Adaptive Model Selection Using Empirical Complexities

Gábor Lugosi   Andrew B. Nobel

June 19, 1998

Gábor Lugosi is with the Department of Economics, Pompeu Fabra University, Barcelona. Email: lugosi@upf.es. His work was supported in part by DGES grant PB96-0300.

Andrew B. Nobel is with the Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260. Email: nobel@stat.unc.edu His work was supported in part by NSF grant DMS-9501926, and was completed in part while he was a Beckman Fellow at the Beckman Institute for Advanced Science and Technology, University of Illinois, U-C.

AMS 1991 subject classifications. Primary 62G07, 62G20. Secondary 62H30.

Key words and phrases. Complexity regularization, classification, pattern recognition, regression estimation, curve fitting, minimum description length.

Abstract

Given n independent replicates of a jointly distributed pair $(X,Y) \in \mathcal{R}^d \times \mathcal{R}$, we wish to select from a fixed sequence of model classes $\mathcal{F}_1, \mathcal{F}_2, \ldots$ a deterministic prediction rule $f : \mathcal{R}^d \to \mathcal{R}$ whose risk is small. We investigate the possibility of empirically assessing the *complexity* of each model class, that is, the actual difficulty of the estimation problem within each class. The estimated complexities are in turn used to define an adaptive model selection procedure, which is based on complexity penalized empirical risk.

The available data are divided into two parts. The first is used to form an empirical cover of each model class, and the second is used to select a candidate rule from each cover based on empirical risk. The covering radii are determined empirically to optimize a tight upper bound on the estimation error. An estimate is chosen from the list of candidates in order to minimize the sum of class complexity and empirical risk. A distinguishing feature of the approach is that the complexity of each model class is assessed empirically, based on the size of its empirical cover.

Finite sample performance bounds are established for the estimates, and these bounds are applied to several non-parametric estimation problems. The estimates are shown to achieve a favorable tradeoff between approximation and estimation error, and to perform as well as if the distribution-dependent complexities of the model classes were known beforehand. In addition, it is shown that the estimate can be consistent, and even possess near optimal rates of convergence, when each model class has an infinite VC or pseudo dimension.

For regression estimation with squared loss we modify our estimate to achieve a faster rate of convergence.

1 Introduction

Let $(X,Y) \in \mathcal{R}^d \times \mathcal{R}$ be a jointly distributed pair, where X represents the outcomes of several real or vector-valued predictors that are related to a real-valued response Y of interest. The relationship between X and Y will generally be stochastic: Y is not assumed to be a function of X. Any measurable function $f : \mathcal{R}^d \to \mathcal{R}$ acts as a deterministic *prediction rule* if $f(X)$ is used to estimate the value of Y.

Let $\ell : \mathcal{R} \times \mathcal{R} \to [0,\infty)$ be a nonnegative loss function having the interpretation that $\ell(y', y)$ measures the loss (or cost) incurred when the true value $Y = y$ is predicted to be y'. The performance of a prediction rule f will be assessed in terms of its expected loss, or risk,

$$L(f) = \mathbf{E}\ell(f(X), Y)\,.$$

The risk of every prediction rule is bounded below by the optimum value

$$L^* = \inf_f L(f) \geq 0\,,$$

where the infimum is taken over all measurable functions $f : \mathcal{R}^d \to \mathcal{R}$. Throughout the paper it is assumed that (X,Y) is such that $\ell(f(X),Y) \in [0,1]$ with probability one.

Constructing a good prediction rule from a finite data set is an important problem in both parametric and non-parametric statistics. Put more precisely, the task is as follows:

> Given a data set $T_n = (X_1,Y_1),\ldots,(X_n,Y_n)$ containing n i.i.d. replicates of the pair (X,Y), select a prediction rule $f : \mathcal{R}^d \to \mathcal{R}$ whose risk is small, in the sense that $L(f) \approx L^*$.

For convenience, the notation $Z = (X,Y)$, $Z_i = (X_i,Y_i)$, and $Z_1^n = T_n$ will be used in what follows.

1.1 Complexity of a model class

Many approaches to the general estimation problem restrict their search for a prediction rule to a constrained collection of functions $\mathcal{F}$ containing a finite or infinite number of prediction rules. In such a cases it is natural to replace the unknown joint distribution of (X,Y) by the empirical distribution of T_n, and to evaluate the performance of each prediction rule $f \in \mathcal{F}$ in terms of its empirical loss

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

Selecting a rule $f \in \mathcal{F}$ in order to minimize $\widehat{L}_n(f)$ is known as empirical risk minimization. To avoid minimization over an infinite set, one may discretize the class $\mathcal{F}$. A simple but suboptimal procedure is the following: Fix a positive number r, and select a finite subset $\mathcal{F}_r = \{f_1,\ldots,f_N\}$ of $\mathcal{F}$ such that for all $f \in \mathcal{F}$ there exists a $g \in \mathcal{F}_r$ with

$$\sup_{x\in\mathcal{R}^d, y\in\mathcal{R}} |\ell(f(x),y) - \ell(g(x),y)| \leq r.$$

(We assume for now that such a finite covering exists.) The smallest N such that this is possible is called the r-covering number of the class of functions

$$\mathcal{H} = \{h(x,y) = \ell(f(x),y) : f \in \mathcal{F}\}$$

with respect to the supremum norm. Denote this quantity by N_r, and assume that $|\mathcal{F}_r| = N_r$.

If f_n is that element of $\mathcal{F}_r$ having minimal empirical risk, then one may readily show that

$$\mathbf{E}L(f_n) - \inf_{f\in\mathcal{F}} L(f) \leq r + 2\mathbf{E}\left\{\max_{f\in\mathcal{F}_r}\left|\widehat{L}_n(f) - L(f)\right|\right\} \leq r + \sqrt{\frac{2\ln N_r}{n}}. \qquad (1)$$

The second inequality follows from the boundedness of the loss function, Hoeffding's (1963) inequality for the moment generating function of a sum of independent bounded random variables, and a standard bounding trick explained, for example, in Pollard (1989).

Since N_r is a monotone decreasing function of r, selecting the covering radius r such that $r \approx \sqrt{(2/n)\log N_r}$ approximately minimizes the upper bound (1). Indeed, if one defines $r' = \inf\left\{r > 0 : r \geq \sqrt{(2/n)\log N_r}\right\}$, then

$$\mathbf{E}L(f_n) - \inf_{f\in\mathcal{F}} L(f) \leq 2r' \leq 2\inf_r\left(r + \sqrt{\frac{2\log N_r}{n}}\right).$$

Thus r' might be called the balanced covering radius of the class $\mathcal{F}$ (with respect to the supremum norm). The quantity $2r'$ is a distribution-free upper bound on the difficulty of estimation in $\mathcal{F}$, and as such, r' may be considered as a measure of the complexity of $\mathcal{F}$. Though bounding the estimation error by r' may seem to be quite crude, it is often close to the best achievable distribution-free upper bound. In fact, the minimax rate of convergence is in many cases proportional to r' (see, e.g., Nicoleris and Yatracos (1997), Yang and Barron (1997)). Nevertheless, one may significantly improve the upper bound above in a distribution-dependent fashion.

Definition: Let $\mathcal{G}$ be a family of functions $g : \mathcal{S} \to \mathcal{R}$, let $s_1^n = s_1,\ldots,s_n$ be a sequence of points in $\mathcal{S}$, and let $r > 0$. A subset $\mathcal{G}_0 \subseteq \mathcal{G}$ is called an r-cover of $\mathcal{G}$ on

s_1^n if for every $g \in \mathcal{G}$ there exists a function $g' \in \mathcal{G}_0$ such that

$$\frac{1}{n}\sum_{j=1}^{n} |g(s_j) - g'(s_j)| \le r\,.$$

The *covering number* $N(s_1^n, r, \mathcal{G})$ is the size of the smallest r-cover of $\mathcal{G}$ on s_1^n. If no finite r-cover exists then $N(s_1^n, r, \mathcal{G}) = \infty$. Let $S_1, \ldots, S_n$ be n random elements taking values in $\mathcal{S}$. If $s_1 = S_1, \ldots, s_n = S_n$, then an r-cover $\mathcal{G}_0$ is called an *empirical cover* of $\mathcal{G}$ on S_1^n with radius r. The covering number $N(S_1^n, r, \mathcal{G})$ is a positive integer-valued random variable.

Replacing the data-independent sup-norm covering number N_r by the (smaller) expected covering numbers $\mathbf{E}N(Z_1^n, r, \mathcal{H})$, one may define an alternative balanced covering radius of $\mathcal{H}$ as follows:

$$\bar{r}_n = \inf\left\{ r : r \ \ge\ \sqrt{\frac{8\log \mathbf{E}N(Z_1^n, r/2, \mathcal{H})}{n}} \right\} \ \vee\ \sqrt{\frac{8}{n}}.$$

Here $a \vee b = \max(a, b)$. Given data T_n, let f_n' denote a function in $\mathcal{F}$ having minimal empirical risk. Then Lemma 2 in Section 6 shows that

$$\mathbf{E}L(f_n') - \inf_{f\in\mathcal{F}} L(f) \ \le\ 2\mathbf{E}\left[\sup_{f\in\mathcal{F}} |\widehat{L}_n(f) - L(f)|\right] \le 8\bar{r}_n,$$

Note that $\bar{r}_n$ depends critically on the (unknown) distribution of $Z = (X, Y)$. For certain "nice" distributions, $\bar{r}_n$ may be significantly smaller than the minimax risk associated with the class $\mathcal{F}$. In other words, the actual complexity of the estimation problem may be much less than the worst-case complexity, as measured by the minimax risk. This implies that adaptive model selection methods which assign a penalty to a model class based on its minimax risk will necessarily perform suboptimally for all such nice distributions. The purpose of this paper is to present a method that assesses the actual (distribution-dependent) balanced covering radius of each model class empirically, and then uses these radii to calculate data-based complexity penalties for adaptive model selection. Our estimates are based on empirical coverings of the model classes. A closely related approach to exploiting nice distributions is elaborated by Shawe-Taylor et al. (1997).

1.2 Adaptive model selection

Empirical risk minimization over a model class $\mathcal{F}$ 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. This trade-off between approximation error and estimation error is best understood by writing

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

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 $\mathcal{F}_1, \mathcal{F}_2, \ldots$ whose union is equal to $\mathcal{F}$. Given data T_n, one wishes to select a good model from *one* of these classes. Denote by $f_n^{(k)}$ a function in $\mathcal{F}_k$ having minimal empirical risk. If the distribution of (X, Y) were known in advance, one would select a model class $\mathcal{F}_K$ such that

$$\begin{aligned}\mathbf{E}L(f_n^{(K)}) - L^* &= \min_k \mathbf{E}L(f_n^{(k)}) - L^* \\ &= \min_k \left[\left(\mathbf{E}L(f_n^{(k)}) - \inf_{f\in\mathcal{F}_k} L(f)\right) + \left(\inf_{f\in\mathcal{F}_k} L(f) - L^*\right)\right].\end{aligned}$$

In the previous section it was shown that for each model class $\mathcal{F}_k$, a quite acceptable upper bound for the estimation error is given by

$$\mathbf{E}L(f_n^{(k)}) - \inf_{f\in\mathcal{F}_k} L(f) \ \le\ 8\bar{r}_n^{(k)}.$$

Here $\bar{r}_n^{(k)}$ denotes the balanced covering radius of the class $\mathcal{H}_k = \{\ell(f(x), y) : f \in \mathcal{F}_k\}$ with respect to Z_1^n, and is defined by

$$\bar{r}_n^{(k)} = \inf\left\{ r : r \ \ge\ \sqrt{\frac{8\log \mathbf{E}N(Z_1^n, r/2, \mathcal{H}_k)}{n}} \right\} \vee \sqrt{\frac{8}{n}}.$$

With this in mind, a slightly less ambitious goal of the model selection problem is to find an estimate g_n such that

$$\mathbf{E}L(g_n) - L^* \ \approx\ \min_k \left[8\bar{r}_n^{(k)} + \left(\inf_{f\in\mathcal{F}_k} L(f) - L^*\right)\right]. \qquad (2)$$

An estimate satisfying (2) achieves an optimal trade-off (over classes $\mathcal{F}_k$) between approximation error and a tight distribution-dependent upper bound on estimation error. The main difficulty in constructing such an estimate is that both $\bar{r}_n^{(k)}$ and the approximation error depend on the unknown distribution of (X, Y), and the optimal k is a complicated function of this distribution. The main result of the paper is the construction of an estimate which achieves this goal. The exact performance bound is given in Theorem 1 below.

Previous approaches to the model selection/prediction problem described above include Grenander's (1981) method of sieves, in which the classes $\mathcal{F}_i$ are nested, finite subsets of a fixed universal collection $\mathcal{F}$. Here, typically, the model class is selected in advance of the data, based only on the sample size n, in such a way that the model class gets richer as n increases, but that this increase of complexity is sufficiently slow so that the estimation error may be controlled.

