Contents

Vol. 18, No. 4, 2009

ISSN 1066-5307, Mathematical Methods of Statistics, 2009, Vol. 18, No. 4, pp. 1–34. © Allerton Press, Inc., 2009.

Estimation of the Density of Regression Errors by Pointwise Model Selection

S. Plancade[1*]

[1]*Université Paris Descartes, Paris, France*

Received February 18, 2009; in final form, November 26, 2009

Abstract—This paper presents two results: a density estimator and an estimator of regression error density. We first propose a density estimator constructed by model selection, which is adaptive for the quadratic risk at a given point. Then we apply this result to estimate the error density in a homoscedastic regression framework $Y_i = b(X_i) + \epsilon_i$ from which we observe a sample (X_i, Y_i). Given an adaptive estimator $\widehat{b}$ of the regression function, we apply the density estimation procedure to the residuals $\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i)$. We get an estimator of the density of ϵ_i whose rate of convergence for the quadratic pointwise risk is the maximum of two rates: the minimax rate we would get if the errors were directly observed and the minimax rate of convergence of $\widehat{b}$ for the quadratic integrated risk.

Key words: density estimation, regression error, pointwise model selection, adaptivity.

2000 Mathematics Subject Classification: 62G07—62G08.

DOI: 10.3103/S1066530709030016

1. INTRODUCTION

Consider a sample (X_i, Y_i) from the homoscedastic regression framework:

$$Y_i = b(X_i) + \epsilon_i, \tag{1}$$

where the (ϵ_i) are unobserved independent identically distributed (i.i.d.) data with common density f, with zero mean and independent of the (X_i). The main goal of this paper is to propose an estimator for the density of ϵ_i, and to provide an upper bound for the quadratic risk of this estimator at a fixed point x_0.

The main issue in regression problems is to predict Y_i by measuring only X_i. The first step in such study is the estimation of the regression function $b(x) = \mathbb{E}[Y \mid X = x]$. This question has already been studied at length. The second step consists in studying the variations of Y_i around its conditional mean, which are characterized by the density of the errors (ϵ_i).

The knowledge of an estimator of the error density has many applications: for example, it allows model validation and, combined with an estimator of the regression function, it provides confidence intervals for future observations Y. The reader is referred to Efromovich [8] for practical applications. Many papers are devoted to density estimation but the difficulty in our problem is to estimate the density from a sample (ϵ_i) which is not observed. The natural approach consists in computing proxies of the (ϵ_i), i.e., quantities based on the data which estimate the true (ϵ_i), and applying to them a density estimation procedure as if they were the true error sample. Observing that $\epsilon_i = Y_i - b(X_i)$, we naturally estimate the errors by the residuals $(\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i))$, where $\widehat{b}$ is an estimator of the regression function. Efromovich applies this strategy with a thresholding density estimation procedure (see, for example, Efromovich [8]). He gets an estimator of the density of the (ϵ_i) whose L^2-risk reaches the same minimax rate of convergence we would obtain if the (ϵ_i) were observed. Nevertheless, this result requires strong regularity conditions on the regression function b, and on the density of the (X_i) and (ϵ_i). Another estimator is built in Plancade [17] by model selection. Its L^2-risk has a rate equal to the maximum of the minimax rates

*E-mail: sandra.plancade@math-info.univ-paris5.fr

of estimation of b and f if the sample (ϵ_i) were observed. Let us also mention the papers by Akritas and Van Keilegom [1] and Kiwitt *et al.* [9] which propose estimators of the regression errors distribution functions. But to the author's knowledge, no paper studies pointwise estimation of the error density by any method.

The estimators presented in this paper are based on a pointwise model selection procedure. Model selection theory has been initiated by Birgé and Massart (see, for example, Birgé and Massart [4]), and adapted to regression function estimation in Baraud [3] in the study of integrated quadratic risks. We will use here the estimator $\widehat{b}$ of b proposed in Baraud [3], constructed by a model selection procedure based on least square estimators. Although the principle of pointwise model selection is the same, the techniques to carry it out are different. In particular, the key tool to prove the adaptivity of classical model selection estimators is the Talagrand inequality, whereas the adaptivity of pointwise model selection estimators comes out of a simpler Bernstein inequality. The techniques developed in this paper are based on Laurent *et al.* [11], in which they develop these methods in a different framework.

This paper presents two results. On the one hand, we build a density estimator which proves to be adaptive for the pointwise risk over some classical classes of regularity. Such estimators have been constructed using kernel methods in Butucea [5], with the same adaptivity properties, along with minimax results over Sobolev classes. Nevertheless, our estimator is completely data driven, whereas the estimation procedure in Butucea [5] brings into play upper bounds on unknown quantities. The second result proceeds from the application of the above density estimation procedure to residuals from the framework (1). We get an estimator of the error density whose pointwise rate of convergence is the maximum of these two rates: the pointwise minimax rate of estimation of f we would get if the errors (ϵ_i) where observed and the L^2-minimax rate of estimation of b.

The paper is organized as follows. In Section 2, we introduce the definitions and notations, in particular, we define spaces of regularity and collections of models. Section 3 presents the density estimator and its convergence properties. This density estimation procedure is used in Section 4 to produce an estimator of the error density. Section 5 is devoted to numerical results. The proofs are gathered in Sections 6, 7 and 8. Section 6 is devoted to the results about density estimator, Section 5 contains the proof of error density Estimation Theorem, and proofs of minor results are gathered in Section 8.

2. DEFINITIONS AND NOTATION

2.1. Notation

Let t be a function defined on an interval I of $\mathbb{R}$ and μ be a density on I. We consider several norms of t:

$$\|t\|_\infty = \sup_{x\in I}|t(x)|, \qquad \|t\| = \left(\int_I t^2(x)\,dx\right)^{1/2}, \qquad \|t\|_\mu = \left(\int_I t^2(x)\mu(x)\,dx\right)^{1/2}.$$

Besides, we consider the following spaces of functions over I:

$$L^2(I) = \{t\colon I \to \mathbb{R}, \|t\| < +\infty\}, \qquad L^\infty(I) = \{t\colon I \to \mathbb{R}, \|t\|_\infty < +\infty\}.$$

Moreover, we denote by $\mathrm{Supp}(t)$ the closure of the set $\{x \in I, t(x) \neq 0\}$. If t is a function k times differentiable, we denote by $t^{(k)}$ its kth derivative.

For every set S, we denote by $\mathbf{1}_S$ the indicator function of S, that is $\mathbf{1}_S(x) = 1$ if $x \in S$ and $\mathbf{1}_S(x) = 0$ otherwise.

For every function $t\colon \mathbb{R} \to \mathbb{R}$, we denote by t^* the Fourier transform of t:

$$t^*(u) = \int_{x\in\mathbb{R}} t(x)e^{-iux}\,dx, \qquad \forall u \in \mathbb{R}.$$

For every linear space S_m we denote by t_m the L^2-orthogonal projection of t onto S_m.

Finally, for every $x \in \mathbb{R}$, we denote by $E(x)$ its integer part, that is $E(x) \in \mathbb{Z}$ and:

$$E(x) \leq x < E(x) + 1.$$

Let $I \subset J$ be two subsets of $\mathbb{R}$, we denote by $J \setminus I = \{x \in J, x \notin I\}$. Finally, we denote by $o(1)$ a quantity such that $\lim_{n\to+\infty} o(1) = 0$.

2.2. Spaces of Functions

We consider the following Sobolev classes, for every $\alpha, L > 0$:

$$W(\alpha, L) = \left\{ F \in L^2(\mathbb{R}), \frac{1}{2\pi} \int_{\mathbb{R}} |F^*(u)|^2 u^{2\alpha}\, du \le L^2 \right\}.$$

The Hölder classes are defined as follows. For every $\beta, L > 0$, and r the largest integer less than β, let

$$\mathcal{H}(\beta, L) = \left\{ F \in L^2(\mathbb{R}),\ |F^{(r)}(x) - F^{(r)}(y)| \le L|x-y|^{\beta - r},\ \forall x, y \in \mathbb{R} \right\}.$$

2.3. Collections of Models

Sine-cardinal basis: Let ϕ be the function defined on $\mathbb{R}$ by:

$$\phi(x) = \frac{\sin(\pi x)}{\pi x}, \qquad \forall x \in \mathbb{R}^*,$$

and $\phi(0) = 1$. For every $m > 0$, $k \in \mathbb{Z}$, set

$$\phi_{m,k}(x) = \sqrt{m}\phi(mx - k), \qquad \forall x \in \mathbb{R},$$

and set

$$A_m = \mathrm{vect}\{\phi_{m,k}, k \in \mathbb{Z}\}. \tag{2}$$

Let $\mathfrak{A}_n$ be the collection of models which incorporates the models (A_m) for m belonging to a grid of step $1/B$, B being a fixed positive integer:

$$\mathfrak{A}_n = \left\{ A_m, m \in \frac{1}{B}\mathbb{N}, m \le M_n \right\}$$

and $M_n \le n$. The following results hold.

Proposition 2.1. (1) *The family $\{\phi_{m,k}, k \in \mathbb{Z}\}$ is orthonormal.*

(2) *For every $m > 0$, $\big\| \sum_{k \in \mathbb{Z}} \phi_{m,k}^2 \big\|_\infty \le m$.*

(3) *For every $0 < m < m'$, $A_m \subset A_{m'}$.*

This result is proved in Section 8.

Wavelet basis: We consider also a collection of functions on $[-1, 1]$ constructed from the compact wavelet decomposition. We only recall here the definition of wavelet bases, the reader is referred to Meyer [16] for more details. Let ψ be an r times differentiable function, called mother wavelet, supported on a compact set $[-B, B]$ and satisfying the following conditions:

(1) $\psi, \ldots, \psi^{(r)}$ are bounded on $[-B, B]$;

(2) For every $0 \le k \le r$ and $\ell \ge 1$ there exists a constant C_ℓ such that $|\psi^{(k)}(x)| \le C_\ell (1 + |x|)^{-\ell}$, $\forall x \in [-B, B]$;

(3) $\int_{-B}^{B} x^k \psi(x)\, dx = 0\ \forall\, 0 \le k \le r$;

(4) The set of functions $\{\psi_{j,k} \colon x \to 2^{j/2}\psi(2^{j/2}x - k), (j,k) \in \mathbb{Z}^2\}$ is an orthonormal basis of $L^2(\mathbb{R})$.

Consider an r times differentiable function φ, called the father wavelet, supported on $[-B, B]$ and satisfying conditions (1) and (2) above, as well as the following conditions:

(3′) $\int_{-B}^{B} \varphi(x)\,dx = 1$;

(4′) The set of functions $\{\varphi_k : x \to \varphi(x-k), k \in \mathbb{Z}\} \cup \{\psi_{j,k}, j \in \mathbb{N}, k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R})$.

See Meyer [16] for examples of such functions ψ and φ. The set $\{\psi_{j,k}, j \geq 0, k \in \mathbb{Z}\} \cup \{\varphi_k, k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2[-1,1]$. As ψ is supported on $[-B,B]$, the restriction of $\psi_{j,k}$ to $[-1,1]$ is identically equal to zero for all $j \in \mathbb{N}$ and $k \notin [-2^j - B, 2^j + B]$. Let us denote $\Gamma(j) = \mathbb{Z} \cap [-2^j - B, 2^j + B]$. Similarly, φ_k is identically equal to zero for all $k \notin [-B-1, B+1] = \Gamma'(0)$. Set

$$B_m = \text{vect}\big(\{\psi_{j,k}, j = 0, \ldots, m-1, k \in \Gamma(j)\} \cup \{\varphi_k, k \in \Gamma'(0)\}\big). \tag{3}$$

It is clear that for every positive integers $m' \geq m$, $B_m \subset B_{m'}$. Now, we define

$$\mathfrak{B}_n = \{B_m, m \in \mathbb{N}^*, 2^m \leq M_n\}$$

with $M_n \leq n$. The following result holds.

Proposition 2.2. *There exists a constant K, which only depends on the father and mother wavelets ψ and φ, such that, for every $m \in \mathbb{N}^*$,*

$$\Big\| \sum_{j=0}^{m-1} \sum_{k \in \Gamma(j)} \psi_{j,k}^2 + \sum_{k \in \Gamma'(0)} \varphi_k^2 \Big\|_\infty \leq K^2 2^m. \tag{4}$$

This result is proved in Section 8.

3. DENSITY ESTIMATION BY POINTWISE MODEL SELECTION

In this section, we present a density estimator which is adaptive for the pointwise risk, over classical classes of regularity. In Section 4, this procedure will be applied to the pseudo observations $\widehat{\epsilon}_i$ of ϵ_i to get an estimator of the error density.

3.1. Framework and Assumptions

Let

$$(V_1, \ldots, V_{2n}) \tag{5}$$

be an i.i.d. sample drawn from a density g supported on $I \subset \mathbb{R}$, which satisfies:

$\mathbf{H_{dens}}$: $\sup_{x \in I} |g(x)| = \nu < +\infty$.

Let $\mathcal{M}_n = \{S_m, m = 1, \ldots, N_n\}$ be a collection of subsets of $L^2(I)$ and $\{D_m, m = 1, \ldots, N_n\}$ a collection of positive integers smaller than or equal to n such that the following assumption holds.

$\mathbf{H_{mod}}$: The collection $\mathcal{M}_n$ is nested, that is:

$$S_1 \subset S_2 \subset \cdots \subset S_{N_n}. \tag{6}$$

Thus, there exists an L^2-orthonormal basis $\{\chi_\lambda, \lambda \in I_n\}$ of S_{N_n} such that, for every model m, S_m is spanned by $\{\chi_\lambda, \lambda \in I_m\}$, where I_m is a subset of I_n. Moreover, we suppose that $D_m \leq D_{m'}$ for every $m \leq m'$.

Moreover, assume that for some positive constant K, the following condition holds:

$$\Big\| \sum_{\lambda \in I_m} \chi_\lambda^2 \Big\|_\infty \leq K^2 D_m, \qquad \forall m \in \{1, \ldots, N_n\}. \tag{7}$$

Finally, we assume that there exists a constant $M \geq 1$ such that for every $n \in \mathbb{N}$ and every $\alpha \in]0,1[$ with $n^\alpha M \leq D_{N_n}$, there exists a model m which satisfies

$$\left(\frac{n}{\log n}\right)^\alpha \leq D_m \leq M \left(\frac{n}{\log n}\right)^\alpha. \tag{8}$$

$\mathbf{H_{bias}}(\beta)$: Let $\beta > 0$. Denoting by g_m the L^2-projection of g on S_m, we assume that for some positive constant C_0,

$$\|g - g_m\|_\infty \leq C_0 D_m^{-\beta}, \qquad \forall m \in \{1, \ldots, N_n\}. \tag{9}$$

3.2. A Preliminary Risk Bound for Non-Adaptive Estimators

We split the sample (5) into two independent sequences:

$$Z_0 = (V_i)_{i\in\{1,\dots,n\}}, \qquad Z_1 = (V_i)_{i\in\{n+1\dots,2n\}}. \tag{10}$$

The sequence Z_0 is used to compute the collection $\{\widehat{g}_m, m = 1,\dots,N_n\}$ of non-adaptive estimators, and the sequence Z_1 to estimate the parameter $\nu = \|g\|_\infty$ that appears in the penalty. Let x_0 be a fixed point in I. For every model $m \in \{1,\dots,N_n\}$, the projection estimator $\widehat{g}_m$ of g on S_m, computed from the sample Z_0 is defined by

$$\widehat{g}_m = \sum_{\lambda\in I_m}\Big(\frac{1}{n}\sum_{i=1}^{n}\chi_\lambda(V_i)\Big)\chi_\lambda. \tag{11}$$

Observing that $\mathbb{E}[\widehat{g}_m(x_0)] = g_m(x_0)$ for every model m, the squared risk of the estimator $\widehat{g}_m$ at the point x_0 can be written as:

$$\mathbb{E}[(\widehat{g}_m - g)^2(x_0)] = \mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)] + (g_m - g)^2(x_0).$$

Moreover,

$$\mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)] = \mathrm{Var}\Big[\sum_{\lambda\in I_m}\Big(\frac{1}{n}\sum_{i=1}^{n}\chi_\lambda(V_i)\Big)\chi_\lambda(x_0)\Big] = \mathrm{Var}\Big[\frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{\lambda\in I_m}\chi_\lambda(V_i)\chi_\lambda(x_0)\Big)\Big].$$

The (V_i) are i.i.d, thus

$$\begin{aligned}\mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)] &= \frac{1}{n}\mathrm{Var}\Big[\sum_{\lambda\in I_m}\chi_\lambda(V_1)\chi_\lambda(x_0)\Big] \le \frac{1}{n}\mathbb{E}\Big[\Big(\sum_{\lambda\in I_m}\chi_\lambda(V_1)\chi_\lambda(x_0)\Big)^2\Big]\\ &= \frac{1}{n}\int_{x\in I}\Big(\sum_{\lambda\in I_m}\chi_\lambda(x)\chi_\lambda(x_0)\Big)^2 g(x)\,dx \le \frac{\nu}{n}\int_{x\in I}\Big(\sum_{\lambda\in I_m}\chi_\lambda(x)\chi_\lambda(x_0)\Big)^2 dx.\end{aligned}$$

By developing the square in the integral, we get

$$\mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)] \le \frac{\nu}{n}\sum_{\lambda,\lambda'\in I_m}\Big[\int_{x\in I}\chi_\lambda(x)\chi_{\lambda'}(x)\,dx\Big]\chi_\lambda(x_0)\chi_{\lambda'}(x_0).$$

Besides, the family $\{\chi_\lambda, \lambda \in I_m\}$ is orthonormal, which ensures that

$$\mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)] \le \frac{\nu}{n}\sum_{\lambda\in I_m}\chi_\lambda^2(x_0),$$

and inequality (7) yields

$$\mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)] \le K^2\nu\frac{D_m}{n}. \tag{12}$$

This bound is standard for a variance term. Finally, for every model $m \in \{1,\dots,N_n\}$ we have the following non-adaptive bound for $\widehat{g}_m$:

$$\mathbb{E}[(\widehat{g}_m - g)^2(x_0)] \le (g - g_m)^2(x_0) + K^2\nu\frac{D_m}{n}. \tag{13}$$

In Section 3.4, we will select a model by a penalized criterion, which requires to estimate the variance term $K^2\nu D_m/n$. Thus, we present an estimator $\widehat{\nu}_n$ of ν.

