Notes on Adaptive Estimation with Lepski's Method

Jan-Christian Hütter and Cheng Mao

October 4, 2017

Consider the model

$$Y_i = f(X_i) + \xi_i \quad \text{for } i \in [n],$$

where $f : [0,1] \to \mathbb{R}$ is an unknown function, $X_i = i/n$ and ξ_i are i.i.d. standard Gaussian variables. For $\beta \in (0,1]$ and $L > 0$, denote by $\Sigma(\beta, L)$ the class of Hölder functions f satisfying

$$|f(x) - f(y)| \le L|x-y|^\beta.$$

We are interested in adaptive estimation of a function $f \in \Sigma(\beta, L)$ at a fixed point x when β is unknown. For simplicity, we view L as a universal constant. In particular, we show that a logarithmic gap exists between nonadaptive and adaptive minimax rates (i.e. Lepski's phenomenon).

Both the method and the lower bounds appeared in [Lepskii, 1991].

1 Regressogram estimators and nonadaptive upper bounds

Let m be a positive integer and let $h = 1/m$. Partition $[0,1]$ into m intervals $\Delta_j = (\frac{j-1}{m}, \frac{j}{m}]$ where $j \in [m]$. Define the regressogram estimator of function f as the piecewise constant function

$$\hat{f}_h(x) = \frac{1}{k_j} \sum_{i=1}^n Y_i \mathbb{1}(X_i \in \Delta_j) \quad \text{for } x \in \Delta_j,\ j \in [m],$$

where $k_j = \sum_{i=1}^n \mathbb{1}(X_i \in \Delta_j)$.

Theorem 1.1. *Fix $x \in [0,1]$. The estimator $\hat{f}_h$ satisfies that*

$$\sup_{f \in \Sigma(\beta,L)} \mathbb{P}\left[|\hat{f}_h(x) - f(x)|^2 \gtrsim (nh)^{-1} s + h^{2\beta}\right] \le e^{-s}, \tag{1.1}$$

and that

$$\sup_{f \in \Sigma(\beta,L)} \mathbb{E}|\hat{f}_h(x) - f(x)|^2 \lesssim (nh)^{-1} + h^{2\beta}.$$

In particular, if β is known and $h = h^*_\beta = n^{-1/(2\beta+1)}$, then

$$\sup_{f \in \Sigma(\beta,L)} \mathbb{E}|\hat{f}_{h^*_\beta}(x) - f(x)|^2 \lesssim n^{-2\beta/(2\beta+1)}.$$

Proof. We use the bias-variance decomposition

$$\begin{aligned}\hat{f}_h(x) - f(x) &= \frac{1}{k_j}\sum_{i=1}^{n}\big[Y_i - f(X_i) + f(X_i) - f(x)\big]\mathbb{1}(X_i \in \Delta_j) \\ &= \frac{1}{k_j}\sum_{i=1}^{n}\xi_i \mathbb{1}(X_i \in \Delta_j) + \frac{1}{k_j}\sum_{i=1}^{n}\big[f(X_i) - f(x)\big]\mathbb{1}(X_i \in \Delta_j).\end{aligned}$$

The variance term has distribution $N(0, 1/k_j)$ since it is an average of k_j i.i.d. standard Gaussian variables. To bound the bias term, note that $|f(X_i) - f(x)| \lesssim h^\beta$ since $f \in \Sigma(\beta, L)$ and $|X_i - x| \le h$. Hence we have

$$|\hat{f}_h(x) - f(x)| \lesssim |g| + h^\beta$$

where $g \sim N(0, 1/k_j)$. Since $k_j \approx nh$, the results easily follow. □

2 Lower bounds for one Hölder class

We want to show that the upper bounds we just derived are the best we can hope for. In particular, we show

Theorem 2.1. *Let*

$$\psi(\beta) = n^{-\frac{2\beta}{2\beta+1}}. \tag{2.1}$$

Then, there exists a constant $c > 0$ such that

$$\inf_{\hat{f}} \sup_{f \in \Sigma(\beta, L)} \mathbb{E}\left[\frac{1}{\psi(\beta)}|\hat{f}(x_0) - f(x_0)|^2\right] \ge c, \tag{2.2}$$

where the infimum ranges over all measurable functions $\hat{f}$ on the data.

