ADAPTIVE NONPARAMETRIC REGRESSION ESTIMATION IN PRESENCE OF RIGHT CENSORING.

E. BRUNEL[1] AND F. COMTE[2]

Abstract. In this paper, we consider the problem of estimating a regression function when the outcome is censored. Two strategies of estimation are proposed: a two-step strategy where the ratio of two projection estimators is used to estimate the regression function; a direct strategy based on a standard mean-square contrast for censored data. For both estimators, non-asymptotic bounds for the integrated mean-square risk are provided and data-driven model selection is performed. In most cases, asymptotically optimal minimax rates of convergence are obtained, when the regression function belongs to a class of Besov functions.

January 31, 2006

AMS Classification (2001): 62G08; 62N01.

Keywords and phrases. Adaptive estimation. Minimax rate. Regression function. Right-Censoring. Nonparametric penalized contrast estimator.

1. Introduction

It is most natural to use regression techniques to explore the relationship between a response and covariates. A vast literature deals with nonparametric methods for completely observed data and provide a very flexible tool in a prospective study. But, some of the responses, typically lifetimes and/or covariates may be censored. Our aim is to introduce new projection estimators of the regression function when the response is censored while the covariate is a multivariate completely observed vector.

First, estimators of the regression function based on Nadaraya-Watson [24] kernel-type estimators were studied and improved by local linear regression smoothers in Fan [13] or by adaptive mean-square methods in Baraud [1]. The optimal rates for such estimators have been established by Stone [28] and are known to be of order $n^{-k/(k+d)}$ when an n sample is available, d-dimensional covariates are considered and the regression function is k times differentiable.

However, in the setting of censored data, the previous methods are not directly applicable. Linear models have been first mainly considered (see Miller [23], Buckley and James [6], Kool et al. [20], Zhou [32], among others). See also Heuchenne and Van Keilegom [16] for a nonlinear semiparametric regression model with censored data. But the flexibility of nonparametric

[1] IUT de Paris V et Laboratoire MAP5, UMR CNRS 8145. email: elodie.brunel@univ-paris5.fr.

[2] Université Paris V, MAP5, UMR CNRS 8145. Email: fabienne.comte@univ-paris5.fr.

regression provides an interesting tool when a general relationship between covariate and response is first to be explored. For instance, Dabrowska [8] studied nonparametric estimators of the conditional survival function. Zheng [31] generalized Stone's regression estimators to the censored case. The consistency of the resulting estimator has also been studied in Györfi *et al.* [14], Chapter 26. Fan and Gijbels [13] studied a local linear smoothers adaptive to the data and in particular to the scarcity at the end of the interval. Park [27] has extended a procedure suggested in Gross and Lai [15] to a general nonparametric model in presence of left-truncation and right-censoring, by using B-splines developments. Recently, Kohler *et al.* [19] have proposed an adaptive mean-square estimator built with polynomial splines. Asymptotic optimal rates of convergence are given up to a logarithmic factor for regression functions belonging to Hölder-type spaces. We consider more general functional spaces (anisotropic Besov spaces).

Our nonparametric method is adaptive in the sense that the dimension of the approximation spaces can be relevantly chosen without knowing the regularity of the unknown function to estimate. A data-driven criterion of selection is introduced for this purpose and then, the estimator achieves automatically the optimal rate without any loss. Two methods are investigated in this paper. First, the projection methods provide a quotient estimator which is easy to compute and can reach the optimal rate in many cases. On the other hand, the mean-square regression methods developed in Baraud [1] are suitable for the censored case: they give a direct adaptive estimator which automatically reaches the optimal rate.

A particular attention is paid to the additive model even if multivariate regression functions can be estimated with both proposed estimators.

The plan of the paper is the following. In Section 2, we first describe the regression model and the censoring mechanism. Then, the approximation spaces with their key properties are introduced and the orders of the bias terms are given. More precisely the $\mathbb{L}_2$-distance between a function belonging to a Besov space and its orthogonal projection on a given approximation space of the collection is evaluated. Section 3 presents the two strategies of estimation: the quotient estimator is studied in term of its Mean Integrated Squared Error (MISE) in Section 4 and the mean-square estimator, whose MISE is studied in Section 5. Most proofs are gathered in Section 6.

2. Preliminaries on the model and on the collection of approximation spaces

2.1. **Nonparametric regression model with censored data.** As mentionned by Gross and Lai [15], when right-censoring is present, functionals of the survival function cannot be estimated on the complete support.

Suppose that $\vec{X}_i$ is a d-dimensional covariate in a compact set, without loss of generality we would assume that $\vec{X}_i$ is taking value into $[0,1]^d$. Let $(\vec{X}_1, Y_1)$, $(\vec{X}_2, Y_2)$, $\ldots$, $(\vec{X}_n, Y_n)$ be independent identically distributed random variables.

We consider a setting analogous to Kohler *et al.*[19] corresponding to a fixed time T for collecting the data. Therefore, the variables before censoring are denoted by $Y_{i,T} = Y_i \wedge T$, where $a \wedge b$ denotes the infimum of a and b. Then the regression model is

$$\mathbb{E}(Y_{i,T}|\vec{X}_i) = r_T(\vec{X}_i),\ i = 1, \dots, n\,.$$

Next, let $C_1, C_2, \dots, C_n$ be n censoring times independent of the $(\vec{X}_i, Y_i)$. Then, the censoring mechanism is as follows: the $\vec{X}_i's$ and the couples (Z_i, δ_i)'s are observed where

$$Z_i = Y_{i,T} \wedge C_i, \quad \delta_i = \mathbf{1}_{\{Y_{i,T} \le C_i\}}$$

δ_i indicates if the observed time Z_i is a lifetime or a censoring time both occuring in the interval [0, T]. Of course, it is often mentionned that the function of interest would be r in the regression model $\mathbb{E}(Y_i|\vec{X}_i) = r(\vec{X}_i)$, but only its biased version r_T is reachable.

Now, set $G(.)$ the cumulative distribution function (c.d.f.) of the C_i's and by F_Y the marginal c.d.f. of the Y_i's with $\bar{F}_Y = 1 - F_Y$ and $\bar{G} = 1 - G$ the associated survival functions. We suppose moreover:

$(\mathcal{A})$ The distribution functions of the Y_i's and C_i's are $\mathbb{R}^+$-supported.

Under $(\mathcal{A})$, the three following conditions are immediately satisfied:

$$\mathbb{P}(Y_i \ge T) = \mathbb{P}(Y_{i,T} = T) > 0 \tag{2.1}$$

Moreover, we suppose that $\mathbb{P}(C_i > T) > 0$ which is satisfied for most well-known parametric survival models where the C_i's are $\mathbb{R}^+$-supported. This implies

$$1 - G(Y_{i,T}) \ge 1 - G(T) := c_G,\ i = 1, \dots, n. \tag{2.2}$$

The c.d.f F_Y is upper bounded on $[0, T]$ so there exists $c_F > 0$,

$$\forall t \in [0, T],\ \ 1 - F_Y(t) \ge 1 - F_Y(T) := c_F > 0. \tag{2.3}$$

Any condition ensuring (2.2) and (2.3) can be substituted to $(\mathcal{A})$.

2.2. **Description of the approximation spaces in the univariate case.** The projection spaces $(S_m)_{m \in \mathcal{M}_n}$ are described hereafter.

[T] *Trigonometric spaces*: S_m is generated by $\{1, \sqrt{2}\cos(2\pi jx), \sqrt{2}\sin(2\pi jx) \text{ for } j = 1, \dots, m\}$, $D_m = 2m + 1$ and $\mathcal{M}_n = \{1, \dots, [n/2] - 1\}$.

[P] *Regular piecewise polynomial spaces*: S_m is generated by $m(r+1)$ polynomials, $r+1$ polynomials of degree $0, 1, \dots, r$ on each subinterval $[(j-1)/m, j/m]$, for $j = 1, \dots m$, $D_m = (r+1)m$, $m \in \mathcal{M}_n = \{1, 2, \dots, [n/(r+1)]\}$. Usual examples are the orthogonal collection in $\mathbb{L}^2([-1, 1])$ of the Legendre polynomials or the histogram basis. Dyadic collection of piecewise polynomials denoted by [DP] correspond to dyadic subdivisions with $m = 2^q$ and $D_m = (r+1)\,2^q$.

[W] *Dyadic wavelet generated spaces* with regularity r and compact support, as described in Donoho and Johnstone [12]. The generating basis is of cardinality $D_m = 2^{m+1}$ and $m \in \mathcal{M}_n =$

$\{1, 2, \ldots, [\ln(n)/2] - 1\}$ see Brunel and Comte [7] for details.

All the spaces above satisfy the same property:

$(\mathcal{H}_1)$ $(S_m)_{m\in\mathcal{M}_n}$ is a collection of finite-dimensional linear sub-spaces of $\mathbb{L}^2([0,1])$, with dimension $\dim(S_m) = D_m$ such that $D_m \leq n$, $\forall m \in \mathcal{M}_n$ and satisfying:

$$\exists \Phi_0 > 0, \forall m \in \mathcal{M}_n, \forall t \in S_m, \|t\|_\infty \leq \Phi_0 \sqrt{D_m}\|t\|. \tag{2.4}$$

where $\|t\|^2 = \int_0^1 t^2(x)dx$, for t in $\mathbb{L}^2([0,1])$.

An orthonormal basis of S_m is denoted by $(\varphi_\lambda)_{\lambda\in\Lambda_m}$ where $|\Lambda_m| = D_m$. Birgé and Massart [4] proved that Property (2.4) in the context of $(\mathcal{H}_1)$ is equivalent to

$$\exists \Phi_0 > 0, \|\sum_{\lambda\in\Lambda_m} \varphi_\lambda^2\|_\infty \leq \Phi_0^2 D_m. \tag{2.5}$$

Moreover, for the results concerning the adaptive estimators, we need the following additional assumption:

$(\mathcal{H}_2)$ $(S_m)_{m\in\mathcal{M}_n}$ is a collection of nested models, we denote by $\mathcal{S}_n$ the space belonging to the collection, such that $\forall m \in \mathcal{M}_n, S_m \subset \mathcal{S}_n$. We denote by N_n the dimension of $\mathcal{S}_n$: $\dim(\mathcal{S}_n) = N_n$ $(\forall m \in \mathcal{M}_n, D_m \leq N_n)$.

