Penalization versus Goldenshluger – Lepski strategies in warped bases regression

Gaëlle Chagny

PENALIZATION VERSUS GOLDENSHLUGER-LEPSKI STRATEGIES IN WARPED BASES REGRESSION

GAËLLE CHAGNY(*)

Abstract. This paper deals with the problem of estimating a regression function f, in a random design framework. We build and study two adaptive estimators based on model selection, applied with warped bases. We start with a collection of finite dimensional linear spaces, spanned by orthonormal bases. Instead of expanding directly the target function f on these bases, we rather consider the expansion of $h = f \circ G^{-1}$, where G is the cumulative distribution function of the design, following Kerkyacharian and Picard [24]. The data-driven selection of the (best) space is done with two strategies: we use both a penalization version of a "warped contrast", and a model selection device in the spirit of Goldenshluger and Lepski [21]. We propose by these methods two functions, $\hat{h}_l$ $(l = 1, 2)$, easier to compute than least-squares estimators. We establish nonasymptotic mean-squared integrated risk bounds for the resulting estimators, $\hat{f}_l = \hat{h}_l \circ G$ if G is known, or $\hat{f}_l = \hat{h}_l \circ \hat{G}$ $(l = 1, 2)$ otherwise, where $\hat{G}$ is the empirical distribution function. We study also adaptive properties, in case the regression function belongs to a Besov or Sobolev space, and compare the theoretical and practical performances of the two selection rules.

Keywords: Adaptive estimator. Model selection. Nonparametric regression estimation. Warped bases.
AMS Subject Classification 2010: 62G05-62G08.

October 2011

1. Introduction

1.1. **Statistical framework.** Consider the observation sample $(X_i, Y_i)_{i\in\{1,\dots,n\}}$ $(n \in \mathbb{N}\backslash\{0\})$ of couples of real random variables following the regression setting,

$$Y_i = f(X_i) + \varepsilon_i, 1 \leq i \leq n, \tag{1}$$

where $f : (a; b) \subset \mathbb{R} \to \mathbb{R}$ is the unknown function that we aim at recovering. The random variables $(\varepsilon_i)_{i\in\{1,\dots,n\}}$ are unobserved, centered, admitting a finite variance σ^2, and independent of the design $(X_i)_{i\in\{1,\dots,n\}}$. We assume that the latter are distributed with a density $g > 0$ with respect to the Lebesgue measure, supported on an interval $(a; b)$, $-\infty \leq a < b \leq +\infty$. We denote by G the associated cumulative distribution function (c.d.f. in the sequel), and G^{-1} its inverse, which exists thanks to the assumption $g > 0$.
The aim of this paper is twofold: first, taking advantage of warped bases, we want to provide an adaptive non parametric strategy to recover the regression function f. Secondly, considering a new development of model selection theory, we are interested in the comparison of two selection strategies, from both theoretical and practical points of view: a classical penalization method and a recent selection device in the spirit of Goldenshluger and Lepski (2011) [21] (shortened by "GL method" in the sequel), applied in an original way to a projection estimator.

(*): MAP5 UMR CNRS 8145, University Paris Descartes, France
email: gaelle.chagny@parisdescartes.fr.

1.2. **Motivation.** Adaptive estimation of the regression function is a well-developed problem, and several procedures have been set. Historical methods are kernel strategies, initiated by Nadaraya (1964) [28] and Watson (1964) [31] who proposed kernel-type estimators, built as the ratio of an estimator of the product fg divided by an estimator of the density g. The data-driven choice of the bandwidth, leading to adaptive estimators, is studied more accurately for example by Fan and Gijbels (1992) [19] and Hardle and Tsybakov (1997) [23], who provide asymptotic results (for methods also involving local polynomials). Nevertheless, estimators resulting of this strategy have the drawback of involving a ratio, with a denominator that can be small: this implies difficulties to study the risk and to implement the method.
In a different direction, estimators based on the expansion of the target function into bases, especially orthogonal-bases, have been proposed: spline bases (Golubev and Nussbaum (1992) [22]), wavelet bases (Donoho *et al.* (1995) [15], Cai and Brown (1998) [13] in the fixed design case, Antoniadis *et al.* (1997) [1] in the random-design case), and also trigonometric bases (Efromovich [18] (1999)). Wavelet thresholding strategies offer a degree of localization leading to almost minimax but asymptotic rate of convergence. To obtain non-asymptotic risk bounds, all these estimators can be studied from the model selection point of view, initiated among others by Barron *et al.* (1999) [5]. The problem is to select a "best" estimator among a collection of projection estimators, for example least-squares estimators, to prove oracle inequalities for the risk. The selection is standardly done by the minimization of a penalized criterion (see for example Kohler and Krzyzack (2001) [26], Wegkamp (2003) [32], Birgé (2004) [7], and Baraud (2002) [4]). But procedures based on the minimization of a least-squares contrast do not provide explicit estimators without matrix invertibility requirements (most of the time implicitly).

1.3. **Estimation strategy.** Adopting this model selection point of view, and using warped bases developed for building wavelet thresholding estimators by Kerkyacharian and Picard (2004) [24], we provide in this paper adaptive estimators. These estimates still satisfy non asymptotic oracle-bounds and reach the exact optimal rate under mild assumptions while being easier to compute and more stable, even in case the amount of data can vary in the estimation domain. More precisely, denoting by $u \circ v$ the composition of functions u and v, we define

$$h = f \circ G^{-1} = f\left(G^{-1}\right). \tag{2}$$

We assume that h is squared integrable, we provide estimators for h of the form

$$\hat{h}_D = \sum_{j=1}^{D} \hat{a}_j \varphi_j,$$

for a collection of possible D, with $(\varphi_j)_j$ a classical orthonormal family, and $\hat{a}_j$ estimator of scalar product $\langle h, \varphi_j \rangle$. Then we define

$$\hat{f}_D = \hat{h}_D \circ G \text{ or } \hat{f}_D = \hat{h}_D \circ \hat{G},$$

as estimators of f, depending on whether we assume that G is known or not (in this last case, $\hat{G}$ is the empirical distribution function). We get thus a development of the estimator in warped bases, that is,

$$\hat{f}_D = \sum_{j=1}^{D} \hat{a}_j \left(\varphi_j \circ G\right), \text{ or } \hat{f}_D = \sum_{j=1}^{D} \hat{a}_j \left(\varphi_j \circ \hat{G}\right).$$

The warping strategy brings a procedure computationally simple, without any matrix inversion (which are costly from practical point of view). The selection of "best" index $\hat{D}$ among all

possible D is done in a second time with two strategies. First, we use a penalized version of a "warped contrast". Next, recent works of Goldenshluger and Lepski (2011) [21], in case of density estimation can be explored to propose a new selection strategy. Thus we have at hand two data-driven estimators of the unknown function.
We prove that they both automatically realize the usual squared-bias/variance compromise, provide non asymptotic oracle-inequalities for each estimator. We give also asymptotic rate of convergence on functional spaces, of Besov or Sobolev type. We find the classical non-parametric estimation rate, that is $n^{-2\alpha/(2\alpha+1)}$ where α is the regularity index. Thus, the equivalence between the two adaptive estimators - one based on penalization, the other on GL method - is obtained from theoretical point of view. However, on our practical examples, the new GL strategy outperforms the penalization device.

1.4. **Organization of the paper.** We begin with the case of known design c.d.f in Section 2. In this simpler framework, we can easily explain how the estimators are built and state their adaptivity, while the general case of unknown design distribution is the subject of Section 3: it requires further technicalities, but similar results are proved. They are illustrated via simulations in Section 4. The proofs are gathered in Section 5.

2. Case of known design c.d.f.

To have a better understanding of the definition and properties of the estimators in the general case, we first focus on the simpler situation of known design distribution. This "toy-case", used also by other authors (see for example Pham Ngoc [29]) allows us to derive very simple results, with few assumptions and short proofs.
We deal first with the estimation of the function h defined by (2). We consider a family of approximation spaces. In a first step, we estimate h or more precisely its projection on these spaces. The second step is to ensure an automatic selection of the space, without any knowledge on f. Finally, we warp the function to estimate f (and not h).

2.1. **Assumptions on the models.** The models are linear spaces of functions included in $L^2([0;1])$, the set of square-integrable real-valued functions on the interval $[0;1]$. We denote the collection $\{S_m,\ m \in \mathcal{M}_n\}$, where $\mathcal{M}_n$ is a finite set of indexes, with cardinality depending on the number of observations n. The assumptions and notations are the following:

- [$\mathcal{M}_1$] All the linear spaces S_m are finite-dimensional. For all $m \in \mathcal{M}_n$, we denote by D_m the dimension of the space S_m and assume $1 \leq D_m \leq n$.
- [$\mathcal{M}_2$] The models are nested, that is, for all $(m_1, m_2) \in \mathcal{M}_n^2$, such that $D_{m_1} \leq D_{m_2}$, $S_{m_1} \subset S_{m_2}$. We denote by $(\varphi_j)_{j\in\{1,\dots,D_m\}}$ an orthonormal basis which spans S_m ($m \in \mathcal{M}_n$), and by $m_{\max}$ the index of the largest model in the collection.
- [$\mathcal{M}_3$] There exists a positive constant ϕ_0 such that for all indexes $m \in \mathcal{M}_n$ and all function $t \in S_m$, $\|t\|_\infty \leq \phi_0\sqrt{D_m}\|t\|$. This useful link between the L^2 norm and the infinite norm is equivalent to a property of the basis $(\varphi_j)_{j\in\{1,\dots,D_m\}}$: $\|\sum_{j=1}^{D_m}\varphi_j^2\|_\infty \leq \phi_0^2 D_m$. See Birgé and Massart [8] for the proof of the equivalence.

The above assumptions are not too restrictive. Indeed, they are verified by the spaces spanned by usual bases: trigonometric basis, regular compactly supported wavelet basis, regular histogram basis and regular polynomial basis (with dyadic subdivisions in the last two examples). We refer to section 3.2.1 for a description of trigonometric models, and to Barron *et al.* [5], and Brunel and Comte [10] for the other examples.

2.2. **Estimation on a fixed model.**

2.2.1. *Contrast and estimator on one model.* We define the contrast function:

$$\forall t \in L^2([0;1]) \mapsto \gamma_n(t,G) := \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} Y_i \left(t \circ G(X_i)\right), \tag{3}$$

where $\|.\|$ is the usual Hilbert norm on the space $L^2([0;1])$, associated to the scalar-product denoted by $\langle .,. \rangle$. Notice that $\gamma_n(.,G)$ represents an empirical counterpart for the quadratic risk: for all $t \in L^2([0;1])$,

$$\begin{aligned}
\mathbb{E}\left[\gamma_n(t,G)\right] - \mathbb{E}\left[\gamma_n(h,G)\right] &= \|t\|^2 - \|h\|^2 - 2\mathbb{E}\left[f(X_1)\left\{(t-h)\circ G\right\}(X_1)\right], \\
&= \|t\|^2 - \|h\|^2 - 2\int_{[a;b]} f(x)\left\{(t-h)\circ G\right\}(x)g(x)dx, \\
&= \|t\|^2 - \|h\|^2 - 2\int_{[0;1]} h(u)(t-h)(u)du, \\
&= \|t\|^2 - \|h\|^2 - 2\langle h, t-h\rangle, \\
&= \|t-h\|^2,
\end{aligned}$$

so that h minimizes $t \mapsto \mathbb{E}[\gamma_n(t,G)]$ over $L^2([0;1])$. This explains why a relevant strategy to estimate h consists in minimizing $\gamma_n(.,G)$ over each set S_m:

$$\hat{h}_m^G = \arg\min_{t \in S_m} \gamma_n(t,G). \tag{4}$$

The unique resulting estimator (for each index m) has a particularly simple expression,

$$\hat{h}_m^G = \sum_{j=1}^{D_m} \hat{a}_j^G \varphi_j, \text{ with } \forall j \in \{1,\dots,D_m\},\ \hat{a}_j^G = \frac{1}{n}\sum_{i=1}^{n} Y_i \varphi_j(G(X_i)). \tag{5}$$

Finally, we set

$$\hat{f}_m^{G,G} = \hat{h}_m^G \circ G$$

as an estimator of f. The explicit formula (5) is an unbiased estimator of the orthogonal projection of h onto S_m. Compare for example to the classical least-squares estimator, which involves a matrix inversion (see Baraud [4] and Section 4 for details). Notice also that our notation for the estimator involves two super-indexes G to underline the dependence on the c.d.f. G through both the coefficient $\hat{a}_j^G$ and the composition by G.

2.2.2. *Risk on one model.* In this section, we fix a model S_m and briefly study the quadratic risk of the estimator $\hat{f}_m^{G,G}$. As for all the results stated in the sequel, we evaluate the risk with respect to the norm $\|.\|_g$ naturally associated to our estimation procedure:

$$\|v\|_g^2 = \int_{(a;b)} v^2(x)g(x)dx, \quad \langle v,w\rangle_g = \int_{(a;b)} v(x)w(x)g(x)dx,$$

for any functions $v, w \in L^2((a;b),g)$, the space of squared-integrable functions on $(a;b)$ with respect to the Lebesgue measure weighted by the density g. However, it is also possible to control the classical L^2 norm on $(a;b)$, under the assumption that g is bounded from below by a strictly positive constant: if, for any $x \in (a;b)$, $g(x) > g_0 > 0$, then

$$\|v\|_g^2 \geq g_0 \int_{(a;b)} v^2(x)dx.$$

Notice besides that the following links hold between this weighted norm and the classical norm on $L^2([0;1])$ previously defined: for $t, s \in L^2([0;1])$, we compute, using $G' = g$,

$$\|t \circ G\|_g = \|t\|, \quad \langle t \circ G, s \circ G \rangle_g = \langle t, s \rangle.$$

Thus, the quadratic risk of $\hat{f}_m^{G,G}$ is given by

$$\begin{aligned} \mathbb{E}\left[\left\|\hat{f}_m^{G,G} - f\right\|_g^2\right] &= \left\|f - f_m^G\right\|_g^2 + \mathbb{E}\left[\left\|f_m^G - \hat{f}_m^{G,G}\right\|_g^2\right], \\ &= \|h - h_m\|^2 + \mathbb{E}\left[\left\|h_m - \hat{h}_m^G\right\|^2\right], \end{aligned} \tag{6}$$

where

$$f_m^G = h_m \circ G \text{ and } h_m \text{ is the orthogonal projection of } h \text{ onto } S_m, \text{ with respect to } \langle ., . \rangle. \tag{7}$$

Hence, we recover the usual decomposition into two terms: a squared bias term, which decreases when the dimension of the model S_m grows (roughly, it is at most of order $D_m^{-2\alpha}$, where α is the index of smoothness of h), and a variance term, proportional to the dimension of the model S_m:

$$\mathbb{E}\left[\left\|f_m^G - \hat{f}_m^{G,G}\right\|_g^2\right] = \sum_{j=1}^{D_m} \mathrm{Var}\left(\hat{a}_j^G\right) = \sum_{j=1}^{D_m} \frac{1}{n} \mathrm{Var}\left(Y_1 \left(\varphi_j \circ G\right)(X_1)\right) \leq \mathbb{E}\left[Y_1^2\right] \phi_0^2 \frac{D_m}{n}, \tag{8}$$

where ϕ_0^2 is defined in Assumption $[\mathcal{M}_3]$ (see section 2.1).
Consequently, the best estimator among the family $(\hat{f}_m^{G,G})_{m \in \mathcal{M}_n}$ (in the sense that it achieves the smallest risk among the collection) is the one which realizes the trade-off between the two terms, without any knowledge of the index of smoothness α.