In order to do so, we reduce the problem to hypothesis testing.

2.1 General lower bound

Start by considering the worst case risk and lower bound it by only considering two candidate functions,

$$\begin{aligned}R = &\sup_{f \in \Sigma(\beta, L)} \mathbb{E}_f[|\hat{f}(x_0) - f(x_0)|^2] &(2.3)\\ \ge &\max_{f \in \{f_1, f_2\}} \mathbb{E}_f[|\hat{f}(x_0) - f(x_0)|^2] &(2.4)\\ \ge &\frac{1}{2}\left(\mathbb{E}_{f_1}[|\hat{f}(x_0) - f_1(x_0)|^2] + \mathbb{E}_{f_2}[|\hat{f}(x_0) - f_2(x_0)|^2]\right). &(2.5)\end{aligned}$$

Given any estimator $\hat{f}$, there is a natural associated test to decide between the two hypothesis

$$H_0 : f = f_1, \quad H_1 : f = f_2, \tag{2.6}$$

namely

$$\hat{T} = \begin{cases} 1, & |\hat{f}(x_0) - f_1(x_0)| \le |\hat{f}(x_0) - f_2(x_0)| \\ 2, & \text{otherwise.}\end{cases}$$

By triangle inequality, and setting $\delta = |f_1(x_0) - f_2(x_0)|$, we can conclude that

$$R \geq \frac{\delta^2}{8}\left(P_{f_1}(\hat{T}=2) + P_{f_2}(\hat{T}=1)\right) \tag{2.7}$$

for this specific choice of $\hat{T}$. Now, we forget about where $\hat{T}$ came from and lower bound it by any hypothesis test $\hat{T}$ that is identified with its rejection set $A = \{\hat{T}=2\}$.

We will lower bound $P_{f_1}(\hat{T}=2) + P_{f_2}(\hat{T}=1)$ by the χ^2 divergence between P_{f_1} and P_{f_2}. For two distributions P and Q, set

$$\chi^2(P,Q) = \int \frac{(dP-dQ)^2}{dQ} = \int \left(\frac{dP}{dQ}\right)^2 dQ - 1 \tag{2.8}$$

We can check that if $P = \prod P_i$, $Q = \prod Q_i$, then

$$\int \left(\frac{dP}{dQ}\right)^2 dQ = \int \left(\frac{\prod dP_i}{\prod dQ_i}\right) d\prod Q_i \tag{2.9}$$

$$= \prod \int \left(\frac{dP_i}{dQ_i}\right)^2 dQ_i. \tag{2.10}$$

Lemma 2.2 (Lemma 8 in [Collier et al., 2016])**.** *Let P_1, P_2 be two probability measures on $(X,\mathcal{U})$. Then, for any $q > 0$,*

$$\inf_{A\in\mathcal{U}}\{P_1(A^c) + qP_2(A)\} \geq \sup_{0<\tau<1}\left\{\frac{q\tau}{1+q\tau}1 - \tau(\chi^2(P_1,P_2)+1)\right\}. \tag{2.11}$$

Proof. We start by estimating the probability for a level set of the likelihood ratio.

$$P_1\left(\frac{dP_2}{dP_1} \geq \tau\right) = \int \mathbb{1}\left\{\frac{dP_2}{dP_1} \geq \tau\right\} dP_1 \tag{2.12}$$

$$= 1 - \int \mathbb{1}\left\{\frac{dP_2}{dP_1} < \tau\right\} dP_1 \tag{2.13}$$

$$= 1 - \int \frac{dP_1}{dP_2}\mathbb{1}\left\{\frac{dP_1}{dP_2} > \frac{1}{\tau}\right\} dP_2 \tag{2.14}$$

$$\geq 1 - \tau\int \left(\frac{dP_1}{dP_2}\right)^2 dP_2 \tag{2.15}$$

$$= 1 - \underbrace{\tau(\chi^2(P_1,P_2)+1)}_{=\alpha}. \tag{2.16}$$

Now, write

$$G = \left\{\frac{dP_2}{dP_1} \geq \tau\right\}$$

Then,

$$P_2(A) = \int \frac{dP_2}{dP_1}\mathbb{1}_A dP_1 \geq \tau P_1(A\cap G) \geq \tau(P_1(A) - \alpha). \tag{2.17}$$

