The Annals of Statistics
2003, Vol. 31, No. 1, 252–273

MODEL SELECTION IN NONPARAMETRIC REGRESSION

BY MARTEN WEGKAMP

Yale University

Model selection using a penalized data-splitting device is studied in the context of nonparametric regression. Finite sample bounds under mild conditions are obtained. The resulting estimates are adaptive for large classes of functions.

1. Introduction. We study the additive regression problem $Y = \eta(X) + \varepsilon$, where $\eta : \mathbb{R}^d \to \mathbb{R}$ is the unknown regression function, ε is the random error and $X \in \mathbb{R}^d$ is the random covariate. The available data consists of independent observations (X_j, Y_j). Our aim is to find an estimator $\widehat{\eta}$ in some class $\mathcal{G}$ such that the mean squared error $\mathbb{E}(\widehat{\eta} - \eta)^2(X)$ is as small as possible. In this paper we consider the least squares estimator $\widehat{g} \in \mathcal{G}$, which satisfies

$$\sum_j \big(Y_j - \widehat{g}(X_j)\big)^2 \le \sum_j \big(Y_j - g(X_j)\big)^2 \qquad \text{for all } g \in \mathcal{G}.$$

van de Geer and Wegkamp (1996) formulate necessary and sufficient local metric entropy conditions of $\mathcal{G}$ for consistency of this estimate in the L_2 norm evaluated at the observation points X_i. van de Geer (1990, 2000) shows how the local metric entropy of $\mathcal{G}$ influences the rate of convergence of the least squares estimator. All these papers, however, assume that η belongs to $\mathcal{G}$.

In the following it will be helpful to decompose the mean squared error into two terms:

$$\mathbb{E}(\widehat{g} - \eta)^2(X) = \inf_{g \in \mathcal{G}} \mathbb{E}(g - \eta)^2(X) + \left\{ \mathbb{E}(\widehat{g} - \eta)^2(X) - \inf_{g \in \mathcal{G}} \mathbb{E}(g - \eta)^2(X) \right\}.$$

It is clear that the *approximation error* $\inf_{g \in \mathcal{G}} \mathbb{E}(g - \eta)^2(X)$ is decreasing in $\mathcal{G}$. However, the more complex $\mathcal{G}$, the more difficult the statistical estimation problem becomes as more and more parameters need to be estimated. That is, the *estimation error* $\mathbb{E}(\widehat{g} - \eta)^2(X) - \inf_{g \in \mathcal{G}} \mathbb{E}(g - \eta)^2(X)$ is an increasing function of the complexity of $\mathcal{G}$. This dilemma is solved by considering a sequence of models $\mathcal{G}_1, \mathcal{G}_2, \ldots$ whose union is equal to $\mathcal{G}$. With the data at hand, we seek the optimal model from one of these classes, which balances the two conflicting errors. Hence we are faced with a selection problem.

One possibility, explored by Barron, Birgé and Massart (1999), is to employ penalized least squares by introducing a penalty of the form $C_{\text{pen}} D_k$, where D_k is

Received January 2001; revised May 2002.

AMS 2000 subject classifications. Primary 60F05, 60F17; secondary 60G15, 62E20.

Key words and phrases. Adaptive estimation, classification, data-splitting, least squares estimation, model selection, penalized least squares, VC-major classes.

the dimension of the finite-dimensional model $\mathcal{G}_k$, and n is the sample size. The penalized least squares estimate $\widehat{g}_{\text{pen}}$ minimizes

$$\min_{g\in\mathcal{G}_k}\frac{1}{n}\sum_{i=1}^{n}\bigl(Y_i-g(X_i)\bigr)^2+\frac{C_{\text{pen}}D_k}{n}$$

over $k\in\mathbb{N}$ and it is shown that for some $C>0$

$$\mathbb{E}(\widehat{g}_{\text{pen}}-\eta)^2(X)\le C\inf_k\biggl(\inf_{g\in\mathcal{G}_k}\mathbb{E}(g-\eta)^2(X)+\frac{C_{\text{pen}}D_k}{n}\biggr).$$

However, the constant C_{pen} depends on unknown quantities related to the error distribution and the regression function. This makes this approach intractable in practice. Also their moment condition $\mathbb{E}\exp(|\varepsilon|/b)\le 4$ for some $b>0$ on the error distribution is strong.

Penalties which are distribution free upper bounds of the estimation error can be found in more restrictive regression settings, such as logistic regression as considered by Hengartner and Wegkamp (1999). Unfortunately, distribution free upper bounds may be too conservative estimates for the actual estimation error at hand. In other words, the upper bounds may be loose for a particular problem, and the model that minimizes the sum of the approximation error and penalty term need not correspond to that which minimizes the mean squared error, so these methods may not give the optimal result.

Bartlett, Boucheron and Lugosi (2002) presented a result in the context of bounded regression ($|Y_j|\le 1$) using random penalties of the form

$$\text{pen}(k)=\mathbb{E}\biggl(\sup_{g\in\mathcal{G}_k}\frac{256}{n}\sum_{j=1}^{n}\sigma_j\bigl(Y_j-g(X_j)\bigr)^2\,\Big|\,(X_1,Y_1),\ldots,(X_n,Y_n)\biggr),$$

where σ_j are independent random signs, that is, $\mathbb{P}\{\sigma=-1\}=\mathbb{P}\{\sigma=+1\}=1/2$. Although this conditional expectation is unknown, it can be approximated with Monte Carlo simulations. These authors propose the penalized least squares estimate $\widehat{g}_{\text{pen}}$ which minimizes

$$\min_{g\in\mathcal{G}_k}\frac{1}{n}\sum_{i=1}^{n}\bigl(Y_i-g(X_i)\bigr)^2+\text{pen}(k)+\frac{8192\log k}{n}$$

over $k\in\mathbb{N}$ and they prove that

$$\mathbb{E}(\widehat{g}_{\text{pen}}-\eta)^2(X)\le\min_k\biggl[\inf_{g\in\mathcal{G}_k}\mathbb{E}(g-\eta)^2(X)+\mathbb{E}\,\text{pen}(k)+\frac{8192\log k}{n}\biggr]+\frac{13096}{n}.$$

However, the above penalty $\text{pen}(k)$ is at best of order $\sqrt{D_k/n}$, and not the desired D_k/n for finite dimensional spaces $\mathcal{G}_k$ of dimension D_k.

Barron (1991) does not consider a sequence of approximation spaces $\mathcal{G}_k$, but rather penalizes each individual function $g \in \mathcal{G}$ with a penalty $\lambda(g)/n$. He obtains for the estimator $\widehat{g}_{\text{pen}}$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - g(X_i)\big)^2 + \frac{\lambda(g)}{n}$$

over $g \in \mathcal{G}$ and some $C > 0$ the inequality

$$\mathbb{E}(\widehat{g}_{\text{pen}} - \eta)^2(X) \le C \inf_{g\in\mathcal{G}}\left[\mathbb{E}(g-\eta)^2(X) + \frac{\lambda(g)}{n}\right]$$

under the restrictions $|Y_i| \le 1$ and $\sum_{g\in\mathcal{G}} \exp(-\lambda(g)) < \infty$. As will become apparent later, we generalize this idea to our selection procedure by considering random functions and allowing for unbounded Y_i.

In this paper we take the approach of splitting the data into two parts, using one part for constructing a least squares estimate $\widehat{g}_k$ for each model $\mathcal{G}_k$, and the second part for selecting one of the previously obtained estimates. In a way, one wishes to see how each estimate $\widehat{g}_k$ behaves on a new data set, and this is precisely what we propose to do. We show that this yields the desired trade-off between estimation and approximation error as we obtain an estimator with mean squared error which is essentially bounded by $\inf_k \mathbb{E}(\widehat{g}_k - \eta)^2(X)$. This is the content of Theorem 2.1 below, and as a consequence of the work of Barron, Birgé and Massart (1999), it leads to adaptive estimation if $\eta \in \mathcal{G}$ for fairly large classes $\mathcal{G}$ such as Besov spaces. Theorems 3.1 and 3.2 make the connection with Barron, Birgé and Massart (1999) by bounding the smallest mean squared error by the same upper bound for the mean squared error of the penalized least squares estimator studied by these authors.

The disadvantage of this method is that not all data is directly used for estimation. However, by repeating the procedure various times, each time splitting the data in a different way, we show that we may take the average of the selected estimates, thereby taking into account all the data. This results in a more balanced estimate which does not depend on a particular split. Plotting a histogram of the selected indices can give valuable insight on the variability induced by taking random partitions of the data.