2.3. **Selection rules and main results.**

2.3.1. *Selection rules.* The aim is to realize a data-driven selection of the space S_m. For that purpose, we give a strategy to choose an estimator among the collection $(\hat{f}_m^{G,G})_{m \in \mathcal{M}_n}$. We propose two different strategies and build consequently two estimators.
First, the selection can be standardly done by

$$\hat{m}^{(1),G} = \arg \min_{m \in \mathcal{M}_n} \left[\gamma_n(\hat{h}_m^G, G) + \mathrm{pen}^G(m)\right],$$

with $\mathrm{pen}^G(.)$ a function to be properly chosen. As, $\gamma_n(\hat{h}_m^G, G) = -\|\hat{h}_m^G\|^2 = -\|\hat{f}_m^{G,G}\|_g^2$, and $\|h - h_m\|^2 = \|h\|^2 - \|h_m\|^2$, we can say that $\gamma_n(\hat{h}_m^G, G)$ estimates the bias term, up to an additive constant. This explains why the order of the penalty can be the upper bound on the variance term, that is

$$\mathrm{pen}^G : m \mapsto c_1 \phi_0^2 \mathbb{E}[Y_1^2] \frac{D_m}{n}, \tag{9}$$

with c_1 a purely numerical constant. In practice, we use a method inspired by the slope heuristic to find the value of this constant (see Section 4).
The second method follows the scheme developed by Goldenshluger and Lepski [21] for density estimation. The adaptive index is also chosen as the value which minimizes a sum of two terms:

$$\hat{m}^{(2),G} = \arg \min_{m \in \mathcal{M}_n} \left[A^G(m) + V^G(m)\right],$$

where V^G is also the order of the variance term:

$$V^G : m \mapsto c_2 \phi_0^2 \mathbb{E}[Y_1^2] \frac{D_m}{n}, \tag{10}$$

where c_2 is a purely numerical constant (adjusted in practice by simulations). Here the function A^G does not depend on the contrast: it is rather based on the comparison of the estimators built in the first stage:

$$A^G(m) = \max_{m' \in \mathcal{M}_n} \left(\left\| \hat{h}^G_{m'} - \hat{h}^G_{m \wedge m'} \right\|^2 - V^G(m') \right)_+,$$

where $x_+ = \max(x, 0)$, $x \in \mathbb{R}$. We will prove besides that $A^G(m)$ has the order of the bias term (see Lemma 6). Thus we get two estimators, explicitly expressed in a warped basis:

$$\tilde{f}^G_1 = \hat{h}^G_{\hat{m}^{(1),G}} \circ G, \quad \tilde{f}^G_2 = \hat{h}^G_{\hat{m}^{(2),G}} \circ G.$$

We stress out the fact that these estimators are simple to compute: their coefficients $\hat{a}^G_j$ are empirical means, and even if the "penalties" (pen^G and V^G) contain the unknown expectation $\mathbb{E}[Y_1^2]$, this term can be easily replaced in practice or theory by the empirical mean $(1/n)\sum_{i=1}^n Y_i^2$ (see Brunel and Comte [10], proof of Theorem 3.4 p.465).
In addition to the advantage of the warped basis, the comparison of these two estimators, from both theoretical and practical point of view is new, and is of interest also for other statistical estimation problems.

2.3.2. *Oracle-inequality.* The first theorem provides non-asymptotic bounds for the risk of each estimator.

Theorem 1. *We assume that the regression function f is bounded on the interval $[a;b]$. We consider models satisfying properties $[\mathcal{M}_1]$, $[\mathcal{M}_2]$ and $[\mathcal{M}_3]$, and finally suppose that there exists a real-number $p > 4$ such that $\mathbb{E}\left[|\varepsilon_1|^{2+p}\right] < \infty$.*
Then, the following inequality holds:

$$\mathbb{E}\left[\left\| \tilde{f}^G_i - f \right\|^2_g \right] \leq \min_{m \in \mathcal{M}_n} \left\{ k_i \left\| f - f^G_m \right\|^2_g + k'_i \phi_0^2 \mathbb{E}\left[Y_1^2\right] \frac{D_m}{n} \right\} + \frac{C_i}{n}, \quad i = 1, 2, \tag{11}$$

where f^G_m is defined by (7), k_i and k'_i, $(i = 1, 2)$ are numerical constants, and C_i $i = 1, 2$ are constants independent of n and m, but depending on $\mathbb{E}[Y_1^2]$, ϕ_0^2, σ^2, $\mathbb{E}[|\varepsilon_1|^{2+p}]$ and $\|f\|_\infty$, where $\|f\|_\infty = \sup_{(a;b)} |f(x)|$.

Let us comment this result.

- These non-asymptotic risk bounds, also called oracle-inequalities prove that both estimators automatically realize the squared bias/variance trade-off under few weak assumptions, up to some multiplicative constants (which are precised in the proof). This enhances the interest of warped bases: the risk of the estimators is smaller (up to the constant) than the risk of the best estimator in the family $(\hat{f}^{G,G}_m)_m$. Moreover, the two estimators (the one selected by the GL method and the one selected by penalization) are theoretically equivalent in this context.
- Note that the assumptions for this result are particularly weak, compared to usual hypotheses in other statistical framework (D_m in only supposed bounded by n). Moreover the proof is short, following the general setting of model selection methods (see for example [8]): it is mainly based on a concentration inequality due to Talagrand. The details can be found in Section 5. Remark also that the choice of $p = 4$ for the integrability of ε_1 (instead of $p > 4$) leads to the same inequality with a remainder of order $\ln^4(n)/n$ (instead of $1/n$). We can still relax this assumption: a moment of order $2+p$, $p > 2$ for ε_1 is enough, if we suppose in compensation $D_m = O(\sqrt{n})$. These moment conditions may probably be improved, but we do not go further in this direction, to avoid additional technicalities. We also point out the fact that other results in regression model hold

under weak conditions on the noise term (in the sense that no exponential moment for the ε_i are required, contrary to the conditions in [5]): see for example recent works of Audibert and Catoni [2] and [3], in a prediction framework, and works of Wegkamp [32] or Baraud [4] for model selection point of view.

2.3.3. *Rate of convergence for the risk.* Even if the novelty of our results is their non-asymptotic characters (compared to other warped-bases estimators in this framework, see for example Kerkyacharian and Picard [24] and Pham Ngoc [29]), we can also deduce from Theorem 1 the rate of convergence of the risk. For that purpose, assume that $h = f \circ G^{-1}$ belongs to the Besov space $\mathcal{B}^{\alpha}_{2,\infty}$, for α a positive number.
Let us recall the definition of this space. First, for r a positive integer and v a positive number, the $r-$th order difference of a real-valued function t on the interval $[0;1]$ is defined by

$$\Delta_v^r t(x) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} t(x+kv),$$

where x is such that the $x+kv$ belongs to $[0;1]$, $k \in \{0, \dots r\}$. Next, for $u > 0$, the modulus of smoothness is given by $\omega_r(t,u)_2 = \sup_{0<v\leq u} \|\Delta_v^r t\|$. We say that the function t belongs to the Besov space $\mathcal{B}^{\alpha}_{2,\infty}$ if t belongs to the space $L^2([0;1])$ and if, for $r = [\alpha]+1$ ([.] is the integer part function), $|t|_{\mathcal{B}^{\alpha}_{2,\infty}} = \sup_{u>0} u^{-\alpha}\omega_r(t,u)_2 < \infty$. We refer to DeVore and Lorentz [16] for general definitions and properties of this space. Finally, for all $L > 0$, we denote by $\mathcal{B}^{\alpha}_{2,\infty}(L)$ the space of functions t which satisfies: $|t|_{\mathcal{B}^{\alpha}_{2,\infty}} \leq L$.
It is well known that for all collections of models described in section 2.1 (trigonometric models, regular polynomial bases, regular and compactly supported wavelet bases), the projection h_m of h on the subspace S_m achieves the rate of approximation for the Besov class of functions $\mathcal{B}^{\alpha}_{2,\infty}(L)$ (see Lemma 12 from Barron *et al.* [5]):

$$\|h - h_m\|^2 \leq C(\alpha) L^2 D_m^{-2\alpha}, \tag{12}$$

where $C(\alpha)$ is a constant depending on α and also on the basis. Therefore, the minimization of the left side of inequality (11) leads to the following corollary:

Corollary 1. *We suppose that the function $h = f \circ G^{-1}$ belongs to the Besov space $\mathcal{B}^{\alpha}_{2,\infty}(L)$, for some fixed $\alpha > 0$ and $L > 0$. We assume also that h is bounded over the interval $[0;1]$. We consider one of the models defined in Section 2.1: trigonometric model, dyadic piecewise polynomials (with a regularity r such that $r \geq \alpha - 1$) or compactly supported regular wavelets. Then, under the assumptions of Theorem 1,*

$$\mathbb{E}\left[\left\|\tilde{f}_i^G - f\right\|_g^2\right] \leq C(L,\alpha) n^{\frac{-2\alpha}{2\alpha+1}}, \ i = 1, 2,$$

with $C(L,\alpha)$ a numerical constant which depends only on L and α.

Thus, the model selection procedure leads not only to a non-asymptotic squared bias/variance trade-off but also to an adaptive estimator: indeed, it automatically reaches the asymptotic rate of order $n^{-2\alpha/(2\alpha+1)}$, the minimax rate, in regression setting.
Theorem 2 in Kerkyacharian and Picard [24] states a rate $(n/\ln(n))^{-2\alpha/(2\alpha+1)}$ for an estimator obtained in the same framework (G known, warped basis) by a thresholding algorithm on wavelet coefficients: thus, the rate we get does not suffer from a loss of a $\ln(n)$ factor. Therefore, our method provides an improvement. Moreover, Theorem 1 and Corollary 1 are valid for several models (wavelets models, but also trigonometric models...) and, contrary to [24], for a noise ε_1 not necessarily gaussian (only weak integrability assumptions are required).

Notice also that the assumptions in Corollary 1 are set on function $h = f \circ G^{-1}$, like Proposition 2 of [24]. Proper regularity conditions on function f can also be used to get the same asymptotic result, by defining "weighted" Besov spaces. We refer to Section 4.3 in [24] in which such spaces are precisely described and their properties stated.

3. Case of unknown design c.d.f.

3.1. **The estimators.** The obvious question resulting of Section 2 is: what is to be done if the c.d.f. is not known? To adapt the previous estimation procedure, we replace G by its empirical counterpart. But instead of estimating G over the whole sample, we assume that we observe $(X_{-i})_{i\in\{1,\dots,n\}}$, a sample of random variables distributed as the $(X_i)_i$, and independent of them, and we define,

$$\hat{G}_n : \ x \ \mapsto \ \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_{-i}\leq x}.$$

The aim is to simplify the proofs. We just set a simple "plug-in" strategy to define the estimators. First, for each index $m \in \mathcal{M}_n$, we set

$$\hat{h}_m^{\hat{G}} = \sum_{j=1}^{D_m} \hat{a}_j^{\hat{G}} \varphi_j, \text{ with } \forall j \in \{1,\dots,D_m\}, \hat{a}_j^{\hat{G}} = \frac{1}{n}\sum_{i=1}^{n} Y_i \varphi_j\left(\hat{G}_n(X_i)\right), \tag{13}$$

which is the minimizer of the contrast function $t \mapsto \gamma_n(t, \hat{G}_n)$ on S_m (see (3)). Note that the $\hat{h}_m^{\hat{G}}$, $m \in \mathcal{M}_n$, are still easily available for the statistician, like the estimators of f: $\hat{f}_m^{\hat{G},\hat{G}} = \hat{h}_m^{\hat{G}} \circ \hat{G}_n$. Then, the selection rules follow exactly the same scheme as previously, and allow us to build two estimators. Define, for each $m \in \mathcal{M}_n$,

$$\begin{aligned} &\text{pen} : \ m \mapsto c_1' \phi_0^2 \mathbb{E}[Y_1^2] D_m/n, \\ &V : \ m \mapsto c_2' \phi_0^2 \mathbb{E}[Y_1^2] D_m/n, A(m) = \max_{m'\in\mathcal{M}_n} \left(\left\| \hat{h}_{m'}^{\hat{G}} - \hat{h}_{m\wedge m'}^{\hat{G}} \right\|^2 - V(m') \right)_+, \end{aligned} \tag{14}$$

with c_1' and c_2' purely numerical constants (adjusted in practice, see Section 4), and set

$$\hat{m}^{(1)} = \arg\min_{m\in\mathcal{M}_n} \left[\gamma_n(\hat{h}_m^{\hat{G}}, \hat{G}_n) + \text{pen}(m) \right], \ \hat{m}^{(2)} = \arg\min_{m\in\mathcal{M}_n} \left[A(m) + 2V(m) \right]. \tag{15}$$

Finally, the selected estimators are

$$\tilde{f}_1^{\hat{G}} = \hat{h}_{\hat{m}^{(1)}}^{\hat{G}} \circ \hat{G}_n, \ \ \tilde{f}_2^{\hat{G}} = \hat{h}_{\hat{m}^{(2)}}^{\hat{G}} \circ \hat{G}_n. \tag{16}$$

3.2. **Main results.**

3.2.1. *Framework.* The goal of this section is to establish adaptive properties for both estimators $\tilde{f}_i^{\hat{G}}$, $i = 1, 2$. As already said, they depend on the empirical c.d.f. $\hat{G}_n$ at two stages, which leads to complexity in the proof. For instance, it requires control of terms of the form $\varphi_j(\hat{G}_n) - \varphi_j(G)$. That is why we select one of the bases only, and not any of the ones used in Section 2. Following the example of Efromovich [18], we use models based on the trigonometric basis, that is $S_m = \text{Span}\{\varphi_1, \dots, \varphi_{D_m}\}$, with $D_m = 2m+1$, $m \in \mathcal{M}_n = \{1, \dots, [n/2]-1\}$, and for all $j \in \{1, \dots, m\}$ and all $x \in [0;1]$,

$$\varphi_1(x) = 1, \ \ \varphi_{2j}(x) = \sqrt{2}\cos(2\pi j x), \ \ \varphi_{2j+1}(x) = \sqrt{2}\sin(2\pi j x).$$

Notice that the assumption $[\mathcal{M}_3]$ is satisfied with $\phi_0 = 1$. This choice is guided among other things by the following property: let h be a function continuously derivable on the interval $[0;1]$, such that $h(0) = h(1)$. The orthogonal projection of the derivative h' of h onto S_m coincides

with the derivative of the projection of h onto S_m. Formally, if we denote by Π_{S_m} the operator of orthogonal projection onto S_m, $\Pi_{S_m}(h') = (\Pi_{S_m}(h))'$.

In this framework, we get a similar result to the one obtained when G was supposed to be known.

Theorem 2. *We assume that the regression function f and the density g admit both continuous derivative on $[a;b]$ (respectively $[0;1]$). We assume also that $\|f\|_g \leq L$ ($L>0$) and that $f(a) = f(b)$. We consider the trigonometric models, and suppose that there exists a real-number $p > 8/3$ such that $\mathbb{E}\left[|\varepsilon_1|^{2+p}\right] < \infty$, and that for any $m \in \mathcal{M}_n$, $D_m = O(n^{1/3}/\ln(n))$.*
Then, the following inequality holds: for all $n \geq n_0 = \exp(\|h'\|^2)$,

$$\mathbb{E}\left[\left\|\tilde{f}_i^{\hat{G}} - f\right\|_g^2\right] \leq \min_{m\in\mathcal{M}_n}\left\{k_i\left\|f - f_m^G\right\|_g^2 + k_i'\phi_0^2\mathbb{E}\left[Y_1^2\right]\frac{D_m}{n}\right\} + \frac{C_i\ln(n)}{n},\ i=1,2, \tag{17}$$

where f_m^G is defined by (7), k_i and k_i', ($i=1,2$) are numerical constants, and C_i ($i=1,2$) are constants independent on n and m, but depending on $\|\varphi_2^{(l)}\|$ ($l=1,3$), $\|h\|$, $\|h'\|$, and $\mathbb{E}[Y_1^2]$.

The theorem proves that warped-bases selected estimators have exactly the same behaviour as least-squares estimator (see for instance Inequality (15), in Baraud [4]): both estimators realize the squared bias/variance compromise. Consequently, a model selection strategy with warped-bases has the advantage of providing estimators easier to compute than least-squares estimators and with analogous theoretical properties.
Notice that the upper bound we provide for the risk holds for any $n \geq n_0$ so it can still be considered as a non-asymptotic result. This is an advantage compared to other procedures based on the thresholding of the estimated coefficients in wavelet bases, even if the bases are also warped (see for example Kerkyacharian and Picard [24]).

3.2.2. *Rate of convergence for the risk.* As a consequence of the choice of trigonometric models, it is natural to consider spaces of periodic functions, that is Sobolev spaces. Following Tsybakov [30], we define first, for α a positive integer and L a positive number, the space $W_2^\alpha(L)$ of real-valued functions h on the interval $[0;1]$ such that $h^{(\alpha-1)}$ is absolutely continuous and

$$\left\|h^{(\alpha)}\right\|^2 = \int_0^1 \left(h^{(\alpha)}(x)\right)^2 dx \leq L^2.$$

Then, we say that a function h belongs to the space $W_{per}^{2,\alpha}(L)$ if it belongs to $W_2^\alpha(L)$ and

$$\forall j = 0,1,\ldots,\alpha-1,\ h^{(j)}(0) = h^{(j)}(1).$$

This definition can be extended to positive real-number α (see [30] for details).
The standard rate of convergence is then achieved if smoothness properties are supposed for h. In fact, the approximation error orders can also be bounded in the case of Sobolev spaces. If h belongs to the space $W_{per}^{2,\alpha}(L)$ for $\alpha \geq 1$ and $L > 0$, and if we denote by h_m its orthogonal projection (for the usal sclar product of $L^2([0;1])$) on the trigonometric model S_m, then Tsybakov [30] (see Lemma A.3 [30]) proves the following inequality:

$$\|h - h_m\|^2 \leq \frac{L^2}{\pi^{2\alpha}} D_m^{-2\alpha}.$$

Consequently, we state the following result, which is similar to Corollary 1:

Corollary 2. *We suppose that the function $h = f \circ G^{-1}$ belongs to the Sobolev space $W^{2,\alpha}_{per}(L)$, for some fixed $\alpha \geq 1$ and $L > 0$. Then, under the assumptions of Theorem 2,*

$$\mathbb{E}\left[\left\|\tilde{f}_i^{\hat{G}} - f\right\|_g^2\right] \leq C(L,\alpha) n^{\frac{-2\alpha}{2\alpha+1}},\ i = 1, 2,$$

where $C(L,\alpha)$ is a constant which depends on L and α.

