A fast adaptive strategy for the estimation of a conditional density

Gaëlle Chagny

A FAST ADAPTIVE STRATEGY FOR THE ESTIMATION OF A CONDITIONAL DENSITY

GAËLLE CHAGNY(*)

Abstract. We consider the problem of estimating the conditional density π of a response vector Y given the predictor X (which is assumed to be a continuous variable). We provide an adaptive nonparametric strategy to estimate π, based on model selection. We start with a collection of finite dimensional product spaces, spanned by orthonormal bases. But instead of expanding directly the target function π on these bases, we rather consider the expansion of $h(x,y) = \pi(F_X^{-1}(x), y)$, where F_X is the cumulative distribution function of the variable X. This 'warping' of the bases allows us to propose a family of projection estimators easier to compute than estimators resulting of the minimization of a regression-type contrast. The data-driven selection of the best estimator $\hat{h}$ for the function h, is done with a model selection device in the spirit of Goldenshluger and Lepski (2011). The resulting estimator is $\hat{\pi}(x,y) = \hat{h}(F_X(x), y)$ if F_X is known, or $\hat{\pi}(x,y) = \hat{h}(\hat{F}(x), y)$ otherwise, where $\hat{F}$ is the empirical distribution function. We prove that it realizes a global squared-bias/variance compromise, in a context of anisotropic function classes: we establish non-asymptotic mean-squared integrated risk bounds and provide also convergence rate for the risk. Simulation experiments illustrate the method.

Keywords: Adaptive estimator. Conditional density. Model selection. Non parametric estimation. Warped bases.

AMS Subject Classification 2010: 62G05; 62G07-62G08.

November 2011

1. Introduction

1.1. **Motivation.** Assume that we observe pairs of real random variables (X, Y) with joint unknown density $f_{(X,Y)}$. The relationship between the predictor X and the response Y is classically described by regression analysis. But this can also be achieved by estimating the entire conditional density, that is

$$\pi(x,y) = \frac{f_{(X,Y)}(x,y)}{f_X(x)}, \text{ if } f_X(x) > 0,$$

where f_X is the marginal density of the X, and is assumed not to vanish on the interval of estimation.
The aim of this paper is to provide a nonparametric strategy to estimate π, which has to be both adaptive, fast and simple to compute. Our main ideas are to use warped bases to build projection estimators and to perform model selection in the spirit of Goldenshluger and Lepski [GL11].

1.2. **State of the art.** Nonparametric conditional density estimation has become only recently a subject of interest, and the adaptive strategies are still rather scarce. To our knowledge, most of the methods to estimate π are based on the principle that it can be seen as a nonparametric

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

weighted regression. This leads mainly to two directions: kernel methods and projection estimators built on regression-type criterions.
For kernel estimation, Fan *et al.* [FYT96] generalize Rosenblatt estimator using local polynomials, while Bashtannyk *et al.* [HBG96] and then Hall *et al.* [HWY99] and Hyndman and Yao [HY02] propose different versions of reweighted Nadaraya-Watson estimator. De Gooijer and Zerom [DGZ03] combine the better side of the last methods, to propose a function which takes only positive values. Kanamori *et al.* [TNK09] study a piecewise linear kernel-estimator using methods based on quantile regression functions: a family of conditional quantile function provides a full description of π. All these estimators involve a ratio: this means both theoretical problems, as studied in Penskaya [Pen95], and numerical problems, due to the denominator which can be close to zero. This leads Faugeras [Fau09] to propose a kernel-type estimator based on the copula function and on the estimation of the marginal cumulative distribution functions of X and Y.
In a different direction, Györfi and Kohler [GK07] consider a partitioning-type estimate. These procedures have in common to be studied with an asymptotic point of view: consistence and asymptotic normality are shown. But the adaptive properties like the choice of the bandwidths for kernel estimator, are studied only in Bashtannyk and Hyndman [BH01] and in Hyndman and Yao [HY02].
Adaptation and minimax results have recently been developed. Efromovich proposes a Fourier basis to build a blockwise-shrinkage Efromovich-Pinsker estimator. The regression setting is first studied in [Efr07], while the general case is the subject of [Efr10a], using characteristic functions to rewrite π. Finally, multidimensionality is taken into account in [Efr10b]. Oracle-inequalities are given.
Such adaptation results are also provided by Brunel *et al.* [BCL07]. They use model selection methods, based on the minimization of a least-squares penalized contrast introduced by Lacour [Lac07]. But this contrast, considered also by Akakpo and Lacour [AL11] to deal with dependent data and inhomogeneous functional classes, does not provide explicit estimator without matrix invertibility requirements (except when using histogram basis). Moreover the penalty given in [BCL07] depends on the unknown infinite norm of π. It can be estimated but it requires then strong regularity assumptions. Notice also that recent works of Cohen and Le Pennec [CLP11] focus on a penalized maximum likelihood estimator leading to risk bounds for a Kullblack-Leibler loss function. In the same way of all recent papers, we provide a data driven estimator but with a new method allowing fast computation, thanks to the fact that we avoid matrix inversion and 'purify' the penalty function.

1.3. **Generality about the estimation method.** The data are pairs of real random variables $(X_i, Y_i)_{i\in\{1,\dots,n\}}$ (with n a positive integer), independent and identically distributed (i.i.d.) with joint density $f_{(X,Y)}$, supported by a subset $A_1 \times A_2$ of $\mathbb{R}^2$ (A_2 a bounded interval). We assume that the marginal density f_X of the X_i does not vanish, and denote by F_X the cumulative distribution function (c.d.f.) of these variables, which consequently admits an inverse.
Our aim is to use model selection point of view with a contrast leading to an explicit estimator and a selection rule which is entirely computable, while satisfying good theoretical properties under weak assumptions. The first point is achieved by the use of warped bases, introduced by Kerkyacharian and Picard [KP04] to provide a wavelet thresholding estimator of a regression function. In our conditional density setting, we precisely define

$$\forall (u,y) \in [0;1] \times A_2,\ h(u,y) = \pi\left(F_X^{-1}(u), y\right), \tag{1}$$

and recover π by estimating both h and F_X. The assumption that h is squared integrable leads to projection estimators of the form

$$\forall (u,y) \in [0;1] \times A_2, \;\; \hat{h}_{D_1,D_2}(u,y) = \sum_{j_1=1}^{D_1} \sum_{j_2=1}^{D_2} \hat{a}_{j_1,j_2} \phi_{j_1} \otimes \psi_{j_2}(u,y),$$

with $\phi_{j_1} \otimes \psi_{j_2}(u,y) = \phi_{j_1}(u)\psi_{j_2}(y)$, for different couples (D_1, D_2) with $(\phi_{j_1} \otimes \psi_{j_2})_{j_1,j_2}$ an orthonormal family of functions and $\hat{a}_{j_1,j_2}$ estimated coefficients. Then, instead of estimating F_X over the whole sample, we assume that we observe $(X_{-i})_{i\in\{1,\dots,n\}}$ a sample of variables with the same distribution than the (X_i) and independent of them. Thus, we can define

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

and propose an estimator of π given by:

$$\forall (x,y) \in A_1 \times A_2, \;\; \hat{\pi}_{D_1,D_2}(x,y) = \hat{h}_{D_1,D_2}\left(\hat{F}_n(x), y\right).$$

We get thus a development of $\hat{\pi}_{D_1,D_2}$ in an orthonormal basis, whose first coordinate is warped by $\hat{F}_n$:

$$\forall (x,y) \in A_1 \times A_2, \;\; \hat{\pi}_{D_1,D_2}(x,y) = \sum_{j_1=1}^{D_1} \sum_{j_2=1}^{D_2} \hat{a}_{j_1,j_2} \phi_{j_1} \otimes \psi_{j_2}\left(\hat{F}_n(x), y\right).$$

The particular case of known c.d.f. F_X is also studied. In the two cases (known or estimated F_X), the procedure is particularly simple and fast to compute, since the coefficients $\hat{a}_{j_1,j_2}$ are just empirical means (they do not involve any matrix inversion). The selection rule of the levels D_1 and D_2 used in a second step is inspired by recent works of Goldenshluger and Lepski [GL11] and is new in the multidimensional framework.
We give both non-asymptotic results such that oracle-inequalities (proving the adaptivity of our estimators) and asymptotic rates of convergence for the quadratic risk if the function h belongs to anisotropic functional spaces. We show that adaptation has no price and that the rate corresponds exactly to the best bias-variance compromize, with assumptions stated on function h instead of π. Moreover, on the practical examples, the strategy we propose outperforms the penalization device of Brunel *et al.* [BCL07]: it is faster and leads to smaller risks in most cases.

1.4. **Organization of the paper.** Section 2 presents the two warped bases estimators (the one built assuming F_X is known, and the one built in the general case). The performances of each estimator are studied in Section 3: the functional spaces are described and global risks bounds and rates of convergence presented. Section 4 is devoted to numerical results. Finally, the proofs are gathered in Section 5.

2. Estimation strategy

All the estimators defined in the sequel are projection estimators. Therefore, we begin with the description of the approximation spaces (Section 2.1). We proceed then in three steps to estimate the conditional density π, on $A_1 \times A_2$. First, we define a collection of estimators for the function h (see its definition (1)), by minimizing a contrast on the models (Section 2.2). The second step is then to ensure the automatic selection of the model, without any knowledge about the regularity of h. This leads to a well defined estimator $\hat{h}$ (Section 2.3). Finally, we partially warp $\hat{h}$ to estimate π.

2.1. **Approximation spaces.** Our estimation procedure is based on the assumption that the function h belongs to $L^2([0;1]\times A_2)$, the set of square-integrable functions on $[0;1]\times A_2$, which is equiped with its usual Hilbert structure: we denote by $\langle .,. \rangle$ the scalar-product and by $\|.\|$ the norm. Consequently, h can be developed in any orthonormal basis, and can be approximated by its orthogonal projections onto the linear subspaces spanned by the first functions of the basis. For the sake of simplicity, we assume $A_2 = [0;1]$ in the sequel. The case of any segment A_2 can be easily obtained by making a scaling change. Following the example of Efromovich [Efr99], we choose the Fourier basis $(\varphi_{j_1}\otimes\varphi_{j_2})_{j_1,j_2\in\mathbb{N}\backslash\{0\}}$ of $L^2([0;1]\times A_2)$, defined for $u,y\in[0;1]$ by

$$\varphi_1(u)=1,\ \ \forall k\in\mathbb{N}\backslash\{0\},\ \varphi_{2k}(u)=\sqrt{2}\cos(2\pi k u),\ \varphi_{2k+1}(u)=\sqrt{2}\sin(2\pi k u), \tag{2}$$

and $\varphi_{j_1}\otimes\varphi_{j_2}(u,y)=\varphi_{j_1}(u)\varphi_{j_2}(y)$. For an index $l=1,2$, we also denote by S_{m_l} the space spanned by $\{\varphi_1,\dots,\varphi_{D_{m_l}}\}$, for $D_{m_l}=2m_l+1$, and m_l an element of the set of indices $\mathcal{I}_n^{(l)}=\{1,\dots,[\sqrt{n}/2]-1\}$ ([.] is the integer part). The approximation spaces are then $\mathbb{S}_m=S_{m_1}\times S_{m_2}$ for $m=(m_1,m_2)\in\mathcal{M}_n$, with $\mathcal{M}_n=\mathcal{I}_n^{(1)}\times\mathcal{I}_n^{(2)}$. Thus, we have

$$\mathbb{S}_m=S_{m_1}\times S_{m_2}=\mathrm{Span}\left\{\varphi_{j_1}\otimes\varphi_{j_2},\ j_1=1,\dots,D_{m_1}\ j_2=1,\dots,D_{m_2}\right\},$$

and the dimension of $\mathbb{S}_m$ is $\mathbb{D}_m=D_{m_1}D_{m_2}$. Notice that for all $m_l\in\mathcal{I}_n^{(l)}$ $(l=1,2)$, $D_{m_l}\leq\sqrt{n}$ and thus $\mathbb{D}_m\leq n$.

Remark 1.

- The basis satisfies $\|\sum_{j1=1}^{D_{m_1}}\sum_{j2=1}^{D_{m_2}}(\varphi_{j_1}\otimes\varphi_{j_2})^2\|_\infty\leq\mathbb{D}_m$, where $\|.\|_\infty$ is the supremum of the function on $[0;1]\times A_2$. This is equivalent to the following useful link between the infinite norm and the L^2 norm (see Birgé and Massart [BM98] for the proof):

$$\forall t\in L^2([0;1]\times A_2),\ \|t\|_\infty\leq\sqrt{D_{m_1}D_{m_2}}\|t\|=\sqrt{\mathbb{D}_m}\|t\|. \tag{3}$$

- For each $m_l,m_l'\in\mathcal{I}_n^{(l)}$ $(l=1,2)$, we have

$$D_{m_l}\leq D_{m_l'}\implies S_{m_l}\subset S_{m_l'}. \tag{4}$$

Notice that other classical models, such as models spanned by regular wavelet basis, histogram basis or dyadic piecewise polynomial basis satisfy similar properties. We refer to Barron *et al.* [BBM99], and Brunel and Comte [BC05] for a precise description. See also Remark 2 below about the extension of our results to these models.

2.2. **Estimation on a fixed model.** We start with the following criterion

$$\forall t\in L^2([0;1]\times A_2)\mapsto\gamma_n(t,\hat{F}_n):=\|t\|^2-\frac{2}{n}\sum_{i=1}^n t\left(\hat{F}_n(X_i),Y_i\right). \tag{5}$$

This contrast is knew and quite far from the regression and density least-squares criterion. The novelty comes both from the L^2 norm which stands in place of the empirical norm used in the classical contrasts (see for example the contrast γ_n^0 in Brunel *et al.* [BCL07]), and from the presence of the empirical c.d.f $\hat{F}_n$. To justify this choice, assume for a moment that the distribution of the X_i is known: we can thus plug the true c.d.f F_X instead of its empirical

counterpart, and compute more easily, for $t \in L^2([0;1] \times A_2)$,

$$\begin{aligned}\mathbb{E}\left[\gamma_n(t,F_X)\right] - \mathbb{E}\left[\gamma_n(h,F_X)\right] &= \|t\|^2 - \|h\|^2 - 2\mathbb{E}\left[(t-h)\left(F_X(X_1),Y_1\right)\right], \\ &= \|t\|^2 - \|h\|^2 - 2\int_{A_1\times A_2}(t-h)(F_X(x),y)\pi(x,y)f_X(x)dxdy, \\ &= \|t\|^2 - \|h\|^2 - 2\int_{[0;1]\times A_2}(t-h)(u,y)h(u,y)dudy, \\ &= \|t\|^2 - \|h\|^2 - 2\langle h, t-h\rangle, \\ &= \|t-h\|^2.\end{aligned}$$

This quantity is minimal when $t = h$. This shows that $\gamma_n(.,F_X)$ in the case of known F_X (or $\gamma_n(.,\hat{F}_n)$ otherwise) suits well for the estimation of h. We set thus, for each model $\mathbb{S}_m$,

$$\hat{h}_m^{\hat{F}} = \arg\min_{t\in\mathbb{S}_m}\gamma_n(t,\hat{F}_n),\ \hat{h}_m^{F_X} = \arg\min_{t\in\mathbb{S}_m}\gamma_n(t,F_X), \tag{6}$$

or equivalently,

$$\hat{h}_m^{\hat{F}} = \sum_{j_1=1}^{D_{m_1}}\sum_{j_2=1}^{D_{m_2}}\hat{a}_{j_1,j_2}^{\hat{F}}\varphi_{j_1}\otimes\varphi_{j_2},\quad \text{with}\quad \hat{a}_{j_1,j_2}^{\hat{F}} = \frac{1}{n}\sum_{i=1}^{n}\varphi_{j_1}(\hat{F}_n(X_i))\varphi_{j_2}(Y_i), \tag{7}$$

and a similar expression for estimator $\hat{h}_m^{F_X}$ with coefficients $a_{j_1,j_2}^{F_X}$ in the case of known c.d.f. F_X. Finally, we set

$$\pi_m^{\hat{F},\hat{F}}(x,y) = \hat{h}_m^{\hat{F}}(\hat{F}_n(x),y)\quad\text{and}\quad \hat{\pi}_m^{F_X,F_X}(x,y) = \hat{h}_m^{F_X}(F_X(x),y),$$

denoted with two super-indexes $\hat{F}$ (or F_X) to underline the double dependence of the estimator on this function, through both the coefficients $\hat{a}_{j,k}^{\hat{F}}$ and the composition of the first variable by F_X. Notice the advantage of the contrast we define: we get an explicit formula for the estimator. The coefficients are empirical means easily computable. They do not involve a matricial inversion compared to the estimator obtained via least-squares criterion (see for example Brunel *et al.* [BCL07]). Moreover, in the case of known c.d.f. F_X, $\hat{h}_m^{F_X}$ is an unbiased estimator of the orthogonal projection of h onto $\mathbb{S}_m$.