Another method which uses a data-splitting device is described in Lugosi and Nobel (1999) in the context of bounded regression ($|Y_j| \le 1$). Essentially they use one half of the data to estimate the VC-dimension of each class $\mathcal{G}_k$. Next, this estimate is plugged into a penalty term, and penalized least squares is performed to find an estimate, which balances the approximation error with the penalty term. Again, as the penalty term only provides a distribution-free upper bound for the estimation error, the trade-off is not optimal. Also, the computational issue of finding the VC-complexity is quite laborious.

Our method is related to adaptive regression by mixing (ARM), proposed by Yang (2001). The ARM algorithm combines general regression estimators $\widehat{f}_k$, which are based on one half of the data, and assigns weights to the candidate estimators via proper assessment of the performance of the likelihood of the estimates utilizing the other part of the data. Under mild conditions Yang shows that the quadratic risk of the resulting estimate is essentially bounded above by the risk of each individual candidate procedure, although a small penalty term of order $1/n$ and risk bounds involving estimates of the conditional variance function $\sigma(x)$ also appear in the upper bounds. Minimizing over procedures automatically renders optimal rate of convergence of ARM. Our selection method is general and applies to various estimates $\hat{f}_k$ as well. For Gaussian errors, the likelihood criterion considered in Yang (2001) is equivalent to the sum of squares considered in this paper. Our main motivation, as outlined earlier, differs in that we want to select a least squares estimate $\widehat{g}_k$ based on a model $\mathcal{G}_k$ with the smallest mean squared error (cf. Section 3). Identifying a good model $\mathcal{G}_k$ allows proper interpretation of certain characteristics of the regression function η. Model mixing or averaging is not suitable for this task. Another difference is that Yang's proofs are based on bounds for the Kullback–Leibler divergences using ideas of Barron (1987) and Yang and Barron (1999). Also our constants in the upper bounds for the squared error risk are much smaller and we use a mild moment condition on ε in lieu of Yang's more complicated condition A2.

In the context of variable selection in linear regression, Shao (1993) uses a data-splitting device as well, and he shows that in order to have an asymptotically correct procedure, which selects only all the relevant variables, one needs that $n/N \to 1$ and $N - n \to \infty$ as $N \to \infty$, where n is the sample size of the data set used for validation, and N is the total sample size. He also proves that the cross-validation case where $n = 1$ is inconsistent.

Devroye and Lugosi (1996, 1997) and Wegkamp (1999) use the data-splitting technique successfully to select the optimal bandwidth for kernel density estimators in an L_1 and L_2 sense, respectively, without any knowledge of the underlying density. Hengartner and Wegkamp (2001) and Hengartner, Matzner-Løber and Wegkamp (2002) explore bandwidth selection of local linear kernel smoothers in nonparametric regression using data-splitting.

The advantages of the method proposed here, besides being conceptually simple, are three-fold. First, it is computationally attractive and there are no unknown constants involved in the algorithm. Second, the method provides tight upper bounds for the mean squared error, as we do not balance the approximation error and a conservative upper bound for the estimation error. In addition, we provide finite sample bounds with explicit constants. Third, only weak moment assumptions on the errors are required.

The paper is organized as follows: We introduce a general selection method and state and prove our main result (Theorem 2.1) in Section 2. We apply the result in Section 3 for least squares estimates using models $\mathcal{G}_k$ which are VC-major classes, thereby including the important special case of finite-dimensional models.

2. Main result. Consider the additive regression model

$$Y_i = \eta(X_i) + \varepsilon_i, \qquad i = 1, \dots, N,$$

where $X_i \in \mathbb{R}^d$ are i.i.d. P distributed, $\varepsilon_i \in \mathbb{R}$ are independent with

$$\mathbb{E}\varepsilon_i = 0, \qquad \mathbb{E}\varepsilon_i^2 \le \sigma^2 < \infty \quad \text{and} \quad \mathbb{E}|\varepsilon_i|^p < \tau_p < \infty \text{ for some } p > 2,$$

and X_i and ε_i are independent. The regression function $\eta : \mathbb{R}^d \to \mathbb{R}$ is completely unknown and to be estimated from the data.

Let us describe the proposed method. First we split the data $(X_i, Y_i), i = 1, \dots, N$, into two parts. The first sample $\{(X_i, Y_i) : i \in \mathcal{I}_m\}$ contains $m \le N$ data points (X_i, Y_i) drawn without replacement from $(X_1, Y_1), \dots, (X_N, Y_N)$. We construct regression estimates $\widehat{g}_k$ based on $\{(X_i, Y_i) : i \in \mathcal{I}_m\}$ using some regression procedure. For instance, $\widehat{g}_k$ is a least squares estimate of η using a model $\mathcal{G}_k$ in Section 3. In order to select the optimal estimate $\widehat{g}$ with the smallest mean squared error among all $\widehat{g}_1, \dots, \widehat{g}_K$, we introduce some positive numbers λ_k associated with each estimate $\widehat{g}_k$. The second data set $\{(X_j, Y_j) : j \in \mathcal{J}_n\}$ consists of the remaining $n = N - m$ observations and is used for validation of the estimates $\widehat{g}_k$. We select the estimate $\widehat{g}$ which minimizes the penalized sum of squares using the second data set $\{(X_j, Y_j) : j \in \mathcal{J}_n\}$, that is, $\widehat{g} \equiv \widehat{g}_{\widehat{k}}$ with

$$\widehat{k} = \operatorname*{arg\,min}_{k=1,\dots,K} \frac{1}{n} \sum_{j \in \mathcal{J}_n} \{Y_j - \widehat{g}_k(X_j)\}^2 + \frac{\lambda_k}{n}.$$

We are ready to state our main result, which bounds the mean squared error of the selected estimate $\widehat{g}$ basically by the smallest mean squared error among the estimates $\widehat{g}_k$ plus a remainder term of order $1/n$. We denote the $L_2(P)$ semi-norm by $\|\cdot\|$.

THEOREM 2.1. *Suppose that $\|\eta - \widehat{g}_k\|_\infty < B$ for some finite constant B. Then for all $a > 0$,*

$$\mathbb{E}\|\widehat{g} - \eta\|^2 \le (1 + a) \inf_{k \le K} \left(\mathbb{E}\|\widehat{g}_k - \eta\|^2 + \frac{\lambda_k}{n} \right) + \frac{C(a, B) + \widetilde{C}(a, p)}{n}.$$

The constant $C(a, B)$ is given by

$$\frac{2(2 + a)}{a} B^2 \log^+ \left(4 \sum_{k=1}^{K} \exp\left\{ -\frac{a(1 + a)}{4(2 + a)B^2} \lambda_k \right\} \right)$$

with $\log^+ x = \max(0, \log x)$, $x > 0$, and the constant $\widetilde{C}(a, p)$, $p > 2$, is related to the error distribution as follows:

$$\widetilde{C}(a, p) = \frac{1 + a}{2} + C_p(\tau_p + \sigma^p) \frac{1 + a}{p - 2} \left(\frac{4(1 + a)}{a} \right)^{p/2} \sum_{k=1}^{K} (1 + \lambda_k)^{-(p-2)/2},$$

for some constant $C_p \le 7.35\, p / \max(1, \log p)$.

The proof is given at the end of this section.

In order to reduce the variability induced by taking a random partition of the data, it is recommended to repeat the above procedure various times. This creates estimates $\widehat{g}_s$, $s = 1, \ldots, S$, and the average estimate $\widetilde{g} = S^{-1} \sum_{s=1}^{S} \widehat{g}_s$ takes all the data into account for the estimation part, and results into a more balanced estimate in finite samples since $\widetilde{g}$ does not depend on a single particular division of the data. By Jensen's inequality and since the bound obtained in Theorem 2.1 does not depend on the particular partitioning of the data,

$$\mathbb{E}\|\widetilde{g} - \eta\|^2 = \mathbb{E}\left\| \frac{1}{S} \sum_{s=1}^{S} \widehat{g}_s - \eta \right\|^2 \le \mathbb{E} \frac{1}{S} \sum_{s=1}^{S} \|\widehat{g}_s - \eta\|^2$$

$$\le (1+a) \inf_{k \le K} \left(\mathbb{E}\|\widehat{g}_k - \eta\|^2 + \frac{\lambda_k}{n} \right) + \frac{C(a,B) + \widetilde{C}(a,p)}{n},$$

that is, we find the same bound for the mean squared error of the average estimate $\widetilde{g}$.