Most of the comments following Corollary 1 also apply to this result. The order of the rate, $n^{(-2\alpha)/(2\alpha+1)}$ in place of the rate $(n/\ln(n))^{(-2\alpha)/(2\alpha+1)}$ achieved by the estimator $\hat{f}^{@}$ in Kerkyacharian and Picard [24] is a consequence of model selection strategy, by penalization or GL method. But the assumptions for their result are different of ours. We decide to concentrate on the trigonometric models (instead of the wavelet setting of [24]). Consequently, the estimators are adaptive for Sobolev regularities. This, and the fact that the index α of regularity has to be larger than 1 can seem to be a little more restrictive than the assumptions of Theorem 2 in [24]: h is there assumed to belong to a Besov space with index $\alpha \geq 1/2$, and to a Hölder space (with regularity $1/2$), and these spaces are larger than the one we use. But contrary to them, and in addition to the convergence rate improvement (no additional $\ln(n)$), our methods allow general noise and not only Gaussian noise. Moreover, trigonometric basis enables us to consider other regularities, and to get faster rates. For example, if h belongs to an analytic space, its Fourier's coefficients decrease with exponential rate: $\|h - h_m\| \leq C\exp(-\epsilon D_m)$, for some $\epsilon > 0$ and C a positive constant, leading to the rate $\ln(n)/n$.
Finally, let us notice that assumptions can probably be stated with regularity conditions directly on f instead of h, by defining "weighted" spaces. But, as our main contribution is to provide non asymptotic results which do not require the control of the bias term (and thus, the regularity assumption), this construction is be beyond the scope of the paper.

4. Simulations

4.1. **Implementation.** The simulation study is mainly conducted in order to compare from practical point of view the penalized estimator $\tilde{f}_1^{\hat{G}}$ and the one defined with the GL method $\tilde{f}_2^{\hat{G}}$, when using the trigonometric basis $(\varphi_j)_j$. This comparison is new and beyond the classical regression setting: the study would be of interest in many other contexts.
We also compute the adaptive least-squares estimator, denoted by $\tilde{f}^{LS}$, to investigate the difference between classical orthonormal bases and warped-bases. Let us recall briefly its definition. First, we set, for $t \in L^2([0;1])$, and $m \in \mathcal{M}_n$:

$$\gamma_n^{LS}(t) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - t(X_i))^2 \text{ and } \mathrm{pen}^{LS}(m) = C\sigma^2\frac{D_m}{n}, \tag{18}$$

with C a numerical constant. We set for each m, $\hat{f}_m^{LS} = \arg\min_{t\in S_m}\gamma_n^{LS}(t)$, and select $\hat{m}^{LS} = \arg\min_{t\in S_m}\gamma_n^{LS}(t) + \mathrm{pen}^{LS}(m)$. Then we have $\hat{f}_{\hat{m}^{LS}}^{LS} = \sum_{j=1}^{D_{\hat{m}^{LS}}}\hat{a}_j^{LS}\varphi_j$, where $\hat{a}^{LS} = (\hat{a}_j^{LS})_j$ is computed by inverting the matrix $M_{\hat{m}} = (M_{\hat{m},j,k})_{j,k\in\{1,\dots,D_{\hat{m}}\}}$, that is $\hat{a}^{LS} = M_{\hat{m}}^{-1}b$, with

$$M_{m,j,k} = \frac{1}{n}\sum_{i=1}^{n}\varphi_j(X_i)\varphi_k(X_i), \text{ and } b = (b_j)_{j\in\{1,\dots,D_m\}},\ \ b_j = \frac{1}{n}\sum_{i=1}^{n}Y_i\varphi_j(X_i). \tag{19}$$

We refer to Baraud [4] for theory and to Comte and Rozenholc [14] for practical computation. We have thus three estimators to compute, from data $(X_i, Y_i)_{i\in\{1,\dots,n\}}$. We first notice that their

common expression is:

$$\hat{f}_{\hat{m}} = \sum_{j=1}^{D_{\hat{m}}} \hat{a}_j \psi_j,$$

with, for $\tilde{f}_1^{\hat{G}}$ and $\tilde{f}_2^{\hat{G}}$, $\hat{a}_j = \hat{a}_j^{\hat{G}}$ defined by equation (13) and $\psi_j = \varphi_j \circ \hat{G}_n$, and for $\tilde{f}^{LS}$, $\hat{a}_j = \hat{a}_j^{LS}$ and $\psi_j = \varphi_j$. In the first case, we generate another sample $(X_{-i})_{i\in\{1,\dots,n\}}$, to find the empirical c.d.f $\hat{G}_n$, and to compute the coefficients $\hat{a}_j^{\hat{G}}$. Concretely, choosing $m_{\max} = 8$, we use the following steps:

• For each $m \in \{1, \dots, m_{\max}\}$, compute crit$(m)$, for the three following definitions:

- crit$(m) = \gamma_n(\hat{h}_m^{\hat{G}}, \hat{G}_n) + \text{pen}(m)$ in the warped-bases case, with penalization. Notice that $\gamma_n(\hat{h}_m^{\hat{G}}) = -\sum_{j=1}^{\hat{D}_m} (\hat{a}_j^{\hat{G}})^2$.
- crit$(m) = A(m) + 2V(m)$ in the warped-bases case, with the GL method. Notice that $A(m) = \max_{m'>m}\{\sum_{j=D_m+1}^{D_{m'}} (a_j^{\hat{G}})^2 - V(m')\}_+$.
- crit$(m) = \gamma_n^{LS}(\hat{f}_m^{LS}) + \text{pen}^{LS}(m)$ in the least-squares case. The least-squares contrast is computed like the warped-bases criterion. The penalty defined by (18) is implemented, with σ^2 replaced by the unbiased estimator,

$$\hat{\sigma}^2 = \frac{1}{n - (2mm + 1)} \sum_{i=1}^{n} (Y_i - \hat{f}_{mm}^{LS}(X_i))^2, \text{ with } mm = [\sqrt{n}].$$

• In the three cases, select $\hat{m}$ (that is $\hat{m} = \hat{m}^{(1)}, \hat{m}^{(2)}, \hat{m}^{LS}$) such that crit$(m)$ is minimum.

• Compute then the three estimators $\tilde{f}_l = \sum_{j=1}^{D_{\hat{m}^{(l)}}} \hat{a}_j^{\hat{G}}(\varphi_j \circ \hat{G}_n)$, $l = 1, 2$ and $\tilde{f}^{LS} = \sum_{j=1}^{D_{\hat{m}^{LS}}} \hat{a}_j^{LS} \varphi_j$, at a sequence of equispaced points in $[a; b]$.

Remark: To implement crit(m), the numerical constants c_1' (of pen), C (of pen^{LS}), and c_2' (of V) have to be calibrated. The constant C is chosen equal to 2.5, which is a value often found in the literature (constants of the C_p criterion of Mallows, for example). We decide to concentrate on the data-driven calibration of the constants involved in the definition of the new estimators, that is c_1' and c_2'. The constant c_1' is useful for the penalized warped bases estimator $\tilde{f}_1^{\hat{G}}$: it can thus be carried out for each simulated sample using a method inspired by the slope heuristic (developed first by Birgé and Massart [9]). But this data-driven solution can not be used for the recent method of GL, leading to the estimator $\tilde{f}_2^{\hat{G}}$. So, to compare in the same way the two estimators, we choose to experiment it with fixed constants, previously stated. The constant c_1' is adjusted prior to the comparison, using however the slope heuristic: we use the graphical interface CAPUSHE developed by Baudry *et al.* [6], to conduct an experimentation over 100 samples (see our examples, Section 4.2), with the so-called "dimension-jump" method. We choose then the largest constant over all attempts proposed by the software, that is $c_1' = 4$ (recall that in penality calibration, it is more secure to overpenalize). For the constant of the GL method, we looked at the quadratic risk with respect to its value c_2', and chose one of the first values leading to reasonable risk and complexity of the selected model, $c_2' = 0.5$ (for the computation of the risk, see Section 4.2 below). Notice finally that the specific factor 2 involved in the definition of $\hat{m}^{(2)}$ (see definition (15)) could be also adjusted: it plays a technical role in the proof but might have been replaced by any other constant larger than 1.

4.2. **Examples.** The procedure is applied for different regression functions, design and noise. To concentrate on the comparison of the three methods, we decide to present the estimation results

for two very smooth functions, on the interval $[0;1]$: a polynomial function, $f_1 : x \mapsto x(x-1)(x-0.6)$, and an exponential function, $f_2 : x \mapsto -\exp(-200(x-0.1)^2) - \exp(-200(x-0.9)^2)$. The sensibility of the method to the underlying design is tested with the following densities, all supported by $[0;1]$. In the definitions, c is a constant adjusted to obtain density function in each case:
$-$ $\mathcal{U}_{[0;1]}$, the classical uniform distribution,
$-$ $\mathcal{DU}_{[0;1]}$, probability distribution with density $x \mapsto cx\mathbf{1}_{[0;1]}$,
$-$ $\mathcal{E}_c(1)$, a truncated exponential distribution with mean 1 that is with density $x \mapsto ce^{-x}\mathbf{1}_{[0;1]}$,
$-$ $\mathcal{N}_c(0.5, 0.01)$, a truncated Gaussian distribution with density $x \mapsto c\exp(-(x-0.5)^2/0.02)\mathbf{1}_{[0;1]}(x)$,
$-$ $\mathcal{NBM}_t$, a truncated bimodal Gaussian distribution, with density $x \mapsto c(\exp(-200(x-0.05)^2) + \exp(-200(x-0.95)^2))\mathbf{1}_{[0;1]}(x)$,
$-$ $\mathcal{CM}$, a distribution with piecewise constant density $2.485\mathbf{1}_{[0;0.2]} + 0.01\mathbf{1}_{]0.1;0.8]} + 2.485\mathbf{1}_{]0.8;1]}$,
Finally, the variables ε_i are generated following either a Gaussian distribution, or a Laplace distribution, with mean 0. They are denoted respectively by $\mathcal{N}(0,v)$ (v the variance) and by $\mathcal{L}(0,b)$ (b a positive real such that the Laplace density is $x \mapsto 1/(2b)\exp(-|x|/b)$). The parameters b and v are adjusted for each of the functions f_1 and f_2: it is natural to choose cases in which there is a little more signal than noise. Precisely, the values are chosen such that the ratio of the variance of the signal ($\mathrm{Var}(f(X_1))$) over the variance of the noise ($\mathrm{Var}(\varepsilon_1)$) belongs to $[1.6; 2.4]$, whatever the design distribution. This ratio, denoted by "s2n", will be precised in Tables 1 and 2.

We compare first the visual quality of the reconstruction, for the three estimators. Figure 1 shows beams of estimated functions versus true functions in four cases. Precisely, for each figure, we plot 20 estimators of each kind, built from i.i.d samples of data of size $n = 500$. The three first plots show that the results are quite good for all the estimators. The noise distribution does not seem to affect significantly the results. Notice that the computation of the estimators $\tilde{f}^{LS}$ requires much more time than the others. It is due to the computation of the inverse of the matrix $M_{\hat{m}}$, while the warped-bases methods are simpler. So one can easily use warped bases for estimation problems with large data samples sizes (see for example domains as fluorescence, physics, neuronal models...). The last plot of Figure 1 shows that the warped-bases estimators behave still correctly if the design density is very inhomogeneous (we obtain the same type of plots when the X_i is distributed with $\mathcal{CM}$). In fact, if we implement the least-squares method without taking additionnal precautions and without numerical approximation for the computation of $M_{\hat{m}}^{-1}$, the estimator can not adapt to a design density which nearly vanishes on a long interval. This highlights the interest for warping the bases: this method seems to be very stable, whatever the design distribution, and even if it is very inhomogeneous: it tends to detect better the hole which can occur in the design density. Let us notice that specific methods exist, taking into account the inhomogeneity of the data to obtain upper bounds for the quadratic pointwise risk, see for example Gaïffas [20].
The beams of estimators seem to enhance the equivalence we found in the theory between the GL method and the penalization method. For more precise results concerning these selection rules, we compare L^2 risk, in the different models (the two functions estimated, the possible design and noise). The ISE (Integrated Squared Error) for one estimator $\tilde{f}$ is $\mathrm{ISE} = \int_a^b (f(x) - \tilde{f}(x))^2 dx$. It

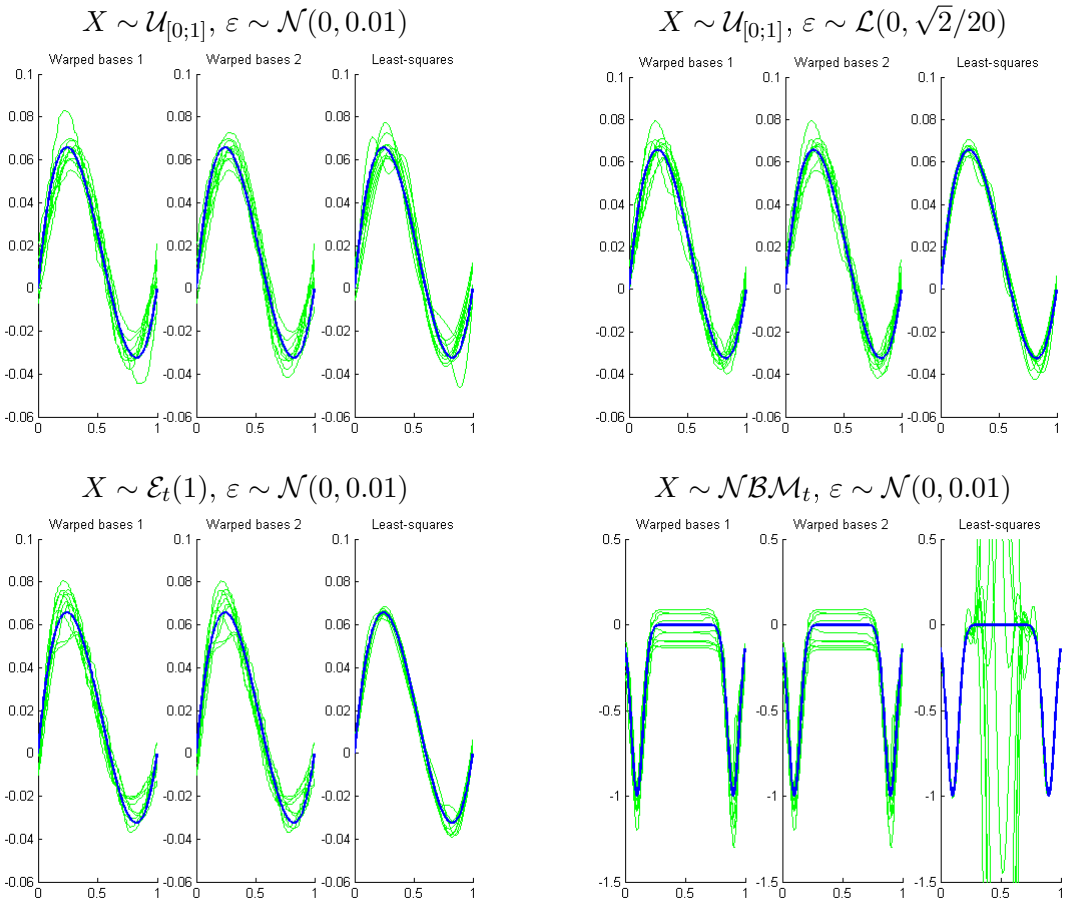

FIGURE 1. Plots of 20 estimators $\tilde{f}_1^{\hat{G}}$ (Warped bases 1), $\tilde{f}_2^{\hat{G}}$ (Warped bases 2) and $\tilde{f}^{LS}$ (Least-squares) of function f_1 or f_2, built from i.i.d. sample in trigonometric bases. Bold line: True function, Thin lines: Estimators.

is computed as follows:

$$ISE = \frac{b-a}{K} \sum_{k=0}^{K} \left[\tilde{f}\left(a + k\frac{b-a}{K}\right) - f\left(a + k\frac{b-a}{K}\right) \right]^2,$$

where K is an integer (we choose $K = 1000$). The mean ISE (MISE) is the mean of those values over $N = 100$ independent simulated samples.

