Model Selection and Error Estimation *

Peter L. Bartlett†
Barnhill Technologies
2030 Addison Street, Suite 102, Berkeley, CA 94704, USA,
bartlett@barnhilltechnologies.com

Stéphane Boucheron
Laboratoire de Recherche en Informatique
Bâtiment 490
CNRS-Université Paris-Sud
91405 Orsay-Cedex
bouchero@lri.fr

Gábor Lugosi‡
Department of Economics,
Pompeu Fabra University
Ramon Trias Fargas 25-27,
08005 Barcelona, Spain,
lugosi@upf.es

August 28, 2001

*A shorter version of this paper was presented at COLT'2000

†Supported by the Australian Research Council. This work was done while the first author was at the Research School of Information Sciences and Engineering, Australian National University.

‡Supported by DGES grant PB96-0300.

Abstract

We study model selection strategies based on penalized empirical loss minimization. We point out a tight relationship between error estimation and data-based complexity penalization: any good error estimate may be converted into a data-based penalty function and the performance of the estimate is governed by the quality of the error estimate. We consider several penalty functions, involving error estimates on independent test data, empirical VC dimension, empirical VC entropy, and margin-based quantities. We also consider the maximal difference between the error on the first half of the training data and the second half, and the expected maximal discrepancy, a closely related capacity estimate that can be calculated by Monte Carlo integration. Maximal discrepancy penalty functions are appealing for pattern classification problems, since their computation is equivalent to empirical risk minimization over the training data with some labels flipped.

1 Introduction

We consider the following prediction problem. Based on a random observation $X \in \mathcal{X}$, one has to estimate $Y \in \mathcal{Y}$. A *prediction rule* is a measurable function $f : \mathcal{X} \to \mathcal{Y}$, with *loss* $L(f) = \mathbb{E}\ell(f(X), Y)$, where $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ is a bounded loss function. The data

$$D_n = (X_1, Y_1), \dots, (X_n, Y_n)$$

consist of a sequence of independent, identically distributed samples with the same distribution as (X, Y) and D_n is independent of (X, Y). The goal is to choose a prediction rule f_n from some restricted class $\mathcal{F}$ such that the *loss* $L(f_n) = \mathbb{E}[\ell(f_n(X), Y)|D_n]$ is as close as possible to the best possible loss, $L^* = \inf_f L(f)$, where the infimum is taken over all prediction rules $f : \mathcal{X} \to \mathcal{Y}$.

Empirical risk minimization evaluates the performance of each prediction rule $f \in \mathcal{F}$ in terms of its empirical loss $\widehat{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)$. This provides an estimate whose loss is close to the optimal loss L^* if the class $\mathcal{F}$ is (i) sufficiently large so that the loss of the best function in $\mathcal{F}$ is close to L^* and (ii) is sufficiently small so that finding the best candidate in $\mathcal{F}$ based on the data is still possible. These two requirements are clearly in conflict. The trade-off is best understood by writing

$$\mathbb{E}L(f_n) - L^* = \left(\mathbb{E}L(f_n) - \inf_{f \in \mathcal{F}} L(f)\right) + \left(\inf_{f \in \mathcal{F}} L(f) - L^*\right).$$

The first term is often called *estimation error*, while the second is the *approximation error*. Often $\mathcal{F}$ is large enough to minimize $L(\cdot)$ for all possible distributions of (X, Y), so that $\mathcal{F}$ is too large for empirical risk minimization. In this case it is common to fix in advance a sequence of smaller model classes $\mathcal{F}_1, \mathcal{F}_2, \dots$ whose union is equal to $\mathcal{F}$. Given the data D_n, one wishes to select a good model from *one* of these classes. This is the problem of model selection.

Denote by $\widehat{f}_k$ a function in $\mathcal{F}_k$ having minimal empirical risk. One hopes to select a model class $\mathcal{F}_K$ such that the excess error $\mathbb{E}L(\widehat{f}_K) - L^*$ is close to

$$\min_k \mathbb{E}L(\widehat{f}_k) - L^* = \min_k \left[\left(\mathbb{E}L(\widehat{f}_k) - \inf_{f \in \mathcal{F}_k} L(f)\right) + \left(\inf_{f \in \mathcal{F}_k} L(f) - L^*\right)\right].$$

The idea of *structural risk minimization* (also known as *complexity regularization*) is to add a complexity penalty to each of the $\widehat{L}_n(\widehat{f}_k)$'s to compensate for the overfitting effect. This penalty is usually closely related to a distribution-free upper bound for $\sup_{f\in\mathcal{F}_k}|\widehat{L}_n(f)-L(f)|$ so that the penalty eliminates the effect of overfitting. Thus, structural risk minimization finds the best trade-off between the approximation error and a distribution-free upper bound on the estimation error. Unfortunately, distribution-free upper bounds may be too conservative for specific distributions. This criticism has led to the idea of using *data-dependent* penalties.

In the next section, we show that any approximate upper bound on error (including a data-dependent bound) can be used to define a (possibly data-dependent) complexity penalty $C_n(k)$ and a model selection algorithm for which the excess error is close to

$$\min_k\left[\mathbb{E}C_n(k)+\left(\inf_{f\in\mathcal{F}_k}L(f)-L^*\right)\right].$$

Section 3 gives several applications of the performance bounds of Section 2: Section 3.1 considers the estimates provided by an independent test sample. These have the disadvantage that they cost data. Section 3.2, considers a distribution-free estimate based on the VC dimension and a data-dependent estimate based on shatter coefficients. Unfortunately, these are difficult to compute. Section 3.3 briefly considers margin-based error estimates, which can be viewed as easily computed estimates of quantities analogous to shatter coefficients. Section 3.4 looks at an estimate provided by maximizing the discrepancy between the error on the first half of the sample and that on the second half. For classification, this estimate can be conveniently computed, simply by minimizing empirical risk with half of the labels flipped. Section 3.5 looks at a more complex estimate: the expected maximum discrepancy. This estimate can be calculated by Monte Carlo integration, and can lead to better performance bounds. In Section 4 we review some concentration inequalities that are central to our proofs. Finally, in Section 5 we offer an experimental comparison of some of the proposed methods.

For clarity, we include in Table 1 notation that we use throughout the paper.

For work on complexity regularization, see Akaike [1], Barron [2],[3] Barron, Birgé, and Massart [4], Barron and Cover [5], Birgé and Massart [8],[9], Buescher and Kumar [11],[12], Devroye, Györfi, and Lugosi, [14], Gallant [16], Geman and Hwang [17], Kearns, Mansour, Ng, and Ron [20], Krzyżak and Linder [23], Lugosi and Nobel [25] Lugosi and Zeger, [27], [26], Mallows [28], Meir [33], Modha and Masry [34], Rissanen [35], Schwarz [37], Shawe-Taylor, Bartlett, Williamson, and Anthony [38], Shen and Wong [39], Vapnik

f	prediction rule, $f : \mathcal{X} \to \mathcal{Y}$
$\mathcal{F}_1, \mathcal{F}_2, \ldots$	sets of prediction rules (model classes)
$\mathcal{F}$	union of model classes $\mathcal{F}_k$
f_k^*	element of F_k with minimal loss
$\widehat{f}_k$	element of $\mathcal{F}_k$ minimizing empirical loss
f_n	prediction rule from $\mathcal{F}$ minimizing $\tilde{L}_n(\widehat{f}_k)$
ℓ	loss function, $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$
L	loss, $L(f) = \mathbb{E}\ell(f(X), Y)$
L_k^*	minimal loss of functions in $\mathcal{F}_k$, $L_k^* = \inf_{f \in \mathcal{F}_k} L(f)$
$\widehat{L}_n$	empirical loss
$R_{n,k}$	estimate (high confidence upper bound) of loss $L(\widehat{f}_k)$
$C_n(k)$	complexity penalty for class $\mathcal{F}_k$
$\tilde{L}_n$	complexity penalized loss estimate, $\tilde{L}_n(\widehat{f}_k) = \widehat{L}_n(\widehat{f}_k) + C_n(k)$
L^*	loss of optimal prediction rule

Table 1: Notation.

[42], Vapnik and Chervonenkis [46], Yang and Barron [50], [51].

Data-dependent penalties are studied by Bartlett [6], Freund [15], Koltchinskii [21], Koltchinskii and Panchenko [22], Lozano [24], Lugosi and Nobel [25], Massart [30] and Shawe-Taylor, Bartlett, Williamson, and Anthony [38].

2 Penalization by error estimates

For each class $\mathcal{F}_k$, let $\widehat{f}_k$ denote the prediction rule that is selected from $\mathcal{F}_k$ based on the data. Our goal is to select, among these rules, one which has approximately minimal loss. The key assumption for our analysis is that the true loss of $\widehat{f}_k$ can be estimated for all k.

Assumption 1 *For every n, there are positive numbers c and m such that for each k an estimate $R_{n,k}$ on $L(\widehat{f}_k)$ is available which satisfies*

$$\mathbb{P}\left[L(\widehat{f}_k) > R_{n,k} + \epsilon\right] \leq ce^{-2m\epsilon^2} \tag{1}$$

for all ϵ.

Notice that c and m might depend on the sample size n.

Now define the data-based complexity penalty by

$$C_n(k) = R_{n,k} - \widehat{L}_n(\widehat{f}_k) + \sqrt{\frac{\log k}{m}}.$$

The last term is required because of technical reasons that will become apparent shortly. It is typically small. The difference $R_{n,k} - \widehat{L}_n(\widehat{f}_k)$ is simply an estimate of the 'right' amount of penalization $L(\widehat{f}_k) - \widehat{L}_n(\widehat{f}_k)$. Finally, define the prediction rule:

$$f_n = \operatorname*{arg\,min}_{k=1,2,\ldots} \tilde{L}_n(\widehat{f}_k),$$

where

$$\tilde{L}_n(\widehat{f}_k) = \widehat{L}_n(\widehat{f}_k) + C_n(k) = R_{n,k} + \sqrt{\frac{\log k}{m}}.$$

The following theorem summarizes the main performance bound for f_n.