Therefore,

$$\begin{aligned} P_1(A^c) + qP_2(A) &\geq \max\{P_1(A^c), qP_2(A)\} && (2.18)\\ &\geq \max\{P_1(A^c), q\tau(P_1(A) - \alpha)\} && (2.19)\\ &\geq \inf_{t\in[0,1]} \max\{1-t, q\tau(t-\alpha)\} && (2.20)\\ &= \frac{q\tau}{1+q\tau}(1-\alpha) && (2.21)\\ &= \frac{q\tau}{1+q\tau}(1-\tau(\chi^2(P_1,P_2)+1)) && \square \end{aligned}$$

2.2 χ^2 between Gaussians

It remains to control the χ^2 divergence between two Gaussians For two Gaussians with variance one, we have

$$\begin{aligned} dQ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)\\ dP &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right), \end{aligned}$$

and consequently, completing the squares,

$$\begin{aligned} \chi^2(P,Q)+1 = \int \left(\frac{dP}{dQ}\right)^2 dQ &= \frac{1}{\sqrt{2\pi}} \int \exp\left(-(x-\mu)^2 + x^2 - \frac{x^2}{2}\right) dx && (2.22)\\ &= \frac{1}{\sqrt{2\pi}} \int \exp\left(-\frac{x^2}{2} + 2\mu x - \mu^2\right) dx && (2.23)\\ &= \frac{1}{\sqrt{2\pi}} \int \exp\left(-\frac{(x-2\mu)^2}{2} + \mu^2\right) dx && (2.24)\\ &= \exp(\mu^2). && (2.25) \end{aligned}$$

2.3 Constructing hypotheses

(See [Korostelev and Korosteleva, 2011]) In our case, set

$$f_1(x) = h^\beta K\left(\frac{x-x_0}{h}\right), \quad f_2(x) \equiv 0, \quad x \in [0,1], \tag{2.26}$$

for h to be chosen later, where

$$K(u) = cK_0(2u), \quad K_0(u) = \exp\left(-\frac{1}{1-u^2}\right) \mathbb{1}\{|u| \leq 1\} \tag{2.27}$$

is a bump function Note that $K(u) = 0$ for $|u| > 1/2$ and $K(u) \in \Sigma(\beta, L)$ if c is chosen small enough (pictures).

By definition,

$$|f_2(x_0) - f_1(x_0)|^2 = K(0)^2 h^{2\beta} \tag{2.28}$$

By (2.25),

$$\chi^2(P_{f_1}, P_{f_2}) + 1 = \exp\left(\sum_{i=1}^n f_1(x_i)^2\right) \tag{2.29}$$

$$= \exp\left(\sum_{i=1}^n h^{2\beta} K\left(\frac{x_i - x_0}{h}\right)^2 \mathbb{1}\{|x_i - x_0| \le h/2\}\right) \tag{2.30}$$

$$\le \exp\left(\|K\|_\infty^2 h^{2\beta+1} n\right), \tag{2.31}$$

which we make an arbitrarily small constant if

$$h = cn^{-\frac{1}{2\beta+1}},$$

and therefore by (2.7) and Lemma 2.2 with $q = 1$, we have a lower bound with rate

$$R \gtrsim n^{-\frac{2\beta}{2\beta+1}}. \tag{2.32}$$

3 Lower bound for adaptivity

3.1 Lower bound for unbalanced rates

Now, assume that we want to attain the rates for two different Hlder classes $\Sigma(\beta_1, L)$ and $\Sigma(\beta_2, L)$, $0 < \beta_1 < \beta_2 \le 1$ simultaneously with one estimator. We will show that this is not possible if we insist on achieving the minimax rates. For the proof, we can actually proceed very similarly to the above, except that we exploit that a function that is in $\Sigma(\beta_2, L)$ is simultaneously in $\Sigma(\beta_1, L)$ and thus needs to be able to be estimated faster than what we were able to achieve with one estimator.

