COMPLEXITY REGULARIZATION VIA LOCALIZED RANDOM PENALTIES

GÁBOR LUGOSI* AND MARTEN WEGKAMP

Abstract. In this paper model selection via penalized empirical loss minimization in nonparametric classification problems is studied. Data-dependent penalties are constructed, which are based on estimates of the complexity of a small subclass of each model class, containing only those functions with small empirical loss. The penalties are novel since those considered in the literature are typically based on the entire model class. Oracle inequalities using these penalties are established, and the advantage of the new penalties over those based on the complexity of the whole model class is demonstrated.

1. Introduction.

In this paper we propose a new complexity-penalized model selection method based on data-dependent penalties. We consider a simple binary classification problem, though most ideas may be extended to a more general framework. Given a random observation $X \in \mathbb{R}^d$, one has to predict $Y \in \{0,1\}$. A *classifier* or *classification rule* is a function $f : \mathbb{R}^d \to \{0,1\}$, with *loss*

$$L(f) \stackrel{\text{def}}{=} \mathbb{P}\{f(X) \neq Y\}.$$

A sample $\mathcal{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ of n independent, identically distributed (i.i.d.) pairs are available. Each pair (X_i, Y_i) has the same distribution as (X, Y) and $\mathcal{D}_n$ is independent of (X, Y). The statistician's task is to select a classification rule f_n based on the data $\mathcal{D}_n$ such that the *probability of error*

$$L(f_n) = \mathbb{P}\{f_n(X) \neq Y \,|\, \mathcal{D}_n\}$$

is small. The Bayes classifier

$$f^*(x) \stackrel{\text{def}}{=} \mathbb{I}\left\{\mathbb{P}[Y = 1|X = x] \geq \mathbb{P}[Y = 0|X = x]\right\}$$

(where $\mathbb{I}$ denotes the indicator function) is the optimal rule as

$$L^* \stackrel{\text{def}}{=} \inf_{f:\mathbb{R}^d \to \{0,1\}} L(f) = L(f^*),$$

Date: March 3, 2003.

1991 *Mathematics Subject Classification.* Primary 62H30, 62G99; secondary 60E15.

Key words and phrases. Classification, complexity regularization, concentration inequalities, oracle inequalities, rademacher averages, random penalties, shatter coefficients.

*Supported by DGI grant BMF2000-0807.

but both f^* and L^* are unknown to the statistician. In this note we study classifiers $f : \mathbb{R}^d \to \{0,1\}$ which minimize the *empirical loss*

$$\widehat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{f(X_i) \neq Y_i\}$$

over a class of rules $\mathcal{F}$. For any $\widehat{f} \in \mathcal{F}$ minimizing the empirical probability of error, we have

$$\begin{aligned}
\mathbb{E}L(\widehat{f}) - L^* &= \mathbb{E}\widehat{L}(\widehat{f}) - L^* + \mathbb{E}(L - \widehat{L})(\widehat{f}) \\
&= \mathbb{E}\inf_{f\in\mathcal{F}} \widehat{L}(f) - L^* + \mathbb{E}(L - \widehat{L})(\widehat{f}) \\
&\leq \inf_{f\in\mathcal{F}} \mathbb{E}\widehat{L}(f) - L^* + \mathbb{E}(L - \widehat{L})(\widehat{f}) \\
&= \inf_{f\in\mathcal{F}} L(f) - L^* + \mathbb{E}(L - \widehat{L})(\widehat{f}).
\end{aligned}$$

Clearly, the *approximation error*

$$\inf_{f\in\mathcal{F}} L(f) - L^*$$

is decreasing as $\mathcal{F}$ becomes richer. However, the more complex $\mathcal{F}$, the more difficult the statistical problem becomes: the *estimation error*

$$\mathbb{E}(L - \widehat{L})(\widehat{f})$$

increases with the complexity of $\mathcal{F}$. In many approaches to the problem described above, one fixes in advance a sequence of model classes $\mathcal{F}_1, \mathcal{F}_2, \ldots$, whose union is $\mathcal{F}$. The problem of penalized model selection is to find a possibly data-dependent penalty $\widehat{C}_k$, assigned to each class $\mathcal{F}_k$, such that minimizing the penalized empirical loss

$$\widehat{L}(f) + \widehat{C}_k, \quad f \in \mathcal{F}_k, \quad k = 1, 2, \ldots$$

leads to a prediction rule $\widehat{f}$ with smallest possible loss. Denote by $\widehat{f}_k$ a function in $\mathcal{F}_k$ having minimal empirical loss and by $L_k^* = \inf_{f\in\mathcal{F}_k} L(f)$ the minimal loss in class $\mathcal{F}_k$.

The main idea is that since $\widehat{f}_k$ minimizes $\widehat{L}(f)$, we find, by the argument described above, that

$$\mathbb{E}L(\widehat{f}_k) - L^* \quad \leq \quad L_k^* - L^* + \mathbb{E}(L - \widehat{L})(\widehat{f}_k).$$

Our goal is to find the class $\mathcal{F}_k$ such that $L(\widehat{f}_k)$ is as small as possible. To this end, a good balance has to be found between the approximation and estimation errors. The approximation error is unknown to us, but the estimation error may be estimated. The key to complexity-regularized model selection is that a tight bound for the estimation error is a good penalty $\widehat{C}_k$. More precisely, we show in Lemma 2.1 below that if for some constant $\kappa > 0$

$$\mathbb{P}\left\{\widehat{C}_k \leq (L - \widehat{L})(\widehat{f}_k)\right\} \leq \frac{\kappa}{n^2k^2},$$

then the oracle inequality

$$\mathbb{E}L(\widehat{f}) - L^* \leq \inf_k \left(L_k^* - L^* + \mathbb{E}\widehat{C}_k \right) + 2\kappa n^{-2}$$

holds, and also, a similar bound,

$$L(\widehat{f}) - L^* \leq \inf_k \left(L_k^* - L^* + 2\widehat{C}_k \right) ,$$

holds with probability greater than $1 - 4\kappa n^{-2}$. This simple result shows that the penalty should be, with large probability, an upper bound on the estimation error, and to guarantee good performance, the bound should be as tight as possible.

Originally, distribution-free bounds, based on uniform-deviation inequalities, were proposed as penalties. For example, the structural risk minimization method of Vapnik and Chervonenkis [27] uses penalties of the form

$$\widehat{C}_k = \kappa \sqrt{\frac{\log \mathbb{S}_k(2n) + \log k}{n}} ,$$

where κ is a constant and $\mathbb{S}_k(2n)$ is the $2n$-maximal shatter coefficient of the class

$$\mathcal{A}_k = \{ \{x : f(x) = 1\},\ f \in \mathcal{F}_k \} ,$$

that is,

$$\begin{aligned} \mathbb{S}_k(2n) &= \max_{x_1,\ldots,x_{2n}} \left|\{ \{x_1, \cdots, x_{2n}\} \cap A,\ A \in \mathcal{A}_k\}\right| \\ &= \max_{x_1,\ldots,x_{2n}} \left|\{ (f(x_1), \cdots, f(x_{2n})),\ f \in \mathcal{F}_k\}\right| , \end{aligned} \tag{1.1}$$

see for example, Vapnik [26], Devroye, Györfi, and Lugosi [9]. The fact that this type of penalty works follows from the Vapnik-Chervonenkis inequality. Such distribution-free bounds are attractive because of their simplicity, but precisely because of their distribution-free nature, they are necessarily loose in many cases.

