



Warped bases for conditional density estimation

Gaëlle Chagny

► To cite this version:

Gaëlle Chagny. Warped bases for conditional density estimation. *Mathematical Methods of Statistics*, 2013, 22 (4), pp.253-282. 10.3103/S1066530713040017 . hal-00641560v2

HAL Id: hal-00641560

<https://hal.science/hal-00641560v2>

Submitted on 26 Jun 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

WARPED BASES FOR CONDITIONAL DENSITY ESTIMATION

GAËLLE CHAGNY^A *

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 prefer to 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 from 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 $\hat{\pi}(x, y) = \hat{h}(\hat{F}(x), y)$ otherwise, where \hat{F} is the empirical distribution function. We prove that it realises a global squared-bias/variance compromise, in a context of anisotropic function classes: we establish non-asymptotic mean-squared integrated risk bounds and also provide risk convergence rates. Simulation experiments illustrate the method.

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

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

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 variable 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, 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 [2011].

1.2. State of the art. Although nonparametric conditional density estimation has increasingly become a subject of interest since the early 1970s', adaptive estimators, which match the performances of an oracle that knows the regularity of the true function, 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 weighted regression. This leads mainly to two directions: kernel methods with

* Corresponding author. Email: gaelle.chagny@parisdescartes.fr

^aLaboratoire MAP5 (UMR CNRS 8145), Université Paris Descartes, Sorbonne Paris Cité, France.

well-chosen bandwidth(s), which have essentially been studied from an asymptotic point of view, and projection estimators built on regression-type criteria minimised on a well-chosen model.

The literature about the asymptotic properties of kernel estimators is vast. Several adjusted forms of the Nadaraya-Watson estimate have been proposed, for conditional distribution function [Stute, 1986b, Hall et al., 1999] and for the conditional density: "double kernel" estimator [Hyndman et al., 1996, Hyndman and Yao, 2002], generalization using local polynomials [Fan et al., 1996], and reweighted kernel estimate [De Gooijer and Zerom, 2003]. Accordingly, data-driven selection rules for the bandwidth are proposed, using methods inspired by Fan and Gijbels [1995], the bootstrap approach [Hall et al., 1999, Bashtannyk and Hyndman, 2001], or cross-validation [Sarda et al., 1994, Fan and Yim, 2004]. All these methods lead to asymptotic results: consistency rates, asymptotic normality, study of the asymptotic variance for example, under various assumptions (such as α -mixing design). A common feature of these estimators is their ratio form. This can be seen as a theoretical difficulty (see Penskaya 1995 for a specific study), which can be bypassed by using quantile regression or the copula function [Carroll et al., 1994, Takeuchi et al., 2009, Faugeras, 2009]: the corresponding estimates still satisfy classical asymptotic properties. Another way of avoiding ratio is to consider a transformation of the input data. This strategy, early studied in Stute [1986a] will be detailed in the next section.

Moreover, projection estimators have been developed: the quality criterion which has thus become classical is the mean integrated squared error, or empirical versions of it. Nonasymptotic results, such as oracle inequalities or lower-bounds for the risk are set for estimators based on orthogonal series. For example, a Fourier basis can be used to build a blockwise-shrinkage Efromovich-Pinsker estimator, using characteristic functions to rewrite π : the regression setting is studied in Efromovich [2007], the general case is the subject of Efromovich [2010a], and multidimensionality is taken into account in Efromovich [2010b]. His estimators match the performances of the oracle under the quadratic risk. The oracle-type inequalities stated permit to establish sharp minimaxity over the bivariate anisotropic Sobolev classes. The problem of dimension reduction is also studied in Fan et al. [2009], in the spirit of single index results. Model selection theory leads also to adaptation results: by minimizing a least-squares penalised contrast introduced by Lacour [2007], Brunel et al. [2007] build an estimator which adapts to an unknown underlying design and is minimax over anisotropic bivariate function classes. But the contrast, considered also by Akakpo and Lacour [2011] to deal with dependent data and inhomogeneous functional classes, does not provide explicit estimator without matrix invertibility requirements (except when using a histogram basis). Moreover the penalty given in Brunel et al. [2007] depends on the unknown infinite norm of π . It can be estimated but it then requires strong regularity assumptions. Notice also that recent works by Cohen and Le Pennec [2011] focus on a penalised maximum likelihood estimator leading to risk bounds for a Jensen-Kullback-Leibler loss function. The maximisation of the likelihood seems to be difficult without additional assumptions on the shape of the conditional density.

The present work is in the spirit of projection methods. We aim at providing an adaptive estimator, which satisfies nonasymptotic risk bounds, but with a simpler expression, thus avoiding matrix inversion and purifying the penalty function. This goal is achieved by using both a "warping" of the data, like in the works of Stute [1986a] and Mehra et al. [2000] (no ratio, no matrix inversion), and by applying in a new and original way the Goldenshluger and Lepski method (the key to avoiding nuisance terms in the penalty).

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.

The fundamental idea to provide a simple explicit estimator is that

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

is the joint density of the random pair $(F_X(X_1), Y_1)$. We provide first an estimator of this function h , and then an estimator for the target function π , by using the reverse formula: $\pi(x, y) = h(F_X(x), y)$. This strategy has also been used by Stute [1986a] and Mehra et al. [2000] to build kernel estimates of cumulative conditional distribution function and conditional density respectively, which are shown to be asymptotically normal. More recently, Kerkycharian and Picard [2004] employed similar ideas to provide wavelet thresholding estimators of a regression function.

We adopt here a nonasymptotic point of view, by using model selection, and we aim at adaptive results. Precisely, the assumption that h is squared integrable leads first to projection estimators of the form

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

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, we propose an estimator of π given by:

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

with \hat{F}_n an empirical counterpart for F_X . To avoid dependency in the proofs, we assume that there exists $(X_{-i})_{i \in \{1, \dots, n\}}$ a sample of variables with the same distribution as the (X_i) and independent of them. Thus, we set

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

However, it is an artefact of the theory: for the practical study, we assume to observe only the pairs $(X_i, Y_i)_i$ and compute successfully the estimator of F_X with the X_i .

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

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

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 advances of Goldenshluger and Lepski [2011] and is particularly well suited to the multidimensional framework.

Our main theorem is an adaptivity nonasymptotic result, an oracle-inequality and permits to deduce 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 compromise, with assumptions stated on the function h instead of π . Moreover, on the practical examples, the strategy we propose outperforms kernel methods summed up in Fan and Yim [2004] and the penalization device of Brunel et al. [2007].

1.4. Organisation 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 risk 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 then proceed 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 definition (1)), by minimizing a contrast on the models (Section 2.2). The second step consists in ensuring 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 equipped with its usual Hilbert structure: we denote by $\langle \cdot, \cdot \rangle$ the scalar-product and by $\|\cdot\|$ 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 theoretical part. The case of any segment A_2 can be easily obtained by making a scaling change, see Section 4. Following the example of Efromovich [1999], we choose the Fourier basis $(\varphi_{j_1} \otimes \varphi_{j_2})_{j_1, j_2 \in \mathbb{N} \setminus \{0\}}$ of $L^2([0; 1] \times A_2)$, defined for $u, y \in [0; 1]$ by

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

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 $\mathcal{I}_n^{(l)} = \{1, \dots, [\sqrt{n}/2] - 1\}$ ($[\cdot]$ 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

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

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 $\| \sum_{j_1=1}^{D_{m_1}} \sum_{j_2=1}^{D_{m_2}} (\varphi_{j_1} \otimes \varphi_{j_2})^2 \|_\infty \leq \mathbb{D}_m$, where $\|\cdot\|_\infty$ 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 1998 for the proof):

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

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

$$(4) \quad D_{m_l} \leq D_{m'_l} \implies S_{m_l} \subset S_{m'_l}.$$

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. [1999], and Brunel and Comte [2005] 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

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

This contrast is new and quite far from the regression and conditional 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. 2007), and from the presence of the empirical c.d.f \hat{F}_n . To justify this choice, plug for a moment the true c.d.f F_X instead of its empirical counterpart, and compute, for $t \in L^2([0; 1] \times A_2)$,

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

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

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

or equivalently,

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

and a similar expression for estimator $\hat{h}_m^{F_X}$ with coefficients $\hat{a}_{j_1, j_2}^{F_X}$. Finally, we set

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

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 easily computable empirical means. They do not involve a matricial inversion compared to the estimator obtained via the least-squares criterion (see for example Brunel et al. 2007). Moreover, in the toy 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 $(\pi_m^{\hat{F}, \hat{F}})_{m \in \mathcal{M}_n}$. The loss function naturally associated to our context is the following L^2 -norm,

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

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 $\langle \cdot, \cdot \rangle_{f_X}$ 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 $F'_X = f_X$,

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

If we want to bound the classical quadratic risk of the estimator, we can assume that f_X is bounded from below by a strictly positive constant f_0 : this assumption, which is standard (see for example Assumption \mathcal{A}_2 in Brunel et al. 2007, or Assumption (H_{Bas}) in Baraud 2002) leads to the inequality $\|v\|_{f_X} \geq f_0 \|v\|$, for all $v \in L^2(A_1 \times A_2, f_X)$. Thus, it is sufficient to bound the weighted risk.

We can give a simple explanation for the choice of an estimator among the collection, considering first the collection of theoretical estimators $(\hat{\pi}_m^{F_X, F_X})_m$. For the weighted L^2 -risk which is used in the rest of the paper, and for each $m \in \mathcal{M}_n$, we get

$$\begin{aligned} \mathbb{E} \left[\|\hat{\pi}_m^{F_X, F_X} - \pi\|_{f_X}^2 \right] &= \|\pi - \pi_m^{F_X}\|_{f_X}^2 + \mathbb{E} \left[\|\pi_m^{F_X} - \hat{\pi}_m^{F_X, F_X}\|_{f_X}^2 \right], \\ (5) \qquad \qquad \qquad &= \|h - h_m\|^2 + \mathbb{E} \left[\|h_m - \hat{h}_m^{F_X}\|^2 \right], \end{aligned}$$

where

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

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 (5) 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 :

$$\begin{aligned} \mathbb{E} \left[\|\hat{\pi}_m^{F_X, F_X} - \pi_m^{F_X}\|_{f_X}^2 \right] &= \sum_{j_1=1}^{D_{m_1}} \sum_{j_2=1}^{D_{m_2}} \text{Var} \left(\hat{a}_{j_1, j_2}^{F_X} \right), \\ (6) \qquad \qquad \qquad &\leq \frac{1}{n} \sum_{j_1=1}^{D_{m_1}} \sum_{j_2=1}^{D_{m_2}} \mathbb{E} \left[(\varphi_{j_1}(F_X(X_i)) \varphi_{j_2}(Y_i))^2 \right] \leq \frac{D_{m_1} D_{m_2}}{n}, \end{aligned}$$

using Property (3). The best model among the collection is the one which minimises the right-hand side in (5), 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 this compromise automatically.