Theorem 1 *Assume that the error estimates $R_{n,k}$ satisfy (1) for some positive constants c and m. Then for all $\epsilon > 0$,*

$$\mathbb{P}\left[L(f_n) - \tilde{L}_n(f_n) > \epsilon\right] \leq 2ce^{-2m\epsilon^2}.$$

Moreover, if for all k, $\widehat{f}_k$ minimizes the empirical loss in the model class $\mathcal{F}_k$, then

$$\mathbb{E}L(f_n) - L^* \leq \min_k \left[\mathbb{E}C_n(k) + \left(\inf_{f\in\mathcal{F}_k} L(f) - L^*\right)\right] + \sqrt{\frac{\log(ce)}{2m}}.$$

The second part of Theorem 1 shows that the prediction rule minimizing the penalized empirical loss achieves an almost optimal trade-off between the approximation error and the expected complexity, provided that the estimate $R_{n,k}$ on which the complexity is based is an approximate upper bound on the loss. In particular, if we knew in advance which of the classes $\mathcal{F}_k$ contained the optimal prediction rule, we could use the error estimates $R_{n,k}$ to obtain an upper bound on $\mathbb{E}L(\widehat{f}_k) - L^*$, and this upper bound would not improve on the bound of Theorem 1 by more than $O\left(\sqrt{\log k/m}\right)$.

If the range of the loss function ℓ is an infinite set, the infimum of the empirical loss might not be achieved. In this case, we could define $\widehat{f}_k$ as a suitably good approximation to the infimum. However, for convenience, we assume throughout that the minimum always exists. It suffices for this, and for various proofs, to assume that for all n and $(x_1, y_1), \ldots, (x_n, y_n)$, the set

$$\{(\ell(f(x_1), y_1), \ldots, \ell(f(x_1), y_1)) : f \in \mathcal{F}_k\}$$

is closed.

Proof. For brevity, introduce the notation

$$L_k^* = \inf_{f\in\mathcal{F}_k} L(f).$$

Then for any $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}\left[L(f_n) - \tilde{L}_n(f_n) > \epsilon\right] &\leq \mathbb{P}\left[\sup_{j=1,2,\ldots}\left(L(\widehat{f}_j) - \tilde{L}_n(\widehat{f}_j)\right) > \epsilon\right] \\
&\leq \sum_{j=1}^{\infty} \mathbb{P}\left[L(\widehat{f}_j) - \tilde{L}_n(\widehat{f}_j) > \epsilon\right] \\
&\quad \text{(by the union bound)}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{\infty} \mathbb{P}\left[L(\widehat{f}_j) - R_{n,j} > \epsilon + \sqrt{\frac{\log j}{m}} \right] \\
&\quad \text{(by definition)} \\
&\le \sum_{j=1}^{\infty} c e^{-2m\left(\epsilon + \sqrt{\frac{\log j}{m}}\right)^2} \quad \text{(by Assumption 1)} \\
&\le \sum_{j=1}^{\infty} c e^{-2m\left(\epsilon^2 + \frac{\log j}{m}\right)} \\
&< 2ce^{-2m\epsilon^2} \quad (\text{since } \textstyle\sum_{j=1}^{\infty} j^{-2} < 2).
\end{aligned}
$$

To prove the second inequality, for each k, we decompose $L(f_n) - L_k^*$ as

$$
L(f_n) - L_k^* = \left(L(f_n) - \inf_j \tilde{L}_n(\widehat{f}_j) \right) + \left(\inf_j \tilde{L}_n(\widehat{f}_j) - L_k^* \right).
$$

The first term may be bounded, by standard integration of the tail inequality shown above (see, e.g., [14, page 208]), as $\mathbb{E}\left[L(f_n) - \inf_j \tilde{L}_n(\widehat{f}_j) \right] \le \sqrt{\log(ce)/(2m)}$. Choosing f_k^* such that $L(f_k^*) = L_k^*$, the second term may be bounded directly by

$$
\begin{aligned}
\mathbb{E} \inf_j \tilde{L}_n(\widehat{f}_j) - L_k^* &\le \mathbb{E}\tilde{L}_n(\widehat{f}_k) - L_k^* \\
&= \mathbb{E}\widehat{L}_n(\widehat{f}_k) - L_k^* + \mathbb{E}C_n(k) \\
&\quad \text{(by the definition of } \tilde{L}_n(\widehat{f}_k)) \\
&\le \mathbb{E}\widehat{L}_n(f_k^*) - L(f_k^*) + \mathbb{E}C_n(k) \\
&\quad \text{(since } \widehat{f}_k \text{ minimizes the empirical loss on } \mathcal{F}_k) \\
&= \mathbb{E}C_n(k),
\end{aligned}
$$

where the last step follows from the fact that $\mathbb{E}\widehat{L}_n(f_k^*) = L(f_k^*)$. Summing the obtained bounds for both terms yields that for each k,

$$
\mathbb{E}L(f_n) \le \mathbb{E}C_n(k) + L_k^* + \sqrt{\log(ce)/(2m)},
$$

which implies the second statement of the theorem. ∎

Sometimes bounds tighter than Assumption 1 are available, as in Assumption 2 below. Such bounds may be exploited to decrease the term $\sqrt{\log k/m}$ in the definition of the complexity penalty.

Assumption 2 *For every n, there are positive numbers c and m such that for each k an estimate $\overline{R}_{n,k}$ of $L(\widehat{f}_k)$ is available which satisfies*

$$\mathbb{P}\left[L(\widehat{f}_k) > \overline{R}_{n,k} + \epsilon\right] \leq ce^{-m\epsilon} \tag{2}$$

for all ϵ.

Define the modified penalty by

$$\overline{C}_n(k) = \overline{R}_{n,k} - \widehat{L}_n(\widehat{f}_k) + \frac{2\log k}{m}$$

and define the prediction rule

$$\overline{f}_n = \mathop{\arg\min}_{k=1,2,\ldots} \overline{L}_n(\widehat{f}_k),$$

where

$$\overline{L}_n(\widehat{f}_k) = \widehat{L}_n(\widehat{f}_k) + \overline{C}_n(k) = \overline{R}_{n,k} + \frac{2\log k}{m}.$$

Then by a trivial modification of the proof of Theorem 1 we obtain the following result.

Theorem 2 *Assume that the error estimates $\overline{R}_{n,k}$ satisfy Assumption 2 for some positive constants c and m. Then for all $\epsilon > 0$,*

$$\mathbb{P}\left[L(f_n) - \overline{L}_n(f_n) > \epsilon\right] \leq 2ce^{-m\epsilon}.$$

Moreover, if for all k, $\widehat{f}_k$ minimizes the empirical loss in the model class $\mathcal{F}_k$, then

$$\mathbb{E}L(\overline{f}_n) - L^* \leq \min_k \left[\mathbb{E}\overline{C}_n(k) + \left(\inf_{f\in\mathcal{F}_k} L(f) - L^*\right)\right] + \frac{\log(2ec)}{m}.$$

So far we have only concentrated on the expected loss of the penalized estimate. However, with an easy modification of the proof we obtain exponential tail inequalities. We work out one such inequality in the scenario of Theorem 1.

Theorem 3 *Assume that the error estimates $R_{n,k}$ satisfy (1) for some positive constants c and m, and that for all k, $\widehat{f}_k$ minimizes the empirical loss in the model class $\mathcal{F}_k$. Then for all $\epsilon > 0$,*

$$\mathbb{P}\left[L(f_n) > \inf_k \left(L_k^* + C_n(k) + \sqrt{\frac{\log k}{n}}\right) + \epsilon\right] \le 2ce^{-m\epsilon^2/2} + 2e^{-n\epsilon^2/2}.$$

Proof. Note that

$$\begin{aligned}
&\mathbb{P}\left[L(f_n) > \inf_k \left(L_k^* + C_n(k) + \sqrt{\frac{\log k}{n}}\right) + \epsilon\right] \\
\le\ & \mathbb{P}\left[L(f_n) > \inf_j \tilde{L}_n(\widehat{f}_j) + \frac{\epsilon}{2}\right] \\
& + \mathbb{P}\left[\inf_j \tilde{L}_n(\widehat{f}_j) > \inf_k \left(L_k^* + C_n(k) + \sqrt{\frac{\log k}{n}}\right) + \frac{\epsilon}{2}\right] \\
\le\ & 2ce^{-m\epsilon^2/2} \\
& + \mathbb{P}\left[\sup_k \left(\tilde{L}_n(\widehat{f}_k) - L_k^* - C_n(k) - \sqrt{\frac{\log k}{n}}\right) > \frac{\epsilon}{2}\right] \\
& \quad \text{(by the first inequality of Theorem 1)} \\
\le\ & 2ce^{-m\epsilon^2/2} \\
& + \sum_{k=1}^{\infty} \mathbb{P}\left[\widehat{L}_n(\widehat{f}_k) - L_k^* > \frac{\epsilon}{2} + \sqrt{\frac{\log k}{n}}\right] \\
& \quad \text{(by the union bound and the definition of } \tilde{L}_n\text{)} \\
\le\ & 2ce^{-m\epsilon^2/2} + \sum_{k=1}^{\infty} \mathbb{P}\left[\widehat{L}_n(f^*) - L_k^* > \frac{\epsilon}{2} + \sqrt{\frac{\log k}{n}}\right] \\
& \quad \text{(since } \widehat{f}_k \text{ minimizes the empirical loss on } \mathcal{F}_k\text{)} \\
\le\ & 2ce^{-m\epsilon^2/2} + \sum_{k=1}^{\infty} e^{-2n\left(\epsilon/2 + \sqrt{\log k/n}\right)^2} \\
& \quad \text{(by Hoeffding's inequality)} \\
\le\ & 2ce^{-m\epsilon^2/2} + 2e^{-n\epsilon^2/2}.
\end{aligned}$$

This concludes the proof. ∎

In the examples shown below we concentrate on the expected loss of penalized empirical error minimizers. Tail probability estimates may be obtained in all cases by a simple application of the theorem above.

3 Applications

3.1 Independent test sample

Assume that m independent sample pairs

$$(X_1', Y_1'), \ldots, (X_m', Y_m')$$

are available. We can simply remove m samples from the training data. Of course, this is not very attractive, but m may be small relative to n. In this case we can estimate $L(\widehat{f}_k)$ by

$$R_{n,k} = \frac{1}{m} \sum_{i=1}^{m} \ell(\widehat{f}_k(X_i'), Y_i'). \tag{3}$$

We apply Hoeffding's inequality to show that Assumption 1 is satisfied with $c = 1$, notice that $\mathbb{E}[R_{n,k}|D_n] = L(\widehat{f}_k)$, and apply Theorem 1 to give the following result.

