Inhomogeneous and Anisotropic Conditional Density Estimation from Dependent Data

Nathalie Akakpo, Claire Lacour

Electronic Journal of Statistics
ISSN: 1935-7524

Inhomogeneous and Anisotropic Conditional Density Estimation from Dependent Data

Nathalie Akakpo

Laboratoire de Probabilités et Modèles Aléatoires
Université Pierre et Marie Curie, Case 188
4, Place Jussieu
75252 Paris Cedex 05
e-mail: nathalie.akakpo@upmc.fr

Claire Lacour

Laboratoire de Mathématiques d'Orsay
Faculté des sciences, bâtiment 425
Université Paris-Sud 11
91405 Orsay Cedex
e-mail: claire.lacour@u-psud.fr

Abstract: The problem of estimating a conditional density is considered. Given a collection of partitions, we propose a procedure that selects from the data the best partition among that collection and then provides the best piecewise polynomial estimator built on that partition. The observations are not supposed to be independent but only β-mixing; in particular, our study includes the estimation of the transition density of a Markov chain. For a well-chosen collection of possibly irregular partitions, we obtain oracle-type inequalities and adaptivity results in the minimax sense over a wide range of possibly anisotropic and inhomogeneous Besov classes. We end with a short simulation study.

AMS 2000 subject classifications: 62G05, 62H12, 62M05, 62M09.
Keywords and phrases: Conditional density, Model selection, Anisotropy, Dependent Data, Adaptive estimation.

Contents

1. Introduction

In this paper, we are concerned with conditional density estimation. Such a model brings more information than the well-studied regression model; for instance, it may reveal multimodality. Yet, references about conditional density estimation are rather scarce, even for nonadaptive procedures. For independent data, we can cite for instance Györfi and Kohler [GK07] for a histogram based procedure, or Faugeras [Fau07] for a copula-based kernel estimator. For mixing data, De Gooijer and Zerom [DGZ03] and Fan and Yim [FY04] propose kernel methods. For Markov chains, nonadaptive estimation of the transition density is considered for instance in [Rou69, Bir83, DG83], and we also refer to [Lac07] for a more complete bibliography. But, in order to reach the optimal rate of convergence, those methods require the smoothness of the function to estimate to be known, so as to choose adequately some tuning parameter.

Adaptive estimators of the conditional density have only recently been proposed. For independent data, Efromovich [Efr07, Efr08] and Brunel, Comte and Lacour [BCL07] give oracle inequalities and adaptivity results in the minimax sense. Efromovich [Efr07, Efr08] uses a Fourier decomposition to build a blockwise-shrinkage Efromovich-Pinsker estimator, whereas Brunel *et al.* [BCL07] perform model selection based on a penalized least-squares criterion. Regarding dependent data, Clémençon [Clé00b] and Lacour [Lac07] study adaptive estimators of the conditional density for Markovian observations, the former via wavelet thresholding, and the latter via model selection. Besides, the procedures proposed by [Efr07, Efr08, BCL07, Lac07] are all able to adapt to anisotropy; otherwise said, the conditional density to estimate is allowed to have unknown and different degrees of smoothness in each direction.

But the smoothness of the function to estimate may also vary spatially. If the risk of the estimator is measured via some $\mathbb{L}_q$-norm, one way to take into account that inhomogeneous behaviour is to consider functions whose smoothness is measured in a $\mathbb{L}_p$-norm, with $p < q$. Among the aforementioned references, only Clémençon [Clé00b] is able to cope with inhomogeneous smoothness. In the

simpler framework of density estimation, without conditioning variables, adaptation to inhomogeneity has been studied in the following works. Thresholding methods, in a univariate framework, are proposed by Hall, Kerkyacharian and Picard [HKP98] for independent data, Clémençon [Clé00a] for Markovian data, and Gannaz and Wintenberger [GW10] for a wide class of weakly dependent data. Piecewise polynomial selection procedures based on a penalized contrast have also been considered, and consist in selecting from the data a best partition and a best piecewise polynomial built on that partition. Thus, Comte and Merlevède [CM02] estimate the univariate density of absolutely regular stationary processes, in discrete or continuous time, selecting a best partition among the collection of all the partitions of $[0,1]$ built on a thin regular grid via a least-squares criterion. Besides, three papers have lately considered density estimators inspired from the "multiresolution histogram" of Engel [Eng94, Eng97], or the "dyadic CART procedure" of Donoho [Don97]. Willett and Nowak [WN07] select best piecewise polynomials built on partitions into dyadic cubes via a penalized maximum likelihood contrast. Klemelä [Kle09] and Blanchard, Schäfer, Rozenholc and Müller [BSRM07] select best histograms based on partitions into dyadic rectangles via a penalized criterion based on the $\mathbb{L}_2$-distance for the first one, and on Kullback-Leibler divergence for the second ones. But all these procedures only reach optimal rates of convergence up to a logarithmic factor, and only [Kle09] is able to prove adaptivity both to anisotropy and inhomogeneity.

In this paper, we provide an estimator of the conditional density via a piecewise polynomial selection procedure based on an adequate least-squares criterion. To deal with the possible dependence of the observations, we mainly use β-mixing coefficients and their coupling properties. Thus, our dependence assumptions, while being satisfied by a wide class of Markov chains, are not restricted to Markovian assumptions. We first prove nonasymptotic oracle type inequalities fulfilled by any collection of partitions satisfying some mild structural conditions. We then consider the collection of partitions into dyadic rectangles, as [Kle09] or [BSRM07]. We obtain oracle-type inequalities and adaptivity results in the minimax sense, without logarithmic factor, over a wide range of Besov smoothness classes that may contain functions with inhomogeneous and anisotropic smoothness, whether the data are independent or satisfy suitable dependence assumptions. The adaptivity of our procedure greatly relies on the approximation result proved in [Aka10]. Moreover, determining in practice the penalized estimator based on that collection only requires a computational complexity linear in the size of the sample.

This paper is organized as follows. We begin by describing the framework and the estimation procedure, and we present an evaluation of the risk on one model. This study allows to understand what bound for the $\mathbb{L}_2$-risk we seek to obtain. The choice of a penalty yielding an oracle-type inequality is the topic of Section 3.1. Section 3.2 is devoted to the collection of partitions into dyadic rectangles, and adaptivity results are proved for an adequate penalty. We show in Section 4 that all these results can be extended to dependent data. In Section 5, the practical implementation of our estimator is explained and some simulations are presented, both for independent and dependent data. Most

proofs are deferred to Section 6.

2. Framework and estimation procedure

In this section we define a contrast and we deduce a collection of estimators $\hat{s}_m$. In order to understand which model m we should choose, we give an evaluation of the risk for each estimator $\hat{s}_m$. This allows us to define the penalized estimator.

2.1. Framework and notation

Let $\{Z_i\}_{i\in\mathbb{Z}} = \{(X_i, Y_i)\}_{i\in\mathbb{Z}}$ be a strictly stationary process, where, for all $i \in \mathbb{Z}$, X_i and Y_i take values respectively in $[0,1]^{d_1}$ and $[0,1]^{d_2}$, with d_1 and d_2 positive integers. We assume that the variables $(X_i)_{i\in\mathbb{Z}}$ admit a bounded marginal density f with respect to the Lebesgue measure. Given some integer $n \geq 2$, our aim is to estimate, on the basis of the observation of $(Z_1, \ldots, Z_n)$, the marginal density s of Y_i conditionally to X_i. Thus, our parameter of interest s is the real-valued function of d variables, where $d = d_1 + d_2$, such that, for all $x \in [0,1]^{d_1}$, $s(x,.) : [0,1]^{d_2} \to \mathbb{R}$ is the density of Y_i conditionally to $X_i = x$. In particular, if $(X_i)_{i\in\mathbb{Z}}$ is a homogeneous Markov chain of order 1, and $Y_i = X_{i+1}$ for all $i \in \mathbb{Z}$, then s is the transition density of the chain $(X_i)_{i\in\mathbb{Z}}$.