2.3. **Model selection.**

2.3.1. *Risk on a fixed model.* In order to explain which model $\mathbb{S}_m$ we should choose, we first study the quadratic risk of each estimator of the collection, in the simpler case of known c.d.f. F_X. The loss function naturally associated to our context is the following L^2-norm,

$$\forall v \in L^2(A_1\times A_2, f_X),\ \|v\|_{f_X}^2 = \int_{A_1\times A_2} v^2(x,y)f_X(x)dxdy,$$

with $L^2(A_1\times A_2, f_X)$, the space of squared-integrable functions on $A_1\times A_2$ with respect to the Lebesgue measure weighted by the density f_X. We denote $\langle .,.\rangle_{f_X}$ the corresponding scalar-product. Notice besides that the following links hold between this norm and the classical norm previously defined: for $t, s \in L^2([0;1]\times A_2)$, we compute, using $F_X' = f_X$,

$$\|t(F_X(.),.)\|_{f_X} = \|t\|,\ \ \langle t(F_X(.),.), s(F_X(.),.)\rangle_{f_X} = \langle t, s\rangle.$$

The classical L^2-norm on $A_1\times A_2$ can be recovered, under the assumption that f_X is bounded from below by a strictly positive constant. This assumption is standard, see for example Assumption $\mathcal{A}_2$ in Brunel *et al.* [BCL07], or Assumption (H_{Bas}) in Baraud [Bar02].
For the weighted L^2-risk which is used in the sequel, and for each $m \in \mathcal{M}_n$, we get

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\pi}_m^{F_X,F_X}-\pi\right\|_{f_X}^2\right] &= \left\|\pi-\pi_m^{F_X}\right\|_{f_X}^2+\mathbb{E}\left[\left\|\pi_m^{F_X}-\hat{\pi}_m^{F_X,F_X}\right\|_{f_X}^2\right],\\ &= \|h-h_m\|^2+\mathbb{E}\left[\left\|h_m-\hat{h}_m^{F_X}\right\|^2\right], \qquad (8)\end{aligned}$$

where

$$\pi_m^{F_X}(x,y)=h_m(F_X(x),y) \text{ and } h_m \text{ is the orthogonal projection of } h \text{ onto } \mathbb{S}_m. \qquad (9)$$

We recover the usual squared-bias/variance decomposition of the risk. The key point is the difference of behaviour of the two terms: they both depend on $\mathbb{D}_m$ but in opposite ways. The first term in the right-hand side of (8) decreases when $\mathbb{D}_m$ grows, since π is better approximated by its projection when the approximation space grows, while the second term grows with $\mathbb{D}_m$:

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\pi}_m^{F_X,F_X}-\pi_m^{F_X}\right\|_{f_X}^2\right] &= \sum_{j_1=1}^{D_{m_1}}\sum_{j_2=1}^{D_{m_2}}\mathrm{Var}\left(\hat{a}_{j_1,j_2}^{F_X}\right),\\ &\leq \frac{1}{n}\sum_{j_1=1}^{D_{m_1}}\sum_{j_2=1}^{D_{m_2}}\mathbb{E}\left[\left(\varphi_{j_1}\left(F_X(X_i)\right)\varphi_{j_2}(Y_i)\right)^2\right]\leq\frac{D_{m_1}D_{m_2}}{n}, \qquad (10)\end{aligned}$$

using Property (3) (see Section 2.1). The best model among the collection is the one which minimizes the right-hand side in (9), making a trade-off between the squared-bias term and the variance term. However, it is unknown since h and h_m are not observed. Therefore, an adaptive estimator of π must make automatically this compromise.

2.3.2. *Selection rule.* We propose to use a scheme proposed by Goldenshluger and Lepski [GL11] for density estimation. The adaptive index is chosen as the value which minimizes the following sum:

$$\hat{m}^{\hat{F}}=\left(\hat{m}_1^{\hat{F}},\hat{m}_2^{\hat{F}}\right)=\arg\min_{m\in\mathcal{M}_n}\left[A(m,\hat{F}_n)+2V^{\hat{F}}(m)\right],$$

where $V^{\hat{F}}$ has the order of the variance term:

$$V^{\hat{F}}: m=(m_1,m_2)\ \mapsto\ c_1\frac{D_{m_1}D_{m_2}}{n}, \qquad (11)$$

with c_1 is a purely numerical constant, adjusted in practice. The function $A(.,\hat{F}_n)$ is based on the comparison of the estimators built in the first stage:

$$A(m,\hat{F}_n)=\max_{m'\in\mathcal{M}_n}\left(\left\|\hat{h}_{m'}^{\hat{F}}-\hat{h}_{m\wedge m'}^{\hat{F}}\right\|^2-V^{\hat{F}}(m')\right)_+, \qquad (12)$$

where $x_+=\max(x,0)$, $x\in\mathbb{R}$. We will prove besides that $A(m,\hat{F}_n)$ has the order of the bias term (see Inequality (22)). Thus we get an estimator, explicitly expressed in a warped basis,

$$\tilde{\pi}(x,y)=\hat{h}_{m^{\hat{F}}}^{\hat{F}}(\hat{F}_n(x),y). \qquad (13)$$

The L^2-norm involved in the definition of $A(.,\hat{m})$ is easy to compute, since the functions $\hat{h}_{m'}^{\hat{F}}$, $m'\in\mathcal{M}_n$ are expressed with a development in an orthonormal basis (see Section 4 for details). This advantage has to be noticed compared to other strategies of model selection using the contrast function or to strategies involving bandwith choice for a kernel.

There are several novelties to underline. First, the warping of the basis for the variable x leads to explicit and simple coefficients $\hat{a}_{j_1,j_2}^{\hat{F}}$ for the estimator. The use of a selection device inspired of

Goldenshluger and Lepski [GL11] is original in the setting of multidimensional model selection. Note also that the specific factor 2 in the definition of $\hat{m}^{\hat{F}}$ plays an important (but technical) role in the proofs. The "penalty" term $V^{\hat{F}}$ is entirely computable with the data (with no term to estimate), up to a purely numerical constant to calibrate. On the opposite, penalization of a regression-type contrast in this context leads to a penalty which depends on the unknown infinite norm of π (see Brunel *et al.* [BCL07], or Lacour [Lac07]).
Finally, let us define also an estimator in the toy case of known c.d.f. F_X:

$$\tilde{\pi}_0(x,y) = \hat{h}^{F_X}_{\hat{m}^{F_X}}(F_X(x),y), \tag{14}$$

with $\hat{m}^{F_X}$ selected as the argument-minimum of $A(m,F_X)+V^{F_X}(m)$, where we denote by $V^{F_X}(m) = c_0 D_{m_1} D_{m_2}/n$, c_0 a numerical constant, which can be different of c_1.

3. Main results

3.1. **Anisotropic Sobolev spaces.** Let us define the functional spaces we consider further for the function h (even if its index of regularity has not to be known). The choice of the trigonometric models leads us to consider spaces of periodic functions, that is Sobolev spaces. We define them directly via Fourier coefficients, keeping in mind that it can also be characterized via weak differentiability (see for example DeVore and Lorentz [DL93] and Härdle *et al.* [HKPT98] for functions of one variable, and Adams [Ada75] for functions of several variables). Precisely, our aim is to extend to functions of two variables the characterization of Tsybakov (Lemma A.3, p.162, [Tsy04]).
Let $t \in L^2([0;1]^2)$. Then there exists a real-valued family $(\theta_{j_1,j_2})_{j_1,j_2\in\mathbb{N}\setminus\{0\}}$ such that

$$t = \sum_{j_1,j_2\in\mathbb{N}\setminus\{0\}} \theta_{j_1,j_2}\varphi_{j_1}\otimes\varphi_{j_2}.$$

Recall that the functions φ_j are defined by (2). We say that t belongs to the partial ball with radius $L>0$ and regularity $\alpha=(\alpha_1,\alpha_2)$ ($\alpha_l\in\mathbb{N}$, $l=1,2$, but not simultaneously equal to zero), if

$$\sum_{j_1,j_2\in\mathbb{N}\setminus\{0\}} \mu^2_{j_1,\alpha_1}\mu^2_{j_2,\alpha_2}\theta^2_{j_1,j_2} \le \frac{L^2}{\pi^{2(\alpha_1+\alpha_2)}}, \tag{15}$$

with $\mu_{j_l,\alpha_l} = j_l^{\alpha_l}$ for even j_l, $\mu_{j_l,\alpha_l} = (j_l-1)^{\alpha_l}$ otherwise. We write $t\in W^2_{per}([0;1]^2,L,\alpha)$, in the spirit of the definition of Tsybakov [Tsy04]. These spaces are anisotropic. The function h can thus have different smoothness properties with respect to different directions.
Let us finally give a useful approximation property of this space. We denote by $t_m = t_{(m_1,m_2)}$ the orthogonal projection of the function t onto the subspace $\mathbb{S}_m = \mathbb{S}_{(m_1,m_2)}$. We have the following rate:

$$\|t-t_m\|^2 \le C(\alpha,L)\left(D_{m_1}^{-2\alpha_1}+D_{m_2}^{-2\alpha_2}\right),$$

where $C(\alpha,L)$ is a constant depending on α and L. This inequality is a particular case of Lemma 9 in Lacour [Lac07], based on papers from Hochmuth [Hoc02] and Nikol'skii [Nik75].

3.2. **Case of known c.d.f. $\mathbf{F_X}$.** We first focus on the simpler situation of known c.d.f. F_X. This allows us to derive the results with few assumptions and short proofs. The first theorem provides non-asymptotic bounds for the risk of the estimator $\tilde{\pi}_0$ (see its definition (14)). We recall that the trigonometric models satisfy properties (3) and (4), and that the dimensions D_{m_l} are bounded by $\sqrt{n}$.

Theorem 1. *We assume that the function h is bounded on the space $[0;1] \times A_2$. Then there exists c a purely numerical constant, and C a constant depending on $\|h\|_\infty$ such that*

$$\mathbb{E}\left[\|\tilde{\pi}_0 - \pi\|_{f_X}^2\right] \leq c \min_{m \in \mathcal{M}_n} \left\{ \frac{D_{m_1} D_{m_2}}{n} + \left\|\pi_m^{F_X} - \pi\right\|_{f_X}^2 \right\} + \frac{C}{n},$$

with $\pi_m^{F_X}$ defined by (9).

The basic outline of model selection (by Goldenshluger-Lepski method in our case) is to estimate the bias-variance sum and to select the model which minimizes it. Theorem 1 shows that it is a good strategy: the right model (in the sense that it realizes the trade-off) has been chosen in a data-driven way and the selected estimator performs as well as the best estimator in the family $\{\pi_m^{F_X,F_X},\ m \in \mathcal{M}_n\}$, up to some multiplicative constants and to a negligible residual term of order $1/n$. The constants are given in the proof, which is deferred to Section 5.2.

Remark 2. This result still holds in a more general setting. The choice of trigonometric models is not a necessary condition. It is sufficient to assume that the models which are used satisfy properties (3) and (4), and have their dimensions bounded by $\sqrt{n}$, which are very weak assumptions.

Theorem 1 enables also us to give a rate of convergence for the estimation of π, under regularity assumptions for function h. Precisely, the minimization of the left-hand-side of the inequality in the case of regular functions leads to the following Corollary.

Corollary 1. *We assume that the function h belongs to the anisotropic Sobolev ball denoted by $W^2_{per}([0;1]^2, L, \alpha)$, for some fixed $L > 0$ and $\alpha = (\alpha_1, \alpha_2)$ ($\alpha_l \in \mathbb{N}$, $l = 1, 2$, but not simultaneously equal to zero), with $\alpha_1 - \alpha_2 + 2\alpha_1\alpha_2 > 0$, and $\alpha_2 - \alpha_1 + 2\alpha_1\alpha_2 > 0$. Then, under the assumptions of Theorem 1,*

$$\mathbb{E}\left[\|\tilde{\pi}_0 - \pi\|_{f_X}^2\right] \leq C(\alpha, L) n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}},$$

with $C(\alpha, L)$ a numerical constant which depends only on α and L, and $\bar{\alpha}$ the harmonic mean of α_1 and α_2.

The harmonic mean of α_1 and α_2 is the real $\bar{\alpha}$ such that $2/\bar{\alpha} = 1/\alpha_1 + 1/\alpha_2$. Note that the condition $\alpha_1 - \alpha_2 + 2\alpha_1\alpha_2 > 0$ is ensured as soon as $\alpha_1 > 1/2$ and $\alpha_2 - \alpha_1 + 2\alpha_1\alpha_2 > 0$ as soon as $\alpha_2 > 1/2$. As the α_l are integers, this implies that they are larger than or equal to 1. In this case, h is bounded.
The corollary means that without knowing α and L (depending on the unknown h), $\tilde{\pi}_0$ does as well as the best possible estimator which knows these quantities. It is thus an adaptive estimator. Since Theorem 1 holds for piecewise polynomials or wavelet basis, the results can be extended to functions h belonging to anisotropic Besov spaces.

Remark 3. We can connect this result to the lower bound established by Lacour [Lac07], over Besov functional classes, for the estimation of the transition density of a Markov chain. The estimation of the conditional density is a particular case of that study. However, the regularity assumptions are set directly on function π in [Lac07], and not on function h. The right framework to relate our result to the one of [Lac07] is to define weighted regularity spaces, such as weighted Besov spaces defined and studied carefully in Kerkyacharian and Picard [KP04]. Since the main goal of our work is to product non-asymptotic bounds for the risk, which do not require such assumptions, we do not go further in that direction. Thus we only conjecture that the rate of convergence $n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}$ is probably optimal in the minimax sense, over Besov classes. The adaptive minimax rate over Sobolev spaces has most likely the same order.

3.3. **Case of unknown c.d.f. $\mathbf{F_X}$.** Coming back to the general case, we set the same result as Theorem 1, with slightly stronger assumptions.

Theorem 2. *We assume that the function h belongs to the anisotropic Sobolev ball denoted by $W^2_{per}([0;1]^2, L, (1,0))$, for some fixed $L > 0$, is bounded on $[0;1]^2$, and is $\mathcal{C}^1$ with respect to its first variable on $[0;1]$. We also assume that, for some constants C_a, C_b, C_c, the trigonometric models satisfy,*

$$\forall m = (m_1, m_2) \in \mathcal{M}_n,\ D_{m_1} \leq C_a \left(\frac{n}{\ln^2(n)}\right)^{1/3} \quad \text{and} \quad C_b \ln^5(n) \leq D_{m_2} \leq C_c\sqrt{n}. \tag{16}$$

Then, there exists numerical constants c and C depending on $\|\varphi_2'\|_{\infty,[0;1]}$, $\|\varphi_2''\|_{\infty,[0;1]}$, $\|\varphi_2^{(3)}\|_{\infty,[0;1]}$, $\|h\|$, $\|\partial_1 h\|$, and L, such that

$$\mathbb{E}\left[\|\tilde{\pi} - \pi\|^2_{f_X}\right] \leq c \min_{m\in\mathcal{M}_n}\left\{\frac{D_{m_1}D_{m_2}}{n} + \left\|\pi_m^{\hat{F}} - \pi\right\|^2_{f_X}\right\} + \frac{C}{n}. \tag{17}$$

Remark 4.
- There exists actually an integer n_0, depending on the function h, such that Inequality (17) holds for all $n \geq n_0$ with a purely numerical constant c. But the result is nonasymptotic, since the inequality holds also for $n < n_0$, taking a constant c which depends on quantities of the problem.
- Up to this result, the models S_{m_1} and S_{m_2} and their respective dimension have played the same role. But in the theorem, the dimension constraints (16) are not the same in each direction. To be totally rigorous, we should denote by $S^{(l)}_{m_l}$ the models and by $D^{(l)}_{m_l}$ their dimension, for each $l = 1, 2$. For the sake of simplicity, we keep the first notations as there is no possible confusion.

