

A NOTE ON PENALIZED MINIMUM DISTANCE ESTIMATION IN NONPARAMETRIC REGRESSION

FLORENTINA BUNEA AND MARTEN WEGKAMP

ABSTRACT. This note introduces a penalized minimum distance regression estimate. The estimate is adaptive in the sense that it optimally balances the L^1 approximation error with the estimation error among a sequence of nested models of increasing complexity.

1. INTRODUCTION

Yatracos (1985) introduced a minimum distance density estimate, which has recently received a lot of attention, cf. Biau and Devroye (2002a and 2002b), Devroye and Lugosi (1996, 1997, 2000) and Nicolieris and Yatracos (1997). For a recent exposition of the theory developed thus far for this estimate and its many applications, we refer to Devroye and Lugosi (2000).

One of the main interesting features of this estimate is that it satisfies an oracle inequality in the L^1 distance. For instance, inspired by this estimate, Devroye and Lugosi (1996, 1997) constructed a data-based bandwidth selection method in kernel density estimation using the Parzen kernel estimate $\hat{f}_{n,h}$. Moreover, they showed that their density estimate \hat{f} satisfies

$$\limsup_{n \rightarrow \infty} \frac{\mathbb{E} \int |\hat{f}(x) - f(x)| dx}{\inf_{h>0} \mathbb{E} \int |\hat{f}_{n,h}(x) - f(x)| dx} \leq 3,$$

regardless of the underlying density f in \mathbb{R}^d , a property which is referred to as *universal*. Until this date, this is the only universal bandwidth selection method. For instance, Wegkamp (1997) established a selection method, which works for any *bounded* density in \mathbb{R}^d and the bound is not needed for implementing his algorithm. The criterion in Wegkamp (1997) is, however, the expected L^2 loss rather than the L^1 loss.

Date: March 4, 2003.

Recently, Biau and Devroye (2002a and 2002b) studied a new *penalized* density method, inspired by Yatracos' minimum density estimate. They consider a sequence of prescribed, nested classes \mathcal{F}_k of densities and their penalty is based on the VC dimension of the Yatracos classes $\{x \in \mathbb{R}^d : f(x) > g(x), f, g \in \mathcal{F}_k\}$. Their method is quite general, but it subsumes that the underlying density belongs to one of the models.

Hengartner and Wegkamp (2001) extended the results obtained by Devroye and Lugosi (1996,1997) to the regression setting. They find an estimate \hat{g} in some class \mathcal{G} such that the L_1 distance between \hat{g} and the unknown regression function is small, and establish an oracle inequality for the expected L^1 loss. The L_1 -norm possesses many attractive properties such as scale invariance and a clear visual interpretation. This note can be viewed as a continuation of their work as we adapt the minimum distance criterion to allow for model selection. Unlike Biau and Devroye (2002a), we do not require that one of the models contains the true regression function.

The rest of the paper is organized as follows: Section 2 first recalls the regression estimate obtained in Hengartner and Wegkamp (1997) and discuss its properties. Next, for any given sequence of nested model classes and suitable penalties, we introduce the penalized minimum distance regression estimator. Our main result is an oracle inequality for this estimate, and the proof is presented in Section 3.

2. PENALIZED MINIMUM DISTANCE ESTIMATION

Preliminaries. We consider the regression problem

$$Y_i = g_0(x_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where the regression function $g_0 : \mathbb{R}^d \rightarrow \mathbb{R}$ is unknown, and the errors ε_i are independent random variables with

$$\mathbb{E}\varepsilon_i = 0, \quad \text{and} \quad \max_i \mathbb{E}\varepsilon_i^2 \leq \sigma^2.$$