Recently various attempts have been made to define the penalties in a data-dependent way to achieve this goal, see, for example, Bartlett, Boucheron, and Lugosi [2], Koltchinskii [11], Koltchinskii and Panchenko [13], Lozano [15], Lugosi and Nobel [17], Massart [19], and Shawe-Taylor, Bartlett, Williamson, and Anthony [22].

For example, in [11] and [2], random complexity penalties based on *Rademacher averages* were proposed and investigated. Rademacher averages are defined as

$$\widehat{R}_{\mathcal{F}_k} = \mathbb{E}\left[\sup_{f \in \mathcal{F}_k} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathbb{I}\{f(X_i) \neq Y_i\} \middle| \mathcal{D}_n \right] ,$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. symmetric $\{-1,1\}$-valued random variables, independent of $\mathcal{D}_n$. The reason why this penalty was introduced is based on the fact that

$$\mathbb{E} \sup_{f \in \mathcal{F}_k} (L - \widehat{L})(f) \asymp \mathbb{E}\widehat{R}_{\mathcal{F}_k}$$

(see, e.g., Van der Vaart and Wellner [25]), and since $\widehat{R}_{\mathcal{F}_k}$ can be shown to be sharply concentrated around its mean. In fact, concentration inequalities have been a key tool in the analysis of data-based penalties (see Massart [19]) and this paper relies heavily on some recent concentration results.

The model selection method based on Rademacher complexities satisfies an oracle inequality of the rough form

$$\mathbb{E}L(\widehat{f}) - L^* \leq \inf_k \left[L_k^* - L^* + \kappa_1 \mathbb{E}\widehat{R}_{\mathcal{F}_k} + \kappa_2 \sqrt{\frac{\log k}{n}} \right] \tag{1.2}$$

(see [2] and [11]) for values of the constants $\kappa_1, \kappa_2 > 0$. The advantage of this bound over the one obtained by the distribution-free penalties mentioned above may perhaps be better understood if we further bound

$$\mathbb{E}\widehat{R}_{\mathcal{F}_k} \leq \sqrt{\frac{\mathbb{E}\log 2\mathbb{S}_k(X_1^n)}{2n}}$$

where

$$\begin{aligned} \mathbb{S}_k(X_1^n) &= |\{\{X_1, \ldots, X_n\} \cap A : A = \{x : f(x) = 1\}, f \in \mathcal{F}_k\}| \\ &= |\{(f(X_1), \ldots, f(X_n)),\ f \in \mathcal{F}_k\}|, \end{aligned} \tag{1.3}$$

is the *random shatter coefficient* of the class $\widehat{\mathcal{F}}_k$, which obviously never exceeds the worst-case shatter coefficient $\mathbb{S}_k(n)$ and may be significantly smaller for certain distributions.

However, this improved penalty is still not completely satisfactory. To see this, recall that by a classical result of Vapnik and Chervonenkis, for any index k,

$$\mathbb{E}L(\widehat{f}_k) - L_k^* \leq c \left(\sqrt{\frac{L_k^* \cdot \mathbb{E}\log \mathbb{S}_k(X_1^n)}{n}} + \frac{\mathbb{E}\log \mathbb{S}_k(X_1^n)}{n} \right) \tag{1.4}$$

which is much smaller than the corresponding expected Rademacher average if L_k^* is small. (For explicit constants we refer to Theorem 1.14 in Lugosi [16].) Since in typical classification problems the minimal error L_k^* in class $\mathcal{F}_k$ is often very small for some k, it is important to find penalties which allow to derive oracle inequalities with the appropriate dependence on L_k^*. In particular, a desirable goal would be to develop classifiers $\widehat{f}$ for which an oracle inequality resembling

$$\mathbb{E}L(\widehat{f}) - L^* \leq \inf_k \left\{ L_k^* - L^* + \kappa_1 \sqrt{\frac{L_k^* \cdot \mathbb{E}\log \mathbb{S}_k(X_1^n)}{n}} + \kappa_2 \frac{\mathbb{E}\log \mathbb{S}_k(X_1^n)}{n} \right\}$$

holds for all distributions. The main results of this note (Theorems 4.1 and 4.2) show that estimates of the desired property are indeed possible to construct in a conceptually simple way.

By the key Lemma 2.1, it suffices to find a data-dependent upper estimate of $(L-\widehat{L})(\widehat{f}_k)$ which has the order of magnitude of the above upper bound. The difficulty is that L_k^* and $\mathbb{E}\log \mathbb{S}_k(X_1^n)$ both depend on the underlying distribution.

The improvement is achieved by decreasing the penalties so that the supremum in the definition of the Rademacher average is not taken over the whole class $\mathcal{F}_k$ but rather over a small subclass $\widehat{\mathcal{F}}_k$ containing only functions which "look good" on the data. More precisely, define the random subclass $\widehat{\mathcal{F}}_k \subset \mathcal{F}_k$ by

$$\widehat{\mathcal{F}}_k = \left\{ f \in \mathcal{F}_k : \widehat{L}(f) \le \kappa_1 \widehat{L}(\widehat{f}_k) + \kappa_2 n^{-1} \log \mathbb{S}_k(X_1^n) + \kappa_3 n^{-1} \log(nk) \right\}$$

for some non-negative constants κ_1, κ_2 and κ_3.

Risk estimates based on localized Rademacher averages have been considered in several recent works. The most closely related procedure is proposed by Koltchinskii and Panchenko [12], who, assuming $\inf_{f\in\mathcal{F}} L(f) = 0$, compute the Rademacher averages of subclasses of $\mathcal{F}$ with empirical loss less than r for different values of r obtained by a recursive procedure, and obtain bounds for the loss of the empirical risk minimizer in terms of the localized Rademacher averages obtained after a certain number of iterations. Our approach of bounding the loss is conceptually simpler: it suffices to compute the Rademacher complexities at only one scale which depends on the smallest empirical loss in the class and a term of a smaller order determined by the shatter coefficients of the whole class. Thus, we use "global" information to determine the scale of localization. Bartlett, Bousquet, and Mendelson [3] also derive closely related generalization bounds, based on localized Rademacher averages. In their approach the performance bounds also depend on Rademacher averages computed at different scales of localization, which are combined by the technique of peeling. For further recent related work we also refer to Bousquet [7], Bousquet, Koltchinskii, and Panchenko [8], and Tsybakov [24].

The rest of the paper is organized as follows. Section 2 presents some basic inequalities on model selection, which generalizes some of the results in Bartlett, Boucheron, and Lugosi [2]. Section 3 proposes a simple but suboptimal penalty which already has some of the main features of the penalties presented in Section 4. It shows, in a transparent way, some of the underlying ideas of the main results. Section 4 introduces a new penalty based on the Rademacher average $\widehat{R}_{\widehat{\mathcal{F}}_k}$ and it is shown that the new estimate yields an improvement of the desired form.