The risks (multiplied by 1000) displayed in Table 1 (estimation of f_1) and 2 (estimation of f_2) for the estimators $\tilde{f}_1^{\hat{G}}$ (WB1) and $\tilde{f}_2^{\hat{G}}$ (WB2) are computed for different sample sizes going from $n = 100$ to 2000. Notice first that the difference of order of size between the values of the two tabulars is explained by the difference of amplitude between the two functions (f_1 takes its values in the interval $[-0.04; 0.07]$ and f_2 in $[-1; 0]$). As expected, the values of MISE get smaller when the sample size increases, and they are similar for both estimators, in most cases. The GL method gives slightly smaller risks in 59% of the cases (in bold-blue in the tables). But it seems that the values are better than those of the penalized estimator in 76% of the cases for the large sample sizes ($n = 500$ to 2000). We have to put this result into perspective: larger

ε	X	n=100	200	500	1000	1500	2000	Estimator
$\mathcal{N}(0, 0.0006)$	$\mathcal{U}_{[0;1]}$	0.238	0.116	0.058	0.029	0.017	0.017	WB1
	s2n=2.07	0.462	0.227	0.087	0.045	0.028	0.024	WB2
	$\mathcal{DU}_{[0;1]}$	0.407	0.254	0.144	0.09	0.069	0.058	WB1
	s2n=1.74	0.55	0.276	**0.141**	**0.084**	**0.064**	**0.054**	WB2
	$\mathcal{E}_t(1)$	0.231	0.152	0.052	0.032	0.021	0.018	WB1
	s2n=1.9	0.501	0.248	0.09	0.042	0.027	0.024	WB2
	$\mathcal{N}_t(0.5, 0.1)$	0.473	0.181	0.089	0.052	0.036	0.028	WB1
	s2n=1.98	0.68	0.243	0.097	0.053	**0.036**	**0.027**	WB2
	$\mathcal{NBM}_t$	0.957	0.788	0.561	0.448	0.436	0.395	WB1
	s2n=1.94	1.037	**0.785**	**0.537**	**0.436**	**0.433**	**0.393**	WB2
	$\mathcal{CM}$	1.012	0.943	0.775	0.718	0.692	0.68	WB1
	s2n=2.07	1.267	0.968	**0.773**	**0.711**	**0.688**	**0.679**	WB2
$\mathcal{L}(0, 0.0173)$	$\mathcal{U}_{[0;1]}$	0.235	0.102	0.051	0.026	0.02	0.016	WB1
		0.44	0.215	0.085	0.04	0.031	0.023	WB2
	$\mathcal{DU}_{[0;1]}$	0.352	0.268	0.13	0.084	0.069	0.059	WB1
		0.494	0.28	**0.122**	**0.074**	**0.062**	**0.054**	WB2
	$\mathcal{E}_t(1)$	0.278	0.133	0.065	0.031	0.024	0.018	WB1
		0.576	0.244	0.099	0.043	0.033	0.023	WB2
	$\mathcal{N}_t(0.5, 0.1)$	0.338	0.2	0.092	0.05	0.036	0.03	WB1
		0.539	0.254	0.101	0.052	**0.036**	**0.028**	WB2
	$\mathcal{NBM}_t$	1.104	0.699	0.562	0.453	0.425	0.412	WB1
		1.221	**0.662**	**0.532**	**0.442**	**0.418**	**0.406**	WB2
	$\mathcal{CM}$	1.078	0.889	0.801	0.716	0.688	0.683	WB1
		1.207	0.919	**0.797**	**0.707**	**0.686**	**0.682**	WB2

TABLE 1. Values of MISE $\times 1000$ averaged over 100 samples, for the estimation of f_1

classes of functions and models would have to be studied to confirm this and we keep in mind that the methods are equivalent from the theoretical point of view.

5. Proofs of the main results

5.1. **A key result.** One of the main argument of the proof of Theorem 1 and Theorem 2 is the control of the centered empirical process defined by

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i \left(t \circ G\right)(X_i) - \langle (t \circ G), f\rangle_g, \quad t \in L^2([0;1]), \tag{20}$$

on the unit sphere

$$\mathcal{S}(m) = \{t \in S_m,\ \|t\| = 1\}$$

of a fixed model S_m. Let us first state the following result, which we use for both theorems.

Proposition 3. *Under the assumptions of Theorem 1, with* $p(m') = 6(1+2\delta)\phi_0^2 \mathbb{E}[Y_1^2]\frac{D_{m'}}{n}$, *(*$\delta > 0$*) for any* $m' \in \mathcal{M}_n$, *there exists a constant* C *depending on* ϕ_0^2, $\|f\|_\infty$, $\mathbb{E}[f^2(X_1)]$, σ^2,

ε	X	n=100	200	500	1000	1500	2000	Estimator
$\mathcal{N}(0, 0.05)$	$\mathcal{U}_{[0;1]}$	73.979	37.574	13.557	6.606	4.088	3.126	WB1
	s2n=2.33	**72.02**	**34.761**	**13.32**	**6.506**	**3.975**	**3.109**	WB2
	$\mathcal{DU}_{[0;1]}$	65.367	54.668	43.972	36.923	32.499	29.707	WB1
	s2n=2.33	73.101	**53.149**	**39.232**	**32.683**	**29.873**	**28.252**	WB2
	$\mathcal{E}_t(1)$	74.224	41.907	17.365	9.384	6.925	5.187	WB1
	s2n=2.37	76.55	**37.431**	**16.401**	**9.074**	**6.842**	5.307	WB2
	$\mathcal{N}_t(0.5, 0.1)$	75.906	53.158	30.022	16.046	13.3	12.119	WB1
	s2n=1.76	88.46	54.046	**27.695**	**15.861**	**13.21**	**12.119**	WB2
	$\mathcal{NBM}_t$	86.712	29.374	14.892	6.949	4.368	3.502	WB1
	s2n=2.06	73.514	32.237	**12.609**	**6.529**	**4.054**	**2.935**	WB2
	$\mathcal{CM}$	125.098	47.224	29.851	20.533	20.016	17.296	WB1
	s2n=1.69	111.872	53.719	31.766	20.593	**18.595**	**16.1**	WB2
$\mathcal{L}(0, 0.1581)$	$\mathcal{U}_{[0;1]}$	77.489	35.98	13.657	6.47	3.808	3.032	WB1
		73.596	**32.823**	13.667	**6.392**	**3.772**	**3.026**	WB2
	$\mathcal{DU}_{[0;1]}$	70.605	55.9	43.65	37.967	33.642	30.021	WB1
		80.886	**54.544**	**38.695**	**32.008**	**29.473**	**27.925**	WB2
	$\mathcal{E}_t(1)$	64.881	44.879	17.774	10.31	6.987	5.856	WB1
		71.622	**38.003**	**16.928**	**9.897**	**6.761**	**5.689**	WB1
	$\mathcal{N}_t(0.5, 0.1)$	82.315	50.384	27.537	15.931	13.474	12.523	WB1
		90.932	**48.743**	**25.119**	16.24	**13.403**	**12.523**	WB2
	$\mathcal{NBM}_t$	98.027	33.034	13.593	7.472	4.697	3.604	WB1
		83.533	**32.761**	**12.162**	**6.437**	**4.484**	**3.119**	WB2
	$\mathcal{CM}$	113.635	48.175	25.483	21.765	18.833	18.229	WB1
		95.868	49.138	**24.812**	**18.662**	**16.717**	**16.011**	WB2

TABLE 2. Values of MISE $\times 1000$ averaged over 100 samples, for the estimation of f_2

$\mathbb{E}[|\varepsilon_1|^p]$ *and* δ *such that,*

$$\mathbb{E}\left[\sum_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m')}(\nu_n(t))^2 - p(m')\right)_+\right] \leq \frac{C}{n}.$$

Proof of Proposition 3

We split the process ν_n into three parts, writing $\nu_n = \nu_n^{(1)} + \nu_n^{(2,1)} + \nu_n^{(2,2)}$, with

$$\nu_n^{(1)}(t) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)\,(t\circ G)\,(X_i) - \langle (t\circ G), f\rangle_g,$$

$$\nu_n^{(2,1)}(t) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathbf{1}_{|\varepsilon_i|\leq\kappa_n}\,(t\circ G)\,(X_i) - \mathbb{E}\left[\varepsilon_i \mathbf{1}_{|\varepsilon_i|\leq\kappa_n}\,(t\circ G)\,(X_i)\right],$$

$$\nu_n^{(2,2)}(t) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathbf{1}_{|\varepsilon_i|>\kappa_n}\,(t\circ G)\,(X_i) - \mathbb{E}\left[\varepsilon_i \mathbf{1}_{|\varepsilon_i|>\kappa_n}\,(t\circ G)\,(X_i)\right],$$

with c a constant depending on the collection of models and where we define

$$\kappa_n = c\frac{\sqrt{n}}{\ln(n)}. \tag{21}$$

We obtain,

$$\begin{aligned}\left(\sup_{t\in\mathcal{S}(m')} \nu_n(t)^2 - p(m')\right)_+ \leq\ & 3\left\{\left(\sup_{t\in\mathcal{S}(m')}\left(\nu_n^{(1)}(t)\right)^2 - \frac{p_1(m')}{3}\right)_+\right. \\ & + \left(\sup_{t\in\mathcal{S}(m')}\left(\nu_n^{(2,1)}(t)\right)^2 - \frac{p_2(m')}{3}\right)_+ \\ & \left. + \sup_{t\in\mathcal{S}(m')}\left(\nu_n^{(2,2)}(t)\right)^2\right\}\end{aligned} \tag{22}$$

with $p_1(.) + p_2(.) = p(.)$.
We upper bound the first two terms by applying the following concentration inequality:

Lemma 4. *Let $\xi_1, \ldots, \xi_n$ be i.i.d. random variables, and define $\nu_n(r) = \frac{1}{n}\sum_{i=1}^n r(\xi_i) - \mathbb{E}[r(\xi_i)]$, for r belonging to a countable class $\mathcal{R}$ of real-valued measurable functions. Then, for $\varepsilon > 0$,*

$$\begin{aligned}\mathbb{E}\left[\left(\sup_{r\in\mathcal{R}}\left(\nu_n\left(r\right)\right)^2 - 2(1+2\varepsilon)H^2\right)_+\right] \leq\ & \frac{4}{K_1}\left\{\frac{v}{n}\exp\left(-K_1\varepsilon\frac{nH^2}{v}\right)\right. \\ & \left. + \frac{49M_1^2}{K_1C^2(\varepsilon)n^2}\exp\left(-\frac{\sqrt{2}K_1C(\varepsilon)\sqrt{\varepsilon}}{7}\frac{nH}{M_1}\right)\right\},\end{aligned} \tag{23}$$

with $C(\varepsilon) = (\sqrt{1+\varepsilon} - 1) \wedge 1$, $K_1 = 1/6$, and

$$\sup_{r\in\mathcal{R}}\|r\|_\infty \leq M_1,\ \mathbb{E}\left[\sup_{r\in\mathcal{R}}|\nu_n(r)|\right] \leq H,\ \text{and}\ \sup_{r\in\mathcal{R}} Var\left(r\left(\xi_1\right)\right) \leq v.$$

Inequality (23) is a classical consequence of Talagrand's inequality given in Klein and Rio [25], see for example Lemma 5 (page 812) in Lacour [27]. Standard density arguments allow to apply it to the unit sphere of a finite dimensional linear space.
We apply Inequality (23) to the first term of equation (22), with function r replaced by r_t : $x \mapsto f(x)(t \circ G)(x)$, $t \in \mathcal{R} = \mathcal{S}_{m'}$, and $\xi_i = X_i$. Let us first compute the constants $M_1^{(1)}$, $H^{(1)}$, and $v^{(1)}$. We observe first that $\|r_t\|_\infty \leq \|f\|_\infty\|t\|_\infty$ and we use assumption $[\mathcal{M}_3]$ to get $\|r_t\|_\infty \leq \phi_0\sqrt{D_{m'}}\|t\|\|f\|_\infty = \phi_0\sqrt{D_{m'}}\|f\|_\infty := M_1^{(1)}$.
Then, noting that $t \in \mathcal{S}(m')$ can be written $t = \sum_{j=1}^{D_{m'}} b_j\varphi_j$ with $\sum_j b_j^2 = 1$, we apply Cauchy-Schwarz's inequality to get $\sup_{t\in\mathcal{S}(m')}\nu_n^{(1)}(t)^2 \leq \sum_{j=1}^{D_{m'}}\nu_n^{(1)}(\varphi_j)^2$. Since assumptions $[\mathcal{M}_2]$ and $[\mathcal{M}_3]$ hold, we obtain

$$\mathbb{E}\left[\sup_{t\in\mathcal{S}(m')}\nu_n^{(1)}(t)^2\right] \leq \sum_{j=1}^{D_{m'}}\frac{1}{n}\mathrm{Var}(f(X_1)\left(\varphi_j \circ G\right)(X_1)) \leq \phi_0^2\mathbb{E}\left[f^2(X_1)\right]\frac{D_{m'}}{n} := \left(H^{(1)}\right)^2.$$

Finally, $\mathrm{Var}(r_t(X_1)) \leq \mathbb{E}[f_t^2(X_1)] \leq \|f\|_\infty^2 := v^{(1)}$. Replacing the quantities $M_1^{(1)}$, $H^{(1)}$ and $v^{(1)}$ by the values derived above, Inequality (23) becomes

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\left(\sup_{t\in\mathcal{S}(m')}\left(\nu_n^{(1)}(t)\right)^2-\frac{p_1(m')}{3}\right)_+\right]$$
$$\leq\frac{4}{K_1}\|f\|_\infty\left\{\frac{1}{n}\sum_{m'\in\mathcal{M}_n}\exp\left(-\bar{k}D_{m'}\right)+\frac{49\phi_0^2\|f\|_\infty}{K_1C^2(\delta)}\frac{1}{n^2}\sum_{m'\in\mathcal{M}_n}D_{m'}\exp\left(-\bar{\bar{k}}\sqrt{n}\right)\right\},$$

with $\bar{k}$ and $\bar{\bar{k}}$ two constants (independent of m' and n) and $p_1(m')=3\times2(1+2\delta)\left(H^{(1)}\right)^2$. Therefore, using that the cardinality of $\mathcal{M}_n$ is bounded by n and also that $D_{m'}\leq n$, the following upper bound holds, for C_1 a constant,

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\left(\sup_{t\in\mathcal{S}(m')}\left(\nu_n^{(1)}(t)\right)^2-\frac{p_1(m')}{3}\right)_+\right]\leq\frac{C_1}{n}. \tag{24}$$

Similarly, we apply Inequality (23) to the second process $\nu_n^{(2,1)}$. We replace r by $r_t:(\varepsilon,x)\mapsto\varepsilon\mathbf{1}_{\varepsilon\leq\kappa_n}t\circ G(x)$, and $\xi_i=(\varepsilon_i,X_i)$. Thus we compute

$$M_1^{(2)}=\kappa_n\phi_0\sqrt{D_{m'}},\quad H^{(2)}=\phi_0\sigma\sqrt{\frac{D_{m'}}{n}},\quad v^{(2)}=\sigma^2.$$

With $p_2(m')=3\times2(1+2\delta)\left(H^{(2)}\right)^2$, we get

$$\mathbb{E}\left[\left(\sup_{t\in\mathcal{S}(m')}\left(\nu_n^{(2,1)}(t)\right)^2-\frac{p_2(m')}{3}\right)_+\right]\leq\frac{C_2}{n}, \tag{25}$$

for C_2 a constant.
Finally, we look for an upper bound for the process $\nu_n^{(2,2)}$. We can not apply the concentration inequality, because it is not bounded. However, following the same line as in computations above, we write

$$\mathbb{E}\left[\sup_{t\in\mathcal{S}(m')}\left(\nu_n^{(2,2)}(t)\right)^2\right]\leq\sum_{j=1}^{D_{m'}}\mathbb{E}\left[\left(\nu_n^{(2,2)}(\varphi_j)^2\right)\right]\leq\frac{1}{n}\mathbb{E}\left[|\varepsilon_1|^{2+p}\mathbf{1}_{|\varepsilon_1|>\kappa_n}\right]\phi_0^2\frac{\kappa_n^{-p}D_{m'}}{n}\leq\frac{C_3}{n}, \tag{26}$$

since κ_n is defined by (21) and $p>4$.
We conclude the proof of Proposition 3 by gathering in the equation (22) the three inequalities (24), (25), and (26).

□

We also set the following technical lemma, which will be useful several times, with ν an empirical process.

Lemma 5. *Let $\nu:L^2([0;1])\mapsto\mathbb{R}$ be a linear functional. Let also m be an index of the collection $\mathcal{M}_n$. Then,*

$$\sup_{t\in\mathcal{S}(m)}\nu^2(t)=\sum_{j=1}^{D_m}\nu^2(\varphi_j).$$

Proof of Lemma 5.
If t belongs to $\mathcal{S}(m)$, it can be written $t=\sum_{j=1}^{D_m}b_j\varphi_j$, with $\sum_{j=1}^{D_m}b_j^2=1$. Thus, by the linearity of ν and the Cauchy-Schwarz Inequality,

$$\nu^2(t)=\left(\sum_{j=1}^{D_m}b_j\nu(\varphi_j)\right)^2\leq\sum_{j=1}^{D_m}\nu^2(\varphi_j).$$

This leads to $\sup_{t\in\mathcal{S}(m)}\nu^2(t)\leq\sum_{j=1}^{D_m}\nu^2(\varphi_j)$. The equality is obtained by choosing $t=\sum_{j=1}^{D_m}b_j\varphi_j\in L^2([0;1])$, with $b_j=\nu(\varphi_j)/(\sum_{k=1}^{D_m}\nu^2(\varphi_k))$.

□

5.2. **Proof of Theorem 1.** We only study the estimator selected with the new GL method, that is $\tilde{f}_2^G$. However, the following proof gives all the ingredients to deal with the other estimator, $\tilde{f}_1^G$ (see a typical sketch in Brunel *et al.* [11], proof of Theorem 3.1 page 185). Moreover, one can refer to [12] to get all the details.