Assumption $(\mathcal{H}_1)$ is satisfied with for instance $\Phi_0 = \sqrt{2}$ for collection [T] and $\Phi_0 = \sqrt{2r+1}$ for collection [P]. Moreover, [T], [DP] and [W] satisfy $(\mathcal{H}_2)$.

2.3. **The general multivariate setting and the particular case of additive models.** Consider the general case of a regression function $r : [0,1]^d \to \mathbb{R}$ where $r(x) = r(x^{(1)}, \ldots, x^{(d)})$. In that case, models S_m are generated by the basis functions $\varphi_{\lambda_1}(x^{(1)}), \ldots, \varphi_{\lambda_d}(x^{(d)})$ for $\lambda = (\lambda_1, \ldots, \lambda_d) \in \Lambda_m = \Lambda_{m_1} \times \cdots \times \Lambda_{m_d}$. For instance for $d = 2$, the corresponding function $t \in S_m$ can be written

$$t(x,y) = \sum_{\lambda\in\Lambda_m} a_\lambda \varphi_\lambda(x,y) = \sum_{\lambda_1\in\Lambda_{m_1}} \sum_{\lambda_2\in\Lambda_{m_2}} a_{\lambda_1,\lambda_2} \varphi_{\lambda_1}(x)\, \varphi_{\lambda_2}(y),$$

with $\forall \lambda \in \Lambda_m$, $a_\lambda = a_{\lambda_1,\lambda_2} \in \mathbb{R}$ and $\varphi_\lambda(x,y) = \varphi_{\lambda_1}(x)\, \varphi_{\lambda_2}(y)$ for $(x,y) \in [0,1]^2$. Clearly, the dimension of such a *product* space $S_m = S_{m_1,\ldots,m_d}$ is the product $D_m = D_{m_1} \times \cdots \times D_{m_d}$ of the cardinalities of the Λ_{m_i}'s. It would not be realistic with standard size of sample to think of more than two or three covariates. If the underlying one-directional spaces $S_{m_1}, \ldots, S_{m_d}$ satisfy $(\mathcal{H}_1)$ and $(\mathcal{H}_2)$, then the resulting *product* space also satisfies these conditions. We also need to denote the space $\mathcal{S}_n = \mathcal{S}_{n_1,\ldots,n_d}$ with the associated spaces $\mathcal{S}_{n_i}$ be such that $\forall m = (m_1, \ldots, m_d) \in \mathcal{M}_n$, $S_{m_i} \subset \mathcal{S}_{n_i}$ for all $i = 1, \ldots, d$. We also denote by N_{n_i} the dimension of each $\mathcal{S}_{n_i}$ and as previously the resulting dimension of $\mathcal{S}_n$ is defined by $N_n = N_{n_1} \times \cdots \times N_{n_d}$.

The particular case of additive models constitutes a way to make the dimension d of the covariate higher. This amounts to consider the following multivariate regression function:

$$r_T(x) = r_T(x^{(1)}, \dots, x^{(d)}) = e_T + r_{T,1}(x^{(1)}) + \dots + r_{T,d}(x^{(d)}). \tag{2.6}$$

For identifiability, we assume moreover that $\int_{[0,1]} r_{T,i}(x^{(i)})\, dx^{(i)} = 0$. It is then possible to build estimators of $r_{T,1}, \dots, r_{T,d}$ on different spaces, when the mean-square estimator is considered. In that case, the models can be described as

$$S_m = \left\{ t(x^{(1)}, \dots, x^{(d)}) = a + \sum_{i=1}^{d} t_i(x^{(i)}),\ (a, t_1, \dots, t_d) \in \mathbb{R} \times \Pi_{i=1}^{d} S_{m^{(i)}}^{g(i)} \right\}$$

where $g(i) = 1$ if the space $S_{m_i}^1$ is chosen as a trigonometric space with dimension D_{m_i} and $g(i) = 2$ if $S_{m_i}^2$ is chosen as a piecewise polynomial space with dimension D_{m_i}, for instance. Those collections also satisfy $(\mathcal{H}_1)$ and $(\mathcal{H}_2)$ with $D_m = 1 + \sum_{i=1}^{d}(D_{m_i} - 1)$ in the inequalities (2.4) or (2.5).

2.4. **Order of the bias in Besov spaces and resulting rates.**
Given a function h (where h stands for the regression function r_T or its product $\psi = r_T f$ with the density f of the $\vec{X}_i$'s) belonging to a class of smooth functions $\mathcal{F}$, the $\mathbb{L}^2$-norm denoted by $\|.\|$ on $\mathcal{F}$ and a collection $(S_m)_{m\in\mathcal{M}_n}$ of linear subspaces of $\mathbb{L}^2(A)$ described in section 2.2 with dimension D_m, elementary approximation theory implies that $\|h - h_m\| = \inf_{t\in S_m} \|h - t\|$ where h_m is the orthogonal projection of h on S_m. The general goal of model selection is to build a collection of estimators $\hat{h}_m$ of h belonging to S_m, then to select a model $\hat{m}$ in $\mathcal{M}_n$ and to bound the quadratic risk of the resulting estimator $\tilde{h} = \hat{h}_{\hat{m}}$ with the following type of inequality:

$$\mathbb{E}\|h - \tilde{h}\|^2 \leq C \inf_{m\in\mathcal{M}_n} \left(\|h - h_m\|^2 + \frac{D_m}{n} \right). \tag{2.7}$$

Both terms in the above upper bound depends on D_m, the former being decreasing and the latter increasing. As a consequence, the infimum automatically makes the usual squared bias/variance compromise and therefore minimizes the risk. The way of illustrating this minimization problem is to put a smoothness assumption on h. This regularity assumption associated to the choice of the spaces S_m leads to a known order of the squared bias term depending on D_m and the index of regularity, and therefore to an explicit rate.

a) Univariate case.
In fact, it is well-known that for all three collections [T], [P] or [W], the $\mathbb{L}^2$ projection h_m on the linear subspace S_m achieves the best rate of approximation in $\mathbb{L}^2$ for the Besov class of functions $\mathcal{F} = \mathcal{B}_{2,\infty}^{\alpha}(A)$ with $A = [0,1]$ (see Lemma 12, Barron et al. [3]) i.e.

$$\|h - h_m\|_2 \leq C(\alpha)\, |h|_{\alpha,2}\, D_m^{-\alpha}, \quad \text{for } h \in \mathcal{B}_{2,\infty}^{\alpha}(A),$$

where $C(\alpha)$ is a constant depending on α and also on the basis. The semi-norm of h in $\mathcal{B}^{\alpha}_{2,\infty}(A)$ is denoted by $|h|_{\alpha,2} = \sup_{y>0} y^{-\alpha}\omega_{\ell}(h,y)_2 < +\infty$ with $\ell = [\alpha]+1$ and with the modulus of smoothness defined by

$$\omega_{\ell}(h,y)_2 = \sup_{0<u\leq y} \| \sum_{k=0}^{\ell} \left(\begin{array}{c} \ell \\ k \end{array} \right) (-1)^{\ell-k}\, h(x+ku)\| \, .$$

For more details, the approximation properties of these spaces can be found in DeVore and Lorentz [10]. Therefore, balancing the approximation and variance terms leads to choose m^* such that $D_{m^*} = O(n^{1/(2\alpha+1)})$ and it provides the optimal rate of order $O(n^{-2\alpha/(2\alpha+1)})$. An inequality such as (2.7) means in that context that the model selection procedure leads not only to a nonasymptotic squared bias/variance compromise but also that the adaptive estimator automatically reaches an asymptotic rate of order $O(n^{-2\alpha/(2\alpha+1)})$ which in most problems is known to be the minimax rate.

b) General multivariate case.
In the case of multivariate functions, the previous definitions of Besov spaces can be generalized in a possibly anisotropic way, i.e. we can consider $\mathcal{F} = \mathcal{B}^{\alpha}_{2,\infty}(A)$ with $A = [0,1]^d$ and $\alpha = (\alpha_1,\dots,\alpha_d)$ standing for the regularity of the function in the different directions. What is known as the isotropic case corresponds to $\alpha_1 = \cdots = \alpha_d$. In that case, general definitions of the corresponding Besov spaces $\mathcal{B}^{\alpha}_{2,\infty}(A)$ with α real and $A = [0,1]^d$ can be found in DeVore [11]. For the general anisotropic case, it is proved in Hochmuth [17] for [P] and [W] and Nikol'skiĭ [26] for [T] that the orthogonal projection h_m on $S_m = S_{m_1,\dots,m_d}$, leads to a squared bias term of order,

$$\|h - h_m\|^2 \leq C_0 \sum_{i=1}^{d} D_{m_i}^{-2\alpha_i}\, , \; \forall m \in \mathcal{M}_n. \tag{2.8}$$

for a positive constant C_0. This leads, still with (2.7), and variance order $\Pi_{i=1}^d D_{m_i}/n$, to the rate :

$$\mathbb{E}\|h - \tilde{h}\|^2 \leq O(n^{-2\bar{\alpha}/(2\bar{\alpha}+d)}) \tag{2.9}$$

where $\bar{\alpha} = d/\sum_{i=1}^d \alpha_i^{-1}$ is the harmonic average of the smoothness coefficients α_i's. Inequality (2.9) is proved in Section 6. Note that this order is achieved for $D_{m_i^*} = O(n^{1/[\alpha_i(d/\bar{\alpha}+2)]})$ and with this choice of the $D_{m_i^*}$'s, so that $D_{m^*} = \Pi_{i=1}^d D_{m_i^*}$, the bias term is of order $O(D_{m^*}^{-2\bar{\alpha}/d})$. The same order is obtained by Neumann [25], who also proves that the resulting rates are minimax.
c) Additive models.
For additive models described in section 2.3, the proof of Proposition 5 in Baraud *et al.* [2] implies that $\|r - r_m\|^2$ is of order $D_m^{-2\underline{\alpha}}$ where $\underline{\alpha} = \min(\alpha_1,\dots,\alpha_d)$ and $D_m = \min(D_{m_1},\dots,D_{m_d})$. In that case, the resulting rate remains a one-dimensional-type rate because the variance is of

order $[1+\sum_{i=1}^{d}(D_{m_i}-1)]/n$ instead of $\Pi_{i=1}^{d} D_{m_i}/n$. This produces a rate of order $n^{-2\underline{\alpha}/(2\underline{\alpha}+1)}$, as if the dimension of the space were equal to one and the function had the worst regularity.