2. Preliminaries

In this section we present two basic auxiliary lemmata on model selection. The first lemma is general in the sense that it does not depend on the particular choice of the penalty $\widehat{C}_k$. This result was mentioned in the introduction and generalizes a result obtained by Bartlett, Boucheron, and Lugosi [2].

Lemma 2.1. *Suppose that the random variables $\widehat{C}_1, \widehat{C}_2, \ldots$ are such that*

$$\mathbb{P}\left\{\widehat{C}_k \leq (L-\widehat{L})(\widehat{f}_k)\right\} \leq \frac{\kappa}{n^2k^2}$$

for some $\kappa > 0$ and for all k. Then we have

$$\mathbb{E}L(\widehat{f}) - L^* \leq \inf_k \left[L_k^* - L^* + \mathbb{E}\widehat{C}_k\right] + \frac{2\kappa}{n^2}.$$

It is clear that we can always take $\widehat{C}_k \leq 1$.

Proof. Observe that

$$\begin{aligned}
\mathbb{E}\sup_k \left\{(L-\widehat{L})(\widehat{f}_k) - \widehat{C}_k\right\} &\leq \mathbb{P}\left\{\sup_k \left[(L-\widehat{L})(\widehat{f}_k) - \widehat{C}_k\right] \geq 0\right\} \\
&\qquad \left(\text{since } \sup_k \left[(L-\widehat{L})(\widehat{f}_k) - \widehat{C}_k\right] \leq 1\right) \\
&\leq \sum_{k=1}^{\infty} \mathbb{P}\left\{(L-\widehat{L})(\widehat{f}_k) - \widehat{C}_k \geq 0\right\} \\
&\qquad \text{(by the union bound)} \\
&\leq \sum_{k=1}^{\infty} \frac{\kappa}{n^2k^2} \\
&\qquad \text{(by assumption)} \\
&\leq \frac{2\kappa}{n^2}.
\end{aligned}$$

Therefore, we may conclude that

$$\begin{aligned}
\mathbb{E}L(\widehat{f}) - L^* &= \mathbb{E}\left[\widehat{L}(\widehat{f}) - L^* + \widehat{C}_{\widehat{k}}\right] + \mathbb{E}\left[(L-\widehat{L})(\widehat{f}) - \widehat{C}_{\widehat{k}}\right] \\
&\quad \text{(where } \widehat{k} \text{ is the selected model index, that is, } \widehat{f} = \widehat{f}_{\widehat{k}}) \\
&\leq \mathbb{E}\inf_k \left[\widehat{L}(\widehat{f}_k) - L^* + \widehat{C}_k\right] + \mathbb{E}\left[(L-\widehat{L})(\widehat{f}) - \widehat{C}_{\widehat{k}}\right] \\
&\quad \text{(by definition of } \widehat{f}) \\
&\leq \mathbb{E}\inf_k \left[\inf_{f\in\mathcal{F}_k} \widehat{L}(f) - L^* + \widehat{C}_k\right] + \mathbb{E}\sup_k \left[(L-\widehat{L})(\widehat{f}_k) - \widehat{C}_k\right] \\
&\quad \text{(by definition of } \widehat{f}_k) \\
&\leq \inf_k \left[\inf_{f\in\mathcal{F}_k} L(f) - L^* + \mathbb{E}\widehat{C}_k\right] + \mathbb{E}\sup_k \left[(L-\widehat{L})(\widehat{f}_k) - \widehat{C}_k\right] \\
&\quad \text{(interchange } \mathbb{E} \text{ and inf)} \\
&\leq \inf_k \left[L_k^* - L^* + \mathbb{E}\widehat{C}_k\right] + \frac{2\kappa}{n^2} \\
&\quad \text{(by the preceding display)}
\end{aligned}$$

and the proof is complete. □

The preceding result is not entirely satisfactory on the following ground. Although it presents a sharp bound, it is a bound for the *average* risk behavior of $\widehat{f}$. However, the penalty is computed on the data at hand, and therefore the proposed criterion should have optimal performance for (almost) all possible sequences of the data. The following result presents a nonasymptotic oracle inequality which holds with large probability and an asymptotic almost sure version.

Lemma 2.2. *Assume that for all $k, n \geq 1$,*

$$\mathbb{P}\left\{\widehat{C}_k \leq (L-\widehat{L})(\widehat{f}_k)\right\} \leq \frac{\kappa}{n^2k^2} \quad \textit{and} \quad \mathbb{P}\left\{\widehat{C}_k \leq (\widehat{L}-L)(f_k^*)\right\} \leq \frac{\kappa}{n^2k^2}.$$

Then for all $n \geq 1$ we have

$$\mathbb{P}\left[L(\widehat{f}) - L^* \leq \inf_k \left(L_k^* - L^* + 2\widehat{C}_k\right)\right] \leq \frac{4\kappa}{n^2}$$

and the asymptotic almost sure bound

$$\mathbb{P}\left[\liminf_{n\to\infty}\left\{L(\widehat{f}) - L^* \leq \inf_k \left(L_k^* - L^* + 2\widehat{C}_k\right)\right\}\right] = 1.$$

Proof. Let $\widehat{k}$ be the selected model index. Notice that

$$\begin{aligned}
L(\widehat{f}) &= \widehat{L}(\widehat{f}) + \widehat{C}_{\widehat{k}} + (L-\widehat{L})(\widehat{f}) - \widehat{C}_{\widehat{k}} \\
&\leq \inf_k \left[\widehat{L}(\widehat{f}_k) + \widehat{C}_k\right] + \sup_k \left[(L-\widehat{L})(\widehat{f}_k) - \widehat{C}_k\right] \\
&\leq \inf_k \left[\widehat{L}(f_k^*) + \widehat{C}_k\right] + \sup_k \left[(L-\widehat{L})(\widehat{f}_k) - \widehat{C}_k\right] \\
&\leq \inf_k \left[L_k^* + 2\widehat{C}_k\right] + \sup_k \left[(\widehat{L}-L)(f_k^*) - \widehat{C}_k\right] + \sup_k \left[(L-\widehat{L})(\widehat{f}_k) - \widehat{C}_k\right].
\end{aligned}$$

By assumption, the last two terms on the right-hand side satisfy

$$\mathbb{P}\left[\sup_k \left[(\widehat{L}-L)(f_k^*) - \widehat{C}_k\right] + \sup_k \left[(L-\widehat{L})(\widehat{f}_k) - \widehat{C}_k\right] \geq 0\right] \leq \sum_{k=1}^{\infty} \frac{2\kappa}{n^2k^2} < \frac{4\kappa}{n^2} ,$$

proving the first inequality. The almost sure statement is a direct consequence of the Borel-Cantelli lemma. □

3. A simple version

The purpose of this short section is to offer a simplified, yet suggestive illustration of the ideas. As discussed in the introduction, an ideal penalty would be a tight upper bound for the expression on the right-hand side of (1.4). Motivated by this bound, we propose the simple penalty

$$\widehat{C}_k = 2\sqrt{2\widehat{L}(\widehat{f}_k) + 8\frac{\log \mathbb{S}_k(2n) + 2\log(nk)}{n}} \cdot \sqrt{\frac{\log \mathbb{S}_k(2n)}{n} + 2\frac{\log(nk)}{n}} ,$$

where $\mathbb{S}_k(2n)$ is the (worst-case) $2n$-shatter coefficient defined in (1.1). Thus, the minimal loss L_k^* in class $\mathcal{F}_k$ is estimated by its natural empirical counterpart $\widehat{L}(\widehat{f}_k) = \inf_{f\in\mathcal{F}_k} \widehat{L}(f)$ and the expected logarithmic shatter coefficient $\mathbb{E}\log \mathbb{S}_k(X_1^n)$ is estimated by the distribution-free upper bound $\log \mathbb{S}_k(2n)$. (This term may be bounded further by $V_k \log(2n+1)$, where V_k is the VC-dimension of the set $\mathcal{A}_k$). The auxiliary terms $n^{-1}\log(nk)$ are necessary to derive the desired oracle inequalities. The next theorem shows that the proposed penalty indeed works.