The bound $\|\eta - \widehat{g}_k\|_\infty \le B$ can be relaxed. The proof of Theorem 2.1 below shows in fact that for all $a > 0$ and $R > 0$,

$$\mathbb{E}\|\widehat{g} - \eta\|^2 \{\widehat{R} \le R\} \le (1+a) \inf_{k \le K} \left(\mathbb{E}\|\widehat{g}_k - \eta\|^2 + \frac{\lambda_k}{n} \right) + \frac{C(a,R) + \widetilde{C}(a,p)}{n},$$

where $\widehat{R}$ is defined as

$$\widehat{R}^2 = \sup_{k \le K} \frac{\int (\widehat{g}_k - \eta)^4 \, dP}{\int (\widehat{g}_k - \eta)^2 \, dP}.$$

The incurred error due to the selection is bounded by the remainder term $\{\widetilde{C}(a,p) + C(a,B)\}/n$, and depends on the number of moments of ε_i and the sequence $\{\lambda_k\}$. The fact that we approximate

$$\|\widehat{g}_k - \eta\|_n \equiv \frac{1}{n} \sum_{j \in \mathcal{J}_n} \left(Y_i - \widehat{g}_k(X_j) \right)^2$$

by $\|\widehat{g}_k - \eta\|^2$ for $k = 1, \ldots, K$, causes the remainder $C(a,B)/n$. This bound does not occur by formulating our result differently.

THEOREM 2.2. *For all $a > 0$,*

$$\mathbb{E}\|\widehat{g} - \eta\|_n^2 \le (1+a) \inf_{k > 0} \left(\mathbb{E}\|\widehat{g}_k - \eta\|^2 + \frac{\lambda_k}{n} \right) + \frac{1+a}{n}$$

$$+ \frac{1+a}{n} C_p (\tau_p + \sigma^p) \frac{2}{p-2} \left(\frac{1+a}{a} \right)^{p/2} \sum_{k=1}^{K} (1 + \lambda_k)^{-(p-2)/2}.$$

We will prove this theorem at the end of this section.

It follows from the proof of Theorem 2.1 below (cf. Lemma 2.5) that the constant $\widetilde{C}(a, p)$ can be improved under more restrictive assumptions. More specifically, if ε_i are Gaussian $\mathcal{N}(0, \sigma^2)$, then

$$\widetilde{C}(a, p) \le \frac{4(1+a)^2\sigma^2}{a}\left[2 + \log \sum_{k=1}^{K} \exp\left(-\frac{a\lambda_k}{8(1+a)\sigma^2}\right)\right].$$

The remainder $\widetilde{C}(a, p)/n$ is due to the fact that we observe Y_i $(= \eta(X_i) + \varepsilon_i)$ instead of $\eta(X_i)$ directly. Indeed, the proof of Lemma 2.5 shows that $\widetilde{C}(a, p) = 0$ if $\varepsilon_i = 0$, and in general, the larger the number of finite moments p, the smaller $\widetilde{C}(a, p)$.

Let us address the important issue of the choice for the λ_k's. If we do not penalize by setting $\lambda_k = 0$ for all k, the number of models $K = K_N$ seriously affects the selection error. The constants $C(a, B)$ and $\widetilde{C}(a, p)$ equal in this case

$$C(a, B) = \frac{2(2+a)}{a} B^2 \log(4eK)$$

and

$$\widetilde{C}(a, p) = \frac{1+a}{2} + KC_p(\tau_p + \sigma^p)\frac{1+a}{p-2}\left(\frac{4(1+a)}{a}\right)^{p/2}.$$

It is clear that we have to restrict the number of models K in order to control the bound for the expected selection error $\{C(a, B) + \widetilde{C}(a, p)\}/n$. Again, in case ε_i are normal $\mathcal{N}(0, \sigma^2)$ random variables, $\widetilde{C}(a, p) \le C_a\sigma^2 \log(K)$, and considering polynomially (in N) many models still leads to a remainder term $\{C(a, B) + \widetilde{C}(a, p)\}/n$ of order $\log(N)/n$.

The advantage of taking $\lambda_k \neq 0$ lies in the fact that the remainder term $\{C(a, p) + \widetilde{C}(a, B)\}/n$ is of order $1/n$ under mild moment conditions on the errors ($\mathbb{E}|\varepsilon_i|^p < \infty$ for some $p > 2$) and we are allowed to consider an infinite number of models. For instance, setting $\lambda_k = k$, we see that

$$\sum_{k=1}^{K} \lambda_k^{-(p-2)/2} \le \sum_{k=1}^{\infty} k^{-(p-2)/2} < \infty \qquad \text{if } p > 4,$$

implying that $\widetilde{C}(a, p)$ is finite regardless of the value K. There is an interesting connection between the size of λ_k and the moment conditions on the noise. For instance, taking $\lambda_k = k^2$ ensures that $\widetilde{C}(a, p) < \infty$ for $p > 2$. It should be noted that the larger λ_k, the smaller the bound for the remainder term, but on the other hand we wish to keep λ_k smaller than $\mathbb{E}\|\widehat{g}_k - \eta\|^2$ so that the first term in the upper bound for $\mathbb{E}\|\widehat{g} - \eta\|^2$ in Theorem 2.1 equals

$$(1+a) \inf_{k \le K}\left(\mathbb{E}\|\widehat{g}_k - \eta\|^2 + \frac{\lambda_k}{n}\right) \le C \inf_{k \le K} \mathbb{E}\|\widehat{g}_k - \eta\|^2$$

for some C close to 1.

PROOF OF THEOREM 2.1. First observe that

$$\mathbb{E}\|\widehat{g} - \eta\|^2 = \mathbb{E}\{\mathbb{E}[\|\widehat{g} - \eta\|^2 \mid (X_i, Y_i),\ j \in \mathcal{I}_m]\}.$$

Since $\widehat{g}_k$ only depends on the training data $\{(X_i, Y_i),\ i \in \mathcal{I}_m\}$, the class of functions $\mathcal{F} = \{\widehat{g}_1, \widehat{g}_2, \ldots\}$ is nonrandom, conditionally given $\{(X_i, Y_i),\ i \in \mathcal{I}_m\}$. The independence between the data sets $\{(X_i, Y_i), i \in \mathcal{I}_m\}$ and $\{(X_j, Y_j), j \in \mathcal{J}_n\}$ ensures that this conditioning does not change the independence structure among the variables (X_j, Y_j), $j \in \mathcal{J}_n$. Using the minimizing property of $\widehat{g}$, we show in Lemma 2.3 below that conditionally given the training data, for every k and $a \geq 0$, the inequality

$$\begin{aligned}
\|\widehat{g} - \eta\|^2 \leq\ & (1+a)\left\{\frac{1}{n}\sum_{j\in\mathcal{J}_n}\big(Y_j - \widehat{g}_k(X_j)\big)^2 - \frac{1}{n}\sum_{j\in\mathcal{J}_n}\big(Y_j - \eta(X_j)\big)^2\right\} \\
& + \sup_k\left[\frac{2(1+a)}{n}\sum_{j\in\mathcal{J}_n}\varepsilon_j(\widehat{g}_k - \eta)(X_j)\right. \\
& \qquad \left. - \frac{a}{2n}\sum_{j\in\mathcal{J}_n}(\widehat{g}_k - \eta)^2(X_j) - \frac{(1+a)\lambda_k}{2n}\right] \\
& + \sup_k\left[\int_{\mathbb{R}^d}(\widehat{g}_k - \eta)^2(x)\,dP(x)\right. \\
& \qquad \left. - \frac{(2+a)}{2n}\sum_{j\in\mathcal{J}_n}(\widehat{g}_k - \eta)^2(X_j) - \frac{(1+a)\lambda_k}{2n}\right]
\end{aligned}$$

holds. Next we observe that the expectation over the assessment data $\{(X_j, Y_j),\ j \in \mathcal{J}_n\}$ of the first term on the right-hand side of the preceding inequality equals $(1+a)\int_{\mathbb{R}^d}(\widehat{g}_k - \eta)^2(x)\,dP(x)$. Lemma 2.5 and Lemma 2.7 establish bounds for the expected values (taken with respect to the assessment data) of both remainder terms, independent of the training data $\{(X_i, Y_i), i \in \mathcal{I}_n\}$. Finally, taking expectations with respect to $\{(X_i, Y_i), i \in \mathcal{I}_n\}$ yields Theorem 2.1. □

Our proofs are novel in the way we derive Lemma 2.3 and handle the remainder terms. In particular, the use of desymmetrized empirical processes $\int_{\mathbb{R}^d}(\widehat{g}_k - \eta)^2(x)\,d(P(x) - (1+a)P_n(x))$, rather than symmetrized empirical processes $\int_{\mathbb{R}^d}(\widehat{g}_k - \eta)^2(x)\,d(P_n(x) - P(x))$, where P_n is the empirical measure putting mass $1/n$ at each X_j, $j \in \mathcal{J}_n$, results in more elegant proofs with sharp, explicit constants as no peeling devices are required as used by, for instance, van de Geer (1990) and Hengartner and Wegkamp (1999). See van de Geer [(2000), pages 70 and 149] for a clear description of the peeling device.