Corollary 1 *Assume that the model selection algorithm of Section 2 is performed with the hold-out error estimate (3). Then*

$$\begin{aligned} &\mathbb{E}L(f_n) - L^* \\ &\leq \min_k \left[\mathbb{E}\left[L(\widehat{f}_k) - \widehat{L}_n(\widehat{f}_k) \right] + \left(\inf_{f \in \mathcal{F}_k} L(f) - L^* \right) + \sqrt{\frac{\log k}{m}} \right] + \frac{1}{\sqrt{2m}}. \end{aligned}$$

In other words, the estimate achieves a nearly optimal balance between the approximation error, and the quantity

$$\mathbb{E}\left[L(\widehat{f}_k) - \widehat{L}_n(\widehat{f}_k) \right],$$

which may be regarded as the amount of overfitting.

With this inequality we recover the main result of Lugosi and Nobel [25], but now with a much simpler estimate. In fact, the bound of the corollary may substantially improve the main result of [25].

The square roots in the bound of Corollary 1 can be removed by increasing the penalty term by a small constant factor and using Bernstein's inequality in place of Hoeffding's as follows: Choose the modified estimate

$$\overline{R}_{n,k} = \frac{1}{1-\alpha}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(\widehat{f}_k(X_i'), Y_i')\right],$$

where $\alpha < 1$ is a positive constant. Then Bernstein's inequality (see, e.g., [14]) yields

$$\mathbb{P}\left[L(\widehat{f}_k) \geq \overline{R}_{n,k} + \epsilon\right] \leq e^{-3m\epsilon\alpha(1-\alpha)/8}\ .$$

Thus, (2) is satisfied with m replaced by $3m\alpha(1-\alpha)/8$. Therefore, defining

$$\overline{C}_{n,k} = \overline{R}_{n,k} - \widehat{L}(\widehat{f}_{n,k}) + \frac{16\log k}{3m\alpha(1-\alpha)},$$

we obtain the performance bound

$$\mathbb{E}L(f_n) - L^* \leq \min_k\left[\mathbb{E}\overline{C}_n(k) + \left(\inf_{f\in\mathcal{F}_k} L(f) - L^*\right)\right] + \frac{16}{3m\alpha(1-\alpha)}\ .$$

3.2 Estimated complexity

In the remaining examples we consider error estimates $R_{n,k}$ which avoid splitting the data.

For simplicity, we concentrate in this section on the case of classification ($\mathcal{Y} = \{0,1\}$ and the *0-1 loss*, defined by $\ell(0,0) = \ell(1,1) = 0$ and $\ell(0,1) = \ell(1,0) = 1$), although similar arguments may be carried out for the general case as well.

Recall the basic Vapnik-Chervonenkis inequality [45], [43],

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}_k}\left(L(f) - \widehat{L}_n(f)\right) > \epsilon\right] \leq 4\mathbb{E}S_k(X_1^{2n})e^{-n\epsilon^2}, \tag{4}$$

where $S_k(X_1^n)$ is the *empirical shatter coefficient* of $\mathcal{F}_k$, that is, the number of different ways the n points $X_1, \ldots, X_n$ can be classified by elements of $\mathcal{F}_k$. It is easy to show that this inequality implies that the estimate

$$R_{n,k} = \widehat{L}_n(\widehat{f}_k) + \sqrt{\frac{\log\mathbb{E}S_k(X_1^{2n}) + \log 4}{n}}$$

satisfies Assumption 1 with $m = n/2$ and $c = 1$. We need to estimate the quantity $\log \mathbb{E} S_k(X_1^{2n})$. The simplest way is to use the fact that $\mathbb{E} S_k(X_1^{2n}) \leq (2n+1)^{V_k}$, where V_k is the VC dimension of $\mathcal{F}_k$. Substituting this into Theorem 1 gives

$$\begin{aligned}
&\mathbb{E}L(f_n) - L^* \\
&\leq \min_k \left[\sqrt{\frac{V_k \log(2n+1) + \log 4}{n}} + \left(\inf_{f \in \mathcal{F}_k} L(f) - L^* \right) + \sqrt{\frac{2\log k}{n}} \right] \\
&\qquad + \sqrt{\frac{1}{n}}. \qquad (5)
\end{aligned}$$

This is the type of distribution-free result we mentioned in the introduction. A more interesting result involves estimating $\mathbb{E} S_k(X_1^{2n})$ by $S_k(X_1^n)$.

Theorem 4 *Assume that the model selection algorithm of Section 2 is used with*

$$R_{n,k} = \widehat{L}_n(\widehat{f}_k) + \sqrt{\frac{12 \log S_k(X_1^n) + \log 4}{n}}$$

and $m = n/80$. Then

$$\begin{aligned}
&\mathbb{E}L(f_n) - L^* \\
&\leq \min_k \left[\sqrt{\frac{12\mathbb{E}\log S_k(X_1^n) + \log 4}{n}} + \left(\inf_{f \in \mathcal{F}_k} L(f) - L^* \right) + 8.95\sqrt{\frac{\log k}{n}} \right] \\
&\qquad + \frac{8.23}{\sqrt{n}}.
\end{aligned}$$

The key ingredient of the proof is a concentration inequality from [10] for the *random* VC *entropy*, $\log_2 S_k(X_1^n)$.

Proof. We need to check the validity of Assumption 1. It is shown in [10] that $f(x_1, \ldots, x_n) = \log_2 S_k(x_1^n)$ satisfies the conditions of Theorem 9 below.

First note that $\mathbb{E} S_k(X_1^{2n}) \leq \mathbb{E}^2 S_k(X_1^n)$, and therefore

$$\begin{aligned}
\log \mathbb{E} S_k(X_1^{2n}) &\leq 2 \log \mathbb{E} S_k(X_1^n) \\
&\leq \frac{2}{\log 2} \mathbb{E} \log S_k(X_1^n) \\
&< 3\mathbb{E} \log S_k(X_1^n)
\end{aligned}$$

by the last inequality of Theorem 9. Therefore,

$$\begin{aligned}
&\mathbb{P}\left[L(\widehat{f}_k) - \widehat{L}_n(\widehat{f}_k) > \epsilon + \sqrt{\frac{3\mathbb{E}\log S_k(X_1^n) + \log 4}{n}}\right] \\
\leq \;& \mathbb{P}\left[\sup_{f\in\mathcal{F}_k}\left(L(f) - \widehat{L}_n(f)\right) > \epsilon + \sqrt{\frac{\log \mathbb{E}S_k(X_1^{2n}) + \log 4}{n}}\;\right] \leq e^{-n\epsilon^2},
\end{aligned}$$

where we used the Vapnik-Chervonenkis inequality (4). It follows that

$$\begin{aligned}
&\mathbb{P}\left[L(\widehat{f}_k) > R_{n,k} + \epsilon\right] \\
= \;& \mathbb{P}\left[L(\widehat{f}_k) - \widehat{L}_n(\widehat{f}_k) > \sqrt{\frac{12\log S_k(X_1^n) + \log 4}{n}} + \epsilon\right] \\
\leq \;& \mathbb{P}\left[L(\widehat{f}_k) - \widehat{L}_n(\widehat{f}_k) > \frac{\epsilon}{4} + \sqrt{\frac{3\mathbb{E}\log S_k(X_1^n) + \log 4}{n}}\;\right] \\
&\quad + \mathbb{P}\left[\sqrt{\frac{12\log S_k(X_1^n) + \log 4}{n}} + \frac{3\epsilon}{4} < \sqrt{\frac{3\mathbb{E}\log S_k(X_1^n) + \log 4}{n}}\;\right] \\
\leq \;& e^{-n\epsilon^2/16} \\
&\quad + \mathbb{P}\left[\sqrt{\frac{12\log S_k(X_1^n) + \log 4}{n}} + \frac{3\epsilon}{4} < \sqrt{\frac{3\mathbb{E}\log S_k(X_1^n) + \log 4}{n}}\;\right].
\end{aligned}$$

The last term may be bounded using Theorem 9 as follows:

$$\begin{aligned}
&\mathbb{P}\left[\sqrt{\frac{12\log S_k(X_1^n) + \log 4}{n}} + \frac{3\epsilon}{4} < \sqrt{\frac{3\mathbb{E}\log S_k(X_1^n) + \log 4}{n}}\;\right] \\
\leq \;& \mathbb{P}\left[\log S_k(X_1^n) < \mathbb{E}\log S_k(X_1^n) - \frac{3}{4}\mathbb{E}\log S_k(X_1^n) - \frac{3}{64}n\epsilon^2\right] \\
\leq \;& \exp\left(-\frac{9}{32}\frac{\left(\mathbb{E}\log S_k(X_1^n) + \frac{n\epsilon^2}{16\log 2}\right)^2}{\mathbb{E}\log S_k(X_1^n)}\right)
\end{aligned}$$

$$\leq \exp\left(-\frac{9}{32}\frac{\left(\mathbb{E}\log S_k(X_1^n)+\frac{n\epsilon^2}{16\log 2}\right)^2}{\mathbb{E}\log S_k(X_1^n)+\frac{n\epsilon^2}{16\log 2}}\right)$$
$$\leq \exp\left(-\frac{9n\epsilon^2}{512\log 2}\right).$$

Summarizing, we have that

$$\begin{aligned}\mathbb{P}\left[L(\widehat{f}_k) > R_{n,k}+\epsilon\right] &\leq e^{-n\epsilon^2/16}+e^{-9n\epsilon^2/512\log 2} \\ &< 2e^{-n\epsilon^2/40}.\end{aligned}$$

Therefore, Assumption 1 is satisfied with $c = 2$ and $m = n/80$. Applying Theorem 1 finishes the proof. ∎

3.3 Effective VC dimension and margin

In practice it may be difficult to compute the value of the random shatter coefficients $S_k(X_1^n)$. An alternative way to assign complexities may be easily obtained by observing that $S_k(X_1^n) \leq (n+1)^{D_k}$, where D_k is the *empirical* VC dimension of class $\mathcal{F}_k$, that is, the VC dimension restricted to the points $X_1, \ldots, X_n$. Now it is immediate that the estimate

$$R_{n,k} = \widehat{L}_n(\widehat{f}_k) + \sqrt{\frac{12D_k\log(n+1)+\log 4}{n}},$$

satisfies Assumption 1 in the same way as the estimate of Theorem 4. (In fact, with a more careful analysis it is possible to get rid of the $\log n$ factor at the price of an increased constant.)

Unfortunately, computing D_k in general is still very difficult. A lot of effort has been devoted to obtain upper bounds for D_k which are simple to compute. These bounds are handy in our framework, since any upper bound may immediately be converted into a complexity penalty. In particular, the margins-based upper bounds on misclassification probability for neural networks [6], support vector machines [38, 7, 44, 13], and convex combinations of classifiers [36, 29] immediately give complexity penalties and, through Theorem 1, performance bounds.