Let us introduce some standard notation. For any real-valued function t defined and bounded on some set $\mathcal{D}$, we set

$$\iota(t) = \inf_{x\in\mathcal{D}} |t(x)| \quad \text{and} \quad \|t\|_\infty = \sup_{x\in\mathcal{D}} |t(x)|.$$

We denote by $\mathbb{L}_2\left([0,1]^{d_1} \times [0,1]^{d_2}\right)$ the set of all real-valued functions which are square integrable with respect to the Lebesgue measure. Since f is bounded, we can also define on $\mathbb{L}_2\left([0,1]^{d_1} \times [0,1]^{d_2}\right)$ the semi-scalar product

$$\langle t, u\rangle_f = \int_{[0,1]^{d_1}\times[0,1]^{d_2}} t(x,y)u(x,y)f(x)\mathrm{d}x\mathrm{d}y$$

and the associated semi-norm $\|.\|_f$.

2.2. Contrast and estimator on one model

In order to estimate the conditional density s, we consider the empirical criterion γ described in [BCL07] and defined on $\mathbb{L}_2\left([0,1]^{d_1} \times [0,1]^{d_2}\right)$ by

$$\gamma(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\int_{[0,1]^{d_2}} t^2(X_i, y)\mathrm{d}y - 2t(X_i, Y_i)\right].$$

Due to the nature of the function to estimate, the contrast used here borrows both from the classical regression and density least-squares contrasts. This contrast verifies:

$$\begin{aligned}\mathbb{E}_s[\gamma(t)-\gamma(s)] &= \mathbb{E}_s\left[\int (t^2-s^2)(X_1,y)\mathrm{d}y - 2(t-s)(X_1,Y_1)\right]\\ &= \iint (t^2-s^2)(x,y)f(x)\mathrm{d}x\mathrm{d}y - 2\iint (t-s)(x,y)s(x,y)f(x)\mathrm{d}x\mathrm{d}y\\ &= \iint (t^2-2ts+s^2)(x,y)f(x)\mathrm{d}x\mathrm{d}y = \|s-t\|_f^2,\end{aligned}$$

so that s minimizes $t \mapsto \mathbb{E}_s[\gamma(t)]$ over $\mathbb{L}_2([0,1]^{d_1}\times[0,1]^{d_2})$. Thus, a natural way to build an estimator of s consists in minimizing γ over some subset of $\mathbb{L}_2([0,1]^{d_1}\times[0,1]^{d_2})$, that we choose here as a space of piecewise polynomial functions with degree smaller than a given nonnegative integer r. More precisely, for a partition m of $[0,1]^{d_1}\times[0,1]^{d_2}$ into rectangles, we denote by S_m the space of all real-valued piecewise polynomial functions on $[0,1]^{d_1}\times[0,1]^{d_2}$ which are polynomial with coordinate degree $\leq r$ on each rectangle of m. We define a best estimator of s with values in the model S_m by setting

$$\hat{s}_m = \operatorname*{argmin}_{t\in S_m} \gamma(t).$$

An explicit formula for computing $\hat{s}_m$ is given in Section 5.

2.3. Risk on one model

In this subsection, we fix some partition m of $[0,1]^{d_1}\times[0,1]^{d_2}$ into rectangles and give some upper-bound for the risk of $\hat{s}_m$ when $Z_1,\dots,Z_n$ are independent. As for all the theorems stated in the sequel, we evaluate that risk in the random semi-norm $\|.\|_n$ naturally associated to our problem, and defined, for all $t \in \mathbb{L}_2([0,1]^{d_1}\times[0,1]^{d_2})$, by

$$\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}\int_{[0,1]^{d_2}} t^2(X_i,y)\mathrm{d}y$$

(remember that our problem is a mixture of regression in the x-direction and of density estimation in the y-direction). However, it is also possible to control the classical $\mathbb{L}_2$-norm, using a truncated estimator (see, for instance, Corollary 3.2 in Section 3.) Besides, for any partition m' of a unit cube into rectangles, we denote by $|m'|$ the number of rectangles in m' and say that the partition m' is regular if all its rectangles have the same dimensions. For the risk of the estimator $\hat{s}_m$, we can prove the following result.

Proposition 2.1. *Let m be a partition of $[0,1]^{d_1}\times[0,1]^{d_2}$ built on a regular partition $m_1^\star \times m_2^\star$, where $m_1^\star$ and $m_2^\star$ are regular partitions of $[0,1]^{d_1}$ and $[0,1]^{d_2}$ into cubes such that*

$$|m_1^\star| \leq \frac{n}{\log^2(n)} \quad \text{and} \quad |m_2^\star| \leq n.$$

Let s_m be the orthogonal projection of s on S_m for the norm $\|.\|$, and D_m denote the dimension of S_m, so that $D_m = (r+1)^d|m|$. Assume that s and f are bounded, and that f is also bounded from below by a positive constant. If the variables $Z_1, \dots, Z_n$ are independent, then

$$\mathbb{E}_s\left[\|s - \hat{s}_m\|_n^2\right] \leq 2\|s - s_m\|_f^2 + 11\|s\|_\infty \frac{D_m}{n} + \frac{C}{n},$$

where C only depends on $r, d, \iota(f), \|f\|_\infty, \|s\|_\infty$.

We recover approximately in the upper-bound stated in Proposition 2.1 the usual decomposition into a squared bias term, of order $\|s-s_m\|_f^2$, and a variance term of order $\|s\|_\infty D_m/n$, proportional to the dimension of the model S_m. A major interest of such a bound is that it allows to understand how to build an optimal estimator from the minimax point of view. Let us first recall that when s belongs to classical classes of functions with isotropic smoothness σ (isotropic Besov classes for instance), a minimax estimator over such a class reaches the estimation rate $n^{-2\sigma/(2\sigma+d)}$. Roughly speaking, when s belongs to a well-chosen class of isotropic functions with smoothness σ measured in a $\mathbb{L}_p$-norm with $p \geq 2$, the bias term $\|s - s_m\|_f^2$ is at most of order $D_m^{-2\sigma/d}$ for any regular partition m into cubes. If we knew at least the smoothness parameter σ, we could choose some regular partition $m_{opt}(\sigma)$ into cubes realizing a good compromise between the bias and the variance terms, *i.e.* such that $D_{m_{opt}(\sigma)}^{-2\sigma/d}$ and $D_{m_{opt}(\sigma)}/n$ are of the same order. We would then obtain with $\hat{s}_{m_{opt}(\sigma)}$ an estimator that reaches the optimal estimation rate $n^{-2\sigma/(2\sigma+d)}$ whatever $p \geq 2$. But when s has isotropic smoothness σ measured in a $\mathbb{L}_p$-norm with $p < 2$, one can only ensure that the bias term $\|s - s_m\|_f^2$ is at most of order $D_m^{-2\sigma/d}$ for some irregular partition m into cubes that does not only depend on σ, but must be adapted to the inhomogeneity of s over the unit cube (see for instance Section 3.2 and [Aka10]). Thus, there exists some well-chosen irregular partition $m_{opt}(s)$ into cubes such that $D_{m_{opt}(s)}$ is of order $n^{d/(2\sigma+d)}$ but that reaches the estimation rate $n^{-2\sigma/(2\sigma+d)}$ only at s, and probably not on the whole class of functions with smoothness σ in a $\mathbb{L}_p$-norm with $p < 2$. Last, if s has anisotropic smoothness, similar properties still hold, with partitions into rectangles - regular or not depending on the homogeneity of s- whose dimensions are adapted to the anisotropy of s.

