REGRESSION FUNCTION ESTIMATION ON NON COMPACT SUPPORT AS A PARTLY INVERSE PROBLEM

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

Abstract. This paper is about nonparametric regression function estimation, first in the independent setting and in a second stage, in the context of an autoregressive model a setting corresponding to dependent variables. 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 all previous results. Upper and lower bounds for new rates are provided. July 11, 2018

MSC2010 *Subject classifications.* 62G08 - 62M05
Key words and phrases. Hermite basis. Laguerre basis. Model selection. Mixing process. 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, the choice of two smoothing parameters.
The second method, proposed by Birgé and Massart (1998), Barron *et al.* (1999), and improved by Baraud (2000, 2002), for fixed and random design, is based on a least squares contrast, analogous to the one used for parametric linear regression:

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

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

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

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. Indeed, it provides directly an estimator of b restricted to the set A, a unique 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, 2000, 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. All these arguments are very favorable to 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.
Our aim in this work is to obtain theoretical results in regression function estimation by a projection method in the case of non compact support A of the basis. Indeed, several bases, such as the Laguerre ($A = \mathbb{R}^+$) or the Hermite ($A = \mathbb{R}$) basis, are not compactly supported. Nonparametric density estimation by a projection method on these bases has been the subject of several recent contributions (see *e.g.* Comte *et al.* 2015, Comte and Genon-Catalot, 2015, 2018, Belomestny *et al.* 2016), showing that theses bases are convenient and easy to handle.

What is new in the present work? Our findings are of three types:

- First, we propose a new definition of the model selection procedure for regression function estimation on a set A whether compact or not and prove that it reaches a bias-variance tradeoff in a way that generalizes part of Baraud's (2002) theorems to the non compact case. In other words, the study presented here encompasses the usual setting, that can be recovered as a particular case (see Section 4.1).
- Second, 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, but not directly as a weight on the variance term, only in the definition of the collection of models. The resulting restriction is related to a property of the design density, but rather to its rates of decrease near infinity than to its regularity (see Proposition 4.3).
- Third, 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.

- Lastly, using non compactly supported bases has the advantage that it does not require a preliminary definition of the support of the basis. The support of a compactly supported basis in an estimation problem is generally considered as fixed in the theoretical part, while it is in practice adjusted on the data.

We also extend the method to dependent models, namely autoregressive models in geometric β-mixing framework (extension of Baraud *et al.* (2001a)).
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 $b_A \in \mathbb{L}^2(A, f(x)dx)$, i.e. $\mathbb{E}b_A^2(X_1) < +\infty$ 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. Section 3 contains our main results. 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. 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. In Section 4, we show how to recover existing results for compactly supported bases and illustrate the case of non compact support with the Hermite and Laguerre bases for estimation on $A = \mathbb{R}$ and $A = \mathbb{R}^+$ respectively. Lastly, Section 5 is devoted to the dependent context of a discrete time autoregressive process. Section 6 gives some concluding remarks. Most proofs are gathered in Section 7 while Section 8 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.

2.1. Definition of the projection estimator.

Consider model (1). Let $A \subset \mathbb{R}$ and let $(\varphi_j, j = 0, \ldots, m-1)$ be an orthonormal system of A-supported functions belonging to $\mathbb{L}^2(A, dx)$. Define $S_m = \text{span}(\varphi_0, \ldots, \varphi_{m-1})$, the linear space spanned by $(\varphi_0, \ldots, \varphi_{m-1})$. 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). Clearly, $\gamma_n(t) = n^{-1}\sum_{i=1}^n [Y_i - t(X_i)]^2 - n^{-1}\sum_{i=1}^n Y_i^2$, so that we recognize a classical least squares contrast. 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

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

We 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 $\hat{b}_m = \sum_{j=0}^{m-1}\hat{a}_j^{(m)}\varphi_j$. Classical computations give, assuming that $\widehat{\Psi}_m$ is invertible almost surely (a.s.), that

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

2.2. **Risk bound on a fixed space.** We now evaluate the risk of the estimator, without any constraint on the basis support. The result hereafter is classical, but 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*

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

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

where f denotes the common density of the X_i's.

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 are in order: is Ψ_m invertible for any m, and does the bias tend to zero when m grows to infinity? The two Lemmas below provide sufficient conditions. Note that 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 and relies on the following remarks. 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 $g \in \mathbb{L}^2(A, dx)$, $\int g\phi_j = 0$, $\forall j \geq 0$ means that $\int \varphi_j(g\sqrt{f}) = 0$ $\forall j \geq 0$ and implies $g = 0$ using our assumptions.
The result stated in Proposition 2.1 is general in the sense that it holds for any basis support, whether compact or not. We want here specifically to stress that (5) is an equality, in particular for the variance order, and this order 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 this 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*

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

The variance term in this case has the order of $\sum_{j=0}^{m-1}\mathbb{E}[\varphi_j^2(X_1)]/n$*. For most classical 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 4.1); but it is proved in Comte and Genon-Catalot (2018) that for Laguerre or Hermite basis (see section 4.2 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 be sure that the variance order does not depend on the basis in regression context.*

A consequence of these facts is that, as the bias is getting small when m grows, while the variance increases, a compromise has to be found, and m has to be relevantly chosen.

3. Main results

We may consider from Proposition 2.1 that the problem is standard. However, 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. We shall show that classical assumptions in nonparametric estimation relying on compactly supported bases and variables are in fact useful to bound from below these eigenvalues independently of m and make the different norms involved in the regression problem equivalent. On the contrary, for non compactly supported bases, the eigenvalues of Ψ_m may be not bounded from below (see Propositions 4.2-4.3 and the discussion thereafter in the Hermite and Laguerre bases cases).

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

$$\| \sum_{j=0}^{m-1} \varphi_j^2 \|_\infty \leq c_\varphi^2 m \quad \text{for} \quad c_\varphi^2 > 0 \quad \text{a constant.} \tag{7}$$

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

For M a matrix, we denote by $\|M\|_{\text{op}}$ 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, $\|MN\|_{\text{op}} \leq \|M\|_{\text{op}} \|N\|_{\text{op}}$.

3.1. **Adaptive procedure.** We present now our first main result, which concerns a model selection procedure in a general setting and associated risk bounds.

To select the most relevant space S_m, we proceed by choosing

$$\hat{m} = \arg \min_{m \in \widehat{\mathcal{M}}_n} \left\{ -\|\hat{b}_m\|_n^2 + \kappa \sigma_\varepsilon^2 \frac{m}{n} \right\} \tag{8}$$

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

$$\widehat{\mathcal{M}}_n = \left\{ m \in \mathbb{N}, \sup_{1 \leq k \leq m} k(\|\widehat{\Psi}_k^{-1}\|_{\text{op}}^2 \vee 1) \leq \mathfrak{c} \frac{n}{\log(n)} \right\}, \quad \mathfrak{c} = \frac{1}{192\, c_\varphi^2 (\|f\|_\infty + (1/3))}. \tag{9}$$

The value of the constant $\mathfrak{c}$ is determined below by Lemma 7.5.

In practice, we set $\hat{b}_m^T = \hat{b}_m$ if $m\|\widehat{\Psi}_m^{-1}\|_{\text{op}}^2 \leq 4\mathfrak{c}n/\log(n)$, and $\hat{b}_m^T = 0$ otherwise. A theoretical counterpart of $\widehat{\mathcal{M}}_n$ is useful:

$$\mathcal{M}_n = \left\{ m \in \mathbb{N}, \sup_{1 \leq k \leq m} k\, (\|\Psi_k^{-1}\|_{\text{op}}^2 \vee 1) \leq \frac{\mathfrak{c}}{4} \frac{n}{\log(n)} \right\}, \tag{10}$$

where $\mathfrak{c}$ is defined in (9).

To justify (8), let us explain how 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. The procedure aims at performing an automatic bias-variance tradeoff.

Theorem 3.1. *Let $(X_i, Y_i)_{1\le i\le n}$ 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 $\kappa \ge \kappa_0$, we have*

$$\mathbb{E}\big[\|\hat{b}_{\hat{m}} - b_A\|_n^2\big] \le C \inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m}\|b_A - t\|_f^2 + \kappa\sigma_\varepsilon^2\frac{m}{n}\right) + \frac{C'}{n}, \tag{11}$$

and

$$\mathbb{E}\big[\|\hat{b}_{\hat{m}} - b_A\|_f^2\big] \le C_1 \inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m}\|b_A - t\|_f^2 + \kappa\sigma_\varepsilon^2\frac{m}{n}\right) + \frac{C_1'}{n} \tag{12}$$

where C, C_1 are a numerical constants and C', C_1' are constants depending on f, b, σ_ε.

Theorem 3.1 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.

Remark 3.1. *The constant $\mathfrak{c}$ 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 $4\mathfrak{c}n/\log(n)$ by $n/\log^2(n)$ in the definitions of the sets $\widehat{\mathcal{M}}_n, \mathcal{M}_n$ 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.

The key tool for proving Theorem 3.1 is Proposition 3.1 which relies on a matricial Bernstein deviation inequality proved in Tropp (2015). The result encompasses all possible classical bases, whether compactly supported or not.

Proposition 3.1. *Assume that* **(A1)** *and* **(A2)** *hold. Let $\widehat{\Psi}_m$ be the $m\times m$ matrix defined in Equation* (3)*. Then for all $u > 0$*

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

3.2. **Rate and optimality.** 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, if we introduce a cutoff and define

$$\tilde{b}_m := \hat{b}_m \mathbf{1}_{m\in\widehat{\mathcal{M}}_n},$$

then we can prove:

Proposition 3.2. *Under* **(A1)** *and* **(A2)** *and $\mathbb{E}(Y_1^4) < +\infty$, for $b_A \in \mathbb{L}^2(A, f(x)dx)$,*

$$\mathbb{E}\big[\|\tilde{b}_m - b_A\|_f^2\big] \le C_0 \inf_{t\in S_m}\|b_A - t\|_f^2 + C_0'\frac{m}{n}, \tag{13}$$

where C_0 is a numerical constant and C_0' a constant depending on σ_ε^2, $\mathbb{E}(Y_1^4)$ and $\|b_A\|_f$.

We do not give the proof of this proposition as it is almost identical to the proof of inequality (12) by changing $\hat{m}$ into m.

