Regression function estimation as a partly inverse problem

Fabienne Comte, Valentine Genon-Catalot

REGRESSION FUNCTION ESTIMATION AS A PARTLY INVERSE PROBLEM

F. COMTE[(1)] AND V. GENON-CATALOT[(2)]

Abstract. This paper is about nonparametric regression function estimation. Our estimator is a one step projection estimator obtained by least-squares contrast minimization. The specificity of our work is to consider a new model selection procedure including a cutoff for the underlying matrix inversion, and to provide theoretical risk bounds that apply to non compactly supported bases, a case which was specifically excluded of most previous results. Upper and lower bounds for resulting rates are provided. October 18, 2018

MSC2010 *Subject classifications.* 62G08 - 62M05
Key words and phrases. Hermite basis. Laguerre basis. Model selection. Non parametric estimation. Regression function.

1. Introduction

Consider observations $(X_i, Y_i)_{1\leq i\leq n}$ drawn from the regression model

$$Y_i = b(X_i) + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i) = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2, \quad i = 1, \ldots, n. \tag{1}$$

The random design variables $(X_i)_{1\leq i\leq n}$ are real-valued, independent and identically distributed (i.i.d.) with common density denoted by f, the noise variables $(\varepsilon_i)_{1\leq i\leq n}$ are i.i.d. real-valued and the two sequences are independent. The problem is to estimate the function $b(.) : \mathbb{R} \to \mathbb{R}$ from observations $(X_i, Y_i)_{1\leq i\leq n}$.
Classical nonparametric estimation strategies are of two types. First, Nadaraya (1964) and Watson (1964) methods rely on quotient estimators of type $\widehat{b} = \widehat{bf}/\widehat{f}$, where $\widehat{bf}$ and $\widehat{f}$ are projection or kernel estimators of bf and f. Those methods are popular, especially in the kernel setting. However, they require the knowledge or the estimation of f (see Efromovich (1999), Tsybakov (2009)) and in the latter case, two smoothing parameters.
The second method, proposed by Birgé and Massart (1998), Barron *et al.* (1999), improved by Baraud (2002), is based on a least squares contrast, analogous to the one used for parametric linear regression:

$$\frac{1}{n}\sum_{i=1}^{n} [Y_i - t(X_i)]^2,$$

minimized over functions t that admit a finite development over some orthonormal A-supported $\mathbb{L}^2(A, dx)$ basis, $A \subset \mathbb{R}$. In other words, this is a projection method where the coefficients of the approximate function in the finite basis play the same role as the regression parameters in the linear model. This strategy solves part of the drawbacks of the first one. It provides directly an estimator of b restricted to the set A, a unique

(1): Université Paris Descartes, Laboratoire MAP5, email: fabienne.comte@parisdescartes.fr.
(2): Université Paris Descartes, Laboratoire MAP5, email: valentine.genon-catalot@parisdescartes.fr.

model selection procedure is required and has been proved to realize an adequate squared bias-variance compromise under weak moment conditions on the noise (see Baraud, 2002). Lastly, there is no quotient to make and the rate only depends on the regularity index of b, while in the quotient method it also generally depends on the one of f. These arguments are in favour of the second strategy. Noting that the least squares contrast can be rewritten

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}[t^2(X_i) - 2Y_i t(X_i)], \tag{2}$$

it can be seen that, for a given function t in a finite dimensional linear space included in $\mathbb{L}^2(A, dx)$, three norms must be compared: the integral $\mathbb{L}^2(A, dx)$-norm, $\|t\|_A^2 = \int_A t^2(x)dx$, associated with the basis, the empirical norm involved in the definition of the contrast, $\|t\|_n^2 = n^{-1}\sum_{i=1}^n t^2(X_i)$, and its expectation, corresponding to a $\mathbb{L}^2(A, f(x)dx)$-norm, $\|t\|_f^2 = \int_A t^2(x)f(x)dx$. Due to this difficulty, only compactly supported bases have been considered i.e. the set A on which estimation is done is generally assumed to be compact. This allows to assume that f is lower bounded on A, a condition which would not hold on non compact A. Then, if f is upper and lower bounded on A, the $\mathbb{L}^2(A, f(x)dx)$ and the $\mathbb{L}^2(A, dx)$ norms are equivalent and this makes the problem simpler. Moreover, the equivalence of the norms $\|t\|_n$ and $\|t\|_f$ for t in a finite dimensional linear space must be handled. This is done by Cohen *et al.* (2013) and we take advantage of their findings. However, Cohen *et al.* (2013)'s work has drawbacks: their stability condition is settled in terms of an unknown quantity; the regression function is assumed to be bounded by a known quantity and the definition of the estimator depends on this known bound; they do not study the model selection problem. Due to their statistically simplified setting, they do not deal with the entire partially inverse problem hidden in the procedure.

Our aim in this work is to obtain theoretical results in regression function estimation by the least squares projection method described above, and we want to handle the case of possibly non compact support A of the basis. This explains why we must avoid boundedness assumption on b. A consequence is that the cutoff which has to be introduced in the definition of the estimator depends on the behaviour of the eigenvalues of a random matrix. This requires a specific study to obtain a bound on the integrated $\mathbb{L}^2$ risk, and makes the model selection question near of an inverse problem with unknown operator.

What is the interest of non compactly supported bases? In general, the estimation set and the bases support are considered as fixed in the theoretical part, while are in practice adjusted on the data. With a non compact support, it is not necessary to fix a preliminary definition. Moreover, we have at disposal non compactly supported bases such as the Laguerre ($A = \mathbb{R}^+$) or the Hermite ($A = \mathbb{R}$) basis which have been used recently for nonparametric estimation by projection (see *e.g.* Comte *et al.* 2015, Comte and Genon-Catalot, 2015, 2018, Belomestny *et al.* 2016), showing that theses bases are both convenient and with specific properties. They are especially useful in certain inverse problems (see Mabon, 2017).

Before giving our plan, let us highlight our main findings.

- First, we propose a new procedure of estimation relying on a random cutoff, and generalize Cohen *et al.* (2013)'s results, with a more statistical flavour.
- We deduce from the bias-variance decomposition upper rates of the estimator on specific Sobolev spaces, for which lower bounds are also established. We recover

the standard rates of the "compact case" but also exhibit non standard ones when considering Laguerre or Hermite bases and spaces.

- We propose a model selection procedure for regression function estimation on a set A whether compact or not, where the collection of models itself is random and prove that it reaches automatically a bias-variance tradeoff. We highlight the regression problem as a partially inverse problem: the eigenvalues of the matrix which must be inverted play a role in the problem not directly as a weight on the variance term but in the definition of the collection of models.

The framework and plan of the paper is the following. We fix a set $A \subset \mathbb{R}$ and concentrate on the estimation of the regression function b restricted to a set A, $b_A := b\mathbf{1}_A$. As A may be unbounded, we do not want to assume that $b_A \in \mathbb{L}^2(A, dx)$ which would exclude linear or polynomial functions. Our main assumption is that $b_A \in \mathbb{L}^4(A, f(x)dx)$, i.e. $\mathbb{E}b_A^4(X_1) < +\infty$ which is rather weak. In Section 2, we define the projection estimator of the regression function b_A and check that the most elementary risk bound holds without any constraint on the support A or the projection basis. In Section 3, we prove a risk-bound for the estimator on one model, borrowing some elements to Cohen *et al.* (1993)'s results to extend them. Then, we study rates and optimality for the integrated $\mathbb{L}^2(A, f(x)dx)$-risk. Introducing regularity spaces linked with f, we prove upper and matching lower bounds for our projection estimator. Then we quickly show how to recover existing results for compactly supported bases and more precisely illustrate the case of non compact support with the Hermite and Laguerre bases for estimation on $A = \mathbb{R}$ and $A = \mathbb{R}^+$ respectively. In Section 4, we propose a model selection strategy on a random collection of models taking into account a possible inversion problem of the matrix allowing a unique definition of the estimator. A risk bound for the adaptive estimator is provided both for the integrated empirical risk and for the integrated $\mathbb{L}^2(A, f(x)dx)$-risk: it generalizes existing results to non compactly supported bases. Section 5 gives some concluding remarks. Most proofs are gathered in Section 6 while Section 7 gives theoretical tools used along the proofs. An appendix is devoted to numerical illustrations.

2. Projection estimator and preliminary results

Recall that f denotes the density of X_1. In the following, $\|.\|_{2,p}$ denotes the euclidean norm in $\mathbb{R}^p$. For $A \subset \mathbb{R}$, $\|.\|_A$ denotes the integral norm in $\mathbb{L}^2(A, dx)$, $\|.\|_f$ the integral norm in $\mathbb{L}^2(A, f(x)dx)$ and $\|.\|_\infty$ the supremum norm on A. For any function h, $h_A = h\mathbf{1}_A$.

2.1. Definition of the projection estimator.

Consider model (1). Let $A \subset \mathbb{R}$ and let $(\varphi_j, j = 0, \dots, m-1)$ be an orthonormal system of A-supported functions belonging to $\mathbb{L}^2(A, dx)$. Define $S_m = \text{span}(\varphi_0, \dots, \varphi_{m-1})$, the linear space spanned by $(\varphi_0, \dots, \varphi_{m-1})$. Note that the φ_j's may depend on m but for simplicity, we omit this in the notation.
We assume that for all j, $\int \varphi_j^2(x) f(x)dx < +\infty$ so that $S_m \subset \mathbb{L}^2(A, f(x)dx)$ and define a projection estimator of the regression function b on A, by

$$\hat{b}_m = \arg\min_{t \in S_m} \gamma_n(t)$$

where $\gamma_n(t)$ is defined in (2). For functions s, t, we set

$$\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^n t^2(X_i) \quad \text{and} \quad \langle s, t\rangle_n := \frac{1}{n}\sum_{i=1}^n s(X_i)t(X_i),$$

and write

$$\langle \vec{u}, t\rangle_n = \frac{1}{n}\sum_{i=1}^{n} u_i t(X_i)$$

when $\vec{u}$ is the vector $(u_1,\dots,u_n)'$, $\vec{u}\,'$ denotes the transpose of $\vec{u}$ and t is a function. We introduce the classical matrices

$$\widehat{\Phi}_m = (\varphi_j(X_i))_{1\le i\le n, 0\le j\le m-1},$$

and

$$\widehat{\Psi}_m = (\langle\varphi_j,\varphi_k\rangle_n)_{0\le j,k\le m-1} = \frac{1}{n}\widehat{\Phi}'_m\widehat{\Phi}_m,\ \Psi_m = \left(\int \varphi_j(x)\varphi_k(x)f(x)dx\right)_{0\le j,k\le m-1} = \mathbb{E}(\widehat{\Psi}_m). \tag{3}$$

Set $\vec{Y} = (Y_1,\dots,Y_n)'$ and define $\vec{\hat{a}}^{(m)} = (\hat{a}_0^{(m)},\dots,\hat{a}_{m-1}^{(m)})'$ as the m-dimensional vector such that $\hat{b}_m = \sum_{j=0}^{m-1}\hat{a}_j^{(m)}\varphi_j$. Assuming that $\widehat{\Psi}_m$ is invertible almost surely (a.s.) yields:

$$\hat{b}_m = \sum_{j=0}^{m-1}\hat{a}_j^{(m)}\varphi_j, \quad \text{with} \quad \vec{\hat{a}}^{(m)} = (\widehat{\Phi}'_m\widehat{\Phi}_m)^{-1}\widehat{\Phi}'_m\vec{Y} = \frac{1}{n}\widehat{\Psi}_m^{-1}\widehat{\Phi}'_m\vec{Y}. \tag{4}$$

2.2. **Bound on the mean empirical risk on a fixed space.** We now evaluate the risk of the estimator, without any constraint on the basis support. Though classical, the result hereafter requires noteworthy comments.

Proposition 2.1. *Let $(X_i,Y_i)_{1\le i\le n}$ be observations drawn from model (1) and denote by $b_A = b\mathbf{1}_A$. Assume that $b_A\in\mathbb{L}^2(A,f(x)dx)$ and that $\widehat{\Psi}_m$ is a.s. invertible. Consider the least squares estimator $\hat{b}_m$ of b, given by (4). Then*

$$\mathbb{E}\big[\|\hat{b}_m - b_A\|_n^2\big] = \mathbb{E}\left(\inf_{t\in S_m}\|t-b_A\|_n^2\right) + \sigma_\varepsilon^2\frac{m}{n}, \tag{5}$$

$$\le \inf_{t\in S_m}\left[\int (b_A-t)^2(x)f(x)dx\right] + \sigma_\varepsilon^2\frac{m}{n}. \tag{6}$$

Note that

$$\inf_{t\in S_m}\left[\int (b_A-t)^2(x)f(x)dx\right] = \|b_A - b_m^f\|_f^2$$

where b_m^f is the $\mathbb{L}^2(A,f(x)dx)$-orthogonal projection of b_A on S_m, i.e. if Ψ_m is invertible, we get $b_m^f = \sum_{j=0}^{m-1}a_j^f(b)\varphi_j$ where

$$(a_0^f(b),\dots,a_{m-1}^f(b))' = \Psi_m^{-1}\overrightarrow{(b\varphi)}_m, \text{ with } \overrightarrow{(b\varphi)}_m = (\langle b,\varphi_0\rangle_f,\dots,\langle b,\varphi_{m-1}\rangle_f)'.$$

This implies that the bias bound is equal to

$$\|b_A - b_m^f\|_f^2 = \|b_A\|_f^2 - \|b_m^f\|_f^2 = \int_A b^2(x)f(x)dx - \overrightarrow{(b\varphi)}'_m\Psi_m^{-1}\overrightarrow{(b\varphi)}_m.$$

It is not obvious from (6) or from the previous formula that the bias term is small when m is large. Therefore, two questions arise: is Ψ_m invertible for any m, and does the bias tend to zero when m grows to infinity? The Lemmas below provide sufficient conditions. These conditions can be refined if the basis is specified.

Lemma 2.1. *Assume that* $\lambda(A \cap \mathrm{supp}(f)) > 0$ *where* λ *is the Lebesgue measure and* $\mathrm{supp}(f)$ *the support of* f*, that the* $(\varphi_j)_{0\leq j\leq m-1}$ *are continuous, and that there exist* $x_0, \ldots, x_{m-1} \in A \cap \mathrm{supp}(f)$ *such that* $\det[(\varphi_j(x_k))_{0\leq j,k\leq m-1}] \neq 0$*. Then,* Ψ_m *is invertible.*

Lemma 2.2. *Assume that* $b_A \in \mathbb{L}^2(A, f(x)dx)$*. Assume that* $(\varphi_j)_{j\geq 0}$ *is an orthonormal basis of* $\mathbb{L}^2(A, dx)$ *such that, for all* $j \geq 0$*,* $\int \varphi_j^2(x)f(x)dx < +\infty$*, that* f *is bounded on* A *and that for all* $x \in A$*,* $f(x) > 0$*.*
Then $\inf_{t\in S_m}\left[\int (b_A - t)^2(x)f(x)dx\right]$ *tends to 0 when* m *tends to infinity.*

Lemma 2.1 follows from the following equality. For all $\vec{u} = (u_0, \ldots, u_{m-1})' \in \mathbb{R}^m \setminus \{\vec{0}\}$, for $t(x) = \sum_{j=0}^{m-1} u_j\varphi_j(x)$, $\vec{u}'\,\Psi_m\,\vec{u} = \|t\|_f^2 = \int_A t^2(x)f(x)dx \geq 0$. Under the assumptions, the result follows.
The proof of Lemma 2.2 is elementary. Note that $\int (b_A - t)^2(x)f(x)dx = \|b_A - t\|_f^2 = \|b_A\sqrt{f} - t\sqrt{f}\|_A^2$. Under the assumptions of Lemma 2.2, the system $\phi_j = \varphi_j\sqrt{f}$, $j \geq 0$ is a complete family of $\mathbb{L}^2(A, dx)$. Indeed, if $g \in \mathbb{L}^2(A, dx)$, $\int g\phi_j = \int \varphi_j(g\sqrt{f}) = 0\ \forall j \geq 0$ implies $g = 0$ using our assumptions.
The result of Proposition 2.1 is general in the sense that it holds for any basis support, whether compact or not. We stress that (5) is an equality, in particular the variance term is **exactly** equal to $\sigma_\varepsilon^2 m/n$. In addition, the result does not depend on the basis.

