Supplement to: Minimax adaptive dimension reduction for regression

Quentin Paris

Supplement to: Minimax adaptive dimension reduction for regression

Quentin Paris

IRMAR, ENS Cachan Bretagne, CNRS, UEB
Campus de Ker Lann
Avenue Robert Schuman, 35170 Bruz, France
quentin.paris@bretagne.ens-cachan.fr

Abstract

In this supplement, we present an oracle inequality for the performance of least-squares estimates in the context of bounded regression. The proof follows the lines devised Koltchinskii (2006). Here, focus has been put on keeping track of the dependence on some constants of the problem, which was crucial for the results of the paper Paris (2013a).

Index Terms – Regression estimation, least-squares estimates, metric entropy.

AMS 2000 Classification – 62H12, 62G08.

Contents

1 Main result

Let $\mathcal{X}$ be a metric space, let P be a probability measure on $\mathcal{X} \times \mathbb{R}$ and let (X, Y) be an $\mathcal{X} \times \mathbb{R}$-valued random variable. The regression function $f^\star$ of Y given X is defined for $x \in \mathcal{X}$ by

$$f^\star(x) = \mathbb{E}\left(Y | X = x\right). \tag{1.1}$$

In this supplement, we study the least-squares estimation of $f^\star$ based on a given class $\mathcal{F}$ of real functions defined on $\mathcal{X}$. For some $L > 0$, it will be assumed that each $f \in \mathcal{F}$ satisfies

$$\sup_{x \in \mathcal{X}} |f(x)| \leq L. \tag{1.2}$$

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of n i.i.d. random variables with same distribution P as (X, Y). The least-squares estimate f_n of $f^\star$ based on $\mathcal{F}$ is defined as any random element in $\mathcal{F}$ satisfying

$$f_n \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(X_i)\right)^2. \tag{1.3}$$

Implicitly, it will be assumed that such an element exists. The performance of f_n will be measured in terms of the mean squared error

$$\mathbb{E}\|f_n - f^\star\|^2 = \mathbb{E} \int_{\mathcal{X}} (f_n - f^\star)^2 \mathrm{d}\mu, \tag{1.4}$$

where μ stands for the distribution of X, and shown to be related to the metric entropy of $\mathcal{F}$. For any probability measure Q on $\mathcal{X}$, and all $u > 0$, we recall that the u-covering number of $\mathcal{F}$ in $\mathbb{L}^2(Q)$, denoted $N(u, \mathcal{F}, \mathbb{L}^2(Q))$, is the minimal number of metric balls with radius u that are needed to cover $\mathcal{F}$ in $\mathbb{L}^2(Q)$. Then, the u-metric entropy of $\mathcal{F}$ in $\mathbb{L}^2(Q)$ is defined by $H(u, \mathcal{F}, \mathbb{L}^2(Q)) = \ln N(u, \mathcal{F}, \mathbb{L}^2(Q))$. We introduce

$$H\left(u, \mathcal{F}\right) = \sup_{Q} \; H\left(u, \mathcal{F}, \mathbb{L}^2(Q)\right), \tag{1.5}$$

where the supremum is taken over all probability measures Q with finite support in $\mathcal{X}$. As far as we know, the idea of considering a uniform version of the metric entropy, such as $H(., \mathcal{F})$, goes back to Koltchinskii (1981) and Pollard (1982). We are now in position to state the main result of this appendix.

Theorem 1.1 (Koltchinskii, 2006)**.** *Suppose that $|Y| \leq T$ and that there exist two constants $A > 0$ and $0 < s < 2$ such that, for all $u > 0$, $H(u, \mathcal{F}) \leq Au^{-s}$. Then, there exists a constant C depending only on s and A such that, for all $\varepsilon \in (0, 1]$,*

$$\mathbb{E}\|f_n - f^\star\|^2 \leq (1 + \varepsilon) \inf_{f \in \mathcal{F}} \|f - f^\star\|^2 + C\varepsilon^{-\frac{2-s}{2+s}} \left(\frac{b}{n}\right)^{\frac{2}{2+s}} + \frac{Cb}{\varepsilon n},$$

where $b = (T + L)^2$.