2.4. Penalized estimator

We give ourselves a finite collection $\mathcal{M}$ of partitions of $[0,1]^{d_1} \times [0,1]^{d_2}$ into rectangles. The aim is to choose the best estimator among the collection $\{\hat{s}_m\}_{m\in\mathcal{M}}$ without assumption on the smoothness of s. To do so, we use the model selection method introduced by [BBM99] which allows us to select an estimator only from the data, by minimizing a penalized criterion. Thus, we consider the random selection procedure

$$\hat{m} = \underset{m\in\mathcal{M}}{\operatorname{argmin}} \{\gamma(\hat{s}_m) + \operatorname{pen}(m)\}$$

and the penalized estimator

$$\tilde{s} = \hat{s}_{\hat{m}},$$

where pen : $\mathcal{M} \to \mathbb{R}^+$ is a so-called penalty function that remains to be chosen so that $\tilde{s}$ performs well. The choice of the collection of partitions $\mathcal{M}$ is discussed is the next section. The practical implementation of the penalized estimator based on the collection of partitions into dyadic rectangles is described in Section 5.

3. Main result

In this section, we study the risk of the penalized estimator $\tilde{s}$ for independent data, first with a general collection of partitions, secondly with a relevant choice of collection that ensures the optimal estimation of a possibly inhomogeneous and anisotropic function s.

3.1. Oracle inequality

Ideally, we would like to choose a penalty pen such that $\tilde{s}$ is almost as good as the best estimator in the collection $\{\hat{s}_m\}_{m\in\mathcal{M}}$, in the sense that

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \leq C \min_{m\in\mathcal{M}} \mathbb{E}_s\left[\|s-\hat{s}_m\|_n^2\right] \tag{3.1}$$

for some positive constant C. Theorem 3.1 below suggests a form of penalty yielding an inequality akin to

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \leq C \min_{m\in\mathcal{M}} \left\{\|s-s_m\|_f^2 + \frac{D_m}{n}\right\}. \tag{3.2}$$

Yet, as recalled in the previous section, for each $m \in \mathcal{M}$, $\mathbb{E}_s\left[\|s-\hat{s}_m\|_n^2\right]$ is expected to be of order $\|s-s_m\|_f^2 + D_m/n$. So, Inequality (3.2) is expected to be almost as good as Inequality (3.1). In order to deal with a large collection $\mathcal{M}$ that may contain irregular partitions, we only impose a minor structural condition on $\mathcal{M}$. That assumption ensures that all the models are included in a biggest model, without imposing that the models be nested as in [BCL07]. We also assume that s and f are bounded.

Assumption (P1) *All the partitions in the collection $\mathcal{M}$ are built on a regular partition $m^\star$ of $[0,1]^d$ into cubes such that*

$$|m^\star|^2 \leq n$$

Assumption (B)

$$s \leq \|s\|_\infty < \infty, \qquad 0 < \iota(f) \leq f \leq \|f\|_\infty < \infty$$

We establish an oracle type inequality for a very general collection of partitions. Thus we state the following model selection theorem.

Theorem 3.1. *Let $\mathcal{M}$ be a collection of partitions satisfying Assumption **(P1)** and $\{L_m\}_{m\in\mathcal{M}}$ be a family of reals greater than or equal to 1, that may depend on n, such that*

$$\sum_{m\in\mathcal{M}} \exp(-L_m|m|) \leq 1. \tag{3.3}$$

*Assume that $(Z_i)_{1\leq i\leq n}$ are independent and s, f satisfy Assumption **(B)**. If the penalty satisfies, for all $m \in \mathcal{M}$,*

$$\mathrm{pen}(m) = \kappa \left(\|s\|_\infty + \frac{(2r+1)^d}{\iota(f)} \right) \frac{L_m^2 D_m}{n}$$

for some large enough positive absolute constant κ, then

$$\mathbb{E}_s \left[\|s - \tilde{s}\|_n^2 \right] \leq C_1 \left(\max_{m\in\mathcal{M}} L_m^2 \right) \min_{m\in\mathcal{M}} \left\{ \|s - s_m\|_f^2 + \frac{D_m}{n} \right\}. \tag{3.4}$$

where C_1 is a positive constant that depends on κ, r, d_1, d_2, $\|s\|_\infty$, $\iota(f)$, $\|f\|_\infty$.

Theorem 3.1 is only proved in its general version for dependent data (see Theorem 4.1 in Section 4.3).

The penalty contains unknown terms, but in practice, $\|s\|_\infty$ and $\iota(f)$ can be replaced with an estimator, as in [BM97] (Proposition 4) for instance, and κ is calibrated via a simulation study. To state a result with the precise replacement, we choose $m_1^\bullet$ and $m_2^\bullet$ regular partitions of $[0,1]^{d_1}$ and $[0,1]^{d_2}$ into cubes such that $m^\bullet = m_1^\bullet \times m_2^\bullet$ verifies Assumption **(P1)**. We define $\hat{f}_{m_1^\bullet} = \arg\min_{t\in F_{m_1^\bullet}} n^{-1} \sum_{i=1}^n [\|t\|^2 - 2t(X_i)]$, where $F_{m_1^\bullet}$ is the space of all functions on $[0,1]^{d_1}$ which are polynomial with coordinate degree $\leq r$ on each rectangle of $m_1^\bullet$, and estimate $\iota(f)$ by $\hat{\iota}(f) = \inf_{x\in[0,1]^{d_1}} \hat{f}_{m_1^\bullet}(x)$. We also impose Besov-type smoothness assumptions on f and s. For $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d) \in (0, r+1)^d$, $R > 0$, $p > 0$, we refer to [Tri06] (Chapter 5) for a definition of the anisotropic Besov space $B_{pp'}^{\boldsymbol{\sigma}}$ and the associated norm $\|.|B_{pp'}^{\boldsymbol{\sigma}}\|$, and we introduce the anisotropic Besov balls

$$\mathcal{B}(\boldsymbol{\sigma}, p, R) = \{t : [0,1]^d \to \mathbb{R} \text{ s.t. } \|t|B_{pp'}^{\boldsymbol{\sigma}}\| \leq R\}, \tag{3.5}$$

where $p' = \infty$ if $0 < p \leq 1$ or $p \geq 2$, and $p' = p$ if $1 < p < 2$. We recall that, due to the continuous embeddings stated for instance in [Tri06], $B_{p\infty}^{\boldsymbol{\sigma}}$ contains all the spaces $B_{pp'}^{\boldsymbol{\sigma}}$, for $p' > 0$, so our choice of p' in the definition of $\mathcal{B}(\boldsymbol{\sigma}, p, R)$ is the less stringent one for $0 < p \leq 1$ or $p \geq 2$. Last, we set $\underline{\boldsymbol{\sigma}} = \min_{1\leq l\leq d} \sigma_l$ and denote by $H(\boldsymbol{\sigma})$ the harmonic mean of $\sigma_1, \ldots, \sigma_d$, *i.e.*

$$\frac{1}{H(\boldsymbol{\sigma})} = \frac{1}{d} \sum_{l=1}^d \frac{1}{\sigma_l}.$$

Corollary 3.1. *Assume that* $s \in \mathcal{B}(\boldsymbol{\sigma}, p, R)$ *and* $f \in \mathcal{B}(\boldsymbol{\alpha}, p, R_1)$ *with*

$$\frac{H(\boldsymbol{\sigma})}{d} > \frac{1}{p} + \frac{1}{2}\frac{H(\boldsymbol{\sigma})}{\underline{\boldsymbol{\sigma}}}, \qquad \frac{H(\boldsymbol{\alpha})}{d_1} > \left(\frac{1}{p} - \frac{1}{2}\right)_+ + \frac{H(\boldsymbol{\alpha})}{\underline{\boldsymbol{\alpha}}}.$$

Assume that $|m_1^\bullet| \geq \ln n$ *and, for all* $m \in \mathcal{M}$,

$$\mathrm{pen}(m) = \bar{\kappa}\left(\|\hat{s}_{m^\bullet}\|_\infty + \frac{(2r+1)^d}{\hat{\iota}(f)}\right)\frac{L_m^2 D_m}{n}$$

for some large enough positive constant $\bar{\kappa}$. *Then, under the assumptions of Theorem 3.1, for* n *large enough,*

$$\mathbb{E}_s\left[\|s - \tilde{s}\|_n^2\right] \leq C_1'\left(\max_{m\in\mathcal{M}} L_m^2\right)\min_{m\in\mathcal{M}}\left\{\|s - s_m\|^2 + \frac{D_m}{n}\right\}.$$

where C_1' *is a positive constant that depends on* $\bar{\kappa}$, r, d_1, d_2, $\|s\|_\infty$, $\iota(f)$, $\|f\|_\infty$.