Theorem 3.1. *Consider the penalized empirical loss minimizer $\widehat{f}$ with the data-based penalty $\widehat{C}_k$ defined above. Then for every n and for all distributions of (X,Y),*

$$\mathbb{E}L(\widehat{f}) - L^* \leq \inf_k \left(L_k^* - L^* + \mathbb{E}\widehat{C}_k\right) + \frac{16}{n^2}.$$

In particular,

$$\mathbb{E}L(\widehat{f}) - L^* \leq \inf_k \left[L_k^* - L^* + 4\sqrt{L_k^* + \frac{2}{n}\{\log \mathbb{S}_k(2n) + 2\log(nk)\}} \cdot \sqrt{\frac{\log \mathbb{S}_k(2n)}{n} + 2\frac{\log(nk)}{n}} \right] + \frac{16}{n^2} .$$

The proof uses Lemma 2.1 and the following uniform deviation bound due to Vapnik and Chervonenkis [27]. (The slighly improved form used here is proved by Anthony, and Shawe-Taylor [1].)

Proposition 3.2. *Let $\mathbb{S}_k(X_1^{2n})$ be the random shatter coefficient of $\mathcal{A}_k$ based on i.i.d. observations $X_1, \cdots, X_{2n}$ defined in (1.3). For all $\varepsilon > 0$ and $n \geq 1$,*

$$\mathbb{P}\left\{ \sup_{f\in\mathcal{F}} L(f) - 2\widehat{L}(f) \geq 2\varepsilon \right\} \leq 4\mathbb{E}\mathbb{S}_k(X_1^{2n}) \exp(-n\varepsilon/4) \tag{3.1}$$

and

$$\mathbb{P}\left\{ \sup_{f\in\mathcal{F}} \widehat{L}(f) - 2L(f) \geq 2\varepsilon \right\} \leq 4\mathbb{E}\mathbb{S}_k(X_1^{2n}) \exp(-n\varepsilon/4). \tag{3.2}$$

Proof. Observe that for all $\varepsilon > 0$ and $n \geq 1$,

$$\left\{ \sup_{f\in\mathcal{F}} L(f) - 2\widehat{L}(f) \geq 2\varepsilon \right\} \subseteq \left\{ \sup_{f\in\mathcal{F}} \frac{L(f) - \widehat{L}(f)}{\sqrt{L(f)}} \geq \sqrt{\varepsilon} \right\}$$

and similarly,

$$\left\{ \sup_{f\in\mathcal{F}} \widehat{L}(f) - 2L(f) \geq 2\varepsilon \right\} \subseteq \left\{ \sup_{f\in\mathcal{F}} \frac{\widehat{L}(f) - L(f)}{\sqrt{\widehat{L}(f)}} \geq \sqrt{\varepsilon} \right\} .$$

The proposition follows by Anthony, and Shawe-Taylor [1]. □

Proof of Theorem 3.1. We start with the proof of the first inequality of Theorem 3.1. In view of Lemma 2.1, it suffices to show that

$$\mathbb{P}\{L(\widehat{f}_k) - \widehat{L}(\widehat{f}_k) \geq \widehat{C}_k\} \leq 8/(nk)^2.$$

Consequently, by (3.1),

$$\begin{aligned}
&\mathbb{P}\left\{ 2\widehat{L}(\widehat{f}_k) + 8\frac{\log \mathbb{S}_k(2n)}{n} + 16\frac{\log(nk)}{n} \leq L(\widehat{f}_k) \right\} \\
&= \mathbb{P}\left\{ L(\widehat{f}_k) - 2\widehat{L}(\widehat{f}_k) \geq 8\frac{\log \mathbb{S}_k(2n)}{n} + 16\frac{\log(nk)}{n} \right\} \\
&\leq 4\mathbb{S}_k(2n) \exp\left\{ -\frac{n}{8}\left(8\frac{\log \mathbb{S}_k(2n)}{n} + 16\frac{\log(nk)}{n} \right) \right\} \\
&= \frac{4}{n^2k^2},
\end{aligned}$$

so that

$$\mathbb{P}\{\widehat{C}_k \geq \widetilde{C}_k\} \geq 1 - 4/(nk)^2,$$

where

$$\widetilde{C}_k = 2\sqrt{L(\widehat{f}_k)} \cdot \sqrt{\frac{\log \mathbb{S}_k(2n)}{n} + 2\frac{\log(nk)}{n}}.$$

Another application of inequality (3.1) yields

$$\begin{aligned}
&\mathbb{P}\{L(\widehat{f}_k) - \widehat{L}(\widehat{f}_k) \geq \widehat{C}_k\} \\
&\leq \mathbb{P}\{L(\widehat{f}_k) - \widehat{L}(\widehat{f}_k) \geq \widetilde{C}_k\} + \frac{4}{(nk)^2} \\
&\leq 4\mathbb{S}_k(2n)\exp\left\{-\frac{n}{4}\cdot 4\left(\frac{\log \mathbb{S}_k(2n)}{n} + 2\frac{\log(nk)}{n}\right)\right\} + \frac{4}{(nk)^2} \\
&= \frac{8}{(nk)^2}.
\end{aligned}$$

Conclude via Lemma 2.1 that

$$\mathbb{E}L(\widehat{f}) \leq \min_k \left(L_k + \mathbb{E}\widehat{C}_k\right) + \frac{16}{n^2}.$$

For the second inequality, deduce that for all $\delta > 0$,

$$\mathbb{E}\sqrt{\widehat{L}(\widehat{f}_k) + \delta} \leq \sqrt{\mathbb{E}\widehat{L}(\widehat{f}_k) + \delta} \leq \sqrt{\mathbb{E}\inf_{f \in \mathcal{F}_k} \widehat{L}(f) + \delta} \leq \sqrt{L_k^* + \delta}.$$

by Jensen's inequality and the definition of $\widehat{f}_k$. □

The bound of Theorem 3.1 has the right dependence on L_k^* as suggested by inequality (1.4) mentioned in the introduction. In particular, if L_k^* happens to equal to zero for some class $\mathcal{F}_k$, then the upper bound has an improved rate of convergence. The disadvantage of the simple penalty defined above is that instead of the expected shatter coefficients, a distribution-free (and therefore suboptimal) upper bound appears for each class $\mathcal{F}_k$.