To simplify the notation a bit, we present the results in the following framework. Without loss of generality, we assume that the index set $\mathcal{J}_n = \{1, 2, \ldots, n\}$. We observe (X_i, Y_i) where the Y_i are related to X_i by

$$Y_i = \eta(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

with X_i independent random variables with common distribution P and ε_i are independent, mean zero random variables with a finite pth moment for some $p > 2$. In addition ε_i and X_i are independent. Let $\mathcal{F}$ be a countable class of functions and $\lambda(f) > 0$ be positive numbers. Define the sum of squares by

$$S_n(f) = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - f(X_i)\big)^2,$$

and our estimate $\widehat{f}$ of η minimizes $S_n(f) + n^{-1}\lambda(f)$ over $f \in \mathcal{F}$, that is,

$$S_n(\widehat{f}) + \frac{\lambda(\widehat{f})}{n} \le S_n(f) + \frac{\lambda(f)}{n} \qquad \text{for all } f \in \mathcal{F}.$$

The idea of introducing $\lambda(f)$ first appeared in Barron (1991) in the context of bounded regression where $|Y_i| \le 1$. In fact, the results in this section generalize Barron's (1991) result to unbounded regression.

Before stating our results we need some more notation. Let P_n be the empirical measure putting mass $1/n$ at each observation X_i, and the empirical $L_2(P_n)$ norm and inner product are defined by

$$\|f\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} f^2(X_i) \quad \text{and} \quad \langle f, g\rangle_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i) g(X_i),$$

respectively. Also, with some abuse of notation we write $\langle \varepsilon, f\rangle_n = (1/n) \sum_{i=1}^{n} \varepsilon_i \times f(X_i)$.

LEMMA 2.3. *For all $a > 0$ and $\bar{f} \in \mathcal{F}$,*

$$\begin{aligned} \|\widehat{f} - \eta\|^2 \le (1+a)\left\{ S_n(\bar{f}) - S_n(\eta) + \frac{\lambda(\bar{f})}{n} \right\} \\ + \sup_{f \in \mathcal{F}} \Bigg[2(1+a)\langle \varepsilon, f - \eta\rangle_n + \|f - \eta\|^2 \\ - (1+a)\|f - \eta\|_n^2 - (1+a)\frac{\lambda(f)}{n} \Bigg]. \end{aligned}$$

PROOF. Define (ignoring the dependence on n in the notation)

$$S(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\varepsilon_i^2 + \|f - \eta\|^2,$$

and observe that

$$\mathbb{E}S_n(f) = S(f) \qquad \text{for any fixed } f \in \mathcal{F}.$$

Also, notice that

$$S_n(f) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 + 2\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i(\eta - f)(X_i) + \|f - \eta\|_n^2.$$

Hence for all $\bar{f} \in \mathcal{F}$ and $a \geq 0$,

$$\begin{aligned}
\|\widehat{f} - \eta\|^2 &= S(\widehat{f}) - S(\eta) \\
&= (1+a)\left[S_n(\widehat{f}) - S_n(\eta) + \frac{\lambda(\widehat{f})}{n}\right] \\
&\quad + \left[\big(S - (1+a)S_n\big)(\widehat{f}) - \big(S - (1+a)S_n\big)(\eta) - (1+a)\frac{\lambda(\widehat{f})}{n}\right] \\
&\leq (1+a)\left[S_n(\bar{f}) - S_n(\eta) + \frac{\lambda(\bar{f})}{n}\right] \\
&\quad + \left[\big(S - (1+a)S_n\big)(\widehat{f}) - \big(S - (1+a)S_n\big)(\eta) - (1+a)\frac{\lambda(\widehat{f})}{n}\right] \\
&\qquad\qquad \text{(by the definition of } \widehat{f}\,) \\
&\leq (1+a)\left[S_n(\bar{f}) - S_n(\eta) + \frac{\lambda(\bar{f})}{n}\right] \\
&\quad + \sup_{f\in\mathcal{F}}\left[\big(S - (1+a)S_n\big)(f) - \big(S - (1+a)S_n\big)(\eta) - (1+a)\frac{\lambda(f)}{n}\right],
\end{aligned}$$

which after some algebra leads to the desired result. □

COROLLARY 2.4. *Define* $h_n(f) = (f-\eta)/\|f-\eta\|_n$ *if* $\|f-\eta\|_n > 0$ *and* $h_n(f) = 0$ *if* $\|f-\eta\|_n = 0$. *For all* $a > 0$, *we have*

$$\begin{aligned}
\mathbb{E}\|\widehat{f} - \eta\|^2 \leq (1+a)\inf_{f\in\mathcal{F}}\left[\|f-\eta\|^2 + \frac{\lambda(f)}{n}\right] \\
+ \mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{2(1+a)^2}{a}\langle\varepsilon, h_n(f)\rangle_n^2 - \frac{(1+a)\lambda(f)}{2n}\right) \\
+ \mathbb{E}\sup_{f\in\mathcal{F}}\left(\|f-\eta\|^2 - \frac{2+a}{2}\|f-\eta\|_n^2 - \frac{(1+a)\lambda(f)}{2n}\right).
\end{aligned}$$

PROOF. First observe that since $\bar{f}$ is fixed,

$$\mathbb{E}[S_n(\bar{f}) - S_n(\eta)] = \|\bar{f} - \eta\|^2.$$

Use the fact that $\bar{f}$ was chosen arbitrarily to take the infimum over $f \in \mathcal{F}$. Second, by the subadditivity of the supremum we find that

$$\begin{aligned}
&\sup_{f\in\mathcal{F}}\left[2(1+a)\langle\varepsilon, f-\eta\rangle_n + \|f-\eta\|^2 - (1+a)\|f-\eta\|_n^2 - (1+a)\frac{\lambda(f)}{n}\right] \\
&\quad\le \sup_{f\in\mathcal{F}}\left[2(1+a)\langle\varepsilon, f-\eta\rangle_n - \frac{a}{2}\|f-\eta\|_n^2 - \frac{1+a}{2}\frac{\lambda(f)}{n}\right] \\
&\qquad + \sup_{f\in\mathcal{F}}\left[\|f-\eta\|^2 - \frac{2+a}{2}\|f-\eta\|_n^2 - \frac{1+a}{2}\frac{\lambda(f)}{n}\right].
\end{aligned}$$

Since the algebraic inequality $2|xy| \le x^2/c + cy^2$ holds for all $x, y \in \mathbb{R}$ and $c > 0$, we find that

$$\begin{aligned}
&2(1+a)\langle\varepsilon, f-\eta\rangle_n - \frac{a}{2}\|f-\eta\|_n^2 \\
&\quad\le (1+a)\left(2\|f-\eta\|_n\left|\left\langle\varepsilon, \frac{f-\eta}{\|f-\eta\|_n}\right\rangle_n\right| - \frac{a/2}{1+a}\|f-\eta\|_n^2\right) \\
&\quad\le \frac{2(1+a)^2}{a}\left\langle\varepsilon, \frac{f-\eta}{\|f-\eta\|_n}\right\rangle_n^2,
\end{aligned}$$

and the proof is complete. □

We will now bound the remainder term

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{2(1+a)^2}{a}\langle\varepsilon, h_n(f)\rangle_n^2 - \frac{(1+a)\lambda(f)}{2n}\right).$$

We will invoke Rosenthal's inequality, which asserts that for independent random variables $Z_1, \dots, Z_n$ with mean zero, the inequality

$$\mathbb{E}\left|\sum_{k=1}^n Z_k\right|^p \le C_p \max\left(\sum_{k=1}^n \mathbb{E}|Z_k|^p, \left(\sum_{k=1}^n \mathbb{E}Z_k^2\right)^{p/2}\right)$$

holds for $p \ge 2$. Recently, Ibragimov and Sharakhmetov (1998) investigated the exact constant C_p and proved the upper bound $C_p \le 7.35p/\max(1, \log p)$.