2.3.2. Selection rule. We propose to use a scheme proposed by Goldenshluger and Lepski [2011] for density estimation. The adaptive index is chosen as follows:

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

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

$$(7) \qquad \qquad \qquad V^{\hat{F}} : m = (m_1, m_2) \mapsto c_1 \frac{D_{m_1} D_{m_2}}{n},$$

with c_1 a purely numerical constant. Its theoretical calibration is precised in the proofs (Section 5.3.1), and, in practice, we adjust it by simulations (see details in Section 4). The function

$A(., \hat{F}_n)$ is based on the comparison of the estimators built in the first stage:

$$(8) \quad A(m, \hat{F}_n) = \max_{m' \in \mathcal{M}_n} \left(\left\| \hat{h}_{m'}^{\hat{F}} - \hat{h}_{m \wedge m'}^{\hat{F}} \right\|^2 - V^{\hat{F}}(m') \right)_+,$$

where $x_+ = \max(x, 0)$, $x \in \mathbb{R}$ and $m \wedge m' = (m_1 \wedge m'_1, m_2 \wedge m'_2)$. We prove that $A(m, \hat{F}_n)$ has the order of the bias term (see Inequality (17)). Thus we get an estimator, explicitly expressed in a warped basis,

$$(9) \quad \tilde{\pi}(x, y) = \hat{h}_{\hat{m}^{\hat{F}}}^{\hat{F}}(\hat{F}_n(x), y).$$

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 must be highlighted compared to other strategies of model selection using the contrast function or to strategies involving bandwidth 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 from Goldenshluger and Lepski [2011] 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. Once the constant c_1 is chosen (through simulation experiments), the "penalty" term $V^{\hat{F}}$ is entirely data driven, while the penalty in the regression-type contrast depends, in addition, on the unknown infinite norm of π (see Brunel et al. 2007, or Lacour 2007).

Finally, let us define an estimator in the theoretical case of known c.d.f. F_X :

$$(10) \quad \tilde{\pi}_0(x, y) = \hat{h}_{\hat{m}^{F_X}}^{F_X}(F_X(x), y),$$

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 needs not 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 they can also be characterised via weak differentiability (see for example DeVore and Lorentz 1993 and Härdle et al. 1998 for functions of one variable, and Adams 1975 for functions of several variables). Precisely, our aim is to extend the characterization of Tsybakov (Lemma A.3, p.162, Tsybakov 2004) to functions of two variables.

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 $\alpha = (\alpha_1, \alpha_2)$ ($\alpha_l \in \mathbb{N}$, $l = 1, 2$, $\alpha \neq (0, 0)$), if

$$(11) \quad \sum_{j_1, j_2 \in \mathbb{N} \setminus \{0\}} \mu_{j_1, \alpha_1}^2 \mu_{j_2, \alpha_2}^2 \theta_{j_1, j_2}^2 \leq \frac{L^2}{\pi^{2(\alpha_1 + \alpha_2)}},$$

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 $t \in W_{per}^2([0; 1]^2, L, \alpha)$, in the spirit of the definition of Tsybakov [2004]. 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:

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

where $C(\alpha, L)$ is a constant depending on α and L . This inequality is a particular case of Lemma 9 in Lacour [2007], based on papers by Hochmuth [2002] and Nikol'skiĭ [1975].

3.2. Nonasymptotic risk bound and consequence. The first theorem provides a nonasymptotic bound for the risk of the estimator $\tilde{\pi}$ (see its definition (9)). We recall that the trigonometric models satisfy properties (3) and (4), and that the dimensions D_{m_l} are bounded by \sqrt{n} . Hereafter we denote by $\|\cdot\|_{\infty, [0;1]}$ the infinite norm of a function over the interval $[0; 1]$.

Theorem 1. *We assume that the function h belongs to the anisotropic Sobolev ball denoted by $W_{per}^2([0; 1]^2, L, (1, 0))$, 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*

$$(12) \quad \forall m = (m_1, m_2) \in \mathcal{M}_n, \quad D_{m_1} \leq C_a \left(\frac{n}{\ln^2(n)} \right)^{1/3} \quad \text{and} \quad C_b \ln^5(n) \leq D_{m_2} \leq C_c \sqrt{n}.$$

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

$$(13) \quad \mathbb{E} \left[\|\tilde{\pi} - \pi\|_{f_X}^2 \right] \leq c \min_{m \in \mathcal{M}_n} \left\{ \frac{D_{m_1} D_{m_2}}{n} + \|\pi_m^{F_X} - \pi\|_{f_X}^2 \right\} + \frac{C}{n}.$$

Remark 2. • There actually exists an integer n_0 , depending on the function h , such that Inequality (13) holds for all $n \geq n_0$ with a purely numerical constant c . But the result is nonasymptotic, since the inequality also holds 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 dimensions have played the same role. But in the theorem, the dimension constraints (12) are not the same in each direction. To be totally rigorous, we should denote by $S_{m_l}^{(l)}$ the models and by $D_{m_l}^{(l)}$ their dimension, for each $l = 1, 2$. For the sake of simplicity, we keep the first notations as there is no possible confusion.
- If we focus on the simpler situation of known c.d.f., we can derive the same result as Inequality (13) for the estimator $\tilde{\pi}_0$ (defined by (10)), with few assumptions: we only assume that the function h is bounded on the space $[0; 1] \times A_2$, and we have no additional condition on the trigonometric model. In this case, the constant c in (13) is purely numerical, and the other constant C depends only on $\|h\|_{\infty}$. This result holds in a more general setting than trigonometric models. It is sufficient to assume that the models satisfy properties (3) and (4), and have dimensions bounded by \sqrt{n} . These assumptions are very weak. Since the proof of this result is both short and representative of the method used to prove Theorem 1, we begin with it in the proof section (see Section 5.2).

The basic outline of model selection (with a method inspired by Goldenshluger and Lepski 2011) is to estimate the bias-variance sum and to select the model which minimises it. Theorem 1 shows that it is a good strategy: the right model (in the sense that it realises 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 $\{\pi_m^{\hat{F}, \hat{F}}, m \in \mathcal{M}_n\}$, 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.

Brunel et al. [2007] provide 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 Brunel et al. [2007], which depends on $\|\pi\|_\infty$. Moreover, replacing this quantity by an estimator requires much more regularity constraints than the one we have, and leads to a semi-asymptotic result (see the appendix of Lacour 2007 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.

Theorem 1 also enables us to give a rate of convergence for the estimation of π , under regularity assumptions for function h . Recall that the bound of Inequality (13) is close to the order of the sum of the variance term and the bias term. The minimization of the left-hand-side of the inequality in the case of regular functions leads to the following corollary, which implies that the obtained rate of convergence is likely to be minimax in most cases.

Corollary 1. *We assume that the function h belongs to the anisotropic Sobolev ball denoted by $W_{per}^2([0; 1]^2, L, \alpha)$, for some fixed $L > 0$, and $\alpha = (\alpha_1, \alpha_2)$ ($\alpha_l \in \mathbb{N}$, $l = 1, 2$, $\alpha \neq (0, 0)$) 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 1,*

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

with $C(\alpha, L)$ a numerical constant which depends 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 .

The harmonic mean of α_1 and α_2 is the real $\bar{\alpha}$ such that $2/\bar{\alpha} = 1/\alpha_1 + 1/\alpha_2$. The corollary signifies that without knowing α and L (depending on the unknown h), $\tilde{\pi}$ does as well as the best possible estimator which knows these quantities. It is thus an adaptive estimator.

4. SIMULATION STUDY

The aim of this section is to illustrate the behaviour of the estimator $\tilde{\pi}$ and to compare it with the regression-type estimator of Brunel et al. [2007] and with kernel estimators [Fan and Yim, 2004].

4.1. Implementation. The estimate $\tilde{\pi}$ is computed by using simulated sample of data $(X_i, Y_i)_{i=1, \dots, n}$. The empirical c.d.f. function \hat{F}_n is the one of the sample $(X_i)_i$: in practice, we do not use additional observations. We actually experimented it, and a second sample does not improve the results.

For each data sample (that is for each computation of the estimators), we calibrate the set $A_1 \times A_2$ for the estimation, over 90% of the variables (X_i, Y_i) : we choose to eliminate the smallest values, and the largest values of the data to avoid side effects.

To implement each estimator $\tilde{\pi}$, we use warped trigonometric basis. Recall that our procedure is based first on the estimation of the function h , which belongs to $L^2([0; 1] \times A_2)$. Accordingly, the orthogonal Fourier basis is $(\varphi_{j_1} \otimes \psi_{j_2})_{j_1, j_2}$, with φ_{j_1} defined by (2), and ψ_{j_2} obtained by:

$$\forall y \in [a_2; b_2], \psi_{j_2}(y) = \frac{1}{\sqrt{b_2 - a_2}} \varphi_{j_2} \left(\frac{y - a_2}{b_2 - a_2} \right),$$

where $A_2 = [a_2; b_2]$.