We omit the proof since it exactly follows the proof of Theorem 12 in [Lac07]. The smoothness conditions arise from the control of $\|s - s_{m^\bullet}\|_\infty$ and $\|f - f_{m_1^\bullet}\|_\infty$, for which we use the results of [Aka10] (Lemma 2). It should be noticed that $m^\bullet$ may differ from $m^\star$. In particular, it may be chosen less fine than $m^\star$ so as to have better estimates of $\|s\|_\infty$ and $\iota(f)$.

Let us now comment on Inequality (3.4), which is similar to (3.2), up to the factors C_1, that does not depend on n, and $\max_{m\in\mathcal{M}} L_m^2$. We have already explained that we need irregular partitions to estimate inhomogeneous functions. However, irregular partitions often form a too rich collection. If L_m only depends on D_m, Condition (3.3) means that L_D have to be large enough to balance the number of models of same dimension D. If the number of model for each dimension is high, the L_m's have to be high too. For instance, [BM97] use weights $(L_m)_{m\in\mathcal{M}}$ of order $\log(n)$ to ensure condition (3.3), which spoils the rates of convergence. We describe in the next section an interesting collection of partitions for which the factor $\max_{m\in\mathcal{M}} L_m^2$ can be bounded by a constant, although the collection is rich enough to have good approximation qualities with respect to functions of inhomogeneous smoothness.

Let us mention that we can define an estimator $\tilde{s}^*$ for which we can control the risk associated to the norm $\|.\|$ instead of $\|.\|_n$.

Corollary 3.2. *Define* $\tilde{s}^* = \tilde{s}\mathbb{1}_{\|\tilde{s}\|\leq n}$. *Then, under assumptions of Theorem 3.1,*

$$\mathbb{E}_s\left[\|s - \tilde{s}^*\|^2\right] \leq C_1''\left(\max_{m\in\mathcal{M}} L_m^2\right)\min_{m\in\mathcal{M}}\left\{\|s - s_m\|^2 + \frac{D_m}{n}\right\}.$$

where C_1'' *is a positive real that depends on* κ, r, d_1, d_2, $\|s\|_\infty$, $\iota(f)$, $\|f\|_\infty$.

The proof exactly follows the proof of Theorem 4 in [Lac07] and then is omitted. (The idea is the following: when $\|\tilde{s}\| \leq n$ then the result is already proved; and $P(\|\tilde{s}\| > n) \leq n^{-2}\mathbb{E}\|\tilde{s}\|^2 \leq 2n^{-2}(\|s\|^2 + \mathbb{E}\|s - \tilde{s}\|^2)$ is low enough to become a remainder term.) Then all the following results (Theorems 2–5) can be stated for the $\mathbb{L}_2$-norm $\|.\|$ replacing $\tilde{s}$ by $\tilde{s}^* = \tilde{s}\mathbb{1}_{\|\tilde{s}\|\leq n}$.

3.2. The penalized estimator based on dyadic partitions

Let us describe the particular collection of partitions that we use here. We call dyadic rectangle of $[0,1]^d$ any set of the form $I_1 \times \ldots \times I_d$ where, for all $1 \leq l \leq d$,

$$I_l = [0, 2^{-j_l}] \quad \text{or} \quad I_l = (k_l 2^{-j_l}, (k_l+1)2^{-j_l}]$$

with $j_l \in \mathbb{N}$ and $k_l \in \{1, \ldots, 2^{j_l} - 1\}$. Otherwise said, a dyadic rectangle of $[0,1]^d$ is defined as a product of d dyadic intervals of $[0,1]$ that may have different lengths. We consider the collection of partitions of $[0,1]^d$ into dyadic rectangles with sidelength $\geq 2^{-J_\star}$, where $J_\star$ is a nonnegative integer chosen according to Proposition 3.1 below. We denote by $\mathcal{M}^{rect}$ such a collection of partitions. Let us underline that a partition of $\mathcal{M}^{rect}$ may be composed of rectangles with different Lebesgue measures, as illustrated by Figure 1.

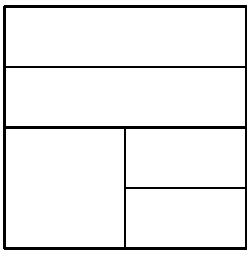

FIG 1. *A partition of* $[0,1]^2$ *into dyadic rectangles.*

For such a collection, we obtain as a straightforward consequence of Theorem 3.1 that the estimator $\tilde{s}$ is almost as good as the best estimator in the collection $\{\hat{s}_m\}_{m \in \mathcal{M}^{rect}}$.

Proposition 3.1. *The notation is that of Theorem 3.1 and Assumption* ***(B)*** *is supposed to be fulfilled. Let*

$$J_\star = \max\left\{k \in \mathbb{N} \ s.t. \ 2^{kd} \leq \sqrt{n}\right\}$$

and let pen *be given on* $\mathcal{M}^{rect}$ *by*

$$\text{pen}(m) = \kappa \left(\|s\|_\infty + \frac{(2r+1)^d}{\iota(f)} \right) \frac{D_m}{n}$$

where κ *is some positive absolute constant. If* κ *is large enough, then*

$$\mathbb{E}_s\left[\|s - \tilde{s}\|_n^2\right] \leq C_2 \min_{m \in \mathcal{M}^{rect}} \left\{ \|s - s_m\|_f^2 + \frac{D_m}{n} \right\} \tag{3.6}$$

where C_2 *is a positive real that depends on* κ, r, d_1, d_2, $\|s\|_\infty$, $\iota(f), \|f\|_\infty$.

Proof. Let D a positive integer. Building a partition of $[0,1]^d$ into D dyadic rectangles amounts to choosing a vector $(l_1,\dots,l_{D-1}) \in \{1,\dots,d\}^{D-1}$ of cutting directions and growing a binary tree with root corresponding to $[0,1]^d$ and with D leaves. For instance, the partition of $[0,1]^2$ represented in Figure 1 can be described by the binary tree structure represented in Figure 2 together with the sequence of cutting directions $(2,1,2,2)$, where 1 stands for a vertical cut, and 2 stands for a horizontal cut. Since the number of binary trees with D leaves