Remark 2.1. *We underline that the latter fact is not obvious. Consider the density estimation setting, where* $\hat{f}_m = \sum_{j=0}^{m-1} \widehat{c}_j\varphi_j$ *with* $\widehat{c}_j = (1/n)\sum_{i=1}^n \varphi_j(X_i)$ *is a projection estimator of* f*. Then the integrated* $\mathbb{L}^2-$*risk bound is*

$$\mathbb{E}(\|\hat{f}_m - f_A\|^2) = \inf_{t\in S_m} \|f_A - t\|^2 + \frac{\sum_{j=0}^{m-1}\mathbb{E}[\varphi_j^2(X_1)]}{n} - \frac{\|f_m\|^2}{n},$$

where $f_m = \sum_{j=0}^{m-1}\langle f, \varphi_j\rangle\varphi_j$ *is the* $\mathbb{L}^2(dx)$*-orthogonal projection of* f *on* S_m*. The variance term has the order of* $\sum_{j=0}^{m-1}\mathbb{E}[\varphi_j^2(X_1)]/n$*. For most compactly supported bases, this term has order* m/n *(for instance, it is equal to* m/n *for histograms or trigonometric polynomial basis, see section 3.3); but it is proved in Comte and Genon-Catalot (2018) that for Laguerre or Hermite basis (see section 3.4 below), this term has exactly the order* $\sqrt{m}/n$ *(lower and upper bound are provided, under weak assumptions). This is why it is important to see that, in regression context, the variance order does not depend on the basis.*

2.3. **Useful inequalities.** For M a matrix, we denote by $\|M\|_{\mathrm{op}}$ the operator norm defined as the square root of the largest eigenvalue of MM'. If M is symmetric, it coincides with $\sup\{|\lambda_i|\}$ where λ_i are the eigenvalues of M. Moreover, if M, N are two matrices with compatible product MN, then, $\|MN\|_{\mathrm{op}} \leq \|M\|_{\mathrm{op}}\|N\|_{\mathrm{op}}$.
The possible values of the dimension m to study the collection $(\hat{b}_m)$ of estimators are subject to restrictions, for which the following property is important:

Proposition 2.2. *Assume that the spaces* S_m *are nested (i.e.* $m \leq m' \Rightarrow S_m \subset S_{m'}$*) and* Ψ_m *(resp.* $\widehat{\Psi}_m$*) is invertible, then* $m \mapsto \|\Psi_m^{-1}\|_{\mathrm{op}}$ *(resp* $m \mapsto \|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}$*) is nondecreasing.*

Let us define

$$L(m) = \sup_{x\in A}\sum_{j=0}^{m-1}\varphi_j^2(x) \ \text{ and assume } \ L(m) < +\infty. \tag{7}$$

This quantity is independent of the choice of the $\mathbb{L}^2(dx)$-orthonormal basis of S_m, and for nested spaces S_m, the map $m \mapsto L(m)$ is increasing. We need to study the set

$$\Omega_m(\delta) = \left\{ \sup_{t\in S_m,\, t\neq 0} \left| \frac{\|t\|_n^2}{\|t\|_f^2} - 1 \right| \leq \delta \right\} \tag{8}$$

where the empirical and the $\mathbb{L}^2(A, f(x)dx)$ norms are equivalent on S_m. Theorem 1 in Cohen *et al.* (1993) provides the adequate inequality. In our context, it takes the following form:

Proposition 2.3. *Let $\widehat{\Psi}_m$, Ψ_m be the $m\times m$ matrices defined in Equation* (3) *and assume that Ψ_m is invertible. Then for all $0 \leq \delta \leq 1$,*

$$\mathbb{P}(\Omega_m(\delta)^c) = \mathbb{P}\left[\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} > \delta\right] \leq 2m\exp\left(-c(\delta)\frac{n}{L(m)(\|\Psi_m^{-1}\|_{\mathrm{op}} \vee 1)}\right).$$

where Id_m denotes the $m \times m$ identity matrix and $c(\delta) = \delta + (1-\delta)\log(1-\delta)$.

As a consequence, we obtain that, choosing $\delta = 1/2$, the set $\Omega_m := \Omega_m(1/2)$ satisfies $\mathbb{P}(\Omega_m^c) \leq 2n^{-4}$ if m is such that

$$L(m)(\|\Psi_m^{-1}\|_{\mathrm{op}} \vee 1) \leq \frac{\mathfrak{c}}{2}\frac{n}{\log(n)}, \quad \mathfrak{c} = \frac{1-\log(2)}{5}. \tag{9}$$

Condition (9) can be understood as ensuring the stability of the least-squares estimator, as underlined in Cohen *et al.* (2013). However, the stability condition therein relies on a theoretical quantity (see $K(m)$ defined in (34) and Lemma 6.2 below). We stress that $L(m)$ is explicitly computable and Ψ_m^{-1} can be estimated by $\widehat{\Psi}_m^{-1}$. Moreover, we can prove:

Proposition 2.4. `(i)` *Assume that f is bounded. Let $\widehat{\Psi}_m$ be the $m \times m$ matrix defined in Equation* (3)*. Then for all $u > 0$*

$$\mathbb{P}\left[\|\Psi_m - \widehat{\Psi}_m\|_{\mathrm{op}} \geq u\right] \leq 2m\exp\left(-\frac{n\,u^2/2}{L(m)\,(\|f\|_\infty + 2\,u/3)}\right).$$

`(ii)` *Assume that $\widehat{\Psi}_m, \Psi_m$ are (a.s.) invertible. Then for $\alpha > 0$,*

$$\left\{\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\mathrm{op}} > \alpha\|\Psi_m^{-1}\|_{\mathrm{op}}\right\} \subset \left\{\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} > \frac{\alpha\wedge 1}{2}\right\}.$$

3. Truncated estimator on a fixed space

We may consider from Proposition 2.1 that the problem is standard. However, it is known that difficulties arise if we want to bound the integrated $\mathbb{L}^2$-risk instead of the empirical risk, even for fixed m. Actually, the general regression problem is an inverse problem since the link between the function of interest b and the density of the observations $(Y_i, X_i)_i$ is of convolution type $f_Y(y) = \int f_\varepsilon(y - b(x))f(x)dx$ where f_Y and f_ε are the densities of Y_1 and ε_1. This can also be seen from the fact that the estimator is computed via the inversion of the matrix $\widehat{\Psi}_m$. Thus we can expect that the procedure depends on the eigenvalues of Ψ_m.

3.1. **Integrated risk bound.** Let us assume as above that $b_A \in \mathbb{L}^2(A, f(x)dx)$. It is not possible to deduce from Proposition 2.1 a bound on $\mathbb{E}[\|\hat{b}_m - b_A\|_f^2]$ for all m such that $\widehat{\Psi}_m$ is invertible. On the other hand, we introduce a cutoff and define

$$\tilde{b}_m := \hat{b}_m \mathbf{1}_{L(m)(\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}} \vee 1) \leq \mathfrak{c}n/\log(n)}, \tag{10}$$

where $L(m)$ is defined by (7) and $\mathfrak{c}$ in (9). On the set $\{L(m)(\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}} \vee 1) \leq \mathfrak{c}n/\log(n)\}$, the matrix $\widehat{\Psi}_m$ is invertible and its eigenvalues $(\lambda_i)_{1\leq i\leq m}$ satisfy $\inf_{1\leq i\leq m}(\lambda_i) \geq m\log(n)/(\mathfrak{c}n)$. Analogously, condition (9) is equivalent to the fact that Ψ_m is invertible and its eigenvalues are lower bounded by $2m\log(n)/(\mathfrak{c}n)$. We have:

Proposition 3.1. *Assume that $\mathbb{E}(\varepsilon_1^4) < +\infty$ and $b_A \in \mathbb{L}^4(A, f(x)dx)$. Then for any m satisfying (9), we have*

$$\mathbb{E}\big[\|\tilde{b}_m - b_A\|_f^2\big] \leq \left(1 + \frac{8\mathfrak{c}}{\log(n)}\right) \inf_{t\in S_m} \|b_A - t\|_f^2 + 8\sigma_\varepsilon^2 \frac{m}{n} + \frac{c}{n}, \tag{11}$$

where c is a constant depending on $\mathbb{E}(\varepsilon_1^4)$ and $\int b_A^4(x)f(x)dx$.

The proof of Proposition 3.1 exploits as a first step the proof of Theorem 3 in Cohen *et al.* (2013). However, the estimator in Cohen *et al.* (2013) is mainly theoretical: indeed they assume that b is bounded and the estimator depends on the bound, which has thus to be known. As A may be unbounded, it is important to get rid of this restriction.

3.2. **Rate and optimality.** So far, the bias rate of the $\mathbb{L}^2(A, f(x)dx)$-risk in (6) and (11) has not been assessed. To this end, we introduce regularity spaces related to f by setting:

$$W_f^s(A, R) = \left\{h \in \mathbb{L}^2(A, f(x)dx), \forall \ell \geq 1, \|h - h_\ell^f\|_f^2 \leq R\ell^{-s}\right\} \tag{12}$$

where we recall that h_ℓ^f is the $\mathbb{L}^2(A, f(x)dx)$-orthogonal projection of h on S_ℓ.
From (11), we easily deduce an upper bound for the risk, which we state below. The risk rate is optimal, as we also prove the following lower bound.

Theorem 3.1. *Assume that $b_A \in W_f^s(A, R)$, condition (25) holds and that $m_{\mathrm{opt}} := n^{1/(s+1)}$ satisfies (9).*

- *Upper bound.* $\mathbb{E}(\|\tilde{b}_{m_{\mathrm{opt}}} - b_A\|_f^2) \leq Cn^{-s/(s+1)}$.
- *Lower bound. Assume in addition that $\varepsilon_1 \sim \mathcal{N}(0, \sigma_\varepsilon^2)$,*

$$\liminf_{n\to+\infty} \inf_{T_n} \sup_{b_A \in W_f^s(A,R)} \mathbb{E}_{b_A}[n^{s/(s+1)}\|T_n - b_A\|_f^2] \geq c$$

where $\inf_{T_n}$ denotes the infimum over all estimators and where the constant $c > 0$ depends on s and R.

The condition that $m_{\mathrm{opt}} = n^{1/(s+1)}$ satisfies (9) is actually mainly a constraint on f, see the discussion at the end of Section 3.4.
The partly inverse problem appears here. The rate of $\|\Psi_m^{-1}\|_{\mathrm{op}}$ as a function of m is to be interpreted as a measure of the degree of ill-posedness of the inverse problem, in the context of regression function estimation.

Proposition 3.2. *Under the assumptions of Theorem 3.1, if moreover* $L(m) \asymp m$ *and* $\|\Psi_m^{-1}\|_{\text{op}} \asymp m^k$, *then*

$$\mathbb{E}[\|\hat{b}_{\hat{m}} - b_A\|_f^2] \leq C(R) n^{-\frac{s}{(s\vee k)+1}}.$$

This result is due to the fact that the constraint $L(m)\|\Psi_m^{-1}\|_{\text{op}} \asymp m^{k+1} \lesssim n/\log(n)$ has to be fulfilled for m_{opt}.

3.3. **Case of compact A and compactly supported bases.** In this section, we assume that A is compact and give examples of bases where, for simplicity, $A = [0,1]$.

Classical compactly supported bases are: histograms $\varphi_j(x) = \sqrt{m}\mathbf{1}_{[j/m,(j+1)/m[}(x)$, for $j = 0, \ldots, m-1$; piecewise polynomials with degree r (rescaled Legendre basis up to degree r on each subinterval $[j/m_r, (j+1)/m_r[$, with $m = (r+1)m_r$); compactly supported wavelets; trigonometric basis with odd dimension m, $\varphi_0(x) = \mathbf{1}_{[0,1]}(x)$ and $\varphi_{2j-1}(x) = \sqrt{2}\cos(2\pi jx)\mathbf{1}_{[0,1]}(x)$, and $\varphi_{2j}(x) = \sqrt{2}\sin(2\pi jx)\mathbf{1}_{[0,1]}(x)$ for $j = 1, \ldots, (m-1)/2$.

For histograms and trigonometric basis, $L(m) = m$, for piecewise polynomials with degree r, $L(m) = (r+1)m$. Compactly supported wavelets also satisfy (7) with $L(m)$ of order m. The trigonometric spaces are nested; for histograms, piecewise polynomials and wavelets, the models are nested if the subdivisions are diadic ($m = 2^k$ for increasing values of k).

Let $P_k(x) = \sqrt{2}L_k(2x-1)\,\mathbf{1}_{[0,1]}(x)$, for $k = 0, \ldots, m-1$ be the Legendre polynomial basis rescaled from $[-1,1]$ to $[0,1]$. It is an $\mathbb{L}^2([0,1], dx)$ orthonormal basis of $S_m = \text{span}(P_0, \ldots, P_{m-1})$. As $\|P_k\|_\infty = \sqrt{2}\sqrt{2k-1}$, we get $L(m) = 2m^2$ (see Cohen *et al.* (2013)).

If A is compact, one can assume that

$$\exists f_0 > 0, \text{ such that } \forall x \in A, \quad f(x) > f_0. \tag{13}$$

This assumption is commonly and crucially used in papers on nonparametric regression. In particular, it implies that Ψ_m is invertible, and more precisely:

Proposition 3.3. *Assume that Assumption (13) is satisfied, then*

$$\forall m \leq n, \quad \|\Psi_m^{-1}\|_{\text{op}} \leq 1/f_0.$$

Indeed (13) implies that, for $\vec{u} = (u_0, \ldots, u_{m-1})'$ a vector of $\mathbb{R}^m$,

$$\vec{u}\,'\,\Psi_m\,\vec{u} = \int_A \left(\sum_{j=0}^{m-1} u_j\varphi_j(x)\right)^2 f(x)dx \;\geq f_0 \int_A \left(\sum_{j=0}^{m-1} u_j\varphi_j(x)\right)^2 dx = f_0\|\vec{u}\|_{2,m}^2. \tag{14}$$

Therefore $\|\Psi_m^{-1}\|_{\text{op}} \leq 1/f_0$ and Proposition 3.3 is proved. A consequence of (13) is that the matrix Ψ_m needs not appear in condition (9), thus the matrix $\widehat{\Psi}_m$ needs not appear in the definition of $\tilde{b}_m$. So we can define, as in Baraud (2002), for c' a constant,

$$\tilde{b}_m = \hat{b}_m \mathbf{1}_{L(m) \leq c' n/\log(n)}. \tag{15}$$

Now, let us discuss about the usual rates in this compact setting. Assume that

$$b_A \in \mathbb{L}^2(A, dx) \text{ and } \|f\|_\infty < +\infty. \tag{16}$$

Then $\forall t \in S_m$, $\|b_A - t\|_f^2 \leq \|f\|_\infty \|b_A - t\|_A^2$ and thus

$$\inf_{t \in S_m} \|b_A - t\|_f^2 \leq \|f\|_\infty \|b_A - b_m\|_A^2 \tag{17}$$

where b_m is the $\mathbb{L}^2(A,dx)$-orthogonal projection of b_A on S_m. Thus we recover a classical bias, and the bias-variance compromise leads to standard rates, typically $n^{-2\alpha/(2\alpha+1)}$ for $b_A \in \mathcal{B}_{\alpha,2,\infty}(A,R)$ a Besov ball with radius R and regularity α (see De Vore and Lorentz (1993), or Baraud (2002, section 2)).

3.4. **Examples of non compact A and non compactly supported bases.** If A is not compact, assumption (13) can not hold, therefore we can not get rid of the matrix Ψ_m. Our contribution is to take into account and enlight the role of Ψ_m and to introduce a new selection procedure involving a random collection of models (see Section 4).

Now we assume that

$$b_A \in \mathbb{L}^2(A, f(x)dx), \quad \lambda(A \cap \mathrm{supp}(f)) > 0, \text{ and } f \text{ is upper bounded.} \tag{18}$$

We give two concrete examples of non compactly supported bases: the Laguerre basis on $A = \mathbb{R}^+$ and the Hermite basis on $A = \mathbb{R}$. See *e.g.* Comte and Genon-Catalot (2018) for density estimation by projection using these bases.

• Laguerre basis, $A = \mathbb{R}^+$. Consider the Laguerre polynomials (L_j) and the Laguerre functions (ℓ_j) given by

$$L_j(x) = \sum_{k=0}^{j} (-1)^k \binom{j}{k} \frac{x^k}{k!}, \qquad \ell_j(x) = \sqrt{2} L_j(2x) e^{-x} \mathbf{1}_{x\geq 0}, \quad j \geq 0. \tag{19}$$

The collection $(\ell_j)_{j\geq 0}$ constitutes a complete orthonormal system on $\mathbb{L}^2(\mathbb{R}^+)$, and is such that (see Abramowitz and Stegun (1964)):

$$\forall j \geq 0, \ \ \forall x \in \mathbb{R}^+, \ \ |\ell_j(x)| \leq \sqrt{2}. \tag{20}$$

Clearly, the collection of models $(S_m = \mathrm{span}\{\ell_0, \ldots, \ell_{m-1}\})$ is nested, and (20) implies that this basis satisfies (7) with $L(m) = 2m$ (the supremum is attained at $x = 0$).

• Hermite basis, $A = \mathbb{R}$. The Hermite polynomial and the Hermite function of order j are given, for $j \geq 0$, by:

$$H_j(x) = (-1)^j e^{x^2} \frac{d^j}{dx^j}(e^{-x^2}), \quad h_j(x) = c_j H_j(x) e^{-x^2/2}, \quad c_j = \left(2^j j! \sqrt{\pi}\right)^{-1/2} \tag{21}$$

The sequence $(h_j, j \geq 0)$ is an orthonormal basis of $\mathbb{L}^2(\mathbb{R}, dx)$. The infinite norm of h_j satisfies (see Abramowitz and Stegun (1964), Szegö (1959) p.242):

$$\|h_j\|_\infty \leq \Phi_0, \qquad \Phi_0 \simeq 1,086435/\pi^{1/4} \simeq 0.8160, \tag{22}$$

so that the Hermite basis satisfies (7) with $L(m) \leq \Phi_0^2 m$. The collection of models is also clearly nested.

Hereafter, we use the notation φ_j to denote ℓ_j in the Laguerre case and h_j in the Hermite case. We denote by $S_m = \mathrm{span}(\varphi_0, \varphi_1, \ldots, \varphi_{m-1})$ the linear space generated by the m functions $\varphi_0, \ldots, \varphi_{m-1}$ and by $f_m = \sum_{j=0}^{m-1} a_j(f)\varphi_j$ the orthogonal projection of f on S_m. Then $a_j(f) = \langle f, \varphi_j \rangle$ will mean the integral of $f\,\varphi_j$ either on $\mathbb{R}$ or on $\mathbb{R}^+$.

As the bases functions are bounded, the terms $\int \varphi_j^2 f$ are finite. Moreover, the assumptions of Lemma 2.2 hold, so that the bias term in Proposition 2.1 tends to zero as $m \to +\infty$.