We recall here some facts which are at the basis of the theory of *support vector machines*, see Bartlett and Shawe-Taylor [7], Cristianini and Shawe-Taylor [13], Vapnik [44] and the references therein.

A model class $\mathcal{F}$ is called a class of (generalized) linear classifiers if there exists a function $\psi : \mathcal{X} \to \mathbb{R}^p$ such that $\mathcal{F}$ is the class of linear classifiers in $\mathbb{R}^p$, that is, the class of all prediction rules of the form

$$f(x) = \begin{cases} 1 & \text{if } \psi(x)^T w \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $w \in \mathbb{R}^p$ is a weight vector satisfying $\|w\| = 1$.

Much of the theory of support vector machines builds on the fact that the "effective" VC dimension of those generalized linear classifiers for which the minimal distance of the correctly classified data points to the separating hyperplane is larger than a certain "margin" may be bounded, independently of the linear dimension p, by a function of the margin. If for some constant $\gamma > 0$, $(2Y_i - 1)\psi(X_i)^T w \geq \gamma$ then we say that the linear classifier *correctly classifies X_i with margin γ*. We recall the following result:

Lemma 1 (Bartlett and Shawe-Taylor [7]). *Let f_n be an arbitrary (possibly data dependent) linear classifier of the form*

$$f_n(x) = \begin{cases} 1 & \text{if } \psi(x)^T w_n \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $w_n \in \mathbb{R}^p$ is a weight vector satisfying $\|w_n\| = 1$. Let $R, \gamma > 0$ be positive random variables and let $K \leq n$ be a positive integer valued random variable such that $\|\psi(X_i)\| \leq R$ for all $i = 1, \ldots, n$ and f_n correctly classifies all but K of the n data points X_i with margin γ, then for all $\delta > 0$,

$$\mathbb{P}\left[L(f_n) > \frac{K}{n} + 27.18\sqrt{\frac{1}{n}\left(\frac{R^2}{\gamma^2}(\log^2 n + 84) + \log\frac{4}{\delta}\right)}\ \right] \leq \delta.$$

Assume now that $\widehat{f}$ minimizes the empirical loss in a class $\mathcal{F}$ of generalized linear classifiers, such that it correctly classifies at least $n - K$ data points with margin γ and $\|\psi(X_i)\| \leq R$ for all $i = 1, \ldots, n$. Choosing $m = n \log 2/8$ and $\delta = 4e^{-2m\epsilon^2}$, an application of the lemma shows that if we take

$$R_n = \frac{K}{n} + 27.18\sqrt{\frac{1}{n}\left(\frac{R^2}{\gamma^2}(\log^2 n + 84)\right)},$$

then we obtain

$$
\begin{aligned}
& \mathbb{P}\left[L(\widehat{f}) > R_n + \epsilon\right] \\
= \;& \mathbb{P}\left[L(\widehat{f}) > \frac{K}{n} + 27.18\sqrt{\frac{1}{n}\left(\frac{R^2}{\gamma^2}(\log^2 n + 84)\right)} + \sqrt{\frac{1}{2m}\log\frac{4}{\delta}}\;\right] \\
\leq \;& \mathbb{P}\left[L(\widehat{f}) > \frac{K}{n}27.18\sqrt{\frac{1}{n}\left(\frac{R^2}{\gamma^2}(\log^2 n + 84) + \log\frac{4}{\delta}\right)}\;\right] \\
& \quad \text{(using the inequality } \sqrt{x+y} \leq \sqrt{x} + \sqrt{y}\text{)} \\
\leq \;& \delta = 4e^{-2m\epsilon^2}.
\end{aligned}
$$

This inequality shows that if all model classes $\mathcal{F}_k$ are classes of generalized linear classifiers and for all classes the error estimate $R_{n,k}$ is defined as above, then condition (1) is satisfied and Theorem 1 may be used. As a result, we obtain the following performance bound:

Theorem 5

$$
\begin{aligned}
\mathbb{E}L(f_n) - L^* \;\leq\; & \min_k \left[\mathbb{E}\left[\frac{K_k}{n} + 27.18\sqrt{\frac{1}{n}\left(\frac{R_k^2}{\gamma_k^2}(\log^2 n + 41)\right)} - \widehat{L}(\widehat{f}_k)\right]\right. \\
& \left. + \left(\inf_{f \in \mathcal{F}_k} L(f) - L^*\right) + 3.4\sqrt{\frac{\log k}{n}}\;\right] + \frac{3.72}{\sqrt{n}},
\end{aligned}
$$

where K_k, γ_k, and R_k are the random variables K, γ, R defined above, corresponding to the class $\mathcal{F}_k$.

The importance of this result lies in the fact that it gives a computationally feasible way of assigning data-dependent penalties to linear classifiers. On the other hand, the estimates $R_{n,k}$ may be much inferior to the estimates studied in the previous section.

3.4 Penalization by maximal discrepancy

In this section we propose an alternative way of computing the penalties with improved performance guarantees. The new penalties may be still difficult to compute efficiently, but there is a better chance to obtain good approximate quantities as they are defined as solutions of an optimization problem.

Assume, for simplicity, that n is even, divide the data into two equal halves, and define, for each predictor f, the empirical loss on the two parts by

$$\widehat{L}_n^{(1)}(f) = \frac{2}{n} \sum_{i=1}^{n/2} \ell(f(X_i), Y_i)$$

and

$$\widehat{L}_n^{(2)}(f) = \frac{2}{n} \sum_{i=n/2+1}^{n} \ell(f(X_i), Y_i).$$

Using the notation of Section 2, define the error estimate $R_{n,k}$ by

$$R_{n,k} = \widehat{L}_n(\widehat{f}_k) + \max_{f \in \mathcal{F}_k} \left(\widehat{L}_n^{(1)}(f) - \widehat{L}_n^{(2)}(f) \right). \tag{6}$$

If $\mathcal{Y} = \{0, 1\}$ and the loss function is the 0-1 loss (i.e., $\ell(0,0) = \ell(1,1) = 0$ and $\ell(0,1) = \ell(1,0) = 1$) then the maximum discrepancy,

$$\max_{f \in \mathcal{F}_k} \left(\widehat{L}_n^{(1)}(f) - \widehat{L}_n^{(2)}(f) \right)$$

may be computed using the following simple trick: first flip the labels of the first half of the data, thus obtaining the modified data set

$$D'_n = (X'_1, Y'_1), \ldots, (X'_n, Y'_n)$$

with $(X'_i, Y'_i) = (X_i, 1 - Y_i)$ for $i \leq n/2$ and $(X'_i, Y'_i) = (X_i, Y_i)$ for $i > n/2$. Next find $f_k^- \in \mathcal{F}_k$ which minimizes the empirical loss based on D'_n,

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} \ell(f(X'_i), Y'_i) \\
&= \frac{1}{2} - \frac{1}{n} \sum_{i=1}^{n/2} \ell(f(X_i), Y_i) + \frac{1}{n} \sum_{i=n/2+1}^{n} \ell(f(X_i), Y_i) \\
&= \frac{1 - \widehat{L}_n^{(1)}(f) + \widehat{L}_n^{(2)}(f)}{2}.
\end{aligned}$$

Clearly, the function f_k^- maximizes the discrepancy. Therefore, the same algorithm that is used to compute the empirical loss minimizer $\widehat{f}_k$ may be used to find f_k^- and compute the penalty based on maximum discrepancy. This is appealing: although empirical loss minimization is often computationally difficult, the same approximate optimization algorithm can be used for both finding prediction rules and estimating appropriate penalties. In particular, if the algorithm only approximately minimizes empirical loss over the class $\mathcal{F}_k$ because it minimizes over some proper subset of $\mathcal{F}_k$, the theorem is still applicable.

Vapnik *et al.* [47] considered a similar quantity for the case of pattern classification. Motivated by bounds (similar to (5)) on $\mathbb{E}L(f_n) - \widehat{L}_n(f)$, they defined an *effective* VC *dimension*, which is obtained by choosing a value of the VC dimension that gives the best fit of the bound to experimental estimates of $\mathbb{E}L(f_n) - \widehat{L}_n(f)$. They showed that for linear classifiers in a fixed dimension with a variety of probability distributions, the fit was good. This suggests a model selection strategy that estimates $\mathbb{E}L(f_n)$ using these bounds. The following theorem justifies a more direct approach (using discrepancy on the training data directly, rather than using discrepancy over a range of sample sizes to estimate effective VC dimension), and shows that an independent test sample is not necessary.

A similar estimate was considered in [49], although the error bound presented in [49, Theorem 3.4] can only be nontrivial when the maximum discrepancy is negative.

Theorem 6 *If the penalties are defined using the maximum-discrepancy error estimates (6), and $m = n/9$, then*

$$\begin{aligned}
&\mathbb{E}L(f_n) - L^* \\
&\quad \leq \min_k \left[\mathbb{E} \max_{f \in \mathcal{F}_k} \left(\widehat{L}_n^{(1)}(f) - \widehat{L}_n^{(2)}(f) \right) \right. \\
&\qquad\qquad \left. + \left(\inf_{f \in \mathcal{F}_k} L(f) - L^* \right) + 3\sqrt{\frac{\log k}{n}} \right] + \frac{2.13}{\sqrt{n}} .
\end{aligned}$$

Proof. Once again, we check Assumption 1 and apply Theorem 1. Introduce the ghost sample $(X_1', Y_1'), \ldots, (X_n', Y_n')$, which is independent of the data and has the same distribution. Denote the empirical loss based on this sample by $L_n'(f) = \frac{1}{n}\sum_{i=1}^n \ell(f(X_i'), Y_i')$. The proof is based on the simple observation that for each k,

$$
\begin{aligned}
&\mathbb{E}\max_{f\in\mathcal{F}_k}\left(L'_n(f)-\widehat{L}_n(f)\right)\\
&= \frac{1}{n}\mathbb{E}\max_{f\in\mathcal{F}_k}\sum_{i=1}^{n}\left(\ell(f(X'_i),Y'_i)-\ell(f(X_i),Y_i)\right)\\
&\le \frac{1}{n}\mathbb{E}\left(\max_{f\in\mathcal{F}_k}\sum_{i=1}^{n/2}\left(\ell(f(X'_i),Y'_i)-\ell(f(X_i),Y_i)\right)\right.\\
&\qquad\left.+\max_{f\in\mathcal{F}_k}\sum_{i=n/2+1}^{n}\left(\ell(f(X'_i),Y'_i)-\ell(f(X_i),Y_i)\right)\right)\\
&= \frac{2}{n}\mathbb{E}\max_{f\in\mathcal{F}_k}\sum_{i=1}^{n/2}\left(\ell(f(X'_i),Y'_i)-\ell(f(X_i),Y_i)\right)\\
&= \mathbb{E}\max_{f\in\mathcal{F}_k}\left(\widehat{L}_n^{(1)}(f)-\widehat{L}_n^{(2)}(f)\right). \qquad (7)
\end{aligned}
$$