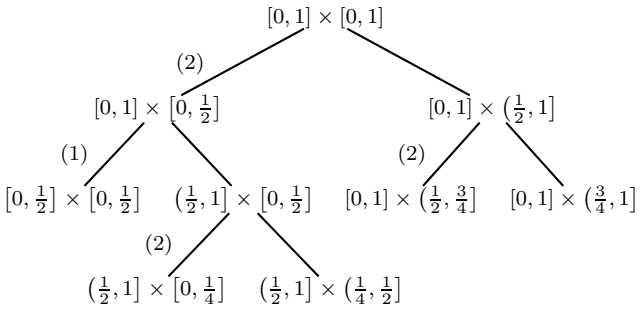

FIG 2. *Binary tree labeled with the sequence of cutting directions* $(2,1,2,2)$ *corresponding with the dyadic partition represented in Figure 1.*

is given by the Catalan number

$$\frac{1}{D}\binom{2(D-1)}{D-1} \leq \frac{4^D}{D}$$

(see for instance [Sta99]), the number of such partitions is at most $(4d)^D$. Therefore, Condition (3.3) is fulfilled for weights L_m all equal to the same constant, and a possible choice is

$$L_m = \log(8d), \text{ for all } m \in \mathcal{M}^{rect}.$$

Inequality (3.6) is then a straightforward consequence of Theorem 3.1. □

We are now able to compute estimation rates for the penalized estimator based on the collection $\mathcal{M}^{rect}$ over the anisotropic Besov balls defined by (3.5), by combining Proposition 3.1 with the approximation results of [Aka10] (Proposition 2 and Theorem 2). Let

$$q(\boldsymbol{\sigma},d,p) = \frac{\underline{\boldsymbol{\sigma}}}{H(\boldsymbol{\sigma})}\frac{d+2H(\boldsymbol{\sigma})}{2H(\boldsymbol{\sigma})}\left(\frac{H(\boldsymbol{\sigma})}{d} - \left(\frac{1}{p}-\frac{1}{2}\right)_+\right),$$

where $(x)_+$ stands for the positive part of a real x. Contrary to [Kle09], we have chosen a parameter $J_\star$ that does not depend on the unknown smoothness of s, hence the factor $\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})$ in the above definition. That factor, which is inferior or equal to 1 with equality only in the isotropic case, may be interpreted as an

index measuring the lack of isotropy. We assume that $q(\boldsymbol{\sigma}, d, p) > 1$, which is equivalent to

$$\frac{H(\boldsymbol{\sigma})}{d} > \begin{cases} \frac{1}{\lambda} - \frac{1}{2} & \text{if } p \geq 2 \\ \frac{1}{2}\left(\frac{1}{p} - 1 + \frac{1}{\lambda} + \sqrt{\left(\frac{1}{p} - 1 + \frac{1}{\lambda}\right)^2 + 2\left(\frac{1}{p} - \frac{1}{2}\right)}\right) & \text{if } 0 < p < 2, \end{cases}$$

where $\lambda = \underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})$. Thus, if $q(\boldsymbol{\sigma}, d, p) > 1$, then $H(\boldsymbol{\sigma})/d > 1/p$, so $\mathcal{B}(\boldsymbol{\sigma}, p, R)$ only contains continuous functions which are uniformly bounded by $C(\boldsymbol{\sigma}, r, d, p)R$.

Theorem 3.2. *The notation is that of Theorem 3.1 and Proposition 3.1, and the assumptions those of Proposition 3.1. Let $p > 0$ and $\boldsymbol{\sigma} \in (0, r+1)^d$ such that $q(\boldsymbol{\sigma}, d, p) > 1$. If $n^{-1} \leq R^2 \leq n^{q(\boldsymbol{\sigma},d,p)-1}$, then there exists some positive real $C(\boldsymbol{\sigma}, r, d, p)$ that only depends on $\boldsymbol{\sigma}, r, d, p$ such that*

$$\sup_{s \in \mathcal{B}(\boldsymbol{\sigma},p,R)} \mathbb{E}_s\left[\|s - \tilde{s}\|_n^2\right] \leq C_2 C(\boldsymbol{\sigma}, r, d, p)\|f\|_\infty \left(Rn^{-H(\boldsymbol{\sigma})/d}\right)^{2d/(d+2H(\boldsymbol{\sigma}))}.$$

The rate $\left(Rn^{-H(\boldsymbol{\sigma})/d}\right)^{2d/(d+2H(\boldsymbol{\sigma}))}$ is the minimax one given the lower bounds proved in [Lac07] for transition density estimation of a Markov chain. We are able to reach that rate not only for functions with homogeneous smoothness, *i.e.* for $p \geq 2$, as [Lac07], but also for functions with inhomogeneous smoothness, *i.e.* for $0 < p < 2$, which is impossible with the collection of regular models considered in [Lac07]. Besides, let us underline that, among the references cited in the introduction, only [Kle09] can deal simultaneously with anisotropy and inhomogeneous smoothness. Theorem 3.2 improves on [Kle09] by allowing to approximately reach the minimax risk up to a factor that does not depend on n and considering smoothness parameters possibly larger than 1.

4. Dependent data

We now show that the previous results can be extended to dependent variables. The case of a Markov chain is of particular interest: if $(X_i)_{i\in\mathbb{Z}}$ is a homogeneous Markov chain of order 1, and $Y_i = X_{i+1}$ for all $i \in \mathbb{Z}$, then s is the transition density of the chain $(X_i)_{i\in\mathbb{Z}}$.

4.1. Definitions and notation

Let us introduce the notions of dependence used in the sequel. For two sub-σ-fields $\mathcal{A}$ and $\mathcal{B}$ of $\mathcal{F}$, the β-mixing (or absolute regularity) coefficient is defined by

$$\beta(\mathcal{A}, \mathcal{B}) = \mathbb{E}\left[\sup_{B\in\mathcal{B}} |\mathbb{P}(B|\mathcal{A}) - \mathbb{P}(B)|\right],$$

and the ρ-mixing (or maximal correlation) coefficient by

$$\rho(\mathcal{A}, \mathcal{B}) = \sup_{X,Y} \frac{|\mathrm{Cov}(X, Y)|}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

where the supremum is taken over all real-valued random variables X and Y that are respectively $\mathcal{A}$ and $\mathcal{B}$-measurable and square integrable. We recall that β and ρ-mixing are among the weakest forms of mixing conditions, in the sense that both β and ρ-mixing are implied by ϕ-mixing (uniform mixing) and imply α-mixing (see for instance [Dou94]). Besides, in general, ρ-mixing does not imply β-mixing, and β-mixing does not imply ρ-mixing. In the sequel, the letter θ stands for β or ρ. For all positive integer j, let

$$\theta_j^{\boldsymbol{Z}} = \theta\left(\sigma(Z_i, i \leq 0), \sigma(Z_i, i \geq j)\right).$$

The process $(Z_i)_{i\in\mathbb{Z}}$ is said to be θ-mixing when $\lim_{j\to+\infty} \theta_j^{\boldsymbol{Z}} = 0$. In particular, $(Z_i)_{i\in\mathbb{Z}}$ is geometrically θ-mixing with rate b, $b > 0$, if there exists a positive constant a such that, for all positive integer j, $\theta_j^{\boldsymbol{Z}} \leq a\exp(-bj)$. We shall also use the 2-mixing coefficients $\theta(\sigma(Z_0), \sigma(Z_j))$, that satisfy, for all $j \geq 1$,

$$\theta(\sigma(Z_0), \sigma(Z_j)) \leq \theta_j^{\boldsymbol{Z}} \tag{4.1}$$

and, if $(Z_i)_{i\in\mathbb{Z}}$ is a Markov chain, $\theta(\sigma(Z_0), \sigma(Z_j)) = \theta_j^{\boldsymbol{Z}}$.

4.2. Dependence assumptions

We consider the following dependence assumptions. Except for the last one, they are related to some rate of mixing. In each case, we also define a real ϑ, that may vary according to the dependence assumption, and will appear in the penalty proposed in the following section.