We have to compute the sum $A(m, \hat{F}_n) + 2V^{\hat{F}}(m)$ for each $m = (m_1, m_2)$. Notice that the quadratic norm in the definition of $A(m, \hat{F}_n)$ (see (8)) is simply equal to a sum of squared-coefficients. For example, if $m \wedge m' = (m_1, m'_2)$,

$$\left\| \hat{h}_{m'}^{\hat{F}} - \hat{h}_{m \wedge m'}^{\hat{F}} \right\|^2 = \sum_{j=D_{m_1}+1}^{D_{m'_1}} \sum_{k=1}^{D_{m'_2}} \left(\hat{a}_{j,k}^{\hat{F}} \right)^2.$$

Finally, we calibrate the numerical constant c_1 involved in the definition of $V^{\hat{F}}$ (see (7)). As in most model selection results, the theoretical value obtained in the proof (Section 5.3.1) is very pessimistic due to rough upper-bounds (for the sake of simplicity of the proof). Thus, specific data-driven calibration has been developed: for example, the so-called "dimension jump" method allows us to apply the slope heuristic (see Baudry et al. 2011) to choose the penalty constant of a classical penalised contrast strategy [Barron et al., 1999]. But this cannot be applied to the recent method of Goldenshluger and Lepski. Consequently, we adjust c_1 prior to the comparison with the other estimates: we look at the quadratic risk with respect to the value of c_1 , and choose one of the values leading to reasonable risk and complexity of the selected models (recall that in penalty calibration, it is more secure to overpenalise): we set thus $c_1 = 0.05$.

4.2. Comparison with kernel estimates. Since kernel methods play a key role in nonparametric conditional density estimation, we begin by comparing the adaptive warped-bases estimate $\tilde{\pi}$ with double-kernel local linear regressions built with a data-driven selection of the bandwidths. Fan and Yim [2004] recall the definition of this estimate. They numerically compare the bandwidths selection rule they propose (a cross-validation strategy based on minimization of the integrated squared error, with three other selection strategies: an ad-hoc empirical bias method [Fan et al., 1996], a bootstrap method (developed first by Hall et al. 1999, adapted by Bashannyk and Hyndman 2001), and a simple rule, based on local parametric models [Hyndman and Yao, 2002]). They show that in most cases, the cross-validation method outperforms the other approaches substantially. Thus, we evaluate $\tilde{\pi}$ by comparison with the cross-validation double-kernel estimate, denoted by $\tilde{\pi}_{FY}$.

We consider the examples corresponding to the independent random sample setting in Fan and Yim [2004]:

- Examples 1': $Y_i = 0.23X_i(16 - X_i) + 0.4\varepsilon_i$, with the X_i uniformly distributed on the interval $[0; 16]$ and the ε_i are independent, (i) standard normal, (ii) from the Student distribution t_2 , (iii) from the Student distribution t_4 .
- Example 2': $Y_i = 20 \cos(\pi X_i/10) + \varepsilon_i$, with the X_i uniformly distributed on the interval $[-20; 20]$ and the ε_i are standard normal, independent.

For sample size $n = 1000$, we compute the root mean squared error (RMSE) with the same formula as in Section 4 of Fan and Yim [2004], for the estimator $\tilde{\pi}$. Figure 1 presents an example of estimate for Example 1' and Table 1 summarises the results: we report the risk obtained here for $\tilde{\pi}$ and the risk of $\tilde{\pi}_{FY}$ obtained by Fan and Yim [2004] (see their Tables 1 and 2). We do not mention the median of the RMSE since it is not significantly different from the mean for the estimator $\tilde{\pi}$.

The values are similar for both estimators in most cases, but slightly better for the warped-bases estimator (in bold in Table 1). This result has to be put into perspective, since more thorough numerical trials have to be conducted to confirm this. The aim was just here to show that a warped-bases adaptive strategy can compete with kernel methods.

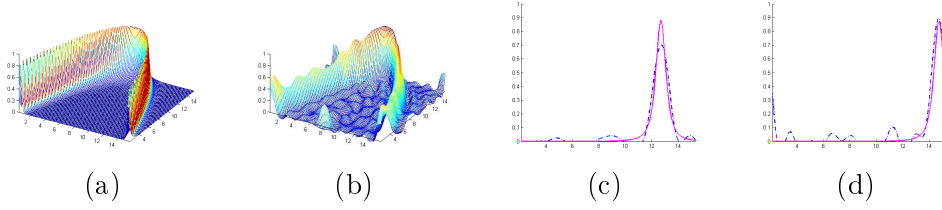


FIGURE 1. Plots of true function versus estimator $\tilde{\pi}$, Example 1' (ii) with $n = 1000$ observations. (a) true function π , (b) estimator $\tilde{\pi}$, (c) and (d) plots of $y \mapsto \pi(x, y)$ (full line), $y \mapsto \tilde{\pi}(x, y)$ (dashed dotted line) for two different fixed x .

	Ex 1' (i)	Ex1' (ii)	Ex1' (iii)	Ex2'
$\tilde{\pi}$	0.7801	0.6345	0.6891	1.0404
$\tilde{\pi}_{FY}$	1.0899	0.7641	1.0143	2.7404

TABLE 1. Values of $\text{RMSE} \times 1000$ averaged over 100 samples of size $n = 1000$ according to Fan and Yim Fan and Yim [2004], in Examples 1' and 2' for the estimators $\tilde{\pi}$ and $\tilde{\pi}_{FY}$.

4.3. Comparison with regression-type estimator. We also compare $\tilde{\pi}$ with another adaptive estimator: the one of Brunel et al. [2007] denoted by $\tilde{\pi}_{BCL}$. The estimator $\tilde{\pi}_{BCL}$ of Brunel et al. [2007] is defined as a penalised least-squares contrast estimator. The penalty is $\text{pen}(m) = K_0 \|\pi\|_{\infty} D_{m_1} D_{m_2} / n$. We implement the method with $K_0 = 0.5$ like in Brunel et al. [2007] but we do not replace $\|\pi\|_{\infty}$ by its theoretical value. 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 $m = \lceil (\ln(n) - 1)/2 \rceil$.

The aim is to investigate at 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.

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

- Example 1: $Y_i = b(X_i) + \varepsilon_i$, with the following possibilities. The X_i follow a uniform distribution on the interval $[0; 1]$ (denoted by $\mathcal{U}_{[0;1]}$), or on the interval $[-1; 1]$ ($\mathcal{U}_{[-1;1]}$), a standard Gaussian distribution ($\mathcal{N}(0, 1)$). The ε_i 's are generated following the standard Gaussian distribution, the Gaussian distribution with variance 4 ($\mathcal{N}(0, 4)$), 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 $b(x) = 2x + 5$ or $b(x) = x^2$. The conditional density π is thus given by

$$\pi(x, y) = f_{\varepsilon}(y - b(x)).$$

- Example 2: $Y_i = b(X_i) + \sigma(X_i)\varepsilon_i$, 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

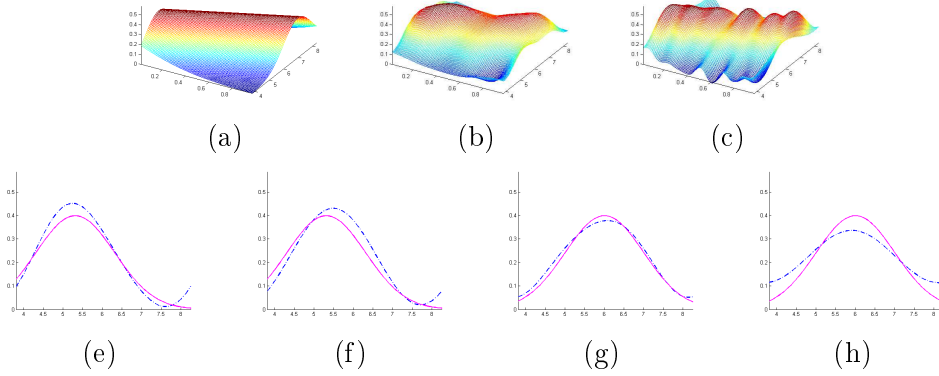


FIGURE 2. Plots of true function versus estimators, Example 1, with X_i i.i.d. $\mathcal{U}_{[0;1]}$, ε_i i.i.d. $\mathcal{N}(0,1)$, and $b(x) = 2x + 5$ with $n = 2000$ observations. First line: (a) true function π , (b) estimator $\tilde{\pi}$, (c) estimator $\tilde{\pi}_{BCL}$. Second line, (d) and (f): plots of $y \mapsto \pi(x, y)$ (full line) and $y \mapsto \tilde{\pi}(x, y)$ (dashed dotted line) for two different fixed x . Second line, (e) and (g): for two fixed x , plots of $y \mapsto \pi(x, y)$ (full line) and $y \mapsto \tilde{\pi}_{BCL}(x, y)$ (dashed dotted line) .

to Example 1, the conditional density is

$$\pi(x, y) = f_\varepsilon(y - b(x)/\sigma(x))/\sigma(x).$$

- Example 3: 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.
- Example 4: X_i follows a uniform distribution $\mathcal{U}_{[0;1]}$, and given $X_i = x$, Y_i follows a Gamma distribution with parameters 3 and $x^2 + 1$. The function π is the corresponding Gamma density.

Examples 2 and 3, and some cases of Example 1 have also been studied by Brunel et al. [2007], and Example 5 by Fan and Yim [2004] (but no risks for this model are given for the last two examples).

Figures 2 and 3 illustrate the visual quality of the reconstruction, for a case of Example 1, and for Example 3. 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...).

For sample sizes $n = 60, 500$ and 2000 , we give in Tables 2 and 3 the estimated values of the risk $\mathbb{E}[\|\hat{\pi} - \pi\|_2^2]$, with $\|\cdot\|_2$ 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 MISE is computed over $N = 100$ replicated samples, and the quadratic norm is approximated using subdivisions of A_1 and A_2 (see Brunel et al. 2007, Section 5.1, for details about the formula).