3.3. Estimation of ν

In this section, we propose an estimator $\widehat{\nu}_n$ of $\nu = \|g\|_\infty$ constructed from the sample Z_1. We consider a collection of models which satisfies the following properties:

$\mathbf{H}_\nu(\beta)$: Let $\mathcal{M}'_n = \{S'_m, m = 1, \ldots, N'_n\}$ be a collection of models. We suppose that for every model m, $\{\xi_\lambda, \lambda \in I'_m\}$ is an L^2-orthonormal basis of S'_m and the (ξ_λ) are continuous on I. Moreover, letting $g_m^{(1)} = \arg\min_{t \in S'_m} \|g - t\|^2$, we assume that $\|g - g_m^{(1)}\|_\infty \leq C_0 {D'_m}^{-\beta}$ for some positive integers $(D'_m)_{m=1,\ldots,N'_n}$.

Let m_0 be a model such that $p_0 = D_{m_0}$ satisfies $\left(\frac{n}{\log n}\right)^\gamma \leq p_0 \leq M\left(\frac{n}{\log n}\right)^\gamma$ for some $\gamma \in]0, 1/2[$. We define:

$$\widehat{g}_m^{(1)} = \sum_{\lambda \in I'_m} \left(\frac{1}{n} \sum_{i=n+1}^{2n} \xi_\lambda(V_i)\right)\xi_\lambda \qquad \text{and} \qquad \widehat{\nu}_n = \|\widehat{g}_{m_0}^{(1)}\|_\infty.$$

Proposition 3.1. *Suppose that Assumptions H_{dens} and $H_\nu(\beta)$ hold. Then for every n such that*

$$(\mathbf{A_1}) \quad C_0 p_0^{-\beta} < \frac{\nu}{6}$$

we have

$$P\Big[\widehat{\nu}_n \leq \frac{1}{2}\nu\Big] \leq 2\exp\Big(-\frac{n\nu}{84K^2 p_0}\Big). \tag{14}$$

If in addition:

$$(\mathbf{A_2}) \quad \frac{p_0}{\sqrt{n}} \leq \frac{\nu}{12K^2},$$

then

$$P[\widehat{\nu}_n \geq 2\nu] \leq \exp\Big(-\frac{n\nu}{456K^2 p_0}\Big). \tag{15}$$

This result is proved in Section 6.2.

Comment 1.

(1) There exists an integer N depending on (K, β, C_0) such that for every $n \geq N$, (A_1) and (A_2) hold.

(2) The collections of models in which $\widehat{\nu}_n$ and $\widehat{g}_{\widehat{m}}$ are computed can be different.

3.4. Construction of the Adaptive Estimator

The model selection procedure developed by Birgé and Massart relies on the following idea: the "best" model among the collection $\mathcal{M}_n$ is the one which minimizes the bias-variance sum in the right hand-side of (13), thus the natural idea consists in building an estimator of this sum and selecting the model $\widehat{m}$ which minimizes it.

On the one hand, the variance term $K^2\nu D_m/n$ is estimated by $K^2\widehat{\nu}_n D_m/n$.

On the other hand, the estimation of the bias term $(g - g_m)^2(x_0)$ is the main distinct point between pointwise and global model selection procedures. In classical L^2-model selection, the bias term $\|g - g_m\|^2$ is estimated, up to a quantity independent of m, by $-\|\widehat{g}_m\|^2$ (see Massart [15]), but this procedure cannot be carried over to the pointwise bias.

We note that, as j tends to infinity, the model S_j grows and g_j tends to g. Therefore, instead of estimating $(g-g_m)^2(x_0)$, we estimate the term $\sup_{j,m\le j\le N_n}(g_j-g_m)^2(x_0)$, which has same order. This heuristic is confirmed as follows. By (9) in Assumption $H_{\text{bias}}(\beta)$,

$$\begin{aligned}\sup_{j,m\le j\le N_n}(g_j-g_m)^2(x_0) &\le 2\sup_{j,m\le j\le N_n}[(g_j-g)^2(x_0)+(g_m-g)^2(x_0)]\\ &\le 2\sup_{j,m\le j\le N_n}[C_0D_j^{-2\beta}+C_0D_m^{-2\beta}]\le 4C_0D_m^{-2\beta}\end{aligned}$$

and $(g-g_m)^2(x_0)$ has order $D_m^{-2\beta}$ as well.

Now, let the best theoretical model m_{opt} be defined by

$$\begin{aligned}m_{opt} &= \arg\min_{m\in\{1,\dots,N_n\}}\Big[\sup_{j,m<j\le N_n}(g_j(x_0)-g_m(x_0))^2+\text{pen}(m)\Big]\\ &= \arg\min_{m\in\{1,\dots,N_n\}}[\text{Crit}(m)],\end{aligned} \tag{16}$$

where $\text{pen}(m)=AK^2x_m\widehat{\nu}_n\frac{D_m}{n}$, A is a constant greater than or equal to 1, and

$$x_m := \frac{45}{2}\log(1+D_m)\max\left\{1,\frac{9K^2}{\widehat{\nu}_n}\log(1+D_m)\frac{D_m}{n}\right\}.$$

Remark about the numerical constant in x_m: If the constant 45/2 is replaced by any constant $B>8$, Theorem 3.1 would still hold, but with different constants (θ_i) (see Section 3.5). Moreover, the condition $B>8$ appears in theoretical upper bounds, but in numerical simulations (see Section 5) the value $B=5$ seems to perform well. Nevertheless, the empirical calibration of this constant, as well as the constant in the penalty below, involves a lot of simulation experiments. This is a general problem in model selection and it is not specific to pointwise model selection.

In view to estimate $\text{Crit}(m)$, the natural idea would be to replace $(g_j-g_m)^2(x_0)$ by $(\widehat{g}_j-\widehat{g}_m)^2(x_0)$, but this proceeding is clearly biased. In fact,

$$\mathbb{E}[(\widehat{g}_m-\widehat{g}_j)^2(x_0)]=(g_j-g_m)^2(x_0)+\mathbb{E}\big[\big((\widehat{g}_j-\widehat{g}_m)(x_0)-(g_j-g_m)(x_0)\big)^2\big].$$

The term $\mathbb{E}\big[\big((\widehat{g}_j-\widehat{g}_m)(x_0)-(g_j-g_m)(x_0)\big)^2\big]$ is upper bounded similarly to inequality (12). More precisely,

$$\begin{aligned}&\mathbb{E}\big[\big((\widehat{g}_j-\widehat{g}_m)(x_0)-(g_j-g_m)(x_0)\big)^2\big]=\text{Var}[(\widehat{g}_j-\widehat{g}_m)(x_0)]\\ &=\text{Var}\Big[\sum_{\lambda\in I_j\setminus I_m}\Big(\frac{1}{n}\sum_{i=1}^n\chi_\lambda(V_i)\Big)\chi_\lambda(x_0)\Big]=\frac{1}{n}\text{Var}\Big[\sum_{\lambda\in I_j\setminus I_m}\chi_\lambda(V_1)\chi_\lambda(x_0)\Big]\\ &\le\frac{\nu}{n}\int_{x\in I}\Big(\sum_{\lambda\in I_j\setminus I_m}\chi_\lambda(x)\chi_\lambda(x_0)\Big)^2dx=\frac{\nu}{n}\sum_{\lambda\in I_j\setminus I_m}\chi_\lambda^2(x_0)\le\frac{\nu}{n}\sum_{\lambda\in I_j}\chi_\lambda^2(x_0)\le\nu K^2\frac{D_j}{n}.\end{aligned}$$

Now, the theoretical criterion $\text{Crit}(m)$ is estimated by

$$\widehat{\text{Crit}}(m)=\sup_{j,m<j\le N_n}\Big[(\widehat{g}_j-\widehat{g}_m)^2(x_0)-K^2\widehat{\nu}_nx_j\frac{D_j}{n}\Big]_+ +\text{pen}(m) \tag{17}$$

and $\widehat{m}=\arg\min_{m\in\{1,\dots,N_n\}}\widehat{\text{Crit}}(m)$.

Our estimator of g is $\widehat{g}_{\widehat{m}}$.

3.5. Results

We can prove the following result about the risk of $\widehat{g}_{\widehat{m}}$ at x_0.

Theorem 3.1. *Suppose that Assumptions $H_{\mathrm{bias}}(\beta)$, $H_\nu(\beta)$, and H_{mod} hold with the constraint $M(n/\log n)^{1/(2\beta+1)} \leq N_n$. Suppose that (A_1), (A_2) and the following condition hold*:

$$(\mathbf{A_3}): \qquad 1 + Mn\log n \leq n \qquad \text{and} \qquad \left(\frac{n}{\log n}\right)^{2\beta/(2\beta+1)} \geq \frac{18MK^2}{\nu}. \tag{18}$$

Then,

$$\mathbb{E}[(\widehat{g}_{\widehat{m}} - g)^2(x_0)] \leq \theta_1 \left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}} + \mathcal{R}_n,$$

where

$$\mathcal{R}_n = \frac{\theta_2}{n} + (\nu + K^2 \max_{m=1,\dots,N_n} D_m)^2 \exp\Big[-\frac{n\nu}{84K^2 p_0}\Big] + \theta_3 p_0^2 \exp\Big[-\frac{n\nu}{456K^2 p_0}\Big]$$

and

$$\theta_1 = \max\Big\{15, 4(3 + \frac{2}{45\log(1+D_1)})\Big\}(2C_0^2 + 45AK^2\nu) + 4C_0^2,$$

$$\theta_2 = 20K^2(\nu + 16K^2)\bigg(\sum_{m=1}^{N_n}(1+D_m)^{1+1/4}\bigg),$$

$$\theta_3 = \frac{45A}{2}(M+1)K^4\nu \max\Big\{15, 4\Big(3 + \frac{2}{45\log(1+D_1)}\Big)\Big\}.$$

Comment 2. Clearly, $\mathcal{R}_n$ is negligible with respect to the rate $(n/\log n)^{-2\beta/(2\beta+1)}$.

Assumption $H_{\mathrm{bias}}(\beta)$ couples the collection of models and the fact that g belongs to a certain space of regularity (through the exponent β). The following Proposition gives examples for which this assumption is satisfied.

Proposition 3.2. (1) *Let (β, L) be two positive numbers, let A_m be the linear subset of $L^2(\mathbb{R})$ defined in* (2)*, and let $h_m = \arg\min_{t\in A_m} \|h - t\|$, for every $h \in L^2(\mathbb{R})$, then there exists a constant $K(\beta)$ such that*

$$\|h - h_m\|_\infty \leq K(\beta)Lm^{-\beta}, \qquad \forall h \in W(\beta + 1/2, L).$$

(2) *Let (β, L) be two positive numbers, and r be an integer greater than β, let B_m be the linear subset of $L^2([-1,1])$ defined in* (3) *and let $h_m = \arg\min_{t\in B_m} \|h - t\|$ for every $h \in L^2([-1,1])$, then there exists a constant $K'(\beta)$ such that*

$$\|h - h_m\|_\infty \leq K'(\beta)L(2^m)^{-\beta}, \qquad \forall h \in \mathcal{H}(\beta, L).$$

This Proposition is proved in Section 8. Moreover, by Propositions 2.1 and 2.2, the collections $\mathfrak{A}_n$ and $\mathfrak{B}_n$ satisfy Assumption H_{mod} for $M = 2$.

Comment 3. It is well known that the minimax rate of convergence for pointwise density estimation over $W(\beta + 1/2, L)$ or $\mathcal{H}(\beta, L)$ is $n^{-2\beta/(2\beta+1)}$ (see, e.g., Tsybakov [19] for Hölder classes and Butucea [5] for Sobolev spaces). Our estimator reaches this rate up to a logarithmic factor. Nevertheless, Lepski [12] defines the adaptive minimax rate, which is the best rate of convergence for adaptive estimators over a range of regularity classes, and proves that the logarithmic loss is unavoidable in adaptive estimation, in several frameworks. Following this line, Butucea [5] proves that the adaptive minimax rate over the classes $\{W(\beta, L), \beta > 0\}$ for pointwise density estimation is $(n/\log n)^{-2\beta/(2\beta+1)}$. Hence if we consider the collection of models $\mathfrak{A}_n$, $\widehat{g}_{\widehat{m}}$ is adaptive minimax over Sobolev classes. Similar results are proved over Hölder classes, for example, in a white noise model (see Lepski and Spokoiny [13]), so we expect that the adaptive minimax rate in pointwise density estimation has the same order. Then if we consider the collection $\mathfrak{B}_n$, our estimator should be adaptive minimax over Hölder classes.

3.6. Comparison with Lepski's Method

The reference method in pointwise estimation is the one originally presented by Lepski [12] and developed in many others papers. In particular, it was adapted to density estimation by Butucea [5]. This procedure provides adaptive rates of convergence and even exact adaptive results (see Butucea [5]). This means that the estimator gets the adaptive rate of convergence and also the best asymptotic constant on given classes of functions. Lepski estimators have better asymptotic properties than the estimator presented in this paper, but the theoretical results remain asymptotic, whereas the results presented here are non-asymptotic. One can object than the large constants which appear in the term $\mathcal{R}_n$ in Theorem 3.1 require large-size samples, but these constants are much larger than the effective ones, as proved by simulations.

In more recent works, Lepski and Goldenshluger [14] prove oracle inequalities in the Gaussian white noise framework, but as far as the author knows, these results have not been developed in density estimation framework.

4. ERROR DENSITY ESTIMATION

4.1. Framework, Outline and Preliminary Results

We consider a $3n$-sample

$$(X_i, Y_i)_{i\in\{-n,\dots,-1\}\cup\{1,\dots,2n\}} \tag{19}$$

from the regression framework (1), where the (X_i) are i.i.d. with density f_X supported on $[0,1]$, the (ϵ_i) are i.i.d. independent of the (X_i) and $\mathbb{E}(\epsilon_1) = 0$. This section presents an estimation procedure of the density f of the (ϵ_i). Let us outline this procedure, which can be decomposed in three steps.

Step 1: From the sequence

$$Z^- = (X_i, Y_i)_{i\in\{-n,\dots,-1\}}, \tag{20}$$

we compute an estimator $\widehat{b}$ of the regression function b.

In Section 4.4, we recall an example of adaptive estimation procedure of the regression function, but the result that we establish in Theorem 4.1 holds for any estimator $\widehat{b}$ of b computed from the sequence Z^-.

Step 2: We compute the residuals of the sequence $(X_i, Y_i)_{\{1,\dots,2n\}}$, namely

$$\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i), \qquad \forall i \in \{1,\dots,2n\}.$$

Noting that $\epsilon_i = Y_i - b(X_i)$, the $\widehat{\epsilon}_i$ are natural proxies for the unobserved (ϵ_i). Given Z^-, the $(\widehat{\epsilon}_i)$ are i.i.d. Let us denote by f^- their common density, which only depends on the sequence $(X_i, Y_i)_{i\in\{-n,\dots,-1\}}$. For every integrable function $t\colon \mathbb{R}\to\mathbb{R}$

$$\mathbb{E}[t(\widehat{\epsilon}_1)\,|\,Z^-] = \mathbb{E}[t((b-\widehat{b})(X_1)+\epsilon_1)\,|\,Z^-] = \int_{x=0}^{1}\int_{y\in\mathbb{R}} t((b-\widehat{b})(x)+y) f_X(x) f(y)\,dy\,dx$$

$$= \int_{x=0}^{1}\int_{z\in\mathbb{R}} t(z) f_X(x) f(z-(b-\widehat{b})(x))\,dz\,dx = \int_{z\in\mathbb{R}} t(z)\left[\int_{x=0}^{1} f(z-(b-\widehat{b})(x)) f_X(x)\,dx\right] dz.$$

Hence,

$$f^-(z) = \int_{x=0}^{1} f\big(z-(b-\widehat{b})(x)\big) f_X(x)\,dx, \qquad \forall z\in\mathbb{R}. \tag{21}$$

Step 3: We apply the density estimation procedure described in Section 3 to the residuals $(\widehat{\epsilon}_i)$.

Thus, the risk of the estimator of f results from two consecutive approximations of different nature: the first one consists in replacing the true (ϵ_i) by the residuals, and the second one is a density estimation error.

These two approximations appear in the following inequality:

$$\mathbb{E}[(\widehat{f_{\widehat{m}}^-} - f)^2(x_0)] \leq 2\Big\{\mathbb{E}[(\widehat{f_{\widehat{m}}^-} - f^-)^2(x_0)] + \mathbb{E}[(f^- - f)^2(x_0)]\Big\}. \quad (22)$$

We suppose that the error density f satisfies the following assumption.

$\mathbf{H_{error}}$: f is Lipschitz with constant $\mathrm{Lip}(f)$, that is:

$$|f(x) - f(y)| \leq \mathrm{Lip}(f)|x - y|, \qquad \forall x, y \in I.$$

Besides, $\sup_{x\in I} |f(x)| = \nu < +\infty$.

We consider a collection of models $\mathcal{M}_n$, which satisfies Assumption H_{mod} and such that one of the following two alternative assumptions holds:

$\mathbf{H^{(1)}_{bias-error}}(\beta)$: $f \in \mathcal{H}(\beta, L)$ and there exists a constant $C_0(\beta, L)$ such that, for every model $S_m \in \mathcal{M}_n$,

$$\|h - h_m\|_\infty \leq C_0(\beta, L)D_m^{-\beta}, \qquad \forall h \in \mathcal{H}(\beta, L).$$

$\mathbf{H^{(2)}_{bias-error}}(\beta)$: $f \in W(\beta + 1/2, L)$ and there exists a constant $C_0(\beta, L)$ such that, for every model $S_m \in \mathcal{M}_n$,

$$\|h - h_m\|_\infty \leq C_0(\beta, L)D_m^{-\beta}, \qquad \forall h \in W(\beta + 1/2, L).$$

Remark 1. According to Proposition 3.2, Assumption $H^{(1)}_{\mathrm{bias-error}}(\beta)$ is satisfied if $f \in \mathcal{H}(\beta, L)$ and the collection $\mathcal{M}_n$ that we consider is the wavelet collection $\mathfrak{B}_n$, and Assumption $H^{(2)}_{\mathrm{bias-error}}(\beta)$ is satisfied if $f \in W(\beta + 1/2, L)$ and $\mathcal{M}_n$ is the sine-cardinal collection $\mathfrak{A}_n$.