The matrices Ψ_m, $\widehat{\Psi}_m$ in these bases have specific properties:

Lemma 3.1. *For all $m \in \mathbb{N}$, for all $m \leq n$, $\widehat{\Psi}_m$ is a.s. invertible.*

The result below on Ψ_m is crucial for understanding our procedure.

Proposition 3.4. *For all m, Ψ_m is invertible and there exists a constant $c^\star$ such that,*

$$\|\Psi_m^{-1}\|_{\text{op}}^2 \geq c^\star m. \tag{23}$$

In the Laguerre and Hermite cases, Inequality (23) clearly implies that $\|\Psi_m^{-1}\|_{\text{op}}$ cannot be uniformly bounded in m contrary to the case of compactly supported bases. This means that the constraint in (9) leads to restrictions on the values m, as illustrated by the next proposition.

Proposition 3.5. *Consider the Laguerre or the Hermite basis. Assume that $f(x) \geq c/(1+x)^k$ for $x \geq 0$ in the Laguerre case or $f(x) \geq c/(1+x^2)^k$ for $x \in \mathbb{R}$ in the Hermite case. Then for m large enough, $\|\Psi_m^{-1}\|_{\text{op}} \leq Cm^k$.*

We performed numerical experiments which seem to indicate that the order m^k is sharp. If f is as in Proposition 3.5, Proposition 3.2 applies: the optimal rate of order $n^{-s/(s+1)}$ can be reached by the adaptive estimator only if $s > k$. Note that in a Sobolev-Laguerre ball:

$$W^s(\mathbb{R}^+, R) = \{h \in \mathbb{L}^2(A, dx), \quad \sum_{j\geq 0} j^s \langle h, \ell_j\rangle^2 \leq R\}, \tag{24}$$

the index s (and not $2s$) is linked with regularity properties of functions (see Section 7 of Comte and Genon-Catalot (2015) and Section 7.2 of Belomestny *et al.* (2016)). The same type of property holds for Sobolev-Hermite balls, see Belomestny *et al.* (2017). Therefore, the rate $n^{-s/(s+1)}$ is non standard[1].

In density estimation using projection methods on Laguerre or Hermite bases, the variance term in the risk bound of projection estimators has order $\sqrt{m}/n$ so that the optimal rate on a Sobolev-Laguerre or Sobolev-Hermite ball for the estimators risk is $n^{-2s/(2s+1)}$ (see Remark 2.1). It seems that, in the regression setting, we cannot have such a gain. Analogous considerations hold with the Hermite basis.

4. Adaptive procedure

Let us consider now the following assumptions.

(A1) The collection of spaces S_m is nested (that is $S_m \subset S_{m'}$ for $m \leq m'$) and such that, for each m, the basis $(\varphi_0, \ldots, \varphi_{m-1})$ of S_m satisfies

$$\forall m \geq 1, \quad L(m) = \|\sum_{j=0}^{m-1} \varphi_j^2\|_\infty \leq c_\varphi^2 m \quad \text{for} \quad c_\varphi^2 > 0 \quad \text{a constant.} \tag{25}$$

(A2) $\|f\|_\infty < +\infty$.

We present now a model selection procedure and associated risk bounds. To select the most relevant space S_m, we proceed by choosing

$$\hat{m} = \arg \min_{m \in \widehat{\mathcal{M}}_n} \left\{ -\|\hat{b}_m\|_n^2 + \kappa \sigma_\varepsilon^2 \frac{m}{n} \right\} \tag{26}$$

[1] If b_A is a combination of Γ-type functions, then the bias term $\inf_{t\in S_m} \|b_A - t\|^2$ is much smaller (exponentially decreasing) and the rate $\log(n)/n$ can be reached by the adaptive estimator (see e.g. Mabon (2017)).

where κ is a numerical constant, and $\widehat{\mathcal{M}}_n$ is a random collection of models defined by

$$\widehat{\mathcal{M}}_n = \left\{ m \in \mathbb{N}, m(\|\widehat{\Psi}_m^{-1}\|_{\text{op}}^2 \vee 1) \leq \mathfrak{d} \frac{n}{\log(n)} \right\}, \quad \mathfrak{d} = \frac{1}{192\, c_\varphi^2 (\|f\|_\infty \vee 1 + (1/3))}. \tag{27}$$

The value of the constant $\mathfrak{d}$ is determined below by Lemma 6.6.
A theoretical counterpart of $\widehat{\mathcal{M}}_n$, with $\mathfrak{d}$ is defined in (27), is useful:

$$\mathcal{M}_n = \left\{ m \in \mathbb{N}, m\ (\|\Psi_m^{-1}\|_{\text{op}}^2 \vee 1) \leq \frac{\mathfrak{d}}{4} \frac{n}{\log(n)} \right\}. \tag{28}$$

Note that the cutoff for defining $\hat{m}$ and $\hat{b}_{\hat{m}}$ is different from the one used in (10). As $m(\|\widehat{\Psi}_m^{-1}\|_{\text{op}} \vee 1) \leq m(\|\widehat{\Psi}_m^{-1}\|_{\text{op}}^2 \vee 1)$, this yields a smaller set of possible values for $\hat{m}$.
The procedure (26) aims at performing an automatic bias-variance tradeoff. Each term is related to the bias or the variance obtained in Proposition 2.1. The squared bias term is equal to $\|b_A - b_m^f\|_f^2 = \|b_A\|_f^2 - \|b_m^f\|_f^2$ where b_m^f is the $\mathbb{L}^2(A, f(x)dx)$-orthogonal projection of b_A on S_m. The first term $\|b_A\|_f^2$ is unknown but does not depend on m; on the other hand, $\|b_m^f\|_f^2 = \mathbb{E}[\|b_m^f\|_n^2]$. Thus, the quantity $-\|\hat{b}_m\|_n^2$ approximates the squared bias, up to an additive constant, while $\sigma_\varepsilon^2 m/n$ has the variance order.

Theorem 4.1. *Let $(X_i, Y_i)_{1 \leq i \leq n}$ be observations from model (1). Assume that* **(A1)**, **(A2)** *hold, that $\mathbb{E}(\varepsilon_1^6) < +\infty$ and $\mathbb{E}\big[b^4(X_1)\big] < +\infty$. Then, there exists a numerical constant κ_0 such that for $\kappa \geq \kappa_0$, we have*

$$\mathbb{E}\big[\|\hat{b}_{\hat{m}} - b_A\|_n^2\big] \leq C \inf_{m \in \mathcal{M}_n} \left(\inf_{t \in S_m} \|b_A - t\|_f^2 + \kappa \sigma_\varepsilon^2 \frac{m}{n} \right) + \frac{C'}{n}, \tag{29}$$

and

$$\mathbb{E}\big[\|\hat{b}_{\hat{m}} - b_A\|_f^2\big] \leq C_1 \inf_{m \in \mathcal{M}_n} \left(\inf_{t \in S_m} \|b_A - t\|_f^2 + \kappa \sigma_\varepsilon^2 \frac{m}{n} \right) + \frac{C_1'}{n} \tag{30}$$

where C, C_1 are a numerical constants and C', C_1' are constants depending on $\|f\|_\infty$, $\mathbb{E}[b^4(X_1)]$, $\mathbb{E}(\varepsilon_1^6)$.

Theorem 4.1 shows that the risk of the estimator $\hat{b}_{\hat{m}}$ automatically realizes the bias-variance tradeoff, up to the multiplicative constants C, C_1, both in term of empirical and of integrated $\mathbb{L}^2(A, f(x)dx)-$risk. The conditions are general, rather weak, and do not impose any support constraint. Theorem 4.1 contains existing results when the bases are regular and compactly supported.

Remark 4.1. *The constant $\mathfrak{d}$ in the definition of $\widehat{\mathcal{M}}_n$ depends on $\|f\|_\infty$ which is unknown. In practice, this quantity has to be replaced by a rough estimator. Otherwise, we can replace the bound $\mathfrak{d}n/\log(n)$ in $\widehat{\mathcal{M}}_n$ by $n/\log^2(n)$ and assume that n is large enough.*
The constant σ_ε^2 is also generally unknown, and must be replaced by an estimator. We simply propose to use the residual least-squares estimator:

$$\widehat{\sigma_\varepsilon^2} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{b}_{m^*}(X_i))^2$$

where m^* is an arbitrarily chosen dimension, which must be neither too large, nor too small; for instance $m^* = \lfloor \sqrt{n} \rfloor$. See e.g. Baraud (2000), section 6.

5. Concluding remarks

In this paper, we study nonparametric regression function estimation by a projection method which was first proposed by Birgé and Massart (1998) and Barron *et al.* (1999). Compared with the popular Nadaraya-Watson approach, the projection method has several advantages. In the Nadaraya-Watson method, one estimates b by a quotient of estimators, namely $\widehat{b} = \widehat{bf}/\widehat{f}$. Dividing by $\widehat{f}$ requires a cutoff or a threshold to avoid too small values in the denominator; determining its level is difficult. It is not clear if bandwidth or model selection must be performed separately or simultaneously for the numerator and the denominator. The rate of the final estimator of b corresponds to the worst rate of the two estimators; in particular, it depends on the regularity index of b, but also on the one of f. Therefore, the rate can correspond to the one associated to the regularity index of b, if f is more regular than b, but it is deteriorated if f is less regular than b.
On the other hand, there is no support constraint for this estimation method.
In the projection method used here, the drawbacks listed above do not perturb the estimation except that the unknown function b is estimated in a restricted domain A. Up to now, this set was mostly assumed to be compact. In the present paper, we show how to eliminate the support constraint by introducing a new selection procedure where the dimension of the projection space is chosen in a random set. The procedure can be applied to non compactly supported bases such as the Laguerre or Hermite bases.
Several extensions of our method can be obtained.
First, note that the result of Proposition 2.1 holds for any sequence $(X_i)_{1\leq i\leq n}$ provided that it is independent of $(\varepsilon_i)_{1\leq i\leq n}$ with i.i.d. centered ε_i.
We also may have considered the heteroskedastic regression the model

$$Y_i = b(X_i) + \sigma(X_i)\varepsilon_i, \quad \mathrm{Var}(\varepsilon_1) = \mathbb{E}(\varepsilon_1^2) = 1$$

and the same contrast. The estimator on S_m is still given by (4). Assuming that $\sigma^2(x)$ is uniformly bounded, we can obtain results similar to those obtained here.
Note that regression strategies have been used in other problems, for instance survival function estimation for interval censored data (see Brunel and Comte (2009)), hazard rate estimation in presence of censoring (see Plancade (2011)): our proposal for classical regression may extend to these contexts, for which it is natural to use $\mathbb{R}^+$-supported bases, see Bouaziz *et al.* (2018). Indeed, the variables are lifetimes and thus nonnegative, and censoring implies that the right-hand bound of the support is unknown and difficult to estimate; it is thus most convenient that the Laguerre basis does not require to choose it.

6. Proofs

6.1. **Proofs of the results of Section 2.**

6.1.1. *Proof of Proposition 2.1.* Let us denote by Π_m the orthogonal projection (for the scalar product of $\mathbb{R}^n$) on the sub-space $\left\{\big(t(X_1),\ldots,t(X_n)\big)', t\in S_m\right\}$ of $\mathbb{R}^n$ and by $\Pi_m b$ the projection of the vector $(b(X_1),\ldots,b(X_n))'$. The following equality holds,

$$\|\hat{b}_m - b_A\|_n^2 = \|\Pi_m b - b_A\|_n^2 + \|\hat{b}_m - \Pi_m b\|_n^2 = \inf_{t\in S_m} \|t - b_A\|_n^2 + \|\hat{b}_m - \Pi_m b\|_n^2 \tag{31}$$

By taking expectation, we obtain

$$\mathbb{E}\big[\|\hat{b}_m - b_A\|_n^2\big] \leq \inf_{t\in S_m} \int (t-b_A)^2(x) f(x)dx + \mathbb{E}\left[\|\hat{b}_m - \Pi_m b\|_n^2\right]. \tag{32}$$

Now we have:

Lemma 6.1. *Under the assumptions of Proposition 2.1,*

$$\mathbb{E}\left[\|\hat{b}_m - \Pi_m b\|_n^2\right] = \sigma_\varepsilon^2 \frac{m}{n}.$$

The result of the previous Lemma can be plugged in (32), thus we obtain Proposition 2.1. □

6.1.2. *Proof of Lemma 6.1.* Denote by $b(X) = (b(X_1), \ldots, b(X_n))'$ and $b_A(X) = (b_A(X_1), \ldots, b_A(X_n))'$. We can write

$$\hat{b}_m(X) = (\hat{b}_m(X_1), \ldots, \hat{b}_m(X_n))' = \widehat{\Phi}_m \vec{\hat{a}}^{(m)},$$

where $\vec{\hat{a}}^{(m)}$ is given by (4), and

$$\Pi_m b = \widehat{\Phi}_m \vec{a}^{(m)}, \quad \vec{a}^{(m)} = (\widehat{\Phi}_m' \widehat{\Phi}_m)^{-1} \widehat{\Phi}_m' b(X).$$

Now, denoting by $\mathbf{P}(X) := \widehat{\Phi}_m (\widehat{\Phi}_m' \widehat{\Phi}_m)^{-1} \widehat{\Phi}_m'$, we get

$$\|\hat{b}_m - \Pi_m b\|_n^2 = \|\mathbf{P}(X)\vec{\varepsilon}\|_n^2 = \frac{1}{n}\vec{\varepsilon}'\mathbf{P}(X)'\mathbf{P}(X)\vec{\varepsilon} = \frac{1}{n}\vec{\varepsilon}'\mathbf{P}(X)\vec{\varepsilon} \tag{33}$$

as $\mathbf{P}(X)$ is the $n \times n$-matrix of the euclidean orthogonal projection on the subspace of $\mathbb{R}^n$ generated by the vectors $\varphi_0(X), \ldots, \varphi_{m-1}(X)$, where $\varphi_j(X) = (\varphi_j(X_1), \ldots, \varphi_j(X_n))'$. Note that $\mathbb{E}(\|\mathbf{P}(X)\vec{\varepsilon}\|_{2,n}^2) \leq \mathbb{E}(\|\vec{\varepsilon}\|_{2,n}^2) < +\infty$. Next, we have to compute, using that $\mathbf{P}(X)$ has coefficients depending on the X_i's only,

$$\mathbb{E}\big[\vec{\varepsilon}'\,\mathbf{P}(X))\vec{\varepsilon}\big] = \sum_{i,j} \mathbb{E}\big[\varepsilon_i\varepsilon_j \mathbf{P}_{i,j}(X)\big] = \sigma_\varepsilon^2 \sum_{i=1}^n \mathbb{E}\big[\mathbf{P}_{i,i}(X)\big] = \sigma_\varepsilon^2 \mathbb{E}\big[\mathrm{Tr}(\mathbf{P}(X))\big],$$

where Tr(.) is the trace of the matrix. So, we find

$$\mathrm{Tr}(\mathbf{P}(X)) = \mathrm{Tr}\big((\widehat{\Phi}_m' \widehat{\Phi}_m)^{-1} \widehat{\Phi}_m' \widehat{\Phi}_m\big) = \mathrm{Tr}(\mathrm{I}_m) = m$$

where I_m is the $m \times m$ identity matrix. Finally, we get $\mathbb{E}\left[\|\hat{b}_m - \Pi_m b\|_n^2\right] = \sigma_\varepsilon^2(m/n)$. This is the result of Lemma 6.1. □

6.1.3. *Proof of Proposition 2.2.* Let $t = \sum_{j=0}^{m-1} a_j \varphi_j$, and $\vec{a} = (a_0, \ldots, a_{m-1})'$, then $\|t\|^2 = \|\vec{a}\|_{2,m} = \vec{a}'\,\vec{a}$ and $\|t\|_f^2 = \vec{a}'\Psi_m\vec{a} = \|\Psi_m^{1/2}\vec{a}\|_{2,m}^2$, where $\Psi_m^{1/2}$ is a symmetric square root of Ψ_m. Thus

$$\sup_{t\in S_m, \|t\|_f=1} \|t\|^2 = \sup_{\vec{a}\in\mathbb{R}^m, \|\Psi_m^{1/2}\vec{a}\|_{2,m}=1} \vec{a}'\,\vec{a}.$$

Set $\vec{b} = \Psi_m^{1/2}\vec{a}$, that is $\vec{a} = \Psi_m^{-1/2}\vec{b}$. Then

$$\sup_{t\in S_m, \|t\|_f=1} \|t\|^2 = \sup_{\vec{b}\in\mathbb{R}^m, \|\vec{b}\|_{2,m}=1} \vec{b}'\Psi_m^{-1}\vec{b} = \|\Psi_m^{-1}\|_{\mathrm{op}}.$$

As, for $m \leq m'$, we assume $S_m \subset S_{m'}$, we also have

$$\|\Psi_m^{-1}\|_{\mathrm{op}} = \sup_{t\in S_m, \|t\|_f=1} \|t\|^2 \leq \sup_{t\in S_{m'}, \|t\|_f=1} \|t\|^2 = \|\Psi_{m'}^{-1}\|_{\mathrm{op}}.$$

Thus $m \mapsto \|\Psi_m^{-1}\|_{\mathrm{op}}$ is non decreasing. The same holds for $\sup_{t\in S_m, \|t\|_n=1} \|t\|^2 = \|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}$. □