Thus, for each k,

$$
\begin{aligned}
&\mathbb{P}\left[L(\widehat{f}_k)>R_{n,k}+\epsilon\right]\\
&= \mathbb{P}\left[L(\widehat{f}_k)-\widehat{L}_n(\widehat{f}_k)>\max_{f\in\mathcal{F}_k}\left(\widehat{L}_n^{(1)}(f)-\widehat{L}_n^{(2)}(f)\right)+\epsilon\right]\\
&\le \mathbb{P}\left[\sup_{f\in\mathcal{F}_k}\left(L(f)-\widehat{L}_n(f)\right)-\max_{f\in\mathcal{F}_k}\left(\widehat{L}_n^{(1)}(f)-\widehat{L}_n^{(2)}(f)\right)>\epsilon\right]\\
&\le \mathbb{P}\left[\sup_{f\in\mathcal{F}_k}\left(L(f)-\widehat{L}_n(f)\right)-\max_{f\in\mathcal{F}_k}\left(\widehat{L}_n^{(1)}(f)-\widehat{L}_n^{(2)}(f)\right)\right.\\
&\qquad\left.>\mathbb{E}\left(\sup_{f\in\mathcal{F}_k}\left(L(f)-\widehat{L}_n(f)\right)-\max_{f\in\mathcal{F}_k}\left(\widehat{L}_n^{(1)}(f)-\widehat{L}_n^{(2)}(f)\right)\right)+\epsilon\right].
\end{aligned}
$$

Now, the difference between the supremum and the maximum satisfies the conditions of McDiarmid's inequality (see Theorem 8 below) with $c_i = 3/n$, so this probability is no more than $\exp(-2\epsilon^2 n/9)$. Thus, Assumption 1 is satisfied with $m = n/9$ and $c = 1$, and the proof is finished. ∎

3.5 A randomized complexity estimator

In this section we introduce an alternative way of estimating the quantity $\mathbb{E}\max_{f\in\mathcal{F}_k}\left(L(f)-\widehat{L}_n(f)\right)$ which may serve as an effective estimate of the complexity of a model class $\mathcal{F}$. The maximum discrepancy estimate of the previous section does this by splitting the data into two halves. Here we offer an alternative which allows us to derive improved performance bounds: we consider the expectation, over a random split of the data into two sets, of the maximal discrepancy. Koltchinskii [21] considers a very similar estimate and proves a bound analogous to Theorem 7 below. We improve this bound further in Theorem **??**.

Let $\sigma_1,\ldots,\sigma_n$ be a sequence of i.i.d. random variables such that $\mathbb{P}\{\sigma_i=1\}=\mathbb{P}\{\sigma_i=-1\}=\frac{1}{2}$ and the σ_i's are independent of the data D_n. Introduce the quantity

$$M_{n,k}=\mathbb{E}\left[\sup_{f\in\mathcal{F}_k}\frac{2}{n}\sum_{i=1}^{n}\sigma_i\ell(f(X_i),Y_i)\middle|D_n\right]. \tag{8}$$

We use $M_{n,k}$ to measure the amount of overfitting in class $\mathcal{F}_k$. Note that $M_{n,k}$ is not known, but it may be computed with arbitrary precision by Monte-Carlo simulation. In the case of pattern classification, each computation in the integration involves minimizing empirical loss on a sample with randomly flipped labels. We offer two different ways of using these estimates for model selection. The first is based on Theorem 1 and the second, with a slight modification, on Theorem 2. We start with the simpler version:

Theorem 7 *Let $m=n/9$, and define the error estimates $R_{n,k}=\widehat{L}_n(\widehat{f}_k)+M_{n,k}$, and choose f_n by minimizing the penalized error estimates*

$$\tilde{L}_n(\widehat{f}_k)=\widehat{L}_n(\widehat{f}_k)+C_n(k)=R_{n,k}+\sqrt{\frac{\log k}{m}}.$$

then

$$\mathbb{E}L(f_n)-L^* \;\leq\; \min_k\left[\mathbb{E}M_{n,k}+\left(\inf_{f\in\mathcal{F}_k}L(f)-L^*\right)+3\sqrt{\frac{\log k}{n}}\;\right]+\frac{2.13}{\sqrt{n}}.$$

Proof. Introduce a ghost sample as in the proof of Theorem 6, and recall that by a symmetrization trick of Giné and Zinn [18],

$$
\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}_k}(L(f)-\widehat{L}_n(f))\right] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}_k}\mathbb{E}\left[L_n'(f)-\widehat{L}_n(f)\Big|D_n\right]\right] \\
&\le \mathbb{E}\left[\sup_{f\in\mathcal{F}_k}(L_n'(f)-\widehat{L}_n(f))\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}_k}\sum_{i=1}^n \sigma_i(\ell(f(X_i'),Y_i')-\ell(f(X_i),Y_i))\right] \\
&\le \frac{2}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}_k}\sum_{i=1}^n \sigma_i\ell(f(X_i),Y_i)\right] \\
&= \mathbb{E}M_{n,k}. \qquad (9)
\end{aligned}
$$

The rest of the proof of Assumption 1 follows easily from concentration inequalities: for each k,

$$
\begin{aligned}
&\mathbb{P}\left[L(\widehat{f}_k)>R_{n,k}+\epsilon\right] \\
&= \mathbb{P}\left[L(\widehat{f}_k)-\widehat{L}_n(\widehat{f}_k)>M_{n,k}+\epsilon\right] \\
&\le \mathbb{P}\left[\sup_{f\in\mathcal{F}_k}\Big(L(f)-\widehat{L}_n(f)\Big)-M_{n,k}>\epsilon\right] \\
&\le \mathbb{P}\Bigg[\sup_{f\in\mathcal{F}_k}\Big(L(f)-\widehat{L}_n(f)\Big)-M_{n,k} \\
&\qquad -\mathbb{E}\left(\sup_{f\in\mathcal{F}_k}\Big(L(f)-\widehat{L}_n(f)\Big)-M_{n,k}\right)>\epsilon\Bigg] \qquad \text{(by (9))} \\
&\le e^{-2n\epsilon^2/9},
\end{aligned}
$$

where at the last step we used McDiarmid's inequality. (It is easy to verify that the difference between the supremum and $M_{n,k}$ satisfies the condition of Theorem 8 with $c_i=3/n$.) Thus, Assumption 1 holds with $c=2$ and $m=n/9$. Theorem 1 implies the result. ∎

4 Concentration inequalities

Concentration-of-measure results are central to our analysis. These inequalities guarantee that certain functions of independent random variables are close to their mean. Here we recall the three inequalities we used in our proofs.

Theorem 8 (McDiarmid [31]). *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to R$ satisfies*

$$\sup_{\substack{x_1,\ldots,x_n,\\ x_i' \in A}} \left| f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \right| \leq c_i$$

for $1 \leq i \leq n$. Then for all $t > 0$

$$\mathbb{P}\left\{ f(X_1, \ldots, X_n) \geq \mathbb{E}f(X_1, \ldots, X_n) + t \right\} \leq e^{-2t^2 / \sum_{i=1}^n c_i^2}$$

and

$$\mathbb{P}\left\{ f(X_1, \ldots, X_n) \leq \mathbb{E}f(X_1, \ldots, X_n) - t \right\} \leq e^{-2t^2 / \sum_{i=1}^n c_i^2} .$$

McDiarmid's inequality is convenient when $f()$ has variance $\Theta(\sum_{i=1}^n c_i^2)$. In other situations when the variance of f is much smaller, the following inequality might be more appropriate.

Theorem 9 (Boucheron, Lugosi, and Massart [10]) *Suppose that $X_1, \ldots, X_n$ are independent random variables taking values in a set A, and that $f : A^n \to R$ is such that there exists a function $g : A^{n-1} \to R$ such that for all $x_1, \ldots, x_n \in A$*

(1) $f(x_1, \ldots, x_n) \geq 0$*;*

(2) $0 \leq f(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \leq 1$ *for all* $i = 1, \ldots, n$*;*

(3) $\sum_{i=1}^n [f(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)] \leq f(x_1, \ldots, x_n)$.

Then for any $t > 0$,

$$\mathbb{P}\left[f(X_1, \ldots, X_n) \geq \mathbb{E}f(X_1, \ldots, X_n) + t \right] \leq \exp\left[-\frac{t^2}{2\mathbb{E}f(X_1, \ldots, X_n) + 2t/3} \right]$$

and

$$\mathbb{P}\left[f(X_1, \ldots, X_n) \leq \mathbb{E}f(X_1, \ldots, X_n) - t \right] \leq \exp\left[-\frac{t^2}{2\mathbb{E}f(X_1, \ldots, X_n)} \right].$$

Moreover,

$$\mathbb{E}f(X_1, \ldots, X_n) \leq \log_2 \mathbb{E}\left[2^{f(X_1, \ldots, X_n)} \right] \leq \frac{1}{\log 2} \mathbb{E}f(X_1, \ldots, X_n).$$

5 Experimental comparison of empirical penalization criteria

5.1 The learning problem

In this section we report experimental comparison of some of the proposed model selection rules in the setup proposed by Kearns, Mansour, Ng, and Ron [20]. In this toy problem, the X_i's are drawn from the uniform distribution on the interval $[0,1]$. The class $\mathcal{F}_k$ is defined as the class of all functions $[0,1] \to \{0,1\}$ such that for each $f \in \mathcal{F}_k$ there exists a partition of $[0,1]$ into $k+1$ intervals such that f is constant over all these intervals. It is straightforward to check that the VC-dimension of $\mathcal{F}_k$ is $k+1$. Following [20], we assume that the "target function" f^* belongs to $\mathcal{F}_k$ for some unknown k and the label Y_i of each example X_i is obtained by flipping the value of $f^*(X_i)$ with probability $\eta \in [0, .5)$ where η denotes the noise level. Then clearly, for any function g:

$$L(g) = \eta + (1-2\eta)\mathbb{P}\{f \neq g\} \ .$$

What makes this simple learning problem especially convenient for experimental study is the fact that the computation of the minima of the empirical loss $\min_{f\in\mathcal{F}_k} \widehat{L}_n(f)$ for all $k \leq n$ can be performed in time $O(n \log n)$ using a dynamic programming algorithm described in [20]. Lozano [24] also reports an experimental comparison of model selection methods for the same problem.