Remark 2. Assumptions $H^{(1)}_{\mathrm{bias-error}}(\beta)$ and $H^{(2)}_{\mathrm{bias-error}}(\beta)$ are less general than Assumption H_{bias} in the density estimation Theorem. In fact, in order to apply the result of Section 3, we need the density f^- of the residuals to satisfy the Assumptions of Theorem 3.1, which is guaranteed under $H^{(1)}_{\mathrm{bias-error}}(\beta)$ or $H^{(2)}_{\mathrm{bias-error}}(\beta)$. This fact comes out of the following proposition.

Proposition 4.1. (1) *For every $x \in \mathbb{R}$, $|f^-(x)| \leq \nu$ a.s.*

(2) *For every β, L positive, $f \in \mathcal{H}(\beta, L) \;\Rightarrow\; f^- \in \mathcal{H}(\beta, L)$ a.s.*

(3) *For every β, L positive, $f \in W(\beta + 1/2, L) \;\Rightarrow\; f^- \in W(\beta + 1/2, L)$ a.s.*

Proposition 4.1 is proved in Section 8.

We consider another collection $\mathcal{M}'_n = \{S'_m, m = 1, \ldots, N_n\}$ (which can be equal to or different from $\mathcal{M}_n$) and for every m, $S'_m = \mathrm{vect}\{\xi_\lambda, \lambda \in I'_m\}$ and the (ξ_λ) are continuous on I. For every $h \in L^2(I)$, let $h^{(1)}_m = \arg\min_{t\in S'_m} \|h - t\|^2$. We suppose that one of the following two alternative assumptions holds:

$\mathbf{H^{(1)}_{\nu-error}}(\beta)$: $f \in \mathcal{H}(\beta, L)$ and for every model $S'_m \in \mathcal{M}_n$,

$$\|h - h^{(1)}_m\|_\infty \leq C_0(\beta, L)D_m^{-\beta}, \qquad \forall h \in \mathcal{H}(\beta, L).$$

$\mathbf{H^{(2)}_{\nu-error}}(\beta)$: $f \in W(\beta + 1/2, L)$ and for every model $S'_m \in \mathcal{M}_n$,

$$\|h - h^{(1)}_m\|_\infty \leq C_0(\beta, L)D_m^{-\beta}, \qquad \forall h \in W(\beta + 1/2, L).$$

4.2. Definition of the Estimator

Let us consider the $3n$-sample (19). Let $\widehat{b}$ be any estimator of b constructed from the sequence Z^- (see 20). Set

$$\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i), \qquad \forall i = 1, \dots, 2n.$$

Let $\mathcal{M}_n = \{S_m, m = 1, \dots, N_n\}$ be a collection of subsets of $L^2(I)$, $\{D_m, m = 1, \dots, N_n\}$ a collection of positive integers, and β a positive number such that Assumptions H_{mod}, and $H^{(1)}_{\mathrm{bias-error}}(\beta)$ or $H^{(2)}_{\mathrm{bias-error}}(\beta)$ hold.

For every model $S_m = \mathrm{vect}\{\chi_\lambda, \lambda \in I_m\}$, let

$$\widehat{f_m^-} = \sum_{\lambda \in I_m} \left(\frac{1}{n} \sum_{i=1}^{n} \chi_\lambda(\widehat{\epsilon}_i) \right) \chi_\lambda. \tag{23}$$

Let $\mathcal{M}'_n = \{S'_m, m = 1, \dots, N'_n\}$ be a collection of subsets of $L^2(I)$, $\{D'_m, m = 1, \dots, N_n\}$ a collection of positive integers such that Assumption $H^{(1)}_{\nu-\mathrm{error}}(\beta)$ or $H^{(2)}_{\nu-\mathrm{error}}(\beta)$ holds. Let m_0 be in $\{1, \dots, N'_n\}$ such that $p_0 = D_{m_0}$ satisfies

$$\left(\frac{n}{\log n}\right)^{\gamma} < p_0 < M \left(\frac{n}{\log n}\right)^{\gamma}$$

for some $\gamma \in]0, 1/2[$, and

$$\widehat{\nu_n^-} = \|(\widehat{f_{m_0}^-})^{(1)}\|_\infty, \qquad \text{where} \quad (\widehat{f_{m_0}^-})^{(1)} = \sum_{\lambda \in I'_m} \Big(\frac{1}{n} \sum_{i=n+1}^{2n} \xi_\lambda(\widehat{\epsilon}_i) \Big) \xi_\lambda.$$

Finally, let

$$\widehat{m} = \arg \min_{m=1,\dots,N_n} \Big\{ \Big[\sup_{j, m \le j \le N_n} (\widehat{f_j^-} - \widehat{f_m^-})^2(x_0) - K^2 x_j \widehat{\nu_n^-} \frac{D_j}{n} \Big]_+ + \mathrm{pen}^-(m) \Big\},$$

where $\mathrm{pen}^-(m) = A K^2 x_m^- \widehat{\nu_n^-} \frac{D_m}{n}$ with

$$x_m^- := \frac{45}{2} \log(1 + D_m) \max \left\{ 1, \frac{9K^2}{\widehat{\nu_n^-}} \log(1 + D_m) \frac{D_m}{n} \right\}.$$

4.3. Result

For the estimator $\widehat{f_{\widehat{m}}^-}$ the following result holds.

Theorem 4.1. *Suppose that Assumptions $H^{(i)}_{\mathrm{bias-error}}(\beta)$ and $H^{(i)}_{\nu-\mathrm{error}}(\beta)$ hold for $i = 1$ or 2 and for some $\beta \ge \beta' > 3/4$, where β' is known. Suppose that Assumption H_{mod} holds with*

$$\left(\frac{n}{\log n}\right)^{1/(2\beta'+1)} \le D_{N_n} \le M \left(\frac{n}{\log n}\right)^{1/(2\beta'+1)}.$$

Consider

$$\gamma \in \left] \frac{1}{\beta'(2\beta'+1)}, \min \left\{ \frac{1}{\beta'+1}, \frac{4\beta'+1}{3(2\beta'+1)} \right\} \right[.$$

Then, for every n such that $1 + Mn/\log n \le n$, we have

$$\mathbb{E}\big[(\widehat{f_{\widehat{m}}^-} - f)^2(x_0)\big] \le \theta'_1 \left(\frac{n}{\log n}\right)^{-2\beta/(2\beta+1)} + \mathcal{C}_n \mathbb{E}\big[\|\widehat{b} - b\|^2_{f_X}\big] + \mathcal{R}_n, \tag{24}$$

where

$$\theta_1' = \Big[(2C_0^2 + 45A\nu K^2)\max\Big(15, 3 + \frac{2}{45\log(1+D_1)}\Big) + C_0^2\Big](M+1),$$

$$\mathcal{C}_n = \mathrm{Lip}(f)^2 + 2\log n\Big[\Big(\nu\Big(\frac{n}{\log n}\Big)^{-1/(2\beta'+1)} + K^2M\Big)^2$$

$$\times\Big(36C_0^2\Big(\frac{n}{\log n}\Big)^{2/(2\beta'+1)-2\beta'\gamma} + (18MK^2)^2\Big(\frac{n}{\log n}\Big)^{(2-4\beta')/(2\beta'+1)}\Big)\Big],$$

$$\mathcal{R}_n = 2\Big(\nu + K^2M\Big(\frac{n}{\log n}\Big)^{1/(2\beta'+1)}\Big)^2\exp\Big(-\frac{\sqrt{n}}{7}\Big) + \frac{\theta_2}{n},$$

and θ_2 is defined in Theorem 3.1.

Remark 3. We have $\mathcal{C}_n = \mathrm{Lip}(f)^2 + o(1)$ and $\mathcal{R}_n \le \kappa_1'/n$, where κ_1' depends on (ν, M, K, β').

Comment 4. By (24), the rate of convergence of our estimator is upper bounded by the maximum of the two following rates:

- the rate of convergence of the estimator $\widehat{b}$ of b.
- the minimax rate of estimation we would obtain for f if the (ϵ_i) were directly observed, that is $(n/\log n)^{-2\beta/(2\beta+1)}$.

According to Comment 3 in Section 3.5, the rate of convergence of $\widehat{f_{\widehat{m}}^-}$ is clearly lower bounded by $(n/\log n)^{-2\beta/(2\beta+1)}$. On the other hand, the term $\mathbb{E}[\|\widehat{b} - b\|_{f_X}^2]$ seems to be avoidable. In an integrated risk context, Efromovich [8] proposes an error density estimator whose rate of convergence does not depend on the risk of $\widehat{b}$. Nevertheless, stronger conditions are required. In particular, the densities of X_i and ϵ_i are supposed to be two times differentiable and the errors (ϵ_i) are supposed to be symmetrical. The convergence results in Efromovich [8] are based on properties of the trigonometric basis and are not easily transposable in a pointwise context.

Besides, in numerical examples, our error density estimator performs nearly as well as the estimator we would obtain if the (ϵ_i) were observed (see Fig. 3, Section 5).

4.4. An Estimator of b

In this section, we briefly exhibit an estimator $\widehat{b}$ of b which suits to our setting. This is the estimator which is implemented in the simulations. The regression function estimator presented here results from Baraud's works (see Baraud [3] and Baraud [2]), gathered in Plancade [17]. Consider the following conditions.

$\mathbf{H_b}$: The density f_X of X_1 is supported on a compact J, and is lower bounded by $m_0 > 0$ and upper bounded by $m_1 < +\infty$.

Let us consider a collection of finite-dimensional models Σ_n which satisfies the following assumption.

$\mathbf{H_{mod-b}}$: Σ_n is included in a global model S_n with dimension smaller than $n^{1/2-d}$ for some $d > 0$. Furthermore, there exists some nonnegative constants Γ and R such that for every integer n,

$$|\{m \in \Sigma_n \colon D_m = D\}| \le \Gamma D^R$$

for every $D \in \mathbb{N}^*$. Finally, there exists a constant K such that

$$\|t\|_\infty \le K\sqrt{N_n}\|t\|, \qquad \forall t \in S_n.$$

For every model $m \in \Sigma_n$, let $\widehat{b}_m$ be the least squares estimator of b:

$$\widehat{b}_m = \arg\min_{t \in S_m} \gamma_n(t), \qquad \text{where} \quad \gamma_n(t) = \frac{1}{n} \sum_{i=-n}^{-1} (Y_i - t(X_i))^2,$$

and the selected model is $\widehat{m} = \arg\min_{m \in \Sigma_n} \left[\gamma_n(\widehat{b}_m) + A' \widehat{\sigma}_n^2 \frac{D_m}{n}\right]$, where $A' > 1$ and $\widehat{\sigma}_n^2$ is an estimator of the variance of ϵ_1: let V_n be a space of dimension $E(n/2)$ which includes the global model S_n, then:

$$\widehat{\sigma}_n^2 = \frac{1}{n - E(n/2)} \inf_{t \in S_n} (Y_i - t(X_i))^2.$$

Let us define $\widehat{b} = \widehat{b}_{\widehat{m}}$ if $\|\widehat{b}_{\widehat{m}}\| \leq n$ and $\widehat{b} = 0$ otherwise.

Proposition 4.2. *Under Assumptions H_{b} and $H_{\mathrm{mod-b}}$,*

$$\mathbb{E}[\|b - \widehat{b}\|_{f_X}^2] \leq C \inf_{m \in \Sigma_n} \left(\|b - b_m\|^2 + \sigma^2 \frac{D_m}{n}\right) + \frac{C'}{n}$$

for some constants C depending on A' and m_1 and C' depending on $(\sigma, \mathbb{E}[\epsilon_1^4], m_0, m_1)$.

Finally, classical results about approximation theory in Besov spaces lead to the following statement: if b belongs to the Besov space $\mathcal{B}_2^{\alpha,\infty}$ (for a definition of Besov space, see DeVore and Lorentz [7]), then $\mathbb{E}[\|\widehat{b} - b\|_{f_X}^2] \leq Cn^{-2\alpha/(2\alpha+1)}$. This entails the following Corollary:

Corollary 4.1. *Suppose that the assumptions of Theorem* 4.1 *hold, as well as Assumptions H_{b} and $H_{\mathrm{mod-b}}$. Then, if b belongs to the Besov space $\mathcal{B}_p^{\alpha,\infty}$ for some $p > 0$ and $\alpha \geq \beta > 1/2$,*

$$\mathbb{E}\big[(\widehat{f}_{\widehat{m}}^{-} - f)^2(x_0)\big] \leq \theta \Big(\frac{n}{\log n}\Big)^{-\frac{2\beta}{2\beta+1}}$$

for some constant θ independent of n.

In other words, if b is smoother than f, the rate of convergence of $\widehat{f}_{\widehat{m}}^{-}$ is the optimal rate we would get if the (ϵ_i) were directly observed. We do not provide a detailed proof of Corollary 4.1, and the reader is referred to the remark at the end of Theorem 5.1 in Plancade [17].

5. SIMULATIONS

5.1. Density Estimation

This section illustrates the density estimation procedure presented in Section 3 with the sine-cardinal collection of models $\mathcal{A}_n$ described in (2). We choose $B = 10$ and $M_n = \sqrt{n}$. We draw 50 samples $(V_1, \ldots, V_n)$ of size $n = 200, 500, 2000$ of i.i.d. variables with Gaussian distribution (denoted by $\mathcal{N}(0,1)$) and with Laplace density $g(x) = \frac{1}{2}\exp(-|x|)$ (denoted by $\mathcal{L}(1)$). Let J be the set of 150 regularly spaced points in $[-5, 5]$. For each sample and for every point $x \in J$ we compute an estimator $\widehat{g}_{\widehat{m}}(x)$ as follows, assuming that the maximum of the density ν is known:

- First we compute the projection density estimators $(\widehat{g}_m(x))$ for every $m \in \frac{1}{10}\mathbb{N}$, $m \leq M_n$, and every $x \in J$ (see (11)).
- Then for every $x \in J$, we select the best model as:

$$\widehat{m} = \arg\min \left\{ \sup_{m \leq j \leq N_n} \left[(\widehat{g}_j - \widehat{g}_m)^2(x) - \alpha\nu \log(1 + j)\frac{j}{n}\right]_+ + \beta\nu\log(1+m)\frac{m}{n} \right\}$$

  with $\alpha = 5$ and $\beta = 10$.
- We plot the set of points $\{(x, \widehat{g}_{\widehat{m}}(x)), x \in J\}$.

$V_i \sim \mathcal{N}(0,1)$

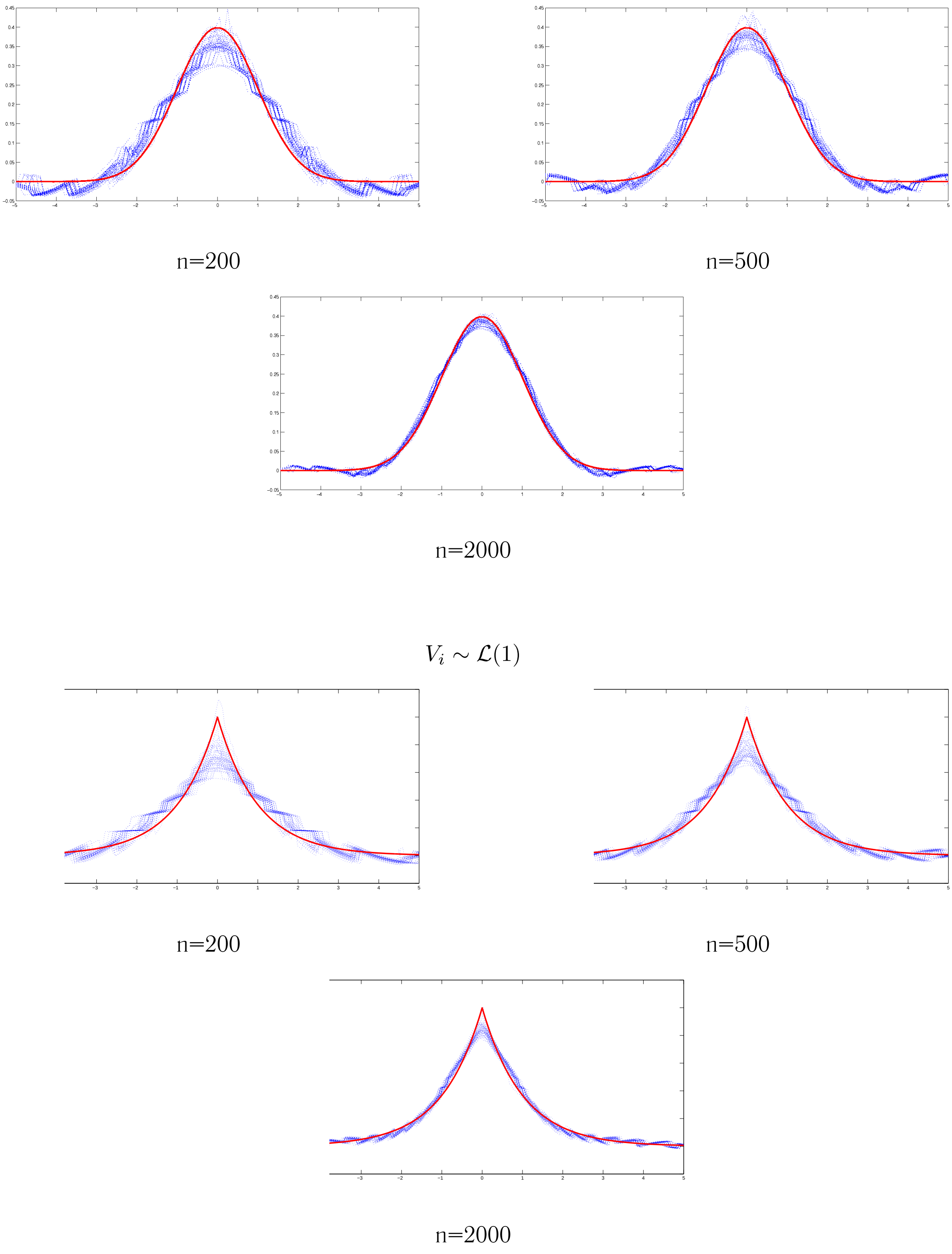
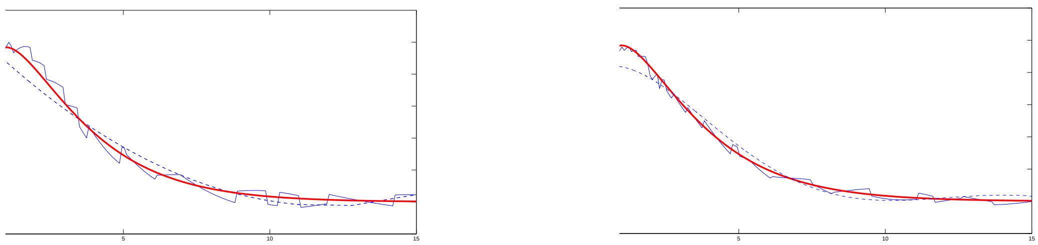


Fig. 1. Beam of 50 density estimators curves (dotted lines) built from i.i.d. samples of size $n = 200$, 500, and 2000 of densities $\mathcal{N}(0,1)$ and $\mathcal{L}(1)$ (thick lines), in sine-cardinal bases.