So far, the bias rate of the $\mathbb{L}^2(A, f(x)dx)$-risk in (6) and (13) has not been assessed. To this end, we introduce regularity spaces related to f by setting:

$$W_f^s(A,R) = \left\{ h \in \mathbb{L}^2(A, f(x)dx), \forall \ell \geq 1, \|h - h_\ell^f\|_f^2 \leq R\ell^{-s} \right\} \tag{14}$$

where we recall that h_ℓ^f is the $\mathbb{L}^2(A, f(x)dx)$-orthogonal projection of h on S_ℓ. The idea behind this definition is the following. Considering an orthonormal basis $(\varphi_j)_{j\geq 0}$ of $\mathbb{L}^2(A, dx)$, associated regularity spaces are standardly defined by:

$$W^s(A,R) = \{h \in \mathbb{L}^2(A, dx), \quad \sum_{j\geq 0} j^s \langle h, \varphi_j\rangle^2 \leq R\}, \tag{15}$$

which describe the rate of decay of the coefficients of the function on the basis. If $h \in W^s(A,R)$, then $\forall \ell \geq 1$, $\|h - h_\ell\|^2 \leq R\ell^{-s}$ where h_ℓ is the $\mathbb{L}^2(A, dx)$-orthogonal projection of h on S_ℓ. And this property is precisely the one used to determine the bias rate of projection estimators built with the basis $(\varphi_j)_{j\geq 0}$. This explains the definition of $W_f^s(A,R)$. Note that, if $h \in W^s(A,R)$ and f is bounded, then

$$\|h - h_\ell^f\|_f^2 \leq \|f\|_\infty \|h - h_\ell\|^2 \leq \|f\|_\infty R\ell^{-s}.$$

Thus, $W^s(A,R) \subset W_f^s(A, R\|f\|_\infty)$.

From (13), we easily deduce an upper bound for the risk, which we state below. The rate obtained is optimal, as we also prove the following lower bound.

Theorem 3.2. *Assume that $b_A \in W_f^s(A,R)$,* **(A1)-(A2)** *hold and that $m_{\rm opt} := n^{1/(s+1)} \in \mathcal{M}_n$.*

- *Upper bound.* $\mathbb{E}(\|\tilde{b}_{m_{\rm opt}} - b_A\|_f^2) \leq Cn^{-s/(s+1)}$.
- *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 $n^{1/(s+1)} \in \mathcal{M}_n$, which imposes $m_{\rm opt}\|\Psi_{m_{\rm opt}}^{-1}\|_{\rm op}^2 \leq \mathfrak{c}n/\log(n)$, is actually mainly a constraint on f, see the discussion in Section 4.2.

The partly inverse problem appears here. The rate of $\|\Psi_m^{-1}\|_{\rm op}^2$ 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 3.3. *Under the assumptions of Theorem 3.2, and if moreover $\|\Psi_m^{-1}\|_{\rm op} \asymp m^k$, then*

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

In the following section, we make the link with previous results, in particular those of the compact support case.

4. Consequences and interpretation

4.1. **Case of compact A and compactly supported bases.** In this section, we assume that A is compact, that

$$b_A \in \mathbb{L}^2(A, dx) \text{ and } (\mathbf{A2}) \text{ holds.} \tag{16}$$

We show that Theorem 3.1 contains and improves existing results when the bases are regular and compactly supported, a case considered by most authors.

Let us give first examples of such bases; for simplicity, we take $A = [0, 1]$. 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 $j = 1, \ldots, (m-1)/2$.

All these collections satisfy (7) with $c_\varphi^2 = 1$ for histograms and trigonometric basis (for which $\sum_{j=0}^{m-1} \varphi_j^2(x) \equiv 1$), $c_\varphi^2 = r+1$ for piecewise polynomials with degree r. The trigonometric spaces are nested; for histograms and piecewise polynomials, the models are nested if the subdivisions are diadic ($m = 2^k$ for increasing values of k).

Clearly, if A is compact, assumption (16) is weak. More importantly, when the basis has compact support A, one can assume that

$$\exists f_0 > 0, \text{ such that } \forall \in A, \quad f(x) > f_0. \tag{17}$$

This assumption is commonly used in papers on nonparametric regression and is crucial in all these results. In particular, this implies that Ψ_m is invertible, and more precisely:

Proposition 4.1. *Assume that Assumption (17) is satisfied, then*

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

Indeed (17) implies that, for $\vec{u} = (u_0, \ldots, u_{m-1})'$ a vector of $\mathbb{R}^m$,

$$\vec{u}\,'\,\Psi_m\,\vec{u} = \int_A \left(\sum_{j=0}^{m-1} u_j\varphi_j(x)\right)^2 f(x)dx \;\geq f_0 \int_A \left(\sum_{j=0}^{m-1} u_j\varphi_j(x)\right)^2 dx = f_0\|\vec{u}\|_{2,m}^2. \tag{18}$$

Therefore $\|\Psi_m^{-1}\|_{\text{op}} \leq 1/f_0$ and Proposition 4.1 is proved.

A consequence of (17) is thus that we can choose

$$\mathcal{M}_n = \{m, m \leq c'n/\log(n)\}, \tag{19}$$

for a constant $c' = \mathfrak{c}f_0^2$. The unknown matrix Ψ_m no more appears in the definition of $\mathcal{M}_n$. The constant c' depends on f_0 but either f_0 is replaced by an estimator, or it is hidden by changing $\log(n)$ into $\log^{1+\delta}(n)$, $\delta > 0$ and taking n large enough.

The results of Theorem 3.1 under assumption (17) correspond to case (K1) of Theorem 1.1 (see also inequality (15)) in Baraud (2002, p.132), under similar moment condition on the noise. Note that our constraint $m \leq c'n/\log(n)$ is better than the one imposed in Baraud (the constraint in Baraud (2002) for a non localized basis such as the trigonometric basis is $m \leq c\sqrt{n/\log^3(n)}$ and thus stronger).

Now, let us discuss about the usual rates in this compact setting. Assume that assumption (16) holds. Then $\forall t \in S_m$, $\|b_A - t\|_f^2 \leq \|f\|_\infty \|b_A - t\|_A^2$ and thus

$$\inf_{t \in S_m} \|b_A - t\|_f^2 \leq \|f\|_\infty \|b_A - b_m\|_A^2 \tag{20}$$

where b_m is the $\mathbb{L}^2(A, dx)$-orthogonal projection of b_A on S_m.

If in addition, A is compact, if the basis is one of the bases described above and f satisfies Assumption (17), then we can compute an upper bound for the convergence rate of the estimator $\tilde{b}_m$. Assume that b_A belongs to a Besov ball $\mathcal{B}_{\alpha,2,\infty}(A, R)$ (see De Vore and Lorentz (1993), or Baraud (2002, section 2)), then we get with (20) that $\inf_{t \in S_m} \|b_A - t\|_f^2 \lesssim m^{-2\alpha}$. So, choosing $m_{\text{opt}} = n^{1/2\alpha+1}$, we find $\mathbb{E}(\|\tilde{b}_{m_{\text{opt}}} - b_A\|^2) \leq C(R) n^{-2\alpha/(2\alpha+1)}$. In other words, we recover the minimax-optimal rate of convergence for the risk bound of both $\tilde{b}_{m_{\text{opt}}}$ and the adaptive estimator $\hat{b}_{\hat{m}}$. Moreover, the dimension m_{opt} belongs to the set $\mathcal{M}_n$ given by (19), and the rate is obtained with respect to the $\mathbb{L}^2(A, dx)$-norm. By (20), it follows that, if $b_A \in \mathcal{B}_{\alpha,2,\infty}(A, R)$, then $b_A \in W_f^{2\alpha}(A, R)$.

4.2. **Examples of non compact A and non compactly supported bases.** In the case of non compact A, assumption (17) 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 $\widehat{\mathcal{M}}_n$.

We assume in this section that

$$b_A \in \mathbb{L}^2(A, f(x)dx), \quad \lambda(A \cap \text{supp}(f)) > 0, \text{ and } f \text{ is upper bounded.} \tag{21}$$

We illustrate our general result through 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

$$L_j(x) = \sum_{k=0}^{j} (-1)^k \binom{j}{k} \frac{x^k}{k!}, \qquad \ell_j(x) = \sqrt{2} L_j(2x) e^{-x} \mathbf{1}_{x \geq 0}, \quad j \geq 0. \tag{22}$$

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)):

$$\forall j \geq 0, \ \ \forall x \in \mathbb{R}^+, \ \ |\ell_j(x)| \leq \sqrt{2}. \tag{23}$$

Clearly, the collection of models $(S_m = \text{span}\{\ell_0, \ldots, \ell_{m-1}\})$ is nested, and (23) implies that this basis satisfies the general assumption (7) with $c_\varphi^2 = 2$. For a function $\theta \in \mathbb{L}^2(\mathbb{R}^+, dx)$, we can develop θ on the Laguerre basis with: $\theta = \sum_{j \geq 0} a_j(\theta) \ell_j$, $a_j(\theta) = \langle \theta, \ell_j \rangle = \int \theta(x) \ell_j(x) dx$.

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

$$H_j(x) = (-1)^j e^{x^2} \frac{d^j}{dx^j}(e^{-x^2}), \quad h_j(x) = c_j H_j(x) e^{-x^2/2}, \quad c_j = \left(2^j j! \sqrt{\pi}\right)^{-1/2} \tag{24}$$

The sequence $(h_j, j \geq 0)$ is an orthonormal basis of $\mathbb{L}^2(\mathbb{R}, dx)$. When a function θ belongs to $\mathbb{L}^2(\mathbb{R}, dx)$, it can be developed in the Hermite basis $\theta = \sum_{j\geq 0} a_j(\theta) h_j$ where $a_j(\theta) = \int_{\mathbb{R}} \theta(x) h_j(x) dx = \langle \theta, h_j \rangle$. The infinite norm of h_j satisfies (see Abramowitz and Stegun (1964), Szegö (1959) p.242):

$$\|h_j\|_\infty \leq \Phi_0, \qquad \Phi_0 \simeq 1,086435/\pi^{1/4} \simeq 0.8160, \tag{25}$$

so that the Hermite basis satisfies the general assumption (7) with $c_\varphi^2 = \Phi_0^2$. 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 and denote by $S_m = \text{span}(\varphi_0, \varphi_1, \ldots, \varphi_{m-1})$ 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 grows to infinity.
The matrices Ψ_m and $\widehat{\Psi}_m$ in these bases have specific properties. The first result concerns $\widehat{\Psi}_m$.

Lemma 4.1. *For all $m \in \mathbb{N}$, for all $m \leq n$, $\widehat{\Psi}_m$ is a.s. invertible.*

Proof ol Lemma 4.1. For all $\vec{u} = (u_0, \ldots, u_{m-1})' \in \mathbb{R}^m \setminus \{\vec{0}\}$, for $t(x) = \sum_{j=0}^{m-1} u_j \varphi_j(x)$, $\vec{u}\,' \widehat{\Psi}_m \vec{u} = \|t\|_n^2 \geq 0$. Thus $\|t\|_n = 0 \Rightarrow t(X_i) = 0$ for $i = 1, \ldots, n$. 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 $\vec{u} = \vec{0}$. □

k	2	3	4	5
$\hat{b}_1$	2.09	3.16	4.21	5.58
$\hat{b}_2$	0.68	1.44	2.05	2.67
$\hat{b}_2(k)/\hat{b}_2(2)$	1.00	2.11	3.02	3.92

TABLE 1. Estimated slope regression coefficients, $\hat{b}_1$ for left curves and $\hat{b}_2$ for right curves, of Figure 1.