6.1.4. *Proof of Proposition 2.3.* The first equality holds by writing

$$\begin{aligned}\sup_{t\in S_m,\|t\|_f=1}\left|\frac{1}{n}\sum_{i=1}^n[t^2(X_i)-\mathbb{E}t^2(X_i)]\right| &= \sup_{\vec{x}\in\mathbb{R}^m,\|\sqrt{\Psi_m}\vec{x}\|_{2,m}=1}\left|\vec{x}'\widehat{\Psi}_m\vec{x}-\vec{x}'\Psi_m\vec{x}\right| \\ = \sup_{\vec{x}\in\mathbb{R}^m,\|\sqrt{\Psi_m}\vec{x}\|_{2,m}=1}\left|\vec{x}'(\widehat{\Psi}_m-\Psi_m)\vec{x}\right| &= \sup_{\vec{u}\in\mathbb{R}^m,\|\vec{u}\|_{2,m}=1}\left|\vec{u}'\sqrt{\Psi_m}^{-1}(\widehat{\Psi}_m-\Psi_m)\sqrt{\Psi_m}^{-1}\vec{u}\right| \\ &= \|\sqrt{\Psi_m}^{-1}(\widehat{\Psi}_m-\Psi_m)\sqrt{\Psi_m}^{-1}\|_{\mathrm{op}}\end{aligned}$$

Now, Theorem 1 in Cohen *et al.* (2013) yields that for $0<\delta<1$, $\mathbb{P}(\Omega_m(\delta)^c)\le 2me^{-c(\delta)n/K(m)}$ where, for $(\theta_j)_{0\le j\le m-1}$ an $\mathbb{L}^2(A,f(x)dx)$-orthonormal basis of S_m,

$$K(m)=\sup_{x\in A}\sum_{j=0}^{m-1}\theta_j^2(x), \tag{34}$$

provided that $K(m)<+\infty$.[2]Note that the quantity $K(m)$ is independent of the choice of the basis $(\theta_j)_{0\le j\le m-1}$. Then, Proposition 2.3 follows from the lemma:

Lemma 6.2. *Assume that Ψ_m is invertible and $L(m)<+\infty$ (see (7)). Then $K(m)<+\infty$, and*

$$K(m)=\sup_{x\in A}\overrightarrow{\varphi_{(m)}}(x)'\Psi_m^{-1}\overrightarrow{\varphi_{(m)}}(x)\le L(m)\|\Psi_m^{-1}\|_{\mathrm{op}},\quad \overrightarrow{\varphi_{(m)}}(x)=(\varphi_0(x),\dots,\varphi_{m-1}(x))'.$$

Proof of Lemma 6.2. Let $\overrightarrow{\theta_{(m)}}(x)=(\theta_0(x),\dots,\theta_{m-1}(x))'$. There exists an $m\times m$ matrix A_m such that $\overrightarrow{\theta_{(m)}}(x)=A_m\overrightarrow{\varphi_{(m)}}(x)$. By definition of the basis $(\theta_j)_{0\le j\le m}$,

$$\int_A\overrightarrow{\theta_{(m)}}(x)\overrightarrow{\theta_{(m)}}(x)'f(x)dx=\mathrm{Id}_m$$

and

$$\int_A\overrightarrow{\theta_{(m)}}(x)\overrightarrow{\theta_{(m)}}(x)'f(x)dx=A_m\Psi_mA_m'.$$

This implies $A_m^{-1}(A_m')^{-1}=(A_m'A_m)^{-1}=\Psi_m$ and $A_m'A_m=\Psi_m^{-1}$. Thus

$$\overrightarrow{\theta_{(m)}}(x)\overrightarrow{\theta_{(m)}}(x)'=\overrightarrow{\varphi_{(m)}}(x)A_m'A_m\overrightarrow{\varphi_{(m)}}(x)=\overrightarrow{\varphi_{(m)}}(x)'\Psi_m^{-1}\overrightarrow{\varphi_{(m)}}(x).$$

This gives the first equality. The bound by $\|\Psi_m^{-1}\|_{\mathrm{op}}\|\overrightarrow{\varphi_{(m)}}(x)\|_{2,m}^2=\|\Psi_m^{-1}\|_{\mathrm{op}}\sum_{j=0}^{m-1}\varphi_j^2(x)$ ends the proof of Lemma 6.2. □

Note that we can see also here that $\mathbf{G}$ in Cohen *et al.* (2013), that we denote here $\widehat{\mathbf{G}}_m$ is such that $\widehat{\mathbf{G}}_m=A_m\widehat{\Psi}_mA_m'$ where A_m' is a square root of Ψ_m^{-1}.

[2]In Cohen *et al.* (2013), the condition $K(m)<+\infty$ is not clearly stated; it is implicit as the result does not hold otherwise. Actually all examples of the paper are for A compact, in which case $K(m)<+\infty$. If A is not compact, then $K(m)$ may be $+\infty$. Therefore our condition (7) and Lemma 6.2 clarify Cohen *et al.*'s result.

6.1.5. *Proof of Proposition 2.4.*

Proof of (i). To get the announced result, we apply again a Bernstein matrix inequality given in Tropp (2012)(see Theorem 7.2). We write $\widehat{\Psi}_m$ as a sum of a sequence of independent matrices $\widehat{\Psi}_m = \frac{1}{n}\sum_{i=1}^n \mathbf{K}_m(X_i)$, with $\mathbf{K}_m(X_i) = (\varphi_j(X_i)\varphi_k(X_i))_{0\le j,k\le m-1}$. We define

$$\mathbf{S}_m = \frac{1}{n}\sum_{i=1}^n \mathbf{K}_m(X_i) - \mathbb{E}\left[\mathbf{K}_m(X_i)\right]. \tag{35}$$

• Bound on $\|\mathbf{K}_m(X_1) - \mathbb{E}\left[\mathbf{K}_m(X_1)\right]\|_{\mathrm{op}}/n$. First we can write that

$$\|\mathbf{K}_m(X_1) - \mathbb{E}\left[\mathbf{K}_m(X_1)\right]\|_{\mathrm{op}} \le \|\mathbf{K}_m(X_1)\|_{\mathrm{op}} + \|\mathbb{E}\left[\mathbf{K}_m(X_1)\right]\|_{\mathrm{op}},$$

and we bound the first term, the other one being similar. As $\mathbf{K}_m(X_1)$ is symmetric and nonnegative a.s., we have a.s.

$$\begin{aligned}\|\mathbf{K}_m(X_1)\|_{\mathrm{op}} &= \sup_{\|\vec{x}\|_{2,m}=1} \sum_{0\le j,k\le m-1} x_j x_k [\mathbf{K}_m(X_1)]_{j,k} = \sup_{\|\vec{x}\|_{2,m}=1} \sum_{0\le j,k\le m-1} x_j x_k \varphi_j(X_1)\varphi_k(X_1) \\ &= \sup_{\|\vec{x}\|_{2,m}=1} \left[\left(\sum_{j=0}^{m-1} x_j \varphi_j(X_1)\right)^2\right] \le L(m).\end{aligned}$$

So we get that, a.s.

$$\|\mathbf{K}_m(X_1) - \mathbb{E}\left[\mathbf{K}_m(X_1)\right]\|_{\mathrm{op}}/n \le \frac{2L(m)}{n} := \mathbf{L}. \tag{36}$$

• Bound on $\nu(\mathbf{S}_m) = \|\sum_{i=1}^n \mathbb{E}\left[(\mathbf{K}_m(X_i) - \mathbb{E}\left[\mathbf{K}_m(X_i)\right])' (\mathbf{K}_m(X_i) - \mathbb{E}\left[\mathbf{K}_m(X_i)\right])\right]\|_{\mathrm{op}}/n^2$. We have

$$\nu(\mathbf{S}_m) = \frac{1}{n}\sup_{\|\vec{x}\|_{2,m}=1} \mathbb{E}\left\|\left(\mathbf{K}_m(X_1) - \mathbb{E}\left[\mathbf{K}_m(X_1)\right]\right)\vec{x}\right\|_{2,m}^2$$

It yields that, for $\vec{x}' = (x_0, \ldots, x_{m-1})$,

$$\begin{aligned}\mathbb{E}_1 := \mathbb{E}\left\|\left(\mathbf{K}_m(X_1) - \mathbb{E}\left[\mathbf{K}_m(X_1)\right]\right)\vec{x}\right\|_{2,m}^2 &= \sum_{j=0}^{m-1} \mathrm{Var}\left[\sum_{k=0}^{m-1} \left(\varphi_j(X_1)\varphi_k(X_1)\right) x_k\right] \\ \le \sum_{j=0}^{m-1} \mathbb{E}\left(\sum_{k=0}^{m-1} \left(\varphi_j(X_1)\varphi_k(X_1)\right) x_k\right)^2 &= \sum_{j=0}^{m-1} \int \left(\sum_{k=0}^{m-1} \left(\varphi_j(u)\varphi_k(u)\right) x_k\right)^2 f(u)du\end{aligned}$$

Therefore as, by **(A2)**, f is bounded,

$$\mathbb{E}_1 \le \|f\|_\infty \sum_{j=0}^{m-1} \int \left(\sum_{k=0}^{m-1} \left(\varphi_j(u)\varphi_k(u)\right) x_k\right)^2 du \le \|f\|_\infty L(m) \sum_{k=0}^{m-1} x_k^2 = \|f\|_\infty L(m).$$

Then we get that $\nu(\mathbf{S}_m) \le \frac{\|f\|_\infty L(m)}{n}$. Applying Theorem 7.2 gives the result (i) of Proposition 2.4.

Proof of (ii). First note that

$$\begin{aligned}\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\rm op} &= \|\Psi_m^{-1/2}\left(\Psi_m^{1/2}\widehat{\Psi}_m^{-1}\Psi_m^{1/2}-{\rm Id}_m\right)\Psi_m^{-1/2}\|_{\rm op})\\ &\leq \|\Psi_m^{-1}\|_{\rm op}\|\Psi_m^{1/2}\widehat{\Psi}_m^{-1}\Psi_m^{1/2}-{\rm Id}_m\|_{\rm op},\end{aligned}$$

so that

$$\left\{\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\rm op}>\alpha\|\Psi_m^{-1}\|_{\rm op}\right\}\subset\left\{\|\Psi_m^{1/2}\widehat{\Psi}_m^{-1}\Psi_m^{1/2}-{\rm Id}_m\|_{\rm op}>\alpha\right\}. \tag{37}$$

Now, we write the decomposition $\left\{\|\Psi_m^{1/2}\widehat{\Psi}_m^{-1}\Psi_m^{1/2}-{\rm Id}_m\|_{\rm op}>\alpha\right\}:=B_1\cup B_2$ with

$$\begin{aligned}B_1 &= \left\{\|\Psi_m^{1/2}\widehat{\Psi}_m^{-1}\Psi_m^{1/2}-{\rm Id}_m\|_{\rm op}>\alpha\right\}\bigcap\left\{\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}-{\rm Id}_m\|_{\rm op}<\frac{1}{2}\right\}\\ B_2 &= \left\{\|\Psi_m^{1/2}\widehat{\Psi}_m^{-1}\Psi_m^{1/2}-{\rm Id}_m\|_{\rm op}>\alpha\right\}\bigcap\left\{\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}-{\rm Id}_m\|_{\rm op}\geq\frac{1}{2}\right\}\end{aligned}$$

Clearly $B_2\subset\left\{\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}-{\rm Id}_m\|_{\rm op}\geq\frac{1}{2}\right\}$.

Applying Theorem 7.1 with $\mathbf{A}={\rm Id}_m$ and $\mathbf{B}=\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}-{\rm Id}_m$, yields

$$\begin{aligned}B_1&\subset\left\{\frac{\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}-{\rm Id}_m\|_{\rm op}}{1-\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}-{\rm Id}_m\|_{\rm op}}>\alpha\right\}\cap\left\{\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}-{\rm Id}_m\|_{\rm op}<\frac{1}{2}\right\}\\ &\subset\left\{\ \|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}-{\rm Id}_m\|_{\rm op}>\alpha/2\right\}\cap\left\{\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}-{\rm Id}_m\|_{\rm op}<\frac{1}{2}\right\}\\ &\subset\left\{\ \|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}-{\rm Id}_m\|_{\rm op}>\alpha/2\right\}\end{aligned}$$

Thus $B_1\cup B_2\subset\left\{\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}-{\rm Id}_m\|_{\rm op}\geq\frac{\alpha\wedge 1}{2}\right\}$, which ends the proof of (ii) and of Proposition 2.4. □

6.2. **Proofs of the results of Section 3.**

6.2.1. *Proof of Proposition 3.1.* We define the sets (see (10)),

$$\Lambda_m=\left\{L(m)(\|\widehat{\Psi}_m^{-1}\|_{\rm op}\vee 1)\leq\mathfrak{c}\frac{n}{\log(n)}\right\},\ \text{and}\ \Omega_m=\left\{\left|\frac{\|t\|_n^2}{\|t\|_f^2}-1\right|\leq\frac{1}{2},\forall t\in S_m\right\}.$$

Below, we prove the following lemma

Lemma 6.3. *Under the assumptions of Proposition 3.1, for m satisfying condition (9), we have*

$$\mathbb{P}(\Lambda_m^c)\leq c/n^4,\quad \mathbb{P}(\Omega_m^c)\leq c/n^4$$

where c is a positive constant.

Now, we write

$$\begin{aligned}\|\widetilde{b}_m-b_A\|_f^2 &= \|\hat{b}_m-b_A\|_f^2\mathbf{1}_{\Lambda_m}+\|b_A\|_f^2\mathbf{1}_{\Lambda_m^c}\\ &= \|\hat{b}_m-b_A\|_f^2\mathbf{1}_{\Lambda_m\cap\Omega_m}+\|\hat{b}_m-b_A\|_f^2\mathbf{1}_{\Lambda_m\cap\Omega_m^c}+\|b_A\|_f^2\mathbf{1}_{\Lambda_m^c}.\end{aligned} \tag{38}$$

From the proof of Theorem 3 in Cohen *et al.* (2013), we get

$$\mathbb{E}\left(\|\hat{b}_m - b_A\|_f^2 \mathbf{1}_{\Lambda_m \cap \Omega_m}\right) \leq \left(1 + \frac{8\mathfrak{c}}{\log(n)}\right) \inf_{t \in S_m} (\|t - b_A\|_f^2) + 8\sigma_\varepsilon^2 \frac{m}{n}. \tag{39}$$

Remark. *For sake of self-containedness, we give a quick and simple proof of a similar bound, with different constants. For any $t \in S_m$, we have using $(x+y)^2 \leq (1+1/\theta)x^2 + (1+\theta)y^2$ with $\theta = 4$,*

$$\begin{aligned} \|\hat{b}_m - b_A\|_f^2 \mathbf{1}_{\Lambda_m \cap \Omega_m} &\leq \frac{5}{4}\|\hat{b}_m - t\|_f^2 \mathbf{1}_{\Lambda_m \cap \Omega_m} + 5\|t - b_A\|_f^2 \mathbf{1}_{\Lambda_m \cap \Omega_m} \\ &\leq \frac{5}{2}\|\hat{b}_m - t\|_n^2 \mathbf{1}_{\Lambda_m \cap \Omega_m} + 5\|t - b_A\|_f^2 \mathbf{1}_{\Lambda_m \cap \Omega_m}, \end{aligned}$$

by using the definition of Ω_m. We insert b_A again and get:

$$\|\hat{b}_m - b_A\|_f^2 \mathbf{1}_{\Lambda_m \cap \Omega_m} \leq 5\|\hat{b}_m - b_A\|_n^2 \mathbf{1}_{\Lambda_m \cap \Omega_m} + 5\|b_A - t\|_n^2 \mathbf{1}_{\Lambda_m \cap \Omega_m} + 5\|t - b_A\|_f^2 \mathbf{1}_{\Lambda_m \cap \Omega_m}.$$

Therefore taking expectation and applying Proposition 2.1 yield

$$\mathbb{E}\left(\|\hat{b}_m - b_A\|_f^2 \mathbf{1}_{\Lambda_m \cap \Omega_m}\right) \leq 15 \inf_{t \in S_m} (\|t - b_A\|_f^2) + 5\sigma_\varepsilon^2 \frac{m}{n}.$$

This just helps to see that Inequality (39) relies on computations w.r.t the empirical norm.