A few remarks are in order. Theorem 1.1 provides an oracle inequality for the performance of the least-squares estimate based on the class $\mathcal{F}$. It shows that up to some error term of order $n^{-2/(2+s)}$, the least-squares estimate f_n performs almost as well as if an oracle gave us the best possible approximation of $f^\star$ in $\mathcal{F}$. In particular, we deduce that when $\mathcal{F}$ is chosen such that the approximation error vanishes, i.e., when

$$\inf_{f\in\mathcal{F}} \|f - f^\star\|^2 = 0,$$

then the rate of convergence of f_n towards $f^\star$ is precisely $n^{-2/(2+s)}$. Regression estimation with respect to the mean squared error has been widely studied and Theorem 1.1 is well known in essence. Important references on that subject may be found in van de Geer (1999); Kohler (2000); Baraud (2002); Györfi et al. (2002); Kohler et al. (2009) and Tsybakov (2009). The problem of the approximation properties of the class $\mathcal{F}$ will not be discussed here and we refer the reader to the papers by Cucker and Smale (2001) and DeVore et al. (2006) for a thorough study of that topic.

2 Outline of the proof

Theorem 1.1 follows from a general bound for the excess risk of empirical risk minimizers exposed in Koltchinskii (2006) and Koltchinskii (2011). In this appendix, we have focused on keeping track of the dependence on the constant T which was crucial in the proof of results in Paris (2013a). In this section we describe the main arguments of the proof.

2.1 Bias-variance decomposition

Let ℓ be the loss function defined by

$$\ell(u, y) = (y - u)^2.$$

For all $f \in \mathbb{L}^2(\mu)$, we denote $\ell \bullet f$ the function defined on $\mathcal{X} \times [-T, T]$ by

$$\ell \bullet f(x, y) = \ell(f(x), y).$$

It may be easily verified that

$$\begin{aligned} \mathbb{E}\|f_n - f^\star\|^2 &= \mathbb{E}P(\ell \bullet f_n - \ell \bullet f^\star) \\ &= \inf_{f\in\mathcal{F}} \|f - f^\star\|^2 + \mathbb{E}\left[P\left(\ell \bullet f_n\right) - \inf_{f\in\mathcal{F}} P\left(\ell \bullet f\right)\right]. \quad (2.1) \end{aligned}$$

The first term on the right side of (2.1) is usually referred to as the approximation error (or bias) and measures how far is $f^\star$ from its best possible approximation in $\mathcal{F}$. The second term, referred to as the estimation error (or

variance), measures how well the estimate f_n behaves compared to the best possible approximation of $f^\star$ in $\mathcal{F}$. Hence, if for all $g \in \mathcal{F}$ we denote its excess risk by

$$\mathcal{R}(g) = P\left(\ell \bullet g\right) - \inf_{f \in \mathcal{F}} P\left(\ell \bullet f\right),$$

we obtain the so called bias-variance decomposition

$$\mathbb{E}\|f_n - f^\star\|^2 = \inf_{f \in \mathcal{F}} \|f - f^\star\|^2 + \mathbb{E}\mathcal{R}(f_n). \tag{2.2}$$

Since $\mathcal{R}(f_n)$ is a positive random variable, we have

$$\mathbb{E}\mathcal{R}(f_n) = \int_0^{+\infty} \mathbb{P}\big(\mathcal{R}(f_n) > u\big)\mathrm{d}u,$$

so that finally, one may reduce the problem of finding a bound on the mean squared error of f_n to that of deriving bounds for the probabilities

$$\mathbb{P}\big(\mathcal{R}(f_n) > u\big), \quad u > 0.$$

2.2 A fixed-point argument

The main observation, made in particular in Koltchinskii (2006), is basically that the excess risk of the least-squares estimate can be linked to the increments of the empirical process in a specific way. To get into more details we will need some notations. For all $\delta > 0$, let $\mathcal{F}(\delta)$ be the δ-minimal set of the excess risk defined by

$$\mathcal{F}(\delta) = \left\{f \in \mathcal{F} : \mathcal{R}(f) \leq \delta\right\},$$

and set

$$\mathcal{L}(\delta) = \left\{(\ell \bullet f - \ell \bullet g) : f, g \in \mathcal{F}(\delta)\right\}.$$

Let

$$P_n = n^{-1} \sum_{i=1}^{n} \delta_{(X_i, Y_i)}$$

be the empirical distribution of the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. For any class $\mathcal{T}$ of function defined on $\mathcal{X} \times \mathbb{R}$, we denote

$$\|P_n - P\|_{\mathcal{T}} = \sup_{\mathbf{t} \in \mathcal{T}} \left|(P_n - P)\mathbf{t}\right|.$$

We are now in position to state the fundamental observation.