3. Two general methods of estimation

As usual in regression problems, two different strategies are available. First, projection contrasts allow to estimate $\psi = r_T f$ and f separately; then the estimator of r_T is obtained as the quotient of those estimators. Secondly, a mean-square contrast can lead to a direct estimator of r_T. The first estimator is in some sense easier to study and in any case very easy to compute; but its theoretical properties are sometimes less satisfactory than those of the mean-square estimator. On the other hand, the latter is more difficult to implement.

3.1. **Useful tools.** In all the following, we consider the standard transformation of the observations:

$$\hat{Y}_{iG} = \frac{\delta_i Z_i}{1-\hat{G}(Z_i)} = \frac{\delta_i Z_i}{\hat{\bar{G}}(Z_i)}$$

where $\hat{G}$ is a relevant estimator of G. Moreover, we denote by

$$Y_{iG} = \frac{\delta_i Z_i}{1-G(Z_i)} = \frac{\delta_i Z_i}{\bar{G}(Z_i)}$$

the (unobserved) theoretical counterpart of the $\hat{Y}_{iG}$'s.

We propose to take the Kaplan-Meier [18] product-limit estimator $\hat{\bar{G}}$, modified in the way suggested by Lo et al. [22], and defined by

$$\hat{\bar{G}}(y) = \prod_{Z_{(i)}\le y}\left(\frac{n-i+1}{n-i+2}\right)^{1-\delta_{(i)}}. \tag{3.1}$$

Then, we have the following useful properties: $\hat{\bar{G}}(y) \ge 1/(n+1)$, $\forall y$ and compared to the standard Kaplan-Meier estimator

$$\hat{\bar{G}}_0 = \prod_{Z_{(i)}\le y}\left(\frac{n-i}{n-i+1}\right)^{1-\delta_{(i)}}$$

we have

$$\sup_{0\le y\le T}|\hat{\bar{G}}_0 - \hat{\bar{G}}| = O(n^{-1}),\ a.s. \tag{3.2}$$

for $0<T<\sup\{t\ge 0\ /\ G(t)=1\}$. The following lemma is useful to control the probability of the uniform deviation of the estimator of the survival distribution function $\hat{\bar{G}}$.

Lemma 3.1. *For all $k\in\mathbb{N}^*$, there exists a constant C_k depending on k, c_G and c_F such that*

$$\mathbb{E}\left(\sup_{y\in[0,T]}|\hat{\bar{G}}(y)-\bar{G}(y)|^{2k}\right) \le \frac{C_k}{n^k}.$$

3.2. **The projection contrasts.** Let $\psi = r_T f$. The minimization of the contrast:

$$\gamma_n(t) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} \hat{Y}_{iG} t(\vec{X}_i) \tag{3.3}$$

by setting $\hat{\psi}_m = \arg\min_{t\in S_m} \gamma_n(t)$ leads to an estimator of ψ. Then we have to determine the adequate penalization function $\mathrm{pen}(m)$ to select the relevant projection space via the standard method:

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n} \left[\gamma_n(\hat{\psi}_m) + \mathrm{pen}(m).\right] \tag{3.4}$$

Let us mention that

$$\hat{\psi}_m = \sum_{\lambda\in\Lambda_m} \hat{a}_\lambda \varphi_\lambda \text{ with } \hat{a}_\lambda = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_{iG}\varphi_\lambda(\vec{X}_i)$$

and that $\gamma_n(\hat{\psi}_m) = -\sum_{\lambda\in\Lambda_m} \hat{a}_\lambda^2$.
A standard estimator of the density f of the $\vec{X}_i$'s is obtained by minimization of the contrast:

$$\breve{\gamma}_n(t) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} t(\vec{X}_i).$$

Then, defining

$$\hat{f}_m = \arg\min_{t\in S_m} \breve{\gamma}_n(t) \tag{3.5}$$

is known to give a good estimator of f. As previously, $\hat{f}_m$ is very easy to compute since $\hat{f}_m = \sum_{\lambda\in\Lambda_m} \hat{\beta}_\lambda\varphi_\lambda$ with $\hat{\beta}_\lambda = (1/n)\sum_{i=1}^{n} \varphi_\lambda(\vec{X}_i)$. Moreover, the penalized estimator $\hat{f}_{\breve{m}}$ by setting $\breve{m}$

$$\breve{m} = \arg\min_{m\in\mathcal{M}_n} \breve{\gamma}_n(\hat{f}_m) + \kappa\Phi_0^2\frac{D_m}{n}, \tag{3.6}$$

reaches the optimal minimax rate (see Barron et al. [3]). Here κ denotes a universal constant. Let $\tilde{f} = \hat{f}_{\breve{m}}$ and $\tilde{\psi} = \hat{\psi}_{\hat{m}}$ be the penalized estimators of f and ψ defined above, then it is natural to consider the following estimator of r:

$$\tilde{r}^P = \left(\frac{\tilde{\psi}}{\tilde{f}}\right)^{(a_n)} \text{ where } \left(\frac{x}{y}\right)^{(\ell)} = \begin{cases} \ell \text{ sign}\left(\frac{x}{y}\right) & \text{if } |x| \geq \ell|y| \\ \frac{x}{y} \text{ else.} \end{cases}$$

with (a_n) a sequence of positive real numbers.

3.3. **The mean-square contrast.** The mean-square strategy leads to study the following contrast:

$$\gamma_n^{MS}(t) = \frac{1}{n}\sum_{i=1}^{n}[t(\vec{X}_i) - \hat{Y}_{iG}]^2. \tag{3.7}$$

In this context, it is useful to consider the empirical norm associated with the design

$$\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^n t^2(\vec{X}_i),\ \langle s,t\rangle_n = \frac{1}{n}\sum_{i=1}^n s(\vec{X}_i)t(\vec{X}_i).$$

Here we define

$$\hat{r}_m = \arg\min_{t\in S_m} \gamma_n^{MS}(t). \tag{3.8}$$

The function $\hat{r}_m$ may be uneasy to define but the vector $(\hat{r}_m(\vec{X}_1),\dots,\hat{r}_m(\vec{X}_n))'$ is always well defined since it is the orthogonal projection in $\mathbb{R}^n$ of the vector $(\hat{Y}_{1G},\dots,\hat{Y}_{nG})'$ onto the subspace of $\mathbb{R}^n$ defined by $\{(t(\vec{X}_1),\dots,t(\vec{X}_n))',\ t\in S_m\}$. This explains why the empirical norms are particularly suitable for the mean-square contrast.

Next, model selection is performed as usual via:

$$m^* = \arg\min_{m\in\mathcal{M}_n}\left\{\gamma_n^{MS}(\hat{r}_m) + \mathrm{pen}^{MS}(m)\right\}, \tag{3.9}$$

where we have to determine the relevant form of pen^{MS} for $\hat{r}_{m^*}$ to be an adaptive estimator of r.

4. Study of the quotient method

4.1. Estimation of $\psi = r_T f$.

The following decomposition holds

$$\begin{aligned}
\gamma_n(t)-\gamma_n(s) &= \|t-\psi\|^2 - \|s-\psi\|^2 + 2\langle t-s,\psi\rangle - \frac{2}{n}\sum_{i=1}^n \hat{Y}_{iG}(t-s)(\vec{X}_i)\\
&= \|t-\psi\|^2 - \|s-\psi\|^2 - \frac{2}{n}\sum_{i=1}^n(\hat{Y}_{iG}-Y_{iG})(t-s)(\vec{X}_i)\\
&\quad -\frac{2}{n}\sum_{i=1}^n[Y_{iG}(t-s)(\vec{X}_i) - \langle t-s,\psi\rangle]\\
&= \|t-\psi\|^2 - \|s-\psi\|^2 - 2\nu_n(t-s) - 2R_n(t-s),
\end{aligned}$$

where

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^n\left[\frac{\delta_i Z_i}{\bar{G}(Z_i)}t(\vec{X}_i) - \langle\psi,t\rangle\right],$$

is a centered empirical process specific to the projection method, and

$$R_n(t) = \frac{1}{n}\sum_{i=1}^n\left[\frac{1}{\hat{\bar{G}}(Z_i)} - \frac{1}{\bar{G}(Z_i)}\right]\delta_i Z_i t(\vec{X}_i) \tag{4.1}$$

is a residual term common to both strategies.

Writing that $\gamma_n(\hat{\psi}_m)\le\gamma_n(\psi_m)$ where ψ_m is the orthogonal projection of ψ on S_m, we obtain

$$\|\hat{\psi}_m-\psi\|^2 \le \|\psi_m-\psi\|^2 + 2\nu_n(\hat{\psi}_m-\psi_m) + 2\,R_n(\hat{\psi}_m-\psi_m).$$

Let $B_m(0,1)=\{t\in S_m,\ \|t\|=1\}$ denote the unit ball of S_m.

$$2\mathbb{E}|\nu_n(\hat\psi_m-\psi_m)| \le \frac{1}{8}\mathbb{E}(\|\hat\psi_m-\psi_m\|^2)+8\mathbb{E}\left(\sup_{t\in B_m(0,1)}\nu_n^2(t)\right)$$

$$\begin{aligned}\mathbb{E}\left(\sup_{t\in B_m(0,1)}\nu_n^2(t)\right) &\le \sum_{\lambda\in\Lambda_m}\mathbb{E}(\nu_n^2(\varphi_\lambda))\le \frac{1}{n^2}\sum_{\lambda\in\Lambda_m}\mathrm{Var}\left(\sum_{i=1}^n\frac{\delta_iZ_i)}{\bar G(Z_i)}\varphi_\lambda(\vec X_i)\right)\\ &\le \frac{1}{n}\sum_{\lambda\in\Lambda_m}\mathbb{E}\left(\frac{\delta_1^2Z_1^2}{(\bar G(Z_1))^2}\varphi_\lambda^2(\vec X_1)\right)\le\frac{\Phi_0^2D_m}{c_G^2n}\mathbb{E}(Y_{1,T}^2).\end{aligned}$$

Thus

$$2\mathbb{E}|\nu_n(\hat\psi_m-\psi_m)|\le\frac{1}{4}\mathbb{E}(\|\hat\psi_m-\psi\|^2)+\frac{1}{4}\|\psi_m-\psi\|^2+\frac{8\Phi_0^2\mathbb{E}(Y_1^2)}{c_G^2}\frac{D_m}{n}.$$

A rough bound for the residual term can be found in an analogous manner:

Lemma 4.1. *There exists a constant $C=2^{10}\sqrt{C_8}$ where C_8 is defined in Lemma 3.1, such that*

$$2\mathbb{E}|R_n(\hat\psi_m-\psi_m)|\le\frac{1}{4}\mathbb{E}(\|\hat\psi_m-\psi\|^2)+\frac{1}{4}\|\psi_m-\psi\|^2+\frac{2^5\Phi_0^2\mathbb{E}^{1/2}(Y_1^4)}{c_G^4}\frac{D_m}{n}+\frac{C\Phi_0^2\mathbb{E}^{1/2}(Y_1^4)}{nc_G^8}.$$

By gathering all terms, we find, under very mild conditions the following result.