In Figure 1, each graph presents 50 estimated curves of $\widehat{g}_{\widehat{m}}$ for a given density g_i and a given n.

Figure 2 presents a comparison between our pointwise model selection estimator and a global model selection estimator computed following the procedure developed by Massart [15], Section 7, for sample of size $n = 500, 2000$ with common density $\chi^2(3)$. The global model selection estimator (dotted line) is computed in a mixed piecewise polynomial and trigonometric polynomial basis using matlab programs available on Yves Rozenholc's web page (http://www.math-info.univ-paris5.fr/ rozen/). The pointwise model selection estimator (solid line) is built following the procedure described above, on the set J of 150 regularly spaced points on $[-1, 15]$. We observe that the pointwise model selection estimator (solid line) fits the true density (thick line) for a smaller sample size better than the global model selection estimator.



n=500 n=2000

Fig. 2. Pointwise model selection estimator (solid line) and global model selection estimator (dotted line) for a sample of size $n = 500, 2000$ of density $\chi^2(3)$ (thick line).

5.2. Error Density Estimation

This section proposes illustrations of the error density estimator described in Section 4, with the following procedure:

- We draw a sample $(X_1, \ldots, X_{2n})$ with common density f_X uniform on $[0, 1]$ and $\chi^2(3)$. We draw also a sample $(\epsilon_1, \ldots, \epsilon_{2n})$ with common density f from a distribution $\mathcal{N}(0, 1)$ and $\mathcal{L}(1)$. We choose a regression function $b(x) = x^3 + 5x$ and $b(x) = \exp(-|x|)$ and compute the sample $(Y_1, \ldots, Y_{2n})$, where $Y_i = b(X_i) + \epsilon_i$.
- From the sample $\{(X_i, Y_i)\}_{i=1\ldots n}$, we compute an estimator $\widehat{b}$ of b following the procedure described in Section 4, using mixed piecewise polynomial and trigonometric polynomial basis (see Comte *et al.* [6]).
- We compute the residuals from the second sample $(\widehat{\epsilon}_i)_{i=n+1,\ldots,2n}$, where $\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i)$.
- Let J be a set of 150 regularly spaced points on $[-5, 5]$. We apply the density estimation procedure described in Section 5.1 to the residuals $(\widehat{\epsilon}_i)_{i=n+1,\ldots,2n}$.

Figure 3 presents the error density estimator (dotted line) and the theoretical estimator we obtain by applying the density estimation procedure of Section 5.1 directly to the sample $(\epsilon_i)_{i=n+1,\ldots,2n}$. The thick line is the true density of ϵ_1.

We have also checked that the error density estimator hardly depends on the design's distribution.

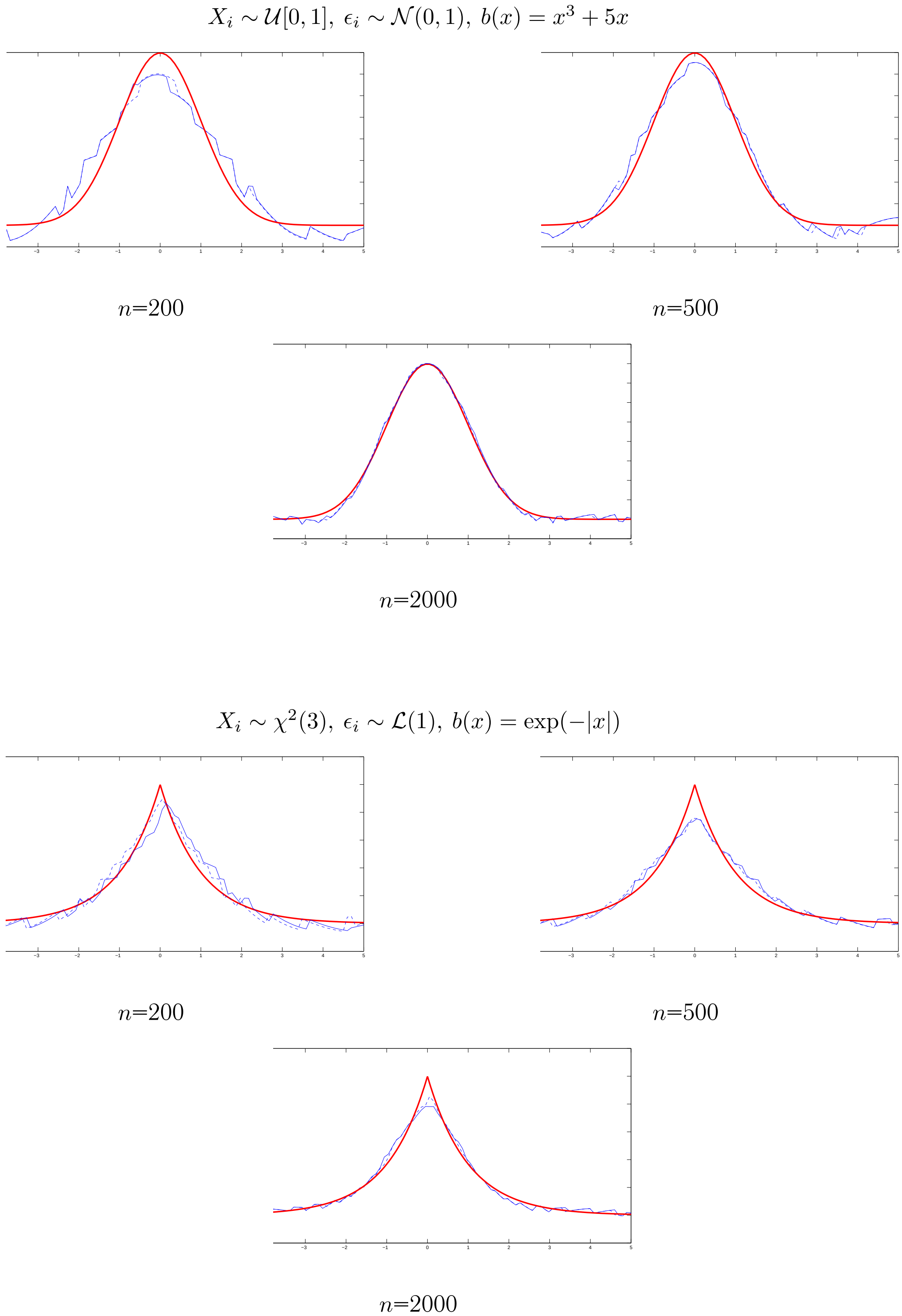


Fig. 3. Error density estimator (solid line), theoretical estimator we would get if the errors were observed (dotted line) and true density (thick line).

6. PROOFS OF SECTION 3

6.1. Proof of Theorem 3.1

The proof is divided in four claims.

Let us denote by $\mathbb{E}_1[\cdot] = \mathbb{E}[\cdot \mid Z_1]$ the conditional expectation given Z_1 and $P_1[\cdot] = P[\cdot \mid Z_1]$ the probability given Z_1.

Claim 1. *Suppose that Assumptions $H_{\rm dens}$ and $H_{\rm mod}$ hold. Then*

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)] \times \mathbf{1}_{\{\widehat{\nu}_n \geq \nu/2\}} \leq 2(g_{m_{opt}} - g)^2(x_0) + 11 \sup_{j, m_{opt} \leq j \leq N_n} (g_j - g_{m_{opt}})^2 + \Big(12 + \frac{4}{x_{m_{opt}}}\Big) \operatorname{pen}(m_{opt}) + \frac{\theta_2}{n}.$$

This entails the following result.

Claim 2. *Under Assumptions $H_{\rm dens}$, $H_{\rm mod}$, $H_{\rm bias}(\beta)$, and (A_3)*

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)] \times \mathbf{1}_{\{\widehat{\nu}_n \geq \nu/2\}} \leq \max(\kappa_2, \widehat{\nu}_n \kappa_3) \inf_{\{m=1,\dots,N_n, (9K^2/\widehat{\nu}_n)\log(1+D_m)D_m/n \leq 1\}} \Big[D_m^{-2\beta} + \log D_m \frac{D_m}{n}\Big] + \frac{\theta_2}{n},$$

where

$$\kappa_1 = 4\Big(3 + \frac{2}{45\log(1+D_1)}\Big), \qquad \kappa_2 = 2C_0^2[\max(15, \kappa_1) + 2], \qquad \kappa_3 = \max(15, \kappa_1)\frac{45AK^2}{2}.$$

We can deduce from Claim 2 the following inequality.

Claim 3. *Suppose that Assumptions $H_{\rm dens}$, $H_{\rm mod}$, $H_{\rm bias}(\beta)$ and (A_3) hold. Moreover, suppose that $M(\log n/n)^{1/(2\beta+1)} \leq D_{N_n}$, then*

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)] \times \mathbf{1}_{\{\widehat{\nu}_n \geq \nu/2\}} \leq \max(\kappa_2, \widehat{\nu}_n \kappa_3)(M+1)\Big(\frac{n}{\log n}\Big)^{\frac{-2\beta}{2\beta+1}} + \frac{\theta_2}{n}.$$

Besides, the following result holds.

Claim 4. *Under Assumptions $H_{\rm dens}$ and $H_{\rm mod}$, for every model $m \in \{1, \dots, N_n\}$ and every $x \in I$,*

$$|\widehat{g}_m(x)| \leq K^2 D_m \quad a.s.$$

The inequalities stated in Claims 3 and 4 allow us to prove Theorem 3.1. Indeed, on the one hand, by Claim 3,

$$\mathbb{E}[(\widehat{g}_{\widehat{m}} - g)^2(x_0)\mathbf{1}_{\{\widehat{\nu}_n \geq \nu/2\}}] \leq \mathbb{E}[\max(\kappa_2, \widehat{\nu}_n \kappa_3)](M+1)\Big(\frac{n}{\log n}\Big)^{-\frac{2\beta}{2\beta+1}} + \frac{\theta_2}{n}.$$

Moreover,

$$\mathbb{E}[\max(\kappa_2, \widehat{\nu}_n \kappa_3)] \leq \mathbb{E}[\kappa_2 + \kappa_3 \widehat{\nu}_n] \leq \kappa_2 + 2\kappa_3 \nu + \kappa_3 \mathbb{E}[\widehat{\nu}_n \mathbf{1}_{\{\widehat{\nu}_n \geq 2\nu\}}].$$

By Claim 4, $\widehat{\nu}_n \leq K^2 p_0$ almost surely, hence $\mathbb{E}[\max(\kappa_2, \widehat{\nu}_n \kappa_3)] \leq \kappa_2 + 2\nu\kappa_3 + \kappa_3 K^2 p_0 P[\widehat{\nu}_n \geq 2\nu]$, and under Assumptions (A_1) and (A_2), inequality (15) of Proposition 3.1 yields

$$\mathbb{E}[\max(\kappa_2, \widehat{\nu}_n \kappa_3)] \leq \kappa_2 + 2\nu\kappa_3 + \kappa_3 K^2 p_0 \exp\Big(-\frac{n\nu}{456K^2 p_0}\Big),$$

which induces that

$$\mathbb{E}[(\widehat{g}_{\widehat{m}} - g)^2(x_0)\mathbf{1}_{\{\widehat{\nu}_n \geq \nu/2\}}]$$

$$\leq \Big(\kappa_2 + 2\nu\kappa_3 + \kappa_3 K^2 p_0 \exp\Big[-\frac{n\nu}{456K^2 p_0}\Big]\Big)(M+1)\Big(\frac{n}{\log n}\Big)^{-\frac{2\beta}{2\beta+1}} + \frac{\theta_2}{n}. \qquad (25)$$

On the other hand, by Claim 4 and inequality (14) in Proposition 3.1,

$$\mathbb{E}[(\widehat{g}_{\widehat{m}} - g)^2(x_0)\mathbf{1}_{\{\widehat{\nu}_n<\nu/2\}}] \leq (\nu + K^2 \max_{m=1,\dots,N_n} D_m)^2 P\Big[\widehat{\nu}_n < \frac{\nu}{2}\Big]$$
$$\leq (\nu + K^2 \max_{m=1,\dots,N_n} D_m)^2 \exp\Big[-\frac{n\nu}{84K^2 p_0}\Big], \qquad (26)$$

and inequalities (25) and (26) provide the result of Theorem 3.1. □

Proof of Claim 1. For every $j \in \{1,\dots,N_n\}$, denote

$$H(j) = K^2 x_j \widehat{\nu}_n \frac{D_j}{n}.$$

The proof of Claim 1 is based on the following steps: we exhibit a quantity $\mathcal{U}_{opt}$ such that

- $\mathbb{E}_1[\mathcal{U}_{opt}]$ has order $\mathrm{Crit}(m_{opt})$.
- $\int_0^{+\infty} P_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt} \geq x]\mathbf{1}_{\{\widehat{\nu}_n\geq\nu/2\}}\,dx$ decreases to 0 with rate $1/n$.

Thus, inequality

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)]\mathbf{1}_{\{\widehat{\nu}_n\geq\nu/2\}} \leq \big(\mathbb{E}_1[((\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt})_+] + \mathbb{E}_1[\mathcal{U}_{opt}]\big)\mathbf{1}_{\{\widehat{\nu}_n\geq/2\nu\}}$$
$$\leq \Big(\int_0^{+\infty} P_1\big[(\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt} \geq x\big]\,dx + \mathbb{E}_1[\mathcal{U}_{opt}]\Big)\mathbf{1}_{\{\widehat{\nu}_n\geq\nu/2\}} \qquad (27)$$

yields the result of Claim 1. Let us consider the first result:

Lemma 6.1. *For every $\delta > 0$, $x > 0$ and for every model m:*

$$P_1[\widehat{\mathrm{Crit}}(m) \geq (1+\delta)\,\mathrm{Crit}(m) + x] \leq 2\sum_{j=m}^{N_n} \exp[-C(x,j,\delta)], \qquad \textit{where}$$

$$C(x,j,\delta) = \min\Big\{\frac{1}{4\nu K^2(1+1/\delta)}\Big(\frac{xn}{D_j} + K^2 x_j\widehat{\nu}_n\Big), \frac{1}{4\sqrt{2(1+1/\delta)}K^2}\Big(\frac{\sqrt{x}n}{D_j} + K\sqrt{x_j\widehat{\nu}_n\frac{n}{D_j}}\Big)\Big\}.$$

Proof of Lemma 6.1. The empirical criterion $\widehat{\mathrm{Crit}}(m)$ (defined in (17)) is built from $\mathrm{Crit}(m)$ (defined in (16)) by replacing the unknown $(g_j - g_m)$ by its empirical counterpart $(\widehat{g}_j - \widehat{g}_m)$, so the deviation between $\widehat{\mathrm{Crit}}(m)$ and $\mathrm{Crit}(m)$ is upper bounded with Bernstein Inequality, which is recalled in Section 9, Theorem 9.1. More precisely:

$$P_1\big[\widehat{\mathrm{Crit}}(m) \geq (1+\delta)\,\mathrm{Crit}(m) + x\big]$$
$$= P_1\Big[\sup_{j,m\leq j\leq N_n} \big((\widehat{g}_j - \widehat{g}_m)^2(x_0) - H(j)\big)_+ \geq (1+\delta)\sup_{j,m\leq j\leq N_n}(g_j - g_m)^2(x_0) + x\Big].$$

As $\sup_{j,m\leq j\leq n_N}(g_j - g_m)^2(x_0) + x$ is positive, we omit the positive part $(\cdot)_+$.

$$P_1\big[\widehat{\mathrm{Crit}}(m) \geq (1+\delta)\,\mathrm{Crit}(m) + x\big]$$
$$= P_1\Big[\sup_{j,m\leq j\leq N_n} \big((\widehat{g}_j - \widehat{g}_m)^2(x_0) - H(j)\big) \geq (1+\delta)\sup_{j,m\leq j\leq N_n}(g_j - g_m)^2(x_0) + x\Big]$$
$$\leq \sum_{j=m}^{N_n} P_1\big[(\widehat{g}_j - \widehat{g}_m)^2(x_0) \geq (1+\delta)(g_j - g_m)^2(x_0) + x + H(j)\big] = \sum_{j=m}^{N_n} P_{j,m} \qquad (28)$$

and for every (j, m),

$$P_{j,m} = P_1\left[(\widehat{g}_j - \widehat{g}_m)^2(x_0) \geq (1+\delta)(g_j - g_m)^2(x_0) + \Big(1+\frac{1}{\delta}\Big)\Bigg(\sqrt{\frac{x+H(j)}{(1+\frac{1}{\delta})}}\Bigg)^2\right].$$

We recall that for every $x, y \in \mathbb{R}$, $(x+y)^2 \leq x^2(1+1/\delta) + y^2(1+\delta)$, thus

$$\begin{aligned}
P_{j,m} &\leq P_1\left[(\widehat{g}_j - \widehat{g}_m)^2(x_0) \geq \Bigg(|(g_j - g_m)(x_0)| + \sqrt{\frac{x+H(j)}{1+\frac{1}{\delta}}}\Bigg)^2\right] \\
&= P_1\left[|(\widehat{g}_j - \widehat{g}_m)(x_0)| \geq |(g_j - g_m)(x_0)| + \sqrt{\frac{x+H(j)}{1+\frac{1}{\delta}}}\ \right] \\
&\leq P_1\left[|(\widehat{g}_j - \widehat{g}_m)(x_0) - (g_j - g_m)(x_0)| + |(g_j - g_m)(x_0)| \geq |(g_j - g_m)(x_0)| + \sqrt{\frac{x+H(j)}{1+\frac{1}{\delta}}}\ \right] \\
&= P_1\left[\Big|\frac{1}{n}\sum_{i=1}^{n}(U_i - \mathbb{E}(U_i))\Big| \geq \sqrt{\frac{x+H(j)}{1+\frac{1}{\delta}}}\ \right], \qquad (29)
\end{aligned}$$

where

$$U_i = \sum_{\lambda \in I_j} \chi_\lambda(V_i)\chi_\lambda(x_0) - \sum_{\lambda \in I_m} \chi_\lambda(V_i)\chi_\lambda(x_0) = \sum_{\lambda \in I_j \setminus I_m} \chi_\lambda(V_i)\chi_\lambda(x_0)$$

and $\mathbb{E}(U_i) = (g_j - g_m)(x_0)$. We are going to upper bound the term (29) with Bernstein's Inequality (Theorem 9.1). Let us compute the terms v and c involved.