Distribution-free consistency and rates of convergence for sieve-type estimates have been investigated, e.g., by Geman and Hwang (1982), Gallant (1987), Shen and Wong (1992), Wong and Shen (1992), Devroye (1988), White (1990), Lugosi and Zeger (1995), and Birgé and Massart (1998).

Complexity regularization, also known as structural risk minimization, extends the methodology of sieve estimates by using the data to choose the class from which the estimate is selected. Complexity regularization seeks to counter optimistic estimates of empirical risk by means of complexity penalties that favor simpler prediction rules, or rules belonging to smaller classes. In other words, the training set T_n is used to adaptively select both a model class $\mathcal{F}_k$ and a suitable prediction rule from that class. The potential advantages of such flexibility are clear. If a function minimizing $L(\cdot)$ lies in $\mathcal{F}_k$, then there is no point in searching for a rule in a larger class, which has a greater estimation error. On the other hand, when no rule f in a non-adaptively chosen class $\mathcal{F}_k$ minimizes $L(\cdot)$, the data may warrant consideration of a larger model class $\mathcal{F}_{k'}$ having better approximation capabilities. Early applications of complexity penalties to the problem of model selection were proposed by Mallows (1973), Akaike (1974), Vapnik and Chervonenkis (1974), and Schwarz (1978).

In the work of Rissanen (1983), Barron (1985), Wallace and Freeman (1987), and Barron and Cover (1991), the complexity penalty assigned to a model class is the length of a binary string describing the class. In this model, minimization of empirical risk plus complexity takes the form of a minimum description length principle. In this paper, as in the earlier work of Vapnik (1982), Barron (1991), Lugosi and Zeger (1996), and the recent work of Barron, Birgé, and Massart (1999), the complexity assigned to a model class does not have the formal interpretation of a description length, but is instead an upper bound on the estimation error of the class. For different applications and extensions of the same ideas we refer to Kearns et al. (1995), Krzyżak and Linder (1998), Meir (1997), Modha and Masry (1996), Shawe-Taylor et al. (1997), and Yang and Barron (1998).

Both the design and the analysis of penalized model fitting procedures rely on bounds for the complexity of the given model classes. As was mentioned above, worst-case assessments of model complexity are vulnerable to the fact that the complexity of a given model class can vary greatly with the underlying distribution of the pair (X, Y). For example, if the random vector X takes values in a finite set $\{x_1, \ldots, x_k\} \subseteq$

$\mathcal{R}^d$, then any model class $\mathcal{F}$ can be viewed as a subset $\{(f(x_1), \ldots, f(x_k)) : f \in \mathcal{F}\}$ of the finite dimensional space $\mathcal{R}^k$, where the dimension k is independent of the sample size n. Under these circumstances worst-case bounds on the complexity of $\mathcal{F}$ will be extremely pessimistic.

As the distribution of (X, Y) is unknown, any procedure that seeks to assess model complexity in a distribution-specific fashion must do so based on the data. In this paper we propose and analyze an adaptive model fitting procedure, which is based on data-dependent complexity penalties.

The available data are divided into two parts. The first is used to form an empirical cover of each model class, and the second is used to select a candidate rule from each cover having minimal empirical risk. The covering radii are determined empirically in order to optimize an upper bound on the estimation error. The empirical complexity of each model class is related to the cardinality of its empirical cover. An estimate g_n is chosen from among the countable list of candidates in order to minimize the sum of class complexity and empirical risk.

Estimates of this sort, based on empirical covering of model classes, were first proposed by Buescher and Kumar (1996a,b), who showed that empirical covering provides consistent learning rules whenever such rules exist.

Below inequalities and rates of convergence for the estimate g_n are established, and application of the estimates to a variety of problems, including nonparametric classification and regression, is considered. The proposed estimates achieve a favorable tradeoff between approximation and estimation error, and they perform as well as if the distribution-dependent complexities of the model classes were known beforehand.

1.3 Summary

Our principal assumptions, and several technical preliminaries are discussed in the next section. In Section 3 the complexity penalized estimator g_n is defined. A general upper bound on the performance of the estimator is given in Theorem 1, after which the relation of the bound to existing results is discussed.

In Section 4, some special cases, including regression function estimation under the L_2 loss, are considered. In these cases, by modifying the complexities assigned to each class, faster rates of convergence are achievable. An upper bound on the performance of the modified estimate is presented in Theorem 2.

Sections 5.1 to 5.5 contain applications of Theorem 1 to curve fitting and classification. In Section 5.6, the complexity-based estimate is employed as a means of fitting piecewise polynomial regression trees to multivariate data. The proofs of Theorems 1 and 2 appear in Section 6.

2 The AMSEC Estimate

2.1 Preliminaries and Assumptions

In what follows, a model class is any family $\mathcal{F}$ of prediction rules $f : \mathcal{R}^d \to \mathcal{R}$. It is assumed that a sequence of model classes

$$\mathcal{F}_1,\ \mathcal{F}_2,\ \mathcal{F}_3, \ldots \tag{3}$$

and a non-negative loss function $l : \mathcal{R} \times \mathcal{R} \to [0, \infty)$ have been fixed in advance. The model classes (3) need not be nested. For each model class $\mathcal{F}_k$ let

$$\mathcal{H}_k = \{h(x, y) = \ell(f(x), y) : f \in \mathcal{F}_k\} \tag{4}$$

be the associated family of error functions. By definition, each error function is non-negative. Each model class $\mathcal{F}_k$ is assumed to contain a countable subclass $\mathcal{F}_k^0$ with the property that every $f \in \mathcal{F}_k$ is a pointwise limit of a sequence of functions from $\mathcal{F}_k^0$. Each family $\mathcal{H}_k$ of error functions is assumed to have the same property. This ensures the measurability of random variables that are defined in terms of suprema or infima over the various classes (see Dudley (1978) for more details).

The data consist of n i.i.d. replicates of a jointly distributed pair $Z = (X, Y) \in \mathcal{R}^d \times \mathcal{R}$. Our principal assumption is that $l(y, y') \leq 1$ for each $y, y' \in \mathcal{R}$, or more generally, that

$$h(Z) \leq 1 \ \text{ with probability one for each error function } \ h \in \cup_{k=1}^{\infty} \mathcal{H}_k\,. \tag{5}$$

By suitably rescaling $\ell(\cdot, \cdot)$, one may ensure that the latter condition holds whenever there is a constant $B < \infty$ such that $h(Z) \leq B$ with probability one for every error function h. In other circumstances, it may be necessary to truncate $\ell(\cdot, \cdot)$, or to assume (e.g. in the case of absolute or squared loss) that the response variable Y is bounded.

If a uniform upper bound B on the error functions exists, but is unknown, one may define a modified estimator that employs a data-dependent rescaling of the loss function. Upper bounds on the performance of the modified estimator will be asymptotic in nature, and will involve distribution dependent constants involving the distribution of $h(Z)$. The condition of uniform boundedness may be replaced by conditions requiring rapidly decreasing tails of $h(Z)$, but for the sake of simplicity such cases are not discussed here.

Beyond boundedness of the error functions, no restrictions are placed on the joint distribution of (X, Y). In particular, the distribution of X is not assumed to be

absolutely continuous, nor is it assumed that the conditional distribution of Y given X is of some parametric form. No regularity or smoothness conditions are placed on the loss function $\ell(\cdot, \cdot)$.

3 Description of the estimate

The estimate is defined by first splitting the available data in half. The first half of the data is used to (i) select a suitable covering radius for each model class, and (ii) construct a suitable empirical cover of each model class using the selected radius. Each model class is assigned an empirical complexity that depends on the size of its empirical cover. The second half of the data is used to assess the empirical risk of a given classification rule. From the empirical cover of each class a candidate rule is selected having minimal empirical risk. The estimate is defined to be a candidate rule for which the sum of empirical risk and class complexity is minimized. A formal description of the estimate follows. Let $F_1, \mathcal{F}_2, \ldots$ be a fixed sequence of model classes.

Data Splitting. Ignoring the last sample point if necessary, assume without loss of generality that the size n of the available data is even. Split the data sequence into two parts of equal size,

$$Z_1, \ldots, Z_m \quad \text{and} \quad Z_{m+1}, \ldots, Z_n\,.$$

where $(n - m) = m = n/2$, and Z_i denotes the pair (X_i, Y_i).

Step 1: For each $k \geq 1$ consider the family $\mathcal{H}_k$ of error functions (4) associated with $\mathcal{F}_k$. Using the first half of the data, evaluate the *balanced empirical covering radius* of $\mathcal{H}_k$ as follows:

$$\hat{r}_m^{(k)} = \inf\left\{r : r \geq \sqrt{\frac{8 \log N\left(Z_1^m, r/2, \mathcal{H}_k\right)}{m}}\right\} \vee \sqrt{\frac{8}{m}}\,. \tag{6}$$

Here $a \vee b = \max(a, b)$. Let $\widehat{\mathcal{H}}_k$ be an empirical cover of $\mathcal{H}_k$ on $Z_1, \ldots, Z_m$ with radius $\hat{r}_m^{(k)}$ and minimal cardinality:

$$|\widehat{\mathcal{H}}_k| \ = \ N(Z_1^m, \hat{r}_m^{(k)}, \mathcal{H}_k)\,.$$

(Covering numbers and empirical covers are defined in Section 1.1 above.) Let $\widehat{\mathcal{F}}_k$ be a corresponding finite subset of $\mathcal{F}_k$ such that $\widehat{\mathcal{H}}_k = \{\ell(f(x), y) : f \in \widehat{\mathcal{F}}_k\}$ and $|\widehat{\mathcal{F}}_k| = |\widehat{\mathcal{H}}_k|$. Assign to the model class $\mathcal{F}_k$ the empirical complexity

$$\widehat{C}_{n-m}(k) = \sqrt{\frac{\log|\widehat{\mathcal{F}}_k| + 2\log k}{2(n-m)}} = \sqrt{\frac{\log N(Z_1^{n/2}, \hat{r}_m^{(k)}, \mathcal{H}_k) + 2\log k}{n}}$$

Note that $\widehat{\mathcal{F}}_k$ may be regarded as an empirical cover of $\mathcal{F}_k$ with respect to a metric that is determined by the loss function $\ell(\cdot,\cdot)$.

Step 2: Define the *empirical risk* of a prediction rule $f : \mathcal{R}^d \to \mathcal{R}$ to be the average loss of f on the second half of the data:

$$\widehat{L}_{n-m}(f) = \frac{1}{n-m} \sum_{i=m+1}^{n} \ell(f(X_i), Y_i) \, .$$

For each $j \geq 1$ let $\widehat{f}_j$ be a member of $\widehat{\mathcal{F}}_j$ having minimal empirical risk,

$$\widehat{f}_j = \arg\min_{f \in \widehat{\mathcal{F}}_j} \widehat{L}_{n-m}(f) \, .$$

Note that $\widehat{f}_j$ depends on $Z_1, \ldots, Z_m$ through the choice of $\widehat{\mathcal{F}}_j$, and on $Z_{m+1}, \ldots, Z_n$ through the definition of $\widehat{L}_{n-m}(\cdot)$.

Step 3: From each model class $\mathcal{F}_j$ there is a candidate rule $\widehat{f}_j$ that is chosen based on the available data. The estimate is chosen from the list of candidates $\widehat{f}_1, \widehat{f}_2, \ldots$ in order to minimize the sum of empirical risk and empirical class complexity. Define $g_n = \widehat{f}_k$ where

$$k = \arg\min_{j \geq 1} \left[\widehat{L}_{n-m}(\widehat{f}_j) + \widehat{C}_{n-m}(j) \right] \, . \tag{7}$$

Thus g_n is defined by means of adaptive model selection, using empirical complexities. It will be referred to as the AMSEC estimator in what follows. Observe that $\widehat{C}_{n-m}(j) \to \infty$ as $j \to \infty$. Since the empirical risks $\widehat{L}_{n-m}(\widehat{f}_j)$ are bounded above by 1, a minimizing index k must exist, and therefore g_n is well-defined.

Remark: We note that the estimate defined above will not, in general, be computationally feasible. This limitation arises principally from the difficulty of evaluating empirical covering numbers, and of selecting a minimal covering of a given radius.

The chosen prediction rule g_n comes from the union of the empirical covers $\widehat{\mathcal{F}} = \bigcup_{j=1}^{\infty} \widehat{\mathcal{F}}_j$. The underlying model classes $\mathcal{F}_j$ may overlap (if they are nested, for example), and therefore the covers $\widehat{\mathcal{F}}_j$ may not be disjoint. With this in mind one may define the complexity of each individual rule $f \in \widehat{\mathcal{F}}$ by

$$\widehat{\Delta}_{n-m}(f) = \min \left\{ \widehat{C}_{n-m}(j) : \text{ all } j \text{ such that } f \in \widehat{\mathcal{F}}_j \right\} \, .$$

Let g_n' be any function in $\widehat{\mathcal{F}}$ achieving an optimum trade-off between performance and complexity:

$$g_n' = \arg\min_{f \in \widehat{\mathcal{F}}} \left[\widehat{L}_{n-m}(f) + \widehat{\Delta}_{n-m}(f) \right] \, . \tag{8}$$

It is easy to show that any function achieving this minimum can be obtained via a two-stage optimization procedure similar to that described in steps 2 and 3 above. Thus the analysis of g_n applies to g_n' as well.