Proposition 4.1. *Under assumptions ($\mathcal{H}_1$) and ($\mathcal{H}_2$) for the collection of models and if Y_1 admits moments of order 4, then the estimator $\hat\psi_m=\arg\min_{t\in S_m}\gamma_n(t)$ for γ_n defined by (3.3) satisfies:*

$$\mathbb{E}(\|\hat\psi_m-\psi\|^2)\le 7\|\psi_m-\psi\|^2+K\frac{\Phi_0^2\,\mathbb{E}^{1/2}(Y_1^4)}{c_G^4}\frac{D_m}{n}+K'\frac{\Phi_0^2\mathbb{E}^{1/2}(Y_1^4)}{c_G^8}\frac{1}{n}.$$

for positive constants K and K'.

This leads to standard rates on Besov spaces provided that the dimension D_m of the projection space is relevantly chosen in function of the index of regularity of the function, by using the bias orders given in Section 2.3. Since this index is unknown, an automatic data-driven choice has to be performed and is obtained for $\hat\psi_{\hat m}$ defined by (3.4).

Theorem 4.1. *Assume that f is upper bounded on $[0,1]^d$ by f_1 and that the Y_i's admit moments of order 8. Consider the collection of models built on [T], [DP] or [W] with N_n defined in Section 2.3 and satisfying $N_n\le n/(16f_1K_\varphi)$ for [DP] and [W] where K_φ is a basis-dependent constant, or $N_n\le\sqrt{n}/(4\sqrt{f_1})$ for [T]. Let $\hat\psi_{\hat m}$ be the adaptive estimator of ψ defined by (3.4) with*

$$\mathrm{pen}(m)=\kappa\Phi_0^2\mathbb{E}\left[\left(\frac{\delta_1Z_1}{\bar G(Z_1)}\right)^2\right]\frac{D_m}{n},$$

where κ is a numerical constant. Then

$$\mathbb{E}(\|\hat\psi_{\hat m}-\psi\|^2)\le C\inf_{m\in\mathcal{M}_n}\left(\|\psi_m-\psi\|^2+\mathrm{pen}(m)\right)+\frac{C'\sqrt{\ln(n)}}{n},\tag{4.2}$$

where C and C' are constants depending on Φ_0, f_1, T and c_G.

Note that the penalty depends on constants which do not have the same status. Indeed, the constant Φ_0 is known, but the constant $\mathbb{E}[(\delta_1 Z_1/\bar{G}(Z_1))^2]$ is unknown and has to be replaced by an estimator. The true penalty is therefore random and equal to

$$\widehat{\mathrm{pen}}(m) = \kappa\hat{\sigma}^2\Phi_0^2\frac{D_m}{n}, \text{ with } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\delta_i Z_i}{\hat{\bar{G}}(Z_i)}\right)^2 \tag{4.3}$$

where the constant κ is usually determined by simulations experiments. It is easy to extend the nonasymptotic bound in (4.2) to the case of the random penalty given by (4.3) (see e.g. Brunel and Comte [7]).

Theorem 4.2. *Assume that assumptions of Theorem 4.1 hold, consider the estimator $\hat{\psi}_{\hat{m}}$ be the adaptive estimator defined by (3.4) with penalty $\widehat{\mathrm{pen}}(m)$ defined by (4.3), where κ is a numerical constant. Then*

$$\mathbb{E}(\|\hat{\psi}_{\hat{m}} - \psi\|^2) \leq C \inf_{m\in\mathcal{M}_n}\left(\|\psi_m - \psi\|^2 + \mathrm{pen}(m)\right) + \frac{C'}{n}, \tag{4.4}$$

where C and C' are constants depending on Φ_0, $\|\psi\|$ and c_G.

The proof of this result, being standard, is omitted.

4.2. **Quotient estimation of r.** On the other hand, the standard adaptive estimator of the density f of the $\vec{X}_i$'s given by (3.5) and (3.6) is known (see Birgé and Massart [4]) to satisfy also an inequality of type (2.7).

The quadratic risk of $\tilde{r}^P$ is bounded by the sum of the risks of $\tilde{\psi}$ and $\tilde{f}$, under adequate conditions. In term of asymptotic rates, this means that the resulting rate for $\tilde{r}^P$ is the worst one between the rate of $\tilde{\psi}$ and the rate of $\tilde{f}$. The optimal minimax rate can then be recovered only if r is more regular than f. More precisely, we can prove for isotropic Besov spaces the following result:

Proposition 4.2. *Assume that $f \in \mathcal{B}^{\alpha_f}_{2,\infty}(A)$ and $\psi \in \mathcal{B}^{\alpha_\psi}_{2,\infty}(A)$, α_f, α_ψ real (isotropic case) and $\alpha_f > 1/2$, $\alpha_\psi > 1/2$ and that $0 < f_0 \leq f(x) \leq f_1 < +\infty$ for all $x \in A = [0,1]^d$. Moreover assume that the Y_i's admit moments of order 8. Consider the collections described in Section 2.2 with dimensions D_m such that $\ln(n) \leq D_m \leq O(\sqrt{n/\ln(n)})$ and such that for all $m \in \mathcal{M}_n$, $\sup_{x\in[0,1]^d}|f_m(x) - f(x)| \leq cD_m^{-\alpha_f+1/2}$, for a positive constant c. Then for $a_n = O(\exp(n^\xi))$ for $0 < \xi < 1/2$ and n large enough,*

$$\mathbb{E}\left(\|\tilde{r}^P - r_T\|^2\right) \leq O\left(n^{-\frac{2\alpha_f}{2\alpha_f+1}} \vee n^{-\frac{2\alpha_\psi}{2\alpha_\psi+1}}\right).$$

Remark 4.1. The assumption $\sup_{x\in[0,1]^d}|f_m(x) - f(x)| \leq cD_m^{-\alpha_f+1/2}, \forall m \in \mathcal{M}_n$ is fulfilled for any collection for $d = 1$ and for collection of piecewise polynomials when $d \geq 1$ (see DeVore [11], Section 6).

This method produces an estimator easy to compute. Moreover, we found out in Brunel and Comte [7] when running practical implementations that a quotient estimator could be surprisingly very competitive as compared to a direct estimator which was expected from the theory to be better.

5. Study of the mean-square estimator

In the mean-square strategy, the contrast decomposition is the following

$$\gamma_n^{MS}(t) - \gamma_n^{MS}(s) = \|t - r_T\|_n^2 - \|s - r_T\|_n^2 - 2R_n(t-s) - 2\nu_n^{MS}(t-s) \tag{5.1}$$

where R_n is defined by (4.1) and

$$\nu_n^{MS}(t) = \frac{1}{n}\sum_{i=1}^n \left[\frac{\delta_i Z_i}{\bar{G}(Z_i)} - r_T(\vec{X}_i)\right] t(\vec{X}_i).$$

Note that

$$\begin{aligned}
\mathbb{E}\left\{\left[\frac{\delta_1 Z_1}{\bar{G}(Z_1)} - r_T(\vec{X}_1)\right] t(\vec{X}_1)\right\} &= \mathbb{E}\left\{\mathbb{E}\left[\left(\frac{\delta_1 Y_{1,T}}{\bar{G}(Y_{1,T})} - r_T(\vec{X}_1)\right) t(\vec{X}_1) |\, X_1, Y_{1,T}\right]\right\} \\
&= \mathbb{E}\left\{\left(\frac{\mathbb{E}(\delta_1|\, X_1, Y_{1,T}) Y_{1,T}}{\bar{G}(Y_{1,T})} - r_T(\vec{X}_1)\right) t(\vec{X}_1)\right\} \\
&= \mathbb{E}\left\{\left(Y_{1,T} - r_T(\vec{X}_1)\right) t(\vec{X}_1)\right\} \\
&= \mathbb{E}\left\{\mathbb{E}\left[Y_{1,T} - r_T(\vec{X}_1)|\vec{X}_1\right] t(\vec{X}_1)\right\} = 0.
\end{aligned}$$

Therefore, $\nu_n^{MS}(t)$ is centered. Then decomposition (5.1) yields to a result similar to Proposition 4.1 with $\nu_n(t)$ replaced by $\nu_n^{MS}(t)$. We give directly the result concerning the adaptive estimator. The automatic selection of the projection space can be performed via penalization:

Theorem 5.1. *Assume that the density f is such that $\forall x \in [0,1]^d, 0 < f_0 \le f(x) < f_1 < +\infty$ and that the Y_i's admit moments of order 8. Consider the collection of models built on [T], [DP] or [W] with $N_n \le n/(16 f_1 K_\varphi)$ for [DP] and [W] where K_φ is a constant depending on the basis, and $N_n \le \sqrt{n}/(4\sqrt{f_1})$ for [T]. Let $\hat{r}_{m^*}$ be the adaptive estimator defined by (3.7) and (3.9) with*

$$\mathrm{pen}(m) = \kappa \Phi_0^2 \mathbb{E}\left[\left(\frac{\delta_1 Z_1}{\bar{G}(Z_1)}\right)^2\right] \frac{D_m}{n},$$

where κ is a numerical constant. Then

$$\mathbb{E}(\|\hat{r}_{m^*} - r_T\|_n^2) \le C \inf_{m \in \mathcal{M}_n} \left(\|r_m - r_T\|^2 + \mathrm{pen}(m)\right) + C' \frac{\sqrt{\ln(n)}}{n}, \tag{5.2}$$

where r_m is the orthogonal projection of r_T onto S_m and C and C' are constants depending on Φ_0, $\|f\|$ and c_G.

Note that the penalty is the same as for estimating ψ. The unknown expectation therein has to be replaced by the same estimator as in (4.3). Again, it can be proved that the estimator obtained by random penalization still satisfies Inequality (5.2) under the assumptions of Theorem 5.1. Since this implies tedious computations, we refer to Birgé and Massart [4] to find the technical elements of this type of proofs in the univariate setting.