Now we bound the two remaining terms. Clearly, with Lemma 6.3,

$$\mathbb{E}(\|b_A\|_f^2 \mathbf{1}_{\Lambda_m^c}) \leq \|b_A\|_f^2 \mathbb{P}(\Lambda_m^c) \leq c/n^4. \tag{40}$$

Next we deal with $\mathbb{E}(\|\hat{b}_m - b_A\|_f^2 \mathbf{1}_{\Lambda_m \cap \Omega_m^c})$. We have $\|\hat{b}_m - b_A\|_f^2 \leq 2(\|\hat{b}_m\|_f^2 + \|b_A\|_f^2)$ and

$$\|\hat{b}_m\|_f^2 = \int \left(\sum_{j=0}^{m-1} \hat{a}_j \varphi_j(x)\right)^2 f(x)dx = (\vec{\hat{a}}^{(m)})' \Psi_m \vec{\hat{a}}^{(m)} \leq \|\Psi_m\|_{\mathrm{op}} \|\vec{\hat{a}}^{(m)}\|_{2,m}^2.$$

First,

$$\begin{aligned} \|\Psi_m\|_{\mathrm{op}} &= \sup_{\|\vec{x}\|_{2,m}=1} \vec{x}' \Psi_{\hat{m}} \vec{x} = \sup_{\|\vec{x}\|_{2,m}=1} \int \left(\sum_{j=0}^{m-1} x_j \varphi_j(u)\right)^2 f(u)du \\ &\leq \sup_{\|\vec{x}\|_{2,m}=1} \int \left(\sum_{j=0}^{m-1} x_j^2 \sum_{j=0}^{m-1} \varphi_j^2(u)\right) f(u)du \leq L(m) \end{aligned}$$

Next, $\|\vec{\hat{a}}^{(m)}\|_{2,m}^2 = (1/n^2)\|\widehat{\Psi}_m^{-1}\widehat{\Phi}_m'\vec{Y}\|_{2,m}^2 \leq (1/n^2)\|\widehat{\Psi}_m^{-1}\widehat{\Phi}_m'\|_{\mathrm{op}}^2 \|\vec{Y}\|_{2,n}^2$ and

$$\|\widehat{\Psi}_m^{-1}\widehat{\Phi}_m'\|_{\mathrm{op}}^2 = \lambda_{\max}\left(\widehat{\Psi}_m^{-1}\widehat{\Phi}_m'\widehat{\Phi}_m\widehat{\Psi}_m^{-1}\right) = n\lambda_{\max}(\widehat{\Psi}_m^{-1}) = n\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}$$

Therefore, for all m satisfying (9),

$$\|\hat{b}_m\|_f^2 \leq \frac{L(m)\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}}{n} \left(\sum_{i=1}^n Y_i^2\right) \leq \frac{\mathfrak{c}}{\log(n)} \left(\sum_{i=1}^n Y_i^2\right), \tag{41}$$

and thus on Λ_m, for $n \geq 3$, $\|\hat{b}_m\|_f^2 \leq C\left(\sum_{i=1}^n Y_i^2\right)$. Then as $\mathbb{E}[(\sum_{i=1}^n Y_i^2)^2] \leq n^2\mathbb{E}(Y_1^4)$, we get

$$\mathbb{E}(\|\hat{b}_m\|_f^2 \mathbf{1}_{\Lambda_m \cap \Omega_m^c}) \leq \sqrt{\mathbb{E}(\|\hat{b}_m\|_f^4)\mathbb{P}(\Omega_m^c)} \leq C\mathbb{E}^{1/2}(Y_1^4) n \mathbb{P}^{1/2}(\Omega_m^c) \leq c'/n.$$

On the other hand $\mathbb{E}(\|b_A\|_f^2 \mathbf{1}_{\Lambda_m \cap \Omega_m^c}) \leq \|b_A\|_f^2 \mathbb{P}(\Omega_m^c) \leq c"/n^4$. Thus

$$\mathbb{E}\left(\|\hat{b}_m - b_A\|_f^2 \mathbf{1}_{\Lambda_m \cap \Omega_m^c}\right) \leq c_1/n. \tag{42}$$

Taking expectation of (38) and plugging (39)-(40)-(42) therein gives the result. □

6.2.2. *Proof of Lemma 6.3.* The bound on $\mathbb{P}(\Omega_m^c)$ follows from Proposition 2.3 under condition (9).

We study now $\mathbb{P}(\Lambda_m^c)$ for m satisfying condition (9). On Λ_m^c, for m satisfying condition (9), we have $L(m)\|\Psi_m^{-1}\|_{\text{op}} \leq \mathfrak{c}n/2\log(n)$ and $L(m)\|\widehat{\Psi}_m^{-1}\|_{\text{op}} > \mathfrak{c}n/\log(n)$. This implies, as

$$\begin{aligned}\mathfrak{c}\frac{n}{\log(n)} < L(m)\|\widehat{\Psi}_m^{-1}\|_{\text{op}} \leq L(m)\|\Psi_m^{-1} - \widehat{\Psi}_m^{-1}\|_{\text{op}} \quad &+ \quad L(m)\|\Psi_m^{-1}\|_{\text{op}} \\ &\leq \quad L(m)\|\Psi_m^{-1} - \widehat{\Psi}_m^{-1}\|_{\text{op}} + \frac{\mathfrak{c}}{2}\frac{n}{\log(n)},\end{aligned}$$

that $L(m)\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\text{op}} \geq \mathfrak{c}n/(2\log(n))$. Therefore, we have

$$\Lambda_m^c \subset \{L(m)\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\text{op}} > \frac{\mathfrak{c}}{2}\frac{n}{\log(n)}\} \subset \{\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\text{op}} > \|\Psi_m^{-1}\|_{\text{op}}\}.$$

Applying Proposition 2.4 (ii) and Proposition 2.3, we get

$$\mathbb{P}(\Lambda_m^c) \leq \mathbb{P}\left(\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \text{Id}_m\|_{\text{op}} \geq \frac{1}{2}\right) \leq \frac{c}{n^4}. \quad \square$$

6.2.3. *Proof of Theorem 3.1.* We use the strategy of proof of Theorem 2.11 in Tsybakov (2009). We define proposals $b_0(x) = 0$ and for $\vec{\theta} = (\theta_0, \ldots, \theta_{m-1})'$ with $\theta_j \in \{0, 1\}$,

$$b_\theta(x) = \delta v_n \sigma_\varepsilon \sum_{j=0}^{m-1} \left[\Psi_m^{-1/2}\vec{\theta}\right]_j \varphi_j(x)$$

where $\Psi_m^{-1/2}$ is a symmetric square-root of the positive definite matrix Ψ_m^{-1}.

We choose $v_n^2 = 1/n$ and $m = n^{1/(s+1)}$.

• We prove that $b_0, b_\theta \in W_f^s(A, R)$.

As $b_\theta \in S_m$, $(b_\theta)_m^f = b_\theta$ and $(b_\theta)_\ell^f = b_\theta$ for all $\ell \geq m$. Indeed, $S_m \subset S_\ell$. Thus, for $\ell \geq m$, $\|b_\theta - (b_\theta)_\ell^f\|_f^2 = 0$.

Next, $\|b_\theta - (b_\theta)_\ell^f\|_f^2 \leq \|b_\theta\|_f^2$ and as $\int \varphi_j\varphi_k f = [\Psi_m]_{j,k}$, we get

$$\|b_\theta\|_f^2 = \delta^2 v_n^2 \sigma_\varepsilon^2 \sum_{0 \leq j,k \leq m-1} \left[\Psi_m^{-1/2}\vec{\theta}\right]_j \left[\Psi_m^{-1/2}\vec{\theta}\right]_k [\Psi_m]_{j,k} = \delta^2 v_n^2 \sigma_\varepsilon^2 \sum_{j=0}^{m-1} \theta_j^2 \leq \delta^2 v_n^2 \sigma_\varepsilon^2 m.$$

Thus for $\ell \leq m$,

$$\ell^s \|b_\theta - (b_\theta)_\ell^f\|_f^2 \leq \ell^s \|b_\theta\|_f^2 \leq \delta^2 v_n^2 \sigma_\varepsilon^2 m \ell^s \leq \delta^2 v_n^2 \sigma_\varepsilon^2 m^{s+1} = \delta^2 \sigma_\varepsilon^2.$$

Choosing δ small enough, we get the result.

• We prove that we can find $\{\theta^{(0)},\dots,\theta^{(M)}\}$, M elements of $\{0,1\}^m$ such that

$$\|b_{\theta^{(j)}} - b_{\theta^{(k)}}\|_f^2 \geq cn^{-s/(s+1)} \text{ for } 0 \leq j < k \leq M.$$

As above, we find

$$\|b_\theta - b_{\theta'}\|_f^2 = \delta^2 v_n^2 \sigma_\varepsilon^2 \sum_{j=0}^{m-1} (\theta_j - \theta'_j)^2 = \delta^2 v_n^2 \sigma_\varepsilon^2 \rho(\theta,\theta'),$$

where $\rho(\theta,\theta') = \sum_{j=0}^{m-1}(\theta_j - \theta'_j)^2 = \sum_{j=0}^{m-1} \mathbf{1}_{\theta_j \neq \theta'_j}$ is the Hamming distance between the two binary sequences θ and θ'. By the Varshamov-Gilbert Lemma (see Lemma 2.9 p.104 in Tsybakov (2009)), for $m \geq 8$, there exists a subset $\{\theta^{(0)},\dots,\theta^{(M)}\}$ such that $\theta^{(0)} = (0,\dots,0)$, $\rho(\theta^{(j)},\theta^{(k)}) \geq m/8$, $0 \leq j < k \leq M$, and $M \geq 2^{m/8}$.

Therefore $\|b_{\theta^{(j)}} - b_{\theta^{(k)}}\|_f^2 \geq \delta^2 v_n^2 \sigma_\varepsilon^2 m/8 = \delta^2 \sigma_\varepsilon^2 n^{-s/(s+1)}/8$.

• Conditional Kullback. Consider first the design $X_1,\dots,X_n$ as fixed. Let $\mathbb{P}^i_{\theta^{(j)}}$ the density of $Y_i = b_{\theta^{(j)}}(X_i) + \varepsilon_i$, i.e. the Gaussian distribution $\mathcal{N}(b_{\theta^{(j)}}(X_i), \sigma_\varepsilon^2)$, and $\mathbb{P}_{\theta^{(j)}}$ the distribution of $(Y_1,\dots,Y_n)$. Then,

$$\frac{1}{M+1}\sum_{j=1}^{M} K(\mathbb{P}_{\theta^{(j)}}, \mathbb{P}_{\theta^{(0)}}) = \frac{1}{M+1}\sum_{j=1}^{M}\sum_{i=1}^{n} \frac{b^2_{\theta^{(j)}}(X_i)}{2\sigma_\varepsilon^2} = \frac{n}{2(M+1)\sigma_\varepsilon^2}\sum_{j=1}^{M} \|b_{\theta^{(j)}}\|_n^2.$$

Then on $\Omega_n = \cup_{m \leq cn/\log(n)} \Omega_m$, we have $\|b_{\theta^{(j)}}\|_n^2 \leq 2\|b_{\theta^{(j)}}\|_f^2$, thus

$$\frac{1}{M+1}\sum_{j=1}^{M} K(\mathbb{P}_{\theta^{(j)}}, \mathbb{P}_{\theta^{(0)}}) \leq \frac{n\delta^2 v_n^2}{M+1}\sum_{j=1}^{M}\sum_{k=0}^{m-1} (\theta_k^{(j)})^2 \leq n\delta^2 v_n^2 m \leq \frac{8\delta^2}{\log(2)}\log(M).$$

For δ^2 small enough so that $8\delta^2/\log(2) := \alpha < 1/8$,

$$\frac{1}{M+1}\sum_{j=1}^{M} K(\mathbb{P}_{\theta^{(j)}}, \mathbb{P}_{\theta^{(0)}})\mathbf{1}_{\Omega_n} \leq \alpha \log(M)\mathbf{1}_{\Omega_n}.$$

Now, following Tsybakov (2009), p.116,

$$\begin{aligned}\sup_{b_A \in W_f^s(A,R)} \mathbb{E}_{b_A}\left[n^{s/(s+1)}\|T_n - b_A\|_f^2\right] &\geq \mathfrak{A}^2 \max_{b_A \in \{b_{\theta^{(j)}}, j=0,\dots,M\}} \mathbb{P}_{b_A}\left(\|T_n - b_A\|_f > \mathfrak{A} n^{-s/[2(s+1)]}\right) \\ &\geq \mathfrak{A}^2\left(\frac{\log(M+1) - \log(2)}{\log(M)} - \alpha\right)\mathbb{P}(\Omega_n).\end{aligned}$$

For n large enough and m satisfying (9), it follows from Lemma 6.3 that $\mathbb{P}(\Omega_n) \geq 1 - (c/n^3) \geq 1/2$. Therefore the lower bound is proved. □

6.2.4. *Proof ol Lemma 3.1.* For all $\vec{u} = (u_0,\dots,u_{m-1})' \in \mathbb{R}^m \backslash \{\vec{0}\}$, for $t(x) = \sum_{j=0}^{m-1} u_j \varphi_j(x)$, $\vec{u}\,'\,\widehat{\Psi}_m\,\vec{u} = \|t\|_n^2 \geq 0$. Thus $\|t\|_n = 0 \Rightarrow t(X_i) = 0$ for $i = 1,\dots,n$. As the X_i are almost surely distinct and $t(x)w(x)$ is a polynomial with degree $m-1$ where $w(x) = e^x$ in the Laguerre case and $w(x) = e^{x^2/2}$ in the Hermite case, for $m \leq n$, we obtain that $t \equiv 0$. This implies $\vec{u} = \vec{0}$. □

6.2.5. *Proof of Proposition 3.4.* The invertibility of Ψ_m follows from Lemma 2.1 under (18). Now we prove (23). First note that, for j large enough,

$$\int \varphi_j^2(x) f(x) dx \leq \frac{c_1}{\sqrt{j}}, \tag{43}$$

where c_1 is a constant. The proof of Inequality (43) in the Hermite case is given in Belomestny *et al.* (2017), Proposition 2.1. and in Comte and Genon-Catalot (2018) in the Laguerre case. As Ψ_m is a symmetric positive definite matrix, $\|\Psi_m^{-1}\|_{\text{op}} = 1/\lambda_{\min}(\Psi_m)$, where $\lambda_{\min}(\Psi_m)$ denotes the smallest eigenvalue of Ψ_m. By (14), we get that for all $j \in \{1, \ldots, m\}$, denoting by $\vec{e}_j$ the jth canonical vector (all coordinates are 0 except the jth which is equal to 1), $\vec{e_j}'\Psi_m \vec{e_j} = \int \varphi_j^2 f$, and

$$\min_{\|\vec{u}\|_{2,m}=1} \vec{u}'\Psi_m \vec{u} \leq \min_{j=1,\ldots,m} \vec{e_j}'\Psi_m \vec{e_j} = \min_{j=1,\ldots,m} \int \varphi_j^2 f \leq \frac{c}{\sqrt{m}}.$$

As a consequence, $\lambda_{\min}(\Psi_m) \leq c/\sqrt{m}$ which implies the result. □

6.2.6. *Proof of Proposition 3.5.* We need results on Laguerre functions with index $\delta > -1$. The Laguerre polynomial with index δ, $\delta > -1$, and degree k is given by

$$L_k^{(\delta)}(x) = \frac{1}{k!} e^x x^{-\delta} \frac{d^k}{dx^k} \left(x^{\delta+k} e^{-x} \right).$$

We consider the Laguerre functions with index δ, given by

$$\ell_k^{(\delta)}(x) = 2^{(\delta+1)/2} \left(\frac{k!}{\Gamma(k+\delta+1)} \right)^{1/2} L_k^{(\delta)}(2x) e^{-x} x^{\delta/2}, \tag{44}$$

and $\ell_k^{(0)} = \ell_k$. The family $(\ell_k^{(\delta)})_{k\geq 0}$ is an orthonormal basis of $\mathbb{L}^2(\mathbb{R}^+)$.

In the following, we use the result of Askey and Wainger (1965) which gives bounds on $\ell_k^{(\delta)}$, depending on k: for $\nu = 4k + 2\delta + 2$, and k large enough, it holds $|\ell_k^{(\delta)}(x/2)| \leq Ce^{-c_0 x}$ for $x \geq 3\nu/2$, where c_0 is a positive fixed constant.

We need similar results for Hermite functions. These can be deduced from the following link between Hermite and Laguerre functions, proved in Comte and Genon-Catalot (2018):

Lemma 6.4. *For $x \geq 0$,*

$$h_{2n}(x) = (-1)^n \sqrt{x/2}\, \ell_n^{(-1/2)}(x^2/2), \quad h_{2n+1}(x) = (-1)^n \sqrt{x/2}\, \ell_n^{(1/2)}(x^2/2).$$

This is completed by the fact that Hermite functions are even for even n, odd for odd n.

We treat the Laguerre basis first. The result of Askey and Wainger (1965) recalled above states that, for j large enough, $\ell_j(x) \leq ce^{-c_0 x}$ for $2x \geq 3(2j+1)$, where $c_0 2$ is a

constant. Thus for $\vec{x} \in \mathbb{R}^m$, $\|\vec{x}\|_{2,m} = 1$, we have

$$\begin{aligned}
\vec{x}'\Psi_m\vec{x} &= \int_0^{+\infty} \left(\sum_{j=0}^{m-1} x_j \ell_j(u)\right)^2 f(u)du \geq \int_0^{3(2m+1)} \left(\sum_{j=0}^{m-1} x_j \ell_j(v/2)\right)^2 f(v/2)dv/2 \\
&\geq \inf_{v\in[0,3(2m+1)]} f(v/2) \int_0^{3(2m+1)/2} \left(\sum_{j=0}^{m-1} x_j \ell_j(u)\right)^2 du \\
&\geq \inf_{u\in[0,3(m+1/2)]} f(u) \left(\int_0^{+\infty} \left(\sum_{j=0}^{m-1} x_j \ell_j(u)\right)^2 du - \int_{3(m+1/2)}^{+\infty} \left(\sum_{j=0}^{m-1} x_j \ell_j(u)\right)^2 du\right)
\end{aligned}$$

Then $\inf_{u\in[0,3(m+1/2)]} f(u) \geq Cm^{-k}$ and $\int_0^{+\infty} \left(\sum_{j=0}^{m-1} x_j \ell_j(u)\right)^2 du = \|\vec{x}\|_{2,m}^2 = 1$ and, for m large enough,

$$\int_{3(m+1/2)}^{+\infty} \left(\sum_{j=0}^{m-1} x_j \ell_j(u)\right)^2 du \leq C'me^{-c_0'm} \leq \frac{1}{2}.$$

It follows that, for m large enough, $\vec{x}'\Psi_m\vec{x} \geq Cm^{-k}/2$.