Lemma 2.1. *Let $\hat{\delta} = \mathcal{R}(f_n)$. Then,*

$$\hat{\delta} \leq \|P_n - P\|_{\mathcal{L}(\hat{\delta})}.$$

This result motivates the following heuristic approach for bounding $\hat{\delta}$. Let $B(\delta)$ be an explicit upper bound for $\|P_n - P\|_{\mathcal{L}(\delta)}$ which holds uniformly for all $\delta > 0$ with high probability. Then, the largest solution to the fixed point equation $B(\delta) = \delta$ is an upper bound for $\hat{\delta}$ which holds with high probability. This heuristic is made rigorous by the following general lemma.

Lemma 2.2. *Let $\{V(\delta) : \delta \geq 0\}$ be nonnegative random variables such that $V(\delta) \leq V(\delta')$ for $\delta \leq \delta'$. Let $\{B(\delta, t) : \delta \geq 0, t \geq 0\}$ be real numbers such that*

$$\mathbb{P}\left(V(\delta) \geq B(\delta, t)\right) \leq e^{-t}.$$

Let $\hat{\delta}$ be a nonnegative random variable, a priori upper bounded by a constant $\bar{\delta} > 0$, and such that $\hat{\delta} \leq V(\hat{\delta})$. Then, if for all $t \geq 0$ we set

$$\sigma_t = \inf\left\{\sigma > 0 : \sup_{\delta \geq \sigma} \frac{B\left(\delta, \frac{t\delta}{\sigma}\right)}{\delta} \leq 1\right\},$$

we obtain, for all $t \geq 0$,

$$\mathbb{P}\left(\hat{\delta} \geq \sigma_t\right) \leq e^{1-t}.$$

Taking $\hat{\delta} = \mathcal{R}(f_n)$ (which is a priori upper bounded by $(T+L)^2$) and denoting $V(\delta) = \|P_n - P\|_{\mathcal{L}(\delta)}$, one would be in position to apply Lemma 2.2 if a set $\{B(\delta, t) : \delta \geq 0, t \geq 0\}$ of real numbers satisfying the requirements of the lemma where available. The two following subsections are devoted to the problem of finding such real numbers.

2.3 Concentration

A powerful means to derive bounds for $\|P_n - P\|_{\mathcal{L}(\delta)}$ is Talagrand's concentration inequality (see Talagrand, 1996). Bousquet's inequality is an important improvement of Talagrand's concentration inequality where constants are explicit (see Bousquet, 2002). Bousquet's concentration inequality states that, for all $b > 0$, for any class $\mathcal{T}$ of $[-b, b]$-valued functions and for all $t > 0$,

$$\mathbb{P}\left(\|P_n - P\|_{\mathcal{T}} \geq \mathbb{E}\|P_n - P\|_{\mathcal{T}} + \sqrt{\frac{2t}{n}\left(\sigma^2(\mathcal{T}) + 4b\mathbb{E}\|P_n - P\|_{\mathcal{T}}\right)} + \frac{2bt}{3n}\right) \leq e^{-t},$$

where we have denoted

$$\sigma^2(\mathcal{T}) = \sup_{\mathbf{t} \in \mathcal{T}} \left(P\mathbf{t}^2 - (P\mathbf{t})^2\right).$$

An application of Bousquet's inequality in our context allows to derive the following result.

Lemma 2.3. *For all $\delta > 0$ and for all $t > 0$ we have*

$$\mathbb{P}\left(\|P_n - P\|_{\mathcal{L}(\delta)} \geq 2\mathbb{E}\|P_n - P\|_{\mathcal{L}(\delta)} + 4\sqrt{b(\delta + D)\frac{2t}{n}} + \frac{8bt}{3n}\right) \leq e^{-t},$$

where $b = (T+L)^2$ and $D = \inf_{f \in \mathcal{F}} \|f - f^\star\|^2$.

2.4 Bounding the supremum of the empirical process

The last step for constructing a set $\{B(\delta,t) : \delta \geq 0, t \geq 0\}$ of real number satisfying the requirements of Lemma 2.2 is to bound the expected supremum of the empirical process

$$\mathbb{E}\|P_n - P\|_{\mathcal{L}(\delta)},$$

appearing in the bound of Lemma 2.3. The main tools for achieving this task are the symmetrization principle combined to the bounds exposed in Giné and Koltchinskii (2006) for the expected supremum of Rademacher processes. An account of these results may also be found in Paris (2013b). The result we will need is the following. In the sequel, the notation $x \vee y$ stands for the maximum of x and y.