The covariates $x_i \in \mathbb{R}^d$ are assumed deterministic, although the results can be readily extended to random designs as well. Let μ_n be the empirical measure based on x_1, \dots, x_n and define the empirical $L_1(\mu_n)$ pseudo metric on \mathbb{R}^d as

$$\|g\|_n = \int |g| d\mu_n = \frac{1}{n} \sum_{i=1}^n |g(x_i)|.$$

Our aim is to find an estimate \hat{g} in some class \mathcal{G} such that the L_1 distance $\mathbb{E}\|\hat{g} - g_0\|_n$ is small. More precisely, we want to establish an oracle inequality of the type

$$(2.1) \quad \mathbb{E}\|\hat{g} - g_0\|_n \leq C_1 \inf_{g \in \mathcal{G}} \|g - g_0\|_n + C_2 n^{-1/2}.$$

Hengartner and Wegkamp (2001) propose to estimate the unknown regression function g_0 by \hat{g} that minimizes

$$\max_{f \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n (Y_i - g(x_i)) \operatorname{sgn}(g - f)(x_i) \right|$$

over functions $g \in \mathcal{G}$, where $\operatorname{sgn}(\cdot)$ is the sign function defined by

$$\operatorname{sgn}(x) = \begin{cases} -1 & \text{if } x < 0; \\ 0 & \text{if } x = 0; \\ +1 & \text{if } x > 0. \end{cases}$$

Indeed, this estimator satisfies (2.1) with $C_1 = 3$ and C_2 depends on the variance σ^2 and the empirical shattering coefficient of the class of differences $\{f - g, f, g \in \mathcal{G}\}$. This number is defined as

$$\mathcal{S}(x_1^n, \mathcal{G}) = \operatorname{Card}(\{(\operatorname{sgn}(f - g)(x_1), \dots, \operatorname{sgn}(f - g)(x_n)) \in \{-1, 1\}^n, f, g \in \mathcal{G}\})$$

The following result directly follows from Hengartner and Wegkamp (1997, Theorem 4).

Theorem 2.1. *Let $\tau_4 = \mathbb{E}\varepsilon^4 < \infty$ and assume that the distribution of ε_i is symmetric. Then,*

$$\begin{aligned} & \mathbb{E} \int_{\mathbb{R}^d} |\hat{g}(x) - g_0(x)| d\mu_n(x) \\ & \leq 3 \inf_{f \in \mathcal{G}} \int_{\mathbb{R}^d} |f(x) - g_0(x)| d\mu_n(x) + 4\sigma \sqrt{\frac{1 + \log 4\mathcal{S}(x_1^n, \mathcal{G})}{n}} + 2\sqrt{\frac{\tau_4 - \sigma^4}{n\sigma^2}}. \end{aligned}$$

Remark. The additional symmetry assumption on the distribution of ε_i is purely technical. A simple symmetrization trick [cf. Hengartner and Wegkamp (2001, p.625)] shows that the results are essentially the same for non-symmetric errors, only the constants are slightly worse.

Model selection. The problem is which class \mathcal{G} to consider *a priori*. On the one hand, \mathcal{G} must be fairly large in order to make the bias $\inf_{g \in \mathcal{G}} \|g - g_0\|_n$ small, but, on the other hand, it should be sparse in order to keep the variance in check. A possible solution is to consider a nested sequence of classes $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{G}_3 \cdots$ and to find the class which has the best trade-off between the bias and variance components. We propose the following two-stage procedure:

Step 1. Compute, for each $k \geq 1$, $\hat{g}_k \in \mathcal{G}_k$, which minimizes

$$\sup_{f \in \mathcal{G}_k} \left| \frac{1}{n} \sum_{i=1}^n (Y_i - g(x_i)) \operatorname{sgn}(g - f)(x_i) \right|$$

over $g \in \mathcal{G}_k$.

