kappa-\beta}} [2\sigma\lambda_n c_\psi \sqrt{N^*}]^{\frac{\kappa}{\kappa-\beta}} + \lambda_n K\sqrt{\frac{\log n}{n}}\right\} \ ,$$

and for $\kappa = \beta$,

$$\hat{d} \leq (1+4\delta)\left\{d^* + \lambda_n K\sqrt{\frac{\log n}{n}}\right\} \ .$$

□

Lemma 5.2 below is the technical lemma applied in the previous lemma. Its proof is straightforward and can be found in Tsybakov and van de Geer (2003).

Lemma 5.2 *We have for all positive v, t, and δ, and $\kappa > \beta$,*

$$vt^{\frac{\beta}{\kappa}} \leq \delta t + v^{\frac{k}{\kappa-\beta}} \delta^{-\frac{\beta}{\kappa-\beta}} \ . \tag{29}$$

We now will show that the set Ω^* has probability close to one. To this end, the following concentration inequality will be applied (Ledoux (1996), Massart (2000)).

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

$$\mathbf{Z} = \sup_{\gamma \in \Gamma} \left| \frac{1}{n} \sum_{i=1}^{n} [\gamma(Z_i) - \mathbb{E}\gamma(Z_i)] \right| \tag{30}$$

and

$$\tau^2 = \sup_{\gamma \in \Gamma} \mathrm{Var}(\gamma(Z)) \ . \tag{31}$$

Then for any positive z,

$$\mathbb{P}\left(\mathbf{Z} \geq 2\mathbb{E}\mathbf{Z} + \tau\sqrt{8z/n} + 69Kz/(2n) \right) \leq \exp(-z) \ . \tag{32}$$

Lemma 5.3 *Define $\mathcal{F}_M = \{f_\alpha \in \mathcal{F} : I(\alpha - \alpha^*) \leq M\}$, 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 $K\sqrt{\log n/n} \leq M \leq K\sqrt{n/\log n}$, we have

$$\mathbb{P}\left(\mathbf{Z}_M \geq 2\mathbb{E}\mathbf{Z}_M + 36K^2 M\sqrt{\log n/n} \right) \leq \exp[-(M^2 \vee K^2)\log n] \ .$$

Proof. In Theorem 5.1, we take

$$\Gamma = \{\gamma_\alpha(x,y) = (1 - yf_\alpha(x))_+ - (1 - yf_{\alpha*}(x))_+ : \ f_\alpha \in \mathcal{F}_M\} \ .$$

We have

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

This implies that $\tau^2 \leq \sup_{f_\alpha \in \mathcal{F}} \|f_\alpha - f_{\alpha^*}\|^2$. So $\tau^2 \leq M^2 \wedge K^2 := \tau_1^2$. We now take $z = (M^2 \vee K^2)\log n$. Then, for $K \leq M \leq K\sqrt{n/\log n}$,

$$\tau_1\sqrt{8z/n} + 69Kz/(2n) = KM\sqrt{8\log n/n} + 69KM^2 \log n/(2n)$$

$$\leq 3KM\sqrt{\log n/n} + (69/2)K^2 M\sqrt{\log n/n} \leq 36K^2 M\sqrt{\log n/n} \ .$$

Moreover, for $K\sqrt{\log n/n} \leq M \leq K$,

$$\tau_1\sqrt{8z/n} + 69Kz/(2n) = KM\sqrt{8\log n/n} + 69KK^2 \log n/(2n)$$

$$\leq 3KM\sqrt{\log n/n} + (69/2)K^2 M\sqrt{\log n/n} \leq 36K^2 M\sqrt{\log n/n} \ .$$

The result thus follows from Theorem 5.1. □

Lemma 5.4 *Suppose Condition D is met. We have for $\mathbf{Z}_M$ defined in Lemma 5.3*

$$\mathbb{E}\mathbf{Z}_M \leq 32M\sqrt{\log n/n} \ . \tag{33}$$

Proof. This follows from the similar arguments as in van de Geer (2003), using the fact that the function $z \mapsto (1-z)_+$, $z \in \mathbf{R}$ 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

$$\mathbb{E}\mathbf{Z}_M \leq 4\mathbb{E}\left(\sup_{f_\alpha \in \mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i Y_i (f_\alpha(X_i) - f_{\alpha^*}(X_i))\right|\right)$$

$$\leq 4M\mathbb{E}\left(\max_{k=1,\ldots,m}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i Y_i \psi_k(X_i)\right|\right) .$$

By Bernstein's inequality (Bernstein (1924), Bennet (1962)), we know that for any $z > 0$,

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n \epsilon_i Y_i \psi_k(X_i)\right| \geq z\right) \leq 2\exp\left[-\frac{nz^2}{2z\|\psi\|_\infty + \|\psi\|_2^2}\right] .$$