In particular, we will show

Theorem 3.1. *Let $0 < \beta_1 < a < \beta_2 \le 1$ and abbreviate*

$$\psi_1 = \left(\frac{\log n}{n}\right)^{\frac{2\beta_1}{2\beta_1+1}}, \quad \psi_2 = \left(\frac{1}{n}\right)^{\frac{2a}{2a+1}}, \tag{3.1}$$

Then, there is a constant $c > 0$ such that

$$\inf_{\hat{f}} \sup_{i\in\{1,2\}} \sup_{f\in\Sigma(\beta_i,L)} \mathbb{E}\left[\frac{1}{\psi_i}|\hat{f}(x_0) - f(x_0)|^2\right] \ge c, \tag{3.2}$$

where the infimum ranges over all measurable functions $\hat{f}$ on the data.

Corollary 3.2. *(i) If $\hat{f}$ is such that*

$$\sup_{f\in\Sigma(\beta_1,L)} \mathbb{E}[|\hat{f}(x_0) - f(x_0)|^2] \lesssim n^{-\frac{2\beta_1}{2\beta_1+1}}, \tag{3.3}$$

then

$$\limsup_n \sup_{f\in\Sigma(\beta_2,L)} n^{\frac{2\beta_2}{2\beta_2+1}} \mathbb{E}[|\hat{f}(x_0) - f(x_0)|^2] = \infty. \tag{3.4}$$

(ii) If $\hat{f}$ is such that

$$\sup_{f\in\Sigma(\beta_2,L)} \mathbb{E}[|\hat{f}(x_0)-f(x_0)|^2] \lesssim n^{-\frac{2\beta_2}{2\beta_2+1}}, \tag{3.5}$$

then

$$\limsup_n \sup_{f\in\Sigma(\beta_1,L)} n^{\frac{2\beta_1}{2\beta_1+1}} \mathbb{E}[|\hat{f}(x_0)-f(x_0)|^2] = \infty. \tag{3.6}$$

Proof of Corollary 3.2. (i) Under (3.5), the supremum for $i=1$ in (3.2) goes to zero, so the supremum for $i=2$ is bounded below by a constant. But $\limsup n^{\frac{2\beta_2}{2\beta_2+1}}\psi_2 = \infty$.

(ii) Similary, the supremum for $i=1$ will go to zero, so the supremum for $i=2$ will be bounded from below and is by a log factor worse than the minimax rate. □

In the remainder, we will prove Theorem 3.1.

Start the same as in (2.3), but now introduce the two rates and assume that

$$|f_2-f_1|^2/\psi_1 = \delta. \tag{3.7}$$

Setting

$$q = \frac{\psi_1}{\psi_2} \tag{3.8}$$

yields

$$R = \max\{\mathbb{E}_1[|\hat{f}-f_1|^2/\psi_1], \mathbb{E}_2[|\hat{f}-f_2|^2/\psi_2]\} \tag{3.9}$$

$$\geq \frac{\delta^2}{8}\left(P_1(\hat{T}=2)+qP_2(\hat{T}=1)\right). \tag{3.10}$$

We will continue along the same lines as before and use Lemma 2.2.

3.2 Construct alternatives

From (3.1),

$$q = n^{\frac{2a}{2a+1}-\frac{2\beta_1}{2\beta_1+1}}(\log n)^{-\frac{2\beta_1}{2\beta_1+1}} \gtrsim n^{c_1}.$$

Define two alternatives very similar to (2.26),

$$f_1(x) = h^{\beta_1}K\left(\frac{x-x_0}{h}\right), \quad f_2(x)\equiv 0, \quad x\in[0,1]$$

Pick

$$h = \left(\frac{c\log n}{n}\right)^{\frac{1}{2\beta_1+1}}$$

and note that $f_1\in\Sigma(L,\beta_1)$, $f_2\in\Sigma(L,\beta_2)$.