In this paper we studied several penalized model selection techniques: a holdout (or cross-validation) method based on independent test sample, penalization based on the empirical VC entropy, a maximum discrepancy estimator, and a randomized complexity estimator. For the investigated learning problem it is easy to see that the empirical VC entropy $\log_2 S_k(X_1^n)$ of class $\mathcal{F}_k$ is almost surely a constant and equal to

$$1 + \log_2 \sum_{i=0}^{k} \binom{n-1}{i},$$

and therefore penalization based on the empirical VC entropy is essentially equivalent to the Guaranteed Risk Minimization (GRM) procedure proposed by Vapnik [44]. Thus, we do not investigate empirically this method. Note that Lozano [24] compares the GRM procedure with a method based on Rademacher penalties, very similar to our randomized complexity estimator and finds that Rademacher penalties systematically outperform the GRM procedure.

In [20], GRM is compared to the Minimum Description Length principle and the independent test sample technique which is regarded as a simplified cross-validation technique. The main message of [20] is that penalization techniques that only take into account the empirical loss and some structural properties of the models cannot compete with cross-validation for all sample sizes. On the contrary, our conclusion (based on experiments) is that data-based penalties perform favorably compared to penalties based on independent test data.

In the figures shown below we report experiments for three methods: (1) the Holdout method (HOLDOUT) bases its selection on $m = n/10$ extra independent samples as described in Section 3.1; (2) the Maximum Discrepancy (MD) method selects a model according to the method of Section 3.4 and (3) Rademacher penalization (RP) performs the randomized complexity method proposed in Section 3.5. When using Maximum Discrepancy (Section 3.4) in experiments, the penalties were:

$$\frac{1}{2} \max_{f \in \mathcal{F}_k} \left(\widehat{L}_n^{(1)}(f) - \widehat{L}_n^{(2)}(f) \right) .$$

We found that multiplying the penalty defined in Section 3.4 by $1/2$ provides superior performance. When using Randomized Complexity Estimators (Section 3.5), the penalties were:

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}_k} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \ell(f(X_i), Y_i) \middle| D_n \right] .$$

Note that in all experiments, the $\sqrt{\frac{\log k}{n}}$ or $\frac{\log k}{n}$ terms were omitted from penalties. For reasons of comparison, the performance of "oracle selection" is also shown on the pictures. This method selects a model by minimizing the true loss $L(\widehat{f}_k)$ among the empirical loss minimizers $\widehat{f}_k$ of all classes $\mathcal{F}_k$, $k = 1, 2, \ldots$.

The training error minimization algorithm described in [20] was implemented using the templates for priority queues and doubly linked lists provided by the LEDA library [32].

5.2 Results

The results are illustrated by the figures below. As a general conclusion, we may observe that the generalization error (i.e., true loss) obtained by methods MDP and RP are favorable compared to HOLDOUT. Even for sample sizes between 500 and 1000, the data-dependent penalization techniques perform as well as HOLDOUT. The data dependent penalization techniques exhibit less variance than HOLDOUT.

The main message of the paper is that good error estimation procedures provide good model selection methods. On the other hand, except for the HOLDOUT method, the data-dependent penalization methods do not try to estimate directly $L(\widehat{f}_k) - \widehat{L}_n(\widehat{f}_k)$, but rather $\sup_{f \in \mathcal{F}_k}(L(f) - \widehat{L}_n(f))$. The figures show that this is accurate when noise level is high and becomes rather inaccurate when noise level decreases. This is a strong incentive to explore further data-dependent penalization techniques that take into account the fact that not all parts of $\mathcal{F}_k$ are equally eligible for minimizing the empirical loss.

Acknowledgements

Thanks to Vincent Mirelli and Alex Smola for fruitful conversations, and thanks to the anonymous reviewers for useful suggestions.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[2] A.R. Barron. Logically smooth density estimation. Technical Report TR 56, Department of Statistics, Stanford University, 1985.

[3] A.R. Barron. Complexity regularization with application to artificial neural networks. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 561–576. NATO ASI Series, Kluwer Academic Publishers, Dordrecht, 1991.

[4] A.R. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related fields*, 113:301–413, 1999.

[5] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034–1054, 1991.

[6] P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, March 1998.

[7] P.L. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In B. Schölkopf, C. J. C. Burges, A. J. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*, pages 43–54. MIT Press, Cambridge, 1999.

[8] L. Birgé and P. Massart. From model selection to adaptive estimation. In E. Torgersen D. Pollard and G. Yang, editors, *Festschrift for Lucien Le Cam: Research papers in Probability and Statistics*, pages 55–87. Springer, New York, 1997.

[9] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4:329–375, 1998.

[10] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications in random combinatorics and learning. *Random Structures and Algorithms*, 16:277-292, 2000.

[11] K.L. Buescher and P.R. Kumar. Learning by canonical smooth estimation, Part I: Simultaneous estimation. *IEEE Transactions on Automatic Control*, 41:545–556, 1996.

[12] K.L. Buescher and P.R. Kumar. Learning by canonical smooth estimation, Part II: Learning and choice of model complexity. *IEEE Transactions on Automatic Control*, 41:557–569, 1996.

[13] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK, 2000.

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

[15] Y. Freund. Self bounding learning algorithms. *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 247–258, 1998.

[16] A.R. Gallant. *Nonlinear Statistical Models.* John Wiley, New York, 1987.

[17] S. Geman and C.R. Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *Annals of Statistics*, 10:401–414, 1982.

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

[19] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[20] M. Kearns, Y. Mansour, A.Y. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. In *Proceedings of the Eighth Annual ACM Workshop on Computational Learning Theory*, pages 21–30. Association for Computing Machinery, New York, 1995.

[21] V. Koltchinskii. Rademacher penalties and structural risk minimization. Manuscript, 1999.

[22] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. Manuscript, 2000.

[23] A. Krzyżak and T. Linder. Radial basis function networks and complexity regularization in function learning. *IEEE Transactions on Neural Networks*, 9:247–256, 1998.

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

[25] G. Lugosi and A. Nobel. Adaptive model selection using empirical complexities. *Annals of Statistics*, vol. 27, no.6, 1999.

[26] G. Lugosi and K. Zeger. Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, 41:677–678, 1995.

[27] G. Lugosi and K. Zeger. Concept learning using complexity regularization. *IEEE Transactions on Information Theory*, 42:48–54, 1996.

[28] C.L. Mallows. Some comments on c_p. *IEEE Technometrics*, 15:661–675, 1997.

[29] L. Mason, P. L. Bartlett, and J. Baxter. Improved generalization through explicit optimization of margins. *Machine Learning*, 38(3): 243–255, 2000.

[30] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la faculté des sciences de l'Université de Toulouse, Mathématiques*, série 6, **IX**, to appear.

[31] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, Cambridge, 1989.

[32] K. Mehlhorn and S. Naher. *Leda : A Platform for Combinatorial and Geometric Computing.* Cambridge University Press, 2000.

[33] R. Meir. Performance bounds for nonlinear time series prediction. In *Proceedings of the Tenth Annual ACM Workshop on Computational Learning Theory*, page 122–129. Association for Computing Machinery, New York, 1997.

[34] D.S. Modha and E. Masry. Minimum complexity regression estimation with weakly dependent observations. *IEEE Transactions on Information Theory*, 42:2133–2145, 1996.

[35] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11:416–431, 1983.

[36] R. E. Schapire, Y. Freund, P. L. Bartlett, and W. S. Lee. Boosting the margin : A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26(5):1651–1686, October 1998.

[37] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[38] 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(5):1926–1940, 1998.

[39] X. Shen and W.H. Wong. Convergence rate of sieve estimates. *Annals of Statistics*, 22:580–615, 1994.

[40] S.J. Szarek. On the best constants in the Khintchine inequality. *Studia Mathematica*, 63:197–208, 1976.

[41] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Etudes Sci. Publ. Math.* , 81:73–205, 1995.

[42] V.N. Vapnik. *Estimation of Dependencies Based on Empirical Data.* Springer-Verlag, New York, 1982.

[43] V.N. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, New York, 1995.

[44] V.N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

[45] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[46] 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.

[47] V. N. Vapnik, E. Levin, and Y. Le Cun. Measuring the VC-dimension of a learning machine. *Neural Computation*, 6(5):851–876, 1994.

[48] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*, Springer-Verlag, New York, 1996.

[49] R. C. Williamson, J. Shawe-Taylor, B. Schölkopf, and A. J. Smola. Sample based generalization bounds. NeuroCOLT Technical Report NC-TR-99-055.

[50] Y. Yang and A.R. Barron. An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, 44:95-116, 1998.

[51] Y. Yang and A.R. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27:1564-1599, 1999.

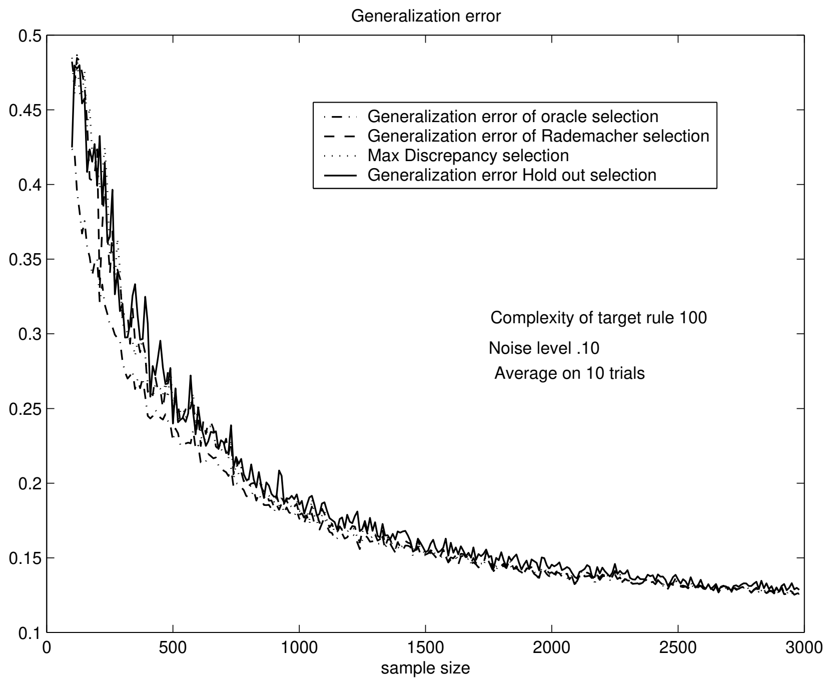

Figure 1: Note that all model selection techniques tend to be indistinguishable from the oracle selection method for samples larger than 1000. However, in contrast to the GRM estimate, the Rademacher and the Maximum Discrepancy selection methods are not outperformed by the HOLDOUT method for sample sizes smaller than 1000.