Recently, Boucheron, Lugosi and Massart [4] proved that $\log \mathbb{S}_k(X_1^n)$ concentrates sharply around its mean. For example, we have the following inequalities:

Proposition 3.3. *For all $\varepsilon > 0$, $n \geq 1$,*

$$\begin{aligned}
\mathbb{P}\left[\mathbb{E}\log \mathbb{S}_k(X_1^n) > 2\log \mathbb{S}_k(X_1^n) + 2\varepsilon\right] &\leq e^{-\varepsilon}, \\
\mathbb{P}\left[\log \mathbb{S}_k(X_1^n) > 2\mathbb{E}\log \mathbb{S}_k(X_1^n) + 2\varepsilon\right] &\leq e^{-\varepsilon}.
\end{aligned}$$

Moreover, for each $n \geq 1$,

$$\mathbb{E}\log \mathbb{S}_k(X_1^n) \leq \log \mathbb{E}\mathbb{S}_k(X_1^n) \leq \frac{1}{\ln 2}\mathbb{E}\log \mathbb{S}_k(X_1^n) \leq 2\mathbb{E}\log \mathbb{S}_k(X_1^n) .$$

This proposition implies that the expected random log shatter coefficients $\mathbb{E}\log \mathbb{S}_k(X_1^n)$ of $\mathcal{F}_k$ may be replaced by a constant times $\log \mathbb{S}_k(X_1^n)$ and vice versa. Hence we may replace the distribution-free bounds $\log \mathbb{S}_k(2n)$ by empirical estimates $\log \mathbb{S}_k(X_1^n)$, at the price of slightly worse constants. The main oracle inequalities in Section 4 are accompanied by asymptotic almost-sure versions of bounds for the expected value. Such bounds are easy to obtain as well, simply by invoking Lemma 2.2 instead of Lemma 2.1. The details are omitted here.

4. Rademacher penalties

The main results of the paper are presented in this section. Assign, to each model class $\mathcal{F}_k$,

$$\widehat{u}_k = 16\frac{4\log \mathbb{S}_k(X_1^n) + 9\log(nk)}{n} \ , \tag{4.1}$$

with $\mathbb{S}_k(X_1^n)$ defined in (1.3), and the class

$$\widehat{\mathcal{F}}_k = \left\{ f \in \mathcal{F}_k : \ \widehat{L}(f) \le 16\widehat{L}(\widehat{f}_k) + 15\widehat{u}_k \right\} \ . \tag{4.2}$$

Observe that the class $\widehat{\mathcal{F}}_k$ contains only those classifiers whose empirical loss is not much larger than that of the empirical minimizer. Note that the constant 16 has no special role, it has been chosen by convenience. Any constant larger than one would lead to similar results, at the price of modifying other constants. The term $\widehat{u}_k$ depends on the shatter coefficient of the whole class $\mathcal{F}_k$ but it is typically small compared to $\widehat{L}(\widehat{f}_k)$.

The penalty is calculated in terms of the Rademacher average of this smaller class. More precisely, define the complexity estimate by

$$\widehat{C}_k = \left(8\widehat{R}_{\widehat{\mathcal{F}}_k} + 20n^{-1}\log(nk) + 2\sqrt{n^{-1}\log(nk)} \cdot \sqrt{8\widehat{L}(\widehat{f}_k) + 7\widehat{u}_k} \right) \wedge 1. \tag{4.3}$$

Again, not too much attention should be paid to the values of the constants involved. We favored simple readable proofs over optimal constants. Note that, through $\mathbb{S}_k(X_1^n)$, the penalty also depends on the random shatter coefficient of the whole class $\mathcal{F}_k$. However, the term involving the shatter coefficient of the entire class $\mathcal{F}_k$,

$$n^{-1}\sqrt{\log(nk) \cdot \log \mathbb{S}_k(X_1^n)},$$

is typically much smaller (by a factor $n^{-1/2}$) than the Rademacher average of the whole class $\mathcal{F}_k$. (For instance, see inequality (4.8) and Proposition 4.6 below.)

We have the following performance bound for the expected loss of the minimizer $\widehat{f}$ of the penalized empirical loss $\widehat{L}(\widehat{f}_k) + \widehat{C}_k$.

Theorem 4.1. *For every n,*

$$\mathbb{E}L(\widehat{f}) - L^* \le \inf_k \left(L_k^* - L^* + \mathbb{E}\widehat{C}_k \right) + \frac{22}{n^2}.$$

In addition, with probability greater than $1 - 44/n^2$,

$$L(\widehat{f}) - L^* \le \inf_k (L_k^* - L^* + 2\widehat{C}_k)$$

and also

$$\mathbb{P}\left[\liminf_{n\to\infty} \left\{ L(\widehat{f}) - L^* \le \inf_k (L_k^* - L^* + 2\widehat{C}_k) \right\} \right] = 1 \ .$$

The next theorem is here to point out that the bound above is indeed a significant improvement over bounds of the type (1.2), and that the dependence on the minimal loss L_k^* and the random shatter coefficient has the form suggested by (1.4). For this purpose, we introduce

$$\overline{u}_k = 16 \frac{8\mathbb{E}\log \mathbb{S}_k(X_1^n) + 17\log(nk)}{n} \tag{4.4}$$

and the class

$$\overline{\mathcal{F}}_k = \{ f \in \mathcal{F}_k : \ L(f) \le 64L_k^* + 63\overline{u}_k \} \ .$$

We also set

$$\varepsilon_k = 2n^{-1}\log(nk) \ .$$

Theorem 4.2. *The following oracle inequality holds*

$$\mathbb{E}L(\widehat{f}) - L^* \quad \le \quad \min_{k\ge 1} \left[L_k^* - L^* + 8\mathbb{E}\widehat{R}_{\overline{\mathcal{F}}_k} + 15\varepsilon_k + 16\sqrt{L_k^* + \overline{u}_k} \cdot \sqrt{2\varepsilon_k} \ \right] + 22n^{-2}.$$

In particular, there exists universal constants κ_1 and κ_2 such that

$$\begin{aligned} &\mathbb{E}L(\widehat{f}) - L^* \\ &\le \ \inf_k \left\{ L_k^* - L^* + \kappa_1 \sqrt{\frac{L_k^* \cdot (\mathbb{E}\log \mathbb{S}_k(X_1^n) \vee \log(nk))}{n}} + \kappa_2 \frac{\mathbb{E}\log \mathbb{S}_k(X_1^n) \vee \log(nk)}{n} \right\} \ . \end{aligned}$$

This oracle inequality has the desired form outlined in the introduction and improves upon the results of [2] and [13]. For example, in the special case when $L_k^* = 0$ for $k \ge k_0$, we obtain, for some numerical constants c_1 and c_2,

$$\mathbb{E}L(\widehat{f}) \ \le \min_{k\ge k_0} c_1 \frac{\mathbb{E}\log \mathbb{S}_k(X_1^n) \vee \log(nk)}{n} + \frac{c_2}{n^2}.$$

which is of a different order of magnitude from the penalties considered by [2] and [13]. Theorem 4.2 is only stated for the expected loss but an inequality which holds with "large" probability may be obtained just as in Theorem 4.1.