Lemma 2.4. *Suppose there exist constants $0 < s < 2$ and $A > 0$ such that, for all $u > 0$, we have $H(u, \mathcal{F}) \leq Au^{-s}$. Then, there exists a constant C depending only on s and A such that, for all $\delta > 0$,*

$$\mathbb{E}\|P_n - P\|_{\mathcal{L}(\delta)} \leq C\left[\frac{\sqrt{b}(\delta+D)^{(2-s)/4}}{\sqrt{n}} \vee \frac{b(\delta+D)^{-s/2}}{n} \vee \frac{b(\delta+D)^{-s/4}}{n}\right],$$

where $b = (T+L)^2$ and $D = \inf_{f\in\mathcal{F}} \|f - f^\star\|^2$.

We mention that an explicit value for C is available in the proof. We are now in position to present the proof of Theorem 1.1 based on this outline.

3 Proof of Theorem A.1.1

First, let us prove that there exists a constant $C > 0$, depending only on s and A, such that, for all $t \geq 0$ and for all $\varepsilon \in (0,1]$,

$$\mathbb{P}\left(\mathcal{R}(f_n) \geq \varepsilon D + C\varepsilon^{-\frac{2-s}{2+s}}\left(\frac{b}{n}\right)^{\frac{2}{2+s}} \vee \frac{Ct}{\varepsilon}\left(\frac{b}{n}\right)\right) \leq e^{1-t}, \tag{3.1}$$

where $b = (T+L)^2$ and $D = \inf_{f\in\mathcal{F}} \|f - f^\star\|^2$. To that aim, let $t > 0$ and $\varepsilon \in (0,1]$ be fixed. In the proof, $C > 0$ will denote a constant depending only on s and A and which value may change from line to line. According to Lemmas 2.2, 2.3 and 2.4, we obtain

$$\mathbb{P}\left(\mathcal{R}(f_n) \geq \sigma_{n,t}\right) \leq e^{1-t},$$

where,

$$\sigma_{n,t} = \inf\left\{\sigma > 0 : \sup_{\delta\geq\sigma} \frac{B_n\left(\delta, \frac{t\delta}{\sigma}\right)}{\delta} \leq 1\right\},$$

with

$$B_n\left(\delta, \frac{t\delta}{\sigma}\right) = C\left[\frac{\sqrt{b}(\delta+D)^{(2-s)/4}}{\sqrt{n}} \vee \frac{b(\delta+D)^{-s/2}}{n} \vee \frac{b(\delta+D)^{-s/4}}{n}\right] + C\left[\sqrt{b(\delta+D)\frac{t\delta}{n\sigma}} + \frac{bt\delta}{n\sigma}\right].$$

If $\sigma_{n,t} \leq \varepsilon D$, then (3.1) is obvious. Therefore, equation (3.1) needs only to proved for $\sigma_{n,t} > \varepsilon D$. Fix $\sigma > \sigma_{n,t} > \varepsilon D$. Then, for all $\delta \geq \sigma$,

$$\begin{aligned}
&\frac{\sqrt{b}(\delta+D)^{(2-s)/4}}{\sqrt{n}} \vee \frac{b(\delta+D)^{-s/2}}{n} \vee \frac{b(\delta+D)^{-s/4}}{n} \\
&= \frac{\sqrt{b}\delta^{(2-s)/4}\left(1+\frac{D}{\delta}\right)^{(2-s)/4}}{\sqrt{n}} \vee \frac{b\delta^{-s/2}\left(1+\frac{D}{\delta}\right)^{-s/2}}{n} \vee \frac{b\delta^{-s/4}\left(1+\frac{D}{\delta}\right)^{-s/4}}{n} \\
&\leq \frac{\sqrt{b}\delta^{(2-s)/4}\left(1+\frac{1}{\varepsilon}\right)^{(2-s)/4}}{\sqrt{n}} \vee \frac{b\delta^{-s/2}\left(1+\frac{1}{\varepsilon}\right)^{-s/2}}{n} \vee \frac{b\delta^{-s/4}\left(1+\frac{1}{\varepsilon}\right)^{-s/4}}{n} \\
&\leq 2\sqrt{\frac{b}{n}}\left(\frac{\delta}{\varepsilon}\right)^{(2-s)/4} \vee \frac{b\delta^{-s/2}}{n} \vee \frac{b\delta^{-s/4}}{n}.
\end{aligned}$$

Also, for all $\delta \geq \sigma$,

$$\sqrt{b(\delta+D)\frac{t\delta}{n\sigma}} \leq \sqrt{\frac{2bt\delta^2}{\varepsilon n\sigma}}.$$