Notice that by Theorem 2.1 each \hat{g}_k satisfies

$$\begin{aligned} & \mathbb{E} \int_{\mathbb{R}^d} |\hat{g}(x) - g_0(x)| d\mu_n(x) \\ & \leq 3 \inf_{f \in \mathcal{G}_k} \int_{\mathbb{R}^d} |f(x) - g_0(x)| d\mu_n(x) + 4\sigma \sqrt{\frac{1 + \log 4\mathcal{S}(x_1^n, \mathcal{G}_k)}{n}} + 2\sqrt{\frac{\tau_4 - \sigma^4}{n\sigma^2}}. \end{aligned}$$

At the second stage of the procedure we select an estimator among the estimates \hat{g}_k , $k \geq 1$ using the penalty

$$\operatorname{pen}(k) = \frac{2\sigma}{\sqrt{n}} \cdot \sqrt{\log \mathcal{S}(x_1^n, \mathcal{G}_k) + 2 \log k}$$

Step 2(a). Select \hat{k} , the minimizer of

$$\begin{aligned} & \sup_k \left\{ \left| \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{g}_m(x_i)) \operatorname{sgn}(\hat{g}_m - \hat{g}_k)(x_i) \right| - \operatorname{pen}(k) \right\} + \operatorname{pen}(m) \\ & \text{over } m = 1, 2, \dots \end{aligned}$$

Step 2(b). Set $\hat{g} = \hat{g}_{\hat{k}}$.

This algorithm computes at the first step the minimum distance regression estimators \hat{g}_k , creating the class $\hat{\mathcal{G}} = \{\hat{g}_k, k \geq 1\}$. Let us explain the reasoning behind the second step. Observe that the naive choice of minimizing

$$\max_{\hat{g}_k \in \hat{\mathcal{G}}} \left| \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{g}_m(x_i)) \operatorname{sgn}(\hat{g}_m - \hat{g}_k)(x_i) \right|,$$

over $\hat{g}_m \in \hat{\mathcal{G}}$, does not work. Namely, after a little reflection, we would expect that \hat{g}_1 based on the sparsest model achieves the maximum over $\hat{g}_k \in \hat{\mathcal{G}}$ and \hat{g}_m based on

the largest model is likely to achieve the minimum. Hence we need to penalize small models when taking the maximum and large models when taking the minimum. This is precisely what the suggested algorithm does.

We have the following result:

Theorem 2.2. *Let $\tau_4 = \mathbb{E}\varepsilon^4 < \infty$ and assume that the distribution of ε_i is symmetric. Then,*

$$\begin{aligned} & \mathbb{E} \int_{\mathbb{R}^d} |\widehat{g}(x) - g_0(x)| d\mu_n(x) \\ & \leq 3 \inf_k \left\{ \mathbb{E} \int_{\mathbb{R}^d} |\widehat{g}_k(x) - g_0(x)| d\mu_n(x) + 3\text{pen}(k) \right\} + 4\sqrt{\frac{3\sigma^2}{n}} + 2\sqrt{\frac{\tau_4 - \sigma^2}{\sigma^2}}. \end{aligned}$$

Theorems 2.1 and 2.2 imply that there exist a universal constant C and a constant C_{σ^2, τ_4} such that

$$\begin{aligned} & \mathbb{E} \int_{\mathbb{R}^d} |\widehat{g}(x) - g_0(x)| d\mu_n(x) \\ & \leq \inf_k \left\{ C \inf_{g \in \mathcal{G}_k} \int_{\mathbb{R}^d} |g(x) - g_0(x)| d\mu_n(x) + C_{\sigma^2, \tau_4} \sqrt{\frac{\mathcal{S}(x_1^n, \mathcal{G}_k) + \log k}{n}} \right\}. \end{aligned}$$

Remark. Assume that ε_i are subgaussian, that is,

$$\mathbb{E} \exp(\lambda \varepsilon_i^2) \leq \Lambda$$

for some $\lambda > 0$ and $\Lambda < \infty$, and define the VC dimension V_k of the set

$$\left(\{ (\text{sgn}(f - g)(x_1), \dots, \text{sgn}(f - g)(x_n)) \in \{-1, 1\}^n, f, g \in \mathcal{G}_k \} \right)$$

as the largest integer N such that

$$\max_{x_1^N \in \mathbb{R}^{d \times N}} \mathcal{S}(x_1^N, \mathcal{G}_k) = 2^N$$