The second result is crucial for understanding our procedure.

Proposition 4.2. *For all m, Ψ_m is invertible and there exists a constant $c^\star$ such that,*

$$\|\Psi_m^{-1}\|_{\text{op}}^2 \geq c^\star m. \tag{26}$$

In the Laguerre and Hermite cases, Inequality (26) 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 the definition (10) of $\mathcal{M}_n$ leads to restrictions on the values m that can be considered in the upper risk bound of Theorem 3.1. This is illustrated by the next proposition.

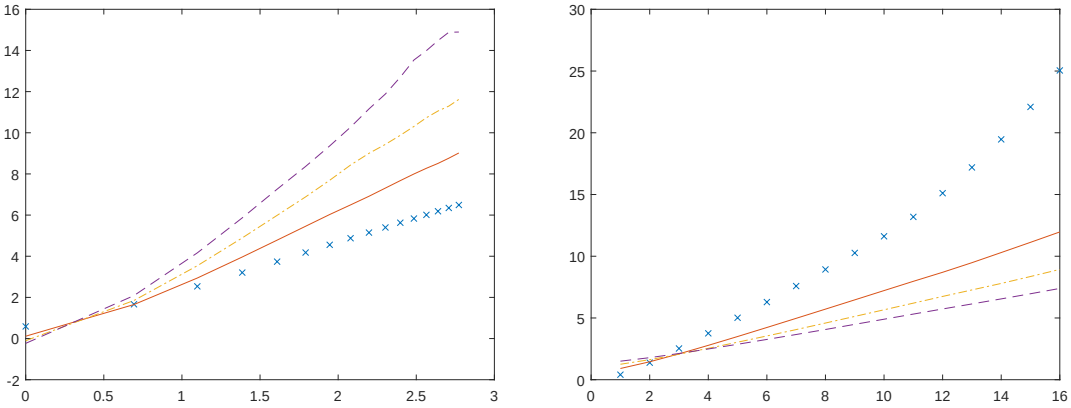


FIGURE 1. Laguerre basis. Left: $\log(m) \mapsto \log(\|\Psi_m^{-1}\|_{\text{op}})$ and density of X given by $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x\geq 0}$. Right: $m \mapsto \log(\|\Psi_m^{-1}\|_{\text{op}})$, and density of X given by $f_k(x) = (k-1)e^{-x/(k-1)}\mathbf{1}_{x\geq 0}$. In both cases: $k = 2$ (blue x marks), $k = 3$ (red solid), $k = 4$ (yellow dashdots) and $k = 5$ (purple dashed).

Proposition 4.3. *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,* $\|\Psi_m^{-1}\|_{\text{op}} \leq Cm^k$.

Discussion.

• The inequality given in Proposition 4.3 seems to give the precise order of $\|\Psi_m^{-1}\|_{\text{op}}$. We illustrate it for the Laguerre basis. Indeed, for the density $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x\geq 0}$, we have computed a Monte-Carlo approximation of $\|\Psi_m^{-1}\|_{\text{op}}$ via 500 samples of size $n = 1000$ and plot in Figure 1 $\log(m) \mapsto \log(\|\Psi_m^{-1}\|_{\text{op}})$ for $m = 1, \ldots, 16$. Then we observe that these curves are linear and with slope approximately equal to k (see Table 1). For the density $(k-1)e^{-x/(k-1)}\mathbf{1}_{x\geq 0}$, from the proof of Proposition 4.3, we conjecture that $m \mapsto \log(\|\Psi_m^{-1}\|_{\text{op}})$ is linear with slope proportional to $1/k$; this is confirmed by Figure 1 and the last two lines of Table 1. Figure 1 also shows that the numerical values of $\|\Psi_m^{-1}\|_{\text{op}}$ and thus of $\|\widehat{\Psi}_m^{-1}\|_{\text{op}}$ are very quickly increasing and thus few elements are considered in $\widehat{\mathcal{M}}_n$. Nevertheless, the selected $\hat{m}$ among these values provides a very satisfactory estimator of b. The procedure is quick and easy. All this is more detailed in the Appendix.

• If f is as in Proposition 4.3, we are in the context of Proposition 3.3, and the result applies: the resulting optimal rate of order $n^{-s/(s+1)}$ can be reached by the adaptive estimator only if $s > 2k$. Note that in a Sobolev-Laguerre ball $W^s(\mathbb{R}^+, R)$ (see definition (15)), the index s 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.

5. Dependent models.

In this section, we extend the previous results to a dependent context, namely an autoregressive model. The general method is the same as in the proof of Theorem 3.1 in Baraud *et al.* (2001b) for autoregressive models, relying on a martingale deviation inequality and a chaining method. The main difficulty here concerns the extension of the deviation inequality stated in Proposition 3.1.

5.1. Mixing deviation inequality.

The deviation inequality of Proposition 3.1 can be extended to the mixing case in the specific case of Laguerre and Hermite bases as follows.

Proposition 5.1. *Assume that $(X_i)_i$ is a strictly stationary and geometrically β-mixing process (i.e. the β-mixing coefficients $(\beta_k)_k$ satisfy $\beta_k \leq ce^{-\theta k}$ for some constants $c > 0, \theta > 0$), with marginal density f. Assume moreover that* **(A1)** *and* **(A2)** *are fulfilled. Let $\widehat{\Psi}_m$ be defined by Equation* (3)*. Then for all $u > 0$*

$$\mathbb{P}\left[\|\Psi_m - \widehat{\Psi}_m\|_{\text{op}} \geq u\right] \leq 4m \exp\left(-\frac{nu^2/2}{(12c_\varphi^2/\theta)m\log(n)(\|f\|_\infty + 2u/3)}\right) + \frac{c}{n^5}.$$

Note that for Laguerre and Hermite bases, Assumption **(A2)** can be replaced by some moment assumptions on X_1, in both dependent and independent results (in Proposition 3.1 and Theorem 3.1 in particular)

5.2. Autoregressive model.

Let us consider the autoregression model:

$$X_{i+1} = b(X_i) + \varepsilon_{i+1}, \quad (\varepsilon_i)_{i\geq 0} \text{ i.i.d., centered with variance } \sigma_\varepsilon^2. \tag{27}$$

We assume that X_0 is independent of the sequence $(\varepsilon_i)_{i\geq 0}$.

Conditions on $b(.)$ and the noise density ensuring that the model (27) admits a strictly stationary and geometrically β-mixing solution are given in *e.g.* Doukhan (1994) (Th. 7 p.102), and recalled in Baraud *et al.* (2001b), section 5.2.

The contrast and the collection of estimators are then defined by

$$\hat{b}_m = \arg\min_{t\in S_m} \bar{\gamma}_n(t), \qquad \text{with } \bar{\gamma}_n(t) = \frac{1}{n}\sum_{i=1}^{n} t^2(X_i) - 2X_{i+1}t(X_i).$$

The elementary computation of Proposition 2.1 can not be generalized here, but the general strategy for selecting m given by (8) can be extended. The sets $\widehat{\mathcal{M}}_n$, $\mathcal{M}_n$ are now

[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)).

given by

$$\widehat{\mathcal{M}}_n = \left\{ m \in \{1,2,\dots,n\},\ \sup_{1\le k\le m} k(\|\widehat{\Psi}_k^{-1}\|_{\mathrm{op}}^2 \vee 1) \le \bar{\mathfrak{c}} \frac{n}{\log^2(n)} \right\}, \tag{28}$$

$$\overline{\mathcal{M}}_n = \left\{ m \in \{1,2,\dots,n\},\ \sup_{1\le k\le m} k\ (\|\Psi_k^{-1}\|_{\mathrm{op}}^2 \vee 1) \le \frac{\bar{\mathfrak{c}}}{4} \frac{n}{\log^2(n)} \right\}, \tag{29}$$

where

$$\bar{\mathfrak{c}} = \frac{\theta}{(12\times 192)c_\varphi^2(\|f\|_\infty + (1/3))}.$$

The additional $\log(n)$ term in the definition of the sets is due to geometric mixing. We set as previously

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

Then we can generalize to the result of Theorem 3.1, thanks to Proposition 5.1.

Theorem 5.1. *Let $(X_i)_{1\le i\le n+1}$ be $n+1$ observations extracted from a strictly stationary and geometrically β-mixing process obtained from model (27), with marginal density f. Assume that* **(A1)** *and* **(A2)** *hold. Assume also that the $(\varepsilon_i)_i$ are i.i.d. centred random variables with $\mathbb{E}(\varepsilon_1^6) < +\infty$). Then, there exists a numerical constant κ_0 such that for $\kappa \ge \kappa_0$, we have*

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

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

Note that in practice, the constant $\bar{\mathfrak{c}}$ is unknown: as mentioned in the independent case (see Remark 3.1), $\|f\|_\infty$ may be replaced by an estimator but the constant θ is difficult to estimate. Thus, for n large enough, it is simpler to take thresholds $n/\log^3(n)$ and $(1/4)(n/\log^3(n))$.

6. 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 always 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

$$Y_i = b(X_i) + \sigma(X_i)\varepsilon_i, \quad \mathrm{Var}(\varepsilon_1) = \mathbb{E}(\varepsilon_1^2) = 1 \tag{30}$$

and the same contrast. Thus the estimator on S_m is still given by (4). Then we can prove that under the assumptions of Proposition 2.1,

$$\begin{aligned} \mathbb{E}\big[\|\hat{b}_m - b_A\|_n^2\big] &\leq \inf_{t\in S_m}\left[\int (b_A - t)^2(x)f(x)dx\right] + \mathbb{E}\big[\max_{1\leq i\leq n}\sigma^2(X_i)\big]\frac{m}{n} \\ &\leq \inf_{t\in S_m}\left[\int (b_A - t)^2(x)f(x)dx\right] + c^2\frac{m}{n}, \end{aligned} \tag{31}$$

if for all x, $\sigma^2(x) \leq c^2$.