As a result, for all $\sigma > \sigma_{n,t}$,

$$\begin{aligned}
\sup_{\delta\geq\sigma}\frac{B_n\left(\delta,\frac{t\delta}{\sigma}\right)}{\delta} &\leq C\left[\frac{\sqrt{b}}{\varepsilon^{(2-s)/4}\sigma^{(2+s)/4}\sqrt{n}} \vee \frac{b}{n\sigma^{(2+s)/2}} \vee \frac{b}{n\sigma^{(4+s)/4}}\right] \\
&\quad + C\left[\sqrt{\frac{bt}{\varepsilon n\sigma}} + \frac{bt}{n\sigma}\right] \\
&=: h_1(\sigma) + h_2(\sigma).
\end{aligned}$$

By monotonicity, this implies

$$\sigma_{n,t} \leq \inf\left\{\sigma > 0 : h_1(\sigma) + h_2(\sigma) \leq 1\right\}.$$

Since both h_1 and h_2 are nonincreasing, it may be easily verified that

$$\inf\left\{\sigma : h_1(\sigma) + h_2(\sigma) \leq 1\right\} \leq \inf\left\{\sigma : h_1(\sigma) \leq \tfrac{1}{2}\right\} \vee \inf\left\{\sigma : h_2(\sigma) \leq \tfrac{1}{2}\right\},$$

where

$$\inf\left\{\sigma : h_1(\sigma) \leq \tfrac{1}{2}\right\} \leq \frac{C}{\varepsilon^{(2-s)/(2+s)}}\left(\frac{b}{n}\right)^{2/(2+s)},$$

and

$$\inf\left\{\sigma : h_2(\sigma) \leq \tfrac{1}{2}\right\} \leq \frac{Cbt}{\varepsilon n}.$$

Finally, for some constant C depending only on s and A,

$$\sigma_{n,t} \leq C \left[\frac{1}{\varepsilon^{(2-s)/(2+s)}} \left(\frac{b}{n} \right)^{2/(2+s)} \vee \frac{bt}{\varepsilon n} \right],$$

which completes the proof of equation (3.1).

To complete the proof of Theorem 1.1, we proceed as follows. The bias-variance decomposition

$$\mathbb{E}\|f_n - f^\star\|^2 = D + \mathbb{E}\mathcal{R}(f_n)$$

shows that we only need to prove that

$$\mathbb{E}\mathcal{R}(f_n) \leq \varepsilon D + C\varepsilon^{-\frac{2-s}{2+s}} \left(\frac{b}{n} \right)^{\frac{2}{2+s}} + \frac{Cb}{\varepsilon n}.$$

According to (3.1), for all $u > 0$,

$$\mathbb{P}\left(\mathcal{R}(f_n) \geq u\right) \leq C \exp\left(-\frac{n\varepsilon}{Cb} \left[u - \varepsilon D - C\varepsilon^{-\frac{2-s}{2+s}} \left(\frac{b}{n} \right)^{\frac{2}{2+s}} \right] \right).$$

Hence, denoting

$$A_{\varepsilon,n} = \varepsilon D + C\varepsilon^{-\frac{2-s}{2+s}} \left(\frac{b}{n} \right)^{\frac{2}{2+s}},$$

we deduce that

$$\begin{aligned}
\mathbb{E}\mathcal{R}(f_n) &= \int_0^{+\infty} \mathbb{P}\left(\mathcal{R}(f_n) \geq u\right) \mathrm{d}u \\
&= \int_0^{A_{\varepsilon,n}} \mathbb{P}\left(\mathcal{R}(f_n) \geq u\right) \mathrm{d}u + \int_{A_{\varepsilon,n}}^{+\infty} \mathbb{P}\left(\mathcal{R}(f_n) \geq u\right) \mathrm{d}u \\
&\leq A_{\varepsilon,n} + \int_{A_{\varepsilon,n}}^{+\infty} \mathbb{P}\left(\mathcal{R}(f_n) \geq u\right) \mathrm{d}u \\
&\leq A_{\varepsilon,n} + C \int_{A_{\varepsilon,n}}^{+\infty} \exp\left(-\frac{n\varepsilon}{Cb} \left\{ u - A_{\varepsilon,n} \right\} \right) \mathrm{d}u \\
&= A_{\varepsilon,n} + C \int_0^{+\infty} \exp\left(-\frac{n\varepsilon u}{Cb} \right) \mathrm{d}u \\
&= \varepsilon D + C\varepsilon^{-\frac{2-s}{2+s}} \left(\frac{b}{n} \right)^{\frac{2}{2+s}} + \frac{Cb}{\varepsilon n}.
\end{aligned}$$