Assumption ($D\beta$) *The process $(Z_i)_{i\in\mathbb{Z}}$ is geometrically β-mixing, with $a \geq 0$ and $b > 0$ such that, for all positive integer j, $\beta_j^{\boldsymbol{Z}} \leq a\exp(-bj)$. Then we denote $\vartheta = 1$ and $\delta = 1$.*

Assumption ($D\beta\rho$) *Assumptions ($D\beta$) is satisfied and, in addition, the series $S_\rho := \sum_{j\in\mathbb{N}} \rho_{2^j}^{\boldsymbol{Z}}$ converges. Then we denote $\vartheta = 250\prod_{j=0}^{\infty}\left(1 + \rho_{\lfloor 2^{j/3}\rfloor+1}^{\boldsymbol{Z}}\right)$ and $\delta = 0$.*

Assumption ($D\beta$2-ρ) *Assumptions ($D\beta$) is satisfied and, in addition, the series $S_{2\text{-}\rho} := \sum_{j\geq 1} \rho(\sigma(Z_0), \sigma(Z_j))$ converges. Then we denote $\vartheta = (1 + 2S_{2\text{-}\rho})$ and $\delta = 0$.*

Assumption ($D\beta_{cond}$) *Assumptions ($D\beta$) is satisfied and, in addition, for all $j \geq 2$, Z_j is independent of Z_1 conditionally to X_j. Then we denote $\vartheta = 1$ and $\delta = 0$.*

Note that ($D\beta\rho$) is in some sense a weaker assumption than ($D\beta$2-ρ), since a logarithmic ρ-mixing is sufficient. In particular, if $(Z_i)_{i\in\mathbb{Z}}$ is a Markov chain,

$(D\beta 2\text{-}\rho)$ implies **$(D\beta\rho)$** (according to (4.1) and since $(Z_i)_{i\in\mathbb{Z}}$ is geometrically ρ-mixing if and only if it is ρ-mixing (*cf.* [Bra05], Theorem 3.3)). On the other hand, Assumption **$(D\beta_{cond})$** does not imply ρ-mixing. For instance, if $(X_i)_{i\in\mathbb{Z}}$ is a Markov chain and $Y_i = X_{i+1}$ for all $i \in \mathbb{Z}$, then Assumption **$(D\beta_{cond})$** is satisfied, but $(X_i)_{i\in\mathbb{Z}}$, and therefore $(Z_i)_{i\in\mathbb{Z}}$, can be chosen non mixing (*cf.* [DP05] for instance).

Let us give sufficient conditions for $(Z_i)_{i\in\mathbb{Z}}$ to be θ-mixing. First, if $(X_i)_{i\in\mathbb{Z}}$ is a strictly stationary θ-mixing process, and $Y_i = X_{i+1}$ for all $i \in \mathbb{Z}$, then $(Z_i)_{i\in\mathbb{Z}}$ also is θ-mixing since, for all $j \geq 2$,

$$\theta_j^{\boldsymbol{Z}} = \theta_{j-1}^{\boldsymbol{X}}.$$

Next, if $(Z_i)_{i\in\mathbb{Z}}$ is a strictly stationary Harris ergodic Markov chain (aperiodic, irreducible, positive Harris recurrent), then $(Z_i)_{i\in\mathbb{Z}}$ is geometrically β-mixing, i.e. Assumption **$(D\beta)$** is verified, if and only if it is geometrically ergodic (*cf.* [Bra05], Theorem 3.7). In the sequel, we will mainly be concerned with mixing assumptions possibly involving ρ-mixing and β-mixing at the same time. Under adequate hypotheses, Markov chains (always assumed to be homogeneous of order 1) provide examples of such processes:

- if $(Z_i)_{i\in\mathbb{Z}}$ is a strictly stationary Harris ergodic Markov chain that is also reversible and geometrically ergodic, then $(Z_i)_{i\in\mathbb{Z}}$ is both geometrically ρ-mixing and geometrically β-mixing (*cf.* [Jon04], Theorem 2);
- if $(Z_i)_{i\in\mathbb{Z}}$ is a strictly stationary, ergodic and aperiodic Markov chain satisfying the Doeblin condition, then $(Z_i)_{i\in\mathbb{Z}}$ is uniformly ergodic, hence both geometrically ρ-mixing and geometrically β-mixing (*cf.* [Bra05], 119–121, or [MT93], Section 16.2).

We refer to [DG83, Mok90, DT93, Dou94, AN98] for examples of stationary processes that are geometrically β-mixing or both geometrically β and ρ-mixing among commonly used time series such as nonlinear ARMA or nonlinear ARCH models.

4.3. Main result

All the results of Section 3 can be extended to the case of dependent data, under slightly more restrictive conditions on the thinest partition.

Assumption (P2) *All the partitions in the collection $\mathcal{M}$ are built on a regular partition $m^\star$ of $[0,1]^d$ into cubes such that*

$$|m^\star|^2 \leq \frac{n}{\log^2(n)}$$

By comparison with Theorem 3.1, a logarithmic factor then appears in the penalty (and then in the rate of estimation) under the sole condition of β-mixing but this term disappears under Assumption **$(D\beta\rho)$**, **$(D\beta 2\text{-}\rho)$** or **$(D\beta_{cond})$**,

hence the factor $\log^{\delta}(n)$ with $\delta \in \{0,1\}$. Let us first present the oracle type inequality.

Theorem 4.1. *Let $\mathcal{M}$ be a collection of partitions satisfying Assumption **(P2)** and $\{L_m\}_{m\in\mathcal{M}}$ be a family of reals, greater than or equal to 1, such that*

$$\sum_{m\in\mathcal{M}} \exp(-L_m|m|) \leq 1.$$

*Assume that $(Z_i)_{i\in\mathbb{Z}}$ satisfies Assumption **($D\beta$)** and s, f satisfy Assumption **(B)**. If the penalty satisfies, for all $m \in \mathcal{M}$,*

$$\mathrm{pen}(m) = \kappa \left(\vartheta (b^{-1}\log(n))^{\delta} \|s\|_{\infty} + \frac{(2r+1)^d}{b^2 \iota(f)} \right) \frac{L_m^2 D_m}{n}$$

for some large enough positive absolute constant κ (where b, δ and ϑ are defined in the assumptions of dependence), then

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \leq C_3 \left(\max_{m\in\mathcal{M}} L_m^2\right) \min_{m\in\mathcal{M}} \left\{ \|s-s_m\|_f^2 + \log^{\delta}(n)\frac{D_m}{n} \right\}.$$

where C_3 is a positive constant that depends on κ, ϑ, δ, a, b, r, d_1, d_2, $\|s\|_{\infty}$, $\iota(f)$, $\|f\|_{\infty}$.

Under Assumptions **($D\beta\rho$)**, **($D\beta$2-ρ)**, the price to pay for avoiding the logarithmic factor despite the dependence of the data is the presence of the term ϑ in the penalty. For practical purposes, it is necessary to include this term in the constant κ to calibrate. Notice that under Assumption **($D\beta_{cond}$)**, for instance when we estimate the transition density of a Markov chain, the logarithmic factor still disappears and $\vartheta = 1$ so that the penalty is almost as simple as in the independent case. Actually it is possible to consider an arithmetical β-mixing instead of a geometrical one. In this case, it is necessary to slightly strengthen assumption **(P2)**, assuming rather $|m^*|^2 \leq n^{1-\zeta}$, with ζ a number in $(0,1)$. Then, if $\beta_q \leq aq^{-b}$ with $b > 5/\zeta - 2$, Theorem 4.1 is still valid in the cases where $\delta = 0$ (ρ-mixing and conditional independence). The penality is identical, except the term b^2 which is removed. The proof is the same as the original statement, see Subsection 6.3, but with $q_n = \lfloor n^{\xi} \rfloor$ where $\xi \in ((5-\zeta)/(2+2b), \zeta/2)$.

Then, for our penalized estimator based on partitions into dyadic rectangles described in Section 3.2, we can state the following theorem.