Our method can be readily extended to the case where the Y_i are not observed but subject to multiplicative noise. More precisely, suppose that the observations are $(Z_i, X_i)_{1\leq i\leq n}$ with

$$Z_i = Y_iU_i, \quad \mathbb{E}(U_i) = 1, \text{ and } (Y_i, X_i) \text{ following model (1).}$$

Assume also that the U_i are i.i.d, and the sequences $(\varepsilon_i)_{1\leq i\leq n}$, $(X_i)_{1\leq i\leq n}$ and $(U_i)_{1\leq i\leq n}$ are independent. In this case, if the matrix $\widehat{\Psi}_m$ is invertible, we define

$$\hat{b}_m = \arg\min_{t\in S_m}\left[\|t\|_n^2 - \frac{2}{n}\sum_{i=1}^n Z_it(X_i)\right].$$

Note that the model can be written $Z_i = b(X_i) + \eta_i$ with $\eta_i = b(X_i)(U_i - 1) + \varepsilon_iU_i$ and is thus of the same type as (30). Then, we can prove that, for any $b_m \in S_m$,

$$\begin{aligned} \mathbb{E}\big[\|\hat{b}_m - b\|_n^2\big] &\leq \int (b_m - b)^2(x)f(x)dx + 2[\mathrm{Var}(U_1)\mathbb{E}(\max_{1\leq i\leq n} b^2(X_i)) + \sigma_\varepsilon^2\mathbb{E}(U_1^2)]\frac{m}{n}, \\ &\leq \int (b_m - b)^2(x)f(x)dx + \frac{1}{3}[\frac{\|b\|_\infty^2}{2} + \sigma_\varepsilon^2]\frac{m}{n} \end{aligned}$$

if $b(.)$ is bounded.

Note that similar 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. 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.

7. Proofs

7.1. Proofs of Section 2.

7.1.1. *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 $\{(t(X_1),\dots,t(X_n))', t\in S_m\}$ of $\mathbb{R}^n$ and by $\Pi_m b$ the projection of the vector $(b(X_1),\dots,b(X_n))'$. The following equality holds,

$$\|\hat b_m - b_A\|_n^2 = \|\Pi_m b - b_A\|_n^2 + \|\hat b_m - \Pi_m b\|_n^2 = \inf_{t\in S_m} \|t - b_A\|_n^2 + \|\hat b_m - \Pi_m b\|_n^2 \tag{32}$$

By taking expectation, we obtain

$$\mathbb{E}\big[\|\hat b_m - b_A\|_n^2\big] \le \inf_{t\in S_m} \int (t-b_A)^2(x) f(x)dx + \mathbb{E}\left[\|\hat b_m - \Pi_m b\|_n^2\right]. \tag{33}$$

Now we have:

Lemma 7.1. *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 (33), thus we obtain Proposition 2.1. □

7.1.2. *Proof of Lemma 7.1.* Denote by $b(X) = (b(X_1),\dots,b(X_n))'$ and $b_A(X) = (b_A(X_1),\dots,b_A(X_n))'$. We can write

$$\hat b_m(X) = (\hat b_m(X_1),\dots,\hat b_m(X_n))' = \widehat\Phi_m \vec{\hat a}^{(m)},$$

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

$$\|\hat b_m - \Pi_m b\|_n^2 = \|\mathbf{P}(X)\vec\varepsilon\|_n^2 = \frac{1}{n}\vec\varepsilon'\mathbf{P}(X)'\mathbf{P}(X)\vec\varepsilon = \frac{1}{n}\vec\varepsilon'\mathbf{P}(X)\vec\varepsilon \tag{34}$$

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 $\varphi_0(X),\dots,\varphi_{m-1}(X)$, where $\varphi_j(X) = (\varphi_j(X_1),\dots,\varphi_j(X_n))'$. Note that

$$\mathbb{E}(\|\mathbf{P}(X)\vec\varepsilon\|_{2,n}^2) \le \mathbb{E}(\|\vec\varepsilon\|_{2,n}^2) < +\infty.$$

Next, we have to compute, using that $\mathbb{P}(X)$ 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

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

This is the result of Lemma 7.1. □

7.1.3. *Proof of Inequality (31).* Let $\overrightarrow{\sigma\varepsilon}$ denotes the $n \times 1$-vector with coordinates $\sigma(X_i)\varepsilon_i$, $i = 1, \ldots, n$. Equality (34) now writes

$$\|\hat{b}_m - \Pi_m b\|_n^2 = \|\mathbf{P}(X)\overrightarrow{\sigma\varepsilon}\|_n^2 = \frac{1}{n}\|\mathbf{P}(X)\overrightarrow{\sigma\varepsilon}\|_{2,n}^2 = \frac{1}{n}(\overrightarrow{\sigma\varepsilon})'\mathbf{P}(X)(\overrightarrow{\sigma\varepsilon}),$$

as $\mathbf{P}(X)'\mathbf{P}(X) = \mathbf{P}(X)$. Thus, we have to bound

$$\begin{aligned}\mathbb{E}\big[(\overrightarrow{\sigma\varepsilon})'\mathbf{P}(X)(\overrightarrow{\sigma\varepsilon})\big] &= \sum_{i,j}\mathbb{E}\big[\varepsilon_i\varepsilon_j\sigma(X_i)\sigma(X_j)[\mathbf{P}(X)]_{i,j}\big] = \sum_{i=1}^{n}\mathbb{E}\big[\sigma^2(X_i)[\mathbf{P}(X)]_{i,i}\big] \\ &\leq \mathbb{E}\big[\max_{1\leq i\leq n}\sigma^2(X_i)\mathrm{Tr}(\mathbf{P}(X))\big] \leq m\mathbb{E}\big[\max_{1\leq i\leq n}\sigma^2(X_i)\big],\end{aligned}$$

where $\mathbf{P}(X)$ is defined in the proof of Lemma 7.1. Finally, we obtain (31). □

7.2. **Proofs of Section 3.**

7.2.1. *Proof of Proposition 3.1.* To get the announced result, we apply a Bernstein matrix inequality (see Theorem 8.2). Thus we write $\widehat{\Psi}_m$ as a sum of a sequence of independent matrices

$$\widehat{\Psi}_m = \frac{1}{n}\sum_{i=1}^{n}\mathbf{K}_m(X_i), \quad \mathbf{K}_m(X_i) = (\varphi_j(X_i)\varphi_k(X_i))_{0\leq j,k\leq m-1}.$$

We put

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

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

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

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.

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

So we get that, a.s.

$$\|\mathbf{K}_m(X_1) - \mathbb{E}\left[\mathbf{K}_m(X_1)\right]\|_{\mathrm{op}}/n \leq \frac{2c_\varphi^2 m}{n}. \tag{36}$$

• Bound on $\nu(\mathbf{S}_m) = \| \sum_{i=1}^n \mathbb{E}\left[(\mathbf{K}_m(X_i) - \mathbb{E}\left[\mathbf{K}_m(X_i)\right])' (\mathbf{K}_m(X_i) - \mathbb{E}\left[\mathbf{K}_m(X_i)\right])\right] \|_{\text{op}}/n^2$. By definition of the operator norm we have

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

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

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

Therefore as f is bounded,

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

Then we get that $\nu(\mathbf{S}_m) \leq \dfrac{c_\varphi^2 \|f\|_\infty m}{n}$. Applying Theorem 8.2 gives the result of Proposition 3.1. □

7.2.2. *Proof of Inequality (11) of Theorem 3.1.* We denote by $\widehat{M}_n$ the maximal element of $\widehat{\mathcal{M}}_n$ defined by (9) and by M_n the maximal element of $\mathcal{M}_n$ defined by (10).

Let us also define also

$$\mathcal{M}_n^+ = \left\{ m \in \mathbb{N}, \sup_{1\leq k\leq m} k\, (\|\Psi_k^{-1}\|_{\text{op}}^2 \vee 1) \leq 4\mathfrak{c}\frac{n}{\log(n)} \right\}, \tag{37}$$

where $\mathfrak{c}$ is still defined in (9). We denote by M_n^+ 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

$$\Xi_n := \left\{ \mathcal{M}_n \subset \widehat{\mathcal{M}}_n \subset \mathcal{M}_n^+ \right\}, \tag{38}$$

and we write the decomposition:

$$\widehat{b}_{\hat{m}} - b_A = \underbrace{(\widehat{b}_{\hat{m}} - b_A)\mathbf{1}_{\Xi_n}}_{:=T_1} + \underbrace{(\widehat{b}_{\hat{m}} - b_A)\mathbf{1}_{\Xi_n^c}}_{:=T_2}. \tag{39}$$

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

Lemma 7.2. *Under the assumptions of Theorem 3.1, there exists κ_0 such that for $\kappa \geq \kappa_0$, we have*

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

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

Lemma 7.3. *We have, for c a positive constant,*

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

Indeed, Lemma 7.2 gives the bound on T_1. For T_2, we use Lemma 7.3 as follows.

Recall that Π_m denotes the orthogonal projection (for the scalar product of $\mathbb{R}^n$) on the sub-space $\big\{\big(t(X_1), \ldots, t(X_n)\big)' t \in S_m\big\}$ 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

$$\|b - \hat{b}_{\hat{m}}\|_n^2 = \|b - \Pi_{\hat{m}} b\|_n^2 + \|\Pi_{\hat{m}} \varepsilon\|_n^2 \leq \|b\|_n^2 + n^{-1} \sum_{k=1}^n \varepsilon_k^2. \tag{40}$$

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 7.2 plugged in decomposition (39), ends the proof of Inequality (11) of Theorem 3.1. □

7.2.3. *Proof of Lemma 7.2.* To begin with, we note that $\gamma_n(\hat{a}_1, \ldots, \hat{a}_m) = -\|\hat{b}_m\|_n^2$. Indeed, using formula (4) and $\widehat{\Phi}_m' \widehat{\Phi}_m = n \widehat{\Psi}_m$, we have

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

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 $b_m \in S_m$,

$$\gamma_n(\hat{b}_{\hat{m}}) + \mathrm{pen}(\hat{m}) \leq \gamma_n(b_m) + \mathrm{pen}(m). \tag{41}$$

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

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

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

$$\Omega_n = \left\{ \left| \frac{\|t\|_n^2}{\|t\|_f^2} - 1 \right| \leq \frac{1}{2},\ \forall t \in \bigcup_{m,m' \in \mathcal{M}_n^+} (S_m + S_{m'}) \setminus \{0\} \right\}. \tag{42}$$

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 $\Xi_n \cap \Omega_n$,

$$\begin{aligned}
\|\hat{b}_{\hat{m}} - b\|_n^2 \leq & \|b_m - b\|_n^2 + \frac{1}{8}\|\hat{b}_{\hat{m}} - b_m\|_f^2 + (8 \sup_{t \in B^f_{\hat{m},m}(0,1)} \nu_n^2(t) + \text{pen}(m) - \text{pen}(\hat{m})) \\
\leq & \Big(1 + \frac{1}{2}\Big)\|b_m - b\|_n^2 + \frac{1}{2}\|\hat{b}_{\hat{m}} - b\|_n^2 + 8\Big(\sup_{t \in B^f_{\hat{m},m}(0,1)} \nu_n^2(t) - p(m,\hat{m})\Big)_+ \\
& + \text{pen}(m) + 8p(m,\hat{m}) - \text{pen}(\hat{m}).
\end{aligned} \tag{43}$$

Here we state the following Lemma:

Lemma 7.4. *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 $8p(m,\hat{m}) - \text{pen}(\hat{m}) \leq \text{pen}(m)$. Thus, by taking expectation in (43) and applying Lemma 7.4, it comes that, for all m in $\mathcal{M}_n$ and b_m in S_m,

$$\mathbb{E}\big[\|\hat{b}_{\hat{m}} - b_A\|_n^2 \mathbf{1}_{\Xi_n \cap \Omega_n}\big] \leq 3\mathbb{E}\big[\|b_m - b_A\|_n^2\big] + 2\text{pen}(m) + \frac{16\,C}{n}. \tag{44}$$

The complement of Ω_n satisfies the following Lemma:

Lemma 7.5. *Assume that* **(A1)-(A2)** *hold. Then, for all* $m \in \mathcal{M}_n$ *(see (10)) and* Ω_n *defined by (42),* $\mathbb{P}(\Omega_n^c) \leq c/n^4$ *where* c *is a positive constant.*

We conclude as above (see equation (40)) 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 (44) ends the proof of Lemma 7.2. □

Proof of Lemma 7.4. We can not apply Talagrand's Inequality to the process ν_n itself, unless we add an assumption imposing that the noise is bounded. This is why we decompose the variables ε_i as follows:

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

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

$$\begin{aligned}\Big(\sup_{t\in B^f_{\hat m,m}(0,1)} \nu_n^2(t) - p(m,\hat m)\Big)_+ \ &\le\ \Big(\sup_{t\in B^f_{\hat m,m}(0,1)} 2\nu_{n,1}^2(t) - p(m,\hat m)\Big)_+ \\ &\quad +2 \sup_{t\in B^f_{\hat m,m}(0,1)} \nu_{n,2}^2(t). \end{aligned} \tag{45}$$

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:

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

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

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

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

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

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,

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

Under condition (37) on $\mathcal{M}_n^+$, we have

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

We can take

$$M_1 := c_\varphi k_n \left(4\mathfrak{c}\frac{n}{\log(n)}\right)^{1/4} (m\vee m')^{1/4}. \tag{46}$$

Consequently, Talagrand Inequality (see Theorem 8.3) implies, for $p(m,m') = 8\frac{\sigma_\varepsilon^2 \max(m,m')}{n}$, and denoting by $m^* := \max(m,m')$,

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

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

$$\begin{aligned}\mathbb{E}\left[\Big(\sup_{t\in B^f_{m,m}(0,1)} [\nu_{n,1}]^2(t) - p(m,\hat m)\Big)_+ \mathbf{1}_{\Xi_n}\right] &\leq \sum_{m'\in\mathcal{M}_n^+} \mathbb{E}\Big(\sup_{t\in B^f_{m',m}(0,1)} [\nu_{n,1}]^2(t) - p(m,m')\Big)_+ \\ &\leq \frac{C}{n}. \qquad (47)\end{aligned}$$

Let us now study the second term in (45). Recall that $M_n^+ \leq 4\mathfrak{c}n/\log(n)$ the dimension of the largest space of the collection. Then we have

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

where the last line follows from the Markov inequality and the choices $k_n = n^{1/4}$ and $p = 4$. This bound together with (47) plugged in (45) gives the result of Lemma 7.4. □

Proof of Lemma 7.5. As the collection of models is nested, we have

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

Then

$$\mathbb{P}\left(\exists t\in S_m, \left|\frac{\|t\|_n^2}{\|t\|_f^2} - 1\right| > \frac{1}{2}\right) = \mathbb{P}\left(\sup_{t\in S_m,\|t\|_f=1}\left|\frac{1}{n}\sum_{i=1}^n[t^2(X_i) - \mathbb{E}t^2(X_i)]\right| > \frac{1}{2}\right).$$

Moreover we have

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

As a consequence,

$$\begin{aligned}\mathbb{P}\left(\exists t\in S_m,\left|\frac{\|t\|_n^2}{\|t\|_f^2}-1\right|>\frac{1}{2}\right) &\le \mathbb{P}\left(\|\Psi_m^{-1}\|_{\rm op}\|\widehat{\Psi}_m-\Psi_m\|_{\rm op}>\frac{1}{2}\right)\\ &= \mathbb{P}\left(\|\widehat{\Psi}_m-\Psi_m\|_{\rm op}>\frac{1}{2\|\Psi_m^{-1}\|_{\rm op}}\right).\end{aligned}\tag{48}$$

We apply Proposition 3.1 and we get

$$\mathbb{P}\left(\exists t\in S_m,\left|\frac{\|t\|_n^2}{\|t\|_f^2}-1\right|>\frac{1}{2}\right)\le 2m\exp\left(-\frac{1}{8c_\varphi^2}\frac{n}{m\|\Psi_m^{-1}\|_{\rm op}}\frac{1}{\|f\|_\infty\|\Psi_m^{-1}\|_{\rm op}+\frac{1}{3}}\right).$$

By the definition of $\mathcal{M}_n^+$ in (37), $m(\|\Psi_m^{-1}\|_{\rm op}\vee 1)\le m(\|\Psi_m^{-1}\|_{\rm op}^2\vee 1)\le 4\mathfrak{c}n/\log n$. Therefore,

$$\mathbb{P}\left(\exists t\in S_m,\left|\frac{\|t\|_n^2}{\|t\|_f^2}-1\right|>\frac{1}{2}\right)\le 2m\exp\left(-\frac{\log n}{8c_\varphi^2\times 4\mathfrak{c}(\|f\|_\infty+\frac{1}{3})}\right)=2\frac{m}{n^6}$$

with the choice of $\mathfrak{c}$ given in (9). Summing up the terms over $\mathcal{M}_n^+$, with the bound on the cardinality implied by (37), gives the result of Lemma 7.5. □.

7.2.4. *Proof of Lemma 7.3.* 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 $k\|\Psi_k^{-1}\|_{\rm op}^2\le\mathfrak{c}n/4\log(n)$ and $k\|\widehat{\Psi}_k^{-1}\|_{\rm op}^2>\mathfrak{c}n/\log(n)$. This implies, as

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

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

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

Now we have

$$\mathbb{P}(\mathcal{M}_n\not\subseteq\widehat{\mathcal{M}}_n)\le\sum_{m\in\mathcal{M}_n}\mathbb{P}(\Delta_m)\le\sum_{m\in\mathcal{M}_n}\mathbb{P}(\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\rm op}^2>\|\Psi_m^{-1}\|_{\rm op}^2).$$

Now, we write the decomposition

$$\begin{aligned}
&\Big\{\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\mathrm{op}}>\|\Psi_m^{-1}\|_{\mathrm{op}}\Big\}\\
&=\Big\{\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\mathrm{op}}>\|\Psi_m^{-1}\|_{\mathrm{op}}\Big\}\bigcap\Big\{\|\Psi_m^{-1}(\widehat{\Psi}_m-\Psi_m)\|_{\mathrm{op}}<\frac{1}{2}\Big\}\\
&\quad\bigcup\Big\{\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\mathrm{op}}>\|\Psi_m^{-1}\|_{\mathrm{op}}\Big\}\bigcap\Big\{\|\Psi_m^{-1}(\widehat{\Psi}_m-\Psi_m)\|_{\mathrm{op}}\geq\frac{1}{2}\Big\}\\
&\subset\Big\{\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\mathrm{op}}>\|\Psi_m^{-1}\|_{\mathrm{op}}\Big\}\bigcap\Big\{\|\Psi_m^{-1}(\widehat{\Psi}_m-\Psi_m)\|_{\mathrm{op}}<\frac{1}{2}\Big\}\\
&\quad\bigcup\Big\{\|\Psi_m^{-1}(\widehat{\Psi}_m-\Psi_m)\|_{\mathrm{op}}\geq\frac{1}{2}\Big\}. \qquad (49)
\end{aligned}$$

To control the second term of the right hand side of (49), we write

$$\Big\{\|\Psi_m^{-1}(\widehat{\Psi}_m-\Psi_m)\|_{\mathrm{op}}\geq\frac{1}{2}\Big\}\subset\Big\{\|\Psi_m^{-1}\|_{\mathrm{op}}\|\widehat{\Psi}_m-\Psi_m\|_{\mathrm{op}}\geq\frac{1}{2}\Big\} \qquad (50)$$

and we recognize the set in (48) for which we already bounded the probability.
Next to control the first term on the right hand side of (49), we apply Theorem 8.1 (with $\mathbf{A}=\Psi_m$ and $\mathbf{B}=\widehat{\Psi}_m-\Psi_m$), which yields

$$\begin{aligned}
&\Big\{\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\mathrm{op}}>\|\Psi_m^{-1}\|_{\mathrm{op}}\Big\}\cap\Big\{\|\Psi_m^{-1}(\widehat{\Psi}_m-\Psi_m)\|_{\mathrm{op}}<\frac{1}{2}\Big\}\\
&=\left\{\frac{\|\widehat{\Psi}_m-\Psi_m\|_{\mathrm{op}}\|\Psi_m^{-1}\|_{\mathrm{op}}^2}{1-\|\Psi_m^{-1}(\widehat{\Psi}_m-\Psi_m)\|_{\mathrm{op}}}>\|\Psi_m^{-1}\|_{\mathrm{op}}\right\}\cap\Big\{\|\Psi_m^{-1}(\widehat{\Psi}_m-\Psi_m)\|_{\mathrm{op}}<\frac{1}{2}\Big\}\\
&\subset\Big\{\|\widehat{\Psi}_m-\Psi_m\|_{\mathrm{op}}>\frac{1}{2}\|\Psi_m^{-1}\|_{\mathrm{op}}^{-1}\Big\}, \qquad (51)
\end{aligned}$$

which corresponds to (48) again and thus the probability is bounded by a term of order $1/n^4$. So starting from (49) and gathering (50) and (51) gives

$$\mathbb{P}(\mathcal{M}_n\not\subset\widehat{\mathcal{M}}_n)\leq\sum_{m\in\mathcal{M}_n}\mathbb{P}(\Delta_m)\leq\frac{c}{n^2},$$

by applying Proposition 3.1 as previously.