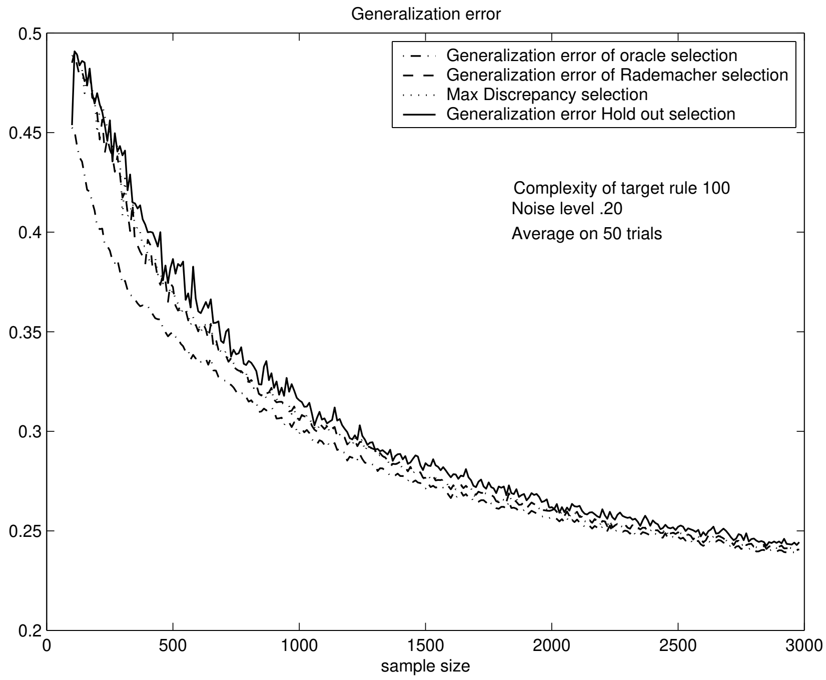

Figure 2: As the noise level is increased, the three model selection methods exhibit more variance and tend to be outperformed by the oracle for larger samples. HOLDOUT exhibits more variance than the other two penalization methods.

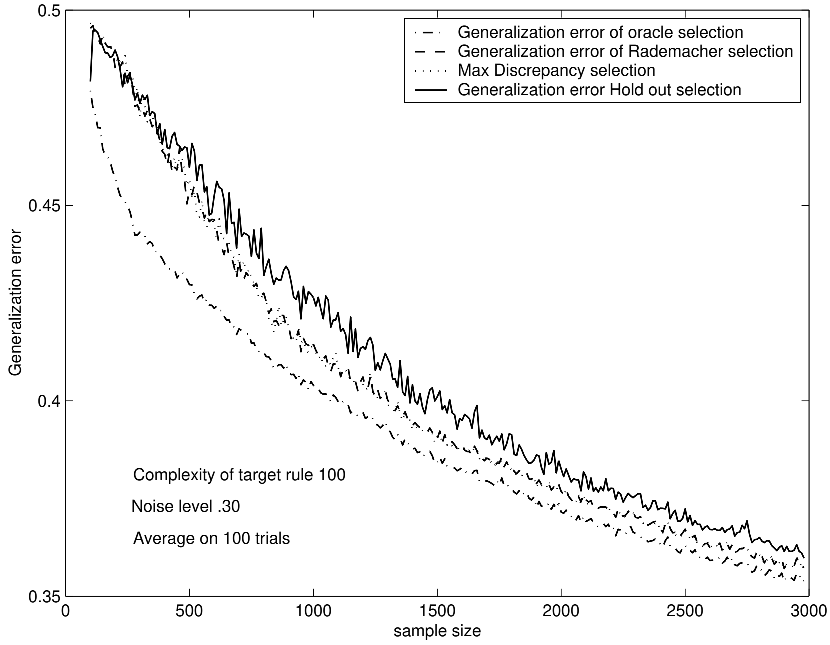

Figure 3: The oracle now has a clear edge on the model selection techniques for sample sizes smaller than 1000.

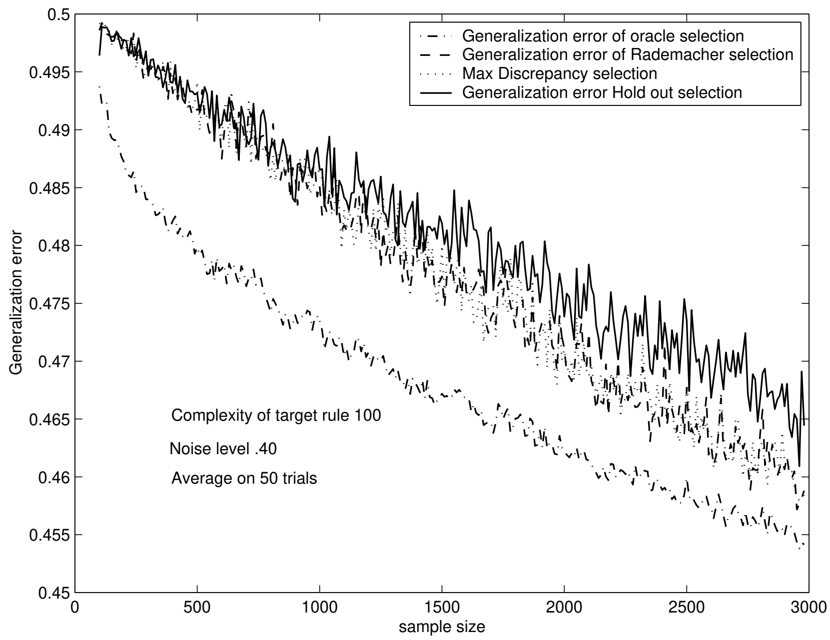

Figure 4: As the noise level becomes large, the three model selection methods remain distinguishable from the oracle for all shown sample sizes.

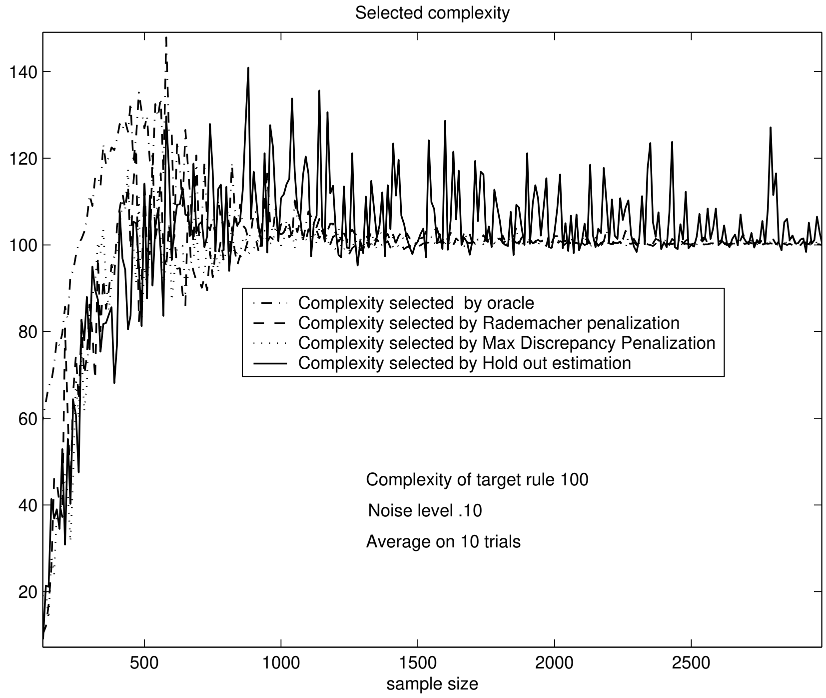

Figure 5: Each point represents the average complexity of the model selected by a given method or oracle at a given sample size. Note that for sample sizes between 500 and 1000, the oracle tends to overcode the sample. This corroborates the fact that liberal penalization methods (like MDL as used in [20]) tend to perform better than conservative methods (like GRM) for that range of sample sizes. Note also that HOLDOUT selection exhibits more variance than the two data-dependent penalty methods.