Similarly to (12) we get:

$$\mathbb{E}_1(U_1^2) \leq \nu \sum_{\lambda \in I_j \setminus I_m} \chi_\lambda^2(x_0) \leq \nu \sum_{\lambda \in I_j} \chi_\lambda^2(x_0) \leq \nu K^2 D_j = v. \qquad (30)$$

Let ℓ be an integer greater than 2, then,

$$\begin{aligned}
\mathbb{E}_1[(U_1)_+^\ell] &\leq \mathbb{E}_1[U_1^2] \times \|U_1\|_\infty^{\ell-2} \leq v\Big\|\sum_{\lambda \in I_j \setminus I_m} \chi_\lambda(V_1)\chi_\lambda(x_0)\Big\|_\infty^{\ell-2} \\
&\leq v\left[\Big\|\sqrt{\sum_{\lambda \in I_j \setminus I_m} \chi_\lambda^2(V_1)}\ \Big\|_\infty \sqrt{\sum_{\lambda \in I_j \setminus I_m} \chi_\lambda^2(x_0)}\ \right]^{\ell-2}
\end{aligned}$$

and according to (7) in H_{mod}, $\mathbb{E}_1[(U_1)_+^l] \leq v[K^2 D_j]^{l-2}$. So, we set

$$c = K^2 D_j. \qquad (31)$$

Finally, we denote by

$$\epsilon = \sqrt{\frac{x+H(j)}{1+1/\delta}} \geq \frac{1}{\sqrt{2(1+1/\delta)}}\big(\sqrt{x} + \sqrt{H(j)}\ \big). \qquad (32)$$

Then by Bernstein's Inequality,

$$P_{j,m} \leq 2\exp\Big[-\min\Big(\frac{n\epsilon^2}{4v}; \frac{n\epsilon}{4c}\Big)\Big].$$

Moreover:

$$\frac{n\epsilon^2}{4v} = \frac{1}{4\nu K^2(1+1/\delta)}\Big(\frac{xn}{D_j} + K^2 x_j \widehat{\nu}_n\Big), \qquad \frac{n\epsilon}{4c} \geq \frac{1}{4\sqrt{2(1+1/\delta)}K^2}\Bigg(\frac{\sqrt{x}n}{D_j} + K\sqrt{x_j\widehat{\nu}_n\frac{n}{D_j}}\ \Bigg).$$

This provides an upper bound of $P_{j,m}$ for every (j, m), which, being inserted in inequality (28), ends the proof of Lemma 6.1. □

We derive from Lemma 6.1 the following result.

Lemma 6.2. *For every positive numbers δ and x and every sequence Z_1,*

(1) $$P_1\Big[\big\{(\widehat{g}_{\widehat{m}}-g)^2(x_0)\geq(1+\delta)\big(\sup_{j,m_{opt}\leq j\leq N_n}(g_j-g)^2(x_0)+\mathrm{Crit}(m_{opt})\big)+x\big\}\cap\{\widehat{m}>m_{opt}\}\Big]$$
$$\leq 4\sum_{m=1}^{N_n}\exp[-C(x,m,\delta)];$$

(2) $$P_1\Big[\big\{(\widehat{g}_{\widehat{m}}-g)^2(x_0)\geq 2(\widehat{g}_{m_{opt}}-g)^2(x_0)+2H(m_{opt})+2(1+\delta)\,\mathrm{Crit}(m_{opt})\big\}\cap\{\widehat{m}\leq m_{opt}\}\Big]$$
$$\leq 2\sum_{j=m_{opt}}^{N_n}\exp[-C(x,j,\delta)].$$

Proof. • Let us prove inequality (1):

$$P_1\Big[\big\{(\widehat{g}_{\widehat{m}}-g)^2(x_0)\geq(1+\delta)\big(\sup_{j,m_{opt}\leq j\leq N_n}(g_j-g)^2(x_0)+\mathrm{Crit}(m_{opt})\big)+x\big\}\cap\{\widehat{m}>m_{opt}\}\Big]$$
$$\leq P_1\Big[\big\{(\widehat{g}_{\widehat{m}}-g)^2(x_0)\geq(1+\delta)\sup_{j,m_{opt}\leq j\leq N_n}(g_j-g)^2(x_0)+\widehat{\mathrm{Crit}}(\widehat{m})+x\big\}\cap\{\widehat{m}>m_{opt}\}\Big]$$
$$+P_1\big[\widehat{\mathrm{Crit}}(\widehat{m})\geq(1+\delta)\,\mathrm{Crit}(m_{opt})\big]. \tag{33}$$

By definition of $\widehat{m}$, $\widehat{\mathrm{Crit}}(\widehat{m})=\inf_{m=1,\dots N_n}\widehat{\mathrm{Crit}}(m)\leq\widehat{\mathrm{Crit}}(m_{opt})$. Hence, by Lemma 6.1,

$$P_1\big[\widehat{\mathrm{Crit}}(\widehat{m})\geq(1+\delta)\,\mathrm{Crit}(m_{opt})\big]\leq P\big[\widehat{\mathrm{Crit}}(m_{opt})\geq(1+\delta)\,\mathrm{Crit}(m_{opt})\big]$$
$$\leq 2\sum_{j=m_{opt}}^{N_n}\exp[-C(x,j,\delta)]\leq 2\sum_{m=1}^{N_n}\exp[-C(x,m,\delta)]. \tag{34}$$

Besides it is clear that for every model m, $\mathrm{Crit}(m)\geq\mathrm{pen}(m)$, and if $\widehat{m}>m_{opt}$, then

$$\sup_{j,m_{opt}\leq j\leq N_n}(g_j-g)^2(x_0)\geq(g_{\widehat{m}}-g)^2(x_0).$$

So,

$$P_1\Big[\big\{(\widehat{g}_{\widehat{m}}-g)^2(x_0)\geq(1+\delta)\sup_{j,m_{opt}\leq j\leq N_n}(g_j-g)^2(x_0)+\widehat{\mathrm{Crit}}(\widehat{m})+x\big\}\cap\{\widehat{m}>m_{opt}\}\Big]$$
$$\leq P_1\Big[(\widehat{g}_{\widehat{m}}-g)^2(x_0)\geq(1+\delta)(g_{\widehat{m}}-g)^2(x_0)+\mathrm{pen}(\widehat{m})+x\Big]$$
$$\leq\sum_{m=1}^{N_n}P_1\Big[(\widehat{g}_m-g)^2(x_0)\geq(1+\delta)(g_m-g)^2(x_0)+\mathrm{pen}(m)+x\Big]=\sum_{m=1}^{N_n}P_m. \tag{35}$$

For every $m\in\{1,\dots,N_n\}$, we have almost surely

$$(\widehat{g}_m-g)^2(x_0)\leq(1+\delta)(g-g_m)^2(x_0)+\Big(1+\frac{1}{\delta}\Big)(\widehat{g}_m-g_m)^2(x_0)$$

and $\mathrm{pen}(m)=AH(m)\geq H(m)$, so

$$P_m\leq P_1\Big[\Big(1+\frac{1}{\delta}\Big)(\widehat{g}_m-g_m)^2(x_0)\geq\mathrm{pen}(m)+x\Big]$$
$$\leq P_1\Bigg[|(\widehat{g}_m-g_m)^2(x_0)|\geq\sqrt{\frac{H(m)+x}{1+1/\delta}}\Bigg]=P_1\Bigg[\Big|\frac{1}{n}\sum_{i=1}^{n}(U_i-\mathbb{E}(U_i))\Big|\geq\sqrt{\frac{H(m)+x}{1+1/\delta}}\Bigg],$$

where $U_i = \sum_{\lambda \in I_m} \chi_\lambda(V_i)\chi_\lambda(x_0)$. Similarly to the proof of Lemma 6.1, we apply Bernstein's Inequality with the parameters defined in (30), (31), and (32), and we get

$$P_m \leq 2\exp(-C(x, m, \delta)). \tag{36}$$

Combining inequalities (33), (34), (35) and (36), the result of Lemma 6.2 (1) follows.

• Let us prove now inequality (2) in Lemma 6.2. By definition of $\widehat{m}$, $\widehat{\mathrm{Crit}}(\widehat{m}) \leq \widehat{\mathrm{Crit}}(m_{opt})$, so

$$\begin{aligned} P_1\big[\widehat{\mathrm{Crit}}(m_{opt}) \geq (1+\delta)\,\mathrm{Crit}(m_{opt})\big] &\geq P_1\big[\widehat{\mathrm{Crit}}(\widehat{m}) \geq (1+\delta)\,\mathrm{Crit}(m_{opt})\big] \\ &\geq P_1\Big[\sup_{j, \widehat{m} \leq j \leq N_n} \big[(\widehat{g}_j - \widehat{g}_{\widehat{m}})^2(x_0) - H(j)\big] + \mathrm{pen}(\widehat{m}) \geq (1+\delta)\,\mathrm{Crit}(m_{opt})\Big] \\ &\geq P_1\Big[\big\{(\widehat{g}_{m_{opt}} - \widehat{g}_{\widehat{m}})^2(x_0) - H(m_{opt}) + \mathrm{pen}(\widehat{m}) \geq (1+\delta)\,\mathrm{Crit}(m_{opt})\big\} \cap \{\widehat{m} \leq m_{opt}\}\Big]. \end{aligned} \tag{37}$$

Besides, $(\widehat{g}_{\widehat{m}} - g)^2(x_0) \leq 2(\widehat{g}_{m_{opt}} - \widehat{g}_{\widehat{m}})^2(x_0) + 2(\widehat{g}_{m_{opt}} - g)^2(x_0)$, therefore

$$(\widehat{g}_{m_{opt}} - \widehat{g}_{\widehat{m}})^2(x_0) \geq \frac{1}{2}(\widehat{g}_{\widehat{m}} - g)^2(x_0) - (\widehat{g}_{m_{opt}} - g)^2(x_0).$$

So we derive from (37) that

$$\begin{aligned} P_1\big[\widehat{\mathrm{Crit}}(m_{opt}) \geq (1+\delta)\,\mathrm{Crit}(m_{opt})\big] \geq P_1\Big[\Big\{\frac{1}{2}(\widehat{g}_{\widehat{m}} - g)^2(x_0) - (\widehat{g}_{m_{opt}} - g)^2(x_0) \\ \geq (1+\delta)\,\mathrm{Crit}(m_{opt}) + H(m_{opt}) - \mathrm{pen}(\widehat{m})\Big\} \cap \{\widehat{m} \leq m_{opt}\}\Big]. \end{aligned}$$

As $\mathrm{pen}(\widehat{m})$ is positive, we get

$$\begin{aligned} &P_1\big[\widehat{\mathrm{Crit}}(m_{opt}) \geq (1+\delta)\,\mathrm{Crit}(m_{opt})\big] \\ &\quad \geq P_1\Big[\big\{(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq 2(\widehat{g}_{m_{opt}} - g)^2(x_0) + 2H(m_{opt}) + 2(1+\delta)\,\mathrm{Crit}(m_{opt})\big\} \cap \{\widehat{m} \leq m_{opt}\}\Big]. \end{aligned}$$

By Lemma 6.1, inequality (2) of Lemma 6.2 follows. □

Let us prove Claim 1. Consider

$$\mathcal{U}_{opt} = 2(\widehat{g}_{m_{opt}} - g)^2(x_0) + 2(1+\delta)\,\mathrm{Crit}(m_{opt}) + 2H(m_{opt}) + \sup_{j, m_{opt} \leq j \leq N_n} (g_j - g)^2(x_0).$$

Then, by inequalities (1) and (2) in Lemma 6.2, we get

$$P_1\big[(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq \mathcal{U}_{opt} + x\big] \leq 4\sum_{m=1}^{N_n} \exp[-C(x, m, \delta)].$$

Take $\delta = 4$, then

$$\begin{aligned} \mathbb{E}_1\big[\big((\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt}\big)_+\big] &\leq \int_0^{+\infty} P_1\big[(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq \mathcal{U}_{opt} + x\big]\,dx \\ &\leq 4\int_0^{+\infty} \Big(\sum_{m=1}^{N_n} \exp[-C(x, m, 4)]\Big)\,dx. \end{aligned} \tag{38}$$

We recall that, for every positive constant C':

$$\int_0^{+\infty} \exp(-C'x)\,dx = \frac{1}{C'}, \qquad \int_0^{+\infty} \exp(-C'\sqrt{x})\,dx = \frac{2}{C'^2}.$$

Therefore, according to the expression of $C(x,m,\delta)$ defined in Lemma 6.1,

$$\int_0^{+\infty}\sum_{m=1}^{N_n}\exp[-C(x,m,4)]\,dx \le \sum_{m=1}^{N_n}\Big[5\nu K^2\frac{D_m}{n}\exp\Big[-\frac{x_m\widehat{\nu}_n}{5\nu}\Big] + 80K^4\frac{D_m^2}{n^2}\exp\Big[-\frac{1}{2\sqrt{10}K}\sqrt{x_m\widehat{\nu}_n\frac{n}{D_m}}\,\Big]\Big].$$

Besides, assuming that $\widehat{\nu}_n \ge \nu/2$,

$$x_m \ge \frac{45}{2}\log(1+D_m) \Rightarrow x_m \ge \frac{45}{4}\frac{\nu}{\widehat{\nu}_n}\log(1+D_m) \Leftrightarrow \exp\Big[-\frac{x_m\widehat{\nu}_n}{5\nu}\Big] \le (1+D_m)^{-(2+1/4)}$$
$$\Leftrightarrow D_m\exp\Big[-\frac{x_m\widehat{\nu}_n}{5\nu}\Big] \le (1+D_m)^{-(1+1/4)} \quad (39)$$

and similarly,

$$x_m \ge \frac{45}{2}\times\frac{9K^2}{\widehat{\nu}_n}\log^2(1+D_m)\frac{D_m}{n} \Leftrightarrow \exp\Big[-\frac{1}{2\sqrt{10}K}\sqrt{x_m\widehat{\nu}_n\frac{n}{D_m}}\,\Big] \le (1+D_m)^{-(2+1/4)}$$
$$\Leftrightarrow D_m\exp\Big[-\frac{1}{2\sqrt{10}K}\sqrt{x_m\widehat{\nu}_n\frac{n}{D_m}}\,\Big] \le (1+D_m)^{-(1+1/4)}. \quad (40)$$

Hence

$$\int_0^{+\infty}\sum_{m=1}^{N_n}\exp[-C(x,m,4)]\,dx \le 5K^2(\nu+16K^2)\Big(\sum_{m=1}^{N_n}(1+D_m)^{1+1/4}\Big)\frac{1}{n}.$$

Plugging these upper bounds in inequality (38) yields

$$\mathbb{E}_1\big[((\widehat{g}_{\widehat{m}}-g)^2(x_0)-\mathcal{U}_{opt})_+\big] \le 20K^2(\nu+16K^2)\Big(\sum_{m=1}^{N_n}(1+D_m)^{1+1/4}\Big)\frac{1}{n} = \frac{\theta_2}{n}. \quad (41)$$

It remains to upper bound $\mathbb{E}_1[\mathcal{U}_{opt}]$:

$$\begin{aligned}\mathbb{E}_1[\mathcal{U}_{opt}] &= 2\mathbb{E}[(\widehat{g}_{m_{opt}}-g)^2(x_0)] + 2\widehat{\nu}_nK^2x_{m_{opt}}\frac{D_{m_{opt}}}{n} \\ &\quad + \sup_{j,m_{opt}\le j\le N_n}(g_j-g_{m_{opt}})^2(x_0) + 10\,\mathrm{Crit}(m_{opt}) \\ &\le 2\Big[(g_{m_{opt}}-g)^2(x_0)+\nu K^2\frac{D_{m_{opt}}}{n}\Big] + 2\widehat{\nu}_nK^2x_{m_{opt}}\frac{D_{m_{opt}}}{n} \\ &\quad + \sup_{j,m_{opt}\le j\le N_n}(g_j-g_{m_{opt}})^2(x_0) + 10\,\mathrm{Crit}(m_{opt}).\end{aligned}$$

Thus on the set $\{\widehat{\nu}_n \ge \nu/2\}$ we have

$$\begin{aligned}\mathbb{E}_1[\mathcal{U}_{opt}]\mathbf{1}_{\{\widehat{\nu}_n\ge\nu/2\}} &\le 2(g_{m_{opt}}-g)^2(x_0) + 4\widehat{\nu}_nK^2\frac{D_{m_{opt}}}{n} \\ &\quad + (2+10A)\widehat{\nu}_nK^2x_{m_{opt}}\frac{D_{m_{opt}}}{n} + 11\sup_{j,m_{opt}\le j\le N_n}(g_j-g_{m_{opt}})^2 \\ &\le 2(g_{m_{opt}}-g)^2(x_0) + 11\sup_{j,m_{opt}\le j\le N_n}(g_j-g_{m_{opt}})^2 \\ &\quad + \Big(12+\frac{4}{x_{m_{opt}}}\Big)\,\mathrm{pen}(m_{opt}). \end{aligned} \quad (42)$$

Putting together inequalities (27), (41), and (42), we get

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)] \times \mathbf{1}_{\{\widehat{\nu}_n \geq \nu/2\}} \leq 2(g_{m_{opt}} - g)^2(x_0) + 11 \sup_{j, m_{opt} \leq j \leq N_n} (g_j - g_{m_{opt}})^2 + \Big(12 + \frac{4}{x_{m_{opt}}}\Big) \operatorname{pen}(m_{opt}) + \frac{\theta_2}{n},$$

which ends the proof of Claim 1. □

Proof of Claim 2. First of all, note that

$$4\Big(3 + \frac{1}{x_{m_{opt}}}\Big) \leq 4\Big(3 + \frac{2}{45\log(1 + D_{m_{opt}})}\Big) \leq 4\Big(3 + \frac{2}{45\log(1 + D_1)}\Big) = \kappa_1.$$

Suppose that $\widehat{\nu}_n \geq \nu/2$ and denote

$$\begin{cases} F(m) = D_m^{-2\beta} + \log(1 + D_m)\frac{D_m}{n}, \\ m_1 = \arg\min\{F(m),\ m = 1, \ldots, N_n,\ (9K^2/\widehat{\nu}_n)\log(1 + D_m)D_m/n \leq 1\}. \end{cases}$$

Thus, $x_{m_1} = (45/2)\log(1 + D_{m_1})$. We consider two cases: $m_{opt} \geq m_1$ and $m_{opt} < m_1$.