Proposition 5.1. *Assume that $r \in \mathcal{B}^{\alpha}_{2,\infty}(A)$ with $\alpha = (\alpha_1, \ldots, \alpha_d)$ and $A = [0,1]^d$ and that an estimator $\tilde{r}^{MS}$ of r defined by (3.8) and (3.9) satisfies Inequality (5.2), then*

$$\mathbb{E}(\|\tilde{r}^{MS} - r_T\|_n^2) = O\left(n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+d}}\right)$$

where $\bar{\alpha}$ is the harmonic mean of the α_i's.

6. Proofs

6.1. Proof of Inequality (2.9).

$$\forall i_0 \in \{1, \ldots, d\}\,,\ \frac{\partial}{\partial m_{i_0}}\left(\sum_{i=1}^{d} D_{m_i}^{-2\alpha_i} + \frac{\Pi_{i=1}^{d} D_{m_i}}{n}\right) = -2\alpha_{i_0} D_{m_{i_0}}^{-2\alpha_{i_0}-1} + \frac{\Pi_{i\neq i_0}^{d} D_{m_i}}{n}\,.$$

Writing that all the derivatives equal zero, implies that $D_{m_k^*} = (\alpha_1/\alpha_k) D_{m_1^*}^{\alpha_1/\alpha_k}$, for all $2 \le k \le d$ and $D_{m_1^*}^{-2\alpha_1} = (1/2\alpha_1) D_{m*}/n$. It follows that for all $i \in \{1, \ldots, d\}$,

$$D_{m_i^*} = O([n]^{1/[\alpha_i(d/\bar{\alpha}+2)]})$$

and $D_{m^*}/n = O(n^{-2\bar{\alpha}/2\bar{\alpha}+d})$. Thus, it ensures that $n = D_{m^*}^{2\bar{\alpha}/d+1}$ and we find that $D_{m_i^*}^{-2\alpha_i} = D_{m^*}/n = D_{m^*}^{1-(2\bar{\alpha}/d+1)} = D_{m^*}^{2\bar{\alpha}/d}$, $\forall i \in \{1, \ldots, d\}$ which gives the announced order of the bias in function of D_{m^*}.

6.2. Proof of Lemma 3.1.

First note that with the remark (3.2), it is enough for $\bar{G}_{n,1}$ to prove the result for $\bar{G}_{n,0}$. We use a nonasymptotic exponential bound for the Kaplan-Meier estimator which can be formulated as follows (see Bitouzé *et al.*, [5]), there exists a constant $c > 0$ such that for any positive λ

$$\mathbb{P}\left(\sqrt{n}\|(1-F_Y)\,(\hat{\bar{G}}_{n,0} - \bar{G})\|_\infty > \lambda\right) \le 2.5\, e^{-2\lambda^2 + c\lambda}$$

and so

$$\begin{aligned}
\mathbb{E}\left[\left(\sup_{y\in[0,T]} |\hat{\bar{G}}_{n,1}(y) - \bar{G}(y)|\right)^{2k}\right] &\le 2k \int_0^{+\infty} u^{2k-1}\, \mathbb{P}(\sup_{y\in[0,T]} |\hat{\bar{G}}_{n,1}(y) - \bar{G}(y)| > u)\, du \\
&= 2k \int_0^{+\infty} u^{2k-1}\, \mathbb{P}(c_F^{-1} \sup_{y\in[0,T]} |(1-F_Y)\,(\hat{\bar{G}}_{n,1} - \bar{G})(y)| > u)\, du \\
&\le 2k \int_0^{+\infty} u^{2k-1}\, \mathbb{P}(\sqrt{n}\|(1-F_Y)\,(\hat{\bar{G}}_{n,1} - \bar{G})\|_\infty > c_F\sqrt{n}\, u)\, du
\end{aligned}$$

$$\leq \quad 5ke^{c^2/8}\int_0^{\infty} u^{2k-1}\exp\left(-2c_F^2 n\left[u-\frac{c}{4\sqrt{n}c_F}\right]^2\right)du$$
$$\leq \quad \frac{5e^{c^2/8}k}{2^k c_F^{2k}}\int_{-c/(2\sqrt{2})}^{+\infty}\left(z+\frac{1}{2\sqrt{2}}\right)^{2k-1} e^{-z^2}dz\, n^{-k} = C_k n^{-k}.\square$$

6.3. **Proof of Lemma 4.1.** Write as in the other cases that

$$2\mathbb{E}|R_n(\hat{\psi}_m-\psi_m)| \quad \leq \quad \frac{1}{8}\mathbb{E}(\|\hat{\psi}_m-\psi_m\|^2)+8\mathbb{E}\left(\sup_{t\in B_m(0,1)} R_n^2(t)\right) \tag{6.1}$$

Let $\Omega_G=\{\omega, 1-\hat{G}(y)\geq c_G/2, \forall y\in[0,T]\}$ and define $\|G-\hat{G}\|_{\infty,T}=\sup_{y\in[0,T]}|G(y)-\hat{G}(y)|$. On the set Ω_G, we have

$$\mathbb{E}\left(\sup_{t\in B_m(0,1)} R_n^2(t)\mathbf{I}_{\Omega_G}\right) \quad \leq \quad \frac{4}{nc_G^4}\sum_{\lambda\in\Lambda_m}\sum_{i=1}^n \mathbb{E}\left(\|G-\hat{G}\|^2_{\infty,T}Y_{i,T}^2\varphi_\lambda^2(\vec{X}_i)\right)$$
$$\leq \quad \frac{4\Phi_0^2 D_m}{nc_G^4}\sum_{i=1}^n \mathbb{E}\left(\|G-\hat{G}\|^2_{\infty,T}Y_{i,T}^2\right)$$

Then by using Lemma 3.2, $\mathbb{E}(\|G-\hat{G}\|^4_{\infty,T})\leq c/n^2$ and since $\mathbb{E}(Y_{1,T}^4)\leq\mathbb{E}(Y_1^4)<+\infty$, we find that

$$\mathbb{E}\left(\sup_{t\in B_m(0,1)} R_n^2(t)\mathbf{I}_{\Omega_G}\right) \quad \leq \quad \frac{4\Phi_0^2\mathbb{E}^{1/2}(Y_1^4)}{c_G^4}\frac{D_m}{n}. \tag{6.2}$$

On the complementary Ω_G^c, we use that $1-\hat{G}(Z_i)\geq 1/(n+1)$ and that $\|G-\hat{G}\|_{\infty,T}>c_G/2$. Then, with Markov Inequality and Lemma 3.1, we obtain

$$\mathbb{E}\left(\sup_{t\in B_m(0,1)} R_n^2(t)\mathbf{I}_{\Omega_G^c}\right) \quad \leq \quad \frac{\Phi_0^2 D_m(n+1)^2}{c_G^2}\mathbb{E}\left(\|G-\hat{G}\|^2_{\infty,T}\mathbf{I}_{\{\|G-\hat{G}\|_{\infty,T}>c_G/2\}}\frac{1}{n}\sum_{i=1}^n Y_{i,T}^2\right)$$
$$\leq \quad \frac{\Phi_0^2 D_m(n+1)^2}{c_G^2}\mathbb{E}^{1/2}\left(\|G-\bar{G}\|^4_{\infty,T}\mathbf{I}_{\{\|G-\hat{G}\|_{\infty,T}>c_G/2\}}\right)\mathbb{E}^{1/2}(Y_1^4)$$
$$\leq \quad 2^6\Phi_0^2 D_m(n+1)^2 c_G^{-8}\mathbb{E}^{1/2}\left(\|G-\hat{G}\|^{16}_{\infty,T}\right)\mathbb{E}^{1/2}(Y_1^4)$$
$$\leq \quad \frac{2^7\sqrt{C_8}\Phi_0^2\mathbb{E}^{1/2}(Y_1^4)}{nc_G^8} \tag{6.3}$$

The result follows by gathering (6.1), (6.2) and (6.3). □

6.4. **A key result.** Most proofs are based on the use of Talagrand's Inequality (see Talagrand [29]):

Lemma 6.1. *Let $U_1,\ldots,U_n$ be independent random variables and define*

$$\nu_n(h)=(1/n)\sum_{i=1}^n[h(U_i)-\mathbb{E}(h(U_i))]$$

for h belonging to a countable class $\mathcal{H}$ of uniformly bounded measurable functions. Then for $\epsilon > 0$

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}|\nu_n(h)|^2 - 2(1+2\epsilon)H^2\right]_+ \leq \frac{6}{K_1}\left(\frac{v}{n}e^{-K_1\epsilon\frac{nH^2}{v}} + \frac{8M_1^2}{K_1 n^2 C^2(\epsilon)}e^{-\frac{K_1 C(\epsilon)\sqrt{\epsilon}}{\sqrt{2}}\frac{nH}{M_1}}\right), \tag{6.4}$$

with $C(\epsilon) = \sqrt{1+\epsilon} - 1$, K_1 is a universal constant, and where

$$\sup_{h\in\mathcal{H}}\sup_{x\in[0,1]^d}|h(x)| \leq M_1, \quad \mathbb{E}\left(\sup_{h\in\mathcal{H}}|\nu_n(h)|\right) \leq H, \quad \sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \mathrm{Var}(h(U_i)) \leq v.$$

The inequality (6.4) is a straightforward consequence of Talagrand's [29] inequality given in Ledoux [21] (or Birgé and Massart [4]) with f replaced by $h = f - \mathbb{E}f(\vec{X}_1)$ and M_1 by $2M_1$, and by taking $\eta = (\sqrt{1+\epsilon} - 1)\wedge 1 = C(\epsilon) \leq 1$. Moreover, standard density arguments allow to apply it to unit balls of finite dimensional spaces, instead of countable sets.
This inequality can be a fortiori applied to identically distributed variables and in that case, v is more simply defined by $\sup_{h\in\mathcal{H}}\mathrm{Var}(h(\vec{X}_1)) \leq v$.