Then, by the same calculation as in (2.31),

$$\chi^2(P_1,P_2)+1 \leq \exp(Cc\log n) = n^{Cc}$$

So, if we pick c small enough, $\tau(\chi^2(P_1,P_2)+1)\gtrsim 1$ and $q\tau\gtrsim 1$, so we get the result.

4 Lepski's method and adaptive upper bounds

Next, we discuss Lepski's method which gives an adaptive estimator achieving a rate only slower than the nonadaptive minimax rate by a logarithmic factor. Suppose there is an unknown $\beta \in [\beta_{\min}, \beta_{\max}] \subset (0, 1]$. Choose a discrete subset

$$\mathcal{B} = \{\beta_{\min} = \beta_1 < \beta_2 < \cdots < \beta_N = \beta_{\max}\}$$

where $\beta_j - \beta_{j-1} \asymp 1/\log n$, and set

$$h_\beta = (n/\log n)^{-1/(2\beta+1)} \quad \text{and} \quad \psi_n(\beta) = h_\beta^{2\beta} = (n/\log n)^{-2\beta/(2\beta+1)}.$$

Lepski's estimator is defined as

$$\hat{f}^*(x) = \hat{f}_{\hat{\beta}}(x),$$

where

$$\hat{\beta} = \max\big\{\beta \in \mathcal{B} : |\hat{f}_{h_\beta}(x) - \hat{f}_{h_{\beta'}}(x)| \le c_0 h_{\beta'}^{\beta'} \text{ for all } \beta' \le \beta,\ \beta' \in \mathcal{B}\big\}.$$

Theorem 4.1. *Fix $x \in [0, 1]$. The estimator $\hat{f}^*$ satisfies that*

$$\sup_{\beta_{\min} \le \beta \le \beta_{\max}} \sup_{f \in \Sigma(\beta, L)} \mathbb{E}\big[|\hat{f}^*(x) - f(x)|^2/\psi_n(\beta)\big] \le C_0,$$

where C_0 is a universal constant.

Proof. First, we claim that it suffices to consider $\beta \in \mathcal{B}$. Indeed, if $\beta \in (\beta_{j-1}, \beta_j)$, then $f \in \Sigma(\beta, L) \subset \Sigma(\beta_{j-1}, L)$. Thus we just need to show that $\psi_n(\beta_{j-1}) \asymp \psi_n(\beta_j)$ (as $\psi_n(\beta)$ is squeezed in between). By the definition of $\psi_n(\beta)$ and that $\beta_j - \beta_{j-1} \asymp 1/\log n$, we obtain

$$\log \frac{\psi_n(\beta_{j-1})}{\psi_n(\beta_j)} = \log \Big(\frac{n}{\log n}\Big)^{\frac{-2\beta_{j-1}}{2\beta_{j-1}+1} + \frac{2\beta_j}{2\beta_j+1}} = \Big(\frac{-2\beta_{j-1}}{2\beta_{j-1}+1} + \frac{2\beta_j}{2\beta_j+1}\Big) \log \frac{n}{\log n} \asymp (\beta_j - \beta_{j-1}) \log n \asymp 1,$$

so the claim follows.

Now fix $f \in \Sigma(\beta_i, L)$ where $i \in [N]$. Let $\mathcal{E}_j$ be the event that $\hat{\beta} = \beta_j$. We have

$$\mathbb{E}\big[|\hat{f}^*(x) - f(x)|^2/\psi_n(\beta_i)\big] = \sum_{j=1}^N \mathbb{E}\big[|\hat{f}_{h_{\beta_j}}(x) - f(x)|^2 \psi_n(\beta_i)^{-1} \mathbb{1}(\mathcal{E}_j)\big]. \tag{4.1}$$