For the Hermite basis, we proceed analogously using that $|h_j(x)| \leq c|x|e^{-c_0x^2}$ for $x^2 \geq (3/2)(4j+3)$. □

6.3. **Proof of the results in Section 4.**

6.3.1. *Proof of Inequality (29) of Theorem 4.1.* We denote by $\widehat{M}_n$ the maximal element of $\widehat{\mathcal{M}}_n$ (see (27)) and by M_n the maximal element of $\mathcal{M}_n$ (see (28)). We need also:

$$\mathcal{M}_n^+ = \left\{ m \in \mathbb{N}, \quad m\left(\|\Psi_m^{-1}\|_{\mathrm{op}}^2 \vee 1\right) \leq 4\mathfrak{d}\frac{n}{\log(n)} \right\}, \tag{45}$$

with $\mathfrak{d}$ give in (27). Let M_n^+ denote the maximal element of $\mathcal{M}_n^+$. Heuristically, with large probability, considering the constants associated with the sets, we should have $M_n \leq \widehat{M}_n \leq M_n^+$ or equivalently $\mathcal{M}_n \subset \widehat{\mathcal{M}}_n \subset \mathcal{M}_n^+$, and on this set, we really bound the risk; otherwise, we bound the probability of the complement. More precisely, we denote by

$$\Xi_n := \left\{ \mathcal{M}_n \subset \widehat{\mathcal{M}}_n \subset \mathcal{M}_n^+ \right\}, \tag{46}$$

and we write the decomposition:

$$\widehat{b}_{\hat{m}} - b_A = \underbrace{(\widehat{b}_{\hat{m}} - b_A)\mathbf{1}_{\Xi_n}}_{:=T_1} + \underbrace{(\widehat{b}_{\hat{m}} - b_A)\mathbf{1}_{\Xi_n^c}}_{:=T_2}. \tag{47}$$

The proof relies on two steps and the two following Lemmas.

Lemma 6.5. *Under the assumptions of Theorem 4.1, there exists κ_0 such that for $\kappa \geq \kappa_0$, we have*

$$\mathbb{E}\left[\|\hat{b}_{\hat{m}} - b_A\|_n^2 \mathbf{1}_{\Xi_n}\right] \leq C \inf_{m\in\mathcal{M}_n} \left(\inf_{t\in S_m} \|t - b_A\|_f^2 \ + \kappa\sigma_\varepsilon^2 \frac{m}{n} \right) + \frac{C'}{n}$$

where C is a numerical constant and C' is a constant depending on f, b, σ_ε.

Lemma 6.6. *We have, for c a positive constant,*

$$\mathbb{P}(\Xi_n^c) = \mathbb{P}\left(\left\{\mathcal{M}_n \not\subset \widehat{\mathcal{M}}_n \text{ or } \widehat{\mathcal{M}}_n \not\subset \mathcal{M}_n^+\right\}\right) \leq \frac{c}{n^2}.$$

Lemma 6.5 gives the bound on T_1.
For T_2, we use Lemma 6.6 as follows. Recall that Π_m denotes the orthogonal projection (for the scalar product of $\mathbb{R}^n$) on the sub-space $\left\{\big(t(X_1),\ldots,t(X_n)\big)', t\in S_m\right\}$ of $\mathbb{R}^n$. We have $\big(\hat{b}_m(X_1),\ldots,\hat{b}_m(X_n)\big)' = \Pi_m Y$. By using the same notation for the function t and the vector $\big(t(X_1),\ldots,t(X_n)\big)'$, we can see that

$$\|b-\hat{b}_{\hat{m}}\|_n^2 = \|b-\Pi_{\hat{m}}b\|_n^2 + \|\Pi_{\hat{m}}\varepsilon\|_n^2 \leq \|b\|_n^2 + n^{-1}\sum_{k=1}^n \varepsilon_k^2. \tag{48}$$

Thus

$$\begin{aligned}\mathbb{E}\big[\|b-\hat{b}_{\hat{m}}\|_n^2 \mathbf{1}_{\Xi_n^c}\big] &\leq \mathbb{E}\big[\|b\|_n^2\mathbf{1}_{\Xi_n^c}\big] + \frac{1}{n}\sum_{k=1}^n \mathbb{E}\big[\varepsilon_k^2\mathbf{1}_{\Xi_n^c}\big] \\ &\leq \left(\sqrt{\mathbb{E}\big[b^4(X_1)\big]} + \sqrt{\mathbb{E}\big[\varepsilon_1^4\big]}\right)\sqrt{\mathbb{P}\left(\Xi_n^c\right)}.\end{aligned}$$

We deduce that

$$\mathbb{E}\big[\|b-\hat{b}_{\hat{m}}\|_n^2\mathbf{1}_{\Xi_n^c}\big] \leq \frac{c'}{n}.$$

This, together with Lemma 6.5 plugged in decomposition (47), ends the proof of Inequality (29) of Theorem 4.1. □

6.3.2. *Proof of Lemma 6.5.* To begin with, we note that $\gamma_n(\hat{b}_m) = -\|\hat{b}_m\|_n^2$. Indeed, using formula (4) and $\widehat{\Phi}_m'\widehat{\Phi}_m = n\widehat{\Psi}_m$, we have

$$\gamma_n\big(\hat{b}_m\big) = \big\|\widehat{\Phi}_m\vec{\hat{a}}^{(m)}\big\|_n^2 - 2\big(\vec{\hat{a}}^{(m)}\big)'\widehat{\Phi}_m'\,\vec{Y} = -\big(\vec{\hat{a}}^{(m)}\big)'\widehat{\Phi}_m'\,\vec{Y} = -\big\|\widehat{\Phi}_m\vec{\hat{a}}^{(m)}\big\|_n^2.$$

Consequently, we can write

$$\hat{m} = \arg\min_{m\in\widehat{\mathcal{M}}_n}\{\gamma_n(\hat{b}_m) + \mathrm{pen}(m)\}, \quad \text{with} \quad \mathrm{pen}(m) = \kappa\sigma_\varepsilon^2\frac{m}{n}.$$

Thus, using the definition of the contrast, we have, for any $m\in\widehat{\mathcal{M}}_n$, and any $b_m\in S_m$,

$$\gamma_n(\hat{b}_{\hat{m}}) + \mathrm{pen}(\hat{m}) \leq \gamma_n(b_m) + \mathrm{pen}(m). \tag{49}$$

Now, on the set $\Xi_n = \left\{\mathcal{M}_n \subset \widehat{\mathcal{M}}_n \subset \mathcal{M}_n^+\right\}$, we have in all cases that $\hat{m}\leq\widehat{M}_n\leq M_n^+$ and either $M_n\leq\hat{m}\leq\widehat{M}_n\leq M_n^+$ or $\hat{m}<M_n\leq\widehat{M}_n\leq M_n^+$. In the first case, $\hat{m}$ is upper and lower bounded by deterministic bounds, and in the second,

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n}\{\gamma_n(\hat{b}_m) + \mathrm{pen}(m)\}.$$

Thus, on Ξ_n, Inequality (49) holds for any $m\in\mathcal{M}_n$ and any $b_m\in S_m$. The decomposition $\gamma_n(t)-\gamma_n(s) = \|t-b\|_n^2 - \|s-b\|_n^2 + 2\nu_n(t-s)$, where $\nu_n(t) = \langle\vec{\varepsilon},t\rangle_n$, yields, for any $m\in\mathcal{M}_n$ and any $b_m\in S_m$,

$$\|\hat{b}_{\hat{m}}-b\|_n^2 \leq \|b_m-b\|_n^2 + 2\nu_n(\hat{b}_{\hat{m}}-b_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}).$$

We introduce, for $\|t\|_f^2 = \int t^2(u)f(u)du$, the unit ball

$$B^f_{m,m'}(0,1) = \{t \in S_m + S_{m'}, \|t\|_f = 1\}$$

and the set

$$\Omega_n = \left\{ \left| \frac{\|t\|_n^2}{\|t\|_f^2} - 1 \right| \leq \frac{1}{2},\ \forall t \in \bigcup_{m,m' \in \mathcal{M}_n^+} (S_m + S_{m'}) \setminus \{0\} \right\}. \tag{50}$$

We start by studying the expectation on Ω_n. On this set, the following inequality holds: $\|t\|_f^2 \leq 2\|t\|_n^2$. We get, on $\Xi_n \cap \Omega_n$,

$$\begin{aligned}
\|\hat{b}_{\hat{m}} - b\|_n^2 \leq & \|b_m - b\|_n^2 + \frac{1}{8}\|\hat{b}_{\hat{m}} - b_m\|_f^2 + (8 \sup_{t \in B^f_{\hat{m},m}(0,1)} \nu_n^2(t) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m})) \\
\leq & \Big(1 + \frac{1}{2}\Big)\|b_m - b\|_n^2 + \frac{1}{2}\|\hat{b}_{\hat{m}} - b\|_n^2 + 8\Big(\sup_{t \in B^f_{\hat{m},m}(0,1)} \nu_n^2(t) - p(m,\hat{m})\Big)_+ \\
& + \mathrm{pen}(m) + 8p(m,\hat{m}) - \mathrm{pen}(\hat{m}).
\end{aligned} \tag{51}$$

Here we state the following Lemma:

Lemma 6.7. *Assume that* **(A1)** *holds, and that* $\mathbb{E}(\varepsilon_1^6) < +\infty$. *Then* $\nu_n(t) = \langle \vec{\varepsilon}, t\rangle_n$ *satisfies*

$$\mathbb{E}\left[\Big(\sup_{t \in B^f_{\hat{m},m}(0,1)} \nu_n^2(t) - p(m,\hat{m})\Big)_+ \mathbf{1}_{\Xi_n \cap \Omega_n} \right] \leq \frac{C}{n}$$

where $p(m,m') = 8\sigma_\varepsilon^2 \max(m,m')/n$.

We see that, for $\kappa \geq \kappa_0 = 32$, we have $8p(m,\hat{m}) - \mathrm{pen}(\hat{m}) \leq \mathrm{pen}(m)$. Thus, by taking expectation in (51) and applying Lemma 6.7, it comes that, for all m in $\mathcal{M}_n$ and b_m in S_m,

$$\mathbb{E}\big[\|\hat{b}_{\hat{m}} - b_A\|_n^2 \mathbf{1}_{\Xi_n \cap \Omega_n}\big] \leq 3\mathbb{E}\big[\|b_m - b_A\|_n^2\big] + 2\mathrm{pen}(m) + \frac{16\,C}{n}. \tag{52}$$

The complement of Ω_n satisfies the following Lemma:

Lemma 6.8. *Assume that* **(A1)**-**(A2)** *hold. Then,* Ω_n *defined by (50) is such that* $\mathbb{P}(\Omega_n^c) \leq c/n^3$ *where* c *is a positive constant.*

We conclude as above (see equation (48)) by writing

$$\mathbb{E}\big[\|b - \hat{b}_{\hat{m}}\|_n^2 \mathbf{1}_{\Xi_n \cap \Omega_n^c}\big] \leq (\sqrt{\mathbb{E}\big[b^4(X_1)\big]} + \sqrt{\mathbb{E}\big[\varepsilon_1^4\big]})\sqrt{\mathbb{P}(\Omega_n^c)}.$$

This result, together with (52) ends the proof of Lemma 6.5. □

Proof of Lemma 6.7. We can not apply Talagrand's Inequality to the process ν_n itself as the noise is not bounded. This is why we decompose the variables ε_i as follows:

$$\varepsilon_i = \eta_i + \xi_i,\ \eta_i = \varepsilon_i \mathbf{1}_{|\varepsilon_i| \leq k_n} - \mathbb{E}\big[\varepsilon_i \mathbf{1}_{|\varepsilon_i| \leq k_n}\big].$$

Then we have

$$\nu_n(t) = \nu_{n,1}(t) + \nu_{n,2}(t), \qquad \nu_{n,1}(t) = \langle \eta, t\rangle_n, \quad \nu_{n,2}(t) = \langle \xi, t\rangle_n,$$

and

$$\Big(\sup_{t\in B^f_{\hat m,m}(0,1)} \nu_n^2(t) - p(m,\hat m)\Big)_+ \leq \Big(\sup_{t\in B^f_{\hat m,m}(0,1)} 2\nu_{n,1}^2(t) - p(m,\hat m)\Big)_+ + 2\sup_{t\in B^f_{\hat m,m}(0,1)} \nu_{n,2}^2(t). \tag{53}$$

We successively bound the two terms.

Let $(\bar\varphi_j)_{j\in\{1,\dots,\max(m,m')\}}$ be an orthonormal basis of $S_m + S_{m'}$ for the weighted scalar product $\langle\cdot,\cdot\rangle_f$. It is easy to see that:

$$\mathbb{E}\Big[\sup_{t\in B^f_{m',m}(0,1)} \nu_{n,1}^2(t)\Big] \leq \sum_{j\leq\max(m,m')} \frac{1}{n}\mathrm{Var}\Big(\eta_1\bar\varphi_j(X_1)\Big) \leq \sum_{j\leq\max(m,m')} \frac{1}{n}\mathbb{E}\left[\Big(\eta_1\bar\varphi_j(X_1)\Big)^2\right]$$
$$\leq \frac{1}{n}\mathbb{E}\big[\varepsilon_1^2\big] \sum_{j\leq\max(m,m')} \mathbb{E}\left[\bar\varphi_j^2(X_1)\right] = \frac{\sigma_\varepsilon^2\max(m,m')}{n} := H^2$$

since the definition of $\bar\varphi_j$ implies that $\int \bar\varphi_j^2(x)f(x)dx = 1$. Next

$$\sup_{t\in B^f_{m',m}(0,1)} \mathrm{Var}(\eta_1 t(X_1)) \leq \mathbb{E}\big[\eta_1^2\big] \sup_{t\in B^f_{m',m}(0,1)} \mathbb{E}\big[t^2(X_1)\big] \leq \sigma_\varepsilon^2 := v$$

since $\mathbb{E}\big[t^2(X_1)\big] = \|t\|_f^2$. Lastly

$$\sup_{t\in B^f_{m',m}(0,1)} \sup_{(u,x)} \big(|u|\mathbf{1}_{|u|\leq k_n}|t(x)|\big) \leq k_n \sup_{t\in B^f_{m',m}(0,1)} \sup_x |t(x)|.$$

For $t = \sum_{j=0}^{m-1} a_j\varphi_j$, we have $\|t\|_f^2 = \vec a'\Psi_m\vec a = \|\sqrt{\Psi_m}\vec a\|_{2,m}^2$. Thus, for any m,

$$\sup_{t\in B^f_m(0,1)} \sup_x |t(x)| \leq c_\varphi\sqrt{m} \sup_{\|\sqrt{\Psi_m}\vec a\|_{2,m}=1} \|\vec a\|_{2,m}$$
$$\leq c_\varphi\sqrt{m} \sup_{\|\vec u\|_{2,m}=1} \|\sqrt{\Psi_m^{-1}}\vec u\|_{2,m} = c_\varphi\sqrt{m}\sqrt{\|\Psi_m^{-1}\|_{\mathrm{op}}}.$$

Under condition (45) on $\mathcal{M}_n^+$, we have

$$\sqrt{m}\sqrt{\|\Psi_m^{-1}\|_{\mathrm{op}}} = \big(m\|\Psi_m^{-1}\|_{\mathrm{op}}^2\big)^{1/4} m^{1/4} \leq \left(4\mathfrak{d}\frac{n}{\log(n)}\right)^{1/4} m^{1/4}.$$

We can take

$$M_1 := c_\varphi k_n \left(4\mathfrak{d}\frac{n}{\log(n)}\right)^{1/4} (m\vee m')^{1/4}. \tag{54}$$

Consequently, the Talagrand Inequality (see Theorem 7.3) implies, for $p(m,m') = 8\frac{\sigma_\varepsilon^2\max(m,m')}{n}$, and denoting by $m^* := \max(m,m')$,

$$\mathbb{E}\left[\Big(\sup_{t\in B^f_{m,m'}(0,1)} [\nu_{n,1}]^2(t) - \frac{1}{2}p(m,m')\Big)_+\right] \leq \frac{C_1}{n}\left(e^{-C_2m^*} + \frac{k_n^2\sqrt{n}(m^*)^{1/2}}{n}e^{-C_3\frac{n^{1/4}(m^*)^{1/4}}{k_n}}\right).$$

So, we choose $k_n = n^{1/4}$ and we get,

$$\mathbb{E}\Big(\sup_{t\in B^f_{m',m}(0,1)} [\nu_{n,1}]^2(t) - \frac{1}{2}p(m,m')\Big)_+ \leq \frac{C_1'}{n}\left(\exp(-C_2 m^*) + (m^*)^{1/2}\exp(-C_3(m^*)^{1/4})\right).$$

By summing up all terms over $m' \in \mathcal{M}_n$, we deduce

$$\begin{aligned}\mathbb{E}\left[\Big(\sup_{t\in B^f_{\hat{m},m}(0,1)} [\nu_{n,1}]^2(t) - p(m,\hat{m})\Big)_+ \mathbf{1}_{\Xi_n}\right] &\leq \sum_{m'\in\mathcal{M}_n^+} \mathbb{E}\Big(\sup_{t\in B^f_{m',m}(0,1)} [\nu_{n,1}]^2(t) - p(m,m')\Big)_+ \\ &\leq \frac{C}{n}. \qquad (55)\end{aligned}$$

Let us now study the second term in (53). Recall that $M_n^+ \leq 4\mathfrak{d}n/\log(n)$ the dimension of the largest space of the collection. Then we have

$$\begin{aligned}\mathbb{E}\Big[\Big(\sup_{t\in B^f_{\hat{m},m}(0,1)} \nu^2_{n,2}(t)\mathbf{1}_{\Xi_n}\Big)_+\Big] &\leq \sum_{j=1}^{M_n^+} \mathbb{E}\big[\langle \xi, \bar{\varphi}_j\rangle_n^2\big] = \sum_{j=1}^{M_n^+} \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^n \xi_i\bar{\varphi}_j(X_i)\Big) \\ &= \frac{1}{n}\sum_{j=1}^{M_n^+} \mathbb{E}\big[\xi_1^2\big]\mathbb{E}\big[\bar{\varphi}_j^2(X_1)\big] \leq \frac{M_n^+}{n}\mathbb{E}\big[\varepsilon_1^2\mathbf{1}_{|\varepsilon_1|>k_n}\big] \\ &\leq \frac{M_n^+}{n}\frac{\mathbb{E}\big[|\varepsilon_1|^{2+p}\big]}{k_n^p} \leq C\frac{\mathbb{E}\big[\varepsilon_1^6\big]}{n},\end{aligned}$$

where the last line follows from the Markov inequality and the choices $k_n = n^{1/4}$ and $p = 4$. This bound together with (55) plugged in (53) gives the result of Lemma 6.7. □

Proof of Lemma 6.8. As the collection of models is nested, we have

$$\mathbb{P}(\Omega_n^c) \leq \sum_{m\in\mathcal{M}_n^+} \mathbb{P}\left(\exists t \in S_m, \left|\frac{\|t\|_n^2}{\|t\|_f^2} - 1\right| > \frac{1}{2}\right) = \sum_{m\in\mathcal{M}_n^+} \mathbb{P}(\Omega_m^c).$$

Now we proved in Lemma 6.3, that $\mathbb{P}(\Omega_m^c) \leq c/n^4$ if $m\|\Psi_m^{-1}\|_{\mathrm{op}} \leq (\mathfrak{c}/2)(n/\log(n))$. Here

$$m(\|\Psi_m^{-1}\|^2_{\mathrm{op}} \vee 1) \leq 4\mathfrak{d}\frac{n}{\log(n)} \Rightarrow m\|\Psi_m^{-1}\|_{\mathrm{op}} \leq 4\mathfrak{d}\frac{n}{\log(n)}.$$

Therefore, the result holds if $4\mathfrak{d} \leq \mathfrak{c}/2$, which is true. With the sum other a set of cardinality less than n, we get that $\mathbb{P}(\Omega_n^c) \leq c/n^3$. □

6.3.3. *Proof of Lemma 6.6.* We study first $\mathbb{P}(\mathcal{M}_n \not\subseteq \widehat{\mathcal{M}}_n) = \mathbb{P}(M_n > \widehat{M}_n)$. On this set, there exists $k \in \mathcal{M}_n$ such that $k \notin \widehat{\mathcal{M}}_n$.