As in the case of known F_X, the theorem shows that the best estimator in the family $\{\pi_m^{\hat{F},\hat{F}},\ m \in \mathcal{M}_n\}$ is found up to some multiplicative constants for the risk, in a data-driven way. Brunel *et al.* [BCL07] provide also the same kind of oracle-inequality for their estimator built by penalization of a regression-type contrast. The assumptions seem first to be slightly less restrictive: it is only assumed that $D_{m_1} \leq n^{1/2}/\ln(n)$. However, the term $V^{\hat{F}}$ does not contain any unknown term and is then entirely computable, contrary to the penalty used in [BCL07], which depends on $\|\pi\|_\infty$. Moreover, replacing this quantity by an estimator requires in fact much more regularity constraints than the one we get, and leads to a semi-asymptotic result (see the appendix of Lacour [Lac07] for an example of these conditions). Consequently, a model selection strategy in the spirit of Goldenshluger-Lepski applied with warped bases has the advantage of providing an estimator easier to compute than a regression-type estimator and with good theoretical properties under quite weak assumptions.
Recall that the bound of Inequality (17) is close to the order of the sum of the variance term and the bias term. It implies that the obtained rate of convergence is likely to be minimax in most cases. More precisely, we prove the following corollary.

Corollary 2. *We assume that the function h belongs to the anisotropic Sobolev ball denoted by $W^2_{per}([0;1]^2, L, (1,0))$, for some fixed $L > 0$, and $\alpha = (\alpha_1, \alpha_2)$ ($\alpha_l \in \mathbb{N}$, $l = 1, 2$, but not simultaneously equal to zero) with $\alpha_1 - 2\alpha_2 + 2\alpha_1\alpha_2 > 0$, and $\alpha_2 - \alpha_1 + 2\alpha_1\alpha_2 > 0$. Then, under the assumptions of Theorem 2,*

$$\mathbb{E}\left[\|\tilde{\pi} - \pi\|^2_{f_X}\right] \leq C(\alpha, L) n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}},$$

with $C(\alpha, L)$ a numerical constant which depends on α and L, $\|\varphi_2'\|_{\infty,[0;1]}$, $\|\varphi_2''\|_{\infty,[0;1]}$, $\|\varphi_2^{(3)}\|_{\infty,[0;1]}$, $\|h\|$, $\|\partial_1 h\|$. The quantity $\bar{\alpha}$ is the harmonic mean of α_1 and α_2.

Even if F_X is unknown, our estimator adapts to the unknown regularity α of the function h. We also refer the reader to Remark 3 concerning the minimax sense of the result.

4. Simulation study

The aim of this section is to illustrate the behaviour of the estimator $\tilde{\pi}$ and to compare it with the estimator of Brunel *et al.* [BCL07] denoted by $\tilde{\pi}_{BCL}$. Thus, we investigate in the same time the difference between the classical bases and the warped bases, and the difference between the Goldenshluger-Lepski method and the penalization device.

4.1. Examples.

We propose a simulation study based on the following examples: we generate samples $(X_i, Y_i, X_{-i})_{i\in\{1,\dots,n\}}$ such that

- Examples 1: $Y_i = b(X_i) + \varepsilon_i$, with the following possibilities. The X_i's follow a uniform distribution on the interval $[0;1]$ (denoted by $\mathcal{U}_{[0;1]}$), or on the interval $[-1;1]$ ($\mathcal{U}_{[-1;1]}$), or a standard Gaussian distribution ($\mathcal{N}(0,1)$). The ε_i's are generated following the standard Gaussian distribution, or the Gamma distribution ($\Gamma(4,1)$) with parameters 4 and 1 (the 1 is the scale parameter). We denote by f_ε their density. The sample (ε_i) is independent of the (X_i). Finally, the regression function b is $b(x) = 2x+5$, $b(x) = \cos(x)$ or $b(x) = x^2$. The conditional density π is thus given by
$$\pi(x,y) = f_\varepsilon(y - b(x)).$$
- Example 2: X_i follows a uniform distribution on $[0;1]$, Y_i a standard Gaussian distribution, and X_i is independent of Y_i. The conditional density is just the density of the variable Y_i.
- Example 3: $Y_i = b(X_i)+\sigma(X_i)\varepsilon_i$, with a uniform distribution on $[0;1]$ for X_i, the previous Gamma distribution for ε_i (which is independent of X_i) and $\sigma(x) = \sqrt{1.3 - |x|}$. Similarly to Examples 1, the conditional density is
$$\pi(x,y) = f_\varepsilon(y - b(x)/\sigma(x))/\sigma(x).$$
- Example 4: The X_i follows a uniform distribution $\mathcal{U}_{[0;1]}$, and given $X_i = x$, Y_i follows the Gaussian mixture $0.5\mathcal{N}(8-4x,1) + 0.5\mathcal{N}(8+4x,1)$. The function π is the density of the mixture.

Examples 3 and 4, and some cases of Examples 1 have also been studied by Brunel *et al.* [BCL07], while Example 2 is proposed by Efromovich [Efr07] (p.2526).

4.2. Remarks about the implementation and results.

To implement each estimator $\tilde{\pi}$ and $\tilde{\pi}_{BCL}$, we use the trigonometric basis. For each sample of data (that is for each computation of the estimators), we calibrate the set A_1 over 95% of the variables X_i: we choose to eliminate the smallest values (2.5%), and the largest values (2.5%) of the data to avoid the side effects. We repeat this method to define A_2 with the variables Y_i.
For our estimator $\tilde{\pi}$, we have to compute the sum $A(m, \hat{F}_n) + 2V(m)$ for each $m = (m_1, m_2)$. Notice that the quadratic norm in the definition of $A(m, \hat{F}_n)$ (see (12)) is simply equal to a sum of squared-coefficients. For example, if $m \wedge m' = (m_1, m_2')$,
$$\left\| \hat{h}^{\hat{F}}_{m'} - \hat{h}^{\hat{F}}_{m\wedge m'} \right\|^2 = \sum_{j=D_{m_1}+1}^{D_{m_1'}} \sum_{k=1}^{D_{m_2'}} \left(\hat{a}^{\hat{F}}_{j,k}\right)^2.$$
A large number of simulations allows us to calibrate the constant in the definition of V: $c_1 = 0.2$. The estimator $\tilde{\pi}_{BCL}$ of Brunel *et al.* [BCL07] is defined as a penalized least-squares contrast

estimator. The penalty is $\text{pen}(m) = K_0\|\pi\|_\infty D_{m_1} D_{m_2}/n$. We put $K_0 = 0.5$ like in [BCL07] but we do not replace $\|\pi\|_\infty$ by an upper bound. To have a real data-driven procedure, we estimate it by taking the supremum of the values of a least-squares estimator on a fixed model $\mathbb{S}_m$ on a rough grid, with $m = [(\ln(n) - 1)/2]$.
Figures 1 and 2 illustrate the visual quality of the reconstruction, for a case of Examples 1, and for Example 4. We do not observe significant differences between the two estimators, which both behave quite well. However, the computation of $\tilde{\pi}_{BCL}$ requires much more time than the one of $\tilde{\pi}$, probably because of the presence of a matricial inversion, consequence of the least-squares contrast. The warped-bases estimator can thus advantageously be used for estimation problems with large data samples (data deriving from domain such as physics, fluorescence, finance...).

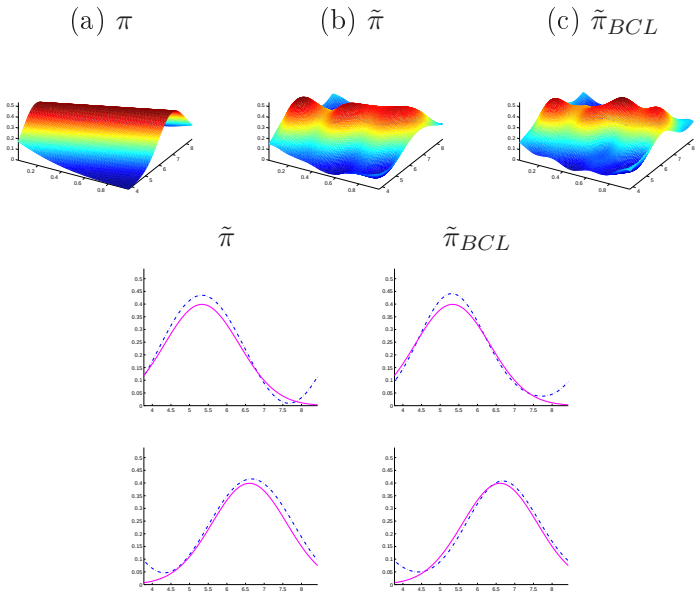

FIGURE 1. Plots of true function versus estimators, Examples 1, with X_i i.i.d. $\mathcal{U}_{[0;1]}$, ε_i i.i.d. $\mathcal{N}(0,1)$, and $b(x) = 2x + 5$ with $n = 2000$ observations. First line: (a) true function π, (b) estimator $\tilde{\pi}$, (c) estimator $\tilde{\pi}_{BCL}$. Second line: plots of $y \mapsto \pi(x,y)$ (full line), $y \mapsto \tilde{\pi}(x,y)$ (left, dashed dotted line) and $y \mapsto \tilde{\pi}_{BCL}(x,y)$ (right, dashed dotted line) for a fixed x. Third line: like the second line, for another value of x.

For sample sizes $n = 200, 500$ and 2000, we give in Tables 1 and 2 the estimated values of the risk $\mathbb{E}[\|\hat{\pi} - \pi\|_2^2]$, with $\|.\|_2$ the quadratic norm on $L^2(A_1 \times A_2)$, and $\hat{\pi} = (\tilde{\pi}_{BCL})_+$ or $(\tilde{\pi})_+$. It is not difficult to see that the choice of the positive part of both estimators can only make their risks decrease. The estimation of the expectation is done over $N = 100$ replicated samples, and the quadratic norm is approximated using subdivisions of A_1 and A_2 (see Brunel *et al.* [BCL07], Section 5.1, for details about the formula).

The risk of our estimator $\tilde{\pi}$ is often better than the one of the penalized least-squares estimator $\tilde{\pi}_{BCL}$. We indicate in those cases (in parenthesis) the percentage of improvement in the two tables: it can be quite important (up to 75%). Precisely, in Table 1, one can notice that for the sample size $n = 200$, there is as many cases where the risk of $\tilde{\pi}$ is better than the one of $\tilde{\pi}_{BCL}$ as the opposite case (risk of $\tilde{\pi}$ larger than the risk of $\tilde{\pi}_{BCL}$). However, for the larger sample sizes $n = 500$ or $n = 2000$, $\tilde{\pi}$ has a smaller risk in 89% of the situations of Example 1 (see Table 1).

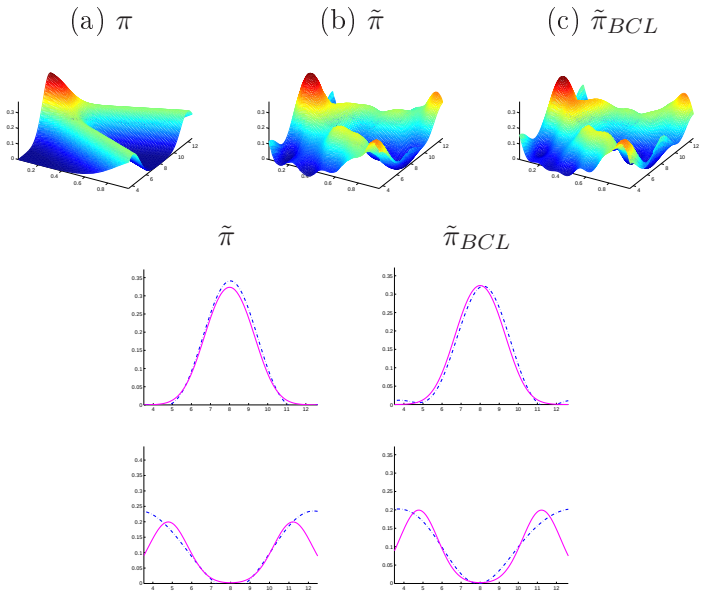

FIGURE 2. Plots of true function versus estimators, Example 4 with $n = 2000$ observations. First line: (a) true function π, (b) estimator $\tilde{\pi}$, (c) estimator $\tilde{\pi}_{BCL}$. Second line: plots of $y \mapsto \pi(x, y)$ (full line), $y \mapsto \tilde{\pi}(x, y)$ (left, dashed dotted line) and $y \mapsto \tilde{\pi}_{BCL}(x, y)$ (right, dashed dotted line) for a fixed x. Third line: like the second line, for another value of x.

To conclude this section, let us stress out two main advantages of building an estimator of π developed in warped bases, and selected with a Goldenshluger-Lepski strategy, in practice: first, its expression is explicit, fast and simple to compute (much faster than the least-squares strategy). Then, on top of its novelty and simplicity, it seems to bring risk values which are smaller than the least-squares method.

5. PROOFS

In all the proofs, the letter C denotes a nonnegative real that may change from line to line. We also denote by $\|t\|_{\infty,A}$ the infinite norm of a function t over a set A, by $\|t\|_A$ its Hilbert norm, and by $\langle ., \rangle_A$ the associated scalar product.

5.1. **Preliminary result.** Let us start by setting a result which is the key argument in the proofs of the two main theorems. We consider the centered empirical process defined by

$$\forall t \in L^2([0;1] \times A_2),\ \nu_n(t) = \frac{1}{n} \sum_{i=1}^{n} t\left(F_X(X_i), Y_i\right) - \mathbb{E}\left[t\left(F_X(X_i), Y_i\right)\right]. \tag{18}$$

The aim of the following proposition is to control the deviations of the supremum of this process on the unit sphere of $\mathbb{S}_m$

$$\mathcal{S}(m) = \{t \in \mathbb{S}_m,\ \|t\| = 1\}. \tag{19}$$

$b(x)$	ε	X	$n=200$	500	2000	Method
$2x+5$	$\mathcal{N}(0,1)$	$\mathcal{U}_{[0;1]}$	3.13	2.89	1.49	$\tilde{\pi}_{BCL}$
			4.12	1.74 (−4%)	0.81 (−45%)	$\tilde{\pi}$
		$\mathcal{U}_{[-1;1]}$	6.63	5.36	3.96	$\tilde{\pi}_{BCL}$
			6.22 (−6%)	4.14 (−23%)	2.58 (−53%)	$\tilde{\pi}$
		$\mathcal{N}(0,1)$	26.94	23.89	22.61	$\tilde{\pi}_{BCL}$
			24.05 (−11%)	20.02 (−16%)	8.40 (−63%)	$\tilde{\pi}$
	$\Gamma(4,1)$	$\mathcal{U}_{[0;1]}$	1.56	1.27	0.77	$\tilde{\pi}_{BCL}$
			1.42 (−9%)	1.01 (−2%)	0.68 (−12%)	$\tilde{\pi}$
		$\mathcal{U}_{[-1;1]}$	3.63	2.98	1.93	$\tilde{\pi}_{BCL}$
			5.18	2.53 (−15%)	1.90 (−2%)	$\tilde{\pi}$
		$\mathcal{N}(0,1)$	14.41	13.39	12.20	$\tilde{\pi}_{BCL}$
			12.55 (−13%)	10.18 (−24%)	6.20 (−49%)	$\tilde{\pi}$
$\cos(x)$	$\mathcal{N}(0,1)$	$\mathcal{U}_{[0;1]}$	2.06	2.36	1.38	$\tilde{\pi}_{BCL}$
			2.34	0.92 (−61%)	0.43 (−69%)	$\tilde{\pi}$
		$\mathcal{U}_{[-1;1]}$	3.61	5.18	2.43	$\tilde{\pi}_{BCL}$
			5.43	1.65 (−68%)	0.81 (−69%)	$\tilde{\pi}$
		$\mathcal{N}(0,1)$	9.87	8.06	4.53	$\tilde{\pi}_{BCL}$
			14.64	6.26 (−22%)	3.20 (−67%)	$\tilde{\pi}$
	$\Gamma(4,1)$	$\mathcal{U}_{[0;1]}$	1.02	0.80	0.45	$\tilde{\pi}_{BCL}$
			0.69 (−32%)	0.49 (−39%)	0.32 (−29%)	$\tilde{\pi}$
		$\mathcal{U}_{[-1;1]}$	1.83	1.86	0.97	$\tilde{\pi}_{BCL}$
			1.27 (−31%)	0.94 (−49%)	0.68 (−30%)	$\tilde{\pi}$
		$\mathcal{N}(0,1)$	5.48	4.92	3.10	$\tilde{\pi}_{BCL}$
			4.43 (−19%)	3.53 (−28%)	2.55 (−18%)	$\tilde{\pi}$
x^2	$\mathcal{N}(0,1)$	$\mathcal{U}_{[0;1]}$	2.49	2.48	1.36	$\tilde{\pi}_{BCL}$
			2.89	1.35 (−46%)	0.60 (−56%)	$\tilde{\pi}$
		$\mathcal{U}_{[-1;1]}$	4.99	5.72	2.45	$\tilde{\pi}_{BCL}$
			5.39	2.03 (−65%)	0.88 (−64%)	$\tilde{\pi}$
		$\mathcal{N}(0,1)$	13.99	9.02	4.35	$\tilde{\pi}_{BCL}$
			23.21	14.92	8.72	$\tilde{\pi}$
	$\Gamma(4,1)$	$\mathcal{U}_{[0;1]}$	0.98	0.98	0.57	$\tilde{\pi}_{BCL}$
			0.88 (−10%)	0.60 (−38%)	0.54 (−5%)	$\tilde{\pi}$
		$\mathcal{U}_{[-1;1]}$	2.13	2.36	1.06	$\tilde{\pi}_{BCL}$
			1.44 (−32%)	1.23 (−48%)	1.02 (−4%)	$\tilde{\pi}$
		$\mathcal{N}(0,1)$	7.98	6.31	3.23	$\tilde{\pi}_{BCL}$
			14.13	7.53	4.61	$\tilde{\pi}$

TABLE 1. Values of MISE $\times 100$ averaged over 100 samples, in Examples 1 (regression models) for the estimators $\tilde{\pi}$ and $\tilde{\pi}_{BCL}$, with percentage of improvement (in parenthesis) of the warped-bases method with Goldenshluger-Lepski selection ($\tilde{\pi}$) compared to the least-squares method ($\tilde{\pi}_{BCL}$).