with the understanding that $V_k = \infty$ if no such integer exists. Then, following Theorem 2 in Hengartner and Wegkamp (1997), we can prove that

$$\mathbb{E} \sup_{f, g \in \mathcal{G}_k} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \text{sgn}(f - g)(x_i) \right| \leq C_{\lambda, \Lambda} \sqrt{\frac{V_k}{n}}$$

for some finite constant $C_{\lambda,\Lambda}$ and

$$\mathbb{E} \int_{\mathbb{R}^d} |\widehat{g}_k(x) - g_0(x)| d\mu_n(x) \leq 3 \inf_{g \in \mathcal{G}_k} \int_{\mathbb{R}^d} |g(x) - g_0(x)| d\mu_n(x) + 2C_{\lambda,\Lambda} \sqrt{\frac{V_k}{n}}.$$

This suggests that the logarithm of the empirical shatter coefficient $\mathcal{S}(x_1^n, \mathcal{G}_k)$ can be replaced by a constant multiple of the VC dimension V_k . This is indeed the case using a more refined chaining technique in Lemma 3.2 below, see Hengartner and Wegkamp (1997, proof of Theorem 2). Since

$$\mathcal{S}(x_1^n, \mathcal{G}_k) \leq (n+1)^{V_k},$$

cf., e.g., Devroye and Lugosi (2000, page 29), using the shatter coefficients introduces a possibly suboptimal $\log n$ term, which seems to be a modest price for assuming only $\mathbb{E}\varepsilon^4 < \infty$.

Remark. As an example for \mathcal{G}_k , we can take the collection of neural networks

$$g(x) = \sum_{i=1}^k c_i f(a_i^T x + b_i) + c_0,$$

where $f : \mathbb{R} \rightarrow \mathbb{R}$ is a sigmoid, $a_i \in \mathbb{R}^d$, $b_i, c_i \in \mathbb{R}$ with $\sum_{i=0}^k |c_i| \leq A_k$. Anthony and Bartlett (1999) show that the VC dimension of \mathcal{G}_k is of order k^2 for many neural networks, implying the bound $\mathcal{S}(x_1^n, \mathcal{G}_k) \lesssim k^2 \log n$. Another source of useful combinatorial calculations is Györfi *et al.* (2000, Chapter 16).

Remark. The results are still valid in case the variables x_i are i.i.d. from a probability measure μ . Moreover, we can, at the cost of more technicalities and worse constants, state the results in terms of the $L_1(\mu)$ distance in lieu of the empirical $L_1(\mu_n)$ distance.

Remark. In general, the variance σ^2 is unknown, but we need it in order to compute each penalty term $\text{pen}(k)$. It may be replaced by $2S^2$ for any estimate S^2 such that

$$\mathbb{P} \left\{ \frac{1}{2} \sigma^2 \leq S^2 \leq 2\sigma^2 \right\} \geq 1 - \frac{C}{n}.$$

This would only affect the constants in Theorem 2.2, and its proof only requires some minor changes.

Remark. Surely it is possible to obtain a similar result for density estimation. Our approach, however, differs in some fundamental ways from Biau and Devroye (2002). First, we use a different type of estimation procedure. Second, our approach is more general as we not assume that the true density belongs to one of the model classes, which is essential in the argument of Biau and Devroye (2002). The proof presented in this note is easier and we do not need that the σ -algebra generated by each model class is the Borel σ -algebra.