The risk of our estimator $\tilde{\pi}$ is often better than the one of the penalised least-squares estimator $\tilde{\pi}_{BCL}$ (in bold in the tables). Precisely, it is always smaller for the sample sizes $n = 500$ and

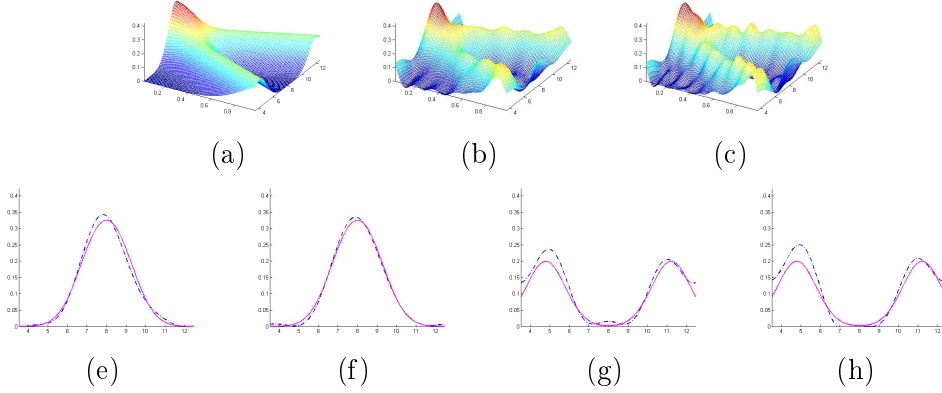


FIGURE 3. Plots of true function versus estimators, Example 3 with $n = 2000$ observations. First line: (a) true function π , (b) estimator $\tilde{\pi}$, (c) estimator $\tilde{\pi}_{BCL}$. Second line, (d) and (f): plots of $y \mapsto \pi(x, y)$ (full line) and $y \mapsto \tilde{\pi}(x, y)$ (dashed dotted line) for two different fixed x . Second line, (e) and (g): for two fixed x , plots of $y \mapsto \pi(x, y)$ (full line) and $y \mapsto \tilde{\pi}_{BCL}(x, y)$ (dashed dotted line) .

$n = 2000$, which confirms that one can easily use the warped-bases estimator for estimation problems with large samples of data, in spite of its bad performances for very small sample sizes.

5. PROOFS

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

5.1. Preliminary results. Let us start by setting a result which is the key argument in the proof of the main theorem. We consider the centred empirical process defined by

$$(14) \quad \forall t \in L^2([0; 1] \times A_2), \nu_n(t) = \frac{1}{n} \sum_{i=1}^n t(F_X(X_i), Y_i) - \mathbb{E}[t(F_X(X_i), Y_i)].$$

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

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

Proposition 2. *If the function h is bounded on $[0; 1] \times A_2$, for all $\delta > 0$, there exists a constant $C > 0$, depending on $\|h\|_{\infty}$, such that,*

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

Proof of Proposition 2. 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.

$b(x)$	ε	X	$n = 60$	500	2000	Method
$2x + 5$	$\mathcal{N}(0, 1)$	$\mathcal{U}_{[0;1]}$	19.81	2.879	1.327	$\tilde{\pi}$
			7.446	2.536	1.409	$\tilde{\pi}_{BCL}$
		$\mathcal{U}_{[-1;1]}$	20.871	4.811	2.971	$\tilde{\pi}$
			9.443	6.384	5.501	$\tilde{\pi}_{BCL}$
		$\mathcal{N}(0, 1)$	38.255	14.833	11.038	$\tilde{\pi}$
			37.374	40.295	38.993	$\tilde{\pi}_{BCL}$
	$\Gamma(4, 1)$	$\mathcal{U}_{[0;1]}$	5.628	0.969	0.479	$\tilde{\pi}$
			2.361	1.417	0.715	$\tilde{\pi}_{BCL}$
		$\mathcal{U}_{[-1;1]}$	9.224	2.255	1.376	$\tilde{\pi}$
			6.666	3.569	2.314	$\tilde{\pi}_{BCL}$
		$\mathcal{N}(0, 1)$	20.094	8.641	6.093	$\tilde{\pi}$
			24.749	20.201	22.571	$\tilde{\pi}_{BCL}$
x^2	$\mathcal{N}(0, 2)$	$\mathcal{U}_{[0;1]}$	5.06	0.288	0.258	$\tilde{\pi}$
			2.986	0.527	0.548	$\tilde{\pi}_{BCL}$
		$\mathcal{U}_{[-1;1]}$	10.381	0.546	0.428	$\tilde{\pi}$
			4.277	0.846	1.033	$\tilde{\pi}_{BCL}$
		$\mathcal{N}(0, 1)$	20.113	2.658	2.379	$\tilde{\pi}$
			14.442	2.754	2.395	$\tilde{\pi}_{BCL}$
	$\Gamma(4, 1)$	$\mathcal{U}_{[0;1]}$	5.845	0.811	0.351	$\tilde{\pi}$
			2.505	0.894	0.547	$\tilde{\pi}_{BCL}$
		$\mathcal{U}_{[-1;1]}$	10.828	0.664	0.599	$\tilde{\pi}$
			4.672	1.144	0.950	$\tilde{\pi}_{BCL}$
		$\mathcal{N}(0, 1)$	22.337	6.277	3.551	$\tilde{\pi}$
			18.723	7.367	3.792	$\tilde{\pi}_{BCL}$

TABLE 2. Values of $\text{MISE} \times 100$ averaged over 100 samples, in Examples 1 (regression models) for the estimators $\tilde{\pi}$ and $\tilde{\pi}_{BCL}$.

Lemma 3. *Let ξ_1, \dots, ξ_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*

$$\begin{aligned}
 (15) \quad \mathbb{E} \left[\left(\sup_{r \in \mathcal{R}} (\nu_n(r))^2 - c(\delta)H^2 \right)_+ \right] &\leq c_1 \left\{ \frac{v}{n} \exp \left(-c_2 \delta \frac{nH^2}{v} \right) \right. \\
 &\quad \left. + \frac{M_1^2}{C^2(\delta)n^2} \exp \left(-c_3 C(\delta) \sqrt{\delta} \frac{nH}{M_1} \right) \right\},
 \end{aligned}$$

Example	$n = 60$	500	2000	Method
Ex 2	7.621	1.163	0.498	$\tilde{\pi}$
	2.739	1.178	0.657	$\tilde{\pi}_{BCL}$
Ex 3	14.451	12.264	12.87	$\tilde{\pi}$
	13.153	12.764	13.175	$\tilde{\pi}_{BCL}$
Ex 4	3.617	0.907	0.557	$\tilde{\pi}$
	2.407	1.142	0.611	$\tilde{\pi}_{BCL}$

TABLE 3. Values of MISE $\times 100$ averaged over 100 samples, in Example 2,3,4,5 for the estimators $\tilde{\pi}$ and $\tilde{\pi}_{BCL}$.

with, $C(\delta) = (\sqrt{1+\delta} - 1) \wedge 1$, $c(\delta) = 2(1 + 2\delta)$ and

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

Inequality (15) is a classical consequence of Talagrand's Inequality given in Klein and Rio [2005]: see for example Lemma 5 (page 812) in Lacour [2008]. 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 :

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

Finally, $\text{Var}(t(F_X(X_1), Y_1)) \leq \mathbb{E}[t^2(F_X(X_1), Y_1)] \leq \|t\|^2 \|h\|_{\infty} = \|h\|_{\infty} := v$. We just replace the quantities M_1, H and v by the values derived above in Inequality (15):

$$\begin{aligned} & \sum_{m' \in \mathcal{M}_n} \mathbb{E} \left[\left(\sup_{t \in \mathcal{S}(m')} \nu_n(t)^2 - c(\delta) \frac{D_{m'_1} D_{m'_2}}{n} \right)_+ \right] \\ & \leq c_1 \left\{ \sum_{m' \in \mathcal{M}_n} \frac{1}{n} \exp(-c_2 D_{m'_1} D_{m'_2}) + \sum_{m' \in \mathcal{M}_n} \frac{D_{m'_1} D_{m'_2}}{n^2} \exp(-c_3 \sqrt{n}) \right\}. \end{aligned}$$

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. \square

We also set the following useful lemma.

Proposition 4. *Let $\nu : L^2([0; 1] \times A_2) \mapsto \mathbb{R}$ be a linear functional. Let also $m = (m_1, m_2)$ be an index of the collection \mathcal{M}_n . Then,*

$$\sup_{t \in \mathcal{S}(m)} \nu^2(t) = \sum_{j=1}^{D_{m_1}} \sum_{k=1}^{D_{m_2}} \nu^2(\varphi_j \otimes \varphi_k).$$

Proof of Proposition 4.

If t belongs to $\mathcal{S}(m)$, 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 $\sum_{j=1}^{D_{m_1}} \sum_{k=1}^{D_{m_2}} b_{j,k}^2 = 1$. Thus, by the linearity of ν and the Cauchy-Schwarz Inequality,

$$\nu^2(t) = \left(\sum_{j=1}^{D_{m_1}} \sum_{k=1}^{D_{m_2}} b_{j,k} \nu(\varphi_j \otimes \varphi_k) \right)^2 \leq \sum_{j=1}^{D_{m_1}} \sum_{k=1}^{D_{m_2}} \nu^2(\varphi_j \otimes \varphi_k).$$

This leads to $\sup_{t \in \mathcal{S}(m)} \nu^2(t) \leq \sum_{j=1}^{D_{m_1}} \sum_{k=1}^{D_{m_2}} \nu^2(\varphi_j \otimes \varphi_k)$. The equality is obtained by choosing $t = \sum_{j=1}^{D_{m_1}} \sum_{k=1}^{D_{m_2}} b_{j,k} \varphi_j \otimes \varphi_k \in L^2([0; 1])$, with $b_{j,k} = \nu(\varphi_j \otimes \varphi_k) / (\sum_{j'=1}^{D_{m_1}} \sum_{k'=1}^{D_{m_2}} \nu^2(\varphi_{j'} \otimes \varphi_{k'}))$. \square

5.2. Proof of Inequality (13) in the theoretical case of known c.d.f F_X . We deal first with the estimator $\tilde{\pi}_0$ (defined by (10)): it satisfies the following inequality:

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

and its proof is a simple example of the scheme we will use to prove the main result, Theorem 1. For the sake of simplicity, in this section, we denote 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 of Inequality (13). We decompose the loss of the estimator as follows:

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

By definition of A ,

$$\|\hat{h}_{\hat{m}}^{F_X} - h\|^2 \leq 3(A(m) + V(\hat{m})) + 3(A(\hat{m}) + V(m)) + 3\|\hat{h}_m^{F_X} - h\|^2,$$

Moreover, by definition of \hat{m} , $A(\hat{m}) + V(\hat{m}) \leq A(m) + V(m)$, which leads to

$$\|\hat{h}_{\hat{m}}^{F_X} - h\|^2 \leq 6(A(m) + V(m)) + 3\|\hat{h}_m^{F_X} - h\|^2.$$

We have already bounded the risk of the estimator on a fixed model (see Section 2.3.1, Inequalities (5) and (6)), therefore, by definition of V , we get

$$(16) \quad \mathbb{E} \left[\|\hat{h}_{\hat{m}}^{F_X} - h\|^2 \right] \leq 6\mathbb{E}[A(m)] + (6c_0 + 3) \frac{D_{m_1} D_{m_2}}{n} + 3\|h_m - h\|^2.$$

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

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

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

Lemma 5. *If the function h is bounded on $[0; 1] \times A_2$, there exists a constant $C > 0$ such that, for $m \in \mathcal{M}_n$,*

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

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

Lemma 6. *If the function h is bounded on $[0; 1] \times A_2$, there exists a constant $C > 0$ such that,*

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

These inequalities imply that

$$(17) \quad \mathbb{E}[A(m)] \leq \frac{C}{n} + 12\|h_m - h\|^2.$$

Gathering this with Inequality (16) ends the proof of the Theorem. \square

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). Using Proposition 4, we first compute,

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

with ν_n the empirical process defined by (14). 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 2.

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 2 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, we have $m \wedge m' = (m'_1, m_2)$. We use $V((m'_1, m_2)) \leq V(m'_1, m'_2)$ 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 2 (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 previous 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, by denoting 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:

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

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 $m'_2 \leq m_2$. By symmetry, this case can be handled similarly to the latter.

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

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

□

5.3. Proof of Theorem 1. 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 Section 5.2. The computations are more technical, since the estimator $\tilde{\pi} = \hat{h}_{\hat{m}}^{\hat{F}}(\hat{F}(\cdot), \cdot)$ depends in two ways on \hat{F} . We denote it by $\hat{\pi}_{\hat{m}}^{\hat{F}, \hat{F}}$, and coherently, we denote by $\hat{\pi}_{\hat{m}}^{F_X, F_X}$ the estimator previously studied, that is $\tilde{\pi}_0$. We also introduce the following intermediate:

$$(18) \quad \forall (x, y) \in A_1 \times A_2, \quad \hat{\pi}^{\hat{F}, F_X}(x, y) = \hat{h}_{\hat{m}}^{\hat{F}}(F_X(x), y).$$

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

$$\begin{aligned}
 (19) \quad T_0^m &= \|\pi - \pi_m^{F_X}\|_{f_X}^2 + \|\pi_m^{F_X} - \hat{\pi}_m^{F_X, F_X}\|_{f_X}^2, \\
 T_1^m &= \left\| \hat{\pi}_m^{F_X, F_X} - \hat{\pi}_m^{\hat{F}, F_X} - \mathbb{E} \left[\hat{\pi}_m^{F_X, F_X} - \hat{\pi}_m^{\hat{F}, F_X} | (X_{-l})_l \right] \right\|_{f_X}^2, \\
 T_2^m &= \left\| \hat{\pi}_m^{\hat{F}, F_X} - \hat{\pi}_m^{\hat{F}, \hat{F}} - \mathbb{E} \left[\hat{\pi}_m^{\hat{F}, F_X} - \hat{\pi}_m^{\hat{F}, \hat{F}} | (X_{-l})_l \right] \right\|_{f_X}^2, \\
 T_3^m &= \left\| \mathbb{E} \left[\hat{\pi}_m^{F_X, F_X} - \hat{\pi}_m^{\hat{F}, F_X} | (X_{-l})_l \right] \right\|_{f_X}^2, \quad T_4^m = \left\| \mathbb{E} \left[\hat{\pi}_m^{\hat{F}, F_X} - \hat{\pi}_m^{\hat{F}, \hat{F}} | (X_{-l})_l \right] \right\|_{f_X}^2.
 \end{aligned}$$

Let us remark that T_0^m is the bias-variance decomposition for the risk of an estimator $\hat{\pi}_m^{F_X, F_X}$, 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 centre 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 (18) in the loss of our estimator:

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

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

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

Furthermore, $\left\| \hat{h}_m^{\hat{F}} - h \right\|^2 = \left\| \hat{\pi}_m^{\hat{F}, F_X} - \pi \right\|_{f_X}^2 \leq 3T_1^m + 3T_3^m + 6T_0^m$. Consequently,

$$(20) \quad \begin{aligned} \left\| \hat{\pi}_{\hat{m}}^{\hat{F}, \hat{F}} - \pi \right\|_{f_X}^2 &\leq 3T_2^{\hat{m}} + 3T_4^{\hat{m}} - 3 \times 2V(\hat{m}) + 3 \times 6(A(m) + 2V(m)) \\ &\quad + 3 \times 3 \times (3T_1^m + 3T_3^m + 6T_0^m), \end{aligned}$$

where the terms T_l^m , $l = 0, \dots, 4$ are defined by (19). We split the term A , first in a similar way as in Section 5.2. 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 (19). By applying Proposition 4, we get, for $p = m$ or $p = m \wedge m'$,

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

for a function $t \in L^2([0; 1] \times A_2)$. We recover the previous empirical process by the decomposition $\tilde{\nu}_n^2(t) \leq 2\nu_n^2(t) + 2R_n^2(t)$, 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 $t = \sum_{j=1}^{D_{p1}} \sum_{k=1}^{D_{p2}} \theta_{j,k} \varphi_j \otimes \varphi_k$, with $\sum_{j=1}^{D_{p1}} \sum_{k=1}^{D_{p2}} \theta_{j,k}^2 = 1$. 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 $\sup_{t \in \mathcal{S}(p)} R_n^2(t) = \sum_{j=1}^{D_{p1}} \sum_{k=1}^{D_{p2}} (\hat{a}_{j,k}^{\hat{F}} - \hat{a}_{j,k}^{F_X})^2$. 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,

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

By subtracting $V(m')$, taking the maximum over $m' \in \mathcal{M}_n$ and integrating give an upper-bound for $\mathbb{E}[A(m)]$. We introduce it into (20) to obtain:

$$\begin{aligned}
 (21) \quad & \mathbb{E} \left[\left\| \hat{\pi}_{\hat{m}}^{\hat{F}, \hat{F}} - \pi \right\|_{f_X}^2 \right] \\
 & \leq 36V(m) + 27\mathbb{E}[2T_0^m + T_1^m + T_3^m] + 18 \left\{ 3 \max_{m' \in \mathcal{M}_n} \|h_{m \wedge m'} - h_{m'}\|^2 \right. \\
 & \quad + 3\mathbb{E} \left[\left(T_2^{\hat{m}} - V(\hat{m}) \right)_+ \right] + 3\mathbb{E} \left[\left(T_4^{\hat{m}} - V(\hat{m}) \right)_+ \right] \\
 & \quad + 6\mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(\sup_{t \in \mathcal{S}(m')} (\nu_n(t))^2 - \frac{V(m')}{18 \times 36} \right)_+ \right] \\
 & \quad + 6\mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(\sup_{t \in \mathcal{S}(m \wedge m')} (\nu_n(t))^2 - \frac{V(m')}{18 \times 36} \right)_+ \right] \\
 & \quad + 12\mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(T_3^{m'} - \frac{V(m')}{18 \times 72} \right)_+ \right] + 12\mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(T_3^{m \wedge m'} - \frac{V(m')}{18 \times 72} \right)_+ \right] \\
 & \quad + 12\mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(T_1^{m'} - \frac{V(m')}{18 \times 72} \right)_+ \right] + 12\mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(T_1^{m \wedge m'} - \frac{V(m')}{18 \times 72} \right)_+ \right] \Big\}. \\
 (22)
 \end{aligned}$$

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

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

using (5) and (6). Moreover, applying twice Proposition 2 shows that

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

with $V_0(m') = 2(1 + 2\delta)D_{m'_1}D_{m'_2}/n$. Choosing c_1 (see the definition (7)) 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 (21) for the risk, we get

$$\begin{aligned}
 (23) \quad & \mathbb{E} \left[\left\| \hat{\pi}_{\hat{m}}^{\hat{F}, \hat{F}} - \pi \right\|_{f_X}^2 \right] \leq 36V(m) + 27 \times 2 \frac{D_{m_1} D_{m_2}}{n} + (12 \times 18 + 27 \times 2)\|h_m - h\|^2 + \frac{C}{n} \\
 & \quad + 3\mathbb{E} \left[\left(T_2^{\hat{m}} - V(\hat{m}) \right)_+ \right] + 3\mathbb{E} \left[\left(T_4^{\hat{m}} - V(\hat{m}) \right)_+ \right] + 27\mathbb{E}[T_1^m + T_3^m] \\
 & \quad + 18 \left\{ 12\mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(T_3^{m'} - \frac{V(m')}{18 \times 72} \right)_+ \right] + 12\mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(T_3^{m \wedge m'} - \frac{V(m')}{18 \times 72} \right)_+ \right] \right. \\
 & \quad \left. + 12\mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(T_1^{m'} - \frac{V(m')}{18 \times 72} \right)_+ \right] + 12\mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(T_1^{m \wedge m'} - \frac{V(m')}{18 \times 72} \right)_+ \right] \right\}.
 \end{aligned}$$

It remains to bound the terms T_l^m , $l = 1, 2, 3, 4$ or their centred 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 $\|\varphi'_2\|_{\infty, [0;1]}$ such that, for $m \in \mathcal{M}_n$,*

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

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 $C > 0$

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

with $V_1(m') = \kappa_1 D_{m'_1} D_{m'_2}/n$, and κ_1 a constant depending only on $\|\varphi'_2\|_{\infty, [0;1]}$.