Theorem 4.2. *The notation is that of Theorems 4.1, Assumption **(B)** is supposed to be fulfilled. Let*

$$J_{\star} = \max\left\{k \in \mathbb{N} \ s.t.\ 2^{kd} \leq \sqrt{n}/\log(n)\right\}.$$

Let $p > 0$ and $\boldsymbol{\sigma} \in (0, r+1)^d$ such that $q(\boldsymbol{\sigma}, d, p) > 1$. If $\log^{\delta}(n)/n \leq R^2 \leq n^{q(\boldsymbol{\sigma},d,p)-1}\log(n)^{\delta - 2q(\boldsymbol{\sigma},d,p)}$, then there exists some positive real $C(\boldsymbol{\sigma}, r, d, p)$ that only depends on $\boldsymbol{\sigma}, r, d, p$ such that

$$\sup_{s\in\mathcal{B}(\boldsymbol{\sigma},p,R)} \mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right] \leq C_4 C(\boldsymbol{\sigma}, r, d, p)\|f\|_{\infty} \left(R\left(\frac{n}{\log^{\delta}(n)}\right)^{-H(\boldsymbol{\sigma})/d} \right)^{2d/(d+2H(\boldsymbol{\sigma}))}.$$

Thus we recover the same rate of estimation as with independent data (*cf.* Theorem 3.2) up to a logarithmic factor that disappears under Assumptions $\boldsymbol{(D\beta\rho)}$, $\boldsymbol{(D\beta 2\text{-}\rho)}$ or $\boldsymbol{(D\beta_{cond})}$.

4.4. Remarks on the dependence assumptions

We can wonder if weaker assumptions of dependence could be used. Another assumption of dependence is used for instance by [Bos98] (Theorem 2.1) to prove that, asymptotically, the quadratic risk of kernel density estimators reaches the minimax rate (see also [CM02]). But we can prove (see [Aka09]) that this assumption is much stronger than Assumption $\boldsymbol{(D\beta 2\text{-}\rho)}$, which is enough for obtaining the optimal estimation rate from the minimax point of view.

It is difficult to bound the risk for $\tilde{s}$ under weaker dependence assumptions but it is possible to weaken the assumptions to bound the risk $\mathbb{E}\left[\|s-\hat{s}_m\|_n^2\right]$ for one model. In [Aka09], a version of Proposition 2.1 is proved under assumptions of geometrical α-mixing. Actually a sufficient condition to ensure that $\mathbb{E}_s\left[\|\hat{s}_m-s_m\|_n^2\right]$ is of the same order as in the independent case is that for some constant C and all $t \in S_m$,

$$\mathrm{Var}\left(\sum_{i=1}^{n} t(Z_i)\right) \leq Cn\mathrm{Var}\left(t(Z_1)\right). \tag{4.2}$$

Assumptions $\boldsymbol{(D\beta\rho)}$ and $\boldsymbol{(D\beta 2\text{-}\rho)}$ are optimal for obtaining such an inequality in the following sense. Let us assume that $(Z_i)_{i\in\mathbb{N}}$ is a strictly stationary Harris ergodic and reversible Markov chain satisfying (4.2) for all real-valued function t defined on $[0,1]^d$. Then the chain is variance bounding in the sense of [RR08], which implies that there is a spectral gap in $\mathbb{L}_2(sf) := \{t : [0,1]^d \to \mathbb{R} \text{ s.t. } \langle t, sf\rangle = 0 \text{ and } \|t\|_{sf} < \infty\}$ (Theorem 14 in [RR08]). This leads to the geometrical ergodicity of the chain (Theorem 2.1 in [RR97]), which, given the reversibility assumption, implies that the chain is ρ-mixing. As a conclusion, a strictly stationary Harris ergodic and reversible Markov chain $(Z_i)_{i\in\mathbb{Z}}$ satisfies (4.2) for all real-valued function t defined on $[0,1]^d$ if and only if it is ρ-mixing.

5. Implementation and simulations

In order to provide useful characterizations for $\hat{s}_m$ and $\hat{m}$ in practice, we need to introduce some adequate basis of each S_m, for $m \in \mathcal{M}$. Let $(Q_j)_{j\in\mathbb{N}}$ be the orthogonal family of the Legendre polynomials in $\mathbb{L}_2([-1,1])$. For all $j \in \mathbb{N}$, we recall that Q_j satisfies

$$\|Q_j\|_\infty = 1 \quad \text{and} \quad \|Q_j\|^2 = \frac{2}{(2j+1)}. \tag{5.1}$$

For $K_1 = \prod_{i=1}^{d_1}[u_i, v_i]$ rectangle of $[0,1]^{d_1}$, $k_1 = (k_1(1), \dots, k_1(d_1)) \in \{0, \dots, r\}^{d_1}$ and $x = (x_1, \dots, x_{d_1}) \in [0,1]^{d_1}$, we set

$$\phi_{K_1,k_1}(x) = \frac{1}{\sqrt{\mu_{d_1}(K_1)}} \prod_{i=1}^{d_1} \sqrt{2k_1(i)+1}\, Q_{k_1(i)}\left(\frac{2x_i - u_i - v_i}{v_i - u_i}\right) \mathbb{1}_{K_1}(x),$$

where μ_{d_1} denotes the Lebesgue measure in $\mathbb{R}^{d_1}$. Therefore, for K_1 rectangle in $[0,1]^{d_1}$, $(\phi_{K_1,k_1})_{k_1 \in \{0,\dots,r\}^{d_1}}$ is a basis of the space of piecewise polynomials functions with support K_1 and coordinate degree $\leq r$, which is orthonormal for the norm $\|.\|$. For K_2 rectangle in $[0,1]^{d_2}$ and $k_2 \in \{0, \dots, r\}^{d_2}$, we define in the same way ψ_{K_2,k_2} on $[0,1]^{d_2}$. For K rectangle in $[0,1]^d$, we shall denote by K_1 and K_2 the rectangles in $[0,1]^{d_1}$ and $[0,1]^{d_2}$ such that $K = K_1 \times K_2$. For $k \in \{0, \dots, r\}^d$, we shall denote by k_1 and k_2 the multi-indices in $\{0, \dots, r\}^{d_1}$ and $\{0, \dots, r\}^{d_2}$ such that $k = (k_1, k_2)$. For any rectangle $K \in [0,1]^d$ and any multi-index $k \in \{0, \dots, r\}^d$, we define $\Phi_{K,k}$ by

$$\Phi_{K,k}(x,y) = \phi_{K_1,k_1}(x)\psi_{K_2,k_2}(y)$$

for $z = (x,y) \in [0,1]^{d_1} \times [0,1]^{d_2}$. Thus, for a partition m of $[0,1]^d$ into rectangles, the family $(\Phi_{K,k})_{K \in m, k \in \{0,\dots,r\}^d}$ is a basis of S_m, orthonormal for the norm $\|.\|$.