3. PROOFS

Lemma 3.1.

$$\|\hat{g} - g_0\|_n \leq \inf_m \{3\|\hat{g}_m - g_0\|_n + 3\text{pen}(m) + 2\Delta_n(m)\},$$

with

$$\Delta_n(m) = \sup_k \left\{ \sup_{f \in \mathcal{G}_m, g \in \mathcal{G}_k} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \text{sgn}(f - g)(x_i) \right| - \text{pen}(k) \right\} - \text{pen}(m)$$

Proof. First, we note that by the triangle inequality, for any m ,

$$(3.1) \quad \|\hat{g} - g_0\|_n \leq \|\hat{g} - \hat{g}_m\|_n + \|\hat{g}_m - g_0\|_n.$$

Next, after adding and subtracting appropriate terms and using the definition of \widehat{k} , we find

$$\begin{aligned}
\|\widehat{g} - \widehat{g}_m\|_n &= \int (\widehat{g} - \widehat{g}_m) \operatorname{sgn}(\widehat{g} - \widehat{g}_m) d\mu_n \\
&= \frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{g}(x_i)) \operatorname{sgn}(\widehat{g} - \widehat{g}_m)(x_i) - \operatorname{pen}(m) + \operatorname{pen}(\widehat{k}) + \\
&\quad \frac{1}{n} \sum_{i=1}^n (\widehat{g}_m(x_i) - Y_i) \operatorname{sgn}(\widehat{g} - \widehat{g}_m)(x_i) - \operatorname{pen}(\widehat{k}) + \operatorname{pen}(m) \\
&\leq \max_k \left[\left| \frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{g}(x_i)) \operatorname{sgn}(\widehat{g} - \widehat{g}_k)(x_i) \right| - \operatorname{pen}(k) \right] + \operatorname{pen}(\widehat{k}) + \\
&\quad \max_k \left[\left| \frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{g}_m(x_i)) \operatorname{sgn}(\widehat{g}_m - \widehat{g}_k)(x_i) \right| - \operatorname{pen}(k) \right] + \operatorname{pen}(m) \\
&\leq 2 \max_k \left[\left| \frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{g}_m(x_i)) \operatorname{sgn}(\widehat{g}_m - \widehat{g}_k)(x_i) \right| - \operatorname{pen}(k) \right] + 2\operatorname{pen}(m) \\
&\quad \text{by definition of } \widehat{g} \text{ and } \widehat{k} \\
&\leq 2 \max_k \left[\left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \operatorname{sgn}(\widehat{g}_m - \widehat{g}_k)(x_i) \right| - \operatorname{pen}(k) - \operatorname{pen}(m) \right] + \\
&\quad 3\operatorname{pen}(m) + 2\|\widehat{g}_m - g_0\|_n,
\end{aligned}$$

as asserted in the lemma. □

Lemma 3.2. *For all m ,*

$$(3.2) \quad \mathbb{P} \{ \Delta_n(m) > \lambda \} \leq 7 \exp \left(-\frac{n\lambda^2}{4\sigma^2} \right) + \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 > 2\sigma^2 \right\}.$$

Proof. Define the set $A_n \stackrel{\text{def}}{=} \{ \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \leq 2\sigma^2 \}$. Let τ_1, \dots, τ_n be a sequence of independent random signs, $\mathbb{P}(\tau_i = -1) = \mathbb{P}(\tau_i = +1) = 1/2$, independent of ε_i , $i \leq n$. Since ε_i has a symmetric distribution, $\tau_i \varepsilon_i$ has the same distribution as ε_i . For any

fixed $m \geq 1$, observe that

$$\begin{aligned}
& \mathbb{P} \left\{ \sup_k \sup_{g \in \mathcal{G}_m} \sup_{f \in \mathcal{G}_k} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \text{sgn}(g - f)(x_i) \right| - \text{pen}(k) - \text{pen}(m) > x, A_n \right\} \\
& \leq \sum_{k=1}^{\infty} \mathbb{P} \left\{ \max_{g \in \mathcal{G}_m} \max_{f \in \mathcal{G}_k} \left| \frac{1}{n} \sum_{i=1}^n \tau_i \varepsilon_i \text{sgn}(g - f)(x_i) \right| - \text{pen}(k) - \text{pen}(m) > x, A_n \right\} \\
& \leq \sum_{k=1}^m \mathbb{P} \left\{ \max_{g \in \mathcal{G}_m} \max_{f \in \mathcal{G}_k} \left| \frac{1}{n} \sum_{i=1}^n \tau_i \varepsilon_i \text{sgn}(g - f)(x_i) \right| - \text{pen}(k) - \text{pen}(m) > x, A_n \right\} \\
& \quad + \sum_{k=m+1}^{\infty} \mathbb{P} \left\{ \max_{g \in \mathcal{G}_m} \max_{f \in \mathcal{G}_k} \left| \frac{1}{n} \sum_{i=1}^n \tau_i \varepsilon_i \text{sgn}(g - f)(x_i) \right| - \text{pen}(k) - \text{pen}(m) > x, A_n \right\}.
\end{aligned}$$