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

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 $V_2(m') = \kappa_2 D_{m'_1}^4 D_{m'_2} \ln^2(n)/n^2$, 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 $\|\varphi_2^{(3)}\|_{\infty, [0;1]}$, $\|h\|$ and $\|\partial_1 h\|$ (∂_1 is the derivation operator with respect to the first variable) such that, for $m \in \mathcal{M}_n$,*

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

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 ,

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

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 $W_{per}^2([0;1]^2, L, (1, 0))$, 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 ($l = 1, \dots, 4$), to have $V(m') \geq V_l(m')$ (or $V_l^b(m')$ for $l = 2$): this enables to apply the inequalities of the lemmas with V and to use it in Inequality (23). We then obtain the result of Theorem 1. \square

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 $U_{-i} = F_X(X_{-i})$ ($i = 1, \dots, n$). 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 $(U_{-i})_{i=1, \dots, n}$. Let us also keep 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

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

We also recall some inequalities to control the deviations of the empirical c.d.f \hat{U}_n . Dvoretzky et al. [1956] established the first one.

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

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

By integration, we then deduce other bounds:

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

$$(24) \quad \mathbb{E} \left[\left\| \hat{U}_n - id \right\|_{\infty, [0;1]}^p \right] \leq \frac{C_p}{n^{p/2}},$$

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

$$(25) \quad \mathbb{E} \left[\left(\left\| \hat{U}_n - id \right\|_{\infty, [0;1]}^p - \kappa \frac{\ln^{p/2}(n)}{n^{p/2}} \right)_+ \right] \leq C n^{-c(p, \kappa)}, \text{ with } c(p, \kappa) = 2^{\frac{2-p}{p}} \kappa^{2/p}.$$

Moreover,

$$(26) \quad \mathbb{E} \left[\left(\left\| \hat{U}_n - id \right\|_{\infty, [0;1]}^2 - \kappa \frac{\ln(n)}{n} \right)^2 \right] \leq C n^{-2-2\kappa}.$$

Inequality (26) is a slightly more precise version of Inequality (25) 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 $T_1^m = \sum_{j,k} \{ (\hat{a}_{j,k}^{F_X} - \hat{a}_{j,k}^{\hat{F}}) - \mathbb{E}[\hat{a}_{j,k}^{F_X} - \hat{a}_{j,k}^{\hat{F}} | (X_{-l})_l] \}^2$. Thus,

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

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

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

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):

$$\begin{aligned} \mathbb{E}[T_1^m | (X_{-l})_l] &\leq \frac{1}{n} \left\| \sum_{k=1}^{D_{m_2}} \varphi_k^2 \right\|_{\infty, [0;1]} \sum_{j=1}^{D_{m_1}} \|\varphi'_j\|_{\infty, [0;1]}^2 \|F_X - \hat{F}_n\|_{\infty, A_1}^2, \\ &\leq \|\varphi'_2\|_{\infty, [0;1]}^2 \frac{D_{m_1}^3 D_{m_2}}{n} \|F_X - \hat{F}_n\|_{\infty, A_1}^2. \end{aligned}$$

It remains to use Inequality (24) of Proposition 12 with $p = 2$ to bound the expectation:

$$\mathbb{E}[T_1^m] \leq C \|\varphi'_2\|_{\infty, [0;1]}^2 \frac{D_{m_1}^3 D_{m_2}}{n^2}$$

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 $p = p_{m'}$. We apply Proposition 4, which leads to $T_1^p = \sup_{t \in \mathcal{S}(p)} (\nu_n^a(t))^2$ 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 centred conditionally to the sample $(X_{-l})_l$. Thus we apply Talagrand's inequality (15), as in the proof of Proposition 2, but conditionally to $(X_{-l})_l$. In this setting the key quantities are such that

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

We compute

$$M_{1,a} = \|\varphi'_2\|_{\infty,[0;1]} D_{p_1}^{3/2} D_{p_2}^{1/2} \left\| \hat{F}_n - F_X \right\|_{\infty, A_1},$$

$$H_{a,p}^2 = \frac{1}{n} \|\varphi'_2\|_{\infty,[0;1]}^2 D_{p_1}^3 D_{p_2} \left\| \hat{F}_n - F_X \right\|_{\infty, A_1}^2, \quad v_a = n H_{a,p}^2,$$

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

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

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 now put $p = m'$ (The case $p = m \wedge m'$ can be handled similarly). We thus have $\mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(T_1^{m'} - 2(1 + 2\kappa \ln(n)) H_{a,m'}^2 \right)_+ |(X_{-l})_l \right]$

$$\leq \sum_{m' \in \mathcal{M}_n} \mathbb{E} \left[\left(T_1^{m'} - 2(1 + 2\kappa \ln(n)) H_{a,m'}^2 \right)_+ |(X_{-l})_l \right],$$

$$\leq C \left\{ n^{-C\kappa} \sum_{m' \in \mathcal{M}_n} \frac{D_{m'_1}^3 D_{m'_2}}{n} + \exp(-C\sqrt{n}) \sum_{m' \in \mathcal{M}_n} \frac{D_{m'_1}^3 D_{m'_2}}{n^2} \right\}.$$

Moreover, we use $D_{m_l} = O(\sqrt{n})$ ($l = 1, 2$), and remark that the cardinal of \mathcal{M}_n is smaller than n , to get $\sum_{m' \in \mathcal{M}_n} D_{m'_1}^3 D_{m'_2} / n \leq \sum_{m' \in \mathcal{M}_n} C n^{3/2} n^{1/2} / n \leq C n^2$. Thus, if we choose κ large enough,

$$\mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(T_1^{m'} - 2(1 + 2\kappa \ln(n)) H_{a,m'}^2 \right)_+ |(X_{-l})_l \right] \leq C \{ n^{2-C\kappa} + n \exp(-C\sqrt{n}) \} \leq C n^{-1}.$$

We then notice that, for any $\alpha_n > 0$

$$2(1 + 2\kappa \ln(n)) H_{a,m'}^2 \leq 6\kappa \|\varphi'_2\|_{\infty,[0;1]}^2 \frac{D_{m'_1}^3 D_{m'_2} \ln(n)}{n} \left\| \hat{F}_n - F_X \right\|_{\infty, A_1}^2,$$

$$\leq 6\kappa \|\varphi'_2\|_{\infty,[0;1]}^2 \frac{D_{m'_1}^3 D_{m'_2} \ln(n)}{n} \left(\alpha_n^2 + \mathbf{1}_{\|\hat{F}_n - F_X\|_{\infty, A_1} \geq \alpha_n} \right).$$

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

$$2(1 + 2\kappa \ln(n)) H_{a,m'}^2 \leq 18\kappa \|\varphi'_2\|_{\infty,[0;1]}^2 \frac{D_{m'_1} D_{m'_2}}{n} + C \frac{n}{\ln^2(n)} \mathbf{1}_{\|\hat{F}_n - F_X\|_{\infty, A_1} \geq \alpha_n},$$

$$= V_1(m') + C \mathbf{1}_{\|\hat{F}_n - F_X\|_{\infty, A_1} \geq \alpha_n},$$

Besides,

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

$$\leq \mathbb{E} \left[\left(T_1^{m'} - 2(1 + 2\kappa \ln(n)) H_{a,m'}^2 \right)_+ \right] + C n^{-2} \ln^{-1}(n),$$

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 split 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 $T_2^{m'}$

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

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

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

Thus, $\mathbb{E}[(T_2^{m'} - V_2(m'))_+] \leq T_{2,a}^{m'} + T_{2,b}^{m'}$, with

$$\begin{aligned}
T_{2,a}^{m'} &= D_{m'_1}^3 \|\varphi'_2\|_{\infty, [0;1]}^2 \mathbb{E} \left[\sum_{j,k} \left(\hat{a}_{j,k}^{\hat{F}} - \mathbb{E} \left[\hat{a}_{j,k}^{\hat{F}} \mid (X_{-l})_l \right] \right)^2 \left(\left\| \hat{U}_n - id \right\|_{\infty, [0;1]}^2 - \kappa' \frac{\ln(n)}{n} \right)_+ \right], \\
T_{2,b}^{m'} &= D_{m'_1}^3 \|\varphi'_2\|_{\infty, [0;1]}^2 \kappa' \frac{\ln(n)}{n} \mathbb{E} \left[\left(\sum_{j,k} \left(\hat{a}_{j,k}^{\hat{F}} - \mathbb{E} \left[\hat{a}_{j,k}^{\hat{F}} \mid (X_{-l})_l \right] \right)^2 - \frac{\kappa}{\|\varphi'_2\|_{\infty, [0;1]}^2} \frac{D_{m'_1} D_{m'_2}}{n} \ln(n) \right)_+ \right].
\end{aligned}$$

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

$$\begin{aligned}
T_{2,a}^{m'} &\leq 2D_{m'_1}^4 D_{m'_2} \|\varphi'_2\|_{\infty, [0;1]}^2 \mathbb{E} \left[\left(\left\| \hat{U}_n - id \right\|_{\infty, [0;1]}^2 - \kappa' \frac{\ln(n)}{n} \right)_+ \right], \\
&\leq Cn^{4/2} n^{1/2} \left(\mathbb{E} \left[\left(\left\| \hat{U}_n - id \right\|_{\infty, [0;1]}^2 - \kappa' \frac{\ln(n)}{n} \right)^2 \right] \right)^{1/2} \text{ using } D_{m_l} \leq \sqrt{n} \ (l = 1, 2), \\
&\leq Cn^{5/2} n^{-1-\kappa'} = Cn^{-3/2-\kappa'} \text{ (Inequality (26))}
\end{aligned}$$

Thus, choosing $\kappa' \geq 7/2$, $\sum_{m' \in \mathcal{M}_n} T_{2,a}^{m'} \leq C/n$. For the second term $T_{2,b}^{m'}$, we notice first that $\sum_{j,k} (\hat{a}_{j,k}^{\hat{F}} - \mathbb{E}[\hat{a}_{j,k}^{\hat{F}} \mid (X_{-l})_l])^2 = \sup_{t \in \mathcal{S}(m')} (\nu_n^b)^2(t)$ (Proposition 4) with