6.5. **Proof of Theorem 4.1.** Let us denote by $B_{m,m'}(0,1) = \{t \in S_m + S_{m'}, \|t\| \leq 1\}$. Since the spaces S_m are nested, $B_{m,m'}(0,1) = B_m(0,1) \vee B_{m'}(0,1)$. By writing that $\forall m \in \mathcal{M}_n$, $\gamma_n(\hat{\psi}_{\hat{m}}) + \mathrm{pen}(\hat{m}) \leq \gamma_n(\psi_m) + \mathrm{pen}(m)$, we obtain that:

$$\begin{aligned}
\frac{1}{2}\|\hat{\psi}_{\hat{m}} - \psi\|^2 &\leq \frac{3}{2}\|\psi_m - \psi\|^2 + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + 8\sup_{t\in B_{m,\hat{m}}(0,1)}[\nu_n(t)]^2 \\
&\quad +8\sup_{t\in B_{m,\hat{m}}(0,1)}R_n^2(t)\mathbb{I}_{\Omega_G} + 8\sup_{t\in B_{m,\hat{m}}(0,1)}R_n^2(t)\mathbb{I}_{\Omega_G^c} \\
&\leq \frac{3}{2}\|\psi_m - \psi\|^2 + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + 8p(m,\hat{m}) \\
&\quad +8\left(\sup_{t\in B_{m,\hat{m}}(0,1)}[\nu_n(t)]^2 - p(m,\hat{m})\right)_+ \\
&\quad +8\sup_{t\in B_{m,\hat{m}}(0,1)}R_n^2(t)\mathbb{I}_{\Omega_G} + 8\sup_{t\in B_{m,\hat{m}}(0,1)}R_n^2(t)\mathbb{I}_{\Omega_G^c}.
\end{aligned}$$

Then the penalty is chosen such that $\forall m' \in \mathcal{M}_n$,

$$8p(m,m') \leq \mathrm{pen}(m) + \mathrm{pen}(m') \tag{6.5}$$

and $p(m,m')$ is determined in order to have

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)}(\nu_n(t))^2 - p(m,m')\right)_+ \leq \frac{c}{n}. \tag{6.6}$$

for a positive contant c. Then (6.6) is obtained by using Talagrand's Theorem recalled in Lemma 6.1, applied in the i.i.d. case to the process ν_n, where it is easy to compute $H^2 =$

$\sigma_T^2\Phi_0^2(D_m \vee D_{m'})/n$ with $\sigma_T^2 = \mathbb{E}(\delta_1^2 Y_{1,T}^2/\bar{G}^2(Y_{1,T}))$. Moreover

$$\sup_{t\in B_{m,m'}(0,1)} \sup_{x\in[0,1]^d} |\delta_1 Y_{1,T} t(x)/\bar{G}(Y_{1,T})| \le \Phi_0\sqrt{D_{m'} \vee D_m} T/c_G := M_1$$

and

$$\begin{aligned}\sup_{t\in B_{m,m'}(0,1)} \mathrm{Var}(\delta_1 Y_{1,T} t(\vec{X}_1)/\bar{G}(Y_{1,T})) &\le (T^2/c_G^2) \sup_{t\in B_{m\vee m'}(0,1)} \int t^2(x)\, f(x)\, dx \\ &\le (T^2/c_G^2)\, f_1 := v.\end{aligned}$$

Finally, with $C(\epsilon) = (\sqrt{1+\epsilon}-1)\wedge 1$ and K_1 standing for a universal constant,

$$\mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)} \nu_n^2(t) - p(m,m')\right)_+ \le \frac{6}{K_1}\left(\frac{v}{n}\, e^{-K_1\epsilon\frac{n H^2}{v}} + \frac{8M_1^2}{K_1\, n^2\, C^2(\epsilon)}\, e^{-\frac{K_1 C(\epsilon)\sqrt{\epsilon}}{\sqrt{2}}\frac{nH}{M_1}}\right),$$

with $p(m,m') = 2(1+2\epsilon)H^2$.

Now replacing M_1, v and H^2 by the values derived above, we obtain for $\epsilon = 1/2$,

$$p(m,m') = 4\Phi_0^2\mathbb{E}\left(\frac{\delta_1^2 Y_{1,T}^2}{\bar{G}^2(Y_{1,T})}\right)\frac{D_m \vee D_{m'}}{n} = 4H^2, \tag{6.7}$$

and the following upper bound,

$$\mathbb{E}\left(\sup_{t\in B_{m\vee m'}(0,1)} \nu_n^2(t) - p(m,m')\right)_+ \le \frac{6/K_1}{n}\left(\alpha_0 e^{-\alpha_1 D_{m'}\vee D_m} + \alpha_2\, e^{-\alpha_3\sqrt{n}}\right)$$

where the constants are $\alpha_0 = T^2 f_1/c_G^2$, $\alpha_1 = K_1\sigma^2\Phi_0^2 c_G^2/(2f_1T^2)$, $\alpha_2 = 8T^2/(K_1 c_G^2 C^2(1/2))$ and $\alpha_3 = K_1 C(1/2)\sigma\Phi_0 c_G/(2T)$. For $|\mathcal{M}_n| \le n$, we obtain

$$\begin{aligned}&\sum_{m'\in\mathcal{M}_n} \mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)} \nu_n^2(t) - p(m,m')\right)_+ \\ \le\ & \frac{6/K_1}{n}\left(\sum_{m'\in\mathcal{M}_n} \alpha_0\, e^{-\alpha_1 D_{m'}} + \alpha_2\, \mathrm{card}(\mathcal{M}_n)\, e^{-\alpha_3\sqrt{n}}\right) \le \frac{c_1}{n}\left(S(\alpha_1) + c_2\right).\end{aligned}$$

for constants c_1 and c_2 and with $S(\alpha_1) = \sum_{k=1}^{+\infty} e^{-\alpha_1 k} < +\infty$. This ends the proof of (6.6).

It results from the proof of Lemma 4.1 that $\mathbb{E}(\sup_{t\in B_{m,\hat{m}}(0,1)} R_n^2(t)\mathbb{I}_{\Omega_G^c}) \le c/n$. This yields

$$\mathbb{E}(\|\hat{\psi}_{\hat{m}} - \psi\|^2) \le 3\|\psi_m - \psi\|^2 + 4\mathrm{pen}(m) + \frac{C}{n} + 16\mathbb{E}\left(\sup_{t\in B_n(0,1)} R_n^2(t)\mathbb{I}_{\Omega_G}\right).$$

Then the result follows if the following Lemma is proved

Lemma 6.2. *If $N_n \le n/(16 f_1 K_\varphi)$ for [DP] and [W], where K_φ is a basis-dependent constant, or if $N_n \le \sqrt{n}/(4\sqrt{f_1})$ for [T], and if Y_1 admits moments of order 8, then*

$$\mathbb{E}\left(\sup_{t\in B_n(0,1)} R_n^2(t)\mathbb{I}_{\Omega_G}\right) \le C\frac{\sqrt{\ln(n)}}{n}$$

for a constant C depending on f_1.

The proof of this lemma is postponed to Section 6.6.

Then by gathering the dimension conditions of Lemma 6.2, the moment condition of Y_1, and the definition of $p(m,m')$ in (6.7), we find the result, where pen is defined by inequality (6.5). □

6.6. **Proof of Lemma 6.2.**

$$\mathbb{E}\left(\sup_{t\in B_n(0,1)} R_n^2(t)\mathbb{I}_{\Omega_G}\right) \leq \frac{4}{c_G^4}\mathbb{E}\left[\|\hat{\bar{G}}-\bar{G}\|^2_{\infty,T}\sup_{t\in B_n(0,1)}\left(\frac{1}{n}\sum_{i=1}^n |Y_{i,T}t(\vec{X}_i)|\right)^2\right]$$

$$\leq \frac{4}{c_G^4}\mathbb{E}^{1/2}\left[\|\hat{\bar{G}}-\bar{G}\|^4_{\infty,T}\right]\mathbb{E}^{1/2}\left[\left(\frac{1}{n}\sum_{i=1}^n Y_{i,T}^2 \sup_{t\in B_n(0,1)}\frac{1}{n}\sum_{i=1}^n t^2(\vec{X}_i)\right)^2\right]$$

$$\leq \frac{4\sqrt{C_2}}{nc_G^4}\mathbb{E}^{1/2}\left\{\left[\frac{1}{n}\sum_{i=1}^n Y_{i,T}^2\left(\sup_{t\in B_n(0,1)}\nu_n'(t^2)+\mathbb{E}(t^2(\vec{X}_1))\right)\right]^2\right\},$$

where

$$\nu_n'(t) = (1/n)\sum_{i=1}^n [t(\vec{X}_i)-\mathbb{E}(t(\vec{X}_i))] \tag{6.8}$$

and $B_n(0,1)=\{t\in\mathcal{S}_n,\ \|t\|=1\}$.

$$\mathbb{E}\left(\sup_{t\in B_n(0,1)} R_n^2(t)\mathbb{I}_{\Omega_G}\right) \leq \frac{4\sqrt{C_2}}{nc_G^4}\mathbb{E}^{1/2}\left\{\left[\frac{1}{n}\sum_{i=1}^n Y_{i,T}^2\left(\sup_{t\in B_n(0,1)}\nu_n'(t^2)+f_1\right)\right]^2\right\},$$

$$\leq \frac{2^{11/4}\sqrt{C_2}}{nc_G^4}\mathbb{E}^{1/4}\left[\left(\frac{1}{n}\sum_{i=1}^n Y_{i,T}^2\right)^4\right]\left\{\left[\mathbb{E}\left(\sup_{t\in B_n(0,1)}\nu_n'(t^2)\right)^4+f_1^4\right]^{1/4}\right\}.$$

Then

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Y_{i,T}^2\right)^4\right] \leq \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^n Y_{i,T}^8\right) \leq \mathbb{E}(Y_1^8).$$

It follows from Baraud [1] that for any $\rho>0$,

$$\mathbb{P}\left(\sup_{t\in B_n(0,1)}|\nu_n'(t^2)|\geq\rho\right) \leq |\Lambda_n|^2\exp\left(-\frac{n\rho^2}{4f_1L_n(\varphi)}\right)$$

where $L_n(\varphi)$ is a quantity associated to the orthonormal basis of the largest space $\mathcal{S}_n$ of the (nested) collection and $(\varphi_\lambda)_{\lambda\in\Lambda_n}$ denote the orthonormal basis of $\mathcal{S}_n$, dim $\mathcal{S}_n = N_n = |\Lambda_n|$. We know from Baraud [1] that $L_n(\varphi)\leq K_\varphi N_n$ for the basis [DP] and [W], and $L_n(\varphi)\leq N_n^2$ for

[T]. Moreover, it can be checked as in Brunel and Comte [7] that, by integration of the previous inequality, we find

$$\mathbb{E}(\sup_{t\in B_n(0,1)}[\nu_n'(t^2)]^4) \le 4\ln(n)^2 + 4\int_{\sqrt{\ln(n)}}^{+\infty} x^3 \mathbb{P}\left(\sup_{t\in B_n(0,1)}|\nu_n'(t^2)| \ge x\right) dx$$

and by integration by parts, we get for a positive constant $\mathcal{K}$

$$\mathbb{E}\left(\sup_{t\in B_n(0,1)}[\nu_n'(t^2)]^4\right) \le \ln^2(n) + \mathcal{K} f_1 N_n \ln(n) L_n(\varphi)\exp\left(-\frac{n\ln(n)}{4f_1 L_n(\varphi)}\right).$$