5.2.1. *Main part of the proof.* In all the proofs, the letter C denotes a nonnegative real that may change from line to line. For the sake of simplicity, we denote in this section by $V=V^G$, $A=A^G$, $\hat{m}=\hat{m}^{(2),G}$. Let S_m be a fixed model in the collection indexed by $\mathcal{M}_n$. We decompose the loss of the estimator as follows:

$$\begin{aligned}\left\|\tilde{f}_2^G-f\right\|_g^2 &= \left\|\hat{h}^G_{\hat{m}}-h\right\|^2,\\ &\leq 3\left\|\hat{h}^G_{\hat{m}}-\hat{h}^G_{m\wedge\hat{m}}\right\|^2+3\left\|\hat{h}^G_{m\wedge\hat{m}}-\hat{h}^G_{m}\right\|^2+3\left\|\hat{h}^G_{m}-h\right\|^2.\end{aligned}$$

By definition of A and $\hat{m}$,

$$\begin{aligned}\left\|\tilde{f}_2^G-f\right\|_g^2 &\leq 3\left(A(m)+V(\hat{m})\right)+3\left(A(\hat{m})+V(m)\right)+3\left\|\hat{h}^G_{m}-h\right\|^2,\\ &\leq 6\left(A(m)+V(m)\right)+3\left\|\hat{h}^G_{m}-h\right\|^2.\end{aligned}$$

We have already bounded the risk of the estimator on a fixed model (see Section 2.2.2, Inequalities (6) and (8)): $\mathbb{E}[\|\hat{h}^G_m-h\|^2]\leq\phi_0^2\mathbb{E}[Y_1^2]D_m/n+\|h_m-h\|^2$. Therefore we get

$$\mathbb{E}\left[\left\|\tilde{f}_2^G-f\right\|_g^2\right]\leq 6\mathbb{E}\left[A(m)\right]+6V(m)+3\phi_0^2\mathbb{E}\left[Y_1^2\right]\frac{D_m}{n}+3\left\|h_m-h\right\|^2.$$

Next, we have to control the term $A(m)$: we use the following lemma, proved just below, to conclude.

Lemma 6. *Under the assumptions of Theorem 1, there exists a constant $C>0$ depending on ϕ_0^2, $\|f\|_\infty$, $\mathbb{E}[f^2(X_1)]$, σ^2, $\mathbb{E}[|\varepsilon_1|^p]$ such that, for each index $m\in\mathcal{M}_n$,*

$$\mathbb{E}\left[A(m)\right]\leq\frac{C}{n}+12\left\|h_m-h\right\|^2.$$

□

5.2.2. *Proof of Lemma 6.* For each index $m\in\mathcal{M}_n$, we decompose,

$$\left\|\hat{h}^G_{m'}-\hat{h}^G_{m\wedge m'}\right\|^2\leq 3\left\|\hat{h}^G_{m'}-h_{m'}\right\|^2+3\left\|h_{m'}-h_{m\wedge m'}\right\|^2+3\left\|h_{m\wedge m'}-\hat{h}^G_{m\wedge m'}\right\|^2.$$

Thus we have

$$\begin{aligned}A(m) &\leq 3\max_{m'\in\mathcal{M}_n}\left[\left\|\hat{h}^G_{m'}-h_{m'}\right\|^2-\frac{V(m')}{6}\right]_+ +3\max_{m'\in\mathcal{M}_n}\left[\left\|h_{m\wedge m'}-\hat{h}^G_{m\wedge m'}\right\|^2-\frac{V(m')}{6}\right]_+\\ &\quad +3\max_{m'\in\mathcal{M}_n}\left\|h_{m'}-h_{m\wedge m'}\right\|^2,\\ &:= 3\left(T_a+T_b^m+T_c^m\right),\end{aligned}\tag{27}$$

and study the terms of the above decomposition.

Upper-bound for T_a

We simplify roughly the problem by writing first

$$\mathbb{E}\left[T_a\right] \leq \sum_{m' \in \mathcal{M}_n} \mathbb{E}\left[\left\{\left\|\hat{h}^G_{m'} - h_{m'}\right\|^2 - \frac{V(m')}{6}\right\}_+\right].$$

Let us notice that

$$\|\hat{h}^G_{m'} - h_{m'}\|^2 = \sum_{j=1}^{D_{m'}} (\hat{a}^G_j - a_j)^2 = \sum_{j=1}^{D_{m'}} \nu_n^2(\varphi_j), \tag{28}$$

with ν_n the empirical process defined by (20). By Lemma 5, this last quantity is equal to $\sup_{t \in \mathcal{S}(m')} \nu_n^2(t)$. Consequently, $\mathbb{E}[T_a] \leq \sum_{m' \in \mathcal{M}_n} \mathbb{E}[\{\sup_{t \in \mathcal{S}(m')} \nu_n^2(t) - \frac{V(m')}{6}\}_+]$. We apply then Proposition 3: the latter is bounded by C/n, for the choice $V(m') = 6 \times p(m')$, which means the choice of $c_2 = 36(1+2\delta)$ in the definition (10).

Upper-bound for T_b^m

To study this term, we write, distinguish whether $m' \leq m$ or $m' > m$,

$$\begin{aligned} T_b^m &= \max\left(\max_{\substack{m' \in \mathcal{M}_n \\ m' \leq m}} \left\{\left\|h_{m'} - \hat{h}^G_{m'}\right\|^2 - \frac{V(m')}{6}\right\}_+, \max_{\substack{m' \in \mathcal{M}_n \\ m' > m}} \left\{\left\|h_m - \hat{h}^G_m\right\|^2 - \frac{V(m')}{6}\right\}_+\right), \\ &\leq \max\left(T_a, \left\{\left\|h_m - \hat{h}^G_m\right\|^2 - \frac{V(m)}{6}\right\}_+\right) \leq T_a + \left\{\left\|h_m - \hat{h}^G_m\right\|^2 - \frac{V(m)}{6}\right\}_+, \end{aligned}$$

using $-V(m') \leq -V(m)$ for $m' > m$. The last computation proves that $\mathbb{E}[T_a] \leq C/n$ and the same bound holds for the second term, as a consequence of Proposition 3. Finally, $\mathbb{E}[T_b^m] \leq C/n$.

Upper-bound for T_c^m

This term is a bias term. We notice that

$$T_c^m = \max_{\substack{m' \in \mathcal{M}_n \\ m \leq m'}} \|h_{m'} - h_m\|^2 \leq 2 \max_{\substack{m' \in \mathcal{M}_n \\ m \leq m'}} \|h_{m'} - h\|^2 + 2\|h - h_m\|^2.$$

But assuming $m \leq m'$, we have $S_m \subset S_{m'}$, thus, the orthogonal projections h_m and $h_{m'}$ of h onto S_m and $S_{m'}$ satisfy $\|h_{m'} - h\|^2 \leq \|h_m - h\|^2$. So we have $T_c^m \leq 4\|h_m - h\|^2$, which conclude the proof.

□

5.3. **Proof of Theorem 2.**

5.3.1. *Notations, and properties of the empirical distribution function.* Let us introduce some useful tools for the sequel. Denoting by $U_{-i} = G(X_{-i})$ the uniform variable associated to X_{-i}, for any $i \in \{1, \ldots, n\}$, we define the empirical distribution function

$$\hat{U}_n : u \mapsto \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{U_{-i} \leq u}. \tag{29}$$

The following equality holds for any coefficient $\hat{a}^{\hat{G}}_j$ of our estimator (see equation (5)):

$$\mathbb{E}\left[\hat{a}^{\hat{G}}_j \,|(X_{-l})_l\right] = \int_0^1 \left(f \circ G^{-1}\right)(u) \left(\varphi_j \circ \hat{U}_n\right)(u) du. \tag{30}$$

Moreover, we will use several inequalities to control the deviations of the empirical c.d.f. $\hat{U}_n$ or $\hat{G}_n$. Recall that the random variable $\|\hat{G}_n - G\|_\infty$ has the same probability distribution as the norm $\|\hat{U}_n - id\|_\infty$ where we denote by $\|\hat{U}_n - id\|_\infty = \sup_{u\in\mathbb{R}} |\hat{U}_n(u) - u|$. The first inequality is the one of Dvoretzky-Kiefer-Wolfowitz (see Dvoretzky *et al.* [17]):

$$\mathbb{P}\left(\left\|\hat{U}_n - id\right\|_\infty \geq \lambda\right) \leq K \exp\left(-2n\lambda^2\right), \tag{31}$$

for any $\lambda > 0$ and a constant K.
By integration, we deduce the following bounds:

- for any integer $p > 0$, there exists a constant $C_p > 0$ such that

$$\mathbb{E}\left[\left\|\hat{U}_n - id\right\|_\infty^p\right] \leq \frac{C_p}{n^{p/2}}, \tag{32}$$

- for any $\kappa > 0$, for any integer $p \geq 2$, there exists a constant C such that

$$\mathbb{E}\left[\left(\left\|\hat{U}_n - id\right\|_\infty^p - \kappa\frac{\ln^{p/2}(n)}{n^{p/2}}\right)_+\right] \leq Cn^{-c(p,\kappa)}, \text{ with } c(p,\kappa) = 2^{\frac{2-p}{p}}\kappa^{2/p}. \tag{33}$$

Moreover,

$$\mathbb{E}\left[\left(\left\|\hat{U}_n - id\right\|_\infty^2 - \kappa\frac{\ln(n)}{n}\right)^2\right] \leq Cn^{-2-2\kappa}. \tag{34}$$

5.3.2. *Preliminary lemmas.* As we have done for Theorem 1, we prove the result for the most original estimator, that is $\tilde{f}_2$ (the proof for the other estimator can be found in [12]). The proof follows almost the same line as the one of Theorem 1. However, further technicalities are required, consequence of the replacement of G by $\hat{G}_n$. Let us introduce some useful notations. We denote by C a numerical constant, which may vary from line to line. In this section, we denote also the estimator by $\hat{f}_{\hat{m}}^{\hat{G},\hat{G}} = \hat{h}_{\hat{m}}^{\hat{G}} \circ \hat{G}_n$ (with shortened $\hat{m}^{(2)}$ in $\hat{m}$), and coherently:

$$\hat{f}_{\hat{m}}^{\hat{G},G} = \hat{h}_{\hat{m}}^{\hat{G}} \circ G,$$

which is an intermediate between the two estimators $\hat{f}_{\hat{m}}^{\hat{G},\hat{G}}$ and $\hat{f}_{\hat{m}}^{G,G}$. We will also use this notation for fixed index $m \in \mathcal{M}_n$. To bound the risk of the target estimator, the following quantities are useful, for any index m:

$$\begin{aligned}
T_0^m &= \|f - f_m^G\|_g^2 + \|f_m^G - \hat{f}_m^{G,G}\|_g^2,\\
T_1^m &= \left\|\hat{f}_m^{G,G} - \hat{f}_m^{\hat{G},G} - \mathbb{E}\left[\hat{f}_m^{G,G} - \hat{f}_m^{\hat{G},G}\,|(X_{-l})_l\right]\right\|_g^2,\\
T_2^m &= \left\|\hat{f}_m^{\hat{G},G} - \hat{f}_m^{\hat{G},\hat{G}} - \mathbb{E}\left[\hat{f}_m^{\hat{G},G} - \hat{f}_m^{\hat{G},\hat{G}}\,|(X_{-l})_l\right]\right\|_g^2,\\
T_3^m &= \left\|\mathbb{E}\left[\hat{f}_m^{G,G} - \hat{f}_m^{\hat{G},G}\,|(X_{-l})_l\right]\right\|_g^2,\ T_4^m = \left\|\mathbb{E}\left[\hat{f}_m^{\hat{G},G} - \hat{f}_m^{\hat{G},\hat{G}}\,|(X_{-l})_l\right]\right\|_g^2.
\end{aligned} \tag{35}$$

They are such that $\mathbb{E}[\|\hat{f}_m^{\hat{G},\hat{G}} - f\|_g^2] \leq \sum_{l=0}^4 T_l^m$. Let us remark that T_0^m is the bias-variance decomposition for the risk of an estimator $\hat{f}_m^{G,G}$ (on the fixed model S_m). The bound for its expectation is given by Inequalities (6) and (8). The lemmas below give bounds for the other terms.

Lemma 7. *Assuming that the models are trigonometric, there exists a constant $C > 0$ (depending on $\|\varphi_2'\|_\infty$ and $\mathbb{E}[Y_1^2]$) such that*

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n} T_1^{m'}\right] \leq C\frac{D_{m_{\max}}^3}{n^2}$$

If $D_{m_{\max}} = O(n^{1/2})$ in particular,

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n} T_1^{m'}\right] \leq C\frac{D_{m_{\max}}}{n}.$$

Lemma 8. *Assuming that the models are trigonometric, that $D_{m_{\max}} = O(n^{1/3})$ and that there exists a real-number $p > 8/3$ such that $\mathbb{E}\left[|\varepsilon_1|^{2+p}\right] < \infty$, there exists a constant $C > 0$ (depending on $\|\varphi_2'\|_\infty$ and $\mathbb{E}[Y_1^2]$) such that*

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_2^{m'} - V_2(m')\right)_+\right] \leq C\frac{\ln(n)}{n},$$

with $V_2(m') = \kappa\kappa' D_{m'}^4 \ln^2(n)/n^2$, and $\kappa' = 7/3$, $\kappa = 96\phi_0^2\mathbb{E}[Y_1^2]\|\varphi_2'\|_\infty^2$.

Assuming that $D_{m'} = O((n/\ln(n)^2)^{1/3})$, we get

$$V_2(m') \leq \kappa\kappa'\frac{D_{m'}}{n} := V_2^{bis}(m').$$

The result of Lemma 8 holds with V_2^{bis} in place of V_2.

Lemma 9. *Assuming that the models are trigonometric, that $D_{m_{\max}} = O(n^{1/3}/\ln(n))$, and that $h \in \mathcal{C}^1([0;1])$, there exists a constant $C > 0$ (depending on $\|\varphi_2'\|_\infty$, $\|\varphi_2^{(3)}\|_\infty$, $\|h\|$, $\|h'\|$, $\mathbb{E}[Y_1^2]$) such that, for all $m \in \mathcal{M}_n$,*

$$\mathbb{E}[T_3^m] \quad \leq \quad C\left(\frac{D_m}{n} + \frac{D_m^4}{n^2} + \frac{D_m^7}{n^3}\right).$$

Moreover, the following inequality holds, for $p_{m'} = m'$ or $p_{m'} = m \wedge m'$:

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{p_{m'},b} - V_3(m')\right)_+\right] \leq \frac{C}{n}.$$

with $V_3(m') = k_3 D_{m'}/n$, and k_3 a numerical constant depending only (and linearly) on $\mathbb{E}[Y_1^2]$.

In particular, if $D_m = O(n^{1/3})$, the first inequality leads to $\mathbb{E}[T_3^m] \leq CD_m/n$.

Lemma 10. *Assuming that the models are trigonometric, that $D_{m_{\max}} = O(n^{1/3}/\ln(n))$, and that $h \in W_{per}^{2,1}(L)$ ($L > 0$), there exists a constant $C > 0$ (depending on $\|\varphi_2'\|_\infty$, $\|\varphi_2^{(3)}\|_\infty$, $\|h\|$, $\|h'\|$, $\mathbb{E}[Y_1^2]$) such that, for all $m \in \mathcal{M}_n$, $n \geq n_0 = \exp\left(\|h'\|^2\right)$,*

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_4^{m'} - V_4(m')\right)_+\right] \leq C\frac{\ln(n)}{n},$$

with $V_4(m') = k_4 D_{m'}/n$, and k_4 a numerical constant depending only (and linearly) on $\mathbb{E}[Y_1^2]$.

Notice that it is also possible to obtain the result for any $n \in \mathbb{N}$. But the price to pay is a penalty V_4 depending on $\|h'\|^2$.