This concludes the proof. □

4 Proofs of the technical results

4.1 Proof of Lemma 2.1

Let $\varepsilon > 0$ be such that $0 < \varepsilon \leq \hat{\delta}$, and let f' be arbitrarily chosen in $\mathcal{F}(\varepsilon)$. Then, we have

$$\begin{aligned}\hat{\delta} &= P(\ell \bullet f_n) - \inf_{f \in \mathcal{F}} P(\ell \bullet f) \\ &\leq P(\ell \bullet f_n - \ell \bullet f') + \varepsilon \\ &= (P - P_n)(\ell \bullet f_n - \ell \bullet f') + P_n(\ell \bullet f_n - \ell \bullet f') + \varepsilon.\end{aligned}$$

We deduce from the definition of f_n that $P_n(\ell \bullet f_n - \ell \bullet f') \leq 0$ which leads to

$$\begin{aligned}\hat{\delta} &\leq (P - P_n)(\ell \bullet f_n - \ell \bullet f') + \varepsilon \\ &\leq \|P - P_n\|_{\mathcal{L}(\hat{\delta})} + \varepsilon.\end{aligned}$$

Since this inequality holds for arbitrary $\varepsilon > 0$, the result follows by taking $\varepsilon \downarrow 0$. □

4.2 Proof of Lemma 2.2

The bound is obvious if $t \leq 1$. Therefore, we will assume $t > 1$. The proof will be divided into two steps.

Step 1. Let $\delta_j, j \geq 0$ be a decreasing sequence of positive numbers with $\delta_0 = \bar{\delta}$ and let $t_j, j \geq 0$ be a sequence of positive numbers. For all $\delta \geq 0$, denote

$$\bar{B}(\delta) = \sum_{j=0}^{+\infty} B(\delta_j, t_j)\, \mathbf{1}\{\delta_{j+1} < \delta \leq \delta_j\},$$

and set

$$\delta^* = \sup\left\{\delta \geq 0 : \delta \leq \bar{B}(\delta)\right\}.$$

The goal of this first step is to prove that

$$\forall \delta \geq \delta^* : \quad \mathbb{P}\left(\hat{\delta} \geq \delta\right) \leq \sum_{\delta_j \geq \delta} e^{-t_j}.$$

Fix $\delta > \delta^*$. If for all $j \geq 0$, we denote $E_j = \left\{V(\delta_j) \leq \bar{B}(\delta_j)\right\}$ and

$$E = \bigcap_{\delta_j \geq \delta} E_j,$$

it may be easily verified that

$$\mathbb{P}(E) \geq 1 - \sum_{\delta_j \geq \delta} e^{-t_j}.$$

On the event E, for all $\delta' \geq \delta$ we have $V(\delta') \leq \bar{B}(\delta')$ by monotonicity of V and by definition of $\bar{B}$. Thus, on the event $\{\hat{\delta} \geq \delta\} \cap E$ we obtain

$$\hat{\delta} \leq V(\hat{\delta}) \leq \bar{B}(\hat{\delta}),$$

which implies that $\delta \leq \hat{\delta} \leq \delta^*$. Since this contradicts $\delta > \delta^*$, we deduce that $\{\hat{\delta} \geq \delta\} \subset E^c$ which implies that

$$\mathbb{P}\left(\hat{\delta} \geq \delta\right) \leq \sum_{\delta_j \geq \delta} e^{-t_j}.$$

By continuity, this also holds for $\delta = \delta^*$.

Step 2. Fix $\sigma > \sigma_t$. Then, for all $\delta \geq 0$, let

$$\bar{B}_\sigma(\delta, t) = \sum_{j=0}^{+\infty} B\left(\frac{\bar{\delta}}{2^j}, \frac{t\bar{\delta}}{\sigma 2^j}\right) \mathbf{1}\left\{\frac{\bar{\delta}}{2^{j+1}} < \delta \leq \frac{\bar{\delta}}{2^j}\right\}.$$

It may be easily verified that

$$\frac{\bar{B}_\sigma(\sigma, t)}{\sigma} \leq \sup_{\delta \geq \sigma} \frac{B\left(\delta, \frac{t\delta}{\sigma}\right)}{\delta} \leq 1,$$

which implies that

$$\sigma \geq \delta_t^* = \sup\left\{\delta \geq 0 : \delta \leq \bar{B}_\sigma(\delta, t)\right\}.$$

Then, according to step 1, we deduce that

$$\mathbb{P}\left(\hat{\delta} \geq \sigma\right) \leq \sum_{\frac{\bar{\delta}}{2^j} \geq \sigma} e^{-\frac{t\bar{\delta}}{\sigma 2^j}}.$$