It follows that if $L_n(\varphi) \le n/(16f_1)$,

$$\mathbb{E}\left(\sup_{t\in B_n(0,1)}[\nu_n'(t^2)]^4\right) \le \ln^2(n) + \frac{\mathcal{K}}{n^2} \le (\mathcal{K}+1)\ln^2(n) \text{ if } n \ge 2.$$

Therefore, if $N_n \le n/(16f_1 K\varphi)$ for [DP] and [W] and if $N_n \le \sqrt{n}/(4\sqrt{f_1})$ for [T], then

$$\mathbb{E}\left(\sup_{t\in \mathcal{B}_n(0,1)} R_n^2(t)\mathbb{I}_{\Omega_G}\right) \le \frac{C_1}{n} + \frac{C_2\sqrt{\ln(n)}}{n}.$$

□

6.7. **Proof of Proposition 4.2.** Let $\Omega_f = \{\omega/\tilde f(x) > f_0/2, \forall x \in [0,1]\}$ and $\tilde\psi = \hat\psi_{\hat m}$. Note that $|r_T(x)| = |\mathbb{E}(Y_{1,T}|\vec X_1)| \le T$. Write that

$$\begin{aligned} \|\tilde r^P - r_T\|^2 &\le 2\left\|\frac{\tilde\psi-\psi}{\tilde f}\mathbb{I}_{\Omega_f}\right\|^2 + 2\left\|\psi\frac{\tilde f - f}{\tilde f f}\mathbb{I}_{\Omega_f}\right\|^2 + 2\sup_{x\in[0,1]^d}|\tilde r(x) - r_T(x)|^2\mathbb{I}_{\Omega_f^c} \\ &\le \frac{8}{f_0^2}\|\tilde\psi-\psi\|^2 + \frac{8T^2}{f_0^2}\|\tilde f - f\|^2 + 2(a_n^2+T^2)\mathbb{I}_{\Omega_f^c}. \end{aligned}$$

Thus,

$$\mathbb{E}(\|\tilde r^P - r_T\|^2) \le \frac{8}{f_0^2}(1\vee T^2)\left(\|\tilde\psi-\psi\|^2 + \|\tilde f - f\|^2\right) + 2(a_n^2+T^2)\mathbb{P}(\Omega_f^c).$$

Therefore, the result follows if $a_n^2\mathbb{P}(\Omega_f^c) = o(1/n)$. Note that $\mathbb{P}(\Omega_f^c) \le \mathbb{P}(\sup_{x\in[0,1]^d}|\tilde f(x) - f(x)| > f_0/2)$. Then for $\alpha_f > 1/2$, $\sup_{x\in[0,1]^d}|f_{\hat m}(x) - f(x)| \le cD_{\hat m}^{-\alpha_f+1/2} \le C(\ln(n)^{-\alpha_f+1/2})$, as $D_m \ge \ln(n)$, $\forall m \in \mathcal{M}_n$. Therefore, for n great enough, $\sup_{x\in[0,1]^d}|f_{\hat m}(x) - f(x)| \le f_0/4$. As $\sup_{x\in[0,1]^d}|\tilde f(x) - f(x)| \le \sup_{x\in[0,1]^d}|\tilde f(x) - f_{\hat m}(x)| + \sup_{x\in[0,1]^d}|f_{\hat m}(x) - f(x)|$, it follows that

$$\begin{aligned} \mathbb{P}(\Omega_f^c) &\le \mathbb{P}(\sup_{x\in[0,1]^d}|\hat f_{\hat m}(x) - f_{\hat m}(x)| > f_0/4) \le \mathbb{P}(\|\hat f_{\hat m} - f_{\hat m}\| > f_0/(4\sqrt{D_{\hat m}})) \\ &\le \sum_{m\in\mathcal{M}_n}\mathbb{P}(\|\hat f_m - f_m\| > f_0/(4\sqrt{D_m})). \end{aligned}$$

Then note that

$$\|\hat f_m - f_m\|^2 = \sum_{\lambda\in\Lambda_m}(\hat a_\lambda(f) - a_\lambda(f))^2 = \sum_{\lambda\in\Lambda}[\nu_n^f(\varphi_\lambda)]^2 = \sup_{t\in B_m(0,1)}[\nu_n^f(t)]^2$$

where $a_\lambda(f) = \langle \varphi_\lambda, f\rangle$ and $\hat a_\lambda(f) = (1/n)\sum_{i=1}^n \varphi_\lambda(\vec X_i)$. Then Talagrand's inequality can be written

$$\mathbb{P}\left(\sup_{t\in B_m(0,1)} |\nu_n^f(t)| \geq 2H_f + \lambda\right) \leq 3\exp\left(K_1 n\left(\frac{\lambda^2}{v_f}\wedge \frac{\lambda}{M_f}\right)\right)$$

where we have $\mathbb{E}(\sup_{t\in B_m(0,1)}[\nu_n^f(t)]^2) \leq \Phi_0^2 D_m/n = H_f^2$, $\sup_{t\in B_m(0,1)} \mathrm{Var}(t(\vec X_1)) \leq f_1 = v_f$ and $\sup_{t\in B_m(0,1)}\sup_{x\in[0,1]^d}|t(x)| \leq \Phi_0\sqrt{D_m} = M_f$. Then choose $\lambda = \Phi_0\sqrt{D_m/n}$ and if $2\Phi_0\sqrt{D_m/n} \leq f_0/(2\sqrt{D_m})$ which holds since $D_m \leq O(\sqrt{n/\ln(n)})$, then the following inequality holds:

$$\mathbb{P}(\Omega_f^c) \leq 3\sum_{m\in\mathcal{M}_n} \exp\left(-K_1 n\left(\frac{\Phi_0^2 D_m}{n f_1}\wedge\frac{1}{\sqrt n}\right)\right) \leq 3|\mathcal{M}_n|\exp(-K_1\sqrt n)$$

so that $a_n^2\mathbb{P}(\Omega_f^c) \leq n a_n^2 \exp(-K_1\sqrt n) = o(1/n^2)$. □

6.8. **Proof of Theorem 5.1.** We start by writing that, $\forall m\in\mathcal{M}_n$,

$$\gamma_n(\hat r_{m^*}) + \mathrm{pen}(m^*) \leq \gamma_n(r_m) + \mathrm{pen}(m)$$

and by using the decomposition (5.1). It follows that

$$\|\hat r_{m^*} - r_T\|_n^2 \leq \|r_m - r_T\|_n^2 + 2R_n(\hat r_{m^*} - r_m) + 2\nu_n^{MS}(\hat r_{m^*} - r_m) + \mathrm{pen}(m) - \mathrm{pen}(m^*).$$

Let us introduce, in the same way as Baraud et al. [2], the set:

$$\Omega_n = \left\{\omega \;/\; \left|\frac{\|t\|_n^2}{\|t\|^2} - 1\right| \leq \frac12, \forall t\in \cup_{m,m'\in\mathcal{M}_n}(S_m + S_{m'})/\{0\}\right\}. \tag{6.9}$$

On the complementary of Ω_n, a separate study leads to the following Lemma:

Lemma 6.3. *$\mathbb{P}(\Omega_n^c) \leq c/n^2$ and $\mathbb{E}(\|\hat r_m - r_T\|_n^2 \mathbb{I}_{\Omega_n^c}) \leq c'/n$, for positive constants c and c'.*

Therefore, we focus on the study of the bounds on Ω_n, where we have $\|t\|^2 \leq 2\|t\|_n^2$.

$$\begin{aligned}
\|\hat r_{m^*} - r_T\|_n^2 \mathbb{I}_{\Omega_n} &\leq \|r_m - r_T\|_n^2 + \frac18\|\hat r_{m^*} - r_m\|^2\mathbb{I}_{\Omega_n} + 16\sup_{t\in B_{m^*,m}(0,1)} R_n^2(t) \\
&\quad +16\sup_{t\in B_{m^*,m}(0,1)}[\nu_n^{MS}]^2(t) + \mathrm{pen}(m) - \mathrm{pen}(m^*) \\
&\leq (1+\frac12)\|r_m - r_T\|_n^2 + \frac12\|\hat r_{m^*} - r_T\|_n^2\mathbb{I}_{\Omega_n} + 16\sup_{t\in B_{m^*,m}(0,1)} R_n^2(t) \\
&\quad +16\left(\sup_{t\in B_{m^*,m}(0,1)}[\nu_n^{MS}]^2(t) - \tilde p(m,m^*)\right)_+ \\
&\quad +\mathrm{pen}(m) + 16\tilde p(m,m^*) - \mathrm{pen}(m^*)
\end{aligned}$$

The supremum of $R_n^2(t)$ has already been studied in Lemma 6.2, and it is easy to see that

$$\begin{aligned}
\mathbb{E}\left(\sup_{t\in B_{m',m}(0,1)}[\nu_n^{MS}]^2(t)\right) &\leq \sum_{\lambda\in\Lambda_m\cup\Lambda_{m'}}\frac{1}{n}\mathrm{Var}\left\{\left[\frac{\delta_1 Y_{1,T}}{\bar{G}(Y_{1,T})}-r_T(\vec{X}_1)\right]\varphi_\lambda(\vec{X}_1)\right\}\\
&\leq \frac{\Phi_0^2(D_m\vee D_{m'})}{n}\mathbb{E}\left[\frac{\delta_1 Y_{1,T}}{\bar{G}(Y_{1,T})}-r_T(\vec{X}_1)\right]^2.
\end{aligned}$$

Therefore, we obtain by applying Talagrand's Inequality

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\sup_{t\in B_{m',m}((0,1)}[\nu_n^{MS}]^2(t)-\tilde{p}(m,m')\right)_+\leq\frac{c}{n}.$$

with

$$\tilde{p}(m,m')=4\frac{\Phi_0^2(D_m\vee D_{m'})}{n}\mathbb{E}\left[\frac{\delta_1 Y_{1,T}}{\bar{G}(Y_{1,T})}-r_T(\vec{X}_1)\right]^2:=4H^2$$