5.3.3. *Main part of the proof.* Let S_m be a fixed model in the collection indexed by $\mathcal{M}_n$. To recover the framework of the proof of Theorem 1, we begin with the decomposition

$$\begin{aligned}\left\|\hat{f}_{\hat{m}}^{\hat{G},\hat{G}}-f\right\|_g^2 &\leq 3\left\|\hat{f}_{\hat{m}}^{\hat{G},\hat{G}}-\hat{f}_{\hat{m}}^{\hat{G},G}-\mathbb{E}\left[\hat{f}_{\hat{m}}^{\hat{G},\hat{G}}-\hat{f}_{\hat{m}}^{\hat{G},G}\,|(X_{-l})_l\right]\right\|_g^2\\ &\quad+3\left\|\mathbb{E}\left[\hat{f}_{\hat{m}}^{\hat{G},\hat{G}}-\hat{f}_{\hat{m}}^{\hat{G},G}\,|(X_{-l})_l\right]\right\|_g^2+3\left\|\hat{f}_{\hat{m}}^{\hat{G},G}-f\right\|_g^2,\\ &= 3T_2^{\hat{m}}+3T_4^{\hat{m}}+3\left\|\hat{h}_{\hat{m}}^{\hat{G}}-h\right\|^2.\end{aligned}$$

Thus, we can introduce A and V, in the last term, in a similar way as previously:
$\left\|\hat{h}_{\hat{m}}^{\hat{G}}-h\right\|^2$

$$\begin{aligned}&\leq 3\left\|\hat{h}_{\hat{m}}^{\hat{G}}-\hat{h}_{m\wedge\hat{m}}^{\hat{G}}\right\|^2+3\left\|\hat{h}_{m\wedge\hat{m}}^{\hat{G}}-\hat{h}_m^{\hat{G}}\right\|^2+3\left\|\hat{h}_m^{\hat{G}}-h\right\|^2,\\ &\leq 3\left(A(m)+V\left(\hat{m}\right)\right)+3\left(A\left(\hat{m}\right)+V\left(\hat{m}\right)\right)+3\left\|\hat{h}_m^{\hat{G}}-h\right\|^2,\\ &= 3\left(A(m)+2V\left(m\right)\right)+3\left(A\left(\hat{m}\right)+2V\left(\hat{m}\right)\right)+3\left\|\hat{h}_m^{\hat{G}}-h\right\|^2-3V\left(\hat{m}\right)-3V\left(m\right),\\ &\leq 6\left(A(m)+2V(m)\right)-2V\left(\hat{m}\right)+3\left\|\hat{h}_m^{\hat{G}}-h\right\|^2,\end{aligned}$$

using the definition of $\hat{m}$. The last term of this decomposition is bounded by:

$$\left\|\hat{h}_m^{\hat{G}}-h\right\|^2 = \left\|\hat{f}_m^{\hat{G},G}-f\right\|_g^2\leq 3T_1^m+3T_3^m+3T_0^m,$$

where T_l^m $(l=0,1,3)$ are defined by (35). As a result, we get

$$\begin{aligned}\left\|\hat{f}_{\hat{m}}^{\hat{G},\hat{G}}-f\right\|_g^2 &\leq 3T_2^{\hat{m}}+3T_4^{\hat{m}}-3\times 2V(\hat{m})+3\times 6\left(A(m)+V(m)\right)\\ &\quad+3\times 3\times\left(3T_1^m+3T_3^m+3T_0^m\right).\end{aligned}$$

Therefore, it follows from Inequalities (6) and (8) that

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{f}_{\hat{m}}^{\hat{G},\hat{G}}-f\right\|_g^2\right] &\leq 18\left(\mathbb{E}\left[A(m)\right]+V(m)\right)+3\mathbb{E}\left[\left(T_2^{\hat{m}}-V(\hat{m})\right)_+\right]+3\mathbb{E}\left[\left(T_4^{\hat{m}}-V(\hat{m})\right)_+\right]\\ &\quad+\mathbb{E}\left[T_1^m\right]+\mathbb{E}\left[T_3^m\right]+27\phi_0^2\mathbb{E}\left[Y_1^2\right]\frac{D_m}{n}+27\left\|f-f_m^G\right\|_g^2.\end{aligned}$$

A bound for $A(m)$ is given by the following lemma, whose proof is deferred to Section 5.3.4.

Lemma 11. *Under the assumptions of Theorem 1, there exists a constant $C>0$ depending on $\|\varphi_2^{(l)}\|$ $(l=1,3)$, $\|h\|$, $\|h'\|$, and $\mathbb{E}[Y_1^2]$, such that, for each index $m\in\mathcal{M}_n$,*

$$\begin{aligned}\mathbb{E}\left[A(m)\right] &\leq 12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m'}-\frac{V(m')}{48}\right)_+\right]+12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m\wedge m'}-\frac{V(m')}{48}\right)_+\right]\\ &\quad+12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}T_1^{m'}\right]+12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}T_1^{m\wedge m'}\right]+12\left\|f_m^G-f\right\|_g^2+\frac{C}{n}.\end{aligned}$$

Then we get

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{f}_{\hat{m}}^{\hat{G},\hat{G}}-f\right\|_g^2\right] &\leq C\left(\mathbb{E}\left[\max_{m'\in\mathcal{M}_n} T_1^{m'}\right]+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n} T_1^{m\wedge m'}\right]+\mathbb{E}\left[T_1^m\right]\right.\\ &+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m'}-\frac{V(m')}{48}\right)_+\right]+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m\wedge m'}-\frac{V(m')}{48}\right)_+\right]\\ &\left.+\mathbb{E}\left[T_3^m\right]+\mathbb{E}\left[\left(T_2^{\hat{m}}-V(\hat{m})\right)_+\right]+\mathbb{E}\left[\left(T_4^{\hat{m}}-V(\hat{m})\right)_+\right]\right)\\ &+C\left(\phi_0^2\mathbb{E}\left[Y_1^2\right]\frac{D_m}{n}+\left\|f-f_m^G\right\|_g^2+\frac{1}{n}\right).\end{aligned}$$

It remains to study the terms $T_l^m, l=1,\dots,4$. Bounding $(T_l^{\hat{m}}-V(\hat{m}))_+\leq\max_{m'}(T_l^{m'}-V(m'))_+$ $(l=2,4)$, it is enough to apply Lemmas 7 to 10 to conclude: we have just to choose the constant in the definition of V larger than the ones of V_l $(l=2,3,4)$.

□

5.3.4. *Proof of Lemma 11.* The following proof is close to the proof of Lemma 6. Fix an index $m'\in\mathcal{M}_n$. We split

$$\left\|\hat{h}_{m'}^{\hat{G}}-\hat{h}_{m\wedge m'}^{\hat{G}}\right\|^2\leq 3\left\|\hat{h}_{m'}^{\hat{G}}-h_{m'}\right\|^2+3\left\|h_{m'}-h_{m\wedge m'}\right\|^2+3\left\|h_{m\wedge m'}-\hat{h}_{m\wedge m'}^{\hat{G}}\right\|^2.$$

Relation (28) still holds for an other empirical process, and by applying Lemma 5, we have, for $p=m'$ or $p=m\wedge m'$ $\|h_p-\hat{h}_p^{\hat{G}}\|^2=\sup_{t\in\mathcal{S}(p)}\tilde{\nu}_n(t)^2$, with, for $t\in L^2([0;1])$,

$$\tilde{\nu}_n(t)=\frac{1}{n}\sum_{i=1}^n Y_i\left(t\circ\hat{G}_n\right)(X_i)-\mathbb{E}\left[Y_i\left(t\circ G\right)(X_i)\right].$$

We split $\tilde{\nu}_n$ into $\tilde{\nu}_n=\nu_n+R_n$, with

$$R_n(t)=\frac{1}{n}\sum_{i=1}^n Y_i t(\hat{G}_n(X_i)-G(X_i)).$$

This yields to $\tilde{\nu}_n^2\leq 2\nu_n^2+2R_n^2$. If t belongs to $\mathcal{S}(p)$, $t=\sum_{j=1}^{D_p}\theta_j\varphi_j$ with $\sum_{j=1}^{D_p}\theta_j^2=1$, so that

$$\begin{aligned}\sup_{t\in\mathcal{S}(p)} R_n^2(t) &= \sup_{\substack{\theta\in\mathbb{R}^p\\ \sum_j\theta_j^2=1}}\left(\sum_{j=1}^{D_p}\theta_j\frac{1}{n}\sum_{i=1}^n Y_i\varphi_j(\hat{G}_n(X_i)-G(X_i))\right)^2,\\ &= \sup_{\substack{\theta\in\mathbb{R}^p\\ \sum_j\theta_j^2=1}}\left(\sum_{j=1}^{D_p}\theta_j\left(\hat{a}_j^{\hat{G}}-\hat{a}_j^G\right)\right)^2=\sum_{j=1}^{D_p}\left(\hat{a}_j^{\hat{G}}-\hat{a}_j^G\right)^2,\end{aligned}$$

by using the same arguments as in the proof of Lemma 5. Introducing the conditional expectation of $\hat{a}_j^{\hat{G}}-\hat{a}_j^G$, we note that $\sup_{t\in\mathcal{S}(p)}R_n^2(t)\leq 2T_1^p+2T_3^p$. We obtain,

$$\left\|h_p-\hat{h}_p^{\hat{G}}\right\|^2\leq 2\sup_{t\in\mathcal{S}(p)}(\nu_n(t))^2+4T_1^p+4T_3^p.$$

and thus, substracting $V(m')$ and taking expectation, this yields
$\mathbb{E}\left[A(m)\right]$

$$\begin{aligned}
\leq\ & 6\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m')}\left(\nu_n(t)\right)^2-\frac{V(m')}{24}\right)_+\right]+6\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m\wedge m')}\left(\nu_n(t)\right)^2-\frac{V(m')}{24}\right)_+\right]\\
&+12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m'}-\frac{V(m')}{48}\right)_+\right]+12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m\wedge m'}-\frac{V(m')}{48}\right)_+\right]\\
&+12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}T_1^{m'}\right]+12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}T_1^{m\wedge m'}\right]+3\max_{m'\in\mathcal{M}_n}\left\|h_{m'}-h_{m\wedge m'}\right\|^2.
\end{aligned}$$

The last term is denoted by T_c^m in (27) and proved to be bounded by $4\left\|h_m-h\right\|^2$ (see the proof of Lemma 6). Moreover, applying Proposition 3 yields to

$$\begin{aligned}
\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m')}\left(\nu_n(t)\right)^2-p(m')\right)_+\right]&\leq\frac{C}{n},\\
\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m\wedge m')}\left(\nu_n(t)\right)^2-p(m')\right)_+\right]&\leq\frac{C}{n},
\end{aligned}$$

using $-p(m')\leq-p(m\wedge m')$ (remember that $p(m')=C\phi_0^2\mathbb{E}\left[Y_1^2\right]D_{m'}/n$). By gathering the last bounds, and noting that the constant c_v' (in the definition of $V(m')$) can be chosen larger than the one of $p(m')$, we obtain the result of Lemma 11.

□

5.3.5. *Proof of Lemmas 7 to 10.* In this section we state upper bounds for T_l^m, $l=1,\dots,4$ (see (35)). Recall that $m_{\max}$ is the index of the largest model in the collection. Notice that $D_{m_{\max}}\geq m_{\max}$, since we work with the trigonometric model. Recall also that we denote by a_j the Fourier coefficients of the function h, that is, $h_m=\sum_{j=1}^{D_m}a_j\varphi_j$, where h_m is the orthogonal projection on the space S_m, $m\in\mathcal{M}_n$.
The sketch of all the proof can be described by the following cases:

- ($\mathcal{A}$) Some of the terms are less than CD_m/n, under the constraint $D_m\leq Cn^{1/3}/\ln(n)$, and so we do not need to center them. For example, they involve expectations of form $\mathbb{E}[\sum_{j=1}^{D_m}(\varphi_j(G(X_1))-\varphi_j(\hat{G}_n(X_1)))^2]$. By using a Taylor formula, we come down to terms of form $\sum_{j=1}^{D_m}(\varphi_j^{(k)})^2\mathbb{E}[\|\hat{U}_n-id\|_\infty^{2k}]$ (k an integer), and bound them with Inequality (32). This is the case for T_1^m (Lemma 7), T_3^m, first inequality (first part of Lemma 9), and for some terms of the decomposition of T_4^m (see proof of Lemma 10).
- ($\mathcal{B}$) The other terms have to be centered to be negligible. There are then two possibilities:
  - ($\mathcal{B}_1$) The first one is to make emerge the supremum of an empirical process (with Lemma 5) and the to use the Talagrand Inequality (23). This is the case for a part of T_2^m and T_3^m (Lemmas 8 and 9, second inequality).
  - ($\mathcal{B}_2$) The second is to bound these terms by quantity of form $C(D_m)\|\hat{U}_n-id\|_\infty^k$ (k an integer, $C(D_m)$ a constant depending on D_m), and to use Inequality (33) or (34).This is the case for the other parts of T_2^m and T_3^m (Lemmas 8 and 9, second inequality).

For sake of conciseness, we do not detail all of the proofs, especially the ones which follow a line already described. However, the lector can find all the details in [12].

• ***Proof of Lemma 7***. Let us note that we can write

$$T_1^{m'} = \left\| \sum_{j=1}^{D_{m'}} \left(\hat{a}_j^G - \hat{a}_j^{\hat{G}} - \mathbb{E}[\hat{a}_j^G - \hat{a}_j^{\hat{G}} | (X_{-l})_l] \right) (\varphi_j \circ G) \right\|_g^2 .$$

As the functions φ_j are orthonormal, it becomes

$$T_1^{m'} = \sum_{j=1}^{D'_m} \left(\hat{a}_j^G - \hat{a}_j^{\hat{G}} - \mathbb{E}\left[\hat{a}_j^G - \hat{a}_j^{\hat{G}} | (X_{-l})_l \right] \right)^2 .$$

This shows that $T_1^{m'} \leq T_1^{m_{\max}}$ and $\mathbb{E}[\max_{m'} T_1^{m'}] \leq \mathbb{E}[T_1^{m_{\max}}]$. Thus it is sufficient to bound $\mathbb{E}[T_1^{m_{\max}}]$. Now, $\mathbb{E}[T_1^{m_{\max}} | (X_{-l})_l] = \sum_{j=1}^{D_{m_{\max}}} \mathrm{Var}(\hat{a}_j^G - \hat{a}_j^{\hat{G}} | (X_{-l})_l)$, where $\mathrm{Var}(.|(X_{-l})_l)$ is the conditional variance with respect to the sample $(X_{-l})_{l \in \{1,\ldots,n\}}$ (we denote by a similar notation the conditional expectation in the sequel). We work out it, for any index $j \in \{1, \ldots, D_{m_{\max}}\}$,

$$\begin{aligned} \mathrm{Var}\left(\hat{a}_j^G - \hat{a}_j^{\hat{G}} | (X_{-l})_l \right) &= \frac{1}{n} \mathrm{Var}\left(Y_1 \left(\varphi_j \left(G(X_1) \right) - \varphi_j \left(\hat{G}_n(X_1) \right) \right) | (X_{-l})_l \right), \\ &\leq \frac{1}{n} \mathbb{E}\left[f(X_1)^2 \left(\varphi_j \left(G(X_1) \right) - \varphi_j \left(\hat{G}_n(X_1) \right) \right)^2 | (X_{-l})_l \right] \\ &\quad + \frac{\sigma^2}{n} \mathbb{E}\left[\left(\varphi_j \left(G(X_1) \right) - \varphi_j \left(\hat{G}_n(X_1) \right) \right)^2 | (X_{-l})_l \right]. \end{aligned}$$

We use the mean value theorem: $(\varphi_j(G(X_1)) - \varphi_j(\hat{G}_n(X_1)))^2 \leq \|\varphi_j'\|_\infty^2 \|G - \hat{G}_n\|_\infty^2$. This leads to

$$\begin{aligned} \mathbb{E}\left[T_1^{m_{\max}} | (X_{-l})_l \right] &\leq \frac{1}{n} \left(\mathbb{E}\left[f^2(X_1) \right] + \sigma^2 \right) \sum_{j=1}^{D_{m_{\max}}} \left\| \varphi_j' \right\|_\infty^2 \left\| G - \hat{G}_n \right\|_\infty^2, \\ &= \frac{1}{n} \mathbb{E}\left[Y_1^2 \right] \sum_{j=1}^{D_{m_{\max}}} \left\| \varphi_j' \right\|_\infty^2 \left\| \hat{U}_n - id \right\|_\infty^2 . \end{aligned}$$

The sum is bounded by $D_{m_{\max}} \times (D_{m_{\max}} \|\varphi_2'\|_\infty)^2$, and we apply Inequality (32) with $p = 2$, to conclude $\mathbb{E}[T_1^{m_{\max}}] \leq C_2 \|\varphi_2'\|_\infty^2 \mathbb{E}[Y_1^2] D_{m_{\max}}^3 / n^2$.