The sum on the right hand side may be bounded as follows. Let

$$j_* = \max\left\{j \geq 0 : \frac{\bar{\delta}}{2^j} \geq \sigma\right\}.$$

Then

$$\sum_{\frac{\bar{\delta}}{2^j} \geq \sigma} e^{-\frac{t\bar{\delta}}{\sigma 2^j}} = \sum_{j=0}^{j_*} e^{-\frac{t\bar{\delta}}{\sigma 2^j}} \leq \sum_{j=0}^{+\infty} e^{-t2^j},$$

and

$$\sum_{j=0}^{+\infty} e^{-t2^j} \leq e^{-t} + \sum_{j=1}^{+\infty} (2^j - 2^{j-1}) e^{-t2^j} \leq e^{-t} + \int_1^{+\infty} e^{-tu} \mathrm{d}u \leq 2e^{-t}.$$

Finally, we have proved that for all $t \geq 0$ and for all $\sigma > \sigma_t$ we have

$$\mathbb{P}\left(\hat{u} \geq \sigma\right) \leq e^{1-t}.$$

The result follows by continuity. □

4.3 Proof of Lemma 2.3

Let $\delta > 0$ be fixed. For all $f \in \mathcal{F}$ we have $0 \leq \ell \bullet f \leq b$. Therefore, $\mathcal{L}(\delta)$ is a class of $[-b, b]$-valued functions. Denote for simplicity $V = \|P_n - P\|_{\mathcal{L}(\delta)}$ and $\sigma^2 = \sigma^2(\mathcal{L}(\delta))$. Using the inequalities $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $2\sqrt{uv} \leq u + v$, for $u, v \geq 0$, we obtain that

$$\mathbb{E}V + \sqrt{\frac{2t}{n}\left(\sigma^2 + 4b\mathbb{E}V\right)} + \frac{2bt}{3n} \leq 2\mathbb{E}V + \sigma\sqrt{\frac{2t}{n}} + \frac{8bt}{3n}.$$

As a result, we deduce from Bousquet's inequality that

$$\mathbb{P}\left(V \geq 2\mathbb{E}V + \sigma\sqrt{\frac{2t}{n}} + \frac{8bt}{3n}\right) \leq e^{-t}.$$

It remains only to prove that $\sigma \leq 4\sqrt{b(\delta + D)}$. To that aim, observe that for any $f \in \mathcal{F}$ we have

$$\begin{aligned}
P(\ell \bullet f - \ell \bullet f^\star)^2 &= \mathbb{E}\left[\left((Y - f(X))^2 - (Y - f^\star(X))^2\right)^2\right] \\
&= \mathbb{E}\left[(2Y - f(X) - f^\star(X))^2 (f(X) - f^\star(X))^2\right] \\
&\leq 4b\ \mathbb{E}\left(f(X) - f^\star(X)\right)^2 \\
&= 4b\ P(\ell \bullet f - \ell \bullet f^\star) \\
&= 4b(\mathcal{R}(f) + D).
\end{aligned}$$

Then, for all $f, g \in \mathcal{F}(\delta)$, we have

$$\begin{aligned}
\sqrt{P(\ell \bullet f - \ell \bullet g)^2} &\leq \sqrt{P(\ell \bullet f - \ell \bullet f^\star)^2} + \sqrt{P(\ell \bullet g - \ell \bullet f^\star)^2} \\
&\leq 4\sqrt{b(\delta + D)}.
\end{aligned}$$

Finally, we have proved that

$$\sigma \leq \sup_{\mathbf{t} \in \mathcal{L}(\delta)} \sqrt{P\mathbf{t}^2} \leq 4\sqrt{b(\delta + D)}.$$

The proof is complete. □

4.4 Proof of Lemma 2.4

For simplicity, we assume that there exists $\bar{f} \in \mathcal{F}$ such that

$$P\left(\ell \bullet \bar{f}\right) = \inf_{f \in \mathcal{F}} P\left(\ell \bullet f\right).$$

From the proof of Lemma 2.3, it results that for all $\delta > 0$ and for all functions $f, g \in \mathcal{F}(\delta)$ we have

$$P(\ell \bullet f - \ell \bullet g)^2 \leq 16b(\delta + D).$$

Therefore, for all $\delta > 0$, we have

$$\begin{aligned}\mathbb{E}\|P_n - P\|_{\mathcal{L}(\delta)} &\leq 2\mathbb{E}\sup\left\{\left|(P_n - P)(\ell \bullet f - \ell \bullet \bar{f})\right| : f \in \mathcal{F}(\delta)\right\} \\ &\leq 2\theta_n\left(16b(\delta + D)\right), \end{aligned} \tag{4.1}$$

where we have denoted

$$\theta_n(\delta) = \mathbb{E}\sup\left\{\left|(P_n - P)(\ell \bullet f - \ell \bullet \bar{f})\right| : \ f \in \mathcal{F},\ P(\ell \bullet f - \ell \bullet \bar{f})^2 \leq \delta\right\}.$$