For $j \ge i$, on the event $\mathcal{E}_j$, it holds that

$$|\hat{f}_{h_{\beta_j}}(x) - \hat{f}_{h_{\beta_i}}(x)| \le c_0 h_{\beta_i}^{\beta_i}, \quad \text{or equivalently} \quad |\hat{f}_{h_{\beta_j}}(x) - \hat{f}_{h_{\beta_i}}(x)|^2/\psi_n(\beta_i) \le c_0^2, \tag{4.2}$$

by the definition of the estimator. Hence

$$\begin{aligned}\sum_{j=i}^N \mathbb{E}\big[|\hat{f}_{h_{\beta_j}}(x) - f(x)|^2 \psi_n(\beta_i)^{-1} \mathbb{1}(\mathcal{E}_j)\big] &\le \sum_{j=i}^N \Big(2c_0^2\, \mathbb{E}\big[\mathbb{1}(\mathcal{E}_j)\big] + 2\, \mathbb{E}\big[|\hat{f}_{h_{\beta_i}}(x) - f(x)|^2 \psi_n(\beta_i)^{-1} \mathbb{1}(\mathcal{E}_j)\big]\Big) \\ &= 2c_0^2 + 2\, \mathbb{E}\big[|\hat{f}_{h_{\beta_i}}(x) - f(x)|^2 \psi_n(\beta_i)^{-1}\big] \le C_0, \end{aligned} \tag{4.3}$$

where the last inequality follows from Theorem 1.1.

Next, we consider $j < i$. It holds that

$$\mathbb{E}\big[|\hat{f}_{h_{\beta_j}}(x) - f(x)|^2 \psi_n(\beta_i)^{-1} \mathbb{1}(\mathcal{E}_j)\big] = \int_0^\infty \mathbb{P}\Big[|\hat{f}_{h_{\beta_j}}(x) - f(x)|^2 \psi_n(\beta_i)^{-1} \mathbb{1}(\mathcal{E}_j) \geq t\Big]\, dt$$
$$\leq t_j \mathbb{P}[\mathcal{E}_j] + \int_{t_j}^\infty \mathbb{P}\Big[|\hat{f}_{h_{\beta_j}}(x) - f(x)|^2/\psi_n(\beta_i) \geq t\Big]\, dt. \quad (4.4)$$

On the event $\mathcal{E}_j$, the definition of the estimator implies that there exists $\beta' \in \mathcal{B}$ with $\beta' < \beta_i$ such that $|\hat{f}_{h_{\beta_i}}(x) - \hat{f}_{h_{\beta'}}(x)| > c_0 h_{\beta'}^{\beta'}$, i.e. $|\hat{f}_{h_{\beta_i}}(x) - \hat{f}_{h_{\beta'}}(x)|^2/\psi_n(\beta') > c_0^2$. Hence we have either $|\hat{f}_{h_{\beta_i}}(x) - f(x)|^2/\psi_n(\beta') > c_0^2/4$ or $|\hat{f}_{h_{\beta'}}(x) - f(x)|^2/\psi_n(\beta') > c_0^2/4$. Thus

$$\mathbb{P}[\mathcal{E}_j] \leq \sum_{\ell=1}^{i-1} \Big(\mathbb{P}\big[|\hat{f}_{h_{\beta_i}}(x) - \hat{f}(x)|^2/\psi_n(\beta_\ell) > c_0^2/4\big] + \mathbb{P}\big[|\hat{f}_{h_{\beta_\ell}}(x) - f(x)|^2/\psi_n(\beta_\ell) > c_0^2/4\big]\Big). \quad (4.5)$$

Note that $f \in \Sigma(\beta_i, L) \subset \Sigma(\beta_\ell, L)$ for all $\ell \leq i$. Thus it follows from (1.1) that for all $s \geq \log n$ and some constant $C > 0$,

$$\mathbb{P}_f\big[|\hat{f}_{h_{\beta_\ell}}(x) - f(x)|^2/\psi_n(\beta_\ell) \geq Cs/\log n\big] \leq e^{-s}. \quad (4.6)$$