• If $m_{opt} \geq m_1$, by $H_{\text{bias}}(\beta)$, $(g_{m_{opt}} - g)^2(x_0) \leq C_0^2 D_{m_{opt}}^{-2\beta} \leq C_0^2 D_{m_1}^{-2\beta}$. Besides, it is obvious that $11 \leq \kappa_1$. Thus by Claim 1, we get

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)]\mathbf{1}_{\{\widehat{\nu}_n \geq \nu/2\}} \leq 2C_0^2 D_{m_1}^{-2\beta} + \kappa_1 \operatorname{Crit}(m_{opt}) + \frac{\theta_2}{n}.$$

As $m_{opt} = \arg\min_{m=1,\ldots,N_n} \operatorname{Crit}(m)$, $\operatorname{Crit}(m_{opt}) \leq \operatorname{Crit}(m_1)$. Then,

$$\begin{aligned} \mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)]\mathbf{1}_{\{\widehat{\nu}_n \geq \nu/2\}} &\leq 2C_0^2 D_{m_1}^{-2\beta} + \kappa_1 \operatorname{Crit}(m_1) + \frac{\theta_2}{n} \\ &\leq 2C_0^2(1 + \kappa_1)D_{m_1}^{-2\beta} + \kappa_1 \frac{45AK^2}{2}\widehat{\nu}_n \log(1 + D_{m_1})\frac{D_{m_1}}{n} + \frac{\theta_2}{n} \\ &\leq \max\Big\{2C_0^2(1 + \kappa_1), \kappa_1\frac{45AK^2}{2}\widehat{\nu}_n\Big\}F(m_1) + \frac{\theta_2}{n}. \end{aligned} \tag{43}$$

• If $m_{opt} < m_1$, then

$$\begin{aligned} (g_{m_{opt}} - g)^2(x_0) &\leq 2(g_{m_{opt}} - g_{m_1})^2(x_0) + 2(g_{m_1} - g)^2(x_0) \\ &\leq 2 \sup_{j, m_{opt} \leq j \leq N_n} (g_j - g_{m_{opt}})^2(x_0) + 2C_0^2 D_{m_1}^{-2\beta}. \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)]\mathbf{1}_{\{\widehat{\nu}_n \geq \nu/2\}} &\leq 15 \sup_{j, m_{opt} \leq j \leq N_n} (g_j - g_{m_{opt}})^2(x_0) + \kappa_1 \operatorname{pen}(m_{opt}) + 4C_0^2 D_{m_1}^{-2\beta} + \frac{\theta_2}{n} \\ &\leq \max(15, \kappa_1)\operatorname{Crit}(m_{opt}) + 4C_0^2 D_{m_1}^{-2\beta} + \frac{\theta_2}{n} \\ &\leq \max(15, \kappa_1)\operatorname{Crit}(m_1) + 4C_0^2 D_{m_1}^{-2\beta} + \frac{\theta_2}{n} \\ &\leq \max(15, \kappa_1)\big[2C_0^2 D_{m_1}^{-2\beta} + \operatorname{pen}(m_1)\big] + 4C_0^2 D_{m_1}^{-2\beta} + \frac{\theta_2}{n} \\ &\leq \max\Big\{2C_0^2[\max(15, \kappa_1) + 2], \max(15, \kappa_1)\frac{45AK^2}{2}\widehat{\nu}_n\Big\}F(m_1) + \frac{\theta_2}{n}. \end{aligned} \tag{44}$$

Moreover, it is clear that

$$2C_0^2(1 + \kappa_1) \leq 2C_0^2[\max(15, \kappa_1) + 2], \qquad \kappa_1\frac{45AK^2}{2} \leq \max(15, \kappa_1)\frac{45AK^2}{2}.$$

Therefore, inequalities (43) and (44) yield the proof of Claim 2. □

Proof of Claim 3. Let m_2 be a model such that

$$\Big(\frac{n}{\log n}\Big)^{\frac{1}{2\beta+1}} \le D_{m_2} \le M\Big(\frac{n}{\log n}\Big)^{\frac{1}{2\beta+1}}. \tag{45}$$

On the set $\{\widehat{\nu}_n \ge \nu/2\}$, by Assumption (A_3),

$$\frac{9K^2}{\widehat{\nu}_n}\log(1+D_{m_2})\frac{D_{m_2}}{n} \le \frac{18K^2}{\nu}\Big(\frac{n}{\log n}\Big)^{-2\beta/(2\beta+1)} \le 1.$$

Hence, by definition of m_1,

$$F(m_1) \le F(m_2) \le M\frac{\log n}{n}\Big(\frac{n}{\log n}\Big)^{\frac{1}{2\beta+1}} + \Big(\frac{n}{\log n}\Big)^{-\frac{2\beta}{2\beta+1}} \le (M+1)\Big(\frac{n}{\log n}\Big)^{-\frac{2\beta}{2\beta+1}}.$$

Thus we derive from Claim 2 that

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)]\mathbf{1}_{\{\widehat{\nu}_n\ge\nu/2\}} \le \max(\kappa_2, \widehat{\nu}_n\kappa_3)(M+1)\Big(\frac{n}{\log n}\Big)^{\frac{-2\beta}{2\beta+1}} + \frac{\theta_2}{n}. \qquad \square$$

Proof of Claim 4. For every model m, $(\widehat{g}_m - g)^2(x_0) \le (|\widehat{g}_m(x_0)| + \nu)^2$ almost surely. Besides,

$$(\widehat{g}_m)^2(x_0) = \sum_{\lambda\in I_m}\Big(\frac{1}{n}\sum_{i=1}^n \chi_\lambda(V_i)\chi_\lambda(x_0)\Big)^2 \le \frac{1}{n}\sum_{i=1}^n\Big(\sum_{\lambda\in I_m}\chi_\lambda(V_i)\chi_\lambda(x_0)\Big)^2$$
$$\le \Big\|\sum_{\lambda\in I_m}\chi_\lambda^2\Big\|_\infty^2 \le K^4D_m^2 \tag{46}$$

which provides the result of Claim 4. $\square$

6.2. Proof of Proposition 3.1

Let us prove inequality (14). Let $x_1 \in I$ be such that $g(x_1) \ge 5\nu/6$, then by definition of $\widehat{\nu}_n$,

$$P\Big[\widehat{\nu}_n \le \frac{\nu}{2}\Big] \le P\Big[\widehat{g}^{(1)}_{m_0}(x_1) \le \frac{\nu}{2}\Big] = P\Big[(\widehat{g}^{(1)}_{m_0} - g_{m_0})(x_1) \le \frac{5\nu}{6} - g_{m_0}(x_1) - \frac{\nu}{3}\Big]$$
$$\le P\Big[(\widehat{g}^{(1)}_{m_0} - g_{m_0})(x_1) \le (g - g_{m_0})(x_1) - \frac{\nu}{3}\Big].$$

By Assumption $H_{\text{bias}}(\beta)$,

$$P\Big[\widehat{\nu}_n \le \frac{\nu}{2}\Big] \le P\Big[(\widehat{g}^{(1)}_{m_0} - g_{m_0})(x_1) \le Lp_0^{-\beta} - \frac{\nu}{3}\Big]$$

and by condition (A_1),

$$P\Big[\widehat{\nu}_n \le \frac{\nu}{2}\Big] \le P\Big[(\widehat{g}^{(1)}_{m_0} - g_{m_0})(x_1) \le -\frac{\nu}{6}\Big] \le P\Big[|(\widehat{g}^{(1)}_{m_0} - g_{m_0})(x_1)| \ge \frac{\nu}{6}\Big]$$
$$= P\Big[\Big|\frac{1}{n}\sum_{i=n+1}^{2n} U_i - \mathbb{E}(U_i)\Big| \ge \frac{\nu}{6}\Big]. \tag{47}$$

Now, apply Bernstein's Inequality (Theorem 9.1) with the following parameters:

$$\mathbb{E}[U_1^2] = \mathbb{E}\Big[\Big(\sum_{\lambda\in I_{m_0}}\xi_\lambda(V_1)\xi_\lambda(x_1)\Big)^2\Big] = \int_I\Big(\sum_{\lambda\in I_{m_0}}\xi_\lambda(x)\xi_\lambda(x_1)\Big)^2 g(x)\,dx$$
$$\le \nu\sum_{\lambda,\lambda'\in I_{m_0}}\Big[\int_I \xi_\lambda(x)\xi_{\lambda'}(x)\,dx\Big]\xi_\lambda(x_1)\xi_{\lambda'}(x_1) = \nu\sum_{\lambda\in I_{m_0}}\xi_\lambda^2(x_1),$$

since the family $\{\xi_\lambda\}$ is orthonormal. Finally, Assumption (7) in H_{mod} yields

$$\mathbb{E}[U_1^2] \leq \nu K^2 p_0 = v.$$

Let l be an integer greater than 2,

$$\mathbb{E}[(X_1)_+^l] \leq \mathbb{E}[U_1^2] \times \|U_1\|_\infty^{l-2} \leq v \Big\| \sum_{\lambda \in I_{m_0}} \xi_\lambda(V_1)\xi_\lambda(x_0) \Big\|_\infty^{l-2}$$

$$\leq v \Bigg[\sqrt{\Big\| \sum_{\lambda \in I_{m_0}} \xi_\lambda^2(V_1) \Big\|_\infty} \sqrt{\sum_{\lambda \in I_{p_0}} \xi_\lambda^2(x_0)} \Bigg]^{l-2} \leq v(K^2 p_0)^{l-2}.$$

Hence we set $c = K^2 p_0$. By Bernstein's Inequality, we derive from inequality (47) that

$$P\Big[\widehat{\nu}_n \leq \frac{\nu}{2}\Big] \leq 2 \exp\Big[- \frac{n\nu}{84 K^2 p_0} \Big],$$

which is the result we wanted to prove.

Let us prove inequality (15). Let $\widehat{x}_1 \in I$ be such that $\widehat{g}_{m_0}(\widehat{x}_1) \geq \frac{5\widehat{\nu}_n}{6}$. Similarly to (47), under condition (A_1),

$$P\Big[\nu \leq \frac{\widehat{\nu}_n}{2}\Big] \leq P\Big[|(\widehat{g}_{m_0}^{(1)} - g_{m_0})(\widehat{x}_1)| \geq \frac{\nu}{6} \Big].$$

Moreover,

$$P\Big[\nu \leq \frac{\widehat{\nu}_n}{2}\Big] \leq P\Big[\sup_{x \in I} |(\widehat{g}_{m_0}^{(1)} - g_{m_0})(x)| \geq \frac{\nu}{6} \Big]$$

$$= P\Big[\sup_{x \in I} \frac{1}{n} \sum_{i=n+1}^{2n} \Big\{ \sum_{\lambda \in I_{m_0}} \big(\xi_\lambda(V_i)\xi_\lambda(x) - \mathbb{E}[\xi_\lambda(V_i)\xi_\lambda(x)] \big) \Big\} \geq \frac{\nu}{6} \Big]$$

$$= P\Big[\sup_{x \in I} \frac{1}{n} \sum_{i=n+1}^{2n} \varphi_x(V_i) \geq \frac{\nu}{6} \Big].$$

We have in view to apply Talagrand's Inequality recalled in Section 9, Theorem 9.2, but the set of functions

$$\mathcal{F} = \Big\{ \varphi_x \colon u \to \sum_{\lambda \in I_{m_0}} \xi_\lambda(x)\xi_\lambda(u) - \mathbb{E}[\xi_\lambda(x)\xi_\lambda(V_1)],\ x \in I \Big\}$$

is not countable. Nevertheless, the (ξ_λ) are continuous, thus for every u the mapping $x \to \varphi_x(u)$ is continuous. Hence, since the set $\mathbb{Q} \cap I$ is dense in I, we have

$$Z = \sup_{x \in I} \frac{1}{n} \sum_{i=n+1}^{2n} \varphi_x(V_i) = \sup_{x \in I \cap \mathbb{Q}} \frac{1}{n} \sum_{i=n+1}^{2n} \varphi_x(V_i),$$

so

$$P\Big[\nu \leq \frac{\widehat{\nu}_n}{2}\Big] \leq P\Big[\sup_{x \in I \cap \mathbb{Q}} \frac{1}{n} \sum_{i=n+1}^{2n} \varphi_x(V_i) \geq \frac{\nu}{6} \Big]$$

and $\mathbb{Q} \cap I$ is countable. Let $x \in I$, by the Cauchy–Schwarz Inequality,

$$\Big(\frac{1}{n} \sum_{i=n+1}^{2n} \varphi_x(V_i) \Big) \leq \frac{1}{n} \sum_{i=n+1}^{2n} (\varphi_x(V_i))^2 = \frac{1}{n} \sum_{i=n+1}^{2n} \Big(\sum_{\lambda \in I_{m_0}} \big(\xi_\lambda(V_i) - \mathbb{E}[\xi_\lambda(V_i)] \big) \xi_\lambda(x) \Big)^2$$

$$\leq \frac{1}{n} \sum_{i=n+1}^{2n} \Big(\sum_{\lambda \in I_{m_0}} \xi_\lambda^2(x) \Big) \Big(\sum_{\lambda \in I_{m_0}} \big(\xi_\lambda(V_i) - \mathbb{E}[\xi_\lambda(V_i)] \big)^2 \Big).$$

Then, by Assumption H_{mod},

$$\Big(\frac{1}{n}\sum_{i=n+1}^{2n}\varphi_x(V_i)\Big) \le K^2 p_0 \frac{1}{n}\sum_{i=n+1}^{2n}\Big(\sum_{\lambda\in I_{m_0}}\big(\xi_\lambda(V_i)-\mathbb{E}[\xi_\lambda(V_i)]\big)^2\Big).$$

Hence,

$$\Big(\mathbb{E}\Big[\Big|\sup_{x\in I\cap\mathbb{Q}}\frac{1}{n}\sum_{i=n+1}^{2n}\varphi_x(V_i)\Big|\Big]\Big)^2 \le \mathbb{E}\Big[\Big(\sup_{x\in I\cap\mathbb{Q}}\frac{1}{n}\sum_{i=n+1}^{2n}\varphi_x(V_i)\Big)^2\Big]$$

$$\le K^2 p_0 \sum_{\lambda\in I_{m_0}}\mathbb{E}\Big[\frac{1}{n}\sum_{i=n+1}^{2n}\big(\xi_\lambda(V_1)-\mathbb{E}[\xi_\lambda(V_1)]\big)^2\Big] = \frac{K^2p_0}{n}\mathbb{E}\Big[\sum_{\lambda\in I_{m_0}}\big(\xi_\lambda(V_1)-\mathbb{E}[\xi_\lambda(V_1)]\big)^2\Big]$$

$$= \frac{K^2p_0}{n}\sum_{\lambda\in I_{m_0}}\mathrm{Var}(\xi_\lambda(V_1)) \le \frac{K^2p_0}{n}\mathbb{E}\Big[\sum_{\lambda\in I_{m_0}}\xi_\lambda^2(V_1))\Big] \le \frac{K^4p_0^2}{n}.$$

Thus,

$$\mathbb{E}\Big[\Big|\sup_{x\in I\cap\mathbb{Q}}\frac{1}{n}\sum_{i=n+1}^{2n}\varphi_x(V_i)\Big|\Big] \le \frac{K^2p_0}{\sqrt{n}} = \mathbb{H}.$$

Let us compute the terms v and c involved in Talagrand's Inequality. For every $x\in I$,

$$\mathrm{Var}\Big(\sum_{\lambda\in I_{m_0}}\xi_\lambda(V_1)\xi_\lambda(x)\Big) \le \mathbb{E}\Big[\Big(\sum_{\lambda\in I_{m_0}}\xi_\lambda(V_1)\xi_\lambda(x)\Big)^2\Big] = \int_I\Big(\sum_{\lambda\in I_{m_0}}\xi_\lambda(u)\xi_\lambda(x)\Big)^2 g(u)\,du$$

$$\le \nu\int_I\Big(\sum_{\lambda\in I_{m_0}}\xi_\lambda(u)\xi_\lambda(x)\Big)^2 du = \nu\sum_{\lambda,\lambda'\in I_{m_0}}\Big[\int_I\xi_\lambda(u)\xi_{\lambda'}(u)\,du\Big]\xi_\lambda(x)\xi_{\lambda'}(x).$$

The family $\{\xi_\lambda,\lambda\in I_{m_0}\}$ is orthonormal, so

$$\mathrm{Var}\Big(\sum_{\lambda\in I_{m_0}}\xi_\lambda(V_1)\xi_\lambda(x)\Big) \le \nu\sum_{\lambda\in I_{m_0}}\xi_\lambda^2(x) \le \nu K^2p_0 = v.$$

Besides,

$$\Big\|\sum_{\lambda\in I_{m_0}}\xi_\lambda(x)\xi_\lambda\Big\|_\infty \le \sqrt{\sum_{\lambda\in I_{m_0}}\xi_\lambda^2(x)}\times\Big\|\sqrt{\sum_{\lambda\in I_{m_0}}\xi_\lambda^2}\Big\|_\infty \le K^2p_0 = b.$$

Moreover, Assumption (A_2) yields

$$P\Big[\nu\le\frac{\widehat{\nu}_n}{2}\Big] \le P\Big[Z\ge\frac{\nu}{6}\Big] = P\Big[Z\ge\mathbb{H}+\Big(\frac{\nu}{6}-\frac{K^2p_0}{\sqrt{n}}\Big)\Big] \le P\Big[Z\ge\mathbb{H}+\frac{\nu}{12}\Big].$$

Finally, Talagrand's Inequality provides the following upper bound:

$$P\Big[Z\ge\mathbb{H}+\frac{a}{3}\nu\Big] \le \exp\Big[-\frac{n(\nu/12)^2}{2(\nu K^2p_0+4(K^2p_0)^2/\sqrt{n}+3K^2p_0(\nu/12))}\Big].$$

Applying once again Assumption (A_2), we get

$$\exp\Big[-\frac{n(\nu/12)^2}{2(\nu K^2p_0+4(K^2p_0)^2/\sqrt{n}+3K^2p_0(\nu/12))}\Big]$$

$$\le \exp\Big[-\frac{n(\nu/12)^2}{2(\nu K^2p_0+4K^2p_0(\nu/12)/\sqrt{n}+3K^2p_0(\nu/12))}\Big]$$

$$= \exp\Big[-\frac{n\nu}{456K^2p_0}\Big]. \qquad \square$$

7. PROOF OF THEOREM 4.1

The proof is based on the decomposition (22).