For this index k, we have $k\|\Psi_k^{-1}\|^2_{\mathrm{op}} \leq \mathfrak{d}n/4\log(n)$ and $k\|\widehat{\Psi}_k^{-1}\|^2_{\mathrm{op}} > \mathfrak{d}n/\log(n)$. This implies, as

$$\mathfrak{d}\frac{n}{\log(n)} < k\|\widehat{\Psi}_k^{-1}\|^2_{\mathrm{op}} \leq 2k\|\Psi_k^{-1} - \widehat{\Psi}_k^{-1}\|^2_{\mathrm{op}} + 2k\|\Psi_k^{-1}\|^2_{\mathrm{op}} \leq 2k\|\Psi_k^{-1} - \widehat{\Psi}_k^{-1}\|^2_{\mathrm{op}} + \frac{\mathfrak{d}}{2}\frac{n}{\log(n)},$$

that $k\|\widehat{\Psi}_k^{-1}-\Psi_k^{-1}\|_{\rm op}^2 \geq \mathfrak{d}n/(4\log(n))$. Let us denote by

$$\Delta_m = \{m\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\rm op}^2 > \frac{\mathfrak{d}}{4}\frac{n}{\log(n)}\},$$

we have,

$$\mathbb{P}(\mathcal{M}_n \not\subseteq \widehat{\mathcal{M}}_n) \leq \sum_{m\in\mathcal{M}_n}\mathbb{P}(\Delta_m) \leq \sum_{m\in\mathcal{M}_n}\mathbb{P}(\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\rm op} > \|\Psi_m^{-1}\|_{\rm op}).$$

We have from (ii) of Proposition 2.4 and Ptoposition 2.3, that $\mathbb{P}(\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\rm op} > \|\Psi_m^{-1}\|_{\rm op}) \leq c/n^4$ for m satisfying (9) with $\mathfrak{c}$ given by (10). Indeed, we can conclude as in the proof of Lemma 6.8 above, because $\mathfrak{d}/4 \leq \mathfrak{c}/2$. Thus we proved that $\mathbb{P}(\mathcal{M}_n \not\subseteq \widehat{\mathcal{M}}_n) \leq c/n^3$.

Now we study $\mathbb{P}(\widehat{\mathcal{M}}_n \not\subseteq \mathcal{M}_n^+)$. On the set $(\widehat{\mathcal{M}}_n \not\subseteq \mathcal{M}_n^+)$, we can find a k satisfying

$$k\|\widehat{\Psi}_k^{-1}\|_{\rm op}^2 \leq \mathfrak{d}\frac{n}{\log(n)} \text{ and } k\|\Psi_k^{-1}\|_{\rm op}^2 > 4\mathfrak{d}\frac{n}{\log(n)},$$

therefore such that

$$k\|\widehat{\Psi}_k^{-1}\|_{\rm op}^2 \leq \mathfrak{d}\frac{n}{\log(n)} \text{ and } k\|\widehat{\Psi}_k^{-1}-\Psi_k^{-1}\|_{\rm op}^2 \geq \mathfrak{d}\frac{n}{\log(n)}.$$

Thus we have

$$\begin{aligned}\mathbb{P}(\widehat{\mathcal{M}}_n \not\subseteq \mathcal{M}_n^+) &\leq \sum_{k\leq \mathfrak{d}n/\log(n)}\mathbb{P}\left(k\|\widehat{\Psi}_m^{-1}\|_{\rm op}^2 \leq \mathfrak{d}\frac{n}{\log(n)} \text{ and } k\|\widehat{\Psi}_k^{-1}-\Psi_k^{-1}\|_{\rm op}^2 \geq \mathfrak{d}\frac{n}{\log(n)}\right)\\ &\leq \sum_{k\leq \mathfrak{d}n/\log(n)}\mathbb{P}\left(k\|\widehat{\Psi}_k^{-1}\|_{\rm op}^2 \leq \mathfrak{d}\frac{n}{\log(n)} \text{ and } \|\widehat{\Psi}_k^{-1}-\Psi_k^{-1}\|_{\rm op} \geq \|\widehat{\Psi}_k^{-1}\|_{\rm op}\right)\end{aligned}$$

Now, proceeding with Proposition 2.4 (ii) and interchanging $\widehat{\Psi}_m$ and Ψ_m, we get

$$\left\{\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\rm op} > \|\widehat{\Psi}_m^{-1}\|_{\rm op}\right\} \subset \left\{\|\widehat{\Psi}_m^{-1/2}\Psi_m\widehat{\Psi}_m^{-1/2}-{\rm Id}_m\|_{\rm op} > \frac{1}{2}\right\}.$$

Using $\|\widehat{\Psi}_m^{-1/2}\Psi_m\widehat{\Psi}_m^{-1/2}-{\rm Id}_m\|_{\rm op} \leq \|\widehat{\Psi}_m^{-1}\|_{\rm op}\|\Psi_m-\widehat{\Psi}_m\|_{\rm op}$, we get

$$\left\{\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\rm op} > \|\widehat{\Psi}_m^{-1}\|_{\rm op}\right\} \subset \left\{\|\widehat{\Psi}_m-\Psi_m\|_{\rm op} > \frac{1}{2}\|\widehat{\Psi}_m^{-1}\|_{\rm op}^{-1}\right\}.$$

Therefore

$$\begin{aligned}\mathbb{P}(\widehat{\mathcal{M}}_n \not\subseteq \mathcal{M}_n^+) &\leq \sum_{k\leq \mathfrak{d}n/\log(n)}\mathbb{P}\left(k\|\widehat{\Psi}_k^{-1}\|_{\rm op}^2 \leq \mathfrak{d}\frac{n}{\log(n)} \text{ and } \|\widehat{\Psi}_k-\Psi_k\|_{\rm op} \geq \frac{1}{2\|\widehat{\Psi}_k^{-1}\|_{\rm op}}\right)\\ &\leq \sum_{k\leq \mathfrak{d}n/\log(n)}\mathbb{P}\left(\|\widehat{\Psi}_k-\Psi_k\|_{\rm op} \geq \frac{1}{2}\sqrt{\frac{k\log(n)}{\mathfrak{d}n}}\right) \leq \frac{c}{n^2},\end{aligned}$$

by applying Proposition 2.4 and using the value of $\mathfrak{d}$ (this is where $\mathfrak{d}$ is chosen). □

6.3.4. *Proof of Inequality (30) of Theorem 4.1.* We have the following sequence of inequalities, for any $m \in \mathcal{M}_n$ and t any element of S_m,

$$\begin{aligned}
\|\hat{b}_{\hat{m}} - b_A\|_f^2 &= \|\hat{b}_{\hat{m}} - b_A\|_f^2 \mathbf{1}_{\Omega_n} + \|\hat{b}_{\hat{m}} - b_A\|_f^2 \mathbf{1}_{\Omega_n^c} \\
&\leq 2\|\hat{b}_{\hat{m}} - t\|_f^2 \mathbf{1}_{\Omega_n} + 2\|t - b_A\|_f^2 \mathbf{1}_{\Omega_n} + \|\hat{b}_{\hat{m}} - b_A\|_f^2 \mathbf{1}_{\Omega_n^c} \\
&\leq 4\|\hat{b}_{\hat{m}} - t\|_n^2 \mathbf{1}_{\Omega_n} + 2\|t - b_A\|_f^2 \mathbf{1}_{\Omega_n} + \|\hat{b}_{\hat{m}} - b_A\|_f^2 \mathbf{1}_{\Omega_n^c} \\
&\leq 8\|\hat{b}_{\hat{m}} - b_A\|_n^2 \mathbf{1}_{\Omega_n} + 8\|t - b_A\|_n^2 \mathbf{1}_{\Omega_n} + 2\|t - b_A\|_f^2 \mathbf{1}_{\Omega_n} + \|\hat{b}_{\hat{m}} - b_A\|_f^2 \mathbf{1}_{\Omega_n^c}
\end{aligned}$$

where Ω_n is defined by (50). Therefore, using the result of Theorem 4.1 and $\mathbb{E}(\|t-b_A\|_n^2) = \|t - b_A\|_f^2$, we get that for all $m \in \mathcal{M}_n$ and for any $t \in S_m$,

$$\mathbb{E}(\|\hat{b}_{\hat{m}} - b_A\|_f^2) \leq C_1 \left(\|t - b_A\|_f^2 + \sigma_\varepsilon^2 \frac{m}{n} \right) + \frac{C_2}{n} + \mathbb{E}\left(\|\hat{b}_{\hat{m}} - b_A\|_f^2 \mathbf{1}_{\Omega_n^c} \right), \tag{56}$$

so only the last term is to be studied. First, recall that Lemma 6.8 implies that $\mathbb{P}(\Omega_n^c) \leq \mathfrak{d}/n^3$. Next, write that $\|\hat{b}_{\hat{m}} - b_A\|_f^2 \leq 2(\|\hat{b}_{\hat{m}}\|_f^2 + \|b_A\|_f^2)$. As f is bounded, we use a slightly improved version of (41). Indeed, for all m,

$$\begin{aligned}
\|\Psi_m\|_{\text{op}} &= \sup_{\|\vec{x}\|_{2,m}=1} \vec{x}' \Psi_m \vec{x} = \sup_{\|\vec{x}\|_{2,m}=1} \int \left(\sum_{j=0}^{m-1} x_j \varphi_j(u) \right)^2 f(u) du \\
&\leq \|f\|_\infty \sup_{\|\vec{x}\|_{2,m}=1} \int \left(\sum_{j=0}^{m-1} x_j \varphi_j(u) \right)^2 du = \|f\|_\infty,
\end{aligned}$$

yields, as for $\hat{m} \in \widehat{\mathcal{M}}_n$, $\|\widehat{\Psi}_{\hat{m}}^{-1}\|_{\text{op}} \vee 1 \leq c\sqrt{n}$,

$$\|\hat{b}_{\hat{m}}\|_f^2 \leq \|f\|_\infty \frac{\|\widehat{\Psi}_{\hat{m}}^{-1}\|_{\text{op}}}{n} \left(\sum_{i=1}^n Y_i^2 \right) \leq \frac{C}{\sqrt{n}} \left(\sum_{i=1}^n Y_i^2 \right).$$

Then as $\mathbb{E}[(\sum_{i=1}^n Y_i^2)^2] \leq n^2 \mathbb{E}(Y_1^4)$, we get

$$\mathbb{E}(\|\hat{b}_{\hat{m}}\|_f^2 \mathbf{1}_{\Omega_n^c}) \leq \sqrt{\mathbb{E}(\|\hat{b}_{\hat{m}}\|_f^4) \mathbb{P}(\Omega_n^c)} \leq C \mathbb{E}^{1/2}(Y_1^4) \sqrt{n} \mathbb{P}^{1/2}(\Omega_n^c) \leq c'/n.$$

On the other hand $\mathbb{E}(\|b_A\|_f^2 \mathbf{1}_{\Omega_n^c}) \leq \|b_A\|_f^2 \mathbb{P}(\Omega_n^c) \leq c"/n^3$. Thus $\mathbb{E}\left(\|\hat{b}_{\hat{m}} - b_A\|_f^2 \mathbf{1}_{\Omega_n^c} \right) \leq c_1/n$ and plugging this in (56) ends the proof of Inequality (30) in Theorem 4.1. □

7. Theoretical tools

A proof of the following theorem can be found in Stewart and Sun (1990).

Theorem 7.1. *Let* $\mathbf{A}$, $\mathbf{B}$ *be* $(m \times m)$ *matrices. If* $\mathbf{A}$ *is invertible and* $\|\mathbf{A}^{-1}\mathbf{B}\|_{\text{op}} < 1$, *then* $\tilde{\mathbf{A}} := \mathbf{A} + \mathbf{B}$ *is invertible and it holds*

$$\|\tilde{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\|_{\text{op}} \leq \frac{\|\mathbf{B}\|_{\text{op}} \|\mathbf{A}^{-1}\|_{\text{op}}^2}{1 - \|\mathbf{A}^{-1}\mathbf{B}\|_{\text{op}}}$$

Theorem 7.2 (Bernstein Matrix inequality)**.** *Consider a finite sequence* $\{\mathbf{S}_k\}$ *of independent, random matrices with common dimension* $d_1 \times d_2$. *Assume that*

$$\mathbb{E}\mathbf{S}_k = 0 \quad \text{and} \quad \|\mathbf{S}_k\|_{\text{op}} \leq L \quad \textit{for each index } k.$$

Introduce the random matrix $\mathbf{Z} = \sum_k \mathbf{S}_k$. *Let* $\nu(\mathbf{Z})$ *be the the variance statistic of the sum:* $\nu(\mathbf{Z}) = \max\{\lambda_{\max}(\mathbb{E}[\mathbf{Z}'\mathbf{Z}]), \lambda_{\max}(\mathbb{E}[\mathbf{Z}\mathbf{Z}'])\}$. *Then*

$$\mathbb{E}\|\mathbf{Z}\|_{\text{op}} \leq \sqrt{2\nu(\mathbf{Z})\log(d_1+d_2)} + \frac{1}{3}L\log(d_1+d_2).$$

Furthermore, for all $t \geq 0$

$$\mathbb{P}\left[\|\mathbf{Z}\|_{\text{op}} \geq t\right] \leq (d_1+d_2)\exp\left(-\frac{t^2/2}{\nu(\mathbf{Z})+Lt/3}\right).$$

A proof can be found in Tropp (2012) or Tropp (2015).
We recall the Talagrand concentration inequality given in Klein and Rio (2005).

Theorem 7.3. *Consider* $n \in \mathbb{N}^*$, $\mathcal{F}$ *a class at most countable of measurable functions, and* $(X_i)_{i\in\{1,\dots,n\}}$ *a family of real independent random variables. Define, for* $f \in \mathcal{F}$, $\nu_n(f) = (1/n)\sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_i)])$, *and assume that there are three positive constants* M, H *and* v *such that* $\sup_{f\in\mathcal{F}}\|f\|_\infty \leq M$, $\mathbb{E}[\sup_{f\in\mathcal{F}}|\nu_n(f)|] \leq H$, *and* $\sup_{f\in\mathcal{F}}(1/n)\sum_{i=1}^n \text{Var}(f(X_i)) \leq v$. *Then for all* $\alpha > 0$,

$$\mathbb{E}\left[\left(\sup_{f\in\mathcal{F}}|\nu_n(f)|^2 - 2(1+2\alpha)H^2\right)_+\right] \leq \frac{4}{b}\left(\frac{v}{n}e^{-b\alpha\frac{nH^2}{v}} + \frac{49M^2}{bC^2(\alpha)n^2}e^{-\frac{\sqrt{2}bC(\alpha)\sqrt{\alpha}}{7}\frac{nH}{M}}\right)$$

with $C(\alpha) = (\sqrt{1+\alpha}-1)\wedge 1$, *and* $b = \frac{1}{6}$.

By density arguments, this result can be extended to the case where $\mathcal{F}$ is a unit ball of a linear normed space, after checking that $f \to \nu_n(f)$ is continuous and $\mathcal{F}$ contains a countable dense family.