Proposition 3. *Under the assumptions of Theorem 1, for all $\delta > 0$, there exists a constant $C > 0$, depending on $\|h\|_\infty$, such that,*

$$\mathbb{E}\left[\max_{m' \in \mathcal{M}_n}\left(\sup_{t \in \mathcal{S}(m')} \nu_n^2(t) - 2(1+2\delta)\frac{D_{m'_1} D_{m'_2}}{n}\right)_+\right] \leq \frac{C}{n}.$$

Example	$n = 200$	500	2000	Method
Ex 2	1.94	1.97	1.07	$\tilde{\pi}_{BCL}$
	2.52	0.65 (−67%)	0.27 (−75%)	$\tilde{\pi}$
Ex 3	1.34	1.19	0.63	$\tilde{\pi}_{BCL}$
	1.48	1.04 (−13%)	0.69	$\tilde{\pi}$
Ex 4	11.72	11.85	10.82	$\tilde{\pi}_{BCL}$
	10.21 (−13%)	10.49 (−11%)	10.13 (−6%)	$\tilde{\pi}$

TABLE 2. Values of MISE ×100 averaged over 100 samples, in Example 2,3,4 for the estimators $\tilde{\pi}$ and $\tilde{\pi}_{BCL}$, with percentage of improvement (in parenthesis) of the warped-bases method with Goldenshluger-Lepski selection ($\tilde{\pi}$) compared to the least-squares method ($\tilde{\pi}_{BCL}$).

Proof of Proposition 3. We first bound the maximum by a sum:

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m')}\nu_n^2(t)-c(\delta)\frac{D_{m'_1}D_{m'_2}}{n}\right)_+\right]\leq\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\left(\sup_{t\in\mathcal{S}(m')}\nu_n^2(t)-c(\delta)\frac{D_{m'_1}D_{m'_2}}{n}\right)_+\right],$$

with the abbreviation $c(\delta)=2(1+2\delta)$ and we apply the following concentration inequality.

Lemma 4. *Let $\xi_1,\dots,\xi_n$ 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 $\delta>0$, there exist three constants c_l, $l=1,2,3$, such that*

$$\mathbb{E}\left[\left(\sup_{r\in\mathcal{R}}\left(\nu_n\left(r\right)\right)^2-c(\delta)H^2\right)_+\right] \leq c_1\left\{\frac{v}{n}\exp\left(-c_2\delta\frac{nH^2}{v}\right)+\frac{M_1^2}{C^2(\delta)n^2}\exp\left(-c_3C(\delta)\sqrt{\delta}\frac{nH}{M_1}\right)\right\}, \tag{20}$$

with, $C(\delta)=(\sqrt{1+\delta}-1)\wedge 1$, $c(\delta)=2(1+2\delta)$ 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 (20) is a classical consequence of Talagrand's Inequality given in Klein and Rio [KR05]: see for example Lemma 5 (page 812) in Lacour [Lac08]. Using density arguments, we can apply it to the unit sphere of a finite dimensional linear space, that is $\mathcal{S}(m')$, for our problem. We replace also the functions r by $r_t:(x,y)\mapsto t(F_X(x),y)$, and compute the constants M_1, H and v. Notice first that $\|r_t\|_\infty\leq\|t\|_\infty$, we deduce from Property (3) that we can set $M_1=\sqrt{D_{m'_1}D_{m'_2}}$. If $t\in\mathcal{S}(m')$, it can be written $t=\sum_{j=1}^{D_{m'_1}}\sum_{k=1}^{D_{m'_2}}b_{j,k}\varphi_j\otimes\varphi_k$, with $\sum_{j,k}b_{j,k}^2=1$. So, using the linearity of the process, and Cauchy-Schwarz's Inequality, we get $\sup_{t\in\mathcal{S}(m')}\nu_n(t)^2\leq\sum_{j=1}^{D_{m'_1}}\sum_{k=1}^{D_{m'_2}}\nu_n^2(\varphi_j\otimes\varphi_k)$. We use anew Property (3) to define H^2:

$$\mathbb{E}\left[\sup_{t\in\mathcal{S}(m')}\nu_n^2(t)\right]\leq\sum_{j=1}^{D_{m'_1}}\sum_{k=1}^{D_{m'_2}}\frac{1}{n}\mathrm{Var}\left(\varphi_j(F_X(X_1))\varphi_k(Y_1)\right)\leq\frac{D_{m'_1}D_{m'_2}}{n}:=H^2.$$

Finally, $\mathrm{Var}(t(F_X(X_1),Y_1))\leq\mathbb{E}[t^2(F_X(X_1),Y_1)]\leq\|t\|^2\|h\|_\infty=\|h\|_\infty:=v$. We just replace the quantities M_1,H and v by the values derived above in Inequality (20):

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\left(\sup_{t\in\mathcal{S}(m')}\nu_n(t)^2-c(\delta)\frac{D_{m'_1}D_{m'_2}}{n}\right)_+\right]$$
$$\leq c_1\left\{\sum_{m'\in\mathcal{M}_n}\frac{1}{n}\exp\left(-c_2D_{m'_1}D_{m'_2}\right)+\sum_{m'\in\mathcal{M}_n}\frac{D_{m'_1}D_{m'_2}}{n^2}\exp\left(-c_3\sqrt{n}\right)\right\}.$$

It remains to remark that the first sum is a constant and that $\sum_{m'\in\mathcal{M}_n}D_{m'_1}D_{m'_2}\leq n^2$ to conclude the proof.

□

5.2. **Proof of Theorem 1.** For the sake of simplicity, we denote in this section by $\hat{m}$ the selected index $\hat{m}^{F_X}$, by V the penalty V^{F_X}, and by A the quantity $A(.,F_X)$. Let $\mathbb{S}_m$ be a fixed model in the collection indexed by $\mathcal{M}_n$.

5.2.1. *Main part of the proof.* We decompose the loss of the estimator as follows:

$$\|\tilde{\pi}_0-\pi\|^2_{f_X}=\left\|\hat{h}^{F_X}_{\hat{m}}-h\right\|^2\leq 3\left\|\hat{h}^{F_X}_{\hat{m}}-\hat{h}^{F_X}_{m\wedge\hat{m}}\right\|^2+3\left\|\hat{h}^{F_X}_{m\wedge\hat{m}}-\hat{h}^{F_X}_{m}\right\|^2+3\left\|\hat{h}^{F_X}_{m}-h\right\|^2.$$

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

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

We have already bounded the risk of the estimator on a fixed model (see Section 2.3.1, Inequalities (8) and (10)), therefore, by definition of V, we get

$$\mathbb{E}\left[\left\|\hat{h}^{F_X}_{\hat{m}}-h\right\|^2\right]\leq 3\mathbb{E}\left[A(m)\right]+(6c_1+3)\frac{D_{m_1}D_{m_2}}{n}+3\left\|h_m-h\right\|^2. \tag{21}$$

To pursue the proof, we have to control the expectation of $A(m)$. By splitting the norm $\|\hat{h}^{F_X}_{m'}-\hat{h}^{F_X}_{m\wedge m'}\|^2$ for $m,m'\in\mathcal{M}_n$, and using the definition of A, we get

$$\begin{aligned}A(m) &\leq 3\max_{m'\in\mathcal{M}_n}\left[\left\|\hat{h}^{F_X}_{m'}-h_{m'}\right\|^2-\frac{V(m')}{6}\right]_+ +3\max_{m'\in\mathcal{M}_n}\left[\left\|h_{m\wedge m'}-\hat{h}^{F_X}_{m\wedge m'}\right\|^2-\frac{V(m')}{6}\right]_+\\ &\quad +3\max_{m'\in\mathcal{M}_n}\left\|h_{m'}-h_{m\wedge m'}\right\|^2.\end{aligned}$$

The three terms of the above decomposition are studied in the following lemmas, proved just below.

Lemma 5. *Under the assumptions of Theorem 1, there exists a constant $C>0$ such that, for $m\in\mathcal{M}_n$,*

$$(a)\ \mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\left\|\hat{h}^{F_X}_{m'}-h_{m'}\right\|^2-\frac{V(m')}{6}\right)_+\right]\leq\frac{C}{n},$$

$$(b)\ \mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\left\|h_{m\wedge m'}-\hat{h}^{F_X}_{m\wedge m'}\right\|^2-\frac{V(m')}{6}\right)_+\right]\leq\frac{C}{n}.$$

Lemma 6. *Under the assumptions of Theorem 1, there exists a constant $C>0$ such that,*

$$\max_{m'\in\mathcal{M}_n}\left\|h_{m'}-h_{m\wedge m'}\right\|^2\leq 4\|h_m-h\|^2.$$

These inequalities show that

$$\mathbb{E}[A(m)] \leq \frac{C}{n} + 4\|h_m - h\|^2. \tag{22}$$

Gathering this with Inequality (21) ends the proof of the Theorem.

□

5.2.2. *Proof of Lemma 5.* To simplify the notations, we denote by $T_p = \|\hat{h}_p^{F_X} - h_p\|^2$ for $p = m'$ or $p = m \wedge m'$, and by $U_p = (T_p - V(m'))_+$.

Inequality (a). We compute first classically

$$\left\|\hat{h}_{m'}^{F_X} - h_{m'}\right\|^2 = \sum_{j=1}^{D_{m'_1}} \sum_{k=1}^{D_{m'_2}} \left(\hat{a}_{j,k}^{F_X} - a_{j,k}\right)^2 = \sum_{j=1}^{D_{m'_1}} \sum_{k=1}^{D_{m'_2}} \nu_n^2(\varphi_j \otimes \varphi_k) = \sup_{t \in \mathcal{S}(m')} \nu_n^2(t), \tag{23}$$

with ν_n the empirical process defined by (18). Thus,

$$\mathbb{E}\left[\max_{m' \in \mathcal{M}_n} U_{m'}\right] = \mathbb{E}\left[\max_{m' \in \mathcal{M}_n} \left(\sup_{t \in \mathcal{S}(m')} \nu_n^2(t) - \frac{V(m')}{6}\right)_+\right],$$

and Inequality (a) of the lemma is proved by applying Proposition 3.

Inequality (b). We have to distinguish several cases, depending on the value of $m \wedge m'$: $\max_{m' \in \mathcal{M}_n} U_{m \wedge m'}$

$$\leq \max_{\substack{m' \in \mathcal{M}_n \\ m'_1 \leq m_1, m'_2 \leq m_2}} U_{m \wedge m'} + \max_{\substack{m' \in \mathcal{M}_n \\ m_1 \leq m'_1, m_2 \leq m'_2}} U_{m \wedge m'} + \max_{\substack{m' \in \mathcal{M}_n \\ m'_1 \leq m_1, m_2 \leq m'_2}} U_{m \wedge m'} + \max_{\substack{m' \in \mathcal{M}_n \\ m_1 \leq m'_1, m'_2 \leq m_2}} U_{m \wedge m'}.$$

- *First term:* $m'_1 \leq m_1$ and $m'_2 \leq m_2$. In this case, $m \wedge m' = m'$. Thus, we bound roughly

$$\mathbb{E}\left[\max_{\substack{m' \in \mathcal{M}_n \\ m'_1 \leq m_1, m'_2 \leq m_2}} U_{m \wedge m'}\right] \leq \mathbb{E}\left[\max_{m' \in \mathcal{M}_n} U_{m'}\right],$$

  and use Inequality (a) to conclude that this term is bounded by C/n.
- *Second term:* $m_1 \leq m'_1$ et $m_2 \leq m'_2$. Here, $m \wedge m' = m$. Using $V(m) \leq V(m')$ (because $D_{m_l} \leq D_{m'_l}$, $l = 1, 2$), we have,

$$\mathbb{E}\left[\max_{\substack{m' \in \mathcal{M}_n \\ m_1 \leq m'_1, m_2 \leq m'_2}} U_{m \wedge m'}\right] \leq \mathbb{E}\left[\max_{\substack{m' \in \mathcal{M}_n \\ m_1 \leq m'_1, m_2 \leq m'_2}} \left(T_m - \frac{V(m)}{6}\right)_+\right] = \mathbb{E}\left[\left(T_m - \frac{V(m)}{6}\right)_+\right],$$

  and it can be seen as a consequence of Proposition 3 and of the beginning of the proof of Inequality (a) that this last term is bounded by C/n.

- *Third term:* $m_1' \leq m_1$ et $m_2 \leq m_2'$. Here, $m \wedge m' = (m_1', m_2)$. We use thus $V((m_1', m_2)) \leq V(m_1', m_2')$ to get

$$\begin{aligned}\mathbb{E}\left[\max_{\substack{m' \in \mathcal{M}_n \\ m_1' \leq m_1, m_2 \leq m_2'}} U_{m \wedge m'}\right] &\leq \mathbb{E}\left[\max_{\substack{m' \in \mathcal{M}_n \\ m_1' \leq m_1, m_2 \leq m_2'}} \left(T_{(m_1', m_2)} - \frac{V((m_1', m_2))}{6}\right)_+\right], \\ &\leq \sum_{m_1' \in \mathcal{I}_n^{(1)}} \mathbb{E}\left[\left(T_{(m_1', m_2)} - \frac{V((m_1', m_2))}{6}\right)_+\right].\end{aligned}$$

  The last term is also bounded by C/n, using a slightly different version of Proposition 3 (take the maximum only over $m_1' \in \mathcal{I}_n^{(1)}$ instead of over $m \in \mathcal{M}_n$, and replace m by $m \wedge m'$).
- *Fourth term:* $m_1 \leq m_1'$ et $m_2' \leq m_2$. We deal with this case by using the same arguments as for the last case.

We conclude that $\mathbb{E}[\max_{m' \in \mathcal{M}_n} U_{m \wedge m'}]$ is upper-bounded by C/n.