$v=f_1T^2/c_G^2$ and $M_1=2(\Phi_0T/c_G)\sqrt{D_{m'}\wedge D_m}$. Note that

$$\begin{aligned}
\tilde{p}(m,m') &= 4\frac{\Phi_0^2(D_m\vee D_{m'})}{n}\mathbb{E}\left[\frac{\delta_1 Y_{1,T}^2}{\bar{G}^2(Y_{1,T})}-2\frac{\delta_1 Y_{1,T}r_T(\vec{X}_1)}{\bar{G}(Y_{1,T})}+r_T(\vec{X}_1)^2\right]\\
&= 4\frac{\Phi_0^2(D_m\vee D_{m'})}{n}\left\{\mathbb{E}\left(\frac{\delta_1 Y_{1,T}^2}{\bar{G}^2(Y_{1,T})}+r_T(\vec{X}_1)^2\right)-2\mathbb{E}\left[\mathbb{E}\left(\frac{\delta_1 Y_{1,T}r_T(\vec{X}_1))}{\bar{G}(Y_{1,T})}\,|\,Y_{1,T},X_1\right)\right]\right\}\\
&= 4\frac{\Phi_0^2(D_m\vee D_{m'})}{n}\mathbb{E}\left[\frac{\delta_1 Y_{1,T}^2}{\bar{G}^2(Y_{1,T})}+r_T^2(\vec{X}_1)-2\,Y_{1,T}r_T(\vec{X}_1)\right]\\
&= 4\frac{\Phi_0^2(D_m\vee D_{m'})}{n}\left\{\mathbb{E}\left[\frac{\delta_1 Y_{1,T}^2}{\bar{G}^2(Y_{1,T})}\right]-\mathbb{E}(r_T^2(\vec{X}_1))\right\}\leq 4\frac{\Phi_0^2(D_m\vee D_{m'})}{n}\mathbb{E}\left[\frac{\delta_1 Z_1^2}{\bar{G}^2(Z_1)}\right]
\end{aligned}$$

which explains that we can choose $\mathrm{pen}(m)=\kappa f_0^{-1}\Phi_0^2\mathbb{E}\left(\delta_1 Z_1^2/\bar{G}^2(Z_1)\right)(D_m/n)$. □

6.9. **Proof of Lemma 6.3.** Let us denote by Π_m the orthogonal projection in $\mathbb{R}^n$ on the subspace $\{(t(\vec{X}_1),\dots,t(\vec{X}_n))',t\in S_m\}$ and by $\hat{Y}_G=(\hat{Y}_{1G},\dots,\hat{Y}_{nG})'$, $r_T(X)=(r_T(\vec{X}_1),\dots,r_T(\vec{X}_n))$ and more generally by $u=(u_1,\dots,u_n)'$. The empirical norm corresponds then to the Euclidean norm in $\mathbb{R}^n$ up to the multiplicative factor $1/n$. We have already mentionned that $\hat{r}_m(X)=\Pi_m\hat{Y}_G$ so that

$$\begin{aligned}
\|\hat{r}_m-r_T\|_n^2 &= \|\Pi_m\hat{Y}_G-r_T(X)\|_n^2\leq 2\|\Pi_m(\hat{Y}_G-Y_G)\|_n^2+2\|\Pi_mY_G-r_T(X)\|_n^2\\
&\leq 2\|\Pi_m(\hat{Y}_G-Y_G)\|_n^2+2\|\Pi_mY_G\|_n^2+2\|r_T(X)\|_n^2\\
&\leq 2\|\hat{Y}_G-Y_G\|_n^2+2\|Y_G\|_n^2+2\|r_T(X)\|_n^2
\end{aligned}$$

by using that $\|\Pi_mu\|_n^2\leq\|u\|_n^2$. First, on the complementary of $\Omega_G=\{\omega,\hat{\bar{G}}(y)\geq c_G/2,\forall y\in[0,T]\}$ defined in Lemma 4.1, $\mathbb{E}(\|\hat{Y}_G-Y_G\|_n^2\mathbb{I}_{\Omega_n^c})$ is bounded by $4T^2(n+1)^2/c_G^2\mathbb{E}(\|\hat{\bar{G}}-G\|_{\infty,T}^2\mathbb{I}_{\Omega_n^c})$

which can be proved to be of order $(1/n)$ mimicking the bound in (6.3). Then, $\|Y_G\|_n^2 \leq T/c_G^2$ implies that $\mathbb{E}[\|Y_G\|_n^2 \mathbb{I}_{\Omega_n^c}]$ is of order $1/n$ as soon as $\mathbb{P}^{1/2}(\Omega_n^c) \leq c/n$. Moreover,

$$\mathbb{E}[\|r_T(X)\|_n^2 \mathbb{I}_{\Omega_n^c}] \leq \sqrt{\mathbb{E}(r_T(\vec{X}_1)^4)}P^{1/2}(\Omega_n^c) \leq \sqrt{\mathbb{E}[\mathbb{E}(r_T(Y_{1,T})|\vec{X}_1)^4]}P^{1/2}(\Omega_n^c)$$

$\leq \sqrt{\mathbb{E}(Y_{1,T}^4)}P^{1/2}(\Omega_n^c) \leq \sqrt{\mathbb{E}(Y_1^4)}P^{1/2}(\Omega_n^c)$ is of order $1/n$ as soon as $\mathbb{P}^{1/2}(\Omega_n^c) \leq c/n$.
Next, by iterative use of the norm equivalence: $\|t\|_f^2/f_1 \leq \|t\|^2 \leq \|t\|_f^2/f_0$ where $\|t\|_f^2 = \int t^2 f$, we obtain:

$$\mathbb{P}(\Omega_n^c) \leq \mathbb{P}\left(\sup_{t\in B_n(0,1)} |\nu_n'(t^2)| \geq c(f_0, f_1)\right)$$

with $\nu_n'(t)$ defined by (6.8) and $c(f_0, f_1)$ a constant depending on f_0 and f_1. This probability is proved to be of order $1/n^2$, as soon as the dimension constraint on N_n given in Lemma 6.2 is satisfied, see the proof of Proposition 7 in [2]. □

References

[1] Y. Baraud, Model selection for regression on a random design, ESAIM Probab. Statist. 6 (2002) 127-146.

[2] Y. Baraud, F. Comte and G. Viennet, Adaptive estimation in autoregression or β-mixing regression via model selection, Ann. Statist 29 (2001) 839-875.

[3] A.R. Barron, L. Birgé and P. Massart, Risk bounds for model selection via penalization, Probab. Theory Relat. Fields 113 (1999) 301-413.

[4] L. Birgé and P. Massart, From model selection to adaptive estimation, in Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics (D. Pollard, E. Torgersen and G. Yang, eds), 1997, 55-87, Springer-Verlag, New-York.

[5] D. Bitouzé, B. Laurent and P. Massart, A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator, Ann. Inst. Henri Poincaré 35 (1999) 735-763.

[6] J. Buckley and I. James, Linear regression with censored data, Biometrika 66 (1979) 429-464.

[7] E. Brunel and F. Comte, Penalized contrast estimation of hazard rate with censored data, Sankhy$\overline{a}$, 67 (2005), 441-475.

[8] D.M. Dabrowska, Nonparametric regression with censored survival regression. Scand. J. Statist. 14 (1987) 181-197.

[9] I. Daubechies, Ten lectures on wavelets, Philadelphia : Society for Industrial and Applied Mathematics, 1992.

[10] R.A. DeVore and G.G. Lorentz, Constructive approximation. Springer-Verlag, 1993.

[11] R.A. DeVore, Nonlinear approximation, Acta Numer. 7 (1998) 51-150.

[12] D.L. Donoho and I.M. Johnstone, Minimax estimation with wavelet shrinkage, Ann. Statist. 26 (1998) 879-921.

[13] J. Fan and I. Gijbels, Censored regression: local linear approximations and their applications, J. of the American Statistical Assoc. 89 (1994) 560-570.

[14] L. Györfi, M. Kohler, A. Krzyzak and H. Walk, A distribution-free theory of nonparametric regression. Springer Series in Statistics, Springer 2002.

[15] S. Gross and T.L. Lai, Nonparametric estimation and regression analysis with left-truncated and right-censored data, J. of the American Statistical Assoc. 91 (1996) 1166-1180.

[16] C. Heuchenne and I. Van Keilegom, Nonlinear regression with censored data, Discussion paper 0512 (2005), Institut de Statistique, Université catholique de Louvain.

[17] R. Hochmuth, Wavelet characterizations for anisotropic Besov spaces, Appl. Comput. Harmon. Anal. 2 (2002) 179-208.

[18] E.L. Kaplan and P. Meier, Nonparametric estimation from incomplete observations, J. Amer. Statist. Assoc. 53 (1958) 457-481.

[19] M. Kohler, S. Kul and K. Màthé, Least squares estimates for censored regression. Preprint 2003. http://www.mathematik.uni-stuttgart.de/mathA/lst3/kohler/hfm-pub-en.html

[20] H. Koul, V. Susarla and J. Van Ryzin, Regression analysis with randomly right-censored data, Ann. Statist. 9 (1981) 1276-1288.

[21] M. Ledoux, On Talagrand's deviation inequalities for product measures, ESAIM Probab. Statist. 1 (1996) 63-87.

[22] S.H. Lo, Y.P. Mack and J.L. Wang, Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator, Probab. Theory Related Fields 80 (1989) 461-473.

[23] R.G. Miller, Least squares regression with censored data, Biometrika 63 (1976) 449-464.

[24] E.A. Nadaraya, On estimating regression. Theory of Probab. and its Appl. 9 (1964) 141-142.

[25] M. Neumann, Multivariate wavelet thresholding in anisotropic function spaces. Statist. Sinica 10 (2000) 399-431.

[26] S.M. Nikol'skiĭ, Approximations of functions of several variables and imbedding theorems, Springer-Verlag, New York, 1975.

[27] J. Park, Optimal global rate of convergence in nonparametric regression with left-truncated and right-censored date, J. of Multivariate Anal. 89 (2004) 70-86.

[28] C.J. Stone, Optimal rates of convergence for nonparametric regression, Ann. Statist. 10 (1982) 1040-1053.

[29] M. Talagrand, New concentration inequalities in product spaces, Invent. Math. 1262 (1996) 505-563.

[30] G.S. Watson, Smooth regression analysis, Sankhy$\overline{a}$ Ser. A, 26 (1964) 359-372.

[31] Z.K. Zheng, Strong consistency of nonparametric regression estimates with censored data, J. Math. Res. Exposition 8 (1988) 307-313.

[32] M. Zhou, M-estimation in censored linear models, Biometrika 79 (1992) 837-841.