References

[Abramowitz and Stegun, 1964] Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edition.

[Aksey and Wainger, 1965] Askey, R. and Wainger, S. (1965) Mean convergence of expansions in Laguerre and Hermite series. *Amer. J. Math.* **87**, 695-708.

[Baraud, 2000] Baraud, Y. (2000) Model selection for regression on a fixed design. *Probab. Theory Related Fields* **117**, 467-493.

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

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

[Belomestny et al., 2016] Belomestny, D., Comte, F., and Genon-Catalot, V. (2016). Nonparametric Laguerre estimation in the multiplicative censoring model. *Electron. J. Statist.*, 10(2):3114–3152.

[Belomestny et al., 2017] Belomestny, D., Comte, F., and Genon-Catalot, V. (2017). Sobolev-Hermite versus Sobolev nonparametric density estimation on R *The Annals of the Institute of Statistical Mathematics*, to appear.

[Birgé and Massart, 1998] Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4**, 329–375.

[Bouaziz et al., 2018] Bouaziz, O., Brunel, E. and Comte, F. (2018). Nonparametric survival function estimation for data subject to interval censoring case 2. Preprint hal-01766456.

[Brunel and Comte, 2009] Brunel, E. and Comte, F. (2009) Cumulative distribution function estimation under interval censoring case 1.*Electron. J. Stat.* **3** , 1-24.

[Cohen et al., 2013] Cohen, A., Davenport, M.A. and Leviatan, D. (2013). On the stability and acuracy of least squares approximations. *Found. Comput. math.* **13**, 819-834.

[Comte and Genon-Catalot, 2015] Comte, F. and Genon-Catalot, V. (2015). Adaptive Laguerre density estimation for mixed Poisson models. *Electron. J. Stat.*, **9**, 1112-1148.

[Comte and Genon-Catalot, 2018] Comte, F. and Genon-Catalot, V. (2018). Laguerre and Hermite bases for inverse problems. *Journal of the Korean Statistical Society*, **47**, 273-296.

[deVore and Lorentz, 1993] DeVore, R.A. and Lorentz, G.G. (1993) *Constructive approximation*, Springer-Verlag, Berlin.

[Efromovich, 1999] Efromovich, S. (1999) *Nonparametric curve estimation. Methods, theory, and applications.* Springer Series in Statistics. Springer-Verlag, New York.

[Klein and Rio, 2005] Klein, T. and Rio, E. (2005) Concentration around the mean for maxima of empirical processes. *Ann. Probab.* **33**, no. 3, 1060-1077.

[Mabon, 2017] Mabon, G. (2017). Adaptive deconvolution on the nonnegative real line. *Scandinavian Journal of Statistics*, 44:707-740.

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

[Plancade, 2011] Plancade, S. (2011) Model selection for hazard rate estimation in presence of censoring. *Metrika* **74**, 313-347.

[Stewart and Sun, 1990] Stewart, G. W. and Sun, J.-G. (1990). *Matrix perturbation theory.* Boston etc.: Academic Press, Inc.

[Szegö, 1975] Szegö, G. (1975) *Orthogonal polynomials.* Fourth edition. American Mathematical Society, Colloquium Publications, Vol. XXIII. American mathematical Society, Providence, R.I.

[Tropp, 2012] Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434.

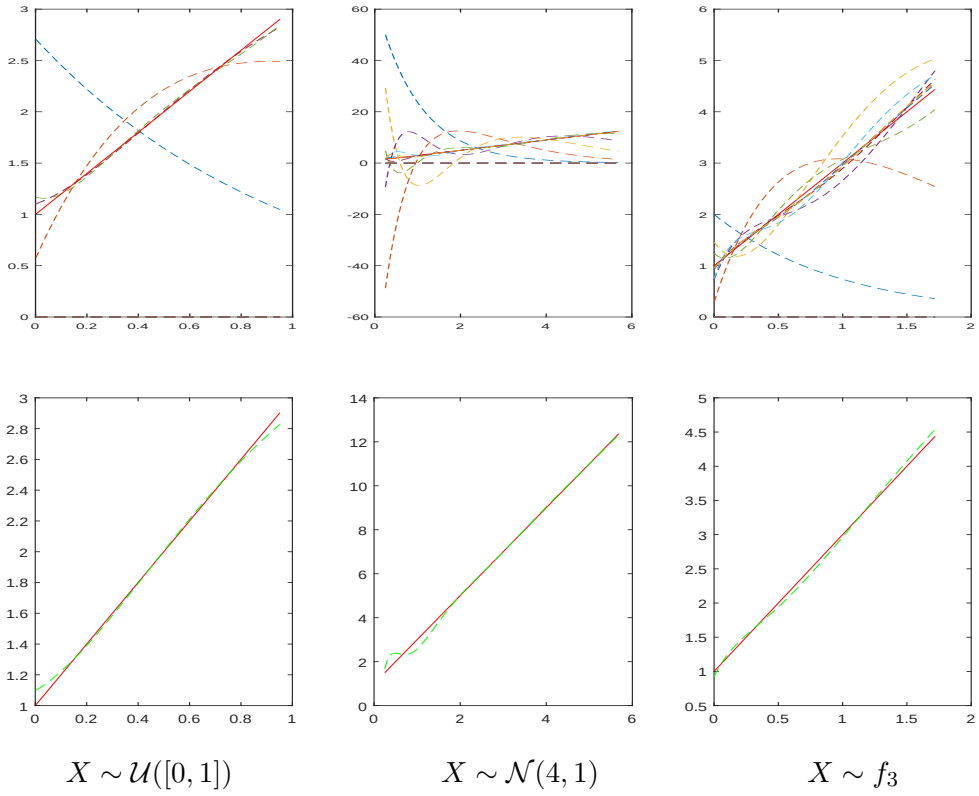
[Tropp, 2015] Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1-2):1–230.

[Tsybakov, 2009] Tsybakov, A. B. (2009) Introduction to nonparametric estimation. Springer Series in Statistics. Springer, New York.

[Watson, 1964] Watson, G.S. (1964) Smooth regression analysis. *Sankhyā, Series A*, **26**, 359-372.

Appendix A. Numerical illustrations

In this section, numerical illustrations of how our method works are presented. The estimation procedure is implemented for the Laguerre (Figures 1 to 4) and the Hermite basis (Figure 5). The $(\varepsilon_i)_{1\leq i\leq n}$ are generated as an i.i.d. sample of Gaussian $\mathcal{N}(0,\sigma^2)$ with $\sigma = 0.5$. Then, we choose different functions $b(.)$ (bounded or not) and different types of distribution of the design $(X_i)_{1\leq i\leq n}$. Typically, a linear function $x \mapsto 2x+1$ is experimented without the information of its linearity, which allows to test moment conditions; on the contrary, $x \mapsto 4x/(1+x^2)$ is bounded and should be easier to reconstruct. For the design density, we consider standard uniform or Gaussian cases, and also different heavy tailed distributions.

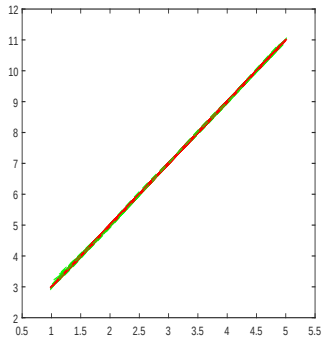


$X \sim \mathcal{U}([0,1])$ $X \sim \mathcal{N}(4,1)$ $X \sim f_3$

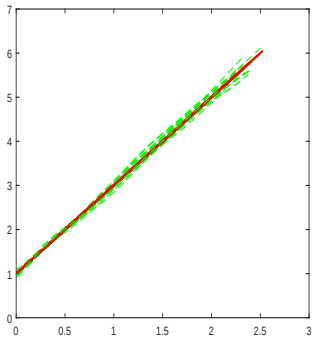
Figure 1. First line: beam of the proposals $\hat{f}_m$ for $m = 1$ to $m_{\max}$ in the Laguerre basis. Second line: the estimator as selected by the procedure, $\hat{f}_{\hat{m}}$. Function $b(x) = 2x+1$, $n = 1000$, density $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x\geq 0}$.

In Figure 1, we plot in the first line the collection of estimators in the Laguerre basis, among which the algorithm makes the selection. The number of computed estimators is different from one example to another, as the collection of models $\widehat{\mathcal{M}}_n$ is random and depends on $\|\widehat{\Psi}_m^{-1}\|_{\text{op}}$. In the practical implementation, we consider the (random) maximum value $m_{\max}$ such that $\|\widehat{\Psi}_m^{-1}\|_{\text{op}} \leq n$, since inversion of the matrix $\widehat{\Psi}_m$ remains possible in such cases. Surprisingly, we can see that very few estimators are sometimes computed

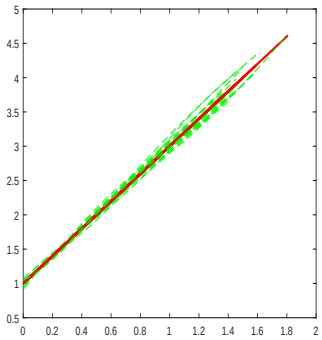
(see the example of uniform distribution on the right). They are also very different from one dimension to another. The second line presents the final estimator, selected by the procedure. In the example of Figure **??**, the curve is linear, and is perfectly estimated, although its particular form is unknown and was not *a priori* easy to obtain with the Laguerre basis.



$\mathcal{N}(3,1)$
$\bar{\hat{m}} = 7.7(0.5), \bar{m}_{\max} = 8.8(0.4)$

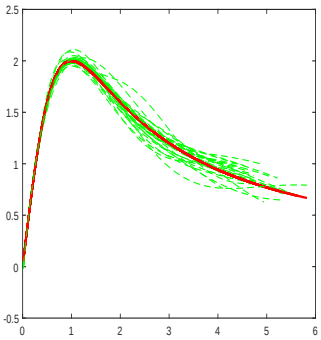


f_k with $k=4$
$\bar{\hat{m}} = 9.9(1.9), \bar{m}_{\max} = 10.4(1.7)$

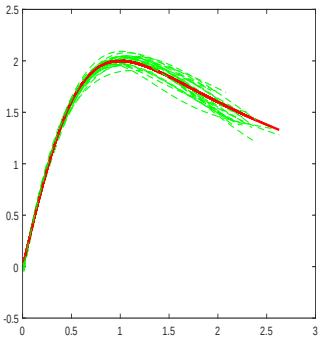


f_k with $k=5$
$\bar{\hat{m}} = 6.7(1.1), \bar{m}_{\max} = 7.6(1.0)$

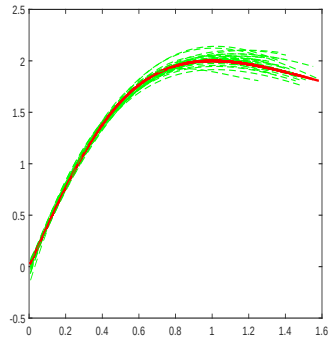
FIGURE 2. 25 estimated curves in Laguerre basis (dotted -green/grey), the true in bold (red), $n = 1000$, $b(x) = 2x+1$ and different laws for the design, $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x\geq 0}$.



$k=3$
$\bar{\hat{m}} = 5.6(1.0), \bar{m}_{\max} = 15.6(0.7)$



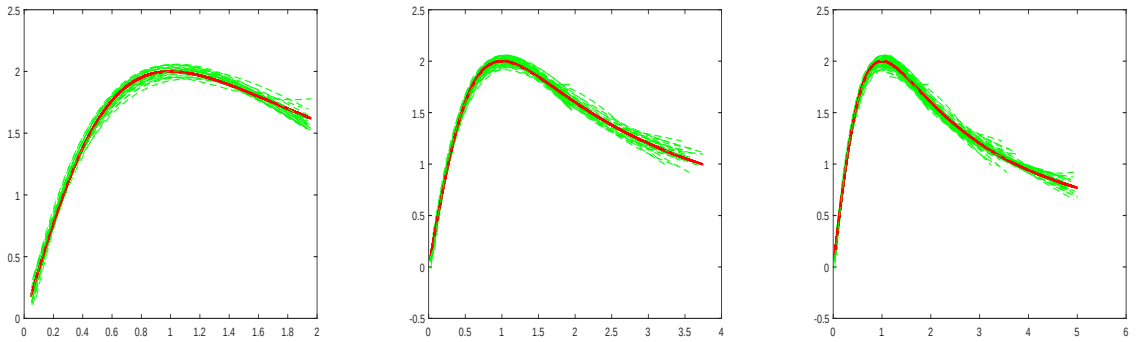
$k=4$
$\bar{\hat{m}} = 4.6(1.1), \bar{m}_{\max} = 10,1(1.6)$



$k=5$
$\bar{\hat{m}} = 3.5(1.3), \bar{m}_{\max} = 7.2(1.0)$

FIGURE 3. 25 estimated curves in the Laguerre basis (dotted -green/grey), the true in bold (red), $n = 1000$, density $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x\geq 0}$ for $k = 3, 4$ and 5, $b(x) = 4x/(1+x^2)\mathbf{1}_{x\geq 0}$.

In Figures 2, 3 and 4, we present beams of 25 estimators computed in the Laguerre basis, they give information about the variability of the procedure. Figure 2 is complementary of Figure 1 and considers the same linear regression function with similar distributions for

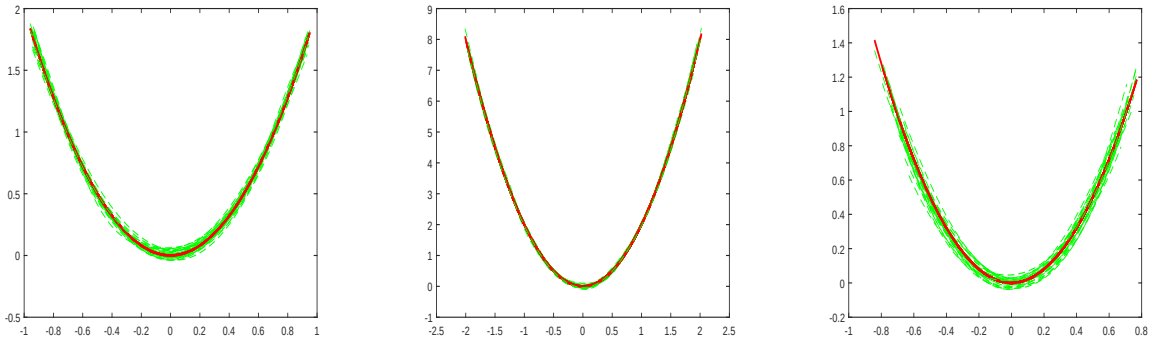


$X \sim \mathcal{U}([0,2])$ $X \sim \mathcal{E}(1)$ $X \sim \mathcal{N}(3,1)$

$\bar{\hat{m}} = 2.5(0.8), \bar{m}_{\max} = 5.0(0)$ $\bar{\hat{m}} = 5.1(0.8), \bar{m}_{\max} = 9.4(0.6)$ $\bar{\hat{m}} = 5.1(0.7), \bar{m}_{\max} = 8.9(0.3)$

FIGURE 4. 25 estimated curves in Laguerre basis (dotted -green/grey), the true in bold (red), $n = 1000$, $b(x) = 4x/(1+x^2)\mathbf{1}_{x\geq 0}$ and different laws for the design.

X, and Figure 3 presents the results for the function $b(x) = 4x/(1+x^2)\mathbf{1}_{x\geq 0}$ and different heavy tailed distributions for X. The beams illustrate the stability of the algorithm, with some design distributions leading to better results, probably due to higher signal-to-noise ratio. The interest of the linear case is also to illustrate the sharpness of the moment conditions: indeed the condition $\mathbb{E}[b^2(X_1)] < +\infty$ for X with density $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x\geq 0}$ is satisfied for $k > 3$ and the condition $\mathbb{E}[b^4(X_1)] < +\infty$ holds for $k > 5$. We checked, in the case of linear $b(.)$, that the method does not work for $k = 2, 3$, but the last two plots of Figure 2 show that it works rather well for $k = 4, 5$. The minimal theoretical condition may thus be weakened from $\mathbb{E}[b^4(X_1)] < +\infty$ to $\mathbb{E}[b^2(X_1)] < +\infty$. The Hermite basis has similar behaviour and an example is provided in Figure 5.



$X \sim \mathcal{U}([-1,1])$ $X \sim \mathcal{N}(0,1)$ $X \sim Laplace/4$

$\bar{\hat{m}} = 5.1(0.4), \bar{m}_{\max} = 7.0(0)$ $\bar{\hat{m}} = 13.4(1.5), \bar{m}_{\max} = 15.0(0.9)$ $\bar{\hat{m}} = 6.2(1.1), \bar{m}_{\max} = 8.1(0.6)$

FIGURE 5. 25 estimated curves in Hermite basis (dotted -green/grey), the true in bold (red), $n = 1000$, $b(x) = 2x^2$ and different laws for the design.

Below each plot, we give the density of the design and the value of $\bar{\hat{m}}$ which is the mean of the selected dimensions for the 25 estimators represented on the figure, with standard deviation in parenthesis. It is associated with the value of $\bar{m}_{\max}$ which is the mean of the maximal dimension for which the estimator is computed, with standard deviation in parenthesis. We can see that the maximal dimension is rather small (less than ten models are compared for selection, in general) but an adequate choice seems always to exist in this small collection. This means that the squared-bias variance compromise in the restricted set $\mathcal{M}_n$ has good performance and that the non compact Laguerre and Hermite bases are very interesting and simple estimation tools. Indeed, the method is very fast and this low complexity, already argued in Belomestny *et al.* (2017), has an important practical interest.