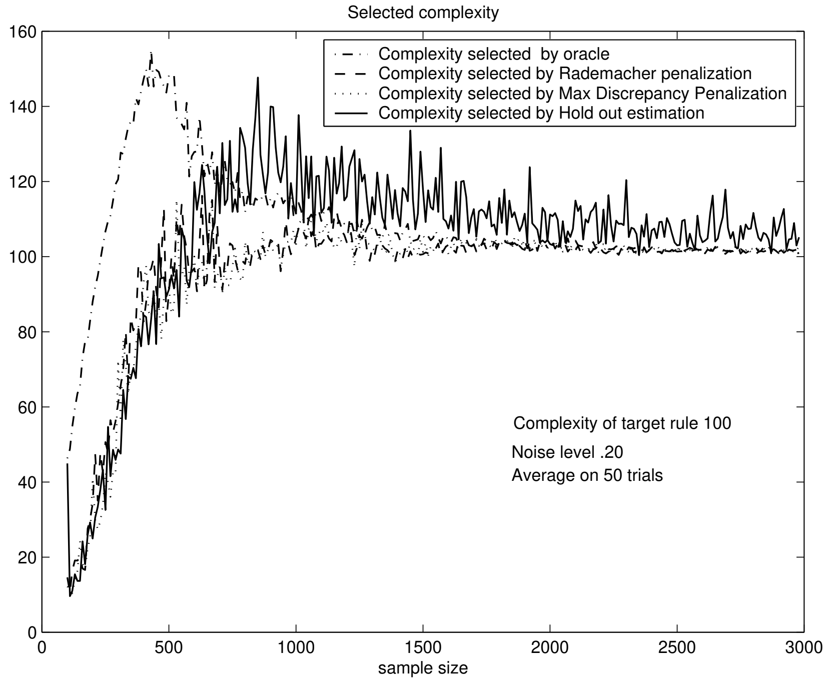

Figure 6: Selected class indices are shown at medium noise level

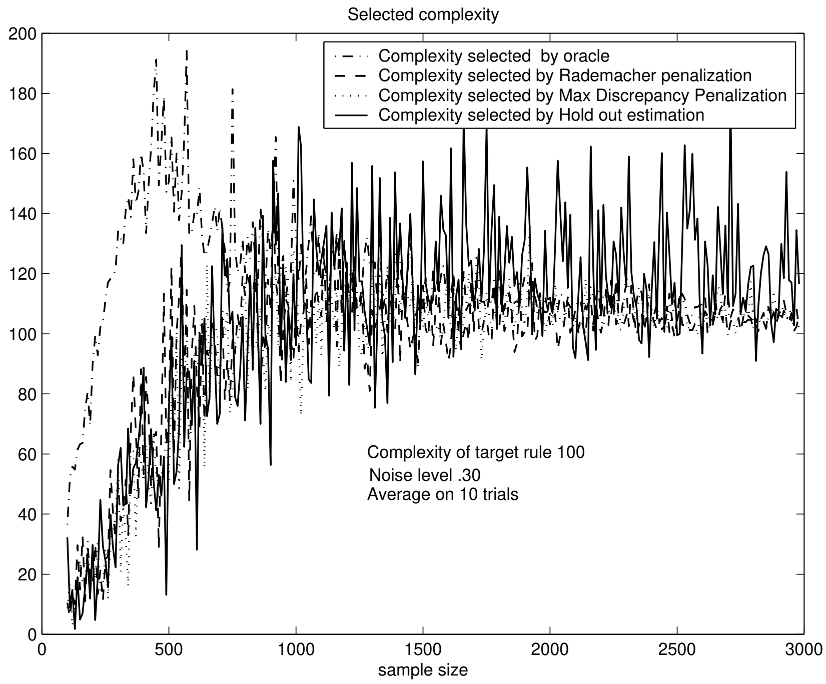

Figure 7: Selected class indices at higher noise level

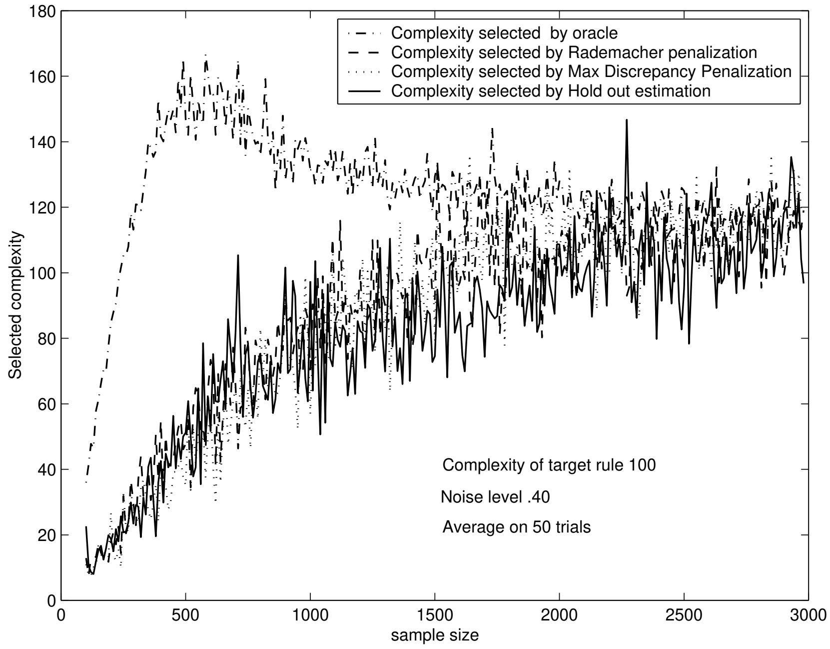

Figure 8: With increasing noise level, the propensity of model selection techniques to undercode and their increasing instability becomes more visible. Note that HOLDOUT is more sensitive to noise than its two competitors.

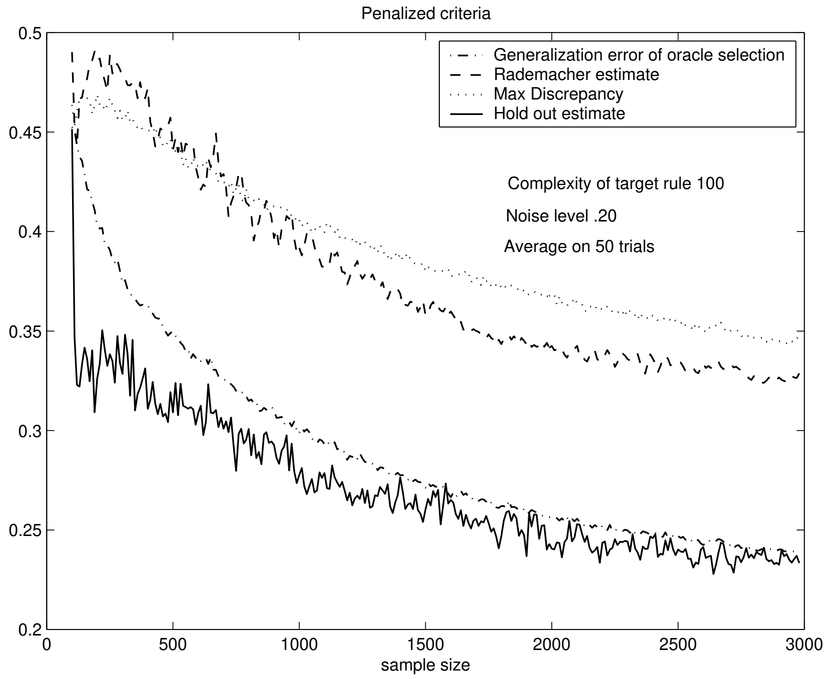

Figure 9: Here and in the next two figures, the minimal penalized empirical loss $\inf_k \tilde{L}_n(\widehat{f}_k)$ is shown as a function of the sample size for different levels of noise. At a low noise level, both Rademacher and Maximum discrepancy estimates overestimate the difference between the training and generalization errors. This phenomenon is due to the fact that these estimates deal with the maximum of the empirical process, which is only an upper bound on the the difference between the training and generalization errors. On the other hand, the HOLDOUT estimate remains optimistic for sample sizes smaller than 1000. Note that although the HOLDOUT estimate is unbiased for each particular class $\mathcal{F}_k$, the infimum of the HOLDOUT estimates over all classes $\mathcal{F}_k$ for $k \leq K = 200$ suffers an optimistic bias of order $\sqrt{\log K/n}$. This provides an explanation for the difference between the generalization error of the oracle selection and the HOLDOUT estimate.

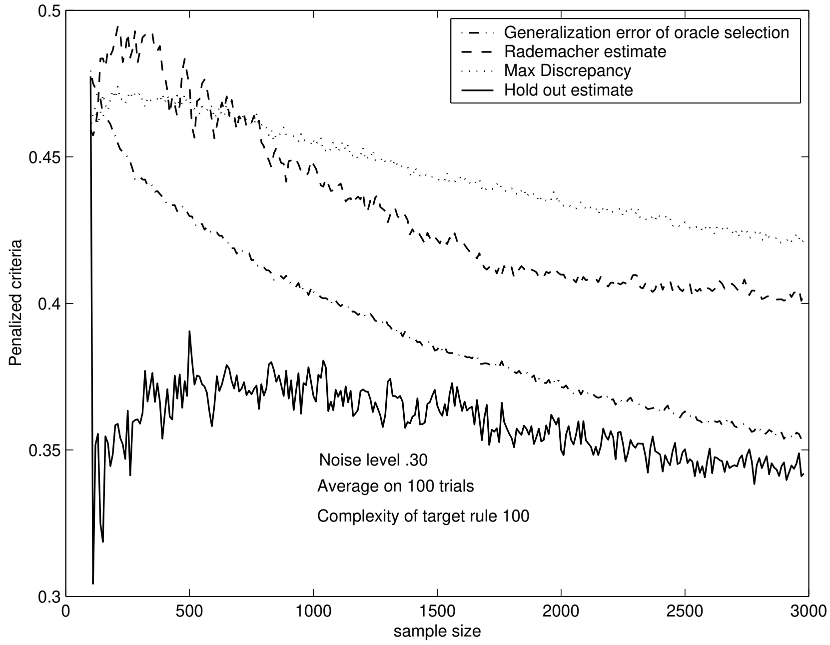

Figure 10: As noise increases, the pessimistic bias of the Rademacher and Maximum Discrepancy estimates becomes smaller.

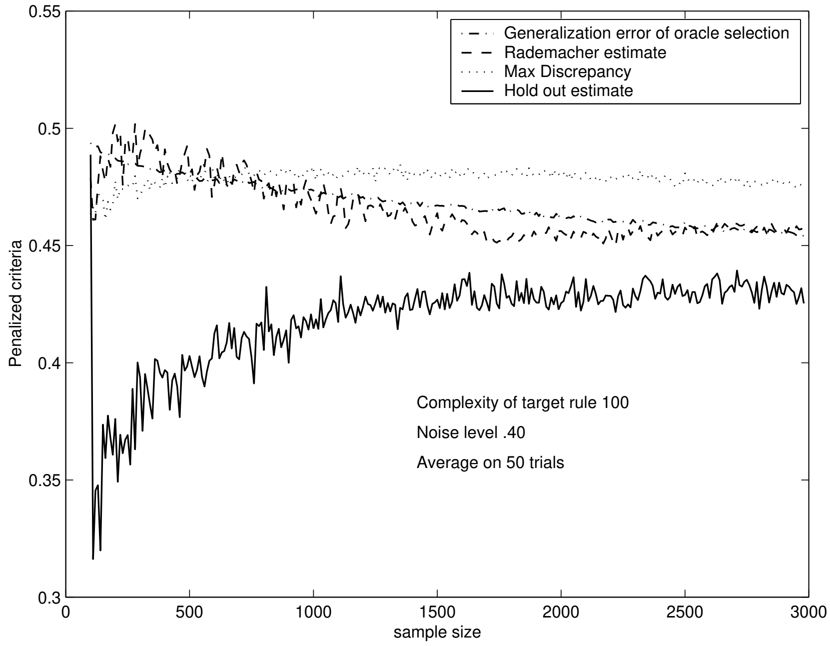

Figure 11: At high noise level, the Rademacher estimate becomes the most accurate approximation to the oracle. The HOLDOUT estimate is unable to catch the true value for samples smaller than 3000.