Let $\sigma_1, \ldots, \sigma_n$ be a sequence of independent Rademacher random variables (i.e. $\mathbb{P}(\varepsilon_i = \pm 1) = 1/2$) independent from our sample. Then, according to the symmetrization inequality (see, e.g., Theorem 2.1 in Koltchinskii, 2011) we have

$$\theta_n(\delta) \leq 2\mathbb{E}\sup_{\mathbf{t}\in\bar{\mathcal{L}}(\delta)} |P_n^\sigma \mathbf{t}|, \tag{4.2}$$

with

$$\bar{\mathcal{L}}(\delta) = \left\{(\ell \bullet f - \ell \bullet \bar{f}) : \ f \in \mathcal{F},\ P(\ell \bullet f - \ell \bullet \bar{f})^2 \leq \delta\right\},$$

and

$$P_n^\sigma \mathbf{t} = \frac{1}{n}\sum_{i=1}^n \sigma_i \mathbf{t}(X_i, Y_i).$$

Since for all functions $f, g \in \mathcal{F}$,

$$P(\ell \bullet f - \ell \bullet g)^2 \leq 4b\|f - g\|^2,$$

we deduce that, for all $u > 0$ and for all $\delta > 0$,

$$\begin{aligned} H\left(u, \bar{\mathcal{L}}(\delta)\right) &= H\left(u, \left\{\ell \bullet f : \ P(\ell \bullet f - \ell \bullet \bar{f})^2 \leq \delta\right\}\right) \\ &\leq H\left(u, \{\ell \bullet f : f \in \mathcal{F}\}\right) \\ &\leq H\left(\tfrac{u}{2\sqrt{b}}, \mathcal{F}\right) \\ &\leq A\left(\tfrac{2\sqrt{b}}{u}\right)^s. \end{aligned}$$

Hence, Theorem 3.12 in Koltchinskii (2011) (or Corollary B.2.2 in Paris, 2013b) yields

$$\mathbb{E}\sup_{\mathbf{t}\in\bar{\mathcal{L}}(\delta)} |P_n^\sigma \mathbf{t}| \leq \frac{Cb^{s/4}\delta^{(2-s)/4}}{\sqrt{n}} \vee \frac{8C^2 b^{(2+s)/2}\delta^{-s/2}}{n} \vee \frac{Cb^{(4+s)/4}\delta^{-s/4}}{3e^2 n}, \tag{4.3}$$

where $C = 576\sqrt{2^s A}/(2-s)$. Combining (4.1), (4.2) and (4.3) we deduce that, for all $\delta > 0$,

$$\mathbb{E}\|P_n - P\|_{\mathcal{L}(\delta)} \leq \frac{c_1\sqrt{b}(\delta + D)^{(2-s)/4}}{\sqrt{n}} \vee \frac{c_2 b(\delta + D)^{-s/2}}{n} \vee \frac{c_3 b(\delta + D)^{-s/4}}{n},$$

where $c_1 = 4C16^{(2-s)/4}$, $c_2 = 32C^2 16^{-s/2}$ and $c_3 = 4C16^{-s/4}/3e^2$. This concludes the proof. □

References

Y. Baraud. Model selection for regression on a random design. *ESAIM: Probability and Statistics*, 6:127–146, 2002.

O. Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus de lAcadémie des Sciences de Paris*, 334:495–500, 2002.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2001.

R. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov. Approximating methods for supervised learning. *Foundations of Computational Mathematics*, 6:3–58, 2006.

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34:1143–1216, 2006.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.

M. Kohler. Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *Journal of Statistical Planning and Inference*, 89:1–23, 2000.

M. Kohler, A. Krzyżak, and H. Walk. Optimal global rates of convergence in nonparametric regression with unbounded data. *Journal of Statistical Planning and Inference*, 123:1286–1296, 2009.

V. Koltchinskii. On the central limit theorem for empirical measures. *Theory of Probability and Mathematical Statistics*, 24:71–82, 1981.

V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34:2593–2656, 2006.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008. Springer, Berlin, 2011.

Q. Paris. Minimax adaptive dimension reduction for regression. *hal-00768911*, 2013a.

Q. Paris. A contribution to complexity reduction in statistics (ph.d. dissertation). 2013b.

D. Pollard. A central limit theorem for empirical processes. *Journal of the Australian Mathematical Society*, 33:235–248, 1982.

M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

S. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, Cambridge, 1999.