3.1 Performance of the estimate

Our initial bounds on the expected loss of the estimate g_n are given in terms of balanced covering radii for the families of error functions $\mathcal{H}_k$. The *balanced covering radius* of $\mathcal{H}_k$ with respect to $Z_1, \ldots, Z_m$ is defined by

$$\bar{r}_m^{(k)} = \inf \left\{ r : r \geq \sqrt{\frac{8 \log \mathbf{E} N\left(Z_1^m, r/2, \mathcal{H}_k\right)}{m}} \right\} \vee \sqrt{\frac{8}{m}}. \tag{9}$$

Recall that the optimal performance obtainable with any prediction rule is given by

$$L^* = \inf_f L(f),$$

where the infimum ranges over all measurable functions $f : \mathcal{R}^d \to \mathcal{R}$. Define also

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

to be the optimal performance of rules in the k'th model class. The following theorem gives expected performance bounds for the estimator g_n defined above.

Theorem 1 *Under the boundedness assumption (5), for each n the AMSEC estimate g_n is such that*

$$\mathbf{E} L(g_n) - L^* \leq \inf_{k \geq 1} \left[13.66 \bar{r}_{n/2}^{(k)} + 5.2 \sqrt{\frac{\log k}{n}} + (L_k^* - L^*) \right] .$$

Remark 1. The bound of Theorem 1 comes quite close to the goal set forth in (2). In addition to a larger constant (13.66 instead of 8), the balanced covering radii are now calculated at sample size $n/2$. The additional term $5.2\sqrt{\log k/n}$ is typically much smaller than the first term. The bounds in Theorem 1 and the corollaries that follow are non-asymptotic. They hold for every fixed sample size n. Thus, in principle, the sequence of model classes may change with sample size, that is each $\mathcal{F}_j$ may be replaced by $\mathcal{F}_{j,n}$.

Remark 2. To evaluate the performance bound in specific examples, one needs upper bounds for $\bar{r}_{n/2}^{(k)}$. Since

$$\bar{r}_{n/2}^{(k)} \leq \inf_r \max \left\{ r, \, 2\sqrt{\frac{\log \mathbf{E} N\left(Z_1^m, r/2, \mathcal{H}_k\right)}{n}}, \, \frac{4}{\sqrt{n}} \right\},$$

we see by taking $r = 4/\sqrt{n}$ for example, that

$$\bar{r}_{n/2}^{(k)} \leq 2\sqrt{\frac{\log \mathbf{E} N\left(Z_1^m, 2n^{-1/2}, \mathcal{H}_k\right)}{n}} \vee \frac{4}{\sqrt{n}}.$$

This inequality will be used in some of the applications below.

Remark 3. (Lipschitz loss.) A loss function $\ell(\cdot,\cdot)$ is called *Lipschitz* if there is a constant $M < \infty$ and a set $C \subset \mathcal{R}$, containing the range of every function in $\cup_{k=1}^{\infty} \mathcal{F}_k$, such that for every $y_1, y_2 \in C$ and every $v \in \mathcal{R}$,

$$|\ell(y_1, v) - \ell(y_2, v)| \leq M|y_1 - y_2|.$$

Note that the absolute loss $\ell(u, v) = |u - v|$ is Lipschitz, and that if ℓ is Lipschitz then for each pair $f, f' \in \mathcal{F}_k$,

$$|L(f) - L(f')| \leq M\mathbf{E}|f(X) - f'(X)|.$$

If $\ell(\cdot,\cdot)$ is Lipschitz, a straightforward argument shows that for every $k \geq 1$ and $r > 0$,

$$N\left(Z_1^{n/2}, r, \mathcal{H}_k\right) \leq N\left(X_1^{n/2}, Mr, \mathcal{F}_k\right).$$

This inequality will be used in some of the applications below.

Remark: For technical reasons, it is necessary to require that both $\hat{r}_m^{(k)}$ and $\bar{r}_m^{(k)}$ be at least $\sqrt{8/m}$. In all interesting situations $N(Z_1^n, r/2, \mathcal{H}_k) \geq 3$, and the maximum in the definition of the covering radii is achieved by the first term.

3.2 Discussion

Theorem 1 is similar in spirit to results of Barron and Cover (1991) and Barron (1991). In their work, there is for each sample size n, a fixed, countable list of candidate rules, each of which is assigned a data-independent complexity. They show that for each n the error of their estimate is bounded by a constant times an index of resolvability, which is the minimum, over all candidates, of the sum of approximation error and complexity. In a similar fashion, the bound of Theorem 1 measures the best possible tradeoff between complexity and approximation ability, and it too may be viewed as an index of resolvability. The crucial improvement here is the appearance of the distribution-dependent quantity $\bar{r}_{n/2}^{(k)}$ in Theorem 1 above.

In applications where the model classes $\mathcal{F}_1, \mathcal{F}_2, \ldots$ contain infinitely many functions, Barron and Cover (1991) and Barron (1991) assume that, for every fixed positive resolution, each class can be covered in supremum norm by finitely many functions.

For each n, their countable list of candidates is the union of the finite ϵ_n-covers of each class. While covering in the supremum norm ensures that the list will have good approximation properties under every distribution, for Lipschitz loss functions the appropriate measure of approximation is the metric of $L_1(P_X)$. Sup-norm covering numbers overestimate L_1 covering numbers, sometimes substantially, and thereby increase the index of resolvability.

In light of its equivalent definition g_n' above (8), it can be seen that our estimate selects, for each n, a countable list of candidate functions from $\mathcal{F}_1, \mathcal{F}_2, \ldots$ in a data-adaptive way. The list contains functions that have good approximation properties in the norm corresponding to the empirical distribution of $X_1, \ldots, X_n$. As a result, our upper bound is expressed in terms the expected L_1 covering numbers, rather than the sup-norm covering numbers.

In recent work, Barron, Birgé and Massart (1999) give an exhaustive review and a wide variety of sharp bounds for estimation procedures based on data-independent complexities. When each of the model classes $\mathcal{F}_k$ is both linear and finite-dimensional, their bounds improve those obtained below, and they obtain rates that differ from ours by a logarithmic factor. In earlier work on linear finite-dimensional model classes, Birgé and Massart (1997) defined a data-dependent complexity penalty different from the one considered here. In their penalty the observations are used to scale a data-independent term that involves the dimension of the model and the sample size. In both papers the complexity penalties derive from distribution-free upper bounds on the estimation error, which are based on the assumption that the individual model classes are finite-dimensional. Our method does not require the availability of such distribution-free bounds, or that each model class be finite dimensional. Indeed, the strength of our method is seen when neither of these conditions holds. Several examples are given in the next two sections.

4 The second estimate

As it was pointed out by Barron (1991), there are special cases, such as regression estimation with squared error loss, in which it may be advantageous to significantly decrease the size of the complexity penalties in order to achieve faster rates of convergence. In this spirit, a modification of the AMSEC estimate is proposed and analyzed below.

Consider the setup of the previous section. Let $\mathcal{F}_1, \mathcal{F}_2, \ldots$ be an arbitrary sequence of model classes and for each k let $r_k > 0$ be a data-independent covering radius for the k'th class. Given data $T_n = Z_1^n$ define classes $\widehat{\mathcal{F}}_k$ as in Step 1 of the previous section by empirically covering the $\mathcal{H}_k$'s with radii r_k, based on the first $m = n/2$ observations. Assign to $\mathcal{F}_k$ the complexity

$$\widehat{C}_{n-m}(k) = 22 \cdot \frac{\log|\widehat{\mathcal{F}}_k| + 2\log k}{n-m}.$$

Select from each $\widehat{\mathcal{F}}_k$ a candidate rule having minimal average risk on the last $n/2$ observations, and choose from among the candidates a rule ψ_n for which the sum of empirical risk and class complexity is minimized. Assume, as before, that $h(Z) \le 1$ with probability one for each $h \in \cup_{k=1}^{\infty}\mathcal{H}_k$.

Theorem 2 *Under the boundedness assumption above, the modified AMSEC estimate satisfies*

$$\mathbf{E}L(\psi_n) \;\le\; c_0 \inf_{k\ge1}\left[r_k + c_1\frac{\log \mathbf{E}N(Z_1^{n/2}, r_k/14, \mathcal{H}_k)}{n} + \frac{c_2\log k}{n} + L_k^*\right] + \frac{c_3}{n},$$

where $c_0, c_1, c_2, c_3 > 1$ are universal constants.

Remark 4. The principal improvement of Theorem 2 over Theorem 1 is that the complexity penalty

$$\bar{r}_{n/2}^{(k)} \approx \inf_r\left[r + 2\sqrt{\frac{\log \mathbf{E}N\left(Z_1^{n/2}, r/2, \mathcal{H}_k\right)}{n}}\right]$$

has now been replaced by $r_k + c_1 n^{-1}\log \mathbf{E}N(Z_1^{n/2}, r_k/14, \mathcal{H}_k)$, which is often much smaller. However, a price is paid for this improvement. Since the constant c_0 is strictly greater than one, subtracting L^* from both sides of the performance bound shows that Theorem 2 provides an asymptotic improvement over Theorem 1 only if $L^* = 0$. If $L^* > 0$ then $\inf_k L_k^*$ is necessarily positive, and the bound of Theorem 2 does not even guarantee consistency: it may happen that $\mathbf{E}L(\psi_n)$ does not converge to L^*. Nevertheless, the case $L^* = 0$ is interesting, and as shown below, Theorem 2 applies to the general situation in the case of squared error loss.

Remark 5. We have not attempted to find the optimal constants for Theorem 2. The values found in the proof below are $c_0 = 10$, $c_1 = 401$, $c_2 = 18$, and $c_3 = 10442$. These may be improved by a more careful analysis.

Remark 6. In the modified AMSEC estimate the covering radii r_k are fixed in advance of the data. As a consequence, the optimal balanced covering radii do not appear in Theorem 2. In certain cases satisfactory approximations can be found by investigating the model classes. For finite-dimensional model classes $r_k \approx n^{-1}$ is generally a good choice.

4.1 Regression function estimation

Consider the squared loss function $\ell(y', y) = (y' - y)^2$. In this case it is well known that for any bounded function $f : \mathcal{R}^d \to \mathcal{R}$,

$$L(f) = \mathbf{E}\left\{(f(X) - Y)^2\right\} = \mathbf{E}\left\{(f(X) - f^*(X))^2\right\} + \mathbf{E}\left\{(f^*(X) - Y)^2\right\},$$

where $f^*(x) = \mathbf{E}\{Y|X = x\}$ is the regression function of Y on X. Note that if Y and each candidate decision rule take values in the unit interval, then the boundedness assumption above is satisfied.

To study regression estimation in the context of Theorem 2 we introduce modified expected and empirical losses

$$J(f) = L(f) - L(f^*) \qquad \text{and} \qquad \widehat{J}_n(f) = \widehat{L}_n(f) - \widehat{L}_n(f^*).$$

If the regression function f^* is unknown, the empirical modified loss $\widehat{J}_n(f)$ cannot be calculated directly. However the AMSEC estimate ψ_n computed with the modified loss is the same as that computed using the unmodified squared loss as the term $\widehat{L}_n(f^*)$ is the same for each candidate rule. It follows from Theorem 2 that

$$\mathbf{E}J(\psi_n) \le c_0 \inf_{k\ge1}\left(r_k + c_1\frac{\log \mathbf{E}N(Z_1^{n/2}, r_k/14, \mathcal{H}_k)}{n} + \frac{c_2\log k}{n} + J_k^*\right) + \frac{c_3}{n},$$

where $J_k^* = \inf_{f\in\mathcal{F}_k} J(f)$. This readily implies the following performance bound for the AMSEC regression estimate: if $(f(X) - Y)^2 \le 1$ for each candidate prediction rule then

$$\mathbf{E}L(\psi_n) - L(f^*) \le c_0 \inf_{k\ge1}\left(r_k + c_1\frac{\log \mathbf{E}N(Z_1^{n/2}, r_k/14, \mathcal{H}_k)}{n} + \frac{c_2\log k}{n} + (L_k^* - L(f^*))\right) + \frac{c_3}{n}.$$

Thus, in the special case of regression function estimation with the squared loss, one may obtain improved rates of convergence even when $L^* = L(f^*) \ne 0$.