7.1. Upper Bound of $\mathbb{E}[(f-f^-)^2(x_0)]$

Proposition 7.1. *Suppose that f is Lipschitz, then*

$$\mathbb{E}\big[(f^- - f)^2(x_0)\big] \le \mathrm{Lip}(f)^2\mathbb{E}\big[\|b-\widehat{b}\|_{f_X}^2\big]. \tag{48}$$

In fact, for every Z^-

$$\begin{aligned}(f-f^-)^2(x_0) &= \Big(\int_0^1 \big[f(x_0) - f(x_0-(b-\widehat{b})(x))\big] f_X(x)\,dx\Big)^2\\ &\le \int_0^1 \big[f(x_0)-f(x_0-(b-\widehat{b})(x))\big]^2 f_X(x)\,dx\\ &\le \mathrm{Lip}(f)\int_0^1 \big[(b-\widehat{b})(x)\big]^2 f_X(x)\,dx = \mathrm{Lip}(f)\|b-\widehat{b}\|_{f_X}^2\end{aligned}$$

and by considering the expectation of the above inequality we get the result of Proposition 7.1. □

7.2. Upper Bound of $\mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)]$

Now, the term $\mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)]$ in (22) is upper bounded with the results of Section 3. By Proposition 4.1, under the assumptions of Theorem 4.1, for every fixed sequence Z^-, f^- satisfies the assumptions of Theorem 3.1. Indeed, let Z^- be fixed, and suppose that Assumption $H^{(1)}_{\text{bias}-\text{error}}(\beta)$ holds, then $f\in\mathcal{H}(\beta,L)$, and by Proposition 4.1, $f^-\in\mathcal{H}(\beta,L)$. Besides, for every $t\in\mathcal{H}(\beta,L)$, $\|t-t_m\|_\infty \le LD_m^{-\beta}$ thus $\|(f^-)_m - f^-\|_\infty \le LD_m^{-\beta}$ and f^- satisfies Assumption $H_{\text{bias}}(\beta)$. The same argument holds with Assumption $H^{(2)}_{\text{bias}-\text{error}}(\beta)$. Similarly, if f satisfies Assumption $H^{(1)}_{\nu-\text{error}}(\beta)$ or $H^{(2)}_{\nu-\text{error}}(\beta)$, then f^- satisfies H_ν. Thus we have,

Proposition 7.2. *Suppose that Assumption $H^{(1)}_{\text{bias}-\text{error}}(\beta)$ or $H^{(2)}_{\text{bias}-\text{error}}(\beta)$ holds for some $\beta\ge\beta'>3/4$. Let p_0 satisfy $(n/\log n)^\gamma \le p_0 \le M(n/\log n)^\gamma$ for some*

$$\gamma\in\Big]\frac{1}{\beta'(2\beta'+1)}, \min\Big\{\frac{1}{\beta'+1}, \frac{4\beta'+1}{3(2\beta'+1)}\Big\}\Big[. \tag{49}$$

Then

$$\mathbb{E}\big[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)\big] \le \theta_1'\Big(\frac{n}{\log n}\Big)^{-2\beta/(2\beta+1)} + \mathcal{C}_n'\mathbb{E}\big[\|\widehat{b}-b\|_{f_X}^2\big] + \mathcal{R}_n$$

with

$$\begin{aligned}\theta_1' &= (\kappa_2+2\nu\kappa_3)(M+1),\\ \mathcal{C}_n' &= 2\log n\Big[\Big(\nu\Big(\frac{n}{\log n}\Big)^{-\frac{1}{2\beta'+1}} + K^2M\Big)^2\Big(36C_0^2\Big(\frac{n}{\log n}\Big)^{\frac{2}{2\beta'+1}-2\beta'\gamma}\\ &\quad + (18MK^2)^2\Big(\frac{n}{\log n}\Big)^{\frac{2-4\beta'}{2\beta'+1}}\Big) + (12K^3)^2\Big(\frac{n}{\log n}\Big)^{3\gamma-\frac{4\beta'+1}{2\beta'+1}}\Big],\\ \mathcal{R}_n &= 2\Big[\nu+K^2M\Big(\frac{n}{\log n}\Big)^{1/(2\beta'+1)}\Big]^2\exp\Big(-\frac{C_0}{14K^2}n^{1-\gamma(1+\beta')}\Big) + \frac{\theta_2}{n}.\end{aligned}$$

Moreover, $\lim_{n\to+\infty}\mathcal{C}_n' = 0$ and $\mathcal{R}_n \le \kappa_1'/n$ for some constant κ_1', which depends on (M,K,β',ν).

Let us define the following sets, which depend on the sequence Z^-:

$$A_1^- = \Big\{C_0 p_0^{-\beta} \le \frac{\nu^-}{6}\Big\}, \quad A_2^- = \Big\{12K^2 \frac{p_0}{\sqrt{n}} \le \nu^-\Big\}, \quad A_3^- = \Big\{\frac{18MK^2}{\nu^-} \le \Big(\frac{n}{\log n}\Big)^{-\frac{2\beta}{2\beta+1}}\Big\}.$$

The proof of Proposition 7.2 comes out of the following decomposition:

$$\begin{aligned}\mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)] &= \mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)\mathbf{1}_{A_1^- \cap A_3^-}] + \mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)\mathbf{1}_{(A_1^-)^c \cup (A_3^-)^c}] \\ &\le \mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)\mathbf{1}_{\{\widehat{\nu}_n^- \ge \nu^-/2\}\cap A_3^-}] + \mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)\mathbf{1}_{\{\widehat{\nu}_n^- < \nu^-/2\}\cap A_1^-}] \\ &\quad + \mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)\mathbf{1}_{(A_1^-)^c \cup (A_3^-)^c}]. \end{aligned} \tag{50}$$

Then, the following claims provide an upper bound for each term in the right-hand side of (50). There exists an integer n_0, which depends on (σ^2, β), such that for every $n \ge n_0$,

Claim 5.

$$\begin{aligned}&\mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)\mathbf{1}_{(A_1^-)^c \cup (A_3^-)^c}] \\ &\quad \le 2\log n\Big(\nu + K^2 M\Big(\frac{n}{\log n}\Big)^{\frac{1}{2\beta'+1}}\Big)^2 \times \Big[\Big(\frac{6C_0}{p_0^\beta}\Big)^2 + (18MK^2)^2\Big(\frac{n}{\log n}\Big)^{-\frac{4\beta}{2\beta+1}}\Big]\mathbb{E}[\|\widehat{b} - b\|_{f_X}^2].\end{aligned}$$

Claim 6.

$$\mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)\mathbf{1}_{\{\widehat{\nu}_n^- < \nu^-/2\}\cap A_1^-}] \le 2\Big(\nu + MK^2\Big(\frac{n}{\log n}\Big)^{\frac{1}{2\beta'+1}}\Big)^2 \exp\Big(-\frac{C_0}{14K^2}\frac{n}{p_0^{1+\beta}}\Big).$$

Claim 7.

$$\begin{aligned}&\mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)\mathbf{1}_{\{\widehat{\nu}_n^- \ge \nu^-/2\}\cap A_3^-}] \\ &\quad \le \Big\{\kappa_2 + \kappa_3\Big(2\nu + 2K^2 p_0 \exp\Big(-\frac{\sqrt{n}}{38}\Big) + 2(12K^3)^2 \log n \frac{p_0^3}{n}\mathbb{E}[\|\widehat{b} - b\|_{f_X}^2]\Big)\Big\} \\ &\quad\quad \times (M+1)\Big(\frac{n}{\log n}\Big)^{-\frac{2\beta}{2\beta+1}} + \frac{\theta_2}{n}.\end{aligned}$$

These claims lead to the proof of Proposition 7.2. Indeed,

$$\begin{aligned}\mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)] &\le (\kappa_2 + 2\nu\kappa_3)(M+1)\Big(\frac{n}{\log n}\Big)^{-\frac{2\beta}{2\beta+1}} \\ &\quad + 2\log n\Big[\Big(\nu + K^2 M\Big(\frac{n}{\log n}\Big)^{\frac{1}{2\beta'+1}}\Big)^2\Big(\frac{36C_0^2}{p_0^{2\beta}} + (18MK^2)^2\Big(\frac{n}{\log n}\Big)^{-\frac{4\beta}{2\beta+1}}\Big) \\ &\quad + (12K^3)^2\frac{p_0^3}{n}(M+1)\Big(\frac{n}{\log n}\Big)^{-\frac{2\beta}{2\beta+1}}\Big]\mathbb{E}[\|\widehat{b} - b\|_{f_X}^2] \\ &\quad + 2\Big(\nu + K^2 M\Big(\frac{n}{\log n}\Big)^{\frac{1}{2\beta'+1}}\Big)^2 \exp\Big(-\frac{\sqrt{n}}{7}\Big) + \frac{\theta_2}{n}.\end{aligned}$$

By the conditions $\beta \ge \beta'$ and $(n/\log)^\gamma \le p_0 \le M(n\log n)^\gamma$, we have

$$\begin{aligned}&\Big(\nu + K^2 M\Big(\frac{n}{\log n}\Big)^{\frac{1}{2\beta'+1}}\Big)^2\Big(\frac{36C_0^2}{p_0^{2\beta}} + (18MK^2)^2\Big(\frac{n}{\log n}\Big)^{-\frac{4\beta}{2\beta+1}}\Big) + (12K^3)^2\frac{p_0^3}{n}(M+1)\Big(\frac{n}{\log n}\Big)^{-\frac{2\beta}{2\beta+1}} \\ &\quad \le \Big(\nu\Big(\frac{n}{\log n}\Big)^{-\frac{1}{2\beta'+1}} + K^2 M\Big)^2\Big(36C_0^2\Big(\frac{n}{\log n}\Big)^{\frac{2}{2\beta'+1} - 2\beta'\gamma} + \Big(18MK^2\Big)^2\Big(\frac{n}{\log n}\Big)^{\frac{2-4\beta'}{2\beta'+1}}\Big) \\ &\quad\quad + (12K^3)^2\Big(\frac{n}{\log n}\Big)^{3\gamma - \frac{4\beta'+1}{2\beta'+1}} = \mathcal{C}_n'.\end{aligned}$$

According to assumption (49), $2/(2\beta'+1)-2\beta'\gamma<0$ and $3\gamma-(4\beta'+1)/(2\beta'+1)<0$, hence $\lim_{n\to+\infty}\mathcal{C}'_n=0$. Moreover,

$$\exp\Big(-\frac{C_0}{14K^2}\frac{n}{p_0^{1+\beta'}}\Big)=\exp\Big(-\frac{C_0}{14K^2}n^{1-\gamma(1+\beta')}\Big)$$

and $1-\gamma(1+\beta')>0$, which entails that $\mathcal{R}_n\le\kappa'_1$ for some constant κ'_1.

Let us prove these Claims. First of all, the probabilities $P[(A_1^-)^c]$, $P[(A_2^-)^c]$, and $P[(A_3^-)^c]$ are upper bounded via the following lemma.

Lemma 7.1. *Consider a sequence (α_n) of positive numbers such that $\alpha_n=o(1/\sqrt{\log n})$. Then for every $n\in\mathbb{N}$ such that*

$$(\mathbf{C})\qquad \frac{2}{\sqrt{\log n}}+\sigma^2\alpha_n^2\log n\le\frac{1}{2},$$

where $\sigma^2=\mathbb{E}[\epsilon_1^2]$, we have

$$P[\nu^-\le\alpha_n]\le 2\log n\,\alpha_n^2\mathbb{E}[\|\widehat{b}-b\|^2_{f_X}].$$

Hence there exists an integer n_0, which depends on (σ^2,β,C_0,K), such that, for every $n\ge n_0$,

$$P[(A_1^-)^c]=P[\nu^-<6C_0p_0^{-\beta}]\le 2\log n\Big(\frac{6C_0}{p_0^\beta}\Big)^2\mathbb{E}[\|\widehat{b}-b\|^2_{f_X}], \tag{51}$$

$$P[(A_2^-)^c]=P\Big[\nu^-\le 12K^2\frac{p_0}{\sqrt{n}}\Big]\le 2\log n(12K^2)^2\frac{p_0^2}{n}\mathbb{E}[\|\widehat{b}-b\|^2_{f_X}], \tag{52}$$

and

$$P[(A_3^-)^c]=P\Big[\nu^-<18MK^2\Big(\frac{n}{\log n}\Big)^{-\frac{2\beta}{2\beta+1}}\Big]\le 2\log n(18MK^2)^2\Big(\frac{n}{\log n}\Big)^{-\frac{4\beta}{2\beta+1}}\mathbb{E}[\|\widehat{b}-b\|^2_{f_X}]. \tag{53}$$

Proof of Lemma 7.1. Given Z^-, ϵ_1 and $(b-\widehat{b})(X_1)$ are independent, which entails

$$\mathbb{E}[\widehat{\epsilon}_1^2\mid Z^-]=\mathbb{E}[\epsilon_1^2\mid Z^-]+\mathbb{E}[(b-\widehat{b})^2(X_1)\mid Z^-]+2\mathbb{E}[\epsilon_1(b-\widehat{b})(X_1)\mid Z^-].$$

Moreover, $\mathbb{E}[\epsilon_1\mid Z^-]=0$, thus

$$\mathbb{E}[\widehat{\epsilon}_1^2\mid Z^-]=\sigma^2+\|b-\widehat{b}\|^2_{f_X}.$$

Then for every $A_n>0$,

$$\int_{|y|>A_n}f^-(y)\,dy\le\frac{1}{A_n^2}\int_{|y|>A_n}y^2f^-(y)\,dy\le\frac{1}{A_n^2}\big(\sigma^2+\|b-\widehat{b}\|^2_{f_X}\big),$$

which entails

$$\int_{|y|\le A_n}f^-(y)\,dy\ge 1-\frac{\sigma^2+\|b-\widehat{b}\|^2_{f_X}}{A_n^2}.$$

On the other hand, $\int_{|y|\le A_n}f^-(y)\,dy\le 2\nu^-A_n$ by definition of ν^-. Hence,

$$\nu^-\ge\frac{1}{2A_n}\Big(1-\frac{\sigma^2+\|b-\widehat{b}\|^2_{f_X}}{A_n^2}\Big)$$

for every $A_n > 0$. Thus,

$$P[\nu^- \le \alpha_n] \le P\Big[1 - \frac{\sigma^2 + \|b - \widehat{b}\|_{f_X}^2}{A_n^2} \le 2A_n\alpha_n\Big] = P\Big[1 - \Big(2A_n\alpha_n + \frac{\sigma^2}{A_n^2}\Big) \le \frac{\|b-\widehat{b}\|_{f_X}^2}{A_n^2}\Big].$$

Let us consider $A_n = 1/(\alpha_n\sqrt{\log n})$, then condition (C) gives

$$P[\nu^- \le \alpha_n] \le P\Big[1 - \Big(\frac{2}{\sqrt{\log n}} + \sigma^2\alpha_n^2 \log n\Big) \le \|b - \widehat{b}\|_{f_X}^2 \log n\alpha_n^2\Big]$$
$$\le P\Big[\frac{1}{2} \le \|b - \widehat{b}\|_{f_X}^2 \log n\alpha_n^2\Big] \le 2\log n\alpha_n^2 \mathbb{E}[\|b - \widehat{b}\|_{f_X}^2].$$

□

Proof of Claim 5. According to Claim 4 in Section 6,

$$(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0) \le (\nu^- + K^2 \max_{m=1,\dots,N_n} D_m)^2 \quad \text{a.s.}$$

and $\nu^- \le \nu$. Besides, by assumption, $\max_{m=1,\dots,N_n} D_m \le M(n/\log n)^{1/3}$,

$$\mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)\mathbf{1}_{(A_1^-)^c \cup (A_3^-)^c}] \le \Big(\nu + K^2 M\Big(\frac{n}{\log n}\Big)^{\frac{1}{2\beta'+1}}\Big)^2 \big(P[(A_1^-)^c] + P[(A_3^-)^c]\big),$$

and inequalities (51) and (53) end the proof of Claim 5. □

Proof of Claim 6. For every Z^-,

$$\mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)\mathbf{1}_{\{\widehat{\nu}_n^- < \nu^-/2\}} \mid Z^-]\mathbf{1}_{A_1^-} \le (\nu^- + K^2 \max_{m=1,\dots,N_n} D_m)^2 P\Big[\widehat{\nu}_n^- < \frac{\nu^-}{2} \mid Z^-\Big]\mathbf{1}_{A_1^-}$$
$$\le 2\Big(\nu + MK^2\Big(\frac{n}{\log n}\Big)^{\frac{1}{2\beta'+1}}\Big)^2 \exp\Big(-\frac{n\nu^-}{84K^2p_0}\Big)\mathbf{1}_{A_1^-}$$
$$\le 2\Big(\nu + MK^2\Big(\frac{n}{\log n}\Big)^{\frac{1}{2\beta'+1}})^2 \exp\Big(-\frac{C_0}{14K^2}\frac{n}{p_0^{1+\beta}}\Big),$$

since $\nu^- \ge 6C_0p_0^{-\beta}$ on A_1^-.