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

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

therefore such that

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

Thus we have

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

We apply the same decomposition as above, interchanging $\widehat{\Psi}_k$ and Ψ_k. We get

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

by applying Proposition 3.1. □

7.2.5. *Proof of Inequality (12) of Theorem 3.1.* 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 (42). Therefore, using the result of Theorem 3.1 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 $t \in S_m$,

$$\mathbb{E}(\|\hat{b}_{\hat{m}}-b_A\|_f^2) \leq C_1\left(\|t-b_A\|_f^2 + \sigma_\varepsilon^2\frac{m}{n}\right) + \frac{C_2}{n} + \mathbb{E}\left(\|\hat{b}_{\hat{m}}-b_A\|_f^2\mathbf{1}_{\Omega_n^c}\right), \tag{52}$$

so only the last term is to be studied. First, recall that Lemma 7.5 implies that $\mathbb{P}(\Omega_n^c) \leq \mathfrak{c}/n^4$. Next, write that $\|\hat{b}_{\hat{m}}-b_A\|_f^2 \leq 2(\|\hat{b}_{\hat{m}}\|_f^2 + \|b_A\|_f^2)$ and

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

First, under $\|\sum_{j=0}^m \varphi_j^2\|_\infty \leq c_\varphi^2 m$, we get

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

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

$$\|\widehat{\Psi}^{-1}_{\hat{m}}\widehat{\Phi}'_{\hat{m}}\|^2_{\rm op} \;=\; \lambda_{\max}\left(\widehat{\Psi}^{-1}_{\hat{m}}\widehat{\Phi}'_{\hat{m}}\widehat{\Phi}_{\hat{m}}\widehat{\Psi}^{-1}_{\hat{m}}\right) = n\lambda_{\max}(\widehat{\Psi}^{-1}_{\hat{m}}) = n\|\widehat{\Psi}^{-1}_{\hat{m}}\|_{\rm op}$$

Therefore, for $\hat{m} \in \widehat{\mathcal{M}}_n$, we get, as $\forall m,\ m(\|\widehat{\Psi}^{-1}_m\|_{\rm op} \vee 1) \leq m(\|\widehat{\Psi}^{-1}_m\|^2_{\rm op} \vee 1)$, that

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

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

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

On the other hand $\mathbb{E}(\|b_A\|^2_f \mathbf{1}_{\Omega_n^c}) \leq \|b_A\|^2_f\mathbb{P}(\Omega_n^c) \leq c"/n^4$. Thus $\mathbb{E}\left(\|\hat{b}_{\hat{m}} - b_A\|^2_f \mathbf{1}_{\Omega_n^c}\right) \leq c_1/n$ and plugging this in (52) ends the proof of Inequality (12) in Theorem 3.1. □

7.2.6. *Proof of Theorem 3.2.* We use the strategy of proof presented in Theorem 2.11 of Tsybakov (2009). To that aim, 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 $b_0, b_\theta \in W^s_f(A, R)$.

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

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

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

Thus for $\ell \leq m$,

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

Choosing δ small enough, we get the result.

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

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

As above, we find

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

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

• 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 Ω_n (see (42)), 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 $m \in \mathcal{M}_n$, it follows from Lemma 7.5 that $\mathbb{P}(\Omega_n) \geq 1 - (c/n^4) \geq 1/2$. Therefore the lower bound is proved. □

7.3. **Proofs of Section 4.2.** 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

$$\ell_k^{(\delta)}(x) = 2^{(\delta+1)/2}\left(\frac{k!}{\Gamma(k+\delta+1)}\right)^{1/2} L_k^{(\delta)}(2x) e^{-x} x^{\delta/2}, \tag{53}$$

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 ℓ_k, depending on k: for $\nu = 4k + 2\delta + 2$, and k large enough, it holds

$$|\ell_k^{(\delta)}(x/2)| \leq C \begin{cases} a) & (x\nu)^{\delta/2} & \text{if } 0 \leq x \leq 1/\nu \\ b) & (x\nu)^{-1/4} & \text{if } 1/\nu \leq x \leq \nu/2 \\ c) & \nu^{-1/4}(\nu - x)^{-1/4} & \text{if } \nu/2 \leq x \leq \nu - \nu^{1/3} \\ d) & \nu^{-1/3} & \text{if } \nu - \nu^{1/3} \leq x \leq \nu + \nu^{1/3} \\ e) & \nu^{-1/4}(x - \nu)^{-1/4} e^{-\gamma_1 \nu^{-1/2}(x-\nu)^{3/2}} & \text{if } \nu + \nu^{1/3} \leq x \leq 3\nu/2 \\ f) & e^{-\gamma_2 x} & \text{if } x \geq 3\nu/2 \end{cases}$$

where γ_1 and γ_2 are positive and fixed constants.

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 7.6. *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.

7.3.1. *Proof of Proposition 4.2.* The invertibility of Ψ_m follows from Lemma 2.1 under (21). Now we prove (26). First note that, for j large enough,

$$\int \varphi_j^2(x) f(x) dx \leq \frac{c_1}{\sqrt{j}}, \tag{54}$$

where c_1 is a constant. The proof of Inequality (54) 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 (18), 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. □

7.3.2. *Proof of Proposition 4.3.* We treat the Laguerre basis first. The result of Askey and Wainger (1965) recalled above states that, for j large enough, $\ell_j(x) \leq ce^{-\gamma_2 x}$ for

$2x \geq 3(2j+1)$, where γ_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,

$$\int_{3(m+1/2)}^{+\infty} \left(\sum_{j=0}^{m-1} x_j \ell_j(u)\right)^2 du \leq C'me^{-\gamma_3 m} \leq \frac{1}{2}.$$

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 $|h_j(x)| \leq c|x|e^{-\gamma_2 x^2}$ for $x^2 \geq (3/2)(4j+3)$. □

7.4. **Proofs of Section 5.**

7.4.1. *Proof of Proposition 5.1.* Consider the coupling method and the associated variables (X_i^*) with Berbee's Lemma, see Berbee (1979), with the method described in Viennet (1997, Prop.5.1 and its proof p.484) .

Assume for simplicity that $n = 2p_nq_n$ for integers p_n, q_n. Then there exist random variables X_i^*, $i = 1, ..., n$ satisfying the following properties:

- For $\ell = 0, ..., p_n - 1$, the random vectors

$$\vec{X}_{\ell,1} = \left(X_{2\ell q_n+1}, ..., X_{(2\ell+1)q_n}\right)' \text{ and } \vec{X}_{\ell,1}^* = \left(X_{2\ell q_n+1}^*, ..., X_{(2\ell+1)q_n}^*\right)'$$

  have the same distribution, and so have the random vectors

$$\vec{X}_{\ell,2} = \left(X_{(2\ell+1)q_n+1}, ..., X_{(2\ell+2)q_n}\right)' \text{and } \vec{X}_{\ell,2}^* = \left(X_{(2\ell+1)q_n+1}^*, ..., X_{(2\ell+2)q_n}^*\right)'.$$

- For $\ell = 0, ..., p_n - 1$,

$$\mathbb{P}\left[\vec{X}_{\ell,1} \neq \vec{X}_{\ell,1}^*\right] \leq \beta_{q_n} \text{ and } \mathbb{P}\left[\vec{X}_{\ell,2} \neq \vec{X}_{\ell,2}^*\right] \leq \beta_{q_n}. \tag{55}$$

- For each $\delta \in \{1, 2\}$, the random vectors $\vec{X}_{0,\delta}^*, ..., \vec{X}_{p_n-1,\delta}^*$ are independent.

Then let $\Omega^*=\{X_i=X_i^*, i=1,\dots,n\}$ and write that

$$\mathbb{P}\left[\|\Psi_m-\widehat{\Psi}_m\|_{\rm op}\geq u\right]\leq \mathbb{P}\left[\{\|\Psi_m-\widehat{\Psi}_m\|_{\rm op}\geq u\}\cap\Omega^*\right]+\mathbb{P}[(\Omega^*)^c].$$

Then using the definition of the variables X_i^*, we get

$$\mathbb{P}[(\Omega^*)^c]\leq 2p_n\beta_{q_n}\leq cne^{-\theta q_n}.$$

Then choosing $q_n=6\log(n)/\theta$ yields $\mathbb{P}[(\Omega^*)^c]\leq c/n^5$.

Now, we have to apply Tropp's result. To that aim, we write $\mathbf{S}_m=(1/2)(\mathbf{S}_{m,1}+\mathbf{S}_{m,2})$ where $\mathbf{S}_m$ is given by (35), $\mathbf{S}_{m,1}$ is built with the $\vec{X}_{\ell,1}$:

$$\mathbf{S}_{m,1}=\frac{1}{p_n}\sum_{\ell=0}^{p_n-1}\frac{1}{q_n}\sum_{r=1}^{q_n}\mathbf{K}_m(X_{2\ell q_n+r})-\mathbb{E}(\mathbf{K}_m(X_{2\ell q_n+r}))$$

and $\mathbf{S}_{m,2}$ is analogously defined with with the $\vec{X}_{\ell,2}$. We have

$$\begin{aligned}\mathbb{P}\left[\{\|\Psi_m-\widehat{\Psi}_m\|_{\rm op}\geq u\}\cap\Omega^*\right] &= \mathbb{P}\left[\{\|\mathbf{S}_{m,1}+\mathbf{S}_{m,2}\|_{\rm op}\geq 2u\}\cap\Omega^*\right]\\ &\leq \mathbb{P}\left[\{\|\mathbf{S}_{m,1}\|_{\rm op}\geq u\}\cap\Omega^*\right]+\mathbb{P}\left[\{\|\mathbf{S}_{m,2}\|_{\rm op}\geq u\}\cap\Omega^*\right]\\ &\leq \mathbb{P}\left[\|\mathbf{S}_{m,1}^*\|_{\rm op}\geq u\right]+\mathbb{P}\left[\|\mathbf{S}_{m,2}^*\|_{\rm op}\geq u\right],\end{aligned}$$

where $\mathbf{S}_{m,\delta}^*$, for $\delta=1,2$ are built on the $\vec{X}_{\ell,\delta}^*$. The two terms are similar, and we treat only the first one.