5 Applications

5.1 Finite dimensional classes

In many applications the model classes $\mathcal{F}_k$ are "finite dimensional," meaning that there exist numbers V_k, w_k such that for every sequence $z_1, \ldots, z_m \in \mathcal{R}^d \times \mathcal{R}$ and every $r > 0$, one has $N(z_1^m, r, \mathcal{H}_k) \le (w_k/r)^{V_k}$. The number V_k may be called the "dimension" of the model class $\mathcal{F}_k$. In this case the performance bound of Theorem 1 together with Remark 2 imply that

$$\mathbf{E}L(g_n) - L^* \le \min_{k\ge 1}\left\{ C\sqrt{\frac{V_k(\log n + \log w_k) + c_k}{n}} + (L_k^* - L^*)\right\}. \qquad (10)$$

For example, if $\mathcal{F}_k$ is a VC-graph class, then it is finite-dimensional in the above sense (see, e.g., Chapter 2. of Pollard (1984)).

When the numbers $V_1, w_1, V_2, w_2, \ldots$ are known in advance of the data, existing complexity-based methods offer similar, and in some specific cases (see Barron, Birgé, and Massart (1999)) better, performance bounds than those in (10) above. One advantage of the adaptive approach taken here is that it may be applied without the knowledge that the model classes are finite-dimensional, and without knowledge of the quantities w_k, V_k. More importantly, however, if for some distribution $\mathbf{E}N(Z_1^m, r, \mathcal{H}_k) \le (w_k'/r)^{V_k'}$ with $w_k' << w_k$ and $V_k' << V_k$, then we may replace w_k and V_k respectively by these smaller values in (10).

One might call V_k' the "effective dimension" of $\mathcal{F}_k$ with respect to the actual distribution of $Z = (X, Y)$. As V_k' is often significantly smaller than V_k, the new method will, in such cases, be superior to methods in which complexity penalties are based on distribution-free quantities. The new method is also able to handle "infinite-dimensional" model classes. One such example is sketched in the following section.

5.2 Piecewise monotone functions

Consider a one-dimensional curve fitting problem in which the k-th model class $\mathcal{F}_k$ contains all those functions $f : \mathcal{R} \to [-1/3, 1/3]$ comprised of k monotone pieces, that is, there exist numbers $u_1 \le \cdots \le u_{k-1}$ such that on each of the intervals $(-\infty, u_1], (u_1, u_2], \ldots, (u_{k-1}, \infty)$, f is either decreasing or increasing. It can be shown that none of the $\mathcal{F}_k$ is finite dimensional in the sense described above. Assume that the response variable is $Y = f^*(X) + W$, where f^* is an unknown function in $\cup_{k=1}^{\infty}\mathcal{F}_k$, and the random variable W is independent of X and such that $\mathbf{P}\{|W| \le 1/3\} = 1$, and the median of W equals zero. Let $\ell(\cdot,\cdot)$ be the absolute-error loss $\ell(y_1, y_2) = |y_1 - y_2|$. Thus the uniform boundedness assumption is satisfied, moreover $L^* = \mathbf{E}|W|$ and $\inf_{f\in\mathcal{F}_k} L(f) = L^*$ if k is so large that $f^* \in \mathcal{F}_k$. Under these assumptions the AMSEC estimator g_n satisfies the following inequality:

Proposition 1 *Let K be the least index k such that $f^* \in \mathcal{F}_k$. Then*

$$\mathbf{E}L(g_n) - L^* \le c\left(\sqrt{\frac{K\log n}{n}} + n^{-1/3}\sqrt{K\log n}\right),$$

where c is a universal constant.

The risk of g_n converges to zero at rate $n^{-1/3}\sqrt{\log n}$. Nemirovksii, Polyak, and Tsybakov (1985) showed that the minimax optimal rate of convergence for the class $\mathcal{F}_1$ is $n^{-1/3}$. Thus, the performance of the estimate g_n is at most a factor of $\sqrt{\log n}$ away from the optimal rate for all $\mathcal{F}_k$.

Proof: As the absolute-error loss is Lipschitz, for every sequence $z_1 = (x_1, y_1), \ldots, z_m = (x_m, y_m)$, every $r > 0$, and every $k \ge 1$ one has $N(z_1^m, r, \mathcal{H}_k) \le N(x_1^m, r, \mathcal{F}_k)$ (see Remark 3 above). To calculate an upper bound for $N(x_1^m, r, \mathcal{F}_k)$, it suffices to count the number of functions restricted to x_1^m, comprised of k monotone pieces, that take at most $N = \lceil 1/r \rceil$ distinct values. Now there are at most $\binom{m+k}{k-1}$ different ways of segmenting $x_1, \ldots, x_m$ into k pieces of lengths $m_1, \ldots, m_k$ with $\sum_{i=1}^k = m$ Since the number of monotone functions on m_i points taking N distinct values is at most $\binom{m_i+N+2}{m_i}$, for each $m \ge k$, and each $r > 0$

$$N(x_1^m, r, \mathcal{F}_k) \le \binom{m+k}{k-1}\cdot \max_{m_1+\cdots+m_k=m}\prod_{i=1}^k \binom{m_i+N+2}{m_i} \le (2m)^{k-1}\cdot(m+N+2)^{k(N+2)},$$

so that

$$\log N(x_1^m, r, \mathcal{F}_k) \le (k-1)\log(2m) + k\left(\frac{1}{r}+3\right)\log\left(m+\frac{1}{r}+3\right).$$

In conjunction with (9), the last bound shows that $r = cm^{-1/3}\sqrt{k\log m}$ is an upper bound for $\bar{r}_m^{(k)}$. As $m = n/2$ the bound stated above follows from Theorem 1. $\square$

5.3 Applications to classification

In the simplest version of the classification problem the response variable Y takes values in $\{0, 1\}$. A (binary) classification rule is any function $f : \mathcal{R}^d \to \{0, 1\}$. Under the absolute loss $\ell(y, y') = |y - y'|$, the risk of f is equal to its probability of error

$$L(f) = \mathbf{P}\{f(X) \ne Y\}.$$

The minimum probability of error L^* is achieved by the Bayes rule $f^*(x) = I\{\mathbf{P}(Y = 1|X = x) \ge 1/2\}$, where $I\{\cdot\}$ is the indicator function of the event in braces. The Bayes rule can be found when the joint distribution of (X, Y) is known.

In the remainder of this section, each model class $\mathcal{F}$ under consideration will be a family of binary classification rules. For each sequence of vectors $x_1, \ldots, x_m \in \mathcal{R}^d$, the shatter coefficient $S(x_1^m, \mathcal{F})$ of $\mathcal{F}$ is defined to be the cardinality of the set

$\{(f(x_1), \ldots, f(x_m)) : f \in \mathcal{F}\}$ of binary m-tuples. One may readily verify that for each $r > 0$,

$$N(x_1^m, r, \mathcal{F}) \leq S(x_1^m, \mathcal{F}) \, . \tag{11}$$

The Vapnik-Chervonenkis (or VC) dimension of $\mathcal{F}$, written $\dim(\mathcal{F})$, is the least integer m such that

$$\max\{S(x_1^m, \mathcal{F}) : x_1, \ldots, x_m \in \mathcal{R}^d\} < 2^m \, ,$$

and $\dim(\mathcal{F}) = \infty$ if no such m exists. It is well known (Vapnik and Chervonenkis, 1971) that for each $m \geq 1$ and each sequence $x_1, \ldots, x_m \in \mathcal{R}^d$,

$$S(x_1^m, \mathcal{F}) \leq m^{\dim(\mathcal{F})} + 1 \, . \tag{12}$$

It follows from (11) and (12) that if the VC-dimension of a model class $\mathcal{F}$ is finite, then its covering numbers are bounded by a polynomial in m that is independent of r.

Fix a sequence $\mathcal{F}_1, \mathcal{F}_2, \ldots$ of families of binary classification rules. If each family $\mathcal{F}_k$ has finite VC-dimension, then Theorem 1 gives useful bounds on the performance of the resulting estimator. Similar bounds were established by Lugosi and Zeger (1996) for an estimator that is based on the method of structural risk minimization proposed by Vapnik and Chervonenkis (1974). In both cases, construction of the corresponding estimator requires that bounds on the dimension of each model class be known in advance of the data. The following performance bound for the AMSEC estimate is an immediate consequence of Theorem 1 and Remarks 2 and 3:

Corollary 1 *Let g_n be the AMSEC estimator for $\mathcal{F}_1, \mathcal{F}_2, \ldots$ based on independent observations $(X_1, Y_1), \ldots (X_n, Y_n)$. If $V_k = \dim(\mathcal{F}_k)$ for each k, then*

$$\mathbf{E}L(g_n) - L^* \leq \min_{k \geq 1} \left\{ 55\sqrt{\frac{V_k \log n}{n}} + 5.2\sqrt{\frac{\log k}{n}} + \left(\inf_{f \in \mathcal{F}_k} L(f) - L^* \right) \right\} ,$$

and the upper bound is non-trivial if some V_k is finite.

Comparison of this result with Theorem 1 of Lugosi and Zeger (1996) shows that the AMSEC estimate, which is based solely on empirical complexities, works as well as the method of structural risk minimization, in which complexity penalties are assigned according to the (known) dimension of each class. More importantly, the arguments above give also an analogous bound with $V_k \log n$ replaced by $\log \mathbf{E}S(X_1^m, \mathcal{F}_k)$. In some cases $\mathbf{E}S(X_1^m, \mathcal{F})$ is significantly smaller than the maximum of $S(x_1^m, \mathcal{F})$ over all m-length vector sequences. Estimates based on data-dependent complexities can perform well even if each model class $\mathcal{F}_k$ has infinite VC-dimension.

5.4 Unions of convex sets

In this section we stay in the framework of classification discussed above, but now we consider certain model classes with infinite VC-dimension, for which the results of the previous section cannot be applied. For $k = 1, 2, \ldots$ let $\mathcal{F}_k$ contain the indicator function of each set $C \subseteq \mathcal{R}^d$ that is equal to the union of at most k convex sets. The VC-dimension of each class $\mathcal{F}_k$ is infinite. However, if the distribution of X has a density with respect to Lebesgue measure then there exist constants $\{b_m\}$, depending on the density of X, such that $b_m/m \to 0$ and $\mathbf{E}S(X_1^m, \mathcal{F}_1) \leq 2^{b_m}$ for each $m \geq 1$, see Devroye, Györfi and Lugosi, 1996, Section 13.4. An inspection of their proof shows, in addition, that for each $k \geq 1$,

$$\mathbf{E}\left\{ S(X_1^m, \mathcal{F}_1)^k \right\} \leq 2^{kb_m} \, . \tag{13}$$

Elementary combinatorial arguments like those in Chapter 2 of Pollard (1984) show that for each k and each sequence $x_1, \ldots, x_m \in \mathcal{R}^d$, $S(x_1^m, \mathcal{F}_k) \leq S(x_1^m, \mathcal{F}_1)^k$. Therefore,

$$\log \mathbf{E}N(X_1^m, m^{-1/2}, \mathcal{F}_k) \;\leq\; \log \mathbf{E}S(X_1^m, \mathcal{F}_k) \;\leq\; kb_m \;=\; ko(m).$$

Moreover, for each distribution of (X, Y), $\inf_{f \in \mathcal{F}_k} L(f) \to L^*$ as $k \to \infty$ since any subset of $\mathcal{R}^d$ can be approximated in the symmetric difference metric by a finite union of convex sets. Combining the last two observations with Theorem 1, one may establish the following result:

Proposition 2 *If the distribution of X is absolutely continuous, the AMSEC estimates g_n for $\mathcal{F}_1, \mathcal{F}_2, \ldots$ are Bayes risk consistent, that is, $\mathbf{E}L(g_n) \to L^*$ as the sample size $n \to \infty$.*

Remark: There is at least one special case in which it is possible to obtain rates of convergence for the estimates of Proposition 2. Suppose that $d = 2$ and that X has a bounded density with bounded support. Then it is known (c.f. Devroye, Györfi, and Lugosi, 1996, Section 13.4) that $b_m \leq c\sqrt{m}$, where $c > 0$ depends only on the distribution of X. Under these additional assumptions Theorem 1 shows that

$$\mathbf{E}L(g_n) - L^* \leq \inf_k \left(c' k^{1/2} n^{-1/4} + (L_k^* - L^*) \right)$$

for some universal constant c'. In computational learning theory it is common to assume that $L^* = 0$ and moreover that $f^* \in \cup_{k=1}^{\infty} \mathcal{F}_k$. In such cases, choosing $r_k = 14/n$, Theorem 2 may be applied to show that the modified estimate ψ_n achieves

$$\mathbf{E}L(\psi_n) \leq c'' \frac{K}{\sqrt{n}},$$

where K is the smallest index k such that $f^* \in \mathcal{F}_k$ and c'' is another universal constant.