Using the assumption that for all k, $\|\psi_k\|_2 \leq 1$ and $\|\psi_k\|_\infty \leq \sqrt{\frac{n}{\log n}}$ (see Condition D), we find for all $z \geq 1$,

$$\mathbb{P}\left(\max_{k=1,\ldots,m}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i Y_i \psi_k(X_i)\right| \geq 2z\sqrt{\frac{\log n}{n}}\right)$$

$$\leq 2n\exp[-\frac{4z\log n}{3}] \leq 2\exp[-\frac{x\log n}{3}] \leq \exp[-\frac{z}{3}].$$

It follows that

$$\mathbb{E}\mathbf{Z}_M \leq (1+3) \times 2 \times 4M\sqrt{\frac{\log n}{n}} = 32M\sqrt{\frac{\log n}{n}} .$$

□

Now, we show that for $\lambda_n = 104K^2\sqrt{\log n/n}$, the set Ω^* has probability at least $1 - c_0\exp[-K^2\log n/2]$ $(\geq 1 - c_0/n)$.

Lemma 5.5 *Suppose Conditions C and D are met. We have for a universal constant* c_0

$$\mathbb{P}\left(\sup_{f_\alpha \in \mathcal{F}}\left|\frac{\nu_n(f_\alpha) - \nu_n(f_{\alpha^*})}{I(\alpha - \alpha^*) + K\sqrt{\log n/n}}\right| > 104K^2\sqrt{\log n}\right) \leq c_0\exp[-\frac{K^2\log n}{2}] . \tag{34}$$

Proof. This follows from the peeling device, which is designed to establish bounds for weighted empirical process, as discussed in van de Geer (2000). We split up $\mathbf{R}^m$ into the sets

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

$$\begin{aligned} S_2 &= \{\alpha : \ K\sqrt{\log n/n} < I(\alpha - \alpha^*) \leq K\} \\ &\subseteq \cup_{j=0}^{j_0}\{\alpha : \ 2^{-(j+1)}K\sqrt{\log n/n} < I(\alpha - \alpha^*) \leq 2^{-j}K\} \end{aligned}$$

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

$$S_3 = \{\alpha : \ I(\alpha - \alpha^*) \geq K\} = \cup_{j=1}^{\infty}\{\alpha : \ 2^{j-1}K < I(\alpha - \alpha^*) \leq 2^j K\} .$$

The combination of Lemma 5.3 and Lemma 5.4 yields that for $K\sqrt{\log n/n} \leq M \leq K\sqrt{n/\log n}$,

$$\mathbb{P}\left(\mathbf{Z}_M \geq 52MK^2\sqrt{\frac{\log n}{n}}\right) \leq \exp[-(M^2 \vee K^2)\log n] . \tag{35}$$

We find on the set S_1,

$$\mathbb{P}\left(\sup_{f_\alpha\in\mathcal{F},\alpha\in S_1}\left|\frac{\nu_n(f_\alpha)-\nu_n(f_{\alpha^*})}{I(\alpha-\alpha^*)+K\sqrt{\log n/n}}\right|\geq 104K^2\sqrt{\log n}\right)\leq$$

$$\mathbb{P}\left(\sup_{f_\alpha\in\mathcal{F},\alpha\in S_1}|\nu_n(f_\alpha)-\nu_n(f_{\alpha^*})|\geq 104[K\sqrt{\frac{\log n}{n}}]K^2\sqrt{\log n}\right)$$

$$\leq \exp[-K^2\log n]\ .$$

Next we consider the set S_2. Take j_0 as the smallest integer such that $2^{-j_0}<K\sqrt{\log n/n}$. Then from (35),

$$\mathbb{P}\left(\sup_{f_\alpha\in\mathcal{F},\alpha\in S_2}\left|\frac{\nu_n(f_\alpha)-\nu_n(f_{\alpha^*})}{I(\alpha-\alpha^*)+2K\sqrt{\log n/n}}\right|\geq 104K^2\sqrt{\log n}\right)$$

$$\leq\sum_{j=0}^{j_0}\mathbb{P}\left(\sup_{f_\alpha\in\mathcal{F},\ I(\alpha-\alpha^*)\leq[2^{-j}K]}|\nu_n(f_\alpha)-\nu_n(f_{\alpha^*})|\geq 52[2^{-j}K]K^2\sqrt{\log n}\right)$$

$$\leq \log n\exp[-K^2\log n]\ .$$

Finally, we consider the set S_3. By Condition C, we know that for any $f_\alpha\in\mathcal{F}$, we have $I(\alpha-\alpha^*)\leq K\sqrt{n/\log n}$. Therefore,

$$\mathbb{P}\left(\sup_{f_\alpha\in\mathcal{F},\alpha\in S_3}\left|\frac{\nu_n(f_\alpha)-\nu_n(f_{\alpha^*})}{I(\alpha-\alpha^*)+K\sqrt{\log n/n}}\right|\geq 104K^2\sqrt{\log n}\right)$$

$$\leq\sum_{j=1}^{\infty}\mathbb{P}\left(\sup_{f_\alpha\in\mathcal{F},\ I(\alpha-\alpha^*)\leq[2^{j}K]}|\nu_n(f_\alpha)-\nu_n(f_{\alpha^*})|\geq 52[2^{j}K]K^2\sqrt{\log n}\right)$$