□

5.2.3. *Proof of Lemma 6.* Following the same lines as in the proof of Lemma 5, we distinguish four cases:

- $m_1' \leq m_1$ and $m_2' \leq m_2$. For such couples (m_1, m_2) and (m_1', m_2'), $\|h_{m'} - h_{m \wedge m'}\|^2 = 0$.
- $m_1 \leq m_1'$ et $m_2 \leq m_2'$. We notice first that $\|h_{m'} - h_{m \wedge m'}\|^2 = \|h_{m'} - h_m\|^2 \leq 2\|h_{m'} - h\|^2 + 2\|h_m - h\|^2$. Since the models are nested in each direction (see Property (4)), we have $\mathbb{S}_m = S_{m_1} \times S_{m_2} \subset S_{m_1'} \times S_{m_2'} = \mathbb{S}_{m'}$. Consequently, $h_m \in \mathbb{S}_{m'}$, and by the definition of the orthogonal projection onto $\mathbb{S}_{m'}$, we get $\|h_{m'} - h\| \leq \|h_m - h\|$. This leads to $\|h_{m'} - h_{m \wedge m'}\|^2 \leq 4\|h_m - h\|^2$.
- $m_1' \leq m_1$ et $m_2 \leq m_2'$. To deal with this case, we use first the following remark: if t belongs to $L^2([0;1] \times A_2)$, then for all $u \in [0;1]$, $y \mapsto t(u,y)$ belongs to $L^2(A_2)$ and $y \in A_2$, $u \mapsto t(u,y)$ belongs to $L^2([0;1])$. Moreover, denoted by G_1 (respectively G_2) a closed linear subspace of $L^2([0;1])$ (respectively of $L^2(A_2)$), and by Π_G the projection operator onto a subspace G, the following equality holds:

$$\Pi_{G_1 \times G_2} t = \Pi_{G_1 \times L^2(A_2)} \left(\Pi_{L^2([0;1]) \times G_2} t\right).$$

  In our setting, we thus compute

$$\begin{aligned}\|h_{m'} - h_{m \wedge m'}\|^2 &= \left\|\Pi_{S_{m_1'} \times L^2(A_2)} \left[\Pi_{L^2([0;1]) \times S_{m_2'}} h - \Pi_{L^2([0;1]) \times S_{m_2}} h\right]\right\|^2, \\ &\leq \left\|\Pi_{L^2([0;1]) \times S_{m_2'}} h - \Pi_{L^2([0;1]) \times S_{m_2}} h\right\|^2, \\ &\leq 2\left\|\Pi_{L^2([0;1]) \times S_{m_2'}} h - h\right\|^2 + 2\left\|\Pi_{L^2([0;1]) \times S_{m_2}} h - h\right\|^2, \\ &\leq 4\left\|\Pi_{L^2([0;1]) \times S_{m_2}} h - h\right\|^2 \leq 4\|h_m - h\|^2,\end{aligned}$$

  where the inequalities of the last line are obtained by noticing that $S_{m_2} \subset S_{m_2'}$ and that $S_{m_1} \subset L^2([0;1])$, and by using the definition of orthogonal projections.

- $m_1 \leq m_1'$ et $m_2' \leq m_2$. This case is the symmetric from the latter, and can be thus handled similarly.

Gathering the bounds of the four cases and taking the maximum of the four upper-bounds lead to the conclusion:

$$\max_{m' \in \mathcal{M}_n} \|h_{m'} - h_{m \wedge m'}\|^2 \leq \max\left\{0, 4\|h_m - h\|^2\right\} = 4\|h_m - h\|^2.$$

□

5.3. **Proof of Theorem 2.** To simplify the notations, we write in this section $A(m)$ to replace $A(m, \hat{F}_n)$, V for $V^{\hat{F}}$, and $\hat{m}$ instead of $\hat{m}^{\hat{F}}$. The main idea of the proof is to recover the framework of the proof of Theorem 1. The computation are more technical, since the estimator $\tilde{\pi} = \hat{h}^{\hat{F}}_{\hat{m}}(\hat{F}(.),.)$ depends doubly on $\hat{F}$. We denote it by $\hat{\pi}^{\hat{F},\hat{F}}_{\hat{m}}$, and coherently, we denote by $\hat{\pi}^{F_X,F_X}_{\hat{m}}$ the estimator previously studied, that is $\tilde{\pi}_0$. We also introduce the following intermediate:

$$\forall (x,y) \in A_1 \times A_2,\ \hat{\pi}^{\hat{F},F_X}(x,y) = \hat{h}^{\hat{F}}_{\hat{m}}(F_X(x), y). \tag{24}$$

These notations suit also well for a fixed index $m \in \mathcal{M}_n$. We denote by $\mathbb{E}[.|(X_{-l})]$ the conditional expectation given the sample $(X_{-l})_{l=1,\dots,n}$ (the conditional expectation will be coherently denoted by $\mathrm{Var}(.|(X_{-l}))$). A key point is the following decomposition which holds for any index m: $\|\hat{\pi}^{\hat{F},\hat{F}}_m - \pi\|^2_{f_X} \leq 6\sum_{l=0}^4 T^m_l$, with

$$\begin{aligned}
T^m_0 &= \|\pi - \pi^{F_X}_m\|^2_{f_X} + \|\pi^{F_X}_m - \hat{\pi}^{F_X,F_X}_m\|^2_{f_X},\\
T^m_1 &= \left\|\hat{\pi}^{F_X,F_X}_m - \hat{\pi}^{\hat{F},F_X}_m - \mathbb{E}\left[\hat{\pi}^{F_X,F_X}_m - \hat{\pi}^{\hat{F},F_X}_m\,|(X_{-l})_l\right]\right\|^2_{f_X},\\
T^m_2 &= \left\|\hat{\pi}^{\hat{F},F_X}_m - \hat{\pi}^{\hat{F},\hat{F}}_m - \mathbb{E}\left[\hat{\pi}^{\hat{F},F_X}_m - \hat{\pi}^{\hat{F},\hat{F}}_m\,|(X_{-l})_l\right]\right\|^2_{f_X},\\
T^m_3 &= \left\|\mathbb{E}\left[\hat{\pi}^{F_X,F_X}_m - \hat{\pi}^{\hat{F},F_X}_m\,|(X_{-l})_l\right]\right\|^2_{f_X},\quad T^m_4 = \left\|\mathbb{E}\left[\hat{\pi}^{\hat{F},F_X}_m - \hat{\pi}^{\hat{F},\hat{F}}_m\,|(X_{-l})_l\right]\right\|^2_{f_X}.
\end{aligned} \tag{25}$$

Let us remark that T^m_0 is the bias-variance decomposition for the risk of an estimator $\hat{\pi}^{F_X,F_X}_m$, and has already been studied (see Section 2.3.1). The sketch of the proof is now to decompose the loss function, using these intermediates and the definition of A and V, and then to bound each of the terms by $CD_{m_1}D_{m_2}/n$ or to center them (so as to show they are negligible).

5.3.1. *Main part of the proof.* We begin by introducing the intermediate estimator defined by (24) in the loss of our estimator:

$$\begin{aligned}
\left\|\hat{\pi}^{\hat{F},\hat{F}}_{\hat{m}} - \pi\right\|^2_{f_X} &\leq 3\left\|\hat{\pi}^{\hat{F},\hat{F}}_{\hat{m}} - \hat{\pi}^{\hat{F},F_X}_{\hat{m}} - \mathbb{E}\left[\hat{\pi}^{\hat{F},\hat{F}}_{\hat{m}} - \hat{\pi}^{\hat{F},F_X}_{\hat{m}}\,|(X_{-l})_l\right]\right\|^2_{f_X}\\
&\quad + 3\left\|\mathbb{E}\left[\hat{\pi}^{\hat{F},\hat{F}}_{\hat{m}} - \hat{\pi}^{\hat{F},F_X}_{\hat{m}}\,|(X_{-l})_l\right]\right\|^2_{f_X} + 3\left\|\hat{\pi}^{\hat{F},F_X}_{\hat{m}} - \pi\right\|^2_{f_X},\\
&= 3T^{\hat{m}}_2 + 3T^{\hat{m}}_4 + 3\left\|\hat{h}^{\hat{F}}_{\hat{m}} - h\right\|^2.
\end{aligned}$$

The last term can be itself decomposed, by construction of A, V, and $\hat{m}$:

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

Furthermore, $\|\hat{h}_m^{\hat{F}} - h\|^2 = \|\hat{\pi}_m^{\hat{F},F_X} - \pi\|_{f_X}^2 = 3T_1^m + 3T_3^m + 3T_0^m$. Consequently,

$$\left\|\hat{\pi}_{\hat{m}}^{\hat{F},\hat{F}} - \pi\right\|_{f_X}^2 \leq 3T_2^{\hat{m}} + 3T_4^{\hat{m}} - 3\times 2V(\hat{m}) + 3\times 6\left(A(m) + 2V(m)\right) + 3\times 3\times\left(3T_1^m + 3T_3^m + 3T_0^m\right), \tag{26}$$

where the terms T_l^m, $l = 0,\dots,4$ are defined by (25). We split the term A, first in a similar way as in Theorem 1. Let $(m,m') \in \mathcal{M}_n^2$,

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

But we immediatly try to recover the splitting terms defined by (25). Let us remark that an analogous relation to (23) holds, for a different empirical process: for $p = m$ or $p = m\wedge m'$,

$$\left\|h_p - \hat{h}_p^{\hat{F}}\right\| = \sup_{t\in\mathcal{S}(p)} \tilde{\nu}_n^2(t), \quad \tilde{\nu}_n(t) = \frac{1}{n}\sum_{i=1}^n t\left(\hat{F}_n(X_i), Y_i\right) - \mathbb{E}\left[t\left(F_X(X_i), Y_i\right)\right],$$

for a function $t \in L^2([0;1]\times A_2)$. We recover the previous empirical process by the decomposition $\tilde{\nu}_n^2(t) \leq 2\nu_n^2(t) + R_n^2(t)$, with $R_n(t) = (1/n)\sum_{i=1}^n t(\hat{F}_n(X_i), Y_i) - t(F_X(X_i), Y_i)$. Moreover, if t belongs to $\mathcal{S}(p)$, we have already written $t = \sum_{j=1}^{D_{p_1}}\sum_{k=1}^{D_{p_2}} \theta_{j,k}\varphi_j\otimes\varphi_k$, with $\sum_{j=1}^{D_{p_1}}\sum_{k=1}^{D_{p_2}}\theta_{j,k}^2 = 1$. Using this expression, Cauchy-Schwarz Inequality, and the definition of the coefficients $\hat{a}_{j,k}^{F_X}$ or $\hat{a}_{j,k}^{\hat{F}}$ yield $\sup_{t\in\mathcal{S}(p)} R_n^2(t) = \sum_{j=1}^{D_{p_1}}\sum_{k=1}^{D_{p_2}}(\hat{a}_{j,k}^{\hat{F}} - \hat{a}_{j,k}^{F_X})^2$. The conditional expectation of $\hat{a}_{j,k}^{\hat{F}} - \hat{a}_{j,k}^{F_X}$ is introduced to get $\sup_{t\in\mathcal{S}(p)} R_n^2(t) \leq 2T_1^p + 2T_3^p$. Consequently,

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

By substracting $V(m')$, taking the maximum over $m' \in \mathcal{M}_n$ and integrating give an upper-bound for $\mathbb{E}[A(m)]$. We introduce it into (26) to obtain:

$$\begin{aligned}
(\Delta)\quad & \mathbb{E}\left[\left\|\hat{\pi}_{\hat{m}}^{\hat{F},\hat{F}} - \pi\right\|_{f_X}^2\right] \\
\leq\ & 36V(m) + 27\mathbb{E}\left[T_0^m + T_1^m + T_3^m\right] + 3\max_{m'\in\mathcal{M}_n}\left\|h_{m\wedge m'} - h_{m'}\right\|^2 \\
& + 3\mathbb{E}\left[\left(T_2^{\hat{m}} - V(\hat{m})\right)_+\right] + 3\mathbb{E}\left[\left(T_4^{\hat{m}} - V(\hat{m})\right)_+\right] \\
& + 6\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m')}\left(\nu_n(t)\right)^2 - \frac{V(m')}{36}\right)_+\right] + 6\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m\wedge m')}\left(\nu_n(t)\right)^2 - \frac{V(m')}{36}\right)_+\right] \\
& + 12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m'} - \frac{V(m')}{72}\right)_+\right] + 12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m\wedge m'} - \frac{V(m')}{72}\right)_+\right] \\
& + 12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_1^{m'} - \frac{V(m')}{72}\right)_+\right] + 12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_1^{m\wedge m'} - \frac{V(m')}{72}\right)_+\right].
\end{aligned}$$

We bound each of these terms. Some of them have already been studied: recall first that

$$\mathbb{E}\left[T_0^m\right] \leq \left\|\pi_m^{F_X} - \pi\right\|^2 + \frac{D_{m_1}D_{m_2}}{n},$$

using (8) and (10). Moreover, applying twice Proposition 3 shows that

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m')}(\nu_n(t))^2-V_0(m')\right)_+\right]\leq\frac{C}{n},$$
$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(\sup_{t\in\mathcal{S}(m\wedge m')}(\nu_n(t))^2-V_0(m')\right)_+\right]\leq\frac{C}{n},$$

with $V_0(m')=2(1+2\delta)D_{m'_1}D_{m'_2}/n$. Choosing c_1 (see the definition (11)) larger than $2(1+2\delta)$, these inequalities hold with V in place of V_0. Finally, we have proved in Lemma 6 that $\max_{m'\in\mathcal{M}_n}\|h_{m'}-h_{m\wedge m'}\|^2\leq 4\|h_m-h\|^2$. Taking into account the previous inequality (Δ) for the risk, we get,

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\pi}_{\hat{m}}^{\hat{F},\hat{F}}-\pi\right\|_{f_X}^2\right] \;\leq\;& 36V(m)+27\frac{D_{m_1}D_{m_2}}{n}+27\mathbb{E}\left[T_1^m+T_3^m\right]+\frac{C}{n}\\ &+3\mathbb{E}\left[\left(T_2^{\hat{m}}-V(\hat{m})\right)_+\right]+3\mathbb{E}\left[\left(T_4^{\hat{m}}-V(\hat{m})\right)_+\right]+(12+27)\|h_m-h\|^2\\ &+12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m'}-\frac{V(m')}{72}\right)_+\right]+12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{m\wedge m'}-\frac{V(m')}{72}\right)_+\right]\\ &+12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_1^{m'}-\frac{V(m')}{72}\right)_+\right]+12\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_1^{m\wedge m'}-\frac{V(m')}{72}\right)_+\right].\end{aligned}\tag{27}$$

It remains to bound the terms T_l^m, $l=1,2,3,4$ or their centering versions, by quantities of order at most $D_{m_1}D_{m_2}/n$. Let us first notice that, for $l=2,4$,

$$\mathbb{E}\left[\left(T_l^{\hat{m}}-V(\hat{m})\right)_+\right]\leq\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_l^{m'}-V(m')\right)_+\right],$$

and then use the lemmas just below, whose proofs are deferred to the following sections.

Lemma 7. *Assuming that the models are trigonometric, there exists a constant C depending only on $\|\varphi_2'\|_\infty$ such that, for $m\in\mathcal{M}_n$,*

$$\mathbb{E}\left[T_1^m\right]\leq C\frac{D_{m_1}^3D_{m_2}}{n^2}.$$

Moreover, the following inequality holds, if $D_{m_1}=O(\sqrt{n}/\ln(n))$, for $p_{m'}=m'$ or $p_{m'}=m\wedge m'$, and for a constant $C>0$

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_1^{p_{m'},b}-V_1(m')\right)_+\right]\leq\frac{C}{n},$$

with $V_1(m')=\kappa_1D_{m'_1}D_{m'_2}/n$, and κ_1 a constant depending only on $\|\varphi_2'\|_\infty$.

If $D_{m_1}=O(n^{1/2})$ in particular, the first inequality of Lemma 7 leads to $\mathbb{E}[T_1^m]\leq CD_{m_1}D_{m_2}$.

Lemma 8. *Assuming that the models are trigonometric, there exists a constant C, which depends on $\|\varphi_2'\|_{\infty,[0;1]}$, 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 $V_2(m')=\kappa_2D_{m'_1}^4D_{m'_2}\ln^2(n)/n^2$, and κ_2 a constant depending also on $\|\varphi_2'\|_{\infty,[0;1]}$.

Assuming that $D_{m'_1}=O(n^{1/3}/\ln^{2/3}(n))$, we have $V_2(m')\leq V_2^b(m'):=\kappa_2'D_{m'_1}D_{m'_2}/n$ (κ_2' a constant independent of h). The inequality of Lemma 8 still holds by replacing V_2 by V_2^b.

Lemma 9. *Assuming that the models are trigonometric, and that h is $\mathcal{C}^1$ with respect to its first variable on $[0;1]$, there exists a constant C depending on $\|\varphi_2^{(3)}\|_\infty$, $\|h\|$ and $\|\partial_1 h\|$ (∂_1 is the derivation operator with respect to the first variable) such that, for $m \in \mathcal{M}_n$,*

$$\mathbb{E}\left[T_3^m\right] \leq C\left(\frac{1}{n} + \frac{D_{m_1}}{n} + \frac{D_{m_1}^4}{n^2} + \frac{D_{m_1}^7}{n^3}\right).$$

Moreover, the following inequality holds, for $p_{m'} = m'$ or $p_{m'} = m \wedge m'$, for $n \geq n_0(h)$, and assuming $D_{m_1} = O(n^{1/3})$ and $D_{m_2} \geq c\ln^4(n)$ (for a constant $c > 0$) for each 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 $V_3(m') = \kappa_3 \frac{D_{m_1'} D_{m_2'}}{n}$, κ_3 a constant independent of h, and $n_0(h)$ a nonnegative integer depending on the function h.

If $D_{m_1} = O(n^{1/3})$ in particular, the first inequality of Lemma 7 leads to $\mathbb{E}[T_3^m] \leq C D_{m_1} D_{m_2}/n$.