5.5 Discrete distributions

If the common distribution of the predictors $X_1, X_2, \ldots$ is discrete then, under mild conditions, simple classification schemes such as empirical minimization are consistent regardless of the model class $\mathcal{F}$ from which prediction rules are selected. Under the same circumstances, the more adaptive procedure considered here exhibit similar behavior. It is shown below that the effective dimension of a model class $\mathcal{F}$ with respect to a sequence $X_1, \ldots, X_m$ is bounded by the number of distinct elements in that sequence. The proposed estimation method exploits this reduction of complexity adaptively, without prior knowledge of X or the model classes $\mathcal{F}_k$. Application of Theorem 1 requires a preliminary result.

Proposition 3 *Let $W_1, W_2, \ldots, W$ be i.i.d. integer-valued random variables, with probabilities $p_k = \mathbf{P}\{W = k\}$ for $k \geq 1$. Let M_n be the number of distinct integers appearing in the sequence $W_1, \ldots, W_n$. Then*

$$\lim_{n\to\infty} n^{-1} \log \mathbf{E} 2^{M_n} = 0. \tag{14}$$

If $\mathbf{E}W_1 = \sum_{k=1}^{\infty} k p_k < \infty$ then

$$\lim_{n\to\infty} n^{-1/2} \log \mathbf{E} 2^{M_n} = 0. \tag{15}$$

Proof: Note that for every integer $k \geq 1$,

$$M_n \leq k + \sum_{i=1}^{n} I\{W_i > k\}\,.$$

From this last inequality and the independence of $W_1, \ldots, W_n$ it follows that

$$\begin{aligned} \mathbf{E}2^{M_n} &\leq 2^k \cdot \left(\mathbf{E}2^{I\{W>k\}}\right)^n \\ &= 2^k \cdot (1 + 2\mathbf{P}\{W > k\})^n \\ &\leq \exp\left[\, k + 2n\mathbf{P}\{W > k\}\,\right]. \end{aligned} \tag{16}$$

Therefore,

$$n^{-1} \log \mathbf{E}2^{M_n} \;\leq\; \frac{k}{n} + 2\mathbf{P}\{W > k\}$$

and letting n tend to infinity,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}2^{M_n} \;\leq\; 2\mathbf{P}\{W > k\}.$$

Suitable choice of k insures that the right-hand side of the last inequality is arbitrarily close to zero, and (14) follows.

To establish (15), note that if $\sum_{k=1}^{\infty} k p_k < \infty$ then $\lim_{k\to\infty} k \cdot \mathbf{P}\{W > k/j\} = 0$ for every fixed positive integer j. Set $N_0 = 1$ and for each $j \geq 1$ let N_j be the least integer $N > N_{j-1}$ such that for every $n \geq N$,

$$n^{1/2}\,\mathbf{P}\left\{W > \frac{n^{1/2}}{j}\right\} \;\leq\; \frac{1}{j}\,.$$

Therefore

$$k_n = \frac{n^{1/2}}{\max\{j : N_j \leq n\}}\,.$$

is such that $k_n = o(n^{1/2})$ and $n^{1/2}\,\mathbf{P}\{W > k_n\} = o(1)$. It follows from (16) with $k = k_n$ that

$$n^{-1/2} \log \mathbf{E}2^{M_n} \leq \frac{k_n}{n^{1/2}} + 2n^{1/2}\,\mathbf{P}\{W > k_n\} \;=\; o(1).$$

□

Proposition 4 *Let g_n be the n-sample AMSEC estimator for an arbitrary sequence $\mathcal{F}_1, \mathcal{F}_2, \ldots$ of families of binary-valued prediction rules. If the distribution of X is supported on a countable set $S \subseteq \mathcal{R}^d$ then the following implications hold.*

1. *If the Bayes rule f^* is in the L_1 closure of $\bigcup_k \mathcal{F}_k$ then $\mathbf{E}L(g_n) \to L^*$.*
2. *If the elements of S may be ordered as $x_1, x_2, \ldots$ in such a way that $\sum_{k=1}^{\infty} kP(x_k)$ is finite, and if $f^* \in \bigcup_k \mathcal{F}_k$, then $\mathbf{E}L(g_n) \leq L^* + O(n^{-1/4})$.*

Proof: Define $W_i = \sum_{j=1}^{\infty} jI\{X_i = x_j\}$ and fix $k \geq 1$. The shatter coefficient of $\mathcal{F}_k$ on X_1^m is at most $\#\{(f(X_1), \ldots, f(X_m)) : f \in \mathcal{F}\} \leq 2^{M_m}$, where M_m is the number of distinct integers among $W_1, \ldots, W_m$. Thus, for every $r > 0$,

$$\mathbf{E}N(X_1^m, r, \mathcal{F}_k) \;\leq\; \mathbf{E}S(X_1^m, \mathcal{F}_k) \;\leq\; \mathbf{E}2^{M_m}\,,$$

and it follows from (14) that $n^{-1} \log \mathbf{E}N(X_1^{n/2}, n^{-1/2}, \mathcal{F}_k) \to 0$. In conjunction with Theorem 1 and Remark 3, this last relation implies

$$\limsup_{n\to\infty} \mathbf{E}L(g_n) - L^* \;\leq\; L_k^* - L^* \;\leq\; \inf_{f\in\mathcal{F}_k} \mathbf{E}|f - f^*|\,.$$

Letting k tend to infinity establishes the first conclusion of the proposition. To establish the second, let K be any index such that $f^* \in \mathcal{F}_K$. By the bound on the expected covering numbers above,

$$\mathbf{E}L(g_n) - L^* \leq c \cdot \left[\sqrt{\frac{\log \mathbf{E}S(X_1^{n/2}, \mathcal{F}_K)}{n}} + \sqrt{\frac{\log K}{n}}\right]$$

for every $n \geq 1$. Equation (15) implies that the first term in brackets is $O(n^{-1/4})$. □

Remark: Note that no conditions have been placed on the model classes $\mathcal{F}_k$, which can be arbitrarily complex.

5.6 Piecewise polynomial regression trees

Here the modified estimate of the previous section is used to fit piecewise polynomial regression trees to multivariate data, when the unknown regression function f^* is smooth, in the sense that it possesses continuous partial derivatives of some unknown order.

Piecewise polynomial regression trees are most naturally described by doubly indexed model classes. The class $\mathcal{F}_{k,p}$ contains functions $f : \mathcal{R}^d \to \mathcal{R}$ that are obtained by (i) forming a hierarchical (tree-structured) partition of $\mathcal{R}^d$ with k cells and then (ii) assigning a truncated multivariate polynomial of degree p to each cell. In selecting a suitable model, the procedure must choose both the number of cells k and the degree of local approximation p. Increasing p enables the procedure to more accurately reproduce the empirical behavior of the data within each cell, while increasing k allows for smaller cells. Balancing these choices against the estimation error of the resulting models, the complexity penalized regression procedure adapts to the unknown regularity of the regression function. Its success is reflected in its rate of convergence, which is within a logarithmic factor of optimal.

A *tree-structured partition* is described by a pair (T, τ), where T is a finite binary tree, and τ is a function that assigns a test vector $\tau(t) \in \mathcal{R}^d$ to every node $t \in T$. Every vector $x \in \mathcal{R}^d$ is associated, through a sequence of binary comparisons, with a descending path in T. Beginning at the root, and at each subsequent internal node of T, x moves to that child of the current node whose test vector is nearest to x in Euclidean distance. In case of ties, x moves to the left child of the current node. The path ends at a terminal node (leaf) of T.

For each node $t \in T$, let U_t be the set of vectors x whose path includes t. If t is the root node of T then $U_t = \mathcal{R}^d$. In general, the region U_t corresponding to an internal node of T is split between the children of that node by the hyperplane that forms the perpendicular bisector of their test vectors. Thus if t is at distance k from the root, then U_t is a polytope having at most k faces. The pair (T, τ) generates a partition π of $\mathcal{R}^d$, whose cells are the regions U_t associated with the terminal nodes of T. Let $\mathcal{T}_k$ contain all those partitions generated by binary trees T having k terminal nodes.

If at each internal node of T the comparison between the test vectors labeling its children involves a *single* coordinate of x, then each cell of the resulting partition is a

d-dimensional rectangle. Partitions of this sort, based on axis-parallel splits, are the basis for the regression trees considered by Breiman, Friedman, Olshen, and Stone (1984).

For each vector $u = (u_1, \ldots, u_d) \in \mathcal{R}^d$ and each sequence $\alpha = (\alpha_1, \ldots, \alpha_d)$ of non-negative integers, let $u^\alpha = u_1^{\alpha_1} \cdots u_d^{\alpha_d}$ and $|\alpha| = \alpha_1 + \cdots + \alpha_d$. For each $p \geq 0$ let

$$\mathcal{G}_p = \left\{ g(x) = \sum_{|\alpha| \leq p} a_\alpha x^\alpha \ : \ a_\alpha \in \mathcal{R} \right\}$$

be the class of multivariate polynomials on $\mathcal{R}^d$ of order p. Assuming that the response variable $Y \in [-1/2, 1/2]$, define the class of truncated polynomials

$$\tilde{\mathcal{G}}_p = \{ (g(\cdot) \wedge 1/2) \vee (-1/2) \ : \ g \in \mathcal{G} \} .$$

A k-node piecewise polynomial regression tree with local order p is a function $f : \mathcal{R}^d \to \mathcal{R}$ of the form

$$f(x) = \sum_{U \in \pi} g_U(x) I_U(x)$$

where $\pi \in \mathcal{T}_k$, and $g_U \in \tilde{\mathcal{G}}_p$ for each $U \in \pi$. In other words, f is obtained by applying a different truncated multivariate polynomial in $\tilde{\mathcal{G}}_p$ within each cell of a partition in $\mathcal{T}_k$. For each pair $k, p \geq 0$ let $\mathcal{F}_{k,p}$ contain all the k-node piecewise polynomial regression trees with local degree p. Let g_n be the complexity penalized regression estimate defined using $\{\mathcal{F}_{k,p} : k, p \geq 0\}$ with covering radii $r_k = 1/n$ as in Section 4 above.

Proposition 5 *If the common distribution P of the measurement vectors X_i is supported on a bounded set $S \subseteq \mathcal{R}^d$, if each $Y_i \in [-1/2, 1/2]$, and if the regression function f^* has continuous partial derivatives of order $s \geq 1$ on some open set containing S, then*

$$\mathbf{E}L(g_n) - L(f^*) \ = \ \mathbf{E}\left[\int |g_n - f^*|^2 dP \right] \ \leq \ C(s, d) \left[\frac{\log n}{n} \right]^{\frac{2s}{2s+d}} ,$$

where the constant $C(s, d)$ is independent of n.

Results of Stone (1982) show that the rate of convergence obtained here is, within a logarithmic factor, minimax optimal simultaneously for all r. Breiman et al. (1984) and Gordon and Olshen (1984) gave sufficient conditions for the L_2 and a.s. consistency of piecewise constant (e.g., $p = 0$) regression trees with rectangular cells. Their conditions stipulate that the cells of the selected partitions must shrink with increasing sample size, and that each cell must contain a minimum number of measurement vectors. Under additional conditions, Chaudhuri et al. (1994) established the consistency of piecewise polynomial regression trees with rectangular cells and fixed local degree p. Each of these results applies to unbounded response variables under suitable moment restrictions. Nobel (1996) considered the consistency of the general polynomial regression trees described above when the approximation degree p is fixed.

Proof: We consider, in turn, the estimation and approximation properties of the model classes $\mathcal{F}_{k,p}$. For each $p \geq 0$, the family $\mathcal{G}_p$ is a finite dimensional vector space of functions on $\mathcal{R}^d$ having dimension

$$\sum_{k=0}^{p} \binom{d+k-1}{d-1} \leq (p+1)\binom{d+p-1}{d-1} \leq (d+p)^{d+1}.$$

Thus $\mathcal{G}_p$ is a VC-graph class, and the same is true of $\tilde{\mathcal{G}}_p$. Standard results concerning VC-graph classes (c.f. Chapter 2 of Pollard (1984)) show that

$$N(x_1^n, r, \tilde{\mathcal{G}}_p) \leq a_p r^{-b_p},$$

where $b_p = 2(d+p)^{d+1} + 4$ and $a_p = e^{2b_p \log b_p}$ are independent of n and $r > 0$. Proposition 1 of Nobel (1996) shows further that

$$N(x_1^n, r, \mathcal{F}_{k,p}) \leq \left(a_p n^d r^{-b_p}\right)^k \tag{17}$$

for each sequence $x_1, \ldots, x_n$ and each $r > 0$.

Assume without loss of generality that X is supported on $S = [0,1]^d$. Let $k = 2^{ld}$ where $l \geq 1$ is an integer, and consider the regular dyadic partition π of $[0,1]^d$ into k cells, each of which is a cube with sides of length 2^{-l}. One can implement π by means of a pair (T, τ), where T is a balanced binary tree of depth ld.