LEMMA 2.5. *Suppose that $\tau_p < \infty$ for some $p > 2$, and define for $a > 0$, $p > 2$,*

$$\widetilde{C}(a,p) = \frac{1+a}{2} + C_p(\tau_p + \sigma^p)\frac{1+a}{p-2}\left(\frac{4(1+a)}{a}\right)^{p/2}\sum_{k=1}^K (1+\lambda_k)^{-(p-2)/2},$$

where $C_p > 0$ is the constant appearing in Rosenthal's inequality. Then for all $a > 0$, we have

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left[\frac{2(1+a)^2}{a} \langle \varepsilon, h_n(f) \rangle_n^2 - \frac{(1+a)\lambda(f)}{2n} \right] \le \frac{\widetilde{C}(a,p)}{n}.$$

Moreover, if the error distribution is Gaussian, then

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left[\frac{2(1+a)^2}{a} \langle \varepsilon, h_n(f) \rangle_n^2 - \frac{(1+a)\lambda(f)}{2n} \right] \le \frac{4(1+a)^2\sigma^2}{na} \left[2 + \log \sum_{f \in \mathcal{F}} \exp\left(-\frac{a\lambda(f)}{8(1+a)\sigma^2} \right) \right].$$

PROOF. First, we observe as in Baraud [(2000), page 484] that for some $C_p > 0$ by the Markov and Rosenthal inequalities, respectively,

$$\begin{aligned}
&\mathbb{P}\{|\langle \varepsilon, h_n(f) \rangle_n| \ge t\} \\
&\quad \le t^{-p} \mathbb{E} |\langle \varepsilon, h_n(f) \rangle_n|^p \\
&\quad \le C_p \mathbb{E} \frac{\tau_p \sum_{i=1}^n |h_n(f)(X_i)|^p + (\sigma^2 \sum_{i=1}^n h_n(f)^2(X_i))^{p/2}}{n^p t^p} \\
&\quad \le C_p \mathbb{E} \frac{\tau_p (\sum_{i=1}^n |h_n(f)(X_i)|^2)^{p/2} + (\sigma^2 \sum_{i=1}^n h_n(f)^2(X_i))^{p/2}}{n^p t^p} \\
&\quad = C_p \frac{\tau_p n^{p/2} + (\sigma^2 n)^{p/2}}{n^p t^p} \\
&\quad = C_p (\tau_p + \sigma^p) n^{-p/2} t^{-p}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\mathbb{P}\left\{ \frac{2(1+a)^2}{a} \langle \varepsilon, h_n(f) \rangle_n^2 - \frac{1+a}{2} \frac{\lambda(f)}{n} \ge \frac{t}{n} \right\} \\
&\qquad \le C_p (\tau_p + \sigma^p) \left(\frac{a}{2(1+a)^2} \left\{ \frac{1+a}{2} \lambda(f) + t \right\} \right)^{-p/2}.
\end{aligned}$$

Integrating out leads to the desired result:

$$\begin{aligned}
&\mathbb{E} \sup_{f \in \mathcal{F}} \left(\frac{2(1+a)^2}{a} \langle \varepsilon, h_n(f) \rangle_n^2 - \frac{1+a}{2} \frac{\lambda(f)}{n} \right) \\
&\quad \le \frac{1+a}{2n} + \sum_{f \in \mathcal{F}} \int_{(1+a)/2n}^{\infty} \mathbb{P}\left\{ \frac{a}{2(1+a)^2} \langle \varepsilon, h_n(f) \rangle_n^2 - \frac{1+a}{2} \frac{\lambda(f)}{n} \ge t \right\} dt \\
&\quad \le \frac{1+a}{2n}
\end{aligned}$$

$$+\frac{1}{n}\sum_{f\in\mathcal{F}}\int_{(1+a)/2}^{\infty} C_p(\tau_p+\sigma^p)\Big(\frac{a}{2(1+a)^2}\Big\{\frac{1+a}{2}\lambda(f)+t\Big\}\Big)^{-p/2}dt$$

$$=\frac{1+a}{2n}$$

$$+\frac{C_p(\tau_p+\sigma^p)}{n}\frac{2}{p-2}\Big(\frac{a}{2(1+a)^2}\Big)^{-p/2}\sum_{f\in\mathcal{F}}\Big(\frac{1+a}{2}\big(\lambda(f)+1\big)\Big)^{-p/2+1}$$

$$=\frac{1+a}{2n}+\frac{C_p(\tau_p+\sigma^p)}{n}\frac{1+a}{p-2}\Big(\frac{4(1+a)}{a}\Big)^{p/2}\sum_{f\in\mathcal{F}}\big(\lambda(f)+1\big)^{-(p-2)/2},$$

and the first claim is proved. The second assertion follows from a similar argument using the tail bounds for Gaussian random variables instead of the Markov and Rosenthal inequalities. We omit the details. □

The following proposition is used in the bound for the remainder term

$$\mathbb{E}\sup_{f\in\mathcal{F}}\Big(\|f-\eta\|^2-\frac{2+a}{2}\|f-\eta\|_n^2-\frac{(1+a)\lambda(f)}{2n}\Big).$$

This is essentially Lemma 2.1 in Einmahl and Mason (1996), which was pointed out to the author by David Mason, who also indicated that the conditions stated in their Lemma 2.1 are too strong.

PROPOSITION 2.6 [Einmahl and Mason (1996)]. *Let $Z_1, Z_2, \ldots, Z_n$ be independent, nonnegative random variables with $\mathbb{E}Z_i=\mu_i$ and $\mathbb{E}Z_i^2\le\sigma^2$. Then*

$$\mathbb{P}\Big\{\frac{1}{n}\sum_{i=1}^{n}(\mu_i-Z_i)\ge\delta\Big\}\le\exp\Big(-\frac{n\delta^2}{2\sigma^2}\Big).$$

PROOF. Since for all $t>0$,

$$\mathbb{E}\exp(-tZ_i)\le 1-t\mu_i+\frac{1}{2}t^2\mathbb{E}Z_i^2\le\exp\Big(-t\mu_i+\frac{t^2\sigma^2}{2}\Big),$$

we have using the independence of Z_i

$$\mathbb{P}\Big\{\frac{1}{n}\sum_{i=1}^{n}(\mu_i-Z_i)\ge\delta\Big\}\le\exp(-tn\delta)\mathbb{E}\exp\Big(t\sum_{i=1}^{n}(\mu_i-Z_i)\Big)$$

$$\le\exp\Big(-tn\delta+\frac{nt^2\sigma^2}{2}\Big).$$

Choosing $t=\delta/\sigma^2$ yields the result. □

LEMMA 2.7. *Define*

$$R = \sup_{f\in\mathcal{F}} \frac{P|f-\eta|^4}{P|f-\eta|^2}$$

and set

$$\Delta = \sum_{f\in\mathcal{F}} \exp\Big(-\frac{a(1+a)}{4(2+a)R}\lambda(f)\Big).$$

For all $a>0$, we have

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left(\|f-\eta\|^2 - \frac{2+a}{2}\|f-\eta\|_n^2 - \frac{(1+a)\lambda(f)}{2n}\right) \le \frac{2(2+a)}{na}R\log(4e\Delta).$$

PROOF. Set $g=(f-\eta)^2$. With some slight abuse of notation we write $\lambda(g)=\lambda(g(f))=\lambda(f)$. Also we use the short-hand notation Pg for the integral $\int_{\mathbb{R}^d} g(x)\,dP(x)$.