Hence we have that if $c_0 \geq 2\sqrt{C}$ then

$$\mathbb{P}\big[|\hat{f}_{h_{\beta_\ell}}(x) - f(x)|^2/\psi_n(\beta_\ell) \geq c_0^2/4\big] \leq \exp\Big(-\frac{c_0^2}{4C}\log n\Big),$$

and

$$\mathbb{P}\big[|\hat{f}_{h_{\beta_i}}(x) - f(x)|^2/\psi_n(\beta_\ell) \geq c_0^2/4\big] \leq \mathbb{P}\big[|\hat{f}_{h_{\beta_i}}(x) - f(x)|^2/\psi_n(\beta_i) \geq c_0^2/4\big] \leq \exp\Big(-\frac{c_0^2}{4C}\log n\Big).$$

Choosing c_0 to be sufficiently large and plugging the above two bounds into (4.5), we obtain that

$$\mathbb{P}[\mathcal{E}_j] \leq n^{-c_0^2/(8C)}.$$

On the other hand, it follows from (4.6) that for $t \geq C\psi_n(\beta_j)/\psi_n(\beta_i)$,

$$\mathbb{P}\big[|\hat{f}_{h_{\beta_j}} - f(x)|^2/\psi_n(\beta_i) \geq t\big] = \mathbb{P}\big[|\hat{f}_{h_{\beta_j}} - f(x)|^2/\psi_n(\beta_j) \geq t\,\psi_n(\beta_i)/\psi_n(\beta_j)\big]$$
$$\leq \exp\Big(-\frac{t\log n}{C}\frac{\psi_n(\beta_i)}{\psi_n(\beta_j)}\Big).$$

Therefore, taking $t_j = c_1\psi_n(\beta_j)/\psi_n(\beta_i)$ in (4.4) where c_1 is a sufficiently large constant and applying the above two bounds, we see that (4.4) is bounded by

$$t_j n^{-c_0^2/(8C)} + \int_{t_j}^\infty \exp\Big(-\frac{t\log n}{C}\frac{\psi_n(\beta_i)}{\psi_n(\beta_j)}\Big)\, dt = \frac{c_1\psi_n(\beta_j)}{\psi_n(\beta_i)} n^{-c_0^2/(8C)} + \frac{C}{\log n}\frac{\psi_n(\beta_j)}{\psi_n(\beta_i)}\exp\Big(-\frac{c_1\log n}{C}\Big)$$
$$\leq 1/\log n, \quad (4.7)$$

if c_1 and c_0 are chosen to be large enough. Since $\beta_j - \beta_{j-1} \asymp 1/\log n$, we have $N \asymp \log n$. Bounding (4.4) by (4.7) and summing over $j < i$, we see that

$$\sum_{j=1}^{i-1} \mathbb{E}\big[|\hat{f}_{h_{\beta_j}}(x) - f(x)|^2 \psi_n(\beta_i)^{-1} \mathbb{1}(\mathcal{E}_j)\big] \lesssim 1. \quad (4.8)$$

Finally, (4.1), (4.3) and (4.8) together yield the theorem. □

5 Extras

1. If $f \in \Sigma(\beta_{\max}, L)$, then $\hat{f}^*$ achieves the optimal rate without the logarithm.
2. [Chichignoud, Lederer, and Wainwright, 2016] "A Practical Scheme and Fast Algorithm to Tune the Lasso With Optimality Guarantees" for an application to the Lasso
3. [Bellec, Lecué, and Tsybakov, 2016] for an application to achieving $\log(p/s)$ with the Lasso instead of the Slope

References

Pierre C. Bellec, Guillaume Lecué, and Alexandre B. Tsybakov. Slope meets lasso: Improved oracle bounds and optimality. *arXiv preprint arXiv:1605.08651*, 2016.

Michaël Chichignoud, Johannes Lederer, and Martin J. Wainwright. A practical scheme and fast algorithm to tune the lasso with optimality guarantees. *Journal of Machine Learning Research*, 17(231):1–20, 2016.

Olivier Collier, Laëtitia Comminges, Alexandre B. Tsybakov, and Nicolas Verzélen. Optimal adaptive estimation of linear functionals under sparsity. *arXiv:1611.09744 [math, stat]*, November 2016.

Aleksandr Petrovich Korostelev and Olga Korosteleva. *Mathematical Statistics: Asymptotic Minimax Theory*, volume 119. American Mathematical Soc., 2011.

O. V. Lepskii. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991.