Fix a cube $U_i \in \pi$ and let z_i be its center, that is, the j'th coordinate of z_i is the midpoint of the j'th interval in the Cartesian product that defines U_i. Let $M < \infty$ bound each partial derivative of f^* on some open set containing S. A multivariate Taylor series expansion of f^* about z_i shows that

$$f^*(z_i + x) = \sum_{|\alpha| \leq s-1} a_\alpha x^\alpha + R(x)$$

where

$$|R(x)| \leq M \sum_{|\alpha| = s} |x^\alpha|.$$

If $z_i + x \in U_i$ then

$$|R(x)| \leq c k^{-s/d}$$

with $c = M2^{-s}(s+d)^d$, and consequently for each $i = 1, \ldots, k$ there is a polynomial $g_i \in \mathcal{G}_{s-1}$ such that

$$\max_{x \in A_i} |f^*(x) - g_i(x)| \leq c k^{-s/d}.$$

As $|f^*| \leq 1/2$, truncating each g_i at $+/-1/2$ leaves the bound unchanged. Piecing together these truncated polynomials produces a function $f \in \mathcal{F}_{k,s-1}$ such that

$$\int |f - f^*|^2 dP \leq c^2 k^{-2s/d}. \tag{18}$$

The upper bound of Theorem 2 is an infimum over indices k, $p \geq 0$. Fixing $p = s$ and applying the bounds (17) and (18) above, one finds that

$$\mathbf{E}L(g_n) - L(f^*) \leq C(s,d) \inf_{k \geq 0} \left[\frac{k \log n}{n} + k^{-2s/d}\right].$$

Optimizing over k gives the desired bound. □

6 Proofs

Our first lemma is a straightforward modification of some arguments in Lugosi and Zeger (1995). Recall that $L(f) = \mathbf{E}\ell(f(X), Y)$, $L^* = \inf_f L(f)$, and $\widehat{L}_n(f) = n^{-1}\sum_{i=1}^n \ell(f(X_i), Y_i)$. Covering numbers and empirical covers are defined in Section 1.1.

Lemma 1 *Let $\mathcal{F}_1, \mathcal{F}_2, \ldots$ be a sequence of finite sets of functions $f : \mathcal{R}^d \to \mathcal{R}$. Let $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{R}^d \times \mathcal{R}$ be independent replicates of a pair (X, Y) such that $\ell(f(X_i), Y_i) \leq 1$ with probability one for all $f \in \cup_{k=1}^\infty \mathcal{F}_k$. Let*

$$f'_k = \arg\min_{f \in \mathcal{F}_k} L(f) \quad \text{and} \quad \widehat{f}_k = \arg\min_{f \in \mathcal{F}_k} \widehat{L}_n(f),$$

be rules in the k'th class having minimal actual and empirical risk, respectively. Let $L'_k = L(f'_k)$. Define non-negative complexities $C_n(1), C_n(2), \ldots$ by

$$C_n(k) = \sqrt{\frac{\log|\mathcal{F}_k| + 2\log k}{2n}},$$

and consider the complexity penalized empirical risks

$$\tilde{L}_n(\widehat{f}_k) = \widehat{L}_n(\widehat{f}_k) + C_n(k) \qquad k = 1, 2, \ldots.$$

If $g_n = \arg\min_{\widehat{f}_k : k \geq 1} \tilde{L}_n(\widehat{f}_k)$ is that function $\widehat{f}_k$ minimizing $\tilde{L}_n$, then

$$\mathbf{E}L(g_n) - L^* \leq \inf_{k \geq 1} \left[3.66 \cdot C_n(k) + (L'_k - L^*)\right].$$

Proof: We begin with the decomposition

$$L(g_n) - L'_k = \left(L(g_n) - \inf_{j\geq 1} \tilde{L}_n(\widehat{f}_j)\right) + \left(\inf_{j\geq 1} \tilde{L}_n(\widehat{f}_j) - L'_k\right),$$

which holds for any $k \geq 1$. Let $\epsilon > 0$ be arbitrary. Then

$$\begin{aligned}
\mathbf{P}\left\{L(g_n) - \inf_{j\geq 1} \tilde{L}_n(\widehat{f}_j) > \epsilon\right\} &\leq \mathbf{P}\left\{\sup_{j\geq 1}\left(L(\widehat{f}_j) - \tilde{L}_n(\widehat{f}_j)\right) > \epsilon\right\} \\
&\leq \sum_{j=1}^{\infty} \mathbf{P}\left\{L(\widehat{f}_j) - \widehat{L}_n(\widehat{f}_j) > \epsilon + C_n(j)\right\} \\
&\leq \sum_{j=1}^{\infty} \mathbf{P}\left\{\max_{f\in\mathcal{F}_j}\left(L(f) - \widehat{L}_n(f)\right) > \epsilon + C_n(j)\right\} \\
&\leq \sum_{j=1}^{\infty} |\mathcal{F}_j| e^{-2n(\epsilon + C_n(j))^2} \qquad (19) \\
&\leq \sum_{j=1}^{\infty} |\mathcal{F}_j| e^{-2n\epsilon^2} e^{-2nC_n(j)^2} \\
&= e^{-2n\epsilon^2} \sum_{j=1}^{\infty} \frac{1}{j^2} \leq 2e^{-2n\epsilon^2},
\end{aligned}$$

where (19) follows from the union bound and Hoeffding's inequality. Standard bounding then shows that

$$\mathbf{E}\left\{L(g_n) - \inf_{j\geq 1} \tilde{L}_n(\widehat{f}_j)\right\} \leq \frac{1}{\sqrt{n}}.$$

On the other hand, if $\epsilon \geq 2C_n(k)$ then

$$\begin{aligned}
\mathbf{P}\left\{\inf_{j\geq 1} \tilde{L}_n(\widehat{f}_j) - L'_k > \epsilon\right\} &\leq \mathbf{P}\left\{\tilde{L}_n(\widehat{f}_k) - L'_k > \epsilon\right\} \\
&\leq \mathbf{P}\left\{\widehat{L}_n(\widehat{f}_k) - L'_k > \frac{\epsilon}{2}\right\} \quad (\text{using } \epsilon \geq 2C_n(k)) \\
&\leq \mathbf{P}\left\{\left(\widehat{L}_n(f'_k) - L(f'_k)\right) > \frac{\epsilon}{2}\right\} \\
&\leq e^{-n\epsilon^2/2},
\end{aligned}$$

where at the last step Hoeffding's inequality is used. Consequently,

$$\begin{aligned}
\mathbf{E}\left\{\left(\inf_{j\geq 1} \tilde{L}_n(\widehat{f}_j) - L'_k\right)^2\right\} &= \int_0^1 \mathbf{P}\left\{\inf_{j\geq 1} \tilde{L}_n(\widehat{f}_j) - L'_k > \sqrt{\epsilon}\right\} d\epsilon \\
&\leq 4C_n(k)^2 + \int_{4C_n(k)^2}^{\infty} e^{-n\epsilon/2} d\epsilon \\
&\leq 4C_n(k)^2 + \frac{2}{nk^2|\mathcal{F}_k|} \leq 5C_n(k)^2.
\end{aligned}$$

Therefore,

$$\mathbf{E}\left\{\inf_{j\geq 1} \tilde{L}_n(\widehat{f}_j) - L^*_k\right\} \leq \sqrt{\mathbf{E}\left\{\left(\inf_{j\geq 1} \tilde{L}_n(\widehat{f}_j) - L^*_k\right)^2\right\}} \leq \sqrt{5}C_n(k).$$

Collecting bounds, we have

$$\mathbf{E}L(g_n) - L'_k \leq \sqrt{5}C_n(k) + \frac{1}{\sqrt{n}} \leq C_n(k)(\sqrt{5} + \sqrt{2}) < 3.66 \cdot C_n(k).$$

Hence

$$\begin{aligned}
\mathbf{E}L(g_n) - L^* &= \inf_{k\geq 1} \left[\mathbf{E}L(g_n) - L'_k + (L'_k - L^*)\right] \\
&\leq \inf_{k\geq 1} \left[3.66 \cdot C_n(k) + (L'_k - L^*)\right].
\end{aligned}$$

□

Let $Z_1, \ldots, Z_m$ be i.i.d. replicates of a random vector $Z \in \mathcal{R}^{d+1}$ and let $\mathcal{H}$ be a family of non-negative functions $h : \mathcal{R}^{d+1} \to [0, \infty)$ such that $h(Z) \leq 1$ with probability one. For each function $h \in \mathcal{H}$, define

$$Ph = \mathbf{E}h(Z) \quad \text{and} \quad \widehat{P}_m h = \frac{1}{m}\sum_{i=1}^{m} h(Z_i).$$

Lemma 2 *If $\bar{r}_m$ is the balanced covering radius (9) of $\mathcal{H}$ then*

$$\mathbf{E}\left[\sup_{h\in\mathcal{H}} |\widehat{P}_m h - Ph|\right] \leq 4\bar{r}_m.$$

Proof: Fix a number $r > \bar{r}_m$. Then by definition of $\bar{r}_m$,

$$r \geq \sqrt{\frac{8 \log \mathbf{E}N\left(Z_1^m, r/2, \mathcal{H}\right)}{m}} \vee \sqrt{\frac{8}{m}} \qquad (20)$$

By standard symmetrization arguments (c.f. Pollard (1989)),

$$\mathbf{E}\left[\sup_{h\in\mathcal{H}} |\widehat{P}_m h - Ph|\right] \leq 2\mathbf{E}\left[\sup_{h\in\mathcal{H}} \left|\frac{1}{m}\sum_{i=1}^{m} \sigma_i h(Z_i)\right|\right],$$

where $\sigma_1, \ldots, \sigma_m$ are independent sign random variables, independent of the Z_i's, such that $\mathbf{P}\{\sigma_i = 1\} = \mathbf{P}\{\sigma_i = -1\} = 1/2$. According to Pollard (1984, Ch2), for every $t > 0$,

$$\mathbf{P}\left\{\sup_{h\in\mathcal{H}} \left|\frac{1}{m}\sum_{i=1}^{m} \sigma_i h(Z_i)\right| > t\right\} \leq 2\mathbf{E}N\left(Z_1^m, \frac{r}{2}, \mathcal{H}\right) e^{-mt^2/8}.$$

Therefore,

$$\begin{aligned}
\mathbf{E}\left[\sup_{h\in\mathcal{H}} |\widehat{P}_m h - Ph|\right] &\leq 2r + 4\int_r^1 \mathbf{E}N\left(Z_1^m, \frac{r}{2}, \mathcal{H}\right) e^{-mt^2/8} dt \\
&\leq 2r + 4\mathbf{E}N\left(Z_1^m, \frac{r}{2}, \mathcal{H}\right) \int_r^1 e^{-mt^2/8} dt \\
&\leq 2r + 4\sqrt{2}\,\mathbf{E}N\left(Z_1^m, \frac{r}{2}, \mathcal{H}\right) \int_{r/\sqrt{8}}^{\infty} e^{-ms^2} \left(2 + \frac{1}{ms^2}\right) ds
\end{aligned}$$

$$
\begin{aligned}
&= 2r + 4\sqrt{2}\,\mathbf{E}N\left(Z_1^m, \frac{r}{2}, \mathcal{H}\right)\left[\frac{-1}{ms}e^{-ms^2}\right]_{r/\sqrt{8}}^{\infty} \\
&= 2r + \frac{16}{mr}\mathbf{E}N\left(Z_1^m, \frac{r}{2}, \mathcal{H}\right)e^{-mr^2/8} \\
&\le 2r + \frac{16}{mr} \\
&\le 4r.
\end{aligned}
$$

The last two inequalities above follow from (20). Taking the infimum over all $r > \bar{r}_m$ establishes the assertion of the Lemma. □

Lemma 3 *If $\bar{r}_m$ and $\widehat{r}_m$ are the balanced covering radius (9) and the balanced empirical covering radius (6) of $\mathcal{H}$ respectively, then*

$$\mathbf{E}\widehat{r}_m \le 2\bar{r}_m.$$

Proof: Fix a radius $r > \bar{r}_m$ and note that the expected value of $\widehat{r}_m$ may be bounded as follows:

$$\mathbf{E}\widehat{r}_m \le r + \int_0^\infty \mathbf{P}\{\widehat{r}_m > r + t\}dt. \tag{21}$$

If $\widehat{r}_m > r + t$, then by definition of $\widehat{r}_m$ and monotonicity of the covering numbers,

$$r + t < \sqrt{\frac{8\log N\left(Z_1^m, (r+t)/2, \mathcal{H}\right)}{m}} \vee \sqrt{\frac{8}{m}} \le \sqrt{\frac{8\log N\left(Z_1^m, r/2, \mathcal{H}\right)}{m}} \vee \sqrt{\frac{8}{m}}$$