First we want to relate $Pg-(1+a)P_ng$ to $(P-P_n)g/\sqrt{Pg}$. For this matter, let $\delta>0$, $C>0$ and suppose that

$$Pg - P_ng \le \sqrt{\delta+C}\sqrt{Pg}.$$

Then for all $\beta>0$,

$$Pg \le (1+\beta)P_ng + (\delta+C)\frac{1+\beta}{\beta}.$$

To appreciate why, simply consider the cases $\sqrt{Pg}\le\frac{1+\beta}{\beta}\sqrt{\delta+C}$ implying $Pg\le P_ng+(\delta+C)\frac{1+\beta}{\beta}$ and its complement $\sqrt{Pg}>\frac{1+\beta}{\beta}\sqrt{\delta+C}$ implying $Pg\le P_ng+\frac{\beta}{1+\beta}Pg$, separately. This argument can be found in Anthony and Bartlett (1999). Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{\sup_g\left(Pg-\frac{2+a}{2}P_ng-\frac{(1+a)\lambda(g)}{2n}\right)\ge\delta\right\}\\
&\quad\le\sum_g\mathbb{P}\left\{Pg-\frac{2+a}{2}P_ng-\frac{(1+a)\lambda(g)}{2n}\ge\delta\right\}\\
&\quad=\sum_g\mathbb{P}\left\{Pg-\left(1+\frac{a}{2}\right)P_ng\ge\frac{1+a/2}{a/2}\left[\frac{a/2}{1+a/2}\frac{1+a}{2n}\lambda(g)+\frac{a/2}{1+a/2}\delta\right]\right\}\\
&\quad\le\sum_g\mathbb{P}\left\{\frac{Pg-P_ng}{\sqrt{Pg}}\ge\sqrt{\frac{a/2}{1+a/2}\left(\frac{1+a}{2n}\lambda(g)+\delta\right)}\right\}\\
&\qquad\qquad\text{(by the contrapositive of the reasoning above taking } \beta=a/2)
\end{aligned}$$

$$\leq \sum_g \exp\biggl(-\frac{n}{2}\frac{Pg}{Pg^2}\frac{a}{2+a}\biggl\{\frac{1+a}{2n}\lambda(g)+\delta\biggr\}\biggr)$$

[by Proposition 2.6 below or Lemma 2.1 in Einmahl and Mason (1996)]

$$\leq \sum_g \exp\biggl(-\frac{a(1+a)}{4(2+a)R}\lambda(g)\biggr)\exp\biggl(-\frac{n}{2R}\frac{a}{2+a}\delta\biggr).$$

Finally, invoke that for any random variable T with $\mathbb{P}\{T \geq t\} \leq A\exp(-Bt)$, for some $A, B > 0$ and all $t > 0$, it follows that

$$\mathbb{E}T \leq \mathbb{E}T^+ = \int_0^\infty \mathbb{P}\{T^+ \geq t\}\,dt = \int_0^\infty \mathbb{P}\{T \geq t\}\,dt \leq \frac{1+\log^+(A)}{B}. \tag{2.1}$$

The proof of both assertions now follows easily. □

PROOF OF THEOREM 2.2. Repeating the same reasoning as in the proof of Lemma 2.3, conditionally given the covariates X_i, we find for all $f \in \mathcal{F}$,

$$\mathbb{E}\|\widehat{f}-\eta\|_n^2 \leq (1+a)\biggl(\mathbb{E}\|f-\eta\|_n^2 + \frac{\lambda(f)}{n}\biggr) + \mathbb{E}\max_{f\in\mathcal{F}}\biggl(\frac{(1+a)^2}{a}\biggl\langle\varepsilon, \frac{f-\eta}{\|f-\eta\|_n}\biggr\rangle_n^2 - \frac{1+a}{n}\lambda(f)\biggr).$$

Now argue as in Lemma 2.5 to obtain

$$\mathbb{P}\biggl\{\frac{(1+a)^2}{a}\biggl\langle\varepsilon, \frac{f-\eta}{\|f-\eta\|_n}\biggr\rangle_n^2 - \frac{1+a}{n}\lambda(f) \geq \frac{t}{n}\biggr\} \leq C_p(\tau_p+\sigma^p)\biggl(\frac{a}{(1+a)^2}\bigl(t+(1+a)\lambda(f)\bigr)\biggr)^{-p/2}$$

so that

$$\begin{aligned}
&\mathbb{E}\max_{f\in\mathcal{F}}\biggl(\frac{(1+a)^2}{a}\biggl\langle\varepsilon, \frac{f-\eta}{\|f-\eta\|_n}\biggr\rangle_n^2 - \frac{1+a}{n}\lambda(f)\biggr)\\
&\quad\leq \frac{1+a}{n} + \sum_{f\in\mathcal{F}}\int_{(1+a)/n}^\infty \mathbb{P}\biggl\{\frac{(1+a)^2}{a}\langle\varepsilon, h_n(f)\rangle_n^2 - \frac{1+a}{n}\lambda(f) \geq t\biggr\}\\
&\quad\leq \frac{1+a}{n} + \sum_{f\in\mathcal{F}}\frac{C_p}{n}(\tau_p+\sigma^p)\biggl(\frac{a}{(1+a)^2}\biggr)^{-p/2}\\
&\qquad\qquad\times\frac{-1}{1-p/2}(1+a)^{1-p/2}\bigl(1+\lambda(f)\bigr)^{1-p/2}\\
&\quad= \frac{1+a}{n} + \frac{C_p}{n}(\tau_p+\sigma^p)\biggl(\frac{1+a}{a}\biggr)^{p/2}\frac{2}{p-2}(1+a)\sum_{f\in\mathcal{F}}\bigl(1+\lambda(f)\bigr)^{-(p-2)/2},
\end{aligned}$$

and the result follows. □

3. Applications. In this section we show how Theorem 2.1 can be applied to adaptive least squares estimation. However, we emphasize that the result of Theorem 2.1 is more general. For instance, we refer to Hengartner, Matzner-Løber and Wegkamp (2002) for a small simulation study in the case where the $\widehat{g}_k$ are local linear regression smoothers and k is the so-called bandwidth. Another interesting possibility is to select among various different types of estimators and find the one which performs best for the given situation (in terms of smallest mean squared error). Also the method readily extends to least absolute deviation regression. In this case, one avoids imposing moment conditions on the error distribution, instead existence of a density of ε_j with some regularity assumptions is needed.

It is known that linear spaces (e.g., splines) have good approximation properties for large classes of functions, for instance Hölderian functions (of order $\alpha > 1/2$). Certain non-linear spaces, such as non-regular histograms and neural networks provide even more flexible approximation. We refer to the work by Barron, Birgé and Massart (1999) for a detailed discussion.

Here we assume the collection of models $\mathcal{G}_k$ at hand consists of uniformly bounded VC-major classes with VC-dimension V_k. We consider least squares estimators $\widehat{g}_k$, which satisfy

$$\frac{1}{m}\sum_{i\in\mathcal{I}_m}\bigl(Y_i-\widehat{g}_k(X_i)\bigr)^2 \le \frac{1}{m}\sum_{i\in\mathcal{I}_m}\bigl(Y_i-g(X_i)\bigr)^2+\frac{1}{m} \qquad \text{for all } g\in\mathcal{G}_k.$$

The least squares estimators are not necessarily unique, which poses a problem for a given sample size n. However, the results of Theorem 3.1 and 3.2 are valid for any function $\widehat{g}_k$ with the above property. We note in passing that $\widehat{g}_k$ is completely determined at the design points $X_i, i\in\mathcal{I}_m$, and Lemma 3.3 below shows that for all $a>0$,

$$\mathbb{E}\sup_{g\in\mathcal{G}_k}\left(\int_{\mathbb{R}^d} g^2(x)\,dP(x)-\frac{1+a}{m}\sum_{i\in\mathcal{I}_m} g^2(X_i)\right)\to 0,$$

provided $V_k\log m/m\to 0$. Hence the least squares estimator will be unique with probability tending to one for all models $\mathcal{G}_k$ with $V_k\log m/m\to 0$. Observe that those cases provide the more interesting upper bounds for $\mathbb{E}\|\widehat{g}-\eta\|^2$ in Theorems 3.1 and 3.2.

THEOREM 3.1. *Let $\{\mathcal{G}_k\}$ be a sequence of VC-major classes with increasing VC-dimensions $\{V_k\}$ and $\sup_{g\in\cup_k\mathcal{G}_k}\|g-\eta\|_\infty\le B<\infty$. Assume further that $\varepsilon_i\overset{\mathcal{D}}{=}\mathcal{N}(0,\sigma^2)$. Then the method described in Section 2 with $m=n$ and $\lambda_k=V_k$ yields an estimator $\widehat{g}$ satisfying for some constants $\kappa_1>1,\kappa_2,\kappa_3,\kappa_4>0$,*

$$\mathbb{E}\|\widehat{g}-\eta\|^2\le\inf_k\left[\kappa_1\inf_{f\in\mathcal{G}_k}\|f-\eta\|^2+\kappa_2\frac{B^4V_k\log(n)}{n}\right]$$
$$+\frac{\kappa_3}{n}\log\sum_k\exp\left(-\frac{\kappa_4}{B^2\vee\sigma^2}V_k\right).$$

Although Theorem 3.1 can be derived from Theorem 2.1 above in combination with Theorem 7 in Barron, Birgé and Massart [(1999), page 357] under the weaker moment assumption $\mathbb{E}\exp(|\varepsilon|/b) \le 4$ for some $b > 0$, we decided to include its proof which is markedly distinct and simpler. Notice that the $\log n$ term also appears in Theorem 7 of Barron, Birgé and Massart (1999).