□

• ***Proof of Lemma 8***. Begining with $\mathbb{E}[\max_{m' \in \mathcal{M}_n} (T_2^{m'} - V_2(m'))_+] \leq \sum_{m' \in \mathcal{M}_n} \mathbb{E}[(T_2^{m'} - V_2(m'))_+]$, we have just to study this quantity for each index m'. We write

$$\begin{aligned} T_2^{m'} &= \int_{[a;b]} \left(\hat{h}_{m'}^{\hat{G}} \left(G(x) \right) - \hat{h}_{m'}^{\hat{G}} \left(\hat{G}_n(x) \right) - \mathbb{E}\left[\hat{h}_{m'}^{\hat{G}} \left(G(x) \right) - \hat{h}_{m'}^{\hat{G}} \left(\hat{G}_n(x) \right) | (X_{-l})_l \right] \right)^2 g(x) dx, \\ &= \int_{[0;1]} \left\{ \sum_{j=1}^{D_{m'}} \left(\hat{a}_j^{\hat{G}} - \mathbb{E}\left[\hat{a}_j^{\hat{G}} | (X_{-l})_l \right] \right) \left(\varphi_j(u) - \varphi_j \left(\hat{U}_n(u) \right) \right) \right\}^2 du, \end{aligned}$$

We use the Cauchy-Schwarz Inequality, and by computations analogous of those of Lemma 7, we get

$$T_2^{m'} \leq \left\| \varphi_2' \right\|_\infty^2 D_{m'}^3 \left\| \hat{U}_n - id \right\|_\infty^2 \sum_{j=1}^{D_{m'}} \left(\hat{a}_j^{\hat{G}} - \mathbb{E}\left[\hat{a}_j^{\hat{G}} | (X_{-l})_l \right] \right)^2 .$$

Thus, we have

$$\begin{aligned}\mathbb{E}\left[\left(T_2^{m'}-V_2(m')\right)_+\right] &\leq D_{m'}^3\left\|\varphi_2'\right\|_\infty^2\mathbb{E}\left[\left(\sum_{j=1}^{D_{m'}}\left(\hat{a}_j^{\hat{G}}-\mathbb{E}\left[\hat{a}_j^{\hat{G}}\,|(X_{-l})_l\right]\right)^2\left\|\hat{U}_n-id\right\|_\infty^2\right.\right.\\ &\qquad\left.\left.-\frac{\kappa\kappa'}{\|\varphi_2'\|_\infty^2}\frac{D_{m'}}{n^2}\ln^2(n)\right)_+\right],\\ &\leq T_{2,a}^{m'}+T_{2,b}^{m'},\end{aligned}$$

denoting by

$$T_{2,a}^{m'}=D_{m'}^3\left\|\varphi_2'\right\|_\infty^2\mathbb{E}\left[\sum_{j=1}^{D_{m'}}\left(\hat{a}_j^{\hat{G}}-\mathbb{E}\left[\hat{a}_j^{\hat{G}}\,|(X_{-l})_l\right]\right)^2\left(\left\|\hat{U}_n-id\right\|_\infty^2-\kappa'\frac{\ln(n)}{n}\right)_+\right],$$

$$T_{2,b}^{m'}=D_{m'}^3\left\|\varphi_2'\right\|_\infty^2\kappa'\frac{\ln(n)}{n}\mathbb{E}\left[\left(\sum_{j=1}^{D_{m'}}\left(\hat{a}_j^{\hat{G}}-\mathbb{E}\left[\hat{a}_j^{\hat{G}}\,|(X_{-l})_l\right]\right)^2-\frac{\kappa}{\|\varphi_2'\|_\infty^2}\frac{D_{m'}}{n}\ln(n)\right)_+\right].$$

For the term $T_{2,a}^{m'}$, we obtain first

$$T_{2,a}^{m'}=D_{m'}^3\left\|\varphi_2'\right\|_\infty^2\sum_{j=1}^{D_{m'}}\mathbb{E}\left[\left(\hat{a}_j^{\hat{G}}-\mathbb{E}\left[\hat{a}_j^{\hat{G}}\,|(X_{-l})_l\right]\right)^4\right]^{1/2}\mathbb{E}\left[\left(\left\|\hat{U}_n-id\right\|_\infty^2-\kappa'\frac{\ln(n)}{n}\right)^2\right]^{1/2},$$

and bound roughly

$$\sum_{j=1}^{D_{m'}}\mathbb{E}\left[\left(\hat{a}_j^{\hat{G}}-\mathbb{E}\left[\hat{a}_j^{\hat{G}}\,|(X_{-l})_l\right]\right)^4\right] \leq 16\phi_0^4\mathbb{E}\left[Y_1^4\right]D_{m'}.$$

Gathering this bound with Inequality (33) leads to,

$$\sum_{m'\in\mathcal{M}_n}T_{2,a}^{m'} \leq C\sum_{m'\in\mathcal{M}_n}D_{m'}^4n^{-1-\kappa'}\leq Cn^{4/3-\kappa'}\leq Cn^{-1}$$

as soon as $D_{m'}\leq Cn^{1/3}$ and for $\kappa'=7/3$. For the second term $T_{2,b}^{m'}$, thanks to Lemma 5, we notice first that $\sum_{j=1}^{D_{m'}}\left(\hat{a}_j^{\hat{G}}-\mathbb{E}\left[\hat{a}_j^{\hat{G}}\,|(X_{-l})_l\right]\right)^2=\sup_{t\in\mathcal{S}(m')}\bar{\nu}_n^2(t)$, with, for $t\in L^2([0;1])$,

$$\bar{\nu}_n(t)=\frac{1}{n}\sum_{i=1}^n Y_it\left(\hat{G}_n(X_i)\right)-\mathbb{E}\left[Y_it\left(\hat{G}_n(X_i)\right)|(X_{-l})_l\right],$$

a process which is centered conditionally to the sample $(X_{-l})_l$. We must now bound its deviations, exactly as we bound the one of the process ν_n, in the proof of Proposition 3, but conditionally to the variables X_{-l}. Let us just recall the sketch of the proof: we split $\bar{\nu}_n$ in three parts, taking into account that $Y_i=f(X_i)+\varepsilon_i(\mathbf{1}_{|\varepsilon|\leq\kappa_n}+\mathbf{1}_{|\varepsilon|>\kappa_n})$. We get thus three terms: the two main are bounded, and are hence controled with the Talagrand Inequality (23). We obtain finally,

$$\sum_{m'\in\mathcal{M}_n}T_{2,b}^{m'}\leq C\frac{\ln(n)}{n},$$

which completes the proof.

□

• ***Proof of Lemma 9, first inequality***. The term $\mathbb{E}[T_3^m]$ requires more computations. Let us first notice that $T_3^m = \sum_{j=1}^{D_m}\{\int_0^1 f(G^{-1}(u))(\varphi_j(u)-\varphi_j(\hat{U}_n(u)))du\}^2$. We apply Taylor formula with Lagrange form for the remainder rest: there exists a random number depending on j, $\hat{\alpha}_{j,n,u}$, such that the following splitting holds:

$$T_3^m \leq 3T_{3,1}^m + 3T_{3,2}^m + 3T_{3,3}^m, \tag{36}$$

with notations

$$\begin{aligned}
T_{3,1}^m &= \sum_{j=1}^{D_m}\left\{\int_0^1 h(u)\left(\hat{U}_n(u)-u\right)\varphi_j'(u)du\right\}^2,\\
T_{3,2}^m &= (1/4)\sum_{j=1}^{D_m}\left\{\int_0^1 h(u)\left(\hat{U}_n(u)-u\right)^2\varphi_j''(u)du\right\}^2,\\
T_{3,3}^m &= (1/6)\sum_{j=1}^{D_m}\left\{\int_0^1 h(u)\left(\hat{U}_n(u)-u\right)^3\varphi_j^{(3)}(\hat{\alpha}_{j,n,u})du\right\}^2.
\end{aligned}$$

Writing the definition of $\hat{U}_n(u)$, and noting that $u = \mathbb{E}[\mathbf{1}_{U_i\leq u}]$ $(i=1,\dots,n)$, we get for the first term

$$T_{3,1}^m = \sum_{j=1}^{D_m}\left(\frac{1}{n}\sum_{i=1}^n A_{i,j} - \mathbb{E}[A_{i,j}]\right)^2, \quad \text{with } A_{i,j} = \int_{U_i}^1 h(u)\varphi_j'(u)du.$$

An integration by parts so as to compute $A_{i,j}$ leads to

$$T_{3,1}^m \leq 2T_{3,1,1}^m + 2T_{3,1,2}^m, \tag{37}$$

with notations

$$\begin{aligned}
T_{3,1,1}^m &= \sum_{j=1}^{D_m}\left\{\frac{1}{n}\sum_{i=1}^n h(U_i)\varphi_j(U_i) - \mathbb{E}\left[h(U_i)\varphi_j(U_i)\right]\right\}^2,\\
T_{3,1,2}^m &= \sum_{j=1}^{D_m}\left\{\int_0^1 h'(u)\left(\hat{U}_n(u)-u\right)\varphi_j(u)du\right\}^2.
\end{aligned} \tag{38}$$

The same study as the one done for T_1^m gives

$$\begin{aligned}
\mathbb{E}\left[T_{3,1,1}^m\right] &\leq \frac{1}{n}\sum_{j=1}^{D_m}\mathbb{E}\left[(h(U_1)\varphi_j(U_1))^2\right] \leq \frac{1}{n}\left\|\sum_{j=1}^{D_m}\varphi_j^2\right\|_\infty\int_0^1 h(u)^2du,\\
&= \int_0^1 h(u)^2du\phi_0^2\frac{D_m}{n} = \phi_0^2\mathbb{E}[f(X_1)^2]\frac{D_m}{n} \leq \phi_0^2\mathbb{E}[Y_1^2]\frac{D_m}{n}.
\end{aligned}$$

Besides, using definition and properties of the orthogonal projection on S_m,

$$T_{3,1,2}^m = \sum_{j=1}^{D_m}\left(\langle h'(\hat{U}_n - id), \varphi_j\rangle\right)^2 = \left\|\Pi_{S_m}(h'(\hat{U}_n - id))\right\|^2 \leq \|h'\|^2\|\hat{U}_n - id\|_\infty^2.$$

Concluding with Inequality (32), $p=2$, we obtain $\mathbb{E}[T_{3,1,2}^m] \leq C_2\|h'\|^2/n$. Hence,

$$\mathbb{E}\left[T_{3,1}^m\right] \leq 2\left(C_2\|h'\|^2\frac{1}{n} + \phi_0^2\mathbb{E}[Y_1^2]\frac{D_m}{n}\right) \leq C\frac{D_m}{n}.$$

Let us deal with $T_{3,2}^m$. We notice that for any $j \geq 2$, $\varphi_j'' = -(\pi\mu_j)^2\varphi_j$, with $\mu_j = j$ for even j, and $\mu_j = j-1$ for odd j. Consequently,

$$\begin{aligned}
\mathbb{E}\left[T_{3,2}^m\right] &= (\pi^4/4)\mathbb{E}\left[\sum_{j=1}^{D_m}\left\{\int_0^1 h(u)\left(\hat{U}_n(u)-u\right)^2\mu_j^2\varphi_j(u)du\right\}^2\right], \\
&\leq (\pi^4/4)D_m^4\mathbb{E}\left[\sum_{j=1}^{D_m}\left\{\int_0^1 h(u)\left(\hat{U}_n(u)-u\right)^2\varphi_j(u)du\right\}^2\right], \\
&= (\pi^4/4)D_m^4\mathbb{E}\left[\sum_{j=1}^{D_m}\left\{\langle h\left(\hat{U}_n-id\right)^2,\varphi_j\rangle\right\}^2\right].
\end{aligned}$$

Proceeding as in the term $T_{3,1,2}$, we get $\mathbb{E}[T_{3,2}^m] \leq C_4(\pi^4/4)\|h\|^2D_m^4/n^2$. Last, we bound roughly

$$\mathbb{E}\left[T_{3,3}^m\right] \leq (1/6)\sum_{j=1}^{D_m}\left\|\varphi_j^{(3)}\right\|_\infty^2\|h\|^2\mathbb{E}\left[\left\|\hat{U}_n-id\right\|_\infty^6\right] \leq \frac{C_6}{6}\left\|\varphi_2^{(3)}\right\|_\infty^2\|h\|^2\frac{D_m^7}{n^3}.$$

Finally, we gather the three bounds for $\mathbb{E}[T_{3,l}]$, $l=1,2,3$, to end the proof of the inequality.

□

• ***Proof of Lemma 9, second inequality***. Let us begin with $V_3(p_{m'}) \leq V_3(m')$. Therefore $\mathbb{E}[\max_{m'\in\mathcal{M}_n}(T_3^{p_{m'},b}-V_3(m'))_+] \leq \mathbb{E}[\max_{m'\in\mathcal{M}_n}(T_3^{p_{m'},b}-V_3(p_{m'}))_+]$. In the sequel, we simplify the notations by setting $p=p_{m'}$. As previously, we get $T_3^{p,b} \leq 6T_{3,1,1}^p+6T_{3,1,2}^p+3T_{3,2}^p+3T_{3,3}^p$. Thus

$$\begin{aligned}
(39)\ \mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{p,b}-V_3(p)\right)_+\right] &\leq \mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(6T_{3,1,1}^p-V_3(p)/3\right)_+\right]+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}6T_{3,1,2}^p\right] \\
&\quad+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(3T_{3,2}^p-V_3(p)/3\right)_+\right] \\
&\quad+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(3T_{3,3}^p-V_3(p)/3\right)_+\right].
\end{aligned}$$

The term that we have not centered is directly negligible: its definition (see (38)) proves that $T_{3,1,2}^p \leq T_{3,1,2}^{m_{\max}}$, thus we obtain

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}6T_{3,1,2}^p\right] \leq \frac{C}{n}. \tag{40}$$

It remains to bound the three other terms. Let us distinguish $T_{3,1,1}^p$ of the two others: Equality (28) and Lemma 5 lead to $T_{3,1,1}^p = \sup_{t\in\mathcal{S}(p)}(\nu_n^{(1)}(t))^2$, for the process defined by

$$\nu_n^{(1)}(t) = \frac{1}{n}\sum_{i=1}^n f(X_i)\left(t\circ G\right)(X_i) - \mathbb{E}\left[f(X_i)\left(t\circ G\right)(X_i)\right].$$

Thus we apply Talagrand Inequality (23), as in the proof of Proposition 3. The useful quantities are the following:

$$M_1^{(1)} = \phi_0\|f\|_\infty\sqrt{D_p},\ \left(H^{(1)}\right)^2 = \frac{D_p}{n}\mathbb{E}\left[f^2(X_1)\right]\phi_0^2,\ v^{(1)} = \|f\|_\infty^2.$$

We have again

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(6T^p_{3,1,1}-V_{3,1,1}(p)\right)_+\right]\leq\frac{C}{n}, \tag{41}$$

with $V_{3,1,1}(p)=6\times 2(1+2\delta)\mathbb{E}\left[f^2(X_1)\right]\phi_0^2 D_p/n$. But as

$$V_{3,1,1}(p)\leq 12(1+2\delta)\mathbb{E}\left[Y_1^2\right]\phi_0^2\frac{D_p}{n}:=V^{bis}_{3,1,1}(p),$$

the result holds with $V^{bis}_{3,1,1}$.
For the two other terms, the strategy is the one described in $(\mathcal{B}_2)$ (beginning of this section). For example, using $T^p_{3,2}\leq(\pi^4/4)\|h\|^2D_p^4\left\|\hat{U}_n-id\right\|_\infty^4$ implies, for $V_{3,2}(p)=\kappa D_p^4\ln^2(n)/n^2$,

$$\begin{aligned}
&\mathbb{E}\left[\left(3T^p_{3,2}-V_{3,2}(p)\right)_+\right]\\
&\qquad\leq\ (3\pi^4/4)\|h\|^2D_p^4\mathbb{E}\left[\left(\left\|\hat{U}_n-id\right\|_\infty^4-\frac{\kappa}{(3\pi^4/4)\|h\|^2}\frac{\ln^2(n)}{n^2}\right)_+\right],\\
&\qquad\leq\ CD_p^4n^{-\kappa_b^{1/2}2^{-1/2}}, \qquad (42)
\end{aligned}$$

for $\kappa_b=\kappa/(3\pi^4/4)\|h\|^2$. Thus, if $D_p\leq Cn^{1/3}$,

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(3T^p_{3,2}-V_{3,2}(p)\right)_+\right]\leq Cn\times n^{4/3}\times n^{-\kappa_b^{1/2}2^{-1/2}}.$$

The choice of $\kappa=50\pi^4/3\|h\|^2$ leads successively to $\kappa_b\geq 200/9$, and to $7/3-\sqrt{\kappa_b/2}\leq -1$, so that the last upper-bound is $O(1/n)$. If $D_p\leq Cn^{1/3}/\ln(n)$, we have

$$V_{3,2}(p)\leq 50\pi^4/3\mathbb{E}\left[Y_1^2\right]\frac{D_p}{n}:=V^{bis}_{3,2}(p),$$

which can also be used. We do not detail the control for the term $T^m_{3,3}$. Similarly, we get

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(3T^p_{3,3}-V^{bis}_{3,3}(p)\right)_+\right]\leq C/n, \tag{43}$$

with $V^{bis}_{3,3}(p)=(13^3\times 2/27)\|\varphi_2^{(3)}\|_\infty^2\mathbb{E}\left[Y_1^2\right]D_p/n$. We conclude the proof of Lemma 9 by gathering Inequalities (40), (41), (42), and (43), in the bound (39), and choosing the constant k_3 such that $V_3\geq 3V^{bis}_{3,1,1}$, $V_3\geq 3V^{bis}_{3,2}$, and $V_3\geq 3V^{bis}_{3,3}$.