Proof of Claim 7. According to Claim 3 in Section 6,

$$\mathbb{E}\big[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)\mathbf{1}_{\{\widehat{\nu}_n^- \ge \nu^-/2\}} \mid Z^-\big]\mathbf{1}_{A_3^-} \le \mathbb{E}[\max(\kappa_2, \widehat{\nu}_n^-\kappa_3) \mid Z^-]\mathbf{1}_{A_3^-}(M+1)\Big(\frac{n}{\log n}\Big)^{-\frac{2\beta}{2\beta+1}} + \frac{\theta_2}{n}$$
$$\le (\kappa_2 + \kappa_3\mathbb{E}[\widehat{\nu}_n^- \mid Z^-])(M+1)\Big(\frac{n}{\log n}\Big)^{-\frac{2\beta}{2\beta+1}} + \frac{\theta_2}{n},$$

which entails that

$$\mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)\mathbf{1}_{\{\widehat{\nu}_n^- \ge \nu^-/2\}\cap A_3^-}] \le (\kappa_2 + \kappa_3\mathbb{E}[\widehat{\nu}_n^-])(M+1)\Big(\frac{n}{\log n}\Big)^{-\frac{2\beta}{2\beta+1}} + \frac{\theta_2}{n}. \tag{54}$$

Besides,

$$\mathbb{E}[\widehat{\nu}_n^- \mid Z^-] \le \mathbb{E}[\widehat{\nu}_n^-\mathbf{1}_{\widehat{\nu}_n^- \le 2\nu^-} \mid Z^-] + \mathbb{E}[\widehat{\nu}_n^-\mathbf{1}_{\{\widehat{\nu}_n^- > 2\nu^-\}} \mid Z^-]\mathbf{1}_{A_2^-} + \mathbb{E}[\widehat{\nu}_n^- \mid Z^-]\mathbf{1}_{(A_2^-)^c}.$$

According to inequality (46), $\widehat{\nu}_n^- = \|\widehat{g}_{m_0}^{(1)}\|_\infty \le K^2p_0$, thus

$$\mathbb{E}[\widehat{\nu}_n^- \mid Z^-] \le 2\nu^- + K^2p_0\exp\Big(-\frac{n\nu^-}{456K^2p_0}\Big) + K^2p_0\mathbf{1}_{(A_2^-)^c}.$$

On A_2^-, $\exp(-n\nu^-/456K^2p_0) \le \exp(-\sqrt{n}/38)$ a.s., so

$$\mathbb{E}[\widehat{\nu}_n^-] \le 2\nu + K^2p_0\Big(\exp\Big(-\frac{\sqrt{n}}{38}\Big) + P[(A_2^-)^c]\Big)$$

and with (52),

$$\mathbb{E}[\widehat{\nu_n^-}] \le 2\nu + K^2 p_0 \exp\Big(-\frac{\sqrt{n}}{38}\Big) + 2\log n(12K^3)^2\frac{p_0^3}{n}\mathbb{E}[\|\widehat{b}-b\|_{f_X}^2]. \tag{55}$$

Then inequalities (54) and (55) provide the proof of Claim 7. □

8. ADDITIONAL PROOFS

8.1. Proof of Proposition 4.1

(1) Let $x \in \mathbb{R}$,

$$|f^-(x)| \le \int_0^1 |f(x-(b-\widehat{b})(y))| f_X(y)\,dy \le \nu \int_0^1 f_X(y)\,dy = \nu \quad a.s.$$

(2) Suppose that $f \in \mathcal{H}(\beta, L)$ and $\beta = r + \alpha$ with $\alpha \in]0,1]$. We have,

$$(f^-)^{(r)}(x) = \frac{\partial^r}{\partial x^r}\Big(\int_0^1 f(x-(b-\widehat{b})(y))f_X(y)\,dy\Big)$$
$$= \int_0^1 \frac{\partial^r}{\partial x^r}(f(x-(b-\widehat{b})(y)))f_X(y)\,dy = \int_0^1 f^{(r)}(x-(b-\widehat{b})(y))f_X(y)\,dy.$$

Hence, for every $x, x' \in \mathbb{R}$,

$$\big|(f^-)^{(r)}(x) - (f^-)^{(r)}(x')\big| \le \int_0^1 |f^{(r)}(x-(b-\widehat{b})(y)) - f^{(r)}(x'-(b-\widehat{b})(y))| f_X(y)\,dy$$
$$\le \int_0^1 L^2|x-x'|^\alpha f_X(y)\,dy = L^2|x-y|^\alpha,$$

which proves that $f^- \in \mathcal{H}(\beta, L)$.

(3) First of all, for every $u \in \mathbb{R}$, the Fourier transform of f^- is

$$(f^-)^*(u) = \int_{x\in\mathbb{R}} f^-(x)e^{-iux}\,dx = \int_{x\in\mathbb{R}}\int_{y=0}^1 f(x-(b-\widehat{b})(y))f_X(y)e^{-iux}\,dx\,dy.$$

Set $z = x - (b-\widehat{b})(y)$, then

$$(f^-)^*(u) = \int_{y=0}^1\int_{z\in\mathbb{R}} f(z)e^{-iuz}e^{-iu(b-\widehat{b})(y)}\,dz f_X(y)\,dy = f^*(u)\int_{y=0}^1 e^{-iu(b-\widehat{b})(y)} f_X(y)\,dy.$$

Hence,

$$|(f^-)^*(u)| \le |f^*(u)| \int_{y=0}^1 |e^{-iu(b-\widehat{b})(y)}| f_X(y)\,dy = |f^*(u)|.$$

Then, if $f \in W(\beta, L)$,

$$\frac{1}{2\pi}\int_{u\in\mathbb{R}} |(f^-)^*(u)|^2 u^{2\beta+1}\,du \le \frac{1}{2\pi}\int_{u\in\mathbb{R}} |f^*(u)|^2 u^{2\beta+1}\,du \le L^2,$$

so $f^- \in W(\beta, L)$.

8.2. Proof of Proposition 2.1 and Proposition 3.2 (1)

Let us prove Proposition 2.1.

(1) A simple calculus proves that the Fourier transform of $\mathbf{1}_{[-\pi,\pi]}$ is $2\pi\phi$, then for every $u \in \mathbb{R}$,

$$\phi^*(u) = \int_{\mathbb{R}} \phi(y)e^{-iuy}\,dy = \frac{1}{2\pi}\int_{\mathbb{R}} \mathbf{1}^*_{[-\pi,\pi]}(y)e^{iuy}\,dy = \frac{1}{2\pi}\mathbf{1}_{[-\pi,\pi]}(-u) = \frac{1}{2\pi}\mathbf{1}_{[-\pi,\pi]}(u).$$

Hence, for every (m,k),

$$\phi^*_{m,k}(u) = (1/\sqrt{m})e^{-iku/m}\mathbf{1}_{[-\pi m,\pi m]}(u). \tag{56}$$

Then, let $m > 0$ and $k, l \in \mathbb{Z}$, according to the Parseval formula, we have

$$\langle \phi_{m,k}, \phi_{m,l}\rangle = \frac{1}{2\pi}\langle \phi^*_{m,k}, \phi^*_{m,l}\rangle = \frac{1}{2\pi m}\int_{-\pi m}^{\pi m} e^{-i(k-l)u/m}\,du = \mathbf{1}_{k=l}.$$

(2) First of all, we recall that for every subset S_m of $L^2(I)$, the two following properties are equivalent (see Birgé and Massart [4], Lemma 1):

$$\Big(\|t\|_\infty \le K\sqrt{D_m}\|t\|,\ \ \forall t \in S_m\Big) \quad \Leftrightarrow \quad \Big\|\sum_{k\in\mathbb{Z}} \phi^2_{m,k}\Big\|_\infty \le K^2 D_m. \tag{57}$$

So, for $t \in S_m$ and $x \in \mathbb{R}$, we prove that $|t(x)| \le \sqrt{m}\|t\|$. As $\mathrm{Supp}(t^*) \subset [-\pi m, \pi m]$, we have

$$(t(x))^2 = \left[\frac{1}{2\pi}\int_{-\pi m}^{\pi m} t^*(u)e^{ixu}\,du\right]^2 \le \frac{1}{(2\pi)^2}\left(\int_{-\pi m}^{\pi m} |t^*(u)|^2\,du\right) \times 2\pi m = m\|t\|^2$$

by Parseval's Equality, which proves the assertion (2) of Proposition 2.1.

(3) By (56), it is obvious that $S_m \subset \{t \in L^2(\mathbb{R}), \mathrm{Supp}(t^*) \subset [-\pi m, \pi m]\}$. Conversely, let $t \in L^2(\mathbb{R})$ be such that $\mathrm{Supp}(t^*) \subset [-\pi m, \pi m]$, then t^* decomposes in Fourier series as

$$t^*(u) = \Big(\sum_{k\in\mathbb{Z}} a_k e^{iku\pi/m}\Big)\mathbf{1}_{[-\pi m,\pi m]} \in \mathrm{vect}\{\phi^*_{m,k}, k \in \mathbb{Z}\}$$

for some numbers $(a_k)_{k\in\mathbb{Z}}$. Thus $t \in S_m$. Hence $S_m = \{t \in L^2(\mathbb{R}), \mathrm{Supp}(t^*) \subset [-\pi m, \pi m]\}$. Then it is obvious that $S_m \subset S_{m'}$ for every $m \le m'$.

• Let us prove (1) of Proposition 3.2. For every $h \in L^2(\mathbb{R})$,

$$h_m = \arg\min_{t\in A_m} \|h - t\|^2 = \arg\min_{\mathrm{Supp}(t^*)\subset[-\pi m,\pi m]} \frac{1}{2\pi}\|h^* - t^*\| = \frac{1}{2\pi}(h^*\mathbf{1}_{[-\pi m,\pi m]})^*.$$

Suppose that $h \in W(\beta + 1/2, L)$, let $x \in \mathbb{R}$,

$$\begin{aligned}(h - h_m)^2(x) &= \left[\frac{1}{2\pi}\int_{\mathbb{R}} (h^* - h^*_m)(u)e^{iux}\,du\right]^2 = \left[\frac{1}{2\pi}\int_{|u|>\pi m} h^*(u)e^{iux}\,du\right]^2 \\ &\le \frac{1}{(2\pi)^2}\int_{|u|>\pi m} |h^*(u)|^2|u|^{2\beta+1}\,du \times \int_{|u|>\pi m} \frac{1}{|u|^{2\beta+1}}\,du \\ &\le \frac{L^2}{2\beta\pi^{2\beta+1}}m^{-2\beta}. \qquad \square\end{aligned}$$

8.3. Proof of Proposition 2.2 and Proposition 3.2 (2)

Let us prove Proposition 2.2. For every $j \in \mathbb{N}$, $x \in \mathbb{R}$,

$$|\{k \in \Gamma(j) \colon \psi_{j,k}(x) \neq 0\}| \leq |\{k \colon -B \leq 2^j x - k \leq B\}| \leq 2B + 1.$$

Thus, for every $m \in \mathbb{N}^*$, $t \in B_m$ and $x \in [-1, 1]$, we have,

$$\begin{aligned}(t(x))^2 &= \Big(\sum_{k \in \Gamma'(0)} \langle \varphi_k, t \rangle \varphi_k(x) + \sum_{j=0}^{m-1} \sum_{k \in \Gamma(j)} \langle \psi_{j,k}, t \rangle \psi_{j,k}(x) \Big)^2 \\ &\leq \Big(\sum_{k \in \Gamma'(0)} \langle \varphi_k, t \rangle^2 + \sum_{j=0}^{m-1} \sum_{k \in \Gamma(j)} \langle \psi_{j,k}, t \rangle^2 \Big) \times \Big(\sum_{k \in \Gamma'(0)} \varphi_k^2(x) + \sum_{j=0}^{m-1} \sum_{k \in \Gamma(j)} \psi_{j,k}^2(x) \Big) \\ &\leq \|t\|^2 \times (2B+1) \Big(\|\varphi\|_\infty^2 + \sum_{j=0}^{m-1} 2^j \|\psi\|_\infty^2 \Big) \leq K^2 \|t\|^2 2^m,\end{aligned}$$

where K depends only on the structure of the mother and father wavelets. According to (57), this proves the result of Proposition 2.2.

• Assertion (2) of Proposition 3.2 follows from Proposition 4 (Section 9, Chapter 2) in Meyer [16]. The result stated there is more general (for Besov spaces and a L^q-norm), and we only recall it in the form we require. Let $h \in L^2(\mathbb{R})$, then

$$h(x) = \sum_{k \in \Gamma'(0)} \langle h, \varphi_k \rangle \varphi_k + \sum_{j \geq 0} \sum_{k \in \Gamma(j)} \langle h, \psi_{j,k} \rangle \psi_{j,k}$$

and

$$h_m(x) = \sum_{k \in \Gamma'(0)} \langle h, \varphi_k \rangle \varphi_k + \sum_{j=0}^{m-1} \sum_{k \in \Gamma(j)} \langle h, \psi_{j,k} \rangle \psi_{j,k}.$$

On the one hand, by Meyer [16], if $h \in \mathcal{H}(\beta, L) = \mathcal{B}_\infty^{\beta,\infty}(L)$ (see Meyer [16] or DeVore and Lorentz [7] for the definition of Besov spaces),

$$\sup_{j \geq 0} 2^{j\beta} \Big\| \sum_{k \in \Gamma(j)} \langle h, \psi_{j,k} \rangle \psi_{j,k} \Big\|_\infty = \| \, |h| \, \| < +\infty.$$

Moreover, there exists a constant, which only depends of ψ and φ, such that $\| \, |h| \, \| \leq CL$ for every $h \in \mathcal{H}(\beta, L)$. Thus, for every $m \geq 1$,

$$\|h - h_m\|_\infty = \Big\| \sum_{j \geq m} \sum_{k \in \Gamma(j)} \langle \psi_{j,k}, h \rangle \psi_{j,k} \Big\|_\infty \leq \sum_{j \geq m} C \| \, |h| \, \| 2^{-j\beta} \leq CL \frac{2^{-m\beta}}{1 - 2^{-\beta}} = \frac{K'(\beta) L}{2^m}. \qquad \square$$

9. APPENDIX : DEVIATION INEQUALITIES FOR EMPIRICAL PROCESSES

The following inequality, called Bernstein's Inequality, is stated in Birgé and Massart [4] (Lemma 8, p. 366).

Theorem 9.1. *Let $(X_1, \ldots, X_n)$ be independent random variables. Suppose that:*

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i^2] \leq v, \qquad \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(X_i)_+^l] \leq \frac{l!}{2} \, v \, c^{l-2}$$

for every $l \geq 2$. Let $S = \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X_i]$. Then, for every $\epsilon > 0$:

$$P[S \geq \epsilon] \leq \exp\Big(- \frac{n\epsilon^2}{2(v + c\epsilon)} \Big), \qquad P[|S| \geq \epsilon] \leq 2 \exp\Big(- \frac{n\epsilon^2}{2(v + c\epsilon)} \Big).$$

The following deviation inequality called Talagrand's Inequality is originally due to works by Talagrand [18] and is stated in this form by Klein and Rio [10].

Theorem 9.2. *Let $(X_1, \ldots, X_n)$ be i.i.d., $\mathcal{F}$ a set of functions, and*

$$Z = \sup_{t \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \big(t(X_i) - \mathbb{E}[t(X_i)]\big).$$

Let $\mathbb{H}$, v, and b be such that

$$\mathbb{E}[|Z|] \leq \mathbb{H}, \qquad \sup_{t \in \mathcal{F}} \mathrm{Var}(t(X_i)) \leq v, \qquad \sup_{t \in \mathcal{F}} \|t\|_\infty \leq b.$$

Then for every $\lambda > 0$

$$P\big[|Z| > \mathbb{H} + \lambda\big] \leq \exp\Big(-\frac{n\lambda^2}{2(v + 4\mathbb{H}b + 3b\lambda)}\Big).$$

REFERENCES

1. M. G. Akritas and I. Van Keilegom, "Non-Parametric Estimation of the Residuals Distribution", Scand. J. Statist. **28** (3), 549–567 (2001).
2. Y. Baraud, "Model Selection for Regression on a Fixed Design", Probab. Theory Related Fields **117** (6), 467–493 (2000).
3. Y. Baraud, "Model Selection for Regression on a Random Design", ESAIM Probab. Statist. **6**, 127–146 (2002).
4. L. Birgé and P. Massart, "Minimum Contrast Estimators on Sieves: Exponential Bounds and Rates of Convergence", Bernoulli **4** (3), 329–475 (1998).
5. C. Butucea, "Exact Adaptive Pointwise Estimation on Sobolev Classes of Densities", ESAIM Probab. Statist. **5**, 1–31 (2001).
6. F. Compte, J. Dedecker, and M. L. Taupin, "Adaptive Density Deconvolution with Dependent Inputs", Math. Methods Statist. **17** (2), 87–112 (2008).
7. R. A. DeVore and G. G. Lorentz, *Constructive Approximation* (Springer, Berlin, 1993).
8. S. Efromovich, "Estimation of the Density of Regression Errors", Ann. Statist. **33** (5), 2194–2227 (2005).
9. S. Kiwitt, E.-R. Nagel, and N. Neumeyer, "Empirical Likelihood for the Error Distribution in Nonparametric Regression Models", Math. Methods Statist. **17** (3), 241–260 (2008).
10. T. Klein and E. Rio, "Concentration around the Mean of Empirical Processes", Ann. Probab. **33** (3), 1060–1077 (2005).
11. C. Laurent, C. Ludena, and C. Prieur, "Adaptive Estimation of Linear Functionals by Model Selection", Electron J. Statist., No. 2, 993–1020 (2008).
12. O. V. Lepski, "Asymptotically Minimax Adaptive Estimation. I. Upper Bounds. Optimally Adaptive Estimates", Theory Probab. Appl. **36** (4), 682–697 (1991).
13. O. V. Lepski and V. G. Spokoiny, "Optimal Pointwise Adaptive Methods in Nonparametric Estimation," Ann. Statist. **25** (6), 2512–2546 (1997).
14. O. V. Lepski and A. Goldenshluger, "Structural Adaptation via l_p-Norm Oracle Inequalities", Probab. Theory Related Fields **126** (1–2), 47–71 (2009).
15. P. Massart, *Concentration Inequalities and Model Selection. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23*, in *Lecture Notes in Math.* (Springer, Berlin, 2007).
16. Y. Meyer, *Ondelettes et operateurs* (Hermann, Paris, 1990).
17. S. Plancade, "Non-Parametric Estimation of the Density of the Regression Noise", C. R. Acad. Sci. Paris **346** (7–8), 461–466 (2008).
18. M. Talagrand, "New Concentration Inequalities in Product Spaces", Invent. Math. **126** (3), 505–563 (1996).
19. A. B. Tsybakov, *Introduction à l'estimation non paramétrique* (Springer, Berlin, 2004).