We can apply Tropp's result as $\mathbf{S}_{m,1}^*$ is a sum of p_n independent matrices. It follows from (36) that

$$\frac{1}{p_nq_n}\|\sum_{r=1}^{q_n}\mathbf{K}_m(X_{2\ell q_n+r}^*)-\mathbb{E}(\mathbf{K}_m(X_{2\ell q_n+r}^*))\|_{\rm op}\leq 2c_\varphi^2\frac{m}{p_n}=\frac{24}{\theta}c_\varphi^2\frac{m\log(n)}{n}.$$

Next, we must bound the variance of $\mathbf{S}_{m,1}^*$. We have

$$\nu(\mathbf{S}_{m,1}^*) = \frac{1}{p_n}\sup_{\|\vec{x}\|_{2,m}=1}\mathbb{E}\left(\frac{1}{q_n^2}\left\|\left[\sum_{r=1}^{q_n}(\mathbf{K}_m(X_r^*)-\mathbb{E}(\mathbf{K}_m(X_r^*)))\right]\vec{x}\right\|_{2,m}^2\right)$$

Next,

$$\begin{aligned}\mathbb{E}_1 &= \mathbb{E}\left(\frac{1}{q_n^2}\left\|\left[\sum_{r=1}^{q_n}(\mathbf{K}_m(X_r^*)-\mathbb{E}(\mathbf{K}_m(X_r^*)))\right]\vec{x}\right\|_{2,m}^2\right)=\frac{1}{q_n^2}\sum_{j=0}^{m-1}{\rm Var}\left[\sum_{r=1}^{q_n}\sum_{k=0}^{m-1}\varphi_j(X_r)\varphi_k(X_r)x_k\right]\\ &\leq \frac{1}{q_n}\sum_{j=0}^{m-1}\sum_{r=1}^{q_n}{\rm Var}\left[\sum_{k=0}^{m-1}\varphi_j(X_r)\varphi_k(X_r)x_k\right]=\sum_{j=0}^{m-1}{\rm Var}\left[\sum_{k=0}^{m-1}\varphi_j(X_1)\varphi_k(X_1)x_k\right]\\ &\leq \sum_{j=0}^{m-1}\mathbb{E}\left[\left(\sum_{k=0}^{m-1}\varphi_j(X_1)\varphi_k(X_1)x_k\right)^2\right]\leq c_\varphi^2m\|f\|_\infty\int\left(\sum_{k=0}^{m-1}\varphi_k(u)x_k\right)^2du=c_\varphi^2m\|f\|_\infty.\end{aligned}$$

This implies that

$$\nu(\mathbf{S}_m)\leq\frac{c_\varphi^2m\|f\|_\infty}{p_n}=\frac{12c_\varphi^2\|f\|_\infty}{\theta}\frac{m\log(n)}{n}.$$

Then, applying Theorem 8.2 gives the announced result. □

7.4.2. *Proof of Theorem 5.1.* We follow the line of the proof of Theorem 3.1, and we have to extend Lemma 7.4 and Lemma 7.5 to the dependent case. We also define

$$\overline{\mathcal{M}}_n^+ = \left\{ m \in \{1,2,\dots,n\},\ \sup_{1\le k\le m} k\left(\|\Psi_k^{-1}\|_{\rm op}^2 \vee 1\right) \le 4\bar{\mathfrak{c}}\frac{n}{\log^2(n)} \right\}. \tag{56}$$

For Lemma 7.5, the extension is the following.

Lemma 7.7. *Assume that $(X_i)_{i\ge1}$ is strictly stationary geometrically β-mixing, with common density f and that* **(A1)** *and* **(A2)** *hold. Then, for $\overline{\Omega}_n$ defined like Ω_n in (42), with $\mathcal{M}_n^+$ replaced by $\overline{\mathcal{M}}_n^+$ (see (56)), we have $\mathbb{P}(\overline{\Omega}_n^c) \le c/n^4$ where c is a positive constant.*

Proof of Lemma 7.7. We start from (48) and apply Proposition 5.1. We get

$$\mathbb{P}\left(\exists t\in S_m, \left|\frac{\|t\|_n^2}{\|t\|_f^2}-1\right|>\frac{1}{2}\right) \le 4m\exp\left(-\frac{1}{96c_\varphi^2}\frac{n\theta}{m\log(n)\|\Psi_m^{-1}\|_{\rm op}(\|f\|_\infty\|\Psi_m^{-1}\|_{\rm op}+\frac{1}{3})}\right)+\frac{c}{n^5}.$$

Using the definition of $\overline{\mathcal{M}}_n^+$ (and specifically of $\bar{\mathfrak{c}}$), we obtain the result. □

Now we can extend Lemma 7.4 as follows.

Lemma 7.8. *Let $(X_i, i=1,\dots,n)$ be observations from model (27) with $\mathbb{E}(\varepsilon_1^6)<+\infty$. Assume that $(X_i)_{i\ge1}$ is strictly stationary geometrically β-mixing, with common density f and that* **(A1)** *and* **(A2)** *hold. Then $\bar\nu_n(t) = n^{-1}\sum_{i=1}^n \varepsilon_{i+1}t(X_i)$ satisfies*

$$\mathbb{E}\Big(\sup_{t\in B^f_{\hat m,m}(0,1)} \bar\nu_n^2(t) - \overline{p}(m,\hat m)\Big)_+ \le \frac{C}{n}$$

where $\overline{p}(m,m') = c\sigma_\varepsilon^2\max(m,m')/n$ for c a numerical constant.

Proof of Lemma 7.8. We start with the same decomposition as in the proof of Lemma 7.4 and split $\bar\nu_n(t)$ into the sum $\overline{\nu}_{n,1}+\overline{\nu}_{n,2}$ as previously. The treatment of $\overline{\nu}_{n,2}$ is identical as it relies on a non-correlation property which is still true. We obtain the same bound with $k_n = (n/\log^2(n))^{1/4}$ and the maximal dimension $N_n \le n/\log^2(n)$.

For $\overline{\nu}_{n,1}$ we proceed by the coupling strategy used in the proof of Proposition 5.1, applied to $u_i = (\varepsilon_{i+1}, X_i)$ which is also a β-mixing sequence with mixing coefficient such that $\beta_k \le ce^{-\theta k}$, as in Baraud *et al.* (2001a). We denote by $\Omega^\star = \{u_i = u_i^\star,\ i=1,\dots,n\}$. We still have $\mathbb{P}((\Omega^\star)^c) \le p_n\beta_{q_n} \le c/n^4$ for $q_n = 5\log(n)/\theta$.

On $\Omega^\star$, we replace the u_i by the $u_i^\star$ and split the term between odd and even blocks. We have to bound, say

$$\mathbb{E}\left[\Big(\sup_{t\in B^f_{\hat m,m}(0,1)} (\bar\nu_{n,1}^{\star,1})^2(t) - \overline{p}(m,\hat m)\Big)_+ \mathbf{1}_{\overline{\Omega}_n\cap\overline{\Xi}_n}\right],$$

where $\overline{\Xi}_n$ is analogous to Ξ_n defined by (38), by using Talagrand inequality applied to mean of p_n independent random variables

$$\bar\nu_{n,1}^{\star,1}(t) = \frac{1}{p_n}\sum_{\ell=0}^{p_n-1}\left(\frac{1}{q_n}\sum_{r=1}^{q_n}\eta^\star_{2\ell q_n+r}t(X^\star_{2\ell q_n+r})\right).$$

Clearly,

$$\mathbb{E}\Big(\sup_{t\in B^f_{m,m'}(0,1)}(\bar\nu^{\star,1}_{n,1})^2(t)\Big)\le\sigma_\varepsilon^2\frac{\max(m,m')}{n}:=H_\star^2$$

still holds. We have

$$\begin{aligned}\sup_{t\in B^f_{m,m'}(0,1)}\mathrm{Var}\left(\frac{1}{q_n}\sum_{r=1}^{q_n}\eta_r^* t(X_r^*)\right) &= \sup_{t\in B^f_{m,m'}(0,1)}\mathrm{Var}\left(\frac{1}{q_n}\sum_{r=1}^{q_n}\eta_r t(X_r)\right)\\ &= \frac{1}{q_n}\sup_{t\in B^f_{m,m'}(0,1)}\mathbb{E}(\eta_1^2)\mathbb{E}(t^2(X_1))\le\frac{\mathbb{E}(\varepsilon_1^2)}{q_n}:=v_\star.\end{aligned}$$

Lastly

$$\sup_{t\in B^f_{m',m}(0,1)}\sup_{\vec u,\vec x\in\mathbb{R}^{q_n}}\left(\frac{1}{q_n}\sum_{r=1}^{q_n}|u_r|\mathbf{1}_{|u_r|\le k_n}|t(x_r)|\right)\le ck_n\left(\frac{n}{\log^2(n)}\right)^{1/4}(m\vee m')^{1/4}:=M_1^\star,$$

where $M_1^\star$ is computed analogously to M_1 given by (46), except that $m\in\overline{\mathcal{M}}_n$ increases the power of the log. Therefore, by applying Theorem 8.3, we obtain

$$\mathbb{E}\Big(\sup_{t\in B^f_{m,m'}(0,1)}(\bar\nu^{\star,1}_{n,1})^2(t)-2H_\star^2\Big)_+\le C_1\left(\frac{1}{n}e^{-C_2(m\vee m')}+\frac{1}{n}(m\vee m')^{1/2}\exp(-C_3(m\vee m')^{1/4})\right),$$

and

$$\mathbb{E}\left[\Big(\sup_{t\in B^f_{\hat m,m}(0,1)}(\bar\nu^{\star,1}_{n,1})^2(t)-\overline{p}(m,\hat m)\Big)_+\mathbf{1}_{\overline{\Omega}_n\cap\overline{\Xi}_n}\right]\le c/n.$$

It remains to bound

$$\mathbb{E}\left[\Big(\sup_{t\in B^f_{\hat m,m}(0,1)}(\bar\nu_{n,1})^2(t)-\overline{p}(m,\hat m)\Big)\mathbf{1}_{(\Omega^\star\cap\overline{\Omega}_n\cap\overline{\Xi}_n)^c}\right]_+.$$

We use the infinite norm computed to evaluate $M_1^\star$ together with the bounds on $\mathbb{P}[(\Omega^\star)^c],\mathbb{P}(\overline{\Omega}_n^c),\mathbb{P}(\overline{\Xi}_n^c)$ to obtain the result. □

8. Theoretical tools

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

Theorem 8.1. *Let* $\mathbf{A}$, $\mathbf{B}$ *be* $(m\times m)$ *matrices. If* $\mathbf{A}$ *is invertible and* $\|\mathbf{A}^{-1}\mathbf{B}\|_{\mathrm{op}}<1$, *then* $\tilde{\mathbf{A}}:=\mathbf{A}+\mathbf{B}$ *is invertible and it holds*

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

Theorem 8.2 (Bernstein Matrix inequality)**.** *Consider a finite sequence* $\{\mathbf{S}_k\}$ *of independent, random matrices with common dimension* $d_1\times d_2$. *Assume that*

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

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 8.3. *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.

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

[Baraud et al., 2001b] Baraud, Y., Comte, F. and Viennet, G. (2001b) Model selection for (auto-)regression with dependent data. *ESAIM Probab. Statist.* **5**, 33-49.

[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 To appear in *The Annals of the Institute of Mathematical Statist.*.

[Berbee, 1979] Berbee, H.C.P. (1979). Random walks with stationary increments and renewal theory. *Math. Tracts. Mathematisch Centrum*, Amsterdam, **112**.

[Bongioanni and Torrea, 2006] Bongioanni, B. and Torrea, J.L. (2006). Sobolev spaces associated to the harmonic oscillator. *Proc.Indian Acad. Sci. (math. Sci.)* **116** (3), 337-360.

[Bongioanni and Torrea, 2009] Bongioanni, B. and Torrea, J. L. (2009). What is a Sobolev space for the Laguerre function systems? *Studia Math.*, 192(2):147–172.

[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.

[Chen et al., 2012] Chen, R. Y., Gittens, A., and Tropp, J. A. (2012). The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Inf. Inference*, 1(1):2–20.

[Comte et al., 2015] Comte, F., Cuenod, C.-A., Pensky, M., and Rozenholc, Y. (2017). Laplace deconvolution and its application to dynamic contrast enhanced imaging. *J. R. Stat. Soc., Ser. B*, **79**, 69-94.