$$\leq\sum_{j=1}^{\infty}\exp[-2^{2j}K^2\log n]\ .$$

We conclude that

$$\mathbb{P}\left(\sup_{f_\alpha\in\mathcal{F}}\left|\frac{\nu_n(f_\alpha)-\nu_n(f_{\alpha^*})}{I(\alpha-\alpha^*)+K\sqrt{\log n/n}}\right|\geq 104K^2\sqrt{\log n}\right)$$

$$\leq\exp[-K^2\log n]+\log n\exp[-K^2\log n]+\sum_{j=1}^{\infty}\exp[-2^{2j}K^2\log n]$$

$$\leq c_0\exp[-\frac{K^2\log n}{2}]\ ,$$

for a universal constant c_0. □

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

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

because a distribution function is right-continuous. Let ϵ_n be defined as in (8), i.e.,

$$\epsilon_n=\inf\{\epsilon_n(f_{\alpha^*}):\ \alpha^*\in\mathcal{A}^*,\ f_{\alpha^*}\in\mathcal{F}\}$$

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

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

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

for some $\alpha_t^*\in\mathcal{A}^*$, $f_{\alpha_t^*}\in\mathcal{F}$, $t=1,2,\ldots$, and with $\epsilon_{n,t}\downarrow\epsilon_n$ as $t\to\infty$. Therefore, by Lemmas 5.1 - 5.5,

$$\mathbb{P}\left(R(\hat{f}_n)-R(f^*)>\epsilon_n\right)=\lim_{t\to\infty}\mathbb{P}\left(R(\hat{f}_n)-R(f^*)>\epsilon_{n,t}\right)$$

$$\leq c_0\exp[-\frac{K^2\log n}{2}]\ .$$

References

[1] Bartlett, P.L., Jordan, M.I. and McAuliffe, J.D. (2003). Convexity, classification and risk bounds. Techn. Report 638, University of California at Berkeley.

[2] Bennet, G. (1962). Probability inequalities for sums of independent random variables. *Journ. Amer. Statist. Acssoc.* **57** 33-45.

[3] Bernstein, S. (1924). Sur un modification de l'inégalité de Tchebichef. *Ann. Sci. Inst. Sav. Ukraine* Sect. Math. I (Russian, French summary).

[4] Boser, , B. Guyon, I. and Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers.*Fifth Annual Conf. on Comp. Learning Theory*, Pittsburgh ACM 142-152.

[5] 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 and Applied Math.* **LVII** 219-266

[6] Donoho, D.L. (1995). Denoising via soft-thresholding. *IEEE Transactions in Information Theory* **41** 613-627.

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

[8] Donoho, D.L. (2004a). For most large underdetermined systems of equations, the minimal ℓ^1-norm near-solution approximates the sparsest near-solution. Techn. Report, Stanford University.

[9] Donoho, D.L. (2004b). For most large underdetermined systems of linear equations, the minimal ℓ^1-norm solution is also the sparsest solution. Techn. Report, Stanford University.

[10] Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets, Approximation and Statistical Applications.* Lecture Notes in Statistics, vol. 129. Springer, New York, Berlin, Heidelberg.

[11] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical learning. Data Mining, Inference and Prediction.* Springer Verlag, New York,

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

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

[14] Koltchinskii, V. (2003) Local Rademacher complexities and oracle inequalities in risk minimization. Manuscript.

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

[16] Ledoux, M. (1996). *Talagrand deviation inequalities for product measures. ESIAM: Probab. Statist.* **1** 63-87. Available at: www.emath.fr/ps/

[17] Lin, Y. (2002). Support vector machines and the Bayes rule in classification. *Data mining knowledge and discovery* **6** 259-275.

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

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

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

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

[22] Schölkopf, B. and Smola, A. (2002). *Learning with Kernels*, MIT Press, Cambridge.

[23] Tibshirani, R. (1996). Regression analysis and selection via the LASSO. *Journal Royal Statist. Soc. B* **58** 267-288.

[24] Tsybakov, A.B. and van de Geer, S.A. (2003). Square root penalty: adaptation to the margin in classification and in edge estimation. Prépublication PMA-820, Lab. de Probab. et Modèles Aléatoires, Université Paris VII. To appear in *Ann. Statist.*

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

[26] van de Geer, S. (2000). *Empirical Processes in M-Estimation*, Cambridge Univ. Press.

[27] van de Geer, S. (2003). Adaptive quantile regression. In: *Recent Advances and Trends in Nonparametric Statistics* (Eds. M.G. Akritas and D.N. Politis), Elsevier, 235-250.

[28] Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag, New York.

[29] Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley, New York.

B. Tarigan
S.A. van de Geer
Mathematical Institute
University of Leiden
P.O. Box 9512
2300 RA Leiden
The netherlands
e-mail: bernadet and geer at math dot leidenuniv dot nl