Lemma 10. *Assuming that the models are trigonometric, that h is $\mathcal{C}^1$ with respect to its first variable on $[0;1]$ and belongs to the anisotropic Sobolev ball denoted by $W^2_{per}([0;1]^2, L, (1,0))$, and that for all $m \in \mathcal{M}_n$, $D_{m_1} = O(n^{1/3}/\ln^{1/3}(n))$ and $D_{m_2} \geq c\ln^5(n)$ (for a constant $c > 0$), there exists a constant C, which depends on $\|\varphi_2'\|_{\infty,[0;1]}$, $\|\varphi_2''\|_{\infty,[0;1]}$, $\|\varphi_2^{(3)}\|_{\infty,[0;1]}$, $\|h\|$, $\|\partial_1 h\|$, and L such that, for $n \geq n_1(h)$,*

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

with $V_4(m') = \kappa_4 D_{m_1'} D_{m_2'}/n$, and κ_4 independent of h, and $n_1(h)$ a nonnegative integer depending on the function h.

To conclude the proof, we choose the constant c_1 larger than κ_l ($l = 1, \ldots, 4$), to have $V(m') \geq V_l(m)$ (or $V_l^b(m')$ for $l = 2$): this allows to apply the inequalities of the lemmas with V and to use it in Inequality (27). We obtain then the result of Theorem 2.

□

5.3.2. *Technical tools for the proof of Lemmas 7 to 10.* Key arguments for the proof of the lemmas are the properties of the empirical cumulative distribution function $\hat{F}_n$ of the sample $(X_{-l})_l$. First, let $U_{-i} = F_X(X_{-i})$ ($i = 1, \ldots, n$). Recall that it is a uniform variable on $[0;1]$. We denote by $\hat{U}_n$ the empirical c.d.f. associated to the sample $(U_{-i})_{i=1,\ldots,n}$. Let us keep also in mind that for all $u \in [0;1]$, $\hat{F}_n(F_X^{-1}(u)) = \hat{U}_n(u)$ and that the random variable $\|\hat{F}_n - F_X\|_{\infty,A_1}$ has the same distribution as $\|\hat{U}_n - id\|_{\infty,[0;1]}$ (with id the function such that $u \mapsto u$). In particular, we get thus

$$\mathbb{E}\left[\hat{a}_{j,k}^{\hat{F}} \,|(X_{-l})_l\right] = \int_{[0;1]\times A_2} \varphi_j \circ \hat{U}_n(u)\varphi_k(y)h(u,y)dudy.$$

We also recall some inequalities to control the deviations of the empirical c.d.f $\hat{U}_n$. Dvoretzky, Kiefer and Wolfowitz [DKW56] established the first one:

Proposition 11. *For any $\lambda > 0$, there exists a constant K such that*

$$\mathbb{P}\left(\left\|\hat{U}_n - id\right\|_{\infty,[0;1]} \geq \lambda\right) \leq K\exp\left(-2n\lambda^2\right).$$

By integration, we deduce then other bounds:

Proposition 12. *For any integer $p > 0$, there exists a constant $C_p > 0$ such that*

$$\mathbb{E}\left[\left\|\hat{U}_n - id\right\|^p_{\infty,[0;1]}\right] \leq \frac{C_p}{n^{p/2}}, \tag{28}$$

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

$$\mathbb{E}\left[\left(\left\|\hat{U}_n - id\right\|^p_{\infty,[0;1]} - \kappa\frac{\ln^{p/2}(n)}{n^{p/2}}\right)_+\right] \leq Cn^{-c(p,\kappa)}, \text{ with } c(p,\kappa) = 2^{\frac{2-p}{p}}\kappa^{2/p}. \tag{29}$$

Moreover,

$$\mathbb{E}\left[\left(\left\|\hat{U}_n - id\right\|^2_{\infty,[0;1]} - \kappa\frac{\ln(n)}{n}\right)^2\right] \leq Cn^{-2-2\kappa}. \tag{30}$$

Inequality (30) is a slightly more precise version of Inequality (29) in the case $p = 2$.

5.3.3. *Proof of Lemma 7.* The first part of the lemma is to bound $\mathbb{E}[T_1^m]$. Using the definition of $\hat{\pi}^{F_X,F_X}$ and $\hat{\pi}^{\hat{F},F_X}$ leads to

$$T_1^m = \left\|\hat{h}_m^{F_X} - \hat{h}_m^{\hat{F}} - \mathbb{E}\left[\hat{h}_m^{F_X} - \hat{h}_m^{\hat{F}}\,|(X_{-l})_l\right]\right\|^2.$$

The decompositions of the estimators in the orthonormal basis $(\varphi_j \otimes \varphi_k)$ yield $T_1^m = \sum_{j,k}\{(\hat{a}_{j,k}^{F_X} - \hat{a}_{j,k}^{\hat{F}}) - \mathbb{E}[\hat{a}_{j,k}^{F_X} - \hat{a}_{j,k}^{\hat{F}}|(X_{-l})]\}^2$. Thus,

$$\mathbb{E}\left[T_1^m\,|(X_{-l})_l\right] = \sum_{j,k}\mathrm{Var}\left(\hat{a}_{j,k}^{F_X} - \hat{a}_{j,k}^{\hat{F}}\,|(X_{-l})_l\right). \tag{31}$$

We work out the conditional variance for any couple (j,k):

$$\begin{aligned}\mathrm{Var}\left(\hat{a}_{j,k}^{F_X} - \hat{a}_{j,k}^{\hat{F}}\,|(X_{-l})_l\right) &= \frac{1}{n}\mathrm{Var}\left(\varphi_j\left(F_X(X_1)\right)\varphi_k(Y_1) - \varphi_j\left(\hat{F}_n(X_1)\right)\varphi_k(Y_1)\,|(X_{-l})_l\right),\\ &\leq \frac{1}{n}\mathbb{E}\left[\varphi_k^2(Y_1)\left\{\varphi_j\left(F_X(X_1)\right) - \varphi_j\left(\hat{F}_n(X_1)\right)\right\}^2|(X_{-l})_l\right].\end{aligned}$$

We apply the mean value theorem, sum over the indices j and k, and remark $\|\varphi_j'\|_{\infty,[0;1]} \leq D_{m_1}\|\varphi_2'\|_{\infty,[0;1]}$ (property of the trigonometric basis):

$$\begin{aligned}\mathbb{E}\left[T_1^m\,|(X_{-l})_l\right] &\leq \frac{1}{n}\left\|\sum_{k=1}^{D_{m_2}}\varphi_k\right\|^2_{\infty,[0;1]}\sum_{j=1}^{D_{m_1}}\|\varphi_j'\|^2_{\infty,[0;1]}\left\|F_X - \hat{F}_n\right\|^2_{\infty,A_1},\\ &\leq \|\varphi_2'\|^2_{\infty,[0;1]}\frac{D_{m_1}^3 D_{m_2}}{n}\left\|F_X - \hat{F}_n\right\|^2_{\infty,A_1}.\end{aligned}$$

It remains to use Inequality (28) of Proposition 12 with $p = 2$ to bound the expectation:

$$\mathbb{E}\left[T_1^m\right] \leq C\|\varphi_2'\|^2_{\infty,[0;1]}\frac{D_{m_1}^3 D_{m_2}}{n^2}$$

This completes the proof of the first inequality. For the second, let us begin with $V_1(p_{m'}) \leq V_1(m')$. Therefore $\mathbb{E}[\max_{m'\in\mathcal{M}_n}(T_1^{p_{m'}} - V_1(m'))_+] \leq \mathbb{E}[\max_{m'\in\mathcal{M}_n}(T_1^{p_{m'}} - V_1(p_{m'}))_+]$. In the sequel, we simplify the notations by setting $p = p_{m'}$. Similar arguments than the ones used to get (23) lead to $T_1^p = \sup_{t\in\mathcal{S}(p)}\left(\nu_n^a(t)\right)^2$ with

$$\nu_n^a(t) = \frac{1}{n}\sum_{i=1}^{n}\left(t(F_X(X_i),Y_i) - t(\hat{F}_n(X_i),Y_i)\right) - \mathbb{E}\left[\left(t(F_X(X_i),Y_i) - t(\hat{F}_n(X_i),Y_i)\right)|(X_{-l})_l\right],$$

a process which is centered conditionally to the sample $(X_{-l})_l$. Thus we apply Talagrand inequality (20), as in the proof of Proposition 3, but conditionally to $(X_{-l})_l$. In this setting the key quantities are such that

$$\sup_{t\in\mathcal{S}(p)}\|r_t\|_\infty \le M_{1,a},\ \mathbb{E}\left[\sup_{t\in\mathcal{S}(p)}|\nu_n^a(t)|\,|(X_{-l})_l\right] \le H_a,$$
$$\text{and}\ \sup_{t\in\mathcal{S}(p)}\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left(r_t(X_i,Y_i)\,|(X_{-l})_l\right) \le v_a.$$

We compute

$$M_{1,a} = \|\varphi_2'\|_{\infty,[0;1]}\, D_{p_1}^{3/2} D_{p_2}^{1/2}\left\|\hat{F}_n - F_X\right\|_{\infty,A_1},$$
$$H_{a,p}^2 = \tfrac{1}{n}\|\varphi_2'\|_{\infty,[0;1]}^2\, D_{p_1}^3 D_{p_2}\left\|\hat{F}_n - F_X\right\|_{\infty,A_1}^2,\quad v_a = nH_{a,p}^2,$$

and obtain thus for $\delta>0$, $\mathbb{E}\left[\left(\sup_{t\in\mathcal{S}(p)}\left(\nu_n^a(t)\right)^2 - 2(1+2\delta)H_{a,p}^2\right)_+ |(X_{-l})_l\right]$

$$\le C_0\left\{H_{a,p}^2\exp\left(-C\delta\right) + \frac{H_{a,p}^2}{C^2(\delta)n}\exp\left(-C\sqrt{\delta}\sqrt{n}\right)\right\}.$$

Here, C_0 is a random constant, which depends on $\|F_X - \hat{F}_n\|_{\infty,A_1}$, and C is purely numerical. But C_0 can be also bounded by a fixed quantity, since the infinite norm is smaller than 1. Thus we write anew C in the sequel. We choose $\delta = \kappa\ln(n)$ ($\kappa>0$), so that $C(\delta)=1$. We put now $p = m'$ (The case $p = m\wedge m'$ can be handled similarly). We have thus $\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_1^{m'} - 2(1+2\kappa\ln(n))H_{a,m'}^2\right)_+ |(X_{-l})_l\right]$

$$\begin{aligned}
&\le \sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\left(T_1^{m'} - 2(1+2\kappa\ln(n))H_{a,m'}^2\right)_+ |(X_{-l})_l\right],\\
&\le C\left\{n^{-C\kappa}\sum_{m'\in\mathcal{M}_n}\frac{D_{m_1'}^3 D_{m_2'}}{n} + \exp\left(-C\sqrt{n}\right)\sum_{m'\in\mathcal{M}_n}\frac{D_{m_1'}^3 D_{m_2'}}{n^2}\right\},\\
&\le C\left\{n^{1-C\kappa} + n\exp\left(-C\sqrt{n}\right)\right\},
\end{aligned}$$

by using just that $D_{m_l} = O(\sqrt{n})$ ($l=1,2$), and that the cardinal of $\mathcal{M}_n$ is smaller than n. The last bound is itself smaller than Cn^{-1}, if we choose κ large enough. We then notice that, for any $\alpha_n>0$

$$\begin{aligned}
2(1+2\kappa\ln(n))H_{a,m'}^2 &\le 6\kappa\left\|\varphi_2'\right\|_{\infty,[0;1]}^2\frac{D_{m_1'}^3 D_{m_2'}\ln(n)}{n}\left\|\hat{F}_n - F_X\right\|_{\infty,A_1}^2,\\
&\le 6\kappa\left\|\varphi_2'\right\|_{\infty,[0;1]}^2\frac{D_{m_1'}^3 D_{m_2'}\ln(n)}{n}\left(\alpha_n^2 + \mathbf{1}_{\|\hat{F}_n - F_X\|_{\infty,A_1}\ge\alpha_n}\right).
\end{aligned}$$

Choosing $\alpha_n = \sqrt{3\ln(n)/n}$, and using $D_{m'_1} = O(\sqrt{n}/\ln(n))$,

$$\begin{aligned} 2(1+2\kappa\ln(n))H^2_{a,m'} &\leq 12\kappa \left\|\varphi'_2\right\|^2_{\infty,[0;1]} \frac{D_{m'_1}D_{m'_2}}{n} + C\mathbf{1}_{\|\hat{F}_n - F_X\|_{\infty,A_1}\geq\alpha_n}, \\ &= V_1(m') + C\mathbf{1}_{\|\hat{F}_n - F_X\|^2_{\infty,A_1}\geq\alpha_n}, \end{aligned}$$

Besides,

$$\begin{aligned} \mathbb{E}\left[\left(T_1^{m'} - V_1(m')\right)_+\right] &\leq \mathbb{E}\left[\left(T_1^{m'} - 2(1+2\kappa\ln(n))H^2_{a,m'}\right)_+\right] + \mathbb{E}\left[C\mathbf{1}_{\|\hat{F}_n - F_X\|_{\infty,A_1}\geq\alpha_n}\right], \\ &\leq \mathbb{E}\left[\left(T_1^{m'} - 2(1+2\kappa\ln(n))H^2_{a,m'}\right)_+\right] + Cn^{-3}, \end{aligned}$$

with the inequality of Proposition 11. To conclude, $\sum_{m'\in\mathcal{M}_n}\mathbb{E}[(T_1^{m'} - V_1(m'))_+] \leq C/n$.

□

5.3.4. *Proof of Lemma 8.* For convenience, the constant κ_2 in the definition of V_2 is splitted into two parts, that is $\kappa_2 = \kappa\kappa'$. The first step is to write $\mathbb{E}[\max_{m'\in\mathcal{M}_n}(T_2^{m'} - V_2(m'))_+] \leq \sum_{m'\in\mathcal{M}_n}\mathbb{E}[(T_2^{m'} - V_2(m'))_+]$. Then it is enough to bound this quantity for each index m'. We write in a shortened form the sum "$\sum_{j=1}^{D_{m'_1}}$": "$\sum_j$" (and the analogous for $\sum_{k=1}^{D_{m'_2}}$). We compute $T_2^{m'}$

$$\begin{aligned} &= \int_{A_1\times A_2} \left(\hat{h}^{\hat{F}}_{m'}\left(F_X(x),y\right) - \hat{h}^{\hat{F}}_{m'}\left(\hat{F}_n(x),y\right)\right. \\ &\quad \left. -\mathbb{E}\left[\hat{h}^{\hat{F}}_{m'}\left(F_X(x),y\right) - \hat{h}^{\hat{F}}_{m'}\left(\hat{F}_n(x),y\right)|(X_{-l})_l\right]\right)^2 f_X(x)dxdy, \\ &= \int_{A_1}\sum_{j,j'}\sum_{k,k'}\left(\hat{a}^{\hat{F}}_{j,k} - \mathbb{E}\left[\hat{a}^{\hat{F}}_{j,k}|(X_{-l})_l\right]\right)\left(\hat{a}^{\hat{F}}_{j',k'} - \mathbb{E}\left[\hat{a}^{\hat{F}}_{j',k'}|(X_{-l})_l\right]\right) \\ &\quad \times\left(\varphi_j\circ F_X(x) - \varphi_j\circ\hat{F}_n(x)\right)\left(\varphi_{j'}\circ F_X(x) - \varphi_{j'}\circ\hat{F}_n(x)\right)\int_{A_2}\varphi_k(y)\varphi_{k'}(y)dy f_X(x)dx, \\ &= \int_{[0;1]}\sum_{k=1}^{D_{m'_2}}\left\{\sum_{j=1}^{D_{m'_1}}\left(\hat{a}^{\hat{F}}_{j,k} - \mathbb{E}\left[\hat{a}^{\hat{F}}_{j,k}|(X_{-l})_l\right]\right)\left(\varphi_j(u) - \varphi_j\circ\hat{U}_n(u)\right)\right\}^2 du, \end{aligned}$$

By Cauchy-Schwarz Inequality, and the mean value theorem,

$$T_2^{m'} \leq \left\|\varphi'_2\right\|^2_{\infty,[0;1]} D^3_{m'_1}\left\|\hat{U}_n - id\right\|^2_{\infty,[0;1]}\sum_{k=1}^{D_{m'_2}}\sum_{j=1}^{D_{m'_1}}\left(\hat{a}^{\hat{F}}_{j,k} - \mathbb{E}\left[\hat{a}^{\hat{F}}_{j,k}|(X_{-l})_l\right]\right)^2.$$