[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. *Preprint MAP5 2017-05, to appear in Journal of the Korean Statistical Society*, DOI information: 10.1016/j.jkss.2018.03.001

[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.

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

[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.

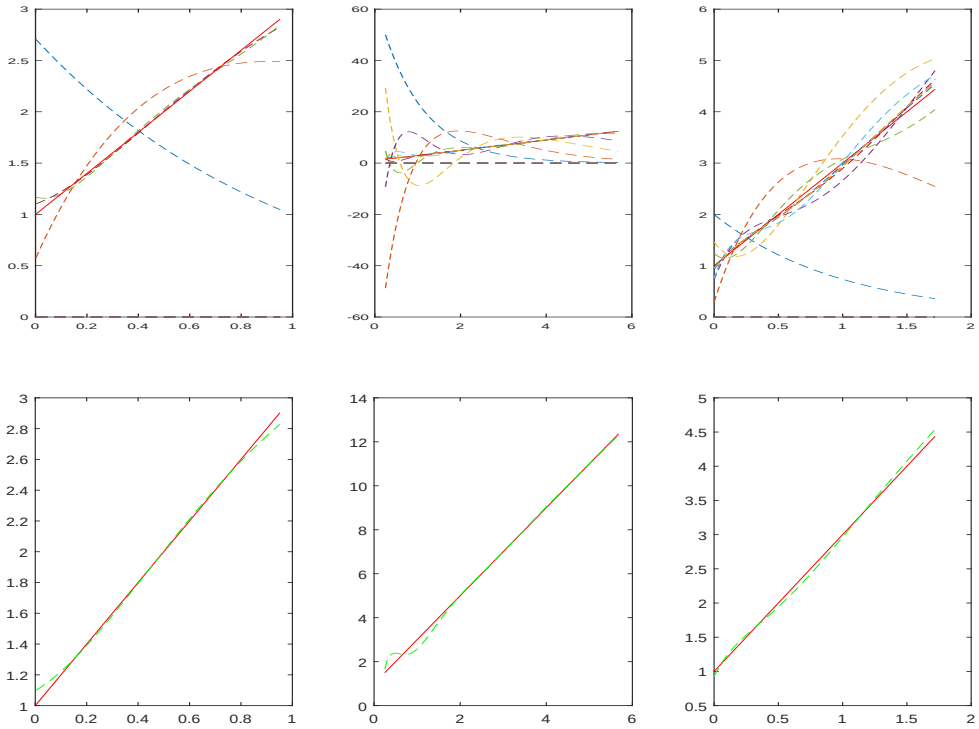
[Tropp, 2015] Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1-2):1–230.

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

[Viennet, 1997] Viennet, G. (1997). Inequalities for absolutely regular processes: application to density estimation. *Probab. Theory Relat. Fields* **107**, 467-492.

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 2 to 5) and the Hermite basis (Figure 6). 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.

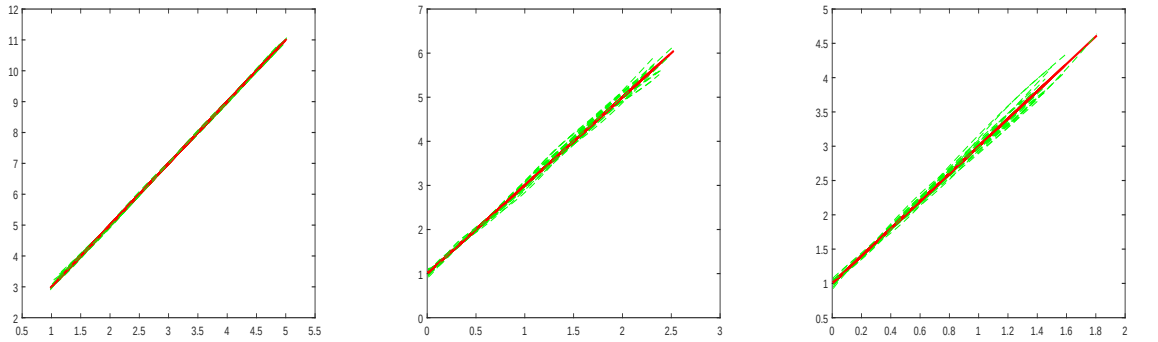


$X \sim \mathcal{U}([0,1])$ $X \sim \mathcal{N}(4,1)$ $X \sim f_3$

Figure 2. 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 2, 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

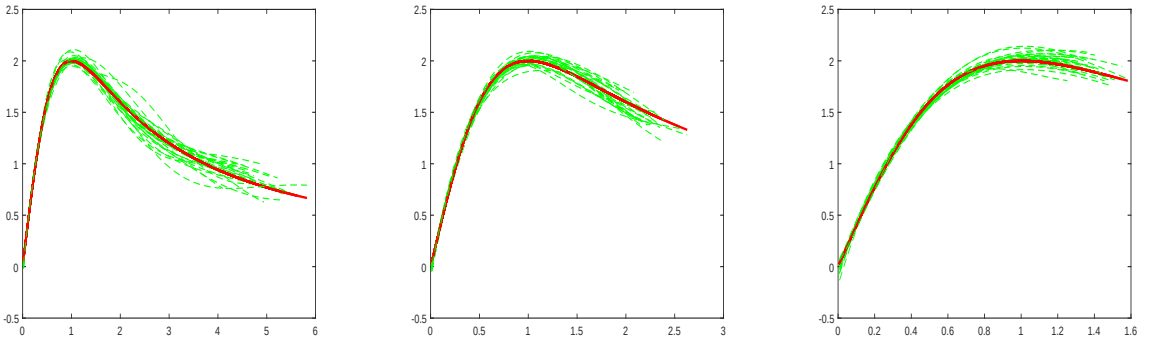
(see the example of uniform distribution on the right). They are also very different from one dimension to another. The second line presents the final estimator, selected by the procedure. In the example of Figure 1, the curve is linear, and is perfectly estimated, although its particular form is unknown and was not *a priori* easy to obtain with the Laguerre basis.



$\mathcal{N}(3,1)$ f_k with $k=4$ f_k with $k=5$

$\bar{\hat{m}} = 7.7(0.5), \bar{m}_{\max} = 8.8(0.4)$ $\bar{\hat{m}} = 9.9(1.9), \bar{m}_{\max} = 10.4(1.7)$ $\bar{\hat{m}} = 6.7(1.1), \bar{m}_{\max} = 7.6(1.0)$

FIGURE 3. 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, $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x\geq 0}$.

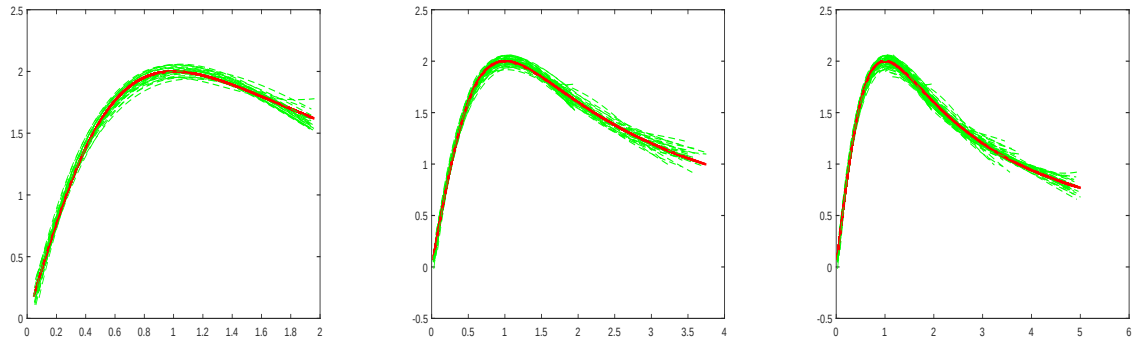


$k=3$ $k=4$ $k=5$

$\bar{\hat{m}} = 5.6(1.0), \bar{m}_{\max} = 15.6(0.7)$ $\bar{\hat{m}} = 4.6(1.1), \bar{m}_{\max} = 10,1(1.6)$ $\bar{\hat{m}} = 3.5(1.3), \bar{m}_{\max} = 7.2(1.0)$

FIGURE 4. 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 3, 4 and 5, we present beams of 25 estimators computed in the Laguerre basis, they give information about the variability of the procedure. Figure 3 is complementary of Figure 2 and considers the same linear regression function with similar distributions for



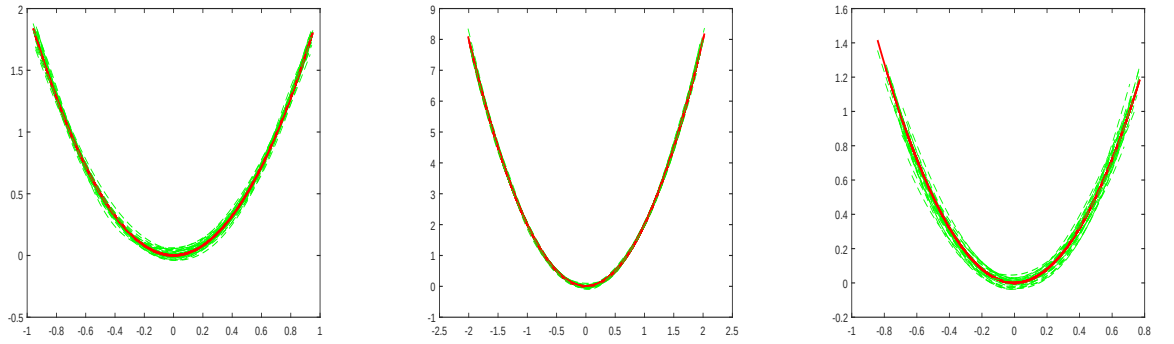
$X \sim \mathcal{U}([0,2])$ $\bar{\hat{m}} = 2.5(0.8), \bar{m}_{\max} = 5.0(0)$

$X \sim \mathcal{E}(1)$ $\bar{\hat{m}} = 5.1(0.8), \bar{m}_{\max} = 9.4(0.6)$

$X \sim \mathcal{N}(3,1)$ $\bar{\hat{m}} = 5.1(0.7), \bar{m}_{\max} = 8.9(0.3)$

FIGURE 5. 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 4 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 that the method does not work for $k = 2, 3$, but the last two plots of Figure 3 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 6.



$X \sim \mathcal{U}([-1,1])$ $\bar{\hat{m}} = 5.1(0.4), \bar{m}_{\max} = 7.0(0)$

$X \sim \mathcal{N}(0,1)$ $\bar{\hat{m}} = 13.4(1.5), \bar{m}_{\max} = 15.0(0.9)$

$X \sim Laplace/4$ $\bar{\hat{m}} = 6.2(1.1), \bar{m}_{\max} = 8.1(0.6)$

FIGURE 6. 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.