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

We now bound the deviations of this empirical process, centred conditionally to (X_{-l}) , exactly as we bound ν_n^a in the proof of Lemma 7: they are controlled by the Talagrand Inequality (15). We finally obtain $\sum_{m' \in \mathcal{M}_n} T_{2,b}^{m'} \leq C \ln(n)/n$, which ends the proof, by gathering this bound with the one of $\sum_{m' \in \mathcal{M}_n} T_{2,a}^{m'}$. \square

5.3.5. Proof of Lemma 9. To compute a bound for $\mathbb{E}[T_3^m]$, let us begin with the definition of the estimators and their coefficients, to get $T_3^m = \sum_{j=1}^{D_{m_1}} \sum_{k=1}^{D_{m_2}} \{\langle \varphi_k, \Lambda_j(y) \rangle_{A_2}\}^2$ 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 $T_3^m = \sum_{j=1}^{D_{m_1}} \|\Pi_{S_{m_2}} \Lambda_j\|_{A_2}^2 \leq \sum_{j=1}^{D_{m_1}} \|\Lambda_j\|_{A_2}^2$, which can be developed as

$$T_3^m \leq \sum_{j=1}^{D_{m_1}} \int_{A_2} \left(\int_{[0;1]} (\varphi_j(\hat{U}_n(u)) - \varphi_j(u)) h(u, y) du \right)^2 dy := \int_{A_2} T_3'^m(y) dy.$$

We apply Taylor's formula with the Lagrange form for the remainder: there exists a random number depending on j , $\hat{\alpha}_{j,n,u}$, such that the following splitting holds:

$$\mathbb{E} [T_3'^m(y)] \leq 3\mathbb{E} [T_{3,1}^m(y)] + 3\mathbb{E} [T_{3,2}^m(y)] + 3\mathbb{E} [T_{3,3}^m(y)],$$

with notations

$$\begin{aligned} T_{3,1}^m(y) &= \sum_{j=1}^{D_{m_1}} \left\{ \int_0^1 h(u, y) (\hat{U}_n(u) - u) \varphi_j'(u) du \right\}^2, \\ T_{3,2}^m(y) &= (1/4) \sum_{j=1}^{D_{m_1}} \left\{ \int_0^1 h(u, y) (\hat{U}_n(u) - u)^2 \varphi_j''(u) du \right\}^2, \\ T_{3,3}^m(y) &= (1/36) \sum_{j=1}^{D_{m_1}} \left\{ \int_0^1 h(u, y) (\hat{U}_n(u) - u)^3 \varphi_j^{(3)}(\hat{\alpha}_{j,n,u}) du \right\}^2. \end{aligned}$$

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

$$\mathbb{E} [T_{3,1}^m(y)] = \mathbb{E} \left[\sum_{j=1}^{D_{m_1}} \left(\frac{1}{n} \sum_{i=1}^n A_{i,j}(y) - \mathbb{E}[A_{i,j}(y)] \right)^2 \right], \quad \text{with } A_{i,j}(y) = \int_{U_i}^1 h(u, y) \varphi_j'(u) du.$$

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 $y \in A_2$:

$$\mathbb{E} [T_{3,1}^m(y)] \leq 2\mathbb{E} [T_{3,1,1}^m(y)] + 2\mathbb{E} [T_{3,1,2}^m(y)],$$

where

$$\begin{aligned} (27) \quad T_{3,1,1}^m(y) &= \sum_{j=1}^{D_{m_1}} \left\{ \frac{1}{n} \sum_{i=1}^n h(U_i, y) \varphi_j(U_i) - \mathbb{E} [h(U_i, y) \varphi_j(U_i)] \right\}^2, \\ T_{3,1,2}^m(y) &= \sum_{j=1}^{D_{m_1}} \left\{ \int_0^1 \partial_1 h(u, y) (\hat{U}_n(u) - u) \varphi_j(u) du \right\}^2. \end{aligned}$$

In the spirit of the bound given for T_1^m , the first term is controlled as follows:

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

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

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

Finally, $T_{3,1,2}^m(y) \leq C \|\partial_1 h(\cdot, y)\|_{[0,1]}^2/n$ by Inequality (24), and thus, by gathering the bounds for $T_{3,1,1}^m(y)$ and $T_{3,1,2}^m(y)$,

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

As regards $T_{3,2}^m(y)$, 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

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

by proceeding with the previous arguments (properties of orthogonal projection and Inequality (24)). So we prove $\int_{A_2} \mathbb{E} [T_{3,2}^m(y)] dy \leq C D_{m_1}^4/n^2$. The computations for the last term are less technical:

$$\mathbb{E} [T_{3,3}^m(y)] \leq (1/36) \sum_{j=1}^{D_{m_1}} \left\| \varphi_j^{(3)} \right\|_{\infty, [0,1]}^2 \|h(\cdot, y)\|_{[0,1]}^2 \mathbb{E} \left[\left\| \hat{U}_n - id \right\|_{\infty, [0,1]}^6 \right],$$

thus $\int_{A_2} \mathbb{E} [T_{3,3}^m(y)] dy \leq C D_{m_1}^7/n^3$. This completes the proof of the first inequality of Lemma 9. With regard to the second inequality, it is enough to bound $\mathbb{E} [\max_{m' \in \mathcal{M}_n} (T_1^p - V_1(p))_+]$, like for the second part of Lemma 7 ($p = m'$ or $p = m \wedge m'$). As previously, we get the splitting

$$(28) \quad T_3^p \leq 6 \int_{A_2} T_{3,1,1}^p(y) dy + 6 \int_{A_2} T_{3,1,2}^p(y) dy + 3 \int_{A_2} T_{3,2}^p(y) dy + 3 \int_{A_2} T_{3,3}^p(y) dy,$$

and

$$\begin{aligned} \mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(T_3^{p,b} - V_3(p) \right)_+ \right] &\leq \mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(6 \int_{A_2} T_{3,1,1}^p(y) dy - V_3(p)/3 \right)_+ \right] \\ &\quad + \mathbb{E} \left[\max_{m' \in \mathcal{M}_n} 6 \int_{A_2} T_{3,1,2}^p(y) dy \right] \\ &\quad + \mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(3 \int_{A_2} T_{3,2}^p(y) dy - V_3(p)/3 \right)_+ \right] \\ &\quad + \mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(3 \int_{A_2} T_{3,3}^p(y) dy - V_3(p)/3 \right)_+ \right]. \end{aligned}$$

The term which is not centred is directly 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 $T_{3,1,2}^p \leq T_{3,1,2}^{m_{\max}}$ (by (27)). Hence, $\mathbb{E}[\max_{m' \in \mathcal{M}_n} 6 \int_{A_2} T_{3,1,2}^p(y) dy] \leq C/n$. Let us briefly study each of the other terms: first $T_{3,1,1}^p(y) = \sup_{s \in S_{p_1}, \|s\|_{[0,1]}=1} \nu_{n,y}^2(s)$ (Proposition 4), with

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

Using once more time Talagrand's Inequality (15) leads to

$$(29) \quad \mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(6 \int_{A_2} T_{3,1,1}^p(y) dy - V_{3,1,1}(p) \right)_+ \right] \leq \frac{C}{n},$$

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

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

since $D_{p_2} \geq c \ln(n)$ ($c > 0$). Inequality (29) holds with $V_{3,1,1}^b$. The last two terms, involving $T_{3,2}^m(y)$ and $T_{3,3}^m(y)$ can be computed with the same strategy: use the proof of the first inequality of Lemma 9 to bound $\int_{A_2} T_{3,l}^m(y) dy$ ($l = 2, 3$) by quantity of the form $C \|\hat{U}_n - id\|_{\infty}^k$, and then apply Inequality (25). The conclusion is that

$$(30) \quad \mathbb{E} \left[\max_{m' \in \mathcal{M}_n} \left(3 \int_{A_2} T_{3,l}^p(y) dy - V_{3,l}(p) \right)_+ \right] \leq C \frac{\ln(n)}{n},$$

for $l = 2, 3$, with $V_{3,2}(p) = CD_{p_1}^4 \ln^2(n)/n^2$, and $V_{3,3}(p) = CD_{p_1}^7 \ln^3(n)/n^3$. 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

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

With the more restrictive low bound $D_{p_2} \geq c \ln^4(n)$, we also get $V_{3,3}(p) \leq CD_{p_1} D_{p_2}/n := V_{3,3}^b(p)$. As usual, Inequalities (30) still hold with $V_{3,l}^b$ 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 $V_3 \geq 3V_{3,1,1}^b$, $V_3 \geq 3V_{3,2}^b$, et $V_3 \geq 3V_{3,3}^b$. \square

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

$$T_4^{m'} = \int_{A_1} \mathbb{E} \left[\sum_{k=1}^{D_{m'_2}} \left(\sum_{j=1}^{D_{m'_1}} \hat{a}_{j,k}^{\hat{F}} \left(\varphi_j \circ F_X(x) - \varphi_j \circ \hat{F}_n(x) \right) \right)^2 |(X_{-l})_l| \right] f_X(x) dx.$$

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 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

$$\mathbb{E} [T_{4,2}^{m'}] = \mathbb{E} \left[\sum_{k=1}^{D_{m'_2}} \sum_{j,j'=1}^{D_{m'_1}} a_{j,k} a_{j',k} \int_0^1 \left(\varphi_j(u) - \varphi_j \circ \hat{U}_n(u) \right) \left(\varphi_{j'}(u) - \varphi_{j'} \circ \hat{U}_n(u) \right) du \right],$$

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

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

The methods use to bound each of these terms have already been detailed for other terms: with regard to the two quantities which are not centred, 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 finally apply Inequality (25), 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

$$(31) \quad T_{4,1,1}^{m'} = \sum_{k=1}^{D_{m'_2}} \sum_{j=1}^{D_{m'_1}} \text{Var} \left(\hat{a}_{j,k}^{\hat{F}} | (X_{-l})_l \right) \int_{[0;1]} \sum_{j=1}^{D_{m'_1}} \left(\varphi_j(u) - \varphi_j \circ \hat{U}_n(u) \right)^2 du.$$