□

• ***Proof of Lemma 10***. The sketch of the proof is the same as the proof of the second inequality of Lemma 9. We split

$$T_4^{m'}\leq 4T^{m'}_{4,1,1}+4T^{m'}_{4,1,2}+2T^{m'}_{4,2,1}+2T^{m'}_{4,2,2}+2T^{m'}_{4,2,3}, \tag{44}$$

where the different terms are defined below, and thus, $\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_4^{m'}-V_4(m')\right)_+\right]$

$$\begin{aligned}
\leq\ &\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(4T^{m'}_{4,1,1}-V_4(m')/3\right)_+\right]+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(4T^{m'}_{4,1,2}-V_4(m')/3\right)_+\right]\\
&+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(2T^{m'}_{4,2,3}-V_4(m')/3\right)_+\right]+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}2T^{m'}_{4,2,1}\right]+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}2T^{m'}_{4,2,2}\right].
\end{aligned}$$

We show that the two terms which we have not centered are negligible (less than $C\ln(n)/n$ if $D_{m_{\max}} = O(n^{1/3})$. For the three others we apply the strategy $(\mathcal{B}_2)$. Let us only detail how $T_4^{m'}$ is split, and the bounds for each $T_{4,l}$. First,

$$\begin{aligned} T_4^{m'} &= \left\| \mathbb{E}\left[\sum_{j=1}^{D_{m'}} \hat{a}_j^{\hat{G}} \left((\varphi_j \circ G) - (\varphi_j \circ \hat{G}_n) \right) |(X_{-l})_l \right] \right\|_g^2, \\ &\leq 2 \left\| \mathbb{E}\left[\sum_{j=1}^{D_{m'}} (\hat{a}_j^{\hat{G}} - a_j) \left((\varphi_j \circ G) - (\varphi_j \circ \hat{G}_n) \right) |(X_{-l})_l \right] \right\|_g^2 \\ &\quad + 2 \left\| \mathbb{E}\left[\sum_{j=1}^{D_{m'}} a_j \left((\varphi_j \circ G) - (\varphi_j \circ \hat{G}_n) \right) |(X_{-l})_l| \right] \right\|_g^2 := 2T_{4,1}^{m'} + 2T_{4,2}^{m'}. \end{aligned}$$

Then,

$$\begin{aligned} T_{4,1}^{m'} &\leq \int_{[a;b]} \mathbb{E}\left[\sum_{j=1}^{D_{m'}} \left(\hat{a}_j^{\hat{G}} - a_j \right)^2 \sum_{j=1}^{D_{m'}} \left(\varphi_j(G(x)) - \varphi_j(\hat{G}_n(x)) \right)^2 |(X_{-l})_l \right] g(x)dx, \\ &= \mathbb{E}\left[\sum_{j=1}^{D_{m'}} \left(\hat{a}_j^{\hat{G}} - a_j \right)^2 \sum_{j=1}^{D_{m'}} \int_{[0;1]} \left(\varphi_j(u) - \varphi_j(\hat{U}_n(u)) \right)^2 du \, |(X_{-l})_l \right], \\ &\leq 2T_{4,1,1}^{m'} + 2T_{4,1,2}^{m'}, \end{aligned}$$

with

$$\begin{aligned} T_{4,1,1}^{m'} &= \int_{[0;1]} \mathbb{E}\left[\left\{ \sum_{j=1}^{D_{m'}} \left(\hat{a}_j^{\hat{G}} - \mathbb{E}\left[\hat{a}_j^{\hat{G}} |(X_{-l})_l \right] \right)^2 \right\} \left\{ \sum_{j=1}^{D_{m'}} \left(\varphi_j(u) - \varphi_j(\hat{U}_n(u)) \right)^2 \right\} |(X_{-l})_l \right] du, \\ T_{4,1,2}^{m'} &= \int_{[0;1]} \mathbb{E}\left[\left\{ \sum_{j=1}^{D_{m'}} \left(\mathbb{E}\left[\hat{a}_j^{\hat{G}} |(X_{-l})_l \right] - a_j \right)^2 \right\} \left\{ \sum_{j=1}^{D_{m'}} \left(\varphi_j(u) - \varphi_j(\hat{U}_n(u)) \right)^2 \right\} |(X_{-l})_l \right] du \end{aligned}$$

Moreover, $T_{4,2}^{m'} = \| \sum_{j=1}^{D_{m'}} a_j((\varphi_j \circ G) - (\varphi_j \circ \hat{G}_n))\|_g^2$, so

$$\begin{aligned} \mathbb{E}\left[T_{4,2}^{m'} \right] &\leq \mathbb{E}\left[\left\| \sum_{j=1}^{D_{m'}} a_j \left((\varphi_j \circ G) - (\varphi_j \circ \hat{G}_n) \right) \right\|_g^2 \right], \\ &= \mathbb{E}\left[\sum_{j,k=1}^{D_{m'}} a_j a_k \int_0^1 (\varphi_j(u) - \varphi_j \circ \hat{U}_n(u))(\varphi_k(u) - \varphi_k \circ \hat{U}_n(u)) du \right]. \end{aligned} \tag{45}$$

This yields, with Taylor formula, $\mathbb{E}[T_{4,2}^{m'}] \leq \mathbb{E}[T_{4,2,1}^{m'} + T_{4,2,2}^{m'} + T_{4,2,3}^{m'}]$, with

$$T_{4,2,1}^{m'} = \sum_{j,k=1}^{D_{m'}} a_j a_k \int_0^1 (u - \hat{U}_n(u))^2 \varphi_j'(u)\varphi_k'(u)du,$$
$$T_{4,2,2}^{m'} = (1/4) \sum_{j,k=1}^{D_{m'}} a_j a_k \int_0^1 (u - \hat{U}_n(u))^4 \varphi_j''(\hat{\alpha}_{j,n,u})\varphi_k''(\hat{\alpha}_{k,n,u})du,$$
$$T_{4,2,3}^{m'} = \sum_{j,k=1}^{D_{m'}} a_j a_k \int_0^1 (u - \hat{U}_n(u))^3 \varphi_j''(\hat{\alpha}_{j,n,u})\varphi_k'(u)du,$$

recalling that $a_l = \langle h, \varphi_l \rangle$. This explains the decomposition (44). Let us now bound each term. The first one is

$$T_{4,1,1}^{m'} = \sum_{j=1}^{D_{m'}} \mathrm{Var}\left(\hat{a}_j^{\hat{G}} | (X_{-l})_l\right) \int_{[0;1]} \sum_{j=1}^{D_{m'}} \left(\varphi_j(u) - \varphi_j(\hat{U}_n(u))\right)^2 du,$$

which is bounded using the mean value theorem:

$$T_{4,1,1}^{m'} \leq \sum_{j=1}^{D_{m'}} \mathrm{Var}\left(\hat{a}_j^{\hat{G}} | (X_{-l})_l\right) D_m^3 \|\varphi_2'\|_\infty^2 \left\|\hat{U}_n - id\right\|_\infty^2.$$

As

$$\begin{aligned} \mathrm{Var}\left(\hat{a}_j^{\hat{G}} | \, (X_{-l})_l\right) &= \frac{1}{n}\mathrm{Var}\left\{Y_1 \varphi_j\left(\hat{G}_n(X_1)\right) | \, (X_{-l})_l\right\}, \\ &\leq \frac{1}{n}\left\|\varphi_j\right\|_\infty^2 \left(\mathbb{E}\left[f^2(X_1)\right] + \sigma^2\right) = \frac{1}{n}\left\|\varphi_j\right\|_\infty^2 \mathbb{E}\left[Y_1^2\right], \end{aligned}$$

we obtain

$$T_{4,1,1}^{m'} \leq \phi_0^2 \mathbb{E}\left[Y_1^2\right] \frac{D_{m'}}{n} \times D_{m'}^3 \|\varphi_2'\|_\infty^2 \left\|\hat{U}_n - id\right\|_\infty^2, \tag{46}$$

which allows us to conclude that as announced, $\mathbb{E}[\max_{m' \in \mathcal{M}_n} \left(4T_{4,1,1}^{m'} - V_4(m')/3\right)_+] \leq C/n$, by Inequality (33). The second term can be written

$$T_{4,1,2}^{m'} = T_3^{m'} \int_{[0;1]} \sum_{j=1}^{D_{m'}} \left(\varphi_j(u) - \varphi_j(\hat{U}_n(u))\right)^2 du,$$

and again by the mean value theorem $T_{4,1,2}^{m'} \leq T_3^{m'} D_{m'}^3 \|\varphi_2'\|_\infty^2 \|\hat{U}_n - id\|_\infty^2$. We replace $T_3^{m'}$ by its detailed bound which we obtain by gathering Inequalities (36) and (37):

$$T_3^{m'} \leq 6T_{3,1,1}^{m'} + 6T_{3,1,2}^{m'} + 3T_{3,2}^{m'} + 3T_{3,3}^{m'}.$$

This leads to $T_{4,1,2}^{m'} \leq \sum_{l=1}^4 T_{4,1,2,l}^{m'}$, and then $T_{4,1,2,l}^{m'} \leq C\|\hat{U}_n - id\|_\infty^{p_l}$ (p_l an integer), so that we can use the method ($\mathcal{B}_2$), for each of this four terms. As announced, the terms $T_{4,2,1}^{m'}$ and $T_{4,2,2}^{m'}$

do not require to be centered: first,

$$\begin{aligned} T_{4,2,1}^{m'} &= \int_0^1 (u-\hat{U}_n(u))^2 \left\{ \left(\Pi_{S_{m'}}(h)\right)'(u) \right\}^2 du, \\ &= \int_0^1 (u-\hat{U}_n(u))^2 \left\{ \Pi_{S_{m'}}(h')(u) \right\}^2 du, \\ &\leq \left\| \hat{U}_n - id \right\|_\infty^2 \left\| \Pi_{S_{m'}}(h') \right\|^2 \leq \left\| \hat{U}_n - id \right\|_\infty^2 \left\| h' \right\|^2, \end{aligned}$$

so that $\mathbb{E}[\max_{m' \in \mathcal{M}_n} T_{4,2,1}^{m'}] \leq C_2 \|h'\|^2/n$. Then, notice that $T_{4,2,2}^{m'} = (1/4)\mathbb{E}[\int_0^1 (u-\hat{U}_n(u))^4 (\sum_{j=1}^{D_{m'}} a_j \varphi_j''(\hat{\alpha}_{j,n,u}))^2 du]$, we bound the Fourier's coefficients of the function h. To that end, we introduce the real numbers μ_j, for $j \in \{1, \dots, D_m\}$, defined by $\mu_j = j$ if j is even, $\mu_j = j-1$ otherwise. We obtain:

$$\left(\sum_{j=1}^{D_{m'}} a_j \varphi_j''(\hat{\alpha}_{j,n,u}) \right)^2 = \|\varphi_2''\|_\infty^2 \left(\sum_{j=1}^{D_{m'}} a_j \mu_j^2 \right)^2 \leq \|\varphi_2''\|_\infty^2 \left(\sum_{j=1}^{D_{m'}} a_j^2 \mu_j^2 \right) \sum_{j=1}^{D_{m'}} \mu_j^2.$$

The function h belongs to the Sobolev space $W_{per}^{2,1}(L)$, because $h(0) = h(1)$, h belongs to $\mathcal{C}^1([0;1])$, and $\|h\|^2 = \|f\|_g^2 \leq L^2$. Thus we use Lemma A.3 (p. 162) from Tsybakov [30]: the sequence $(a_j)_j$ belongs to the ellipsoid $\Theta(1, L^2/\pi^2)$, so

$$T_{4,2,2}^{m'} \leq C\mathbb{E}\left[\left\| \hat{U}_n - id \right\|_\infty^4 D_{m'}^3 \right] \leq C\mathbb{E}\left[\left\| \hat{U}_n - id \right\|_\infty^2 D_{m_{\max}}^3 \right] \leq C \frac{D_{m_{\max}}^3}{n^2}.$$

Following the same line of computations, we write,

$$T_{4,2,3}^{m'} = \mathbb{E}\left[\int_0^1 (u-\hat{U}_n(u))^3 \left(\sum_{j=1}^{D_{m'}} a_j \varphi_j''(\hat{\alpha}_{j,n,u}) \right) \left(\sum_{k=1}^{D_{m'}} a_k \varphi_k'(u) \right) du \right],$$

and bound as follows, for $u \in [0;1]$

$$\left| \sum_{j=1}^{D_{m'}} a_j \varphi_j''(\hat{\alpha}_{j,n,u}) \right| \leq \|\varphi_2''\|_\infty \frac{L}{\pi} D_{m'}^{3/2}, \quad \left| \sum_{k=1}^{D_{m'}} a_k \varphi_k'(u) \right| \leq \|\varphi_2'\|_\infty \frac{L}{\pi} D_{m'}^{1/2}.$$

Consequently, $T_{4,2,3}^{m'} \leq \mathbb{E}[\|\hat{U}_n - id\|_\infty^3 D_{m'}^2]$, and we apply again the usual tools to end the proof. □

Acknowledgement

I would like to thank Fabienne Comte for helpful advices and carefully readings of this work.

References

[1] Antoniadis, A.; Grégoire, G.; Vial, P. Random design wavelet curve smoothing. *Statist. Probab. Lett.* 35 (1997), no. 3, 225-232.

[2] Audibert, J.Y.; Catoni, O. Robust linear least squares regression. *Ann. Statist.* (2011) (to appear), arXiv:1010.0074.

[3] Audibert, J.Y.; Catoni, O. Robust linear regression through PAC-Bayesian truncation. Preprint, arXiv:1010.0072.

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

[5] Barron, A.; Birgé, L.; Massart, P. Risk bounds for model selection via penalization. *Probab. Theory Related Fields* 113 (1999), no. 3, 301-413.

[6] Baudry, J.P.; Maugis, C.; Michel, B. Slope heuristics: overview and implementation. *Statistics and Computing* (2011) (to appear).

[7] Birgé, L. Model selection for Gaussian regression with random design. *Bernoulli* 10 (2004), no. 6, 1039-1051.
[8] Birgé, L.; Massart, P. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* 4 (1998), no. 3, 329-375.
[9] Birgé, L.; Massart, P. Minimal penalties for gaussian model selection. *Probability Theory and Related Fields* 138 (2006) no. 1-2, 33-73.
[10] Brunel, E.; Comte, F. Penalized contrast estimation of density and hazard rate with censored data. *Sankhya* 67 (2005), no. 3, 441-475.
[11] Brunel, E.; Comte, F.; Guilloux, A. Nonparametric density estimation in presence of bias and censoring. *TEST* 18 (2009), no. 1, 166-194.
[12] Chagny, G. Régression: bases déformées et sélection de modèles par pénalisation et méthode de Lepski. Preprint, hal 00519556 v2.
[13] Cai, T.T.; Brown, L.D. Wavelet shrinkage for nonequispaced samples. *Ann. Statist.* 26 (1998), no. 5, 1783-1799.
[14] Comte, F.; Rozenholc, Y. A new algorithm for fixed design regression and denoising. *Ann. Inst. Statist. Math.* 56 (2004), no. 3, 449-473.
[15] Donoho, D.L.; Johnstone, I.M.; Kerkyacharian, G.; Picard, D. Wavelet shrinkage: asymptopia? With discussion and a reply by the authors. *J. Roy. Statist. Soc. Ser. B* 57 (1995), no. 2, 301-369.
[16] DeVore, R.A.; Lorentz, G. Constructive approximation. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 303. *Springer-Verlag, Berlin*, 1993.
[17] Dvoretzky, A.; Kiefer, J.; Wolfowitz, J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* 27 (1956), 642-669.
[18] Efromovich, S. Nonparametric curve estimation. Methods, theory, and applications. Springer Series in Statistics. *Springer-Verlag, New York*, 1999. xiv+411 pp. ISBN: 0-387-98740-1
[19] Fan, J.; Gijbels, I. Variable bandwidth and local linear regression smoothers. *Ann. Statist.* 20 (1992), no. 4, 2008-2036.
[20] Gaïffas, S. On pointwise adaptive curve estimation based on inhomogeneous data. *ESAIM Probab. Stat.* 11 (2007), 344-364.
[21] Goldenshluger, A.; Lepski, O. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.* 39 (2011), no. 3, 1608-1632.
[22] Golubev, G. K.; Nussbaum, M. Adaptive spline estimates in a nonparametric regression model. (Russian) *Teor. Veroyatnost. i Primenen.* 37 (1992), no. 3, 554-561; *translation in Theory Probab. Appl.* 37 (1992), no. 3, 521-529.
[23] Hardle, W.; Tsybakov, A. Local polynomial estimators of the volatility function in nonparametric autoregression. *J. Econometrics* 81 (1997), no. 1, 223-242.
[24] Kerkyacharian, G.; Picard, D. Regression in random design and warped wavelets. *Bernoulli* 10 (2004), no. 6, 1053-1105.
[25] Klein, T.; Rio, E. Concentration around the mean for maxima of empirical processes. *Ann. Probab.* 33 (2005), no. 3, 1060-1077.
[26] Kohler, M.; Krzyzak, A. Nonparametric regression estimation using penalized least squares. *IEEE Trans. Inform. Theory* 47 (2001), no. 7, 3054-3058.
[27] Lacour, C. Adaptive estimation of the transition density of a particular hidden Markov chain. *J. Multivariate Anal.* 99 (2008), no. 5, 787-814.
[28] Nadaraya E. On estimating regression. *Theory of Probability and its Application.* 9 (1964), 141-142.
[29] Pham Ngoc T-M. Regression in random design and Bayesian warped wavelets estimators. *Electron. J. Stat.* 3 (2009), 1084-1112.
[30] Tsybakov, A.B. Introduction à l'estimation non-paramétrique. Mathématiques & Applications (Berlin), 41. *Springer-Verlag, Berlin*, 2004.
[31] Watson G.S. Smooth regression analysis. *Sankhya Ser. A* 26 (1964), 359-372.
[32] Wegkamp, M. Model selection in nonparametric regression. *Ann. Statist.* 31 (2003), no. 1, 252-273.