We begin with bounding the first term on the right. Since \mathcal{G}_k are nested classes, we increase the probability by taking supremum over both $f, g \in \mathcal{G}_m$. Thus

$$\begin{aligned}
& \sum_{k=1}^m \mathbb{P} \left\{ \max_{g \in \mathcal{G}_m} \max_{f \in \mathcal{G}_k} \left| \frac{1}{n} \sum_{i=1}^n \tau_i \varepsilon_i \text{sgn}(g - f)(x_i) \right| - \text{pen}(k) - \text{pen}(m) > x, A_n \right\} \\
& \leq \sum_{k=1}^m \mathbb{P} \left\{ \max_{f, g \in \mathcal{G}_m} \left| \frac{1}{n} \sum_{i=1}^n \tau_i \varepsilon_i \text{sgn}(g - f)(x_i) \right| - \text{pen}(k) - \text{pen}(m) > x, A_n \right\} \\
& = \sum_{k=1}^m \mathbb{E} \mathbb{P} \left\{ \max_{f, g \in \mathcal{G}_m} \left| \frac{1}{n} \sum_{i=1}^n \tau_i \varepsilon_i \text{sgn}(g - f)(x_i) \right| - \text{pen}(k) - \text{pen}(m) > x, A_n \mid \varepsilon_1, \dots, \varepsilon_n \right\} \\
& \leq \sum_{k=1}^m \mathcal{S}(x_1^n, \mathcal{G}_m) \max_{f, g \in \mathcal{G}_m} \mathbb{E} \mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n \tau_i \varepsilon_i \text{sgn}(g - f)(x_i) \right| - \text{pen}(k) - \text{pen}(m) > x, A_n \mid \varepsilon_1, \dots, \varepsilon_n \right\} \\
& \quad \text{by the union bound} \\
& \leq \sum_{k=1}^m 2 \mathcal{S}_m(x_1^n) \exp \left(-\frac{n}{4\sigma^2} \{ \text{pen}(k) + \text{pen}(m) + x \}^2 \right) \\
& \quad \text{by Hoeffding's inequality} \\
& \leq \sum_{k=1}^m 2 \exp \left(-\frac{nx^2}{4\sigma^2} \right) \\
& \quad \text{by the definition of } \text{pen}(m).
\end{aligned}$$

Similarly, we can argue for the second term that

$$\begin{aligned}
& \sum_{k=m+1}^{\infty} \mathbb{P} \left\{ \max_{g \in \mathcal{G}_m} \max_{f \in \mathcal{G}_k} \left| \frac{1}{n} \sum_{i=1}^n \tau_i \varepsilon_i \operatorname{sgn}(g-f)(x_i) \right| - \operatorname{pen}(k) - \operatorname{pen}(m) > x, A_n \right\} \\
& \leq \sum_{k=m+1}^{\infty} \mathbb{P} \left\{ \max_{f, g \in \mathcal{G}_k} \left| \frac{1}{n} \sum_{i=1}^n \tau_i \varepsilon_i \operatorname{sgn}(g-f)(x_i) \right| - \operatorname{pen}(k) - \operatorname{pen}(m) > x, A_n \right\} \\
& \leq \sum_{k=m+1}^{\infty} 2 \exp \left(-\frac{nx^2}{4\sigma^2} \right).
\end{aligned}$$