Thus, $\mathbb{E}[(T_2^{m'} - V_2(m'))_+] \leq T^{m'}_{2,a} + T^{m'}_{2,b}$, with

$$T^{m'}_{2,a} = D^3_{m'_1}\left\|\varphi'_2\right\|^2_{\infty,[0;1]}\mathbb{E}\left[\sum_{j,k}\left(\hat{a}^{\hat{F}}_{j,k} - \mathbb{E}\left[\hat{a}^{\hat{F}}_{j,k}|(X_{-l})_l\right]\right)^2\left(\left\|\hat{U}_n - id\right\|^2_{\infty,[0;1]} - \kappa'\frac{\ln(n)}{n}\right)_+\right],$$

$$T^{m'}_{2,b} = D^3_{m'_1}\left\|\varphi'_2\right\|^2_{\infty,[0;1]}\kappa'\frac{\ln(n)}{n}\mathbb{E}\left[\left(\sum_{j,k}\left(\hat{a}^{\hat{F}}_{j,k} - \mathbb{E}\left[\hat{a}^{\hat{F}}_{j,k}|(X_{-l})_l\right]\right)^2 - \frac{\kappa}{\|\varphi'_2\|^2_{\infty,[0;1]}}\frac{D_{m'_1}D_{m'_2}}{n}\ln(n)\right)_+\right].$$

Bounding roughly $\sum_{j,k}(\hat{a}^{\hat{F}}_{j,k} - \mathbb{E}[\hat{a}^{\hat{F}}_{j,k}(X_{-l})_l])^2$ leads to

$$T^{m'}_{2,a} \leq 2D^4_{m'_1} D_{m'_2} \left\|\varphi'_2\right\|^2_{\infty,[0;1]} \mathbb{E}\left[\left(\left\|\hat{U}_n - id\right\|^2_{\infty,[0;1]} - \kappa' \frac{\ln(n)}{n}\right)_+\right].$$

Inequality (30) and the assumptions $D_{m_l} \leq \sqrt{n}$ $(l = 1, 2)$ allow to conclude that $T^{m'}_{2,a} \leq Cn^{3/2-2\kappa'}$, and thus, choosing $\kappa' \geq 7/4$, $\sum_{m' \in \mathcal{M}_n} T^{m'}_{2,a} \leq C/n$. For the second term $T^{m'}_{2,b}$, we notice first that $\sum_{j,k}(\hat{a}^{\hat{F}}_{j,k} - \mathbb{E}[\hat{a}^{\hat{F}}_{j,k}|(X_{-l})_l])^2 = sup_{t \in \mathcal{S}(m')}(\nu^b_n)^2(t)$, with

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

We now bound the deviations of this empirical process, centered conditonally to (X_{-l}), exactly as we bound ν^a_n in the proof of Lemma 7: they are controled by Talagrand Inequality (20). We obtain finally $\sum_{m' \in \mathcal{M}n} T^{m'}_{2,b} \leq C \ln(n)/n$, which ends the proof, by gathering this bound with the one of $\sum_{m' \in \mathcal{M}n} T^{m'}_{2,a}$.

□

5.3.5. *Proof of Lemma 9.* To compute a bound for $\mathbb{E}[T^m_3]$, let us begin with the definition of the estimators and their coefficients, to get $T^m_3 = \sum_{j=1}^{D_{m_1}} \sum_{k=1}^{D_{m_2}} \{\langle \varphi_k, \Lambda_j(y)\rangle_{A_2}\}^2$ with $\Lambda_j(y) = \int_{A_1}(\varphi_j(\hat{F}_n(x)) - \varphi_j(F_X(x)))f_{(X,Y)}(x, y)dx$. Thus we can write $T^m_3 = \sum_{j=1}^{D_{m_1}} \|\Pi_{S_{m_2}}\Lambda_j\|^2_{A_2} \leq \sum_{j=1}^{D_{m_1}} \|\Lambda_j\|^2_{A_2}$, which can be developed as

$$T^m_3 = \sum_{j=1}^{D_{m_1}} \int_{A_2} \left(\int_{[0;1]} (\varphi_j(\hat{U}_n(u)) - \varphi_j(u))h(u, y)du\right)^2 dy := \int_{A_2} T'^m_3(y)dy.$$

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:

$$\mathbb{E}\left[T'^m_3(y)\right] \leq 3\mathbb{E}\left[T^m_{3,1}(y)\right] + 3\mathbb{E}\left[T^m_{3,2}(y)\right] + 3\mathbb{E}\left[T^m_{3,3}(y)\right],$$

with notations

$$T^m_{3,1}(y) = \sum_{j=1}^{D_{m_1}} \left\{\int_0^1 h(u, y)\left(\hat{U}_n(u) - u\right)\varphi'_j(u)du\right\}^2,$$

$$T^m_{3,2}(y) = (1/4)\sum_{j=1}^{D_{m_1}} \left\{\int_0^1 h(u, y)\left(\hat{U}_n(u) - u\right)^2\varphi''_j(u)du\right\}^2,$$

$$T^m_{3,3}(y) = (1/6)\sum_{j=1}^{D_{m_1}} \left\{\int_0^1 h(u, y)\left(\hat{U}_n(u) - u\right)^3\varphi^{(3)}_j(\hat{\alpha}_{j,n,u})du\right\}^2.$$

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

$$\mathbb{E}\left[T^m_{3,1}(y)\right] = \mathbb{E}\left[\sum_{j=1}^{D_{m_1}}\left(\frac{1}{n}\sum_{i=1}^{n} A_{i,j}(y) - \mathbb{E}[A_{i,j}(y)]\right)^2\right], \quad \text{with } A_{i,j}(y) = \int_{U_i}^1 h(u, y)\varphi'_j(u)du.$$

We integrate by parts in $A_{i,j}$ (h is assumed to be $\mathcal{C}^1$ with respect to its first variable). This leads to another splitting, for each $y \in A_2$:

$$\mathbb{E}\left[T^m_{3,1}(y)\right] \leq 2\mathbb{E}\left[T^m_{3,1,1}(y)\right] + 2\mathbb{E}\left[T^m_{3,1,2}(y)\right],$$

where

$$\begin{aligned} T^m_{3,1,1}(y) &= \sum_{j=1}^{D_{m_1}} \left\{ \frac{1}{n}\sum_{i=1}^{n} h(U_i,y)\varphi_j(U_i) - \mathbb{E}\left[h(U_i,y)\varphi_j(U_i)\right] \right\}^2, \\ T^m_{3,1,2}(y) &= \sum_{j=1}^{D_{m_1}} \left\{ \int_0^1 \partial_1 h(u,y)\left(\hat{U}_n(u) - u\right)\varphi_j(u)du \right\}^2. \end{aligned} \tag{32}$$

In the spirit of the bound given for T^m_1, the first term is controlled as follows:

$$\mathbb{E}\left[T^m_{3,1,1}(y)\right] \leq \frac{1}{n}\sum_{j=1}^{D_{m_1}} \mathbb{E}\left[\left(h(U_1,y)\varphi_j(U_1)\right)^2\right] \leq \frac{D_{m_1}}{n}\int_0^1 h(u,y)^2 du.$$

Thus, $\int_{A_2}\mathbb{E}[T^m_{3,1,1}(y)]dy \leq \|h\|^2 D_{m_1}/n$. Then, by definition and properties of the orthogonal projection on $\mathbb{S}_m$,

$$\mathbb{E}\left[T^m_{3,1,2}(y)\right] = \mathbb{E}\left[\sum_{j=1}^{D_{m_1}}\left(\langle \partial_1 h(.,y)(\hat{U}_n - id), \varphi_j\rangle_{[0;1]}\right)^2\right] \leq \mathbb{E}\left[\left\|\partial_1 h(.,y)(\hat{U}_n - id)\right\|^2_{[0;1]}\right].$$

Finally, $T^m_{3,1,2}(y) \leq C\|\partial_1 h(.,y)\|^2_{[0;1]}/n$ by Inequality (28), and thus, by gathering the bounds for $T^m_{3,1,1}(y)$ and $T^m_{3,1,2}(y)$,

$$\int_{A_2}\mathbb{E}\left[T^m_{3,1}(y)\right]dy \leq C\left(\frac{1}{n} + \frac{D_{m_1}}{n}\right).$$

As regards $T^m_{3,2}(y)$, we remark first that for $j \geq 2$, $\varphi_j'' = -(\pi\mu_j)^2\varphi_j$, with $\mu_j = j$ for even j, and $\mu_j = j-1$ otherwise, so that μ_j is bounded by D_{m_1}. Hence,

$$\begin{aligned} \mathbb{E}\left[T^m_{3,2}(y)\right] &\leq (\pi^4/4)D^4_{m_1}\mathbb{E}\left[\sum_{j=2}^{D_{m_1}}\left\{\int_0^1 h(u,y)\left(\hat{U}_n(u)-u\right)^2\varphi_j(u)du\right\}^2\right], \\ &\leq (\pi^4/4)D^4_{m_1}\mathbb{E}\left[\left\|h(.,y)\left(\hat{U}_n - id\right)^2\right\|^2_{[0;1]}\right] \leq C\int_{[0;1]} h^2(u,y)du\frac{D^4_{m_1}}{n^2}, \end{aligned}$$

by proceeding with the previous arguments (properties of orthogonal projection and Inequality (28)). So we prove $\int_{A_2}\mathbb{E}[T^m_{3,2}(y)]dy \leq CD^4_{m_1}/n^2$. The computations for the last term are less technical:

$$\mathbb{E}\left[T^m_{3,3}(y)\right] \leq (1/6)\sum_{j=1}^{D_{m_1}}\left\|\varphi_j^{(3)}\right\|^2_{\infty,[0;1]}\|h(.,y)\|^2_{[0;1]}\mathbb{E}\left[\left\|\hat{U}_n - id\right\|^6_{\infty,[0;1]}\right],,$$

thus $\int_{A_2}\mathbb{E}[T^m_{3,3}(y)]dy \leq CD^7_{m_1}/n^3$. This completes the proof of the first inequality of Lemma 9. With regard to the second inequality, it is enough to bound $\mathbb{E}[\max_{m'\in\mathcal{M}_n}(T^p_1 - V_1(p))_+]$, like for the second part of Lemma 7 ($p = m'$ or $p = m \wedge m'$). As previously, we get the splitting

$$T^p_3 \leq 6\int_{A_2} T^p_{3,1,1}(y)dy + 6\int_{A_2} T^p_{3,1,2}(y)dy + 3\int_{A_2} T^p_{3,2}(y)dy + 3\int_{A_2} T^p_{3,3}(y)dy, \tag{33}$$

and

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_3^{p,b}-V_3(p)\right)_+\right]\leq \mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(6\int_{A_2}T^p_{3,1,1}(y)dy-V_3(p)/3\right)_+\right]$$
$$+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}6\int_{A_2}T^p_{3,1,2}(y)dy\right]$$
$$+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(3\int_{A_2}T^p_{3,2}(y)dy-V_3(p)/3\right)_+\right]$$
$$+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(3\int_{A_2}T^p_{3,3}(y)dy-V_3(p)/3\right)_+\right].$$

The term which is not centered is direcly negligible : denoting by $m_{\max}$ the largest couple of index (maximum is taken term by term) in the collection $\mathcal{M}_n$, we remark that $T^p_{3,1,2}\leq T^{m_{\max}}_{3,1,2}$ (by (32)). Hence, $\mathbb{E}[\max_{m'\in\mathcal{M}_n}6\int_{A_2}T^p_{3,1,2}(y)dy]\leq C/n$. Let us briefly study each of the other terms: first $T^p_{3,1,1}(y)=\sup_{s\in S_{p_1},\|s\|_{[0;1]}=1}\nu^2_{n,y}(s)$, with

$$\nu_{n,y}(s)=\frac{1}{n}\sum_{i=1}^n\pi(X_i,y)s\circ F_X(X_i)-\mathbb{E}\left[\pi(X_i,y)s\circ F_X(X_i)\right].$$

Using once more time Talagrand Inequality (20) leads to

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(6\int_{A_2}T^p_{3,1,1}(y)dy-V_{3,1,1}(p)\right)_+\right]\leq\frac{C}{n}, \tag{34}$$

with $V_{3,1,1}(p)=6\times 2(1+2\delta)\|h\|^2_\infty D_{p_1}/n$, $(\delta>0)$. Besides, for $n\geq n_0=\exp(\|h\|^2_\infty)$,

$$V_{3,1,1}(p)\leq 12(1+2\delta)\ln(n)\frac{D_{p_1}}{n}\leq C\frac{D_{p_1}D_{p_2}}{n}:=V^b_{3,1,1}(p),$$

since $D_{p_2}\geq c\ln(n)$ $(c>0)$. Inequality (34) holds with $V^b_{3,1,1}$. The two last terms, involving $T^m_{3,2}(y)$ and $T^m_{3,3}(y)$ can be compute with the same strategy: use the proof of the first inequality of Lemma 9 to bound $\int_{A_2}T^m_{3,l}(y)dy$ $(l=2,3)$ by quantity of the form $C\|\hat{U}_n-id\|^k_\infty$, and then apply Inequality (29). The conclusion is that

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(3\int_{A_2}T^p_{3,l}(y)dy-V_{3,l}(p)\right)_+\right]\leq C\frac{\ln(n)}{n}, \tag{35}$$

for $l=2,3$, with $V_{3,2}(p)=CD^4_{p_1}\ln^2(n)/n^2$, and $V_{3,3}(p)=CD^7_{p_1}\ln^3(n)/n^3$. Assuming both $n\geq n_1=\exp(\|h\|^2)$, and $D_{p_1}=O(n^{1/3})$, $D_{p_2}\geq c\ln^3(n)$, we have

$$V_{3,2}(p)\leq C\frac{D_{p_1}D_{p_2}}{n}:=V^b_{3,2}(p).$$

With the more restrictive low bound $D_{p_2}\geq c\ln^4(n)$, we get also $V_{3,3}(p)\leq CD_{p_1}D_{p_2}/n:=V^b_{3,3}(p)$. As usual, Inequalities (35) still hold with $V^b_{3,l}$ instead of $V_{3,l}$. The proof is complete if we gather all these bounds and if we choose the constant κ_3 such that $V_3\geq 3V^b_{3,1,1}$, $V_3\geq 3V^b_{3,2}$, et $V_3\geq 3V^b_{3,3}$.