Proofs of Theorems 4.1 and 4.2. First, recall the definitions of $\widehat{u}_k$ and $\overline{u}_k$ in (4.1) and (4.4), respectively, and in addition, define

$$u_k = 8\frac{2\log \mathbb{E}\mathbb{S}_k(X_1^n) + 2\log(nk)}{n}$$

and the event

$$B_k \stackrel{\text{def}}{=} \{u_k \le \widehat{u}_k \le \overline{u}_k\} \ .$$

Observe Proposition 3.3 above yields that, with probability at least $1 - 1/(nk)^2$,

$$\begin{aligned}
u_k &= \frac{16}{n}\left\{\log \mathbb{E}\mathbb{S}_k(X_1^n) + \log(nk)\right\} \\
&\le \frac{16}{n}\left\{2\mathbb{E}\log \mathbb{S}_k(X_1^n) + \log(nk)\right\} \\
&\le \frac{16}{n}\left\{2[2\log \mathbb{S}_k(X_1^n) + 4\log(nk)] + \log(nk)\right\} \\
&= \widehat{u}_k \\
&\le \frac{16}{n}\left\{4[2\mathbb{E}\log \mathbb{S}_k(X_1^n) + 4\log(nk)] + 9\log(nk)\right\} \\
&= \frac{16}{n}\left\{8\mathbb{E}\log \mathbb{S}_k(X_1^n) + 17\log(nk)\right\} = \overline{u}_k
\end{aligned}$$

and therefore

$$\mathbb{P}B_k^c \le (nk)^{-2}. \tag{4.5}$$

Finally, we introduce the event

$$A_k = \left\{\sup_{f\in\mathcal{F}_k} L(f) - 2\widehat{L}(f) \le u_k\right\} \cap \left\{\sup_{f\in\mathcal{F}_k} \widehat{L}(f) - 2L(f) \le u_k\right\}$$

and the class

$$\mathcal{F}_k^* = \{f \in \mathcal{F}_k : \ L(f) \le 4L_k^* + 3u_k\} \,.$$

The following intermediate result will be useful in the proofs of both theorems.

Lemma 4.3. *We have*

$$\mathbb{P}\{A_k \cap B_k\} \ge 1 - \frac{9}{(nk)^2}, \tag{4.6}$$

and on the set $A_k \cap B_k$, the following holds:

(i) $\widehat{f}_k \in \mathcal{F}_k^*$.
(ii) $\mathcal{F}_k^* \subseteq \widehat{\mathcal{F}}_k$, *and in particular,* $\widehat{R}_{\mathcal{F}_k^*} \le \widehat{R}_{\widehat{\mathcal{F}}_k}$.
(iii) $L_k^* \le 2\widehat{L}(\widehat{f}_k) + u_k$.

Proof. To begin with, notice that

$$\mathbb{E}\mathbb{S}_k(X_1^{2n}) \le \mathbb{E}\mathbb{S}_k(X_1^n)\mathbb{S}_k(X_{n+1}^{2n}) = \mathbb{E}^2\mathbb{S}_k(X_1^n)$$

by the definition of the shatter coefficient and by the independence of the X_i. Thus, by Proposition 3.2,

$$\mathbb{P}A_k^c \;\le\; 8\mathbb{E}\mathbb{S}_k(X_1^{2n})\exp\left(-\frac{nu_k}{8}\right) \le \frac{8}{n^2k^2}.$$

This bound and (4.5) imply assertion (4.6). To prove claim (i), observe that on A_k,

$$\begin{aligned}
L(\widehat{f}_k) &\le 2\widehat{L}(\widehat{f}_k) + u_k \\
&\qquad \text{(by definition of } A_k\text{)} \\
&\le 2\widehat{L}(f_k^*) + u_k \\
&\qquad \text{(by definition of } \widehat{f}_k\text{)} \\
&\le 2\,(2L_k^* + u_k) + u_k \\
&\qquad \text{(by definition of } A_k\text{)} \\
&= 4L_k^* + 3u_k.
\end{aligned}$$

For claim (ii), notice that for any $f \in \mathcal{F}_k^*$,

$$\begin{aligned}
\widehat{L}(f) &\le 2L(f) + u_k \\
&\qquad \text{(by definition of } A_k\text{)} \\
&\le 2\,[4L_k^* + 3u_k] + u_k \\
&\qquad \text{(by definition of } \mathcal{F}_k^*\text{)} \\
&= 8L_k^* + 7u_k \\
&\le 8L(\widehat{f}_k) + 7u_k \\
&\qquad \text{(by definition of } L_k^*\text{)} \\
&\le 16\widehat{L}(\widehat{f}_k) + 15u_k \\
&\qquad \text{(by definition of } A_k\text{)} \\
&\le 16\widehat{L}(\widehat{f}_k) + 15\widehat{u}_k \\
&\qquad \text{(by definition of } B_k\text{)}\ .
\end{aligned}$$

Claim (ii) now follows. Claim (iii) is immediate from the definition of A_k since both $\widehat{f}_k$ and f_k^* belong to $\mathcal{F}_k$. □

Next we link the Rademacher average $\widehat{R}_{\mathcal{F}_k^*}$ to $\mathbb{E}\sup_{f\in\mathcal{F}_k^*}|L(f) - \widehat{L}(f)|$. By a classical symmetrization device (cf. Giné and Zinn [10] or Van der Vaart and Wellner [25])

$$\mathbb{E}\sup_{f\in\mathcal{F}_k^*}\left|\widehat{L}(f) - L(f)\right| \le 2\mathbb{E}\widehat{R}_{\mathcal{F}_k^*}\ . \tag{4.7}$$

Also, $\widehat{R}_{\mathcal{F}_k^*}$ is known to concentrate sharply around its mean. For example, we have, by results of Boucheron, Lugosi, and Massart [4], [5], the following bounds.

Proposition 4.4. *For all $\epsilon > 0$, $n \geq 1$,*

$$\mathbb{P}\left[\widehat{R}_{\mathcal{F}_k} \geq 2\mathbb{E}\widehat{R}_{\mathcal{F}_k} + \epsilon\right] \leq e^{-6n\epsilon/5} \quad \text{and} \quad \mathbb{P}\left[\widehat{R}_{\mathcal{F}_k} \leq \frac{1}{2}\mathbb{E}\widehat{R}_{\mathcal{F}_k} - \epsilon\right] \leq e^{-n\epsilon}.$$

Proof. Define $Z \stackrel{\text{def}}{=} n\widehat{R}_{\mathcal{F}_k}$, then it follows from Boucheron, Lugosi, and Massart [4] that

$$\log \mathbb{E}\exp(\lambda(Z - \mathbb{E}Z)) \leq \mathbb{E}Z(e^{\lambda} - 1 - \lambda),$$

which implies further that for $0 \leq \lambda < 3$

$$\log \mathbb{E}\exp(\lambda(Z - \mathbb{E}Z)) \leq \frac{\lambda \mathbb{E}Z}{2(1 - \lambda/3)}.$$

After an application of Markov's inequality we find

$$\mathbb{P}\left[Z \geq \mathbb{E}Z + \sqrt{2\mathbb{E}Zx} + x/3\right] \leq e^{-x}.$$

We obtain the desired upper-tail bound by inserting $Z = n\widehat{R}_{\mathcal{F}_k}$ in the preceding display and invoking the inequality $2\sqrt{xy} \leq x + y$. The bound for the lower tail follows from the inequality

$$\mathbb{P}\left[Z \leq \mathbb{E}Z - \sqrt{2x\mathbb{E}Z}\right] \leq e^{-x}$$

(see [4]) and since $x + \frac{1}{2}y \geq \sqrt{2xy}$. $\square$

Finally, we make key use of the following concentration inequality for the supremum of an empirical process, recently established by Talagrand [23], see also Ledoux [14], Massart [19], Rio [21]. The best known constants reported here have been obtained by Bousquet [6].