The conditional variance is

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

By Property (3) applied to the sum over j, k of the last quantity, $\sum_{j,k} \text{Var}(\hat{a}_{j,k}^{\hat{F}} | (X_{-l})_l) \leq D_{m'_1} D_{m'_2} / n$. Besides, we use the mean value theorem to bound the integral of (31) so that

$$T_{4,1,1}^{m'} \leq \frac{D_{m'_1} D_{m'_2}}{n} \times D_{m'_1}^3 \|\varphi'_2\|_{\infty,[0;1]}^2 \left\| \hat{U}_n - id \right\|_{\infty,[0;1]}^2,$$

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,

$$T_{4,1,2}^{m'} = T_3^{m'} \int_{[0;1]} \sum_{j=1}^{D_{m'_1}} \left(\varphi_j(u) - \varphi_j \circ \hat{U}_n(u) \right)^2 du,$$

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 (28), 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 $l = 1, \dots, 4$. Let us consider now the terms $T_{4,2,1}^{m'}$ and $T_{4,2,2}^{m'}$ which do not require to be centred. It is usefull to remark that the Fourier 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 $T_{4,2,1}^{m'}$ 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

$$T_{4,2,1}^{m'} \leq \left\| \hat{U}_n - id \right\|_{\infty,[0;1]}^2 \sum_{k=1}^{D_{m'_2}} \|\xi'_k\|_{[0;1]}^2.$$

Let us compute then the derivative of ξ_k to bound roughly

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

We thus have $\mathbb{E}[\max_{m' \in \mathcal{M}_n} T_{4,2,1}^{m'}] \leq \|\partial_1 h\|^2 \mathbb{E}[\|\hat{U}_n - id\|_{\infty, [0;1]}^2] \leq C/n$ with Inequality (24). Recall now that

$$T_{4,2,2}^{m'} = (1/4) \sum_{k=1}^{D_{m'_2}} \int_0^1 (u - \hat{U}_n(u))^4 \left(\sum_{j=1}^{D_{m'_1}} a_{j,k} \varphi_j''(\hat{\alpha}_{j,n,u}) \right)^2 du.$$

We introduce $\mu_j = j$ for even j and $\mu_j = j-1$ for odd j . Since h belongs to $W_{per}^2([0;1]^2, L, (1,0))$ and according to (11)

$$\begin{aligned} \sum_{k=1}^{D_{m'_2}} \left(\sum_{j=1}^{D_{m'_1}} a_{j,k} \varphi_j''(\hat{\alpha}_{j,n,u}) \right)^2 &\leq \|\varphi_2''\|_{\infty, [0;1]}^2 \sum_{k=1}^{D_{m'_2}} \left(\sum_{j=1}^{D_{m'_1}} |a_{j,k}| \mu_j^2 \right)^2, \\ &\leq \|\varphi_2''\|_{\infty, [0;1]}^2 \sum_{k=1}^{D_{m'_2}} \sum_{j=1}^{D_{m'_1}} a_{j,k}^2 \mu_j^2 \sum_{j=1}^{D_{m'_1}} \mu_j^2, \\ &\leq \|\varphi_2''\|_{\infty, [0;1]}^2 \frac{L^2}{\pi^2} D_{m'_1}^3 \leq C D_{m_{1,\max}}^3. \end{aligned}$$

Hence, $\mathbb{E}[\max_{m' \in \mathcal{M}_n} T_{4,2,2}^{m'}] \leq \mathbb{E}[\|\hat{U}_n - id\|_{\infty, [0;1]}^4] C D_{m_{1,\max}}^3 \leq C D_{m_{1,\max}}^3 / n^2 \leq C/n$ 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

$$T_{4,2,3}^{m'} = \int_{[0;1]} (u - \hat{U}_n)^3 \sum_{k=1}^{D_{m'_2}} \left(\sum_{j=1}^{D_{m'_1}} a_{j,k} \varphi_j''(\hat{\alpha}_{j,n,u}) \right) \left(\sum_{j=1}^{D_{m'_1}} a_{j,k} \varphi_j'(u) \right).$$

and compute like the term $T_{4,2,2}^{m'}$:

$$\sum_{k=1}^{D_{m'_2}} \left(\sum_{j=1}^{D_{m'_1}} a_{j,k} \varphi_j''(\hat{\alpha}_{j,n,u}) \right)^2 \leq \|\varphi_2''\|_{\infty, [0;1]}^2 \frac{L^2}{\pi^2} D_{m'_1}^3, \quad \sum_{k=1}^{D_{m'_2}} \left(\sum_{j=1}^{D_{m'_1}} a_{j,k} \varphi_j'(u) \right)^2 \leq \|\varphi_2'\|_{\infty, [0;1]}^2 \frac{L^2}{\pi^2} D_{m'_1}.$$

This leads to

$$T_{4,2,3}^{m'} \leq \|\varphi_2'\|_{\infty, [0;1]} \|\varphi_2''\|_{\infty, [0;1]} \frac{L^2}{\pi^2} D_{m'_1} \left\| \hat{U}_n - id \right\|_{\infty, [0;1]}^3,$$

and we apply tools already used to complete the proof. □

ACKNOWLEDGEMENTS

I would like to thank Fabienne Comte for her decisive suggestions throughout this work, and the referees for their helpful remarks.

REFERENCES

- Robert A. Adams. *Sobolev spaces*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Pure and Applied Mathematics, Vol. 65.
 Nathalie Akakpo and Claire Lacour. Inhomogeneous and anisotropic conditional density estimation from dependent data. *Electronic Journal of Statistics*, 5:1618–1653, 2011.

- Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6: 127–146 (electronic), 2002.
- Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.
- David M. Bashtannyk and Rob J. Hyndman. Bandwidth selection for kernel conditional density estimation. *Comput. Statist. Data Anal.*, 36(3):279–298, 2001.
- Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2011.
- Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.
- Elodie Brunel and Fabienne Comte. Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā*, 67(3):441–475, 2005.
- Elodie Brunel, Fabienne Comte, and Claire Lacour. Adaptive estimation of the conditional density in the presence of censoring. *Sankhyā*, 69(4):734–763, 2007.
- Raymond J. Carroll, David Ruppert, and A. H. Welsh. Fitting heteroscedastic regression models. *J. Amer. Statist. Assoc.*, 89(425):100–116, 1994.
- Serge Cohen and Erwan Le Pennec. Conditional density estimation by penalized likelihood model selection and applications. *Submitted arXiv:1103.2021*, 2011.
- Jan G. De Gooijer and Dawit Zerom. On conditional density estimation. *Statist. Neerlandica*, 57(2):159–176, 2003.
- 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.
- 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.
- Sam Efromovich. *Nonparametric curve estimation*. Springer Series in Statistics. Springer-Verlag, New York, 1999. Methods, theory, and applications.
- Sam Efromovich. Conditional density estimation in a regression setting. *Ann. Statist.*, 35(6): 2504–2535, 2007.
- Sam Efromovich. Oracle inequality for conditional density estimation and an actuarial example. *Ann. Inst. Statist. Math.*, 62(2):249–275, 2010a.
- Sam Efromovich. Dimension reduction and adaptation in conditional density estimation. *J. Amer. Statist. Assoc.*, 105(490):761–774, 2010b.
- Jian-qing Fan, Liang Peng, Qi-wei Yao, and Wen-yang Zhang. Approximating conditional density functions using dimension reduction. *Acta Math. Appl. Sin. Engl. Ser.*, 25(3):445–456, 2009.
- Jianqing Fan and Irène Gijbels. Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B*, 57(2):371–394, 1995.
- Jianqing Fan and Tsz Ho Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834, 2004.
- Jianqing Fan, Howell Tong, and Qiwei Yao. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206, 1996.
- Olivier P. Faugeras. A quantile-copula approach to conditional density estimation. *J. Multivariate Anal.*, 100(9):2083–2099, 2009.
- Alexander Goldenshluger and Oleg Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632, 2011.
- 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.

- Wolfgang Härdle, Gérard Kerkycharian, Dominique Picard, and Alexander Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1998.
- Reinhard Hochmuth. Wavelet characterizations for anisotropic Besov spaces. *Appl. Comput. Harmon. Anal.*, 12(2):179–208, 2002.
- Rob J. Hyndman and Qiwei Yao. Nonparametric estimation and symmetry tests for conditional density functions. *J. Nonparametr. Stat.*, 14(3):259–278, 2002.
- Rob J. Hyndman, David M. Bashtannyk, and Gary K. Grunwald. Estimating and visualizing conditional densities. *J. Comput. Graph. Statist.*, 5(4):315–336, 1996.
- Gérard Kerkycharian and Dominique Picard. Regression in random design and warped wavelets. *Bernoulli*, 10(6):1053–1105, 2004.
- Thierry Klein and Emmanuel Rio. Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077, 2005.
- Claire Lacour. Adaptive estimation of the transition density of a Markov chain. *Ann. Inst. H. Poincaré Probab. Statist.*, 43(5):571–597, 2007.
- Claire Lacour. Adaptive estimation of the transition density of a particular hidden Markov chain. *J. Multivariate Anal.*, 99(5):787–814, 2008.
- K. L. Mehra, Y. S. Ramakrishnaiah, and P. Sashikala. Laws of iterated logarithm and related asymptotics for estimators of conditional density and mode. *Ann. Inst. Statist. Math.*, 52(4):630–645, 2000.
- 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.
- Marianna Penskaya. Mean square consistent estimation of a ratio. *Scand. J. Statist.*, 22(1):129–137, 1995.
- P. Sarda, P. Vieu, and É. Youndjé. Validation croisée pour l'estimation non-paramétrique de la densité conditionnelle. *Publ. Inst. Statist. Univ. Paris*, 38(1):57–80, 1994.
- Winfried Stute. Conditional empirical processes. *Ann. Statist.*, 14(2):638–647, 1986a.
- Winfried Stute. On almost sure convergence of conditional empirical distribution functions. *Ann. Probab.*, 14(3):891–901, 1986b.
- 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.
- Alexandre B. Tsybakov. *Introduction à l'estimation non-paramétrique*, volume 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2004.