Specialized to finite dimensional models, we can considerably weaken our assumption on the error distribution (at the cost of a multiplicative $\log n$ term). Suppose that we have a collection of closed, convex subsets of finite dimensional spaces $\mathcal{G}_k$ with dimension D_k, where k is bounded by D_k. For instance, nested models with $D_k = k$ satisfy this requirement. A more specific example is the case where $\mathcal{G}_k$ is the linear space of piecewise polynomials of degree less than r on the dyadic grid on $[0, 1]$. That is, $\mathcal{G}_k$ is the linear space consisting of functions

$$g(x) = \sum_{j=1}^{2^k} \pi_j(x) \mathbb{1}_{((j-1)2^{-k}, j2^{-k}]}(x), \qquad 0 \le x \le 1,$$

where π_j are polynomials of degree less than some fixed integer $r \ge 1$. Notice that here $D_k = r2^k$ and $D_k \ge k$ is satisfied. This setting using finite-dimensional spaces is probably the most interesting one in practice, and has received much attention; see, for instance, the papers by Baraud (2000) and Barron, Birgé and Massart (1999).

THEOREM 3.2. *Let $\{\mathcal{G}_k\}_k$ be a finite sequence of closed convex finite dimensional subspaces with dimensions $D_k \ge k$ and with $\sup_{f \in \bigcup_k \mathcal{G}_k} \|f - \eta\|_\infty^2 < B < \infty$. Assume further that $\tau_p < \infty$ for some $p > 4$ and let $\sigma^2 \ge \mathbb{E}\varepsilon_i^2$. For each sample size n, consider only those $\mathcal{G}_k$ with $D_k \le n$. Then the method described in Section 2 with $m = n$ and $\lambda_k = k$ yields an estimator $\widehat{g}$ satisfying, for some numerical constants $\kappa_1, \kappa_2 > 1$ and some $C(B, \sigma^2, p) > 0$,*

$$\mathbb{E}\|\widehat{g} - \eta\|^2 \le \inf_k \left[\kappa_1 \inf_{f \in \mathcal{G}_k} \|f - \eta\|^2 + \kappa_2 \frac{B^4 D_k}{n} \log\left(\frac{2n}{D_k}\right) \right] + \frac{C(B, \sigma^2, p)}{n}.$$

Observe that the $\log n$ term is not present in Barron, Birgé and Massart [(1999) Theorem 4, page 331], but they require that the moment generating function of ε exists.

PROOF OF THEOREM 3.1. First, observe that the constants $C(a, B)$ and $\widetilde{C}(a, p)$ defined in Theorem 2.1 are finite since $\{V_k\}$ is an increasing sequence and ε_i has a Gaussian distribution. (Alternatively, we can drop the monotonicity assumption on $\{V_k\}$ and require that the number of models $K \le N^\beta$ for some $\beta > 0$.) In order to apply Theorem 2.1, we need to compute $\mathbb{E}\|\widehat{g}_k - \eta\|^2$ for each k.

Let P_n be the empirical measure putting mass $1/n$ at each $X_i, i \in \mathcal{I}_n$. Let $\|\cdot\|_n$ be the empirical $L_2(P_n)$ norm, that is, $\|f\|_n^2 = \frac{1}{n}\sum_{i \in \mathcal{I}_n} f^2(X_i)$, and we

write $\langle \varepsilon, f\rangle_n = \frac{1}{n}\sum_{i\in\mathcal{I}_n}\varepsilon_i f(X_i)$. First we link $\mathbb{E}\|\widehat{g}_k-\eta\|^2$ with $\mathbb{E}\|\widehat{g}_k-\eta\|_n^2$. It is essentially due to Vapnik (1998).

LEMMA 3.3. *Assume that*

$$R(k,q)=\min\left(\sup_{f\in\mathcal{G}_k}\frac{(P|f-\eta|^{2q})^{1/q}}{(P|f-\eta|^2)^{1/2}}\left(\frac{1}{2}\left(\frac{q-1}{q-2}\right)^{q-1}\right)^{1/q},\ \sup_{f\in\mathcal{G}_k}\|f-\eta\|_\infty^2\right)$$

is finite for some $q>2$. Let V_k be the VC dimension of the VC-major class $\mathcal{G}_k$. Then for all $a>0$,

$$\mathbb{E}\{\|\widehat{g}_k-\eta\|^2-(1+a)\|\widehat{g}_k-\eta\|_n^2\}\le\frac{4(1+a)R^2(k,q)}{na}\left\{\log(4e)+V_k\left\{1+\log^+\left(\frac{2n}{V_k}\right)\right\}\right\}.$$

PROOF. Observe that

$$\mathbb{E}\{\|\widehat{g}_k-\eta\|^2-(1+a)\|\widehat{g}_k-\eta\|_n^2\}\le\mathbb{E}\sup_{g\in\mathcal{G}_k}(P-(1+a)P_n)(g-\eta)^2.$$

Next, observe that

$$\begin{aligned}&\mathbb{P}\left\{\sup_{g\in\mathcal{G}_k}(P-(1+a)P_n)(g-\eta)^2\ge\delta\right\}\\&\quad\le\mathbb{P}\left\{\sup_{g\in\mathcal{G}_k}\frac{(P-P_n)(g-\eta)^2}{\sqrt{P(g-\eta)^2}}\ge\sqrt{\frac{a}{1+a}}\delta\right\}\\&\qquad\text{(by the same argument as in the proof of Lemma 2.7)}\\&\quad\le 4\mathcal{S}(2n,k)\exp\left(-\frac{a}{1+a}\frac{n\delta}{4R^2(k,q)}\right)\end{aligned}$$

[by Vapnik (1998), Theorem 4.2, page 139, and Theorem 5.2, page 194],

where $\mathcal{S}(2n,k)$ is the $2n$-shattering coefficient of the class

$$\{x:\ g(x)>\beta,\ g\in\mathcal{G}_k,\beta>0\}.$$

This number is related to the VC-dimension V_k by the inequality:

$$\log(\mathcal{S}(n,k))\le V_k\left\{1+\log^+\left(\frac{n}{V_k}\right)\right\};$$

see Vapnik (1998), page 192. Again invoke inequality (2.1) to conclude the proof. □

Now we focus on the analysis of $\mathbb{E}\|\widehat{g}_k-\eta\|_n^2$.

LEMMA 3.4. *For all $a > 0$,*

$$\mathbb{E}\|\widehat{g}_k - \eta\|_n^2 \le (1+a)\Big(\inf_{g\in\mathcal{G}_k}\|g-\eta\|^2 + \frac{1}{n}\Big) + \mathbb{E}\sup_{g\in\mathcal{G}_k}\big(2(1+a) < \varepsilon, g-\eta >_n - a\|g-\eta\|_n^2\big).$$

PROOF. The proof is the same as that of Lemma 2.3, but one conditions on the covariates $X_i, i \in \mathcal{I}_n$. $\square$