Combining this last inequality with (20) gives the bound

$$
\begin{aligned}
\mathbf{P}\{\widehat{r}_m > r + t\} &\le \mathbf{P}\left\{\sqrt{\frac{8\log N\left(Z_1^m, r/2, \mathcal{H}\right)}{m}} \vee \sqrt{\frac{8}{m}}\right. \\
&\qquad \left. > \left(\sqrt{\frac{8\log \mathbf{E}N\left(Z_1^m, r/2, \mathcal{H}\right)}{m}} \vee \sqrt{\frac{8}{m}}\right) + t\right\} \\
&\le \mathbf{P}\left\{\sqrt{\frac{8\log N\left(Z_1^m, r/2, \mathcal{H}\right)}{m}} > \sqrt{\frac{8\log \mathbf{E}N\left(Z_1^m, r/2, \mathcal{H}\right)}{m}} + t\right\}.
\end{aligned}
$$

Let $\psi(x) = e^{mx^2/8}$. As ψ is monotone increasing, Markov's inequality implies that the last probability above is at most

$$
\begin{aligned}
&\mathbf{E}\psi\left(\sqrt{\frac{8\log N\left(Z_1^m, r/2, \mathcal{H}\right)}{m}}\right) \cdot \left[\psi\left(\sqrt{\frac{8\log \mathbf{E}N\left(Z_1^m, r/2, \mathcal{H}\right)}{m}} + t\right)\right]^{-1} \\
&= \mathbf{E}N\left(Z_1^m, \frac{r}{2}, \mathcal{H}\right)\exp\left\{\frac{-m}{8}\left(\sqrt{\frac{8\log \mathbf{E}N\left(Z_1^m, r/2, \mathcal{H}\right)}{m}} + t\right)^2\right\} \\
&\le e^{-mt^2/8}.
\end{aligned}
$$

Now $\int_0^\infty e^{-mt^2/8}dt = \sqrt{2\pi/m} \le \bar{r}_m$, and therefore (21) shows that $\mathbf{E}\widehat{r}_m \le r + \bar{r}_m$. Taking the infimum over all $r > \bar{r}_m$ completes the proof. □

Proof of Theorem 1: (See Section 3 above for the definition of quantities appearing in the proof.) If $Z_1, \ldots, Z_m$ are held fixed, then the empirical covers $\widehat{\mathcal{F}}_1$, $\widehat{\mathcal{F}}_2$, ... may be treated as fixed, finite model classes with cardinalities given by $|\widehat{\mathcal{F}}_k| = N(Z_1^m, \widehat{r}_m^{(k)}, \mathcal{H}_k)$. Conditional application of Lemma 1 gives the following bound:

$$
\begin{aligned}
\mathbf{E}L(g_n) - L^* &= \mathbf{E}\,\mathbf{E}\left\{L(g_n) - L^* | Z_1^m\right\} \\
&\le \mathbf{E}\left\{\inf_{k\ge 1}\left(3.66 \cdot \widehat{C}_{n-m}(k) + (L_k' - L^*)\right)\right\} \\
&= \mathbf{E}\left\{\inf_{k\ge 1}\left(3.66 \cdot \widehat{C}_{n-m}(k) + (L_k' - L_k^*) + (L_k^* - L^*)\right)\right\}.
\end{aligned}
$$

Now observe that

$$
\begin{aligned}
L_k' - L_k^* &= \min_{f\in\widehat{\mathcal{F}}_k} L(f) - \inf_{f\in\mathcal{F}_k} L(f) \\
&\le \min_{f\in\widehat{\mathcal{F}}_k} \widehat{L}_m(f) - \inf_{f\in\mathcal{F}_k} \widehat{L}_m(f) + 2\sup_{f\in\mathcal{F}_k}|\widehat{L}_m(f) - L(f)| \\
&\le \widehat{r}_m^{(k)} + 2\sup_{h\in\mathcal{H}_k}|\widehat{P}_m h - Ph|.
\end{aligned}
$$

Therefore, by an applications of Lemmas 2 and 3,

$$
\begin{aligned}
&\mathbf{E}L(g_n) - L^* \\
&\le \mathbf{E}\left\{\inf_{k\ge 1}\left[3.66 \cdot \widehat{C}_{n-m}(k) + \widehat{r}_m^{(k)} + 2\sup_{h\in\mathcal{H}_k}|\widehat{P}_m h - Ph| + (L_k^* - L^*)\right]\right\} \\
&\le \inf_{k\ge 1}\left[\mathbf{E}\left\{3.66 \cdot \widehat{C}_{n-m}(k)\right\} + \mathbf{E}\widehat{r}_m^{(k)} + 2\sup_{h\in\mathcal{H}_k}|\widehat{P}_m h - Ph| + (L_k^* - L^*)\right] \\
&\le \inf_{k\ge 1}\left[\mathbf{E}\left\{3.66 \cdot \widehat{C}_{n-m}(k)\right\} + 10\bar{r}_m^{(k)} + (L_k^* - L^*)\right]
\end{aligned}
$$

It remains to consider the expectation of the empirical complexities. As $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and $m = n - m = n/2$,

$$
\begin{aligned}
\widehat{C}_{n-m}(k) &\le \sqrt{\frac{\log N(Z_1^m, \widehat{r}_m^{(k)}, \mathcal{H}_k)}{2m}} + \sqrt{\frac{2\log k}{n}} \\
&= \frac{1}{4}\sqrt{\frac{8\log N(Z_1^m, 2\widehat{r}_m^{(k)}/2, \mathcal{H}_k)}{m}} + \sqrt{\frac{2\log k}{n}} \\
&\le \frac{2}{4}\widehat{r}_m^{(k)} + \sqrt{\frac{2\log k}{n}},
\end{aligned}
$$

where the last inequality follows from the definition of $\hat{r}_m^{(k)}$ and the fact that $2\hat{r}_m^{(k)} > \bar{r}_m^{(k)}$. By another application of Lemma 3,

$$\mathbf{E}\left\{3.66 \cdot \widehat{C}_{n-m}(k)\right\} \leq 3.66 \cdot \bar{r}_m^{(k)} + 5.2\sqrt{\frac{\log k}{n}}$$

and the proof is complete. □

To prove Theorem 2, we need the following technical lemma:

Lemma 4 *Let $\mathcal{H}$ be a finite class of functions $h : \mathcal{X} \to \mathcal{R}$ and let $Z \in \mathcal{X}$ be a random variable such that $0 \leq h(Z) \leq 1$ with probability one for all $h \in \mathcal{H}$. If $Z_1, \ldots, Z_n$ are i.i.d. copies of Z, and η, γ, ϵ are positive numbers, then*

$$\mathbf{P}\left\{\min_{h\in\mathcal{H}:Ph>\eta+(1+\gamma)\epsilon} \widehat{P}_n h \leq \eta + \gamma\epsilon\right\} \leq |\mathcal{H}| \exp\left[-\frac{3n}{8} \cdot \frac{\epsilon^2}{\eta + (1+\gamma)\epsilon}\right].$$

Proof: If

$$\max_{h\in\mathcal{H}} \frac{Ph - \widehat{P}_n h}{\sqrt{Ph}} \leq \frac{\epsilon}{\sqrt{\eta + (1+\gamma)\epsilon}},$$

then for every $h \in \mathcal{H}$,

$$\widehat{P}_n h \geq Ph - \epsilon\sqrt{\frac{Ph}{\eta + (1+\gamma)\epsilon}}.$$

As the function $x - c\sqrt{x}$ is monotone increasing for $x \geq c^2/4$, if in addition $Ph > \eta + (1+\gamma)\epsilon$, then

$$\widehat{P}_n h \geq \eta + (1+\gamma)\epsilon - \epsilon\sqrt{\frac{\eta + (1+\gamma)\epsilon}{\eta + (1+\gamma)\epsilon}} = \eta + \gamma\epsilon.$$

Hence

$$\begin{aligned}\mathbf{P}\left\{\min_{h\in\mathcal{H}:Ph>\eta+(1+\gamma)\epsilon} \widehat{P}_n h \leq \eta + \gamma\epsilon\right\} &\leq \mathbf{P}\left\{\max_{h\in\mathcal{H}} \frac{Ph - \widehat{P}_n h}{\sqrt{Ph}} \geq \frac{\epsilon}{\sqrt{\eta + (1+\gamma)\epsilon}}\right\} \\ &\leq |\mathcal{H}| \max_{h\in\mathcal{H}} \mathbf{P}\left\{\frac{Ph - \widehat{P}_n h}{\sqrt{Ph}} \geq \frac{\epsilon}{\sqrt{\eta + (1+\gamma)\epsilon}}\right\}.\end{aligned}$$

It therefore suffices to show that for every $h \in \mathcal{H}$,

$$\mathbf{P}\left\{\frac{Ph - \widehat{P}_n h}{\sqrt{Ph}} \geq \theta\right\} \leq \exp\left[-\frac{3n}{8} \cdot \frac{\epsilon^2}{\eta + (1+\gamma)\epsilon}\right],$$

where $\theta = \epsilon(\eta + (1+\gamma)\epsilon)^{-1/2}$. Note that the probability on the left-hand side is zero whenever $\theta \geq \sqrt{Ph}$, so we may assume without loss of generality that $\theta < \sqrt{Ph}$. Then

$$\mathbf{P}\left\{Ph - \widehat{P}_n h \geq \theta\sqrt{Ph}\right\} \leq \exp\left[\frac{-n\theta^2 Ph}{2Ph + (2/3)\theta\sqrt{Ph}}\right] \leq \exp\left[\frac{-3n\theta^2}{8}\right].$$

The first inequality above follows from Bernstein's inequality (see, e.g., Pollard, 1984, p.193) and the fact that $\mathbf{Var}h(Z) \leq Ph$. The second follows from the assumption that $\theta < \sqrt{Ph}$. □

Lemma 5 *Consider the same situation as in Lemma 1 but now with complexities*

$$C_n(k) = 22 \cdot \frac{\log|\mathcal{F}_k| + 2\log k}{n}.$$

Let ψ_n be the candidate rule $\widehat{f}_j$ minimizing the sum of class complexity and empirical risk. Then

$$\mathbf{E}L(\psi_n) \leq \inf_{k\geq 1}(2C_n(k) + 5L'_k) + \frac{106}{n}.$$

Proof: Let $\widehat{f}_k$ and f'_k be defined as in Lemma 1. In order to establish the stated inequality, we first derive a probabilistic bound for the difference between $L(\psi_n)$ and L'_k. For any number $\epsilon > 0$,

$$\mathbf{P}\left\{L(\psi_n) - L'_k \geq \epsilon\right\} = \sum_{j=1}^{\infty} \mathbf{P}\left\{L(\widehat{f}_j) - L'_k \geq \epsilon \text{ and } \psi_n = \widehat{f}_j\right\}$$

Set $\epsilon = 4L'_k + 2\delta$ and consider a single term in the sum. If $\psi_n = \widehat{f}_j$ and $L(\widehat{f}_j) - L'_k \geq \epsilon$, then there is a function $f \in \mathcal{F}_j$ such that

$$\tilde{L}_n(f) \leq \tilde{L}_n(\widehat{f}_k) \leq \tilde{L}_n(f'_k) \quad \text{and} \quad L(f) \geq 5L'_k + 2\delta\,.$$

Therefore,

$$\begin{aligned}&\mathbf{P}\left\{L(\widehat{f}_j) - L'_k > 4L'_k + 2\delta \text{ and } \psi_n = \widehat{f}_j\right\} \\ &\leq \mathbf{P}\left\{\inf_{f\in\mathcal{F}_j:L(f)\geq 5L'_k+2\delta} \tilde{L}_n(f) \leq \tilde{L}_n(f'_k)\right\} \\ &= \mathbf{P}\left\{\inf_{f\in\mathcal{F}_j:L(f)\geq 5L'_k+2\delta} \widehat{L}_n(f) \leq \widehat{L}_n(f'_k) + (C_n(k) - C_n(j))\right\}.\end{aligned}$$

Let A be the event in the last line above. Define additional events

$$B = \left\{\widehat{L}_n(f'_k) + C_n(k) - C_n(j) \geq 0\right\},$$

and

$$C = \left\{2\widehat{L}_n(f'_k) < 3L'_k + C_n(j) - C_n(k) + \delta\right\}.$$

Clearly $\mathbf{P}(A \cap B^c) = 0$, and consequently

$$\mathbf{P}(A) \leq \mathbf{P}(A \cap B \cap C) + \mathbf{P}(C^c)\,. \tag{22}$$

A straightforward calculation shows that $\mathbf{Var}(\widehat{L}_n(f'_k)) \le L'_k/n$. It then follows from Bernstein's inequality that

$$\begin{aligned}\mathbf{P}(C^c) &= \mathbf{P}\left\{2(\widehat{L}_n(f'_k) - L'_k) \ge L'_k + \delta + C_n(j) - C_n(k)\right\} \\ &\le \exp\left(\frac{-n(\delta + C_n(j) - C_n(k))}{28/3}\right) \\ &= \frac{k^2|\mathcal{F}_k|}{j^2|\mathcal{F}_j|}\exp\left(\frac{-n\delta}{28/3}\right). \qquad (23)\end{aligned}$$