□

5.3.6. *Proof of Lemma 10.* Let us first split the term $T_4^{m'}$ in several parts. Similarly to the bound obtained for T_3^m, we use the definitions of the estimators and their coefficients, and the fact that the basis $(\varphi_k)_k$ is orthonormal: hence,

$$T_4^{m'} \leq \int_{A_1} \mathbb{E}\left[\sum_{k=1}^{D_{m_2'}} \left(\sum_{j=1}^{D_{m_1'}} \hat{a}_{j,k}^{\hat{F}} \left(\varphi_j \circ F_X(x) - \varphi_j \circ \hat{F}_n(x)\right)\right)^2 |(X_{-l})_l\right] f_X(x)dx.$$

We write it $T_4^{m'} \leq 2T_{4,1}^{m'} + 2T_{4,2}^{m'}$, with

$$T_{4,1}^{m'} = \int_{A_1} \mathbb{E}\left[\sum_{k=1}^{D_{m_2'}} \left(\sum_{j=1}^{D_{m_1'}} \left(\hat{a}_{j,k}^{\hat{F}} - a_{j,k}\right) \left(\varphi_j \circ F_X(x) - \varphi_j \circ \hat{F}_n(x)\right)\right)^2 |(X_{-l})_l\right] f_X(x)dx,$$
$$T_{4,2}^{m'} = \int_{A_1} \mathbb{E}\left[\sum_{k=1}^{D_{m_2'}} \left(\sum_{j=1}^{D_{m_1'}} a_{j,k} \left(\varphi_j \circ F_X(x) - \varphi_j \circ \hat{F}_n(x)\right)\right)^2 |(X_{-l})_l\right] f_X(x)dx,$$

where we denote by $a_{j,k} = \langle h, \varphi_j \otimes \varphi_k \rangle$, the Fourier's coefficients of the function h. Then we have also $T_{4,1}^{m'} \leq 2T_{4,1,1}^{m'} + 2T_{4,1,2}^{m'}$ with the notations

$$T_{4,1,1}^{m'} = \int_{[0;1]} \mathbb{E}\left[\sum_{k=1}^{D_{m_2'}} \left\{\sum_{j=1}^{D_{m_1'}} \left(\hat{a}_{j,k}^{\hat{F}} - \mathbb{E}\left[\hat{a}_{j,k}^{\hat{F}} |(X_{-l})_l\right]\right)^2\right\} \left\{\sum_{j=1}^{D_{m_1'}} \left(\varphi_j(u) - \varphi_j \circ \hat{U}_n(u)\right)^2\right\} |(X_{-l})_l\right] du,$$
$$T_{4,1,2}^{m'} = \int_{[0;1]} \mathbb{E}\left[\sum_{k=1}^{D_{m_2'}} \left\{\sum_{j=1}^{D_{m_1'}} \left(\mathbb{E}\left[\hat{a}_{j,k}^{\hat{F}} |(X_{-l})_l\right] - a_{j,k}\right)^2\right\} \left\{\sum_{j=1}^{D_{m_1'}} \left(\varphi_j(u) - \varphi_j \circ \hat{U}_n(u)\right)^2\right\} |(X_{-l})_l\right] du.$$

As

$$\mathbb{E}\left[T_{4,2}^{m'}\right] = \mathbb{E}\left[\sum_{k=1}^{D_{m_2'}} \sum_{j,j'=1}^{D_{m_1'}} a_{j,k}a_{j',k} \int_0^1 \left(\varphi_j(u) - \varphi_j \circ \hat{U}_n(u)\right) \left(\varphi_{j'}(u) - \varphi_{j'} \circ \hat{U}_n(u)\right) du\right],$$

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

$$T_{4,2,1}^{m'} = \sum_{k=1}^{D_{m_2'}} \sum_{j,j'=1}^{D_{m_1'}} a_{j,k}a_{j',k} \int_0^1 (u - \hat{U}_n(u))^2 \varphi_j'(u)\varphi_{j'}'(u)du,$$
$$T_{4,2,2}^{m'} = (1/4) \sum_{k=1}^{D_{m_2'}} \sum_{j,j'=1}^{D_{m_1'}} a_{j,k}a_{j',k} \int_0^1 (u - \hat{U}_n(u))^4 \varphi_j''(\hat{\alpha}_{j,n,u})\varphi_{j'}''(\hat{\alpha}_{j',n,u})du,$$
$$T_{4,2,3}^{m'} = \sum_{k=1}^{D_{m_2'}} \sum_{j,j'=1}^{D_{m_1'}} a_{j,k}a_{j',k} \int_0^1 (u - \hat{U}_n(u))^3 \varphi_j''(\hat{\alpha}_{j,n,u})\varphi_{j'}'(u)du.$$

Hence, the decomposition of the studied term is $T_4^{m'} \leq 4T_{4,1,1}^{m'}+4T_{4,1,2}^{m'}+2T_{4,2,1}^{m'}+2T_{4,2,2}^{m'}+2T_{4,2,3}^{m'}$, and consequently

$$\begin{aligned}\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(T_4^{m'}-V_4(m')\right)_+\right] &\leq \mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(4T_{4,1,1}^{m'}-V_4(m')/3\right)_+\right]\\ &+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(4T_{4,1,2}^{m'}-V_4(m')/3\right)_+\right]\\ &+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}\left(2T_{4,2,3}^{m'}-V_4(m')/3\right)_+\right]\\ &+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}2T_{4,2,1}^{m'}\right]+\mathbb{E}\left[\max_{m'\in\mathcal{M}_n}2T_{4,2,2}^{m'}\right].\end{aligned}$$

The methods use to bound each of these terms have already been detailed for other terms: with regard to the two qunatities which are not centered, we bound it to show that they are negligible (that is of order at most C/n). For the others, we first bound each $T_{4,l}^{m'}$ by a quantity of the form $C\|\hat{U}_n - id\|_{\infty,[0;1]}$, and we apply finally Inequality (29), as we have already done for $T_{2,a}^m$ for example. That is why we only give the bounds for each $T_{4,l}^{m'}$. To begin, the term $T_{4,1,1}^{m'}$ can be written

$$T_{4,1,1}^{m'} = \sum_{k=1}^{D_{m_2'}}\sum_{j=1}^{D_{m_1'}}\mathrm{Var}\left(\hat{a}_{j,k}^{\hat{F}}|(X_{-l})_l\right)\int_{[0;1]}\sum_{j=1}^{D_{m_1'}}\left(\varphi_j(u)-\varphi_j\circ\hat{U}_n(u)\right)^2du. \tag{36}$$

The conditional variance is

$$\begin{aligned}\mathrm{Var}\left(\hat{a}_{j,k}^{\hat{F}}|(X_{-l})_l\right) &= \mathrm{Var}\left\{\frac{1}{n}\sum_{i=1}^n\varphi_k(Y_i)\varphi_j\circ\hat{F}_n(X_i)|(X_{-l})_l\right\},\\ &\leq \frac{1}{n}\mathbb{E}\left[\varphi_k(Y_1)^2\left(\varphi_j\circ\hat{F}_n(X_1)\right)^2|(X_{-l})_l\right].\end{aligned}$$

By Property (3) applied to the sum over j,k of the last quantity, $\sum_{j,k}\mathrm{Var}(\hat{a}_{j,k}^{\hat{F}}|(X_{-l})_l) \leq D_{m_1'}D_{m_2'}/n$. Besides, we use the mean value theorem to bound the integral of (36) so that

$$T_{4,1,1}^{m'} \leq \frac{D_{m_1'}D_{m_2'}}{n}\times D_{m_1'}^3\|\varphi_2'\|_{\infty,[0;1]}^2\left\|\hat{U}_n-id\right\|_{\infty,[0;1]}^2, \tag{37}$$

which allows us to control $\mathbb{E}[\max_{m'\in\mathcal{M}_n}(4T_{4,1,1}^{m'}-V_4(m')/3)_+]$ as explained previously. Furthermore,

$$T_{4,1,2}^{m'} = T_3^{m'}\int_{[0;1]}\sum_{j=1}^{D_{m_1'}}\left(\varphi_j(u)-\varphi_j\circ\hat{U}_n(u)\right)^2du,$$

which leads to $T_{4,1,2}^{m'} \leq T_3^{m'}D_{m_1'}^3\|\varphi_2'\|_{\infty,[0;1]}^2\|\hat{U}_n-id\|_{\infty,[0;1]}^2$. The term $T_3^{m'}$ is replaced by its detailed upper-bound (33), and as a result, $T_{4,1,2}^{m'} \leq \sum_{l=1}^4 T_{4,1,2,l}^{m'}$. Roughly speaking, we get $T_{4,1,2,l}^{m'} \leq C\|\hat{U}_n-id\|_{\infty,[0;1]}$ and apply the previous strategy for each $l=1,\ldots,4$. Let us consider now the terms $T_{4,2,1}^{m'}$ and $T_{4,2,2}^{m'}$ which do not require to be centered. It is usefull to remark that the Fourier's coefficients of h can be written

$$a_{j,k} = \langle\xi_k,\varphi_j\rangle_{[0;1]} = \int_{[0;1]}\xi_k(u)\varphi_j(u)du, \quad \text{with } \xi_k(u)=\int_{A_2}h(u,y)\varphi_k(y)dy.$$

Since the term $T^{m'}_{4,2,1}$ involves the derivative of the projection of ξ_k onto $S_{m'_1}$, we use a specific property of the trigonometric basis: $\sum_{j=1}^{D_{m'_1}} a_{j,k}\varphi'_j = \left(\Pi_{S_{m'_1}}(\xi_k)\right)' = \Pi_{S_{m'_1}}(\xi'_k)$, so

$$T^{m'}_{4,2,1} \leq \left\|\hat{U}_n - id\right\|^2_{\infty,[0;1]} \sum_{k=1}^{D_{m'_2}} \left\|\xi'_k\right\|^2_{[0;1]}.$$

Let us compute then the derivative of ξ_k to bound roughly

$$\sum_{k=1}^{D_{m'_2}} \left\|\xi'_k\right\|^2_{[0;1]} = \sum_{k=1}^{D_{m'_2}} \int_{[0;1]} \left(\int_{A_2} \partial_1 h(u,y)\varphi_k(y)dy\right)^2 du \leq \int_{[0;1]} \|\partial_1 h(u,.)\|^2_{A_2}\, du = \|\partial_1 h\|^2.$$

We have thus $\mathbb{E}[\max_{m'\in\mathcal{M}_n} T^{m'}_{4,2,1}] \leq \|\partial_1 h\|^2 \mathbb{E}[\|\hat{U}_n - id\|^2_{\infty,[0;1]}] \leq C/n$ with Inequality (28). Recall now that

$$T^{m'}_{4,2,2} = (1/4)\sum_{k=1}^{D_{m'_2}} \int_0^1 (u-\hat{U}_n(u))^4 \left(\sum_{j=1}^{D_{m'_1}} a_{j,k}\varphi''_j(\hat{\alpha}_{j,n,u})\right)^2 du.$$

We introduce $\mu_j = j$ for even j and $\mu_j = j-1$ for odd j. Since h belongs to $W^2_{per}([0;1]^2, L, (1,0))$ and according to (15),

$$\begin{aligned}
\sum_{k=1}^{D_{m'_2}} \left(\sum_{j=1}^{D_{m'_1}} a_{j,k}\varphi''_j(\hat{\alpha}_{j,n,u})\right)^2 &\leq \left\|\varphi''_2\right\|^2_{\infty,[0;1]} \sum_{k=1}^{D_{m'_2}} \left(\sum_{j=1}^{D_{m'_1}} a_{j,k}\mu_j^2\right)^2, \\
&\leq \left\|\varphi''_2\right\|^2_{\infty,[0;1]} \sum_{k=1}^{D_{m'_2}} \sum_{j=1}^{D_{m'_1}} a^2_{j,k}\mu_j^2 \sum_{j=1}^{D_{m'_1}} \mu_j^2, \\
&\leq \left\|\varphi''_2\right\|^2_{\infty,[0;1]} \frac{L^2}{\pi^2} D^3_{m'_1} \leq C D^3_{m_{1,\max}}.
\end{aligned}$$

Hence, $\mathbb{E}[\max_{m'\in\mathcal{M}_n} T^{m'}_{4,2,2}] \leq \mathbb{E}[\|\hat{U}_n - id\|^4_{\infty,[0;1]}] C D^3_{m_{1,\max}} \leq C D^3_{m_{1,\max}}/n^2 \leq C/n$ as soon as $D_{m_{1,\max}} \leq n^{1/3}$ (we denote by $D_{m_{1,\max}}$ the largest index on the collection (D_{m_1})). Following the same sketch for the last term, we write

$$T^{m'}_{4,2,3} = \int_{[0;1]} (u-\hat{U}_n)^3 \sum_{k=1}^{D_{m'_2}} \left(\sum_{j=1}^{D_{m'_1}} a_{j,k}\varphi''_j(\hat{\alpha}_{j,n,u})\right)\left(\sum_{j=1}^{D_{m'_1}} a_{j,k}\varphi'_j(u)\right).$$

and compute as in the term $T^{m'}_{4,2,2}$:

$$\sum_{k=1}^{D_{m'_2}} \left(\sum_{j=1}^{D_{m'_1}} a_{j,k}\varphi''_j(\hat{\alpha}_{j,n,u})\right)^2 \leq \|\varphi''_2\|^2_{\infty,[0;1]} \frac{L^2}{\pi^2} D^3_{m'_1}, \quad \sum_{k=1}^{D_{m'_2}} \left(\sum_{j=1}^{D_{m'_1}} a_{j,k}\varphi'_j(u)\right)^2 \leq \|\varphi'_2\|^2_{\infty,[0;1]} \frac{L^2}{\pi^2} D_{m'_1}.$$

This leads to

$$T^{m'}_{4,2,3} \leq \|\varphi'_2\|_{\infty,[0;1]} \|\varphi''_2\|_{\infty,[0;1]} \frac{L^2}{\pi^2} D_{m'_1} \left\|\hat{U}_n - id\right\|^3_{\infty,[0;1]},$$

and we apply tools already used to complete the proof.

□

Acknowledgement

I would like to thank Fabienne Comte for her helpful suggestions throughout this work.

References

[Ada75] Robert A. Adams. *Sobolev spaces*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Pure and Applied Mathematics, Vol. 65.

[AL11] Nathalie Akakpo and Claire Lacour. Inhomogeneous and anisotropic conditional density estimation from dependent data. *Submitted hal-00557307*, 2011.

[Bar02] Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002.

[BBM99] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[BC05] Elodie Brunel and Fabienne Comte. Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā*, 67(3):441–475, 2005.

[BCL07] Elodie Brunel, Fabienne Comte, and Claire Lacour. Adaptive estimation of the conditional density in the presence of censoring. *Sankhyā*, 69(4):734–763, 2007.

[BH01] David M. Bashtannyk and Rob J. Hyndman. Bandwidth selection for kernel conditional density estimation. *Comput. Statist. Data Anal.*, 36(3):279–298, 2001.

[BM98] Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

[CLP11] Serge Cohen and Erwan Le Pennec. Conditional density estimation by penalized likelihood model selection and applications. *Submitted arXiv:1103.2021*, 2011.

[DGZ03] Jan G. De Gooijer and Dawit Zerom. On conditional density estimation. *Statist. Neerlandica*, 57(2):159–176, 2003.

[DKW56] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27:642–669, 1956.

[DL93] Ronald A. DeVore and George G. Lorentz. *Constructive approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993.

[Efr99] Sam Efromovich. *Nonparametric curve estimation*. Springer Series in Statistics. Springer-Verlag, New York, 1999. Methods, theory, and applications.

[Efr07] Sam Efromovich. Conditional density estimation in a regression setting. *Ann. Statist.*, 35(6):2504–2535, 2007.

[Efr10a] Sam Efromovich. Dimension reduction and adaptation in conditional density estimation. *J. Amer. Statist. Assoc.*, 105(490):761–774, 2010.

[Efr10b] Sam Efromovich. Oracle inequality for conditional density estimation and an actuarial example. *Ann. Inst. Statist. Math.*, 62(2):249–275, 2010.

[Fau09] Olivier P. Faugeras. A quantile-copula approach to conditional density estimation. *J. Multivariate Anal.*, 100(9):2083–2099, 2009.

[FYT96] Jianqing Fan, Qiwei Yao, and Howell Tong. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206, 1996.

[GK07] László Györfi and Michael Kohler. Nonparametric estimation of conditional distributions. *IEEE Trans. Inform. Theory*, 53(5):1872–1879, 2007.

[GL11] Alexander Goldenshluger and Oleg Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632, 2011.

[HBG96] Rob J. Hyndman, David M. Bashtannyk, and Gary K. Grunwald. Estimating and visualizing conditional densities. *J. Comput. Graph. Statist.*, 5(4):315–336, 1996.

[HKPT98] Wolfgang Härdle, Gérard Kerkyacharian, Dominique Picard, and Alexander Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1998.

[Hoc02] Reinhard Hochmuth. Wavelet characterizations for anisotropic Besov spaces. *Appl. Comput. Harmon. Anal.*, 12(2):179–208, 2002.

[HWY99] Peter Hall, Rodney C.L. Wolff, and Qiwei Yao. Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.*, 94(445):154–163, 1999.

[HY02] Rob J. Hyndman and Qiwei Yao. Nonparametric estimation and symmetry tests for conditional density functions. *J. Nonparametr. Stat.*, 14(3):259–278, 2002.

[KP04] Gérard Kerkyacharian and Dominique Picard. Regression in random design and warped wavelets. *Bernoulli*, 10(6):1053–1105, 2004.

[KR05] Thierry Klein and Emmanuel Rio. Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077, 2005.

[Lac07] Claire Lacour. Adaptive estimation of the transition density of a Markov chain. *Ann. Inst. H. Poincaré Probab. Statist.*, 43(5):571–597, 2007.

[Lac08] Claire Lacour. Adaptive estimation of the transition density of a particular hidden Markov chain. *J. Multivariate Anal.*, 99(5):787–814, 2008.

[Nik75] S. M. Nikol′skiĭ. *Approximation of functions of several variables and imbedding theorems*. Springer-Verlag, New York, 1975. Translated from the Russian by John M. Danskin, Jr., Die Grundlehren der Mathematischen Wissenschaften, Band 205.

[Pen95] Marianna Penskaya. Mean square consistent estimation of a ratio. *Scand. J. Statist.*, 22(1):129–137, 1995.

[TNK09] Ichiro Takeuchi, Kaname Nomura, and Takafumi Kanamori. Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Comput.*, 21(2):533–559, 2009.

[Tsy04] Alexandre B. Tsybakov. *Introduction à l'estimation non-paramétrique*, volume 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2004.