It remains to bound the remainder term

$$\mathbb{E}\sup_{g\in\mathcal{G}_k}\big(2(1+a) < \varepsilon, g-\eta >_n - a\|g-\eta\|_n^2\big)$$

on the right-hand side of the preceding display. We employ an approximation argument which avoids the usual chaining argument [cf., e.g., van de Geer (1990)]. Let $\mathcal{G}_k^*$ be an $1/n$ covering net of $\mathcal{G}_k$ with respect to the $L_2(P_n)$ semidistance. Hence for each $g \in \mathcal{G}_k$, there is a $g^* \in \mathcal{G}_k^*$ such that $\|g - g^*\|_n \le 1/n$. Observe that since

$$\begin{aligned}
&\mathbb{E}\big|2(1+a)\langle\varepsilon, g-\eta\rangle_n - a\|g-\eta\|_n^2 - 2(1+a)\langle\varepsilon, g^*-\eta\rangle_n - a\|g^*-\eta\|_n^2\big| \\
&\quad\le 2(1+a)\mathbb{E}\|\varepsilon\|_n\|g-g^*\|_n + a\mathbb{E}\|g-g^*\|_n\{\|g-\eta\|_n + \|g^*-\eta\|_n\} \\
&\quad\le \frac{2(1+a)\sigma}{n} + \frac{2a}{n}\Big\{\sup_g\|g\| + \|\eta\|\Big\},
\end{aligned}$$

we have

$$\begin{aligned}
&\mathbb{E}\sup_{g\in\mathcal{G}_k} 2(1+a)\langle\varepsilon, g-\eta\rangle_n - a\|g-\eta\|_n^2 \\
&\quad\le \mathbb{E}\max_{g\in\mathcal{G}_k^*}\big(2(1+a)\langle\varepsilon, g-\eta\rangle_n - a\|g-\eta\|_n^2\big) \\
&\qquad + \frac{2(1+a)\sigma}{n} + \frac{2a}{n}\Big\{\sup_g\|g\| + \|\eta\|\Big\}.
\end{aligned}$$

Since $\mathcal{G}_k$ is assumed to be a VC-major class, the cardinality $|\mathcal{G}_k| \le n^{V_k}$ for some $V_k > 0$. This, combined with the normality assumption of the errors ε_i, yields the expectation on the right in the preceding display that is of order $CV_k\log(n)/n$ for some $C > 0$ by a standard calculation. This concludes the proof of Theorem 3.1. $\square$

PROOF OF THEOREM 3.2. Since $\lambda_k = k$,

$$C(a,B) = \frac{2(2+a)B^2}{a}\left\{1 + \log^+\left(4\Big/\left\{\exp\left(\frac{a(1+a)}{4(2+a)B^2}\right) - 1\right\}\right)\right\}$$

and

$$\widetilde{C}(a,p)=\frac{1+a}{2}+C_p(\tau_p+\sigma^p)\frac{1+a}{p-2}\left(\frac{4(1+a)}{a}\right)^{p/2}\sum_{k=1}^{\infty}(1+k)^{(2-p)/2},$$

so that both $C(a,B)$ and $\widetilde{C}(a,p)$ are finite for $p>4$.

LEMMA 3.5. *We have, for all* $D_k \le n$,

$$\mathbb{E}\|\widehat{g}_k-\eta\|_n^2 \le \inf_{g\in\mathcal{G}_k}\|g-\eta\|^2+\frac{\sigma^2 D_k}{n}.$$

PROOF. Since $\mathcal{G}_k$ is a closed, convex subset of a finite dimensional space, the projection of the vector $(Y_1,\dots,Y_n)$ onto $\mathcal{G}_k$ exists. The result follows from the usual bias and variance decomposition

$$\mathbb{E}\|\widehat{g}_k-\eta\|_n^2=\mathbb{E}\inf_{g\in\mathcal{G}_k}\|g-\eta\|_n^2+\frac{1}{n^2}\sum_{i\in\mathcal{I}_n}\mathbb{E}\varepsilon_i^2 D_k,$$

and the fact that the expected value of an infimum is less than the infimum of the expected values. □

Combining Theorem 2.1, Lemma 3.3 and Lemma 3.5 yields Theorem 3.2. □

Using the (combined) data-splitting device, we recover the results obtained by Baraud (2000), Barron, Birgé and Massart (1999) and Hengartner and Wegkamp (1999). Our assumptions on the error distribution and regression function are weaker, and the implementation does not involve unknown penalties, though a careful choice of the λ_k is needed. Again we stress that the bounds for $\mathbb{E}\|\widetilde{g}-\eta\|^2$ in Theorems 3.1 and 3.2 are upper bounds of the mean squared errors. Barron, Birgé and Massart (1999) consider certain finite dimensional models $\mathcal{G}_k$ with dimension D_k and they show that for large enough penalties of the form $C_{b,B}D_k/N$, the squared error risk of the penalized least squares estimator balances the approximation error and the penalty term of each model $\mathcal{G}_k$, which generally differs from the mean squared error $\mathbb{E}\|\widehat{g}_k-\eta\|^2$, where $\widehat{g}_k$ is the least squares estimate of model $\mathcal{G}_k$. The constant $C_{b,B}$ depends on $B=\|\eta\|_\infty$ and $b>0$ implicitly defined by $\mathbb{E}\exp(|\varepsilon|/b)\le 4$, which are unknown. Taking (or "guessing") $C_{b,B}$ too large may result in choosing a model which does not achieve the smallest mean squared error among all considered models for a fixed sample size at hand. The upper bounds may be loose for a particular problem, and the model that minimizes the upper bound

$$\inf_k\left\{\inf_{f\in\mathcal{G}_k}\|f-\eta\|^2+\frac{C_{b,B}D_k}{N}\right\}$$

need not correspond to that which minimizes the mean squared error, so the penalized least squares estimator may not give the optimal result. In contrast, Theorem 2.1 shows that the risk of the estimator based on the data-splitting method is bounded by a small multiple (larger than 1) of the smallest mean squared error $\mathbb{E}\|\widehat{g}_k - \eta\|^2$ among the considered models.

Acknowledgments. I am grateful to Yale University and its Social Science Research Fund for support of this research. I would like to thank Jan Beirlant (Katholieke Universiteit Leuven) and Jon Wellner (University of Washington, Seattle) for their hospitality during my sabbatical stay, and Andrew Barron, Florentina Bunea and David Mason for discussions. I appreciate all the remarks and constructive criticism of the referees.

REFERENCES

ANTHONY, M. and BARTLETT, P. (1999). *Neural Network Learning*: *Theoretical Foundations*. Cambridge Univ. Press.

BARAUD, Y. (2000). Model selection for regression on a fixed design. *Probab. Theory Related Fields* **117** 467–493.

BARRON, A. (1987). Are Bayes rules consistent in information? In *Open Problems in Communication and Computation* (T. Cover and B. Gopinath, eds.) 85–91. Springer, Berlin.

BARRON, A. (1991). Complexity regularization with applications to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics* (G. Roussas, ed.) 561–576. Kluwer, Dordrecht.

BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413.

BARTLETT, P., BOUCHERON, S. and LUGOSI, G. (2002). Model selection and error estimation. *Machine Learning* **48** 85–113.

DEVROYE, L. and LUGOSI, G. (1996). A universally acceptable smoothing factor for kernel density estimates. *Ann. Statist.* **24** 2499–2512.

DEVROYE, L. and LUGOSI, G. (1997). Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes. *Ann. Statist.* **25** 2626–2635.

EINMAHL, U. and MASON, D. (1996). Some universal results on the behavior of increments of partial sums. *Ann. Probab.* **24** 1388–1407.

HENGARTNER, N., WEGKAMP, M. and MATZNER-LØBER, E. (2002). Bandwidth selection for local linear regression smoothers. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 791–804.

HENGARTNER, N. and WEGKAMP, M. (1999). A note on model selection procedures in nonparametric classification. Preprint, Dept. Statistics, Yale Univ.

HENGARTNER, N. and WEGKAMP, M. (2001). Estimation and selection procedures in regression: The L_1-approach. *Canad. J. Statist.* **29** 621–632.

IBRAGIMOV, R. and SHARAKHMETOV, SH. (1998). On an exact constant for the Rosenthal inequality. *Theory Probab. Appl.* **42** 294–302.

LUGOSI, G. and NOBEL, A. (1999). Adaptive model selection using empirical complexities. *Ann. Statist.* **27** 1830 – 1864.

SHAO, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88** 486–494.

VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924.

VAN DE GEER, S. (2000). *Applications of Empirical Process Theory*. Cambridge Univ. Press.

VAN DE GEER, S. and WEGKAMP, M (1996). Consistency for the least squares estimator in nonparametric regression. *Ann. Statist.* **24** 2513–2523.

VAPNIK, V. (1998). *Statistical Learning Theory*. Wiley, New York.

WEGKAMP, M. H. (1999). Quasi-universal bandwidth selection for kernel density estimators. *Canad. J. Statist.* **27** 409–420

YANG, Y. (2001). Adaptive regression by mixing. *J. Amer. Statist. Assoc.* **96** 574–588.

YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564 – 1599.

DEPARTMENT OF STATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06520
E-MAIL: marten.wegkamp@yale.edu