Proposition 4.5. *Set $\Sigma_{\mathcal{F}_k^*} = \sup_{f \in \mathcal{F}_k^*} L(f)(1 - L(f))$. For all $\epsilon > 0$, $n \geq 1$*

$$\mathbb{P}\left[\sup_{f \in \mathcal{F}_k^*}\left|\widehat{L}(f) - L(f)\right| \geq 2\mathbb{E}\sup_{f \in \mathcal{F}_k^*}\left|\widehat{L}(f) - L(f)\right| + \Sigma_{\mathcal{F}_k^*}\sqrt{2\epsilon} + \frac{4\epsilon}{3}\right] \leq e^{-n\epsilon}.$$

We are now ready to prove Theorems 4.1 and 4.2.

Proof of Theorem 4.1. Deduce, using (i), (ii) and (iii) of Lemma 4.3, the following string of inequalities:

$$\begin{aligned}
&\mathbb{P}\left[\left\{L(\widehat{f}_k) \geq \widehat{L}(\widehat{f}_k) + \widehat{C}_k\right\} \cap A_k \cap B_k\right] \\
&= \mathbb{P}\left[\left\{L(\widehat{f}_k) \geq \widehat{L}(\widehat{f}_k) + 8\widehat{R}_{\widehat{\mathcal{F}}_k} + 10\varepsilon_k + \sqrt{8\widehat{L}(\widehat{f}_k) + 7\widehat{u}_k}\sqrt{2\varepsilon_k}\right\} \cap A_k \cap B_k\right] \\
&\leq \mathbb{P}\left[\left\{\exists f \in \mathcal{F}_k^* :\ L(f) \geq \widehat{L}(f) + 8\widehat{R}_{\widehat{\mathcal{F}}_k} + 10\varepsilon_k + \sqrt{8\widehat{L}(\widehat{f}_k) + 7\widehat{u}_k}\sqrt{2\varepsilon_k}\right\} \cap A_k \cap B_k\right] \\
&\qquad \text{(by property (i))} \\
&\leq \mathbb{P}\left[\left\{\exists f \in \mathcal{F}_k^* :\ L(f) \geq \widehat{L}(f) + 8\widehat{R}_{\mathcal{F}_k^*} + 10\varepsilon_k + \sqrt{8\widehat{L}(\widehat{f}_k) + 7u_k}\sqrt{2\varepsilon_k}\right\} \cap A_k \cap B_k\right] \\
&\qquad \text{(by property (ii) and definition of } B_k\text{)} \\
&\leq \mathbb{P}\left[\left\{\exists f \in \mathcal{F}_k^* :\ L(f) \geq \widehat{L}(f) + 8\widehat{R}_{\mathcal{F}_k^*} + 10\varepsilon_k + \sqrt{4L_k^* + 3u_k}\sqrt{2\varepsilon_k}\right\} \cap A_k \cap B_k\right] \\
&\qquad \text{(by property (iii))} \\
&\leq \mathbb{P}\left\{\sup_{f \in \mathcal{F}_k^*} |L(f) - \widehat{L}(f)| \geq 8\widehat{R}_{\mathcal{F}_k^*} + 10\varepsilon_k + \Sigma_{\mathcal{F}_k^*}\sqrt{2\varepsilon_k}\right\}
\end{aligned}$$

where the last inequality follows from

$$\Sigma^2_{\mathcal{F}_k^*} = \sup_{f \in \mathcal{F}_k^*} \mathrm{Var}(\mathbb{I}\{f(X) \neq Y\}) \leq \sup_{f \in \mathcal{F}_k^*} L(f) \leq 4L_k^* + 3u_k.$$

Invoke inequality (4.7), inequality (4.6) and Propositions 4.4 and 4.5 above to conclude that

$$\begin{aligned}
&\mathbb{P}\left\{L(\widehat{f}_k) \geq \widehat{L}(\widehat{f}_k) + \widehat{C}_k\right\} \\
&\leq \mathbb{P}\left\{\sup_{f \in \mathcal{F}_k^*} |L(f) - \widehat{L}(f)| \geq 8\widehat{R}_{\mathcal{F}_k^*} + 10\varepsilon_k + \Sigma_{\mathcal{F}_k^*}\sqrt{2\varepsilon_k}\right\} + \frac{9}{n^2k^2} \\
&\qquad \text{(since } \mathbb{P}(A_k \cap B_k)^c \leq 9/(n^2k^2) \text{ by (4.6) in Lemma 4.3)} \\
&\leq \mathbb{P}\left\{\sup_{f \in \mathcal{F}_k^*} |L(f) - \widehat{L}(f)| \geq 4\mathbb{E}\widehat{R}_{\mathcal{F}_k^*} + 2\varepsilon_k + \Sigma_{\mathcal{F}_k^*}\sqrt{2\varepsilon_k}\right\} + \frac{10}{n^2k^2} \\
&\qquad \text{(by Proposition 4.4)} \\
&\leq \mathbb{P}\left\{\sup_{f \in \mathcal{F}_k^*} |L(f) - \widehat{L}(f)| \geq 2\mathbb{E}\sup_{f \in \mathcal{F}_k^*} |\widehat{L}(f) - L(f)| + \frac{4\varepsilon_k}{3} + \Sigma_{\mathcal{F}_k^*}\sqrt{2\varepsilon_k}\right\} + \frac{10}{n^2k^2} \\
&\qquad \text{(by (4.7))} \\
&\leq \frac{11}{n^2k^2} \text{ (by Proposition 4.5).}
\end{aligned}$$

This proves the first assertion. The almost sure statement follows by invoking Lemma 2.2 and the preceding argument (which also shows that

$$\mathbb{P}\left\{\widehat{C}_k \leq (L - \widehat{L})(f_k^*)\right\} \leq \frac{11}{n^2k^2}$$

although the last assertion could be shown in a much easier way as it only involves a single function f_k^*). Theorem 4.1 follows from Lemma 2.1 and Lemma 2.2. □

In the proof of Theorem 4.2 we need the symmetrization device

$$\mathbb{E}\widehat{R}_{\mathcal{F}_k} \le 2\mathbb{E} \sup_{f\in\mathcal{F}_k} \left|\widehat{L}(f) - L(f)\right| + \frac{\sup_{f\in\mathcal{F}_k} L(f)}{\sqrt{n}} \ . \tag{4.8}$$

(see, e.g., Mendelson [20, p.18]), and also the following result due to Massart [18]. (The version stated here is taken from Lugosi [16].)