If $B \cap C$ occurs then

$$2(C_n(j) - C_n(k)) \le 2\widehat{L}_n(f'_k) \ \le\ 3L(f'_k) + (C_n(j) - C_n(k)) + \delta$$

which implies

$$C_n(j) - C_n(k) \ \le\ 3L'_k + \delta\,.$$

Thus $B \cap C$ implies that

$$5L'_k + 2\delta \ge 2L'_k + (\delta + C_n(j) - C_n(k))\,.$$

It follows from these considerations that

$$\begin{aligned}&\mathbf{P}(A \cap B \cap C) \\ &\le \mathbf{P}\left\{\inf_{f\in\mathcal{F}_j:\,L(f)\ge 2L'_k+(\delta+C_n(j)-C_n(k))}\widehat{L}_n(f) \le \frac{3}{2}L'_k + \frac{1}{2}\left(\delta + (C_n(j) - C_n(k))_+\right)\right\} \\ &\le |\mathcal{F}_j|\cdot\exp\left(\frac{-n(\delta + C_n(j) - C_n(k))}{22}\right) \\ &\le \frac{k^2}{j^2}|\mathcal{F}_k|\cdot\exp\left(\frac{-n\delta}{22}\right)\end{aligned}$$

where the second inequality follows from Lemma 4 with $\eta = L'_k$, $\gamma = 1$, and $\epsilon = (L'_k + \delta + (C_n(j) - C_n(k))_+)/2$. Combining the inequality above with (22) and (23) shows that

$$\mathbf{P}\{L(\widehat{f}_j) - L(f'_k) > 4L'_k + 2\delta\} \ \le\ 2\,\frac{k^2}{j^2}|\mathcal{F}_k|\cdot\exp\left(\frac{-n\delta}{22}\right)$$

and therefore, by the union bound and by replacing δ with $\delta/2$,

$$\begin{aligned}\mathbf{P}\left\{L(\psi_n) - L'_k \ge 4L'_k + \delta\right\} &\le 2k^2|\mathcal{F}_k|\exp\left(\frac{-n\delta}{44}\right)\cdot\sum_{j=1}^{\infty}\frac{1}{j^2} \\ &\le 4k^2\cdot|\mathcal{F}_k|\exp\left(\frac{-n\delta}{44}\right). \qquad (24)\end{aligned}$$

Using the last inequality, the expected difference between $L(\psi_n)$ and L'_k may be bounded as follows:

$$\begin{aligned}\mathbf{E}\left[L(\psi_n) - L'_k\right] &\le \mathbf{E}\left[L(\psi_n) - L'_k\right]_+ \\ &\le 4L'_k + u + \int_u^\infty \mathbf{P}\left\{L(\psi_n) - L'_k \ge 4L'_k + t\right\}dt \\ &\le 4L'_k + u + 4k^2|\mathcal{F}_k|\int_u^\infty \exp\left(\frac{-nt}{44}\right)dt \\ &= 4L'_k + u + \frac{176k^2|\mathcal{F}_k|}{n}\int_{nu/44}^\infty e^{-v}dv \\ &\le 4L'_k + \frac{44\log\left(4ek^2|\mathcal{F}_k|\right)}{n},\end{aligned}$$

where in the last step u is set equal to $44n^{-1}\log(4k|\mathcal{F}_k|)$. It follows that for every $k \ge 1$,

$$\mathbf{E}L(\psi_n) \le 5L'_k + \frac{44\log\left(4ek^2|\mathcal{F}_k|\right)}{n} \le 5L'_k + 2C_n(k) + \frac{106}{n},$$

as desired. □

The following inequality is due to Pollard (1986), see also Haussler (1992) for the proof.

Lemma A *Let $\mathcal{H}$ be a family of functions $h : \mathcal{R}^{d+1} \to [0,1]$, and let $Z_1, \ldots, Z_m \in \mathcal{R}^{d+1}$ be i.i.d. random vectors. For each $u > 0$ and each $\alpha \in (0,1)$,*

$$\mathbf{P}\left\{\sup_{h\in\mathcal{H}}\frac{|P_m h - Ph|}{P_m h + Ph + u} > \alpha\right\} \le 4\mathbf{E}N\left(Z_1^m, \frac{\alpha u}{8}, \mathcal{G}\right)e^{-m\alpha^2 u/16}\,.$$

Lemma 6 *Let $\mathcal{F}_k$ be a model class, $f^*_k = \arg\min_{f\in\mathcal{F}_k} L(f)$ and $L^*_k = L(f^*_k)$. For each $r > 0$,*

$$\mathbf{E}\left\{L'_k - 2L^*_k\right\} \le 2r + \frac{392\log\mathbf{E}N(Z_1^m, r/14, \mathcal{H}_k)}{m} + \frac{1023}{m}.$$

Proof: We first derive a probabilistic bound for the difference between L'_k and $2L^*_k$. If $L'_k - 2L^*_k > 2r + t$ for some $t > 0$, then there exists a prediction rule $f \in \widehat{\mathcal{F}}_k \subseteq \mathcal{F}_k$ such that

$$\frac{1}{m}\sum_{i=1}^m |\, l(f(X_i), Y_i) - l(f^*_k(X_i), Y_i)\,| < r \quad\text{and}\quad L(f) \ \ge\ L'_k \ \ge\ 2L^*_k + 2r + t.$$

The first inequality implies $\widehat{L}_m(f) < \widehat{L}_m(f^*_k) + r$, and it follows that

$$\begin{aligned}&\mathbf{P}\{L'_k - 2L^*_k > t + 2r\} \\ &\le \mathbf{P}\left\{\inf_{f\in\mathcal{F}_k:L(f)>2L^*_k+2r+t}\widehat{L}_m(f) < \widehat{L}_m(f^*_k) + r\right\}\end{aligned}$$

$$
\begin{aligned}
&\le \mathbf{P}\left\{\inf_{f\in\mathcal{F}_k: L(f)>2L_k^*+2r+t} \widehat{L}_m(f) < \frac{3}{2}L_k^* + r + \frac{t}{2}\right\} \\
&\quad + \mathbf{P}\left\{\widehat{L}_m(f_k^*) + r \ge \frac{3}{2}L(f_k^*) + r + \frac{t}{2}\right\} \\
&\le \mathbf{P}\left\{\inf_{f\in\mathcal{F}_k: L(f)>2L_k^*+2r+t} \widehat{L}_m(f) < \frac{3}{2}L_k^* + r + \frac{t}{2}\right\} \\
&\quad + \mathbf{P}\left\{\widehat{L}_m(f_k^*) \ge \frac{3}{2}L(f_k^*) + \frac{t}{2}\right\}. \qquad (25)
\end{aligned}
$$

Bernstein's inequality implies that the second probability in (25) is at most $e^{-mt/10}$. To bound the first, let $f \in \mathcal{F}_k$ be any prediction rule such that $L(f) \ge 2L_k^* + 2r + t$. If in addition

$$
\frac{L(f) - \widehat{L}_m(f)}{L(f) + \widehat{L}_m(f) + 2(2r+t)} \le \frac{1}{7}
$$

then by a straighforward calculation,

$$
\widehat{L}_m(f) \ \ge\ (2L_k^* + 2r + t)\frac{3}{4} - \frac{2r+t}{4}. \ = \ \frac{3}{2}L(f_k^*) + r + \frac{t}{2}.
$$

It follows from Lemma A that the first inequality in (25) is at most

$$
\mathbf{P}\left\{\sup_{f\in\mathcal{F}_k} \frac{L(f) - \widehat{L}_m(f)}{L(f) + \widehat{L}_m(f) + 2(2r+t)} > \frac{1}{7}\right\} \ \le\ 4\mathbf{E}N(Z_1^m, r/14, \mathcal{H}_k)e^{-mt/392}.
$$

Summarizing, for each $t > 0$,

$$
\mathbf{P}\{L_k' - 2L_k^* > t + 2r\} \le 5\mathbf{E}N(Z_1^m, r/14, \mathcal{H}_k)e^{-mt/392}.
$$

Thus, for every $u > 0$,

$$
\begin{aligned}
\mathbf{E}L_k' - 2L_k^* &\le 2r + \int_0^\infty \mathbf{P}\{L_k' - L_k^* > t + 2r\}dt \\
&\le 2r + u + \int_u^\infty 5\mathbf{E}N\left(Z_1^m, \frac{r}{14}, \mathcal{H}_k\right)\exp\left[\frac{-mt}{392}\right]dt
\end{aligned}
$$

The desired inequality follows by setting $u = 392\log\left(5\mathbf{E}N(Z_1^m, r/14, \mathcal{H}_k)\right)/m$. □

Proof of Theorem 2: By conditioning on Z_1^m and applying Lemma 5 one obtains the bound

$$
\mathbf{E}L(\psi_n) \le \mathbf{E}\left\{\inf_k\left(2C_{n-m}(k) + 5L_k'\right)\right\} + \frac{212}{n}.
$$

By Lemma 6 and the definition of the complexities $\widehat{C}_n(k)$ the first term on the right hand side is at most

$$
\begin{aligned}
&\inf_k\left(\frac{88\mathbf{E}\log N(Z_1^m, r_k, \mathcal{H}_k)}{n} + \frac{176\log k}{n} + 5\mathbf{E}\left\{L_k' - 2L_k^*\right\} + 10L_k^*\right) \\
&\le \inf_k\left(\frac{88\mathbf{E}\log N(Z_1^m, r_k, \mathcal{H}_k)}{n} + \frac{176\log k}{n} + 10r_k + \frac{3920\log\left(\mathbf{E}N(Z_1^m, r_k/14, \mathcal{H}_k)\right)}{n}\right. \\
&\qquad\left. + \frac{10230}{n} + 10L_k^*\right) \\
&\le \inf_k\left(\frac{4008\log\left(\mathbf{E}N(Z_1^m, r_k/14, \mathcal{H}_k)\right)}{n} + \frac{176\log k}{n} + 10r_k + 10L_k^*\right) + \frac{10230}{n},
\end{aligned}
$$

and the result follows. □

Acknowledgements

The authors wish to thank Peter Bartlett for several illuminating conversations on the subject of the paper, and Amir Dembo whose comments led to a shorter proof of Proposition 3. We also thank the referees for their many helpful comments and suggestions, which improved the presentation of the paper.

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. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

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

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

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

[8] S. N. Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow, 1946.

[9] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth International, Belmont, CA., 1984.

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

[11] 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.

[12] P. Chaudhuri, M.-C. Huang, W.-Y. Loh, and R. Yao Piecewise polynomial regression trees. *Statistica Sinica*, 4:143-167, 1994.

[13] L. Devroye. Automatic pattern recognition: A study of the probability of error. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:530–543, 1988.

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

[15] R. M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6:899–929, 1978.

[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] L. Gordon and R. Olshen. Almost sure consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 15:147-163, 1984.

[19] U. Grenander. *Abstract Inference*. John Wiley, New York, 1981.

[20] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Information and Computation*, vol. 100, pp. 78-150, 1992.

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

[22] 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.

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

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

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

[26] C.L. Mallows. Some comments on C_p. *IEEE Technometrics*, 15:661–675, 1973.

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

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

[29] T. Nicoleris and Y. Yatracos. Rates of convergence of estimates, Kolmogorov's entropy and the dimensionality reduction principle in regression. *Annals of Statistics*, 25:2493–2511, 1997.

[30] A. S. Nemirovskii, B. T. Polyak, and A. B. Tsybakov. Rate of convergence of nonparametric estimates of maximum-likelihood type. *Problems of Information Transmission*, 21:258–272, 1985.

[31] A.B. Nobel. Histogram regression estimation using data-dependent partitions. *Annals of Statistics*, 24:1084-1105, 1996.

[32] D. Pollard. *Convergence of Stochastic Processes.* Springer-Verlag, New York, 1984.

[33] D. Pollard. Rates of uniform almost sure convergence for empirical processes indexed by unbounded classes of functions, 1986. Manuscript.

[34] D. Pollard. Asymptotics via empirical processes. *Statistical Science*, 4:341–366, 1989.

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

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

[37] 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.

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

[39] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10:1040–1053, 1982.

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

[41] 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.

[42] 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.

[43] C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding *J. Roy. Statist. Soc. B*, 49:240-265, 1987.

[44] H. White. Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, 3:535–549, 1990.

[45] W. H. Wong and X. Shen. Probability inequalities for likelihood ratios and convergence rates of sieve MLE's. Technical Report 346, Department of Statistics, University of Chicago, Chicago, IL, 1992.

[46] Y. Yang and A.R. Barron. Information-theoretic determination of minimax rates of convergence. Submitted to *Annals of Statistics*, 1997.

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