Therefore, combining the two previous bounds, we conclude that

$$\begin{aligned}
& \mathbb{P} \left\{ \sup_k \sup_{g \in \mathcal{G}_m} \sup_{f \in \mathcal{G}_k} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \operatorname{sgn}(g-f)(x_i) \right| - \operatorname{pen}(k) - \operatorname{pen}(m) > x, A_n \right\} \\
& \leq 4 \frac{\pi^2}{6} \exp \left(-\frac{nx^2}{4\sigma^2} \right),
\end{aligned}$$

and the proof is complete. \square

Proof of Theorem 2.2. We are now in the position to prove Theorem 2.2. By Chebyshev's inequality, we have that

$$\mathbb{P}(A_n^C) \leq \frac{\tau_4 - \sigma^4}{n\sigma^4}.$$

Next, observe that for any m , by Cauchy-Schwarz,

$$\begin{aligned}
\mathbb{E}\Delta_n(m) &= \mathbb{E}\Delta_n(m)I_{A_n} + \mathbb{E}\Delta_n(m)I_{A_n^C} \\
&\leq \sqrt{\frac{1 + \log 7}{n/(4\sigma^2)}} + \sqrt{\frac{\tau_4 - \sigma^4}{\sigma^2}},
\end{aligned}$$

invoking that $\mathbb{E}\Delta_n^2(m) \leq \sigma^2$ and after a standard calculation involving Lemma 3.2 [cf. Devroye, Györfi and Lugosi (1996, Problem 12.1)]. This bound and Lemma 3.1 yield

$$\mathbb{E}\|\hat{g} - g_0\|_n \leq \inf_k \{3\mathbb{E}\|\hat{g}_k - g_0\|_n + 3\operatorname{pen}(k)\} + 2\sqrt{\frac{1 + \log 7}{n/(4\sigma^2)}} + 2\sqrt{\frac{\tau_4 - \sigma^4}{n\sigma^2}}$$

Combination of this bound and Theorem 2.1 gives the desired result. \square

REFERENCES

1. M. Anthony and P. Bartlett (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge: Cambridge University Press.
2. G. Biau and L. Devroye (2002a). Density Estimation by the penalized combinatorial method. *Preprint*, Université Paris VI.
3. G. Biau and L. Devroye (2002b). A note on density model size testing. *Preprint*, Université Paris VI.
4. L. Devroye, L. Györfi, and G. Lugosi (1996). *A probabilistic Theory of Pattern Recognition*. New York: Springer.
5. L. Devroye and G. Lugosi (1996). A universally acceptable smoothing factor for kernel density estimation. *Annals of Statistics*, 24, 2499 – 2512.
6. L. Devroye and G. Lugosi (1997). Nonasymptotic smoothing factors, kernel complexity and Yatracos classes. *Annals of Statistics*, 25, 2626 – 2637.
7. L. Devroye and G. Lugosi (2000). *Combinatorial methods in density estimation*. New York: Springer.
8. L. Györfi, M. Kohler, A. Krzyżak, and H. Walk (2002). *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer.
9. N. Hengartner and M. Wegkamp (2001). Estimation and Selection Procedures in Regression: an L1 approach. *Canadian Journal of Statistics*, 29 (4), 621- 632.
10. T. Nicolieris and Y.G. Yatracos (1997). Rates of convergence of estimates, Kolmogorov’s entropy and the dimensionality reduction principle in regression. *Annals of Statistics*, 25, 2493 – 2511.
11. M. Wegkamp (1999). Quasi-Universal Bandwidth Selection for Kernel Density Estimators. *Canadian Journal of Statistics*, 27(2), 409 - 420.
12. Y.G. Yatracos (1985). Rates of convergence of minimum distance estimators and Kolmogorov entropy. *Annals of Statistics*, 13, 768 – 774.

DEPARTMENT OF STATISTICS, FLORIDA STATE UNIVERSITY, TALLAHASSEE, FLORIDA 32306
E-mail address: bunea@stat.fsu.edu

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