Proposition 4.6. *Set* $\Sigma_k = \sup_{f\in\mathcal{F}_k} \sqrt{L(f)(1-L(f))}$. *Then for all* $n \ge 1$,

$$\mathbb{E} \sup_{f\in\mathcal{F}_k} \left|\widehat{L}(f) - L(f)\right| \le \frac{8\mathbb{E}\log 2\mathbb{S}_k(X_1^{2n})}{n} + 4\sqrt{\frac{2\Sigma_k^2 \ \mathbb{E}\log 2\mathbb{S}_k(X_1^{2n})}{n}} \ .$$

Proof. The statement follows almost immediately from Theorem 1.10 in Lugosi [16] by noting that the worst-case shatter coefficients may be replaced with impunity by the random shatter coefficients. □

Proof of Theorem 4.2. Observe that on the event $A_k \cap B_k$, $\widehat{\mathcal{F}}_k \subseteq \overline{\mathcal{F}}_k$, where $\overline{\mathcal{F}}_k$ is as defined in Theorem 4.2. Indeed, for any $f \in \widehat{\mathcal{F}}_k$,

$$\begin{aligned} L(f) &\le 2\widehat{L}(f) + u_k \\ &\qquad \text{(by definition of } A_k) \\ &\le 2\left[16\widehat{L}(\widehat{f}_k) + 15\widehat{u}_k\right] + u_k \\ &\qquad \text{(by definition of } \widehat{\mathcal{F}}_k) \\ &\le 32\widehat{L}(\widehat{f}_k) + 31\overline{u}_k \\ &\qquad \text{(by definition of } B_k) \\ &\le 32\widehat{L}(f_k^*) + 31\overline{u}_k \\ &\qquad \text{(by definition of } \widehat{f}_k) \\ &\le 32[2L_k^* + u_k] + 31\overline{u}_k \\ &\qquad \text{(by definition of } A_k) \\ &= 64L_k^* + 63\overline{u}_k. \end{aligned}$$

Also, we notice that on the event A_k,

$$\widehat{L}(\widehat{f}_k) \le \widehat{L}(f_k^*) \le 2L_k^* + u_k.$$

These observations imply that

$$\begin{aligned} \widehat{C}_k I_{A_k \cap B_k} &\leq 8\widehat{R}_{\overline{\mathcal{F}}_k} + 10\varepsilon_k + 2\sqrt{64L_k^* + 63\overline{u}_k}\sqrt{2\varepsilon_k} \\ &\leq 8\widehat{R}_{\overline{\mathcal{F}}_k} + 10\varepsilon_k + 16\sqrt{L_k^* + \overline{u}_k}\sqrt{2\varepsilon_k}. \end{aligned}$$

Consequently, it follows from Lemma 4.3 above that

$$\begin{aligned} \mathbb{E}\widehat{C}_k &\leq \mathbb{E}\widehat{C}_k I_{A_k} + \mathbb{P}(A_k \cap B_k)^c \\ &\leq 8\mathbb{E}\widehat{R}_{\overline{\mathcal{F}}_k} + 10\varepsilon_k + 16\sqrt{L_k^* + \overline{u}_k}\sqrt{2\varepsilon_k} + 9(nk)^{-2} \\ &\leq 8\mathbb{E}\widehat{R}_{\overline{\mathcal{F}}_k} + 15\varepsilon_k + 16\sqrt{L_k^* + \overline{u}_k}\sqrt{2\varepsilon_k}. \end{aligned}$$

This bound and Theorem 4.1 yield the first inequality of Theorem 4.2. The second inequality follows from the symmetrization (4.8) and Proposition 4.6 above. □

Acknowledgements. We thank Olivier Bousquet for his invaluable remarks and advices. We also appreciate the helpful remarks by two referees.

References

[1] M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47:207–217, 1993.

[2] P. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2001.

[3] P. Bartlett, O. Bousquet, and S. Mendelson. Localized Rademacher complexities. In *Proceedings of the 15th annual conference on Computational Learning Theory*, pages 44–48, 2002.

[4] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16:277–292, 2000.

[5] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *Annals of Probability*, to appear, 2003.

[6] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Acad. Sci. Paris*, 334:495–500, 2002.

[7] O. Bousquet. New approaches to statistical learning theory. *Annals of the Institute of Statistical Mathematics*, 2003.

[8] O. Bousquet, V. Koltchinskii, and D. Panchenko. Some local measures of complexity of convex hulls and generalization bounds. In *Proceedings of the 15th Annual Conference on Computational Learning Theory*, pages 59–73. Springer, 2002.

[9] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

[10] E. Giné and J. Zinn. Some limit theorems for empirical processes. *Annals of Probability*, 12:929–989, 1984.

[11] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47:1902–1914, 2001.

[12] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In E. Giné, D.M. Mason, and J.A. Wellner, editors, *High Dimensional Probability II*, pages 443–459, 2000.

[13] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.

[14] M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, 1:63–87, 1997. `http://www.emath.fr/ps/`.

[15] F. Lozano. Model selection using rademacher penalization. In *Proceedings of the Second ICSC Symposia on Neural Computation (NC2000)*. ICSC Adademic Press, 2000.

[16] G. Lugosi. Pattern classification and learning theory. In L. Györfi, editor, *Principles of Nonparametric Learning*, pages 5–62. Springer, Vienna, 2002.
[17] G. Lugosi and A. Nobel. Adaptive model selection using empirical complexities. *Annals of Statistics*, 27:1830–1864, 1999.
[18] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Annals of Probability*, 28:863–884, 2000.
[19] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciencies de Toulouse*, IX:245–303, 2000.
[20] S. Mendelson. A few notes on statistical learning theory. In S. Mendelson and A. Smola, editors, *Advanced Lectures in Machine Learning*, LNCS 2600, pages 1–40. Springer, 2003.
[21] E. Rio. Inégalités de concentration pour les processus empiriques de classes de parties. *Probability Theory and Related Fields*, 119:163–175, 2001.
[22] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44:1926–1940, 1998.
[23] M. Talagrand. A new look at independence. *Annals of Probability*, 24:1–34, 1996. (Special Invited Paper).
[24] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *C. R. Acad. Sci. Paris*, to appear, 2001.
[25] A.W. van der Waart and J.A. Wellner. *Weak convergence and empirical processes.* Springer-Verlag, New York, 1996.
[26] V.N. Vapnik. *Statistical Learning Theory.* John Wiley, New York, 1998.
[27] V.N. Vapnik and A.Ya. Chervonenkis. *Theory of Pattern Recognition.* Nauka, Moscow, 1974. (in Russian); German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin, 1979.

Department of Economics, Pompeu Fabra University, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain
E-mail address: lugosi@upf.es

Department of Statistics, Yale University, P.O. Box 208290, New Haven, CT 06520-8290, United States of America
E-mail address: marten.wegkamp@yale.edu