We denote by

$$\hat{s}_m = \sum_{K \in m} \sum_{k \in \{0,\dots,r\}^d} \hat{a}_{K,k} \Phi_{K,k}$$

the decomposition of $\hat{s}_m$ in the basis $(\Phi_{K,k})_{K \in m, k \in \{0,\dots,r\}^d}$. For all $K \in m$, we define the matrices

$$A_K = \left(\hat{a}_{K,(k_1,k_2)}\right)_{(k_1,k_2) \in \{0,\dots,r\}^{d_1} \times \{0,\dots,r\}^{d_2}},$$

$$\Upsilon_K = \left(\frac{1}{n}\sum_{i=1}^{n} \phi_{K_1,k_1}(X_i)\psi_{K_2,k_2}(Y_i)\right)_{(k_1,k_2) \in \{0,\dots,r\}^{d_1} \times \{0,\dots,r\}^{d_2}},$$

and

$$G_{K_1} = \left(\frac{1}{n}\sum_{i=1}^{n} \phi_{K_1,k_1}(X_i)\phi_{K_1,l_1}(X_i)\right)_{(k_1,l_1) \in \{0,\dots,r\}^{d_1} \times \{0,\dots,r\}^{d_1}}.$$

Since ϕ_{K_1,k_1} and ϕ_{L_1,l_1} (resp. ψ_{K_2,k_2} and ψ_{L_2,l_2}) have disjoint supports when $K_1 \neq L_1$ (resp. $K_2 \neq L_2$) and $(\psi_{K_2,k_2})_{k_2 \in \{0,\dots,r\}^{d_2}}$ is orthonormal, we obtain after some computation that, for all $K \in m$, A_K is given by

$$G_{K_1} A_K = \Upsilon_K. \tag{5.2}$$

Let us mention that when $r = 0$, we can write, for all rectangle K (we do not mention any index k),

$$\hat{s}_m \mathbb{1}_K = \frac{1}{\mu_{d_2}(K_2) \sum_{i=1}^{n} \mathbb{1}_{K_1}(X_i)} \sum_{i=1}^{n} \mathbb{1}_K(Z_i) \quad \text{if some } X_i \in K_1,$$

and $\hat{s}_m \mathbb{1}_K = 0$ otherwise, where μ_{d_2} denotes the Lebesgue measure in $\mathbb{R}^{d_2}$.

Thanks to Formula (5.2), one can check that, for all $m \in \mathcal{M}^{rect}$,

$$\gamma(\hat{s}_m) = -\sum_{K\in m}\sum_{k\in\{0,\dots,r\}^d} (A_K)_{(k_1,k_2)}(\Upsilon_K)_{(k_1,k_2)}.$$

We shall consider a penalty pen of the form

$$\text{pen}(m) = c\|\hat{s}_{m\bullet}\|_\infty \frac{D_m}{n}, \tag{5.3}$$

where c is some positive constant, as in Theorem 4.1. With such a penalty, $\hat{m}$ is given by

$$\hat{m} = \underset{m\in\mathcal{M}^{rect}}{\text{argmin}} \sum_{K\in m} \mathcal{L}(K)$$

where, for all rectangle K,

$$\mathcal{L}(K) = \sum_{k\in\{0,\dots,r\}^d} \left(-(A_K)_{(k_1,k_2)}(\Upsilon_K)_{(k_1,k_2)} + c\frac{\|\hat{s}_{m\bullet}\|_\infty}{n}\right).$$

That characterization allows to determine $\hat{m}$ without having to compute all the estimators of the collection $\{\hat{s}_m\}_{m\in\mathcal{M}^{rect}}$. Indeed, we can for instance adapt to our estimation framework the algorithm proposed by [Don97], which requires a computational complexity of order $2^{dJ_\star}$. Thus, choosing $2^{dJ_\star}$ at most of order n, which allows for a larger choice of $J_\star$ than prescribed by our theoretical results (*cf.* Proposition 3.1 and Theorem 4.2), the computational complexity is at most linear in the number of observations. Let us also mention that the algorithm proposed by [BSRM07] allows for instance the slightly larger choice $J_\star = \lfloor \log(n) \rfloor$, that does not depend on d, while keeping an almost linear computational complexity, that is of order $nd\log^{d+1}(n)$.

We propose a simulation study based on the 4 following examples.

Example 1.

$$Y_i = 0.5X_i + 1 + \epsilon_i, i = 1,\dots,n,$$

where $(X_i)_{1\le i\le n}$ are i.i.d. Gaussian variables with mean 6 and variance $4/3$, $(\epsilon_i)_{1\le i\le n}$ are i.i.d. reduced and centered Gaussian variables, independent of the X_i's.

Example 2.

$$Y_i = \sin(X_i) + (\cos(X_i) + 3)\epsilon_i, i = 1,\dots,n,$$

where $(X_i)_{1\le i\le n}$ are i.i.d. uniformly distributed over $[-6, 6]$, $(\epsilon_i)_{1\le i\le n}$ are i.i.d. reduced and centered Gaussian variables, independent of the X_i's.

Example 3. Let $\beta(., a, b)$ be the density of the β distribution with parameters a and b,

$$Y_i = \frac{1}{3}(X_i + 1) + \left(\frac{1}{9} - \frac{1}{23}\left(\frac{1}{2}\beta(5X_i/3, 4, 4) + \frac{1}{20}\beta((5X_i - 2)/3, 400, 400)\right)\right)\epsilon_i$$

where $(X_i)_{1\leq i\leq n}$ are i.i.d. uniformy distributed in $[0,1]$, $(\epsilon_i)_{1\leq i\leq n}$ are i.i.d. reduced and centered Gaussian variables, independent of the X_i's.

Example 4.

$$Y_i = \frac{1}{4}(g(X_i)+1) + \frac{1}{8}\epsilon_i, i = 1,\dots,n$$

where $(X_i)_{1\leq i\leq n}$ are i.i.d. uniformy distributed in $[0,1]$, $(\epsilon_i)_{1\leq i\leq n}$ are i.i.d. Gaussian reduced and centered, independent of the X_i's, and g is the density of

$$\frac{3}{4}N_1 + \frac{1}{4}N_2$$

where N_1 is Gaussian with mean $1/2$ and standard error $1/6$, N_2 is Gaussian with mean $3/4$ and standard error $1/18$, N_1 and N_2 are independent.

Each model is of the form

$$Y_i = \mu(X_i) + \sigma(X_i)\epsilon_i,$$

where ϵ_i is a reduced and centered Gaussian variable, so the conditional density of Y_i given X_i is given by

$$s(x,y) = \phi((y-\mu(x))/\sigma(x))/\sigma(x),$$

where ϕ is the density of ϵ_1. Besides, this allows us to consider Markovian counterparts of Examples 1 to 4, that we will call Example 1 (Markov),..., Example 4 (Markov). More precisely, we also estimate the transition density of the Markov chain $(X_i)_{i\geq 1}$ that satisfies

$$X_{i+1} = \mu(X_i) + \sigma(X_i)\epsilon_i,$$

with X_1 that follows the stationary distribution of the chain. Thus, for Example 1 (Markov), X_1 has the same distribution as in Example 1, but in the other examples, the distribution of X_1 differs between the independent and the Markovian cases. In practice, we simulate the chain long enough so that it finally reaches the stationary regime. We estimate s respectively on $[4,8]^2$ for Example 1, $[-6,6]^2$ for Example 2, and $[0,1]^2$ for Examples 3 and 4, both for independent and Markovian data. The four conditional densities are represented on these rectangles in Figure 5. We may say that the first two examples are rather homogeneous functions, whereas the last two are rather inhomogeneous.

We implement $\tilde{s}$ for $r=0$ and choose the following parameters. The supremum norm of s is estimated by $\|\hat{s}_{m^\bullet}\|$, where $m^\bullet$ is the regular partition of $[0,1]^2$ into cubes with sidelength $2^{J_\bullet}$. We select a best partition among those into dyadic rectangles with sidelength $\geq 2^{-J_\star}$, with $2^{J_\star}$ as close as possible to $\sqrt{n}$. For $n=250$, we set $J_\bullet = 2$ and $J_\star = 4$, and for $n=1000$, we set $J_\bullet = 3$ and $J_\star = 5$. Let us denote by $\tilde{s}(c)$ the penalized estimator obtained with the penalty (5.3) for the penalty constant c. For the sample sizes $n=250$ and $n=1000$, we give respectively in Tables 1 and 2 the estimated values of $\|\hat{s}_{m^\bullet}\|_\infty$, $\mathbb{E}_s\left[\|s-\tilde{s}(3)\|_n^2\right]$ and $\min_c \mathbb{E}_s\left[\|s-\tilde{s}(c)\|_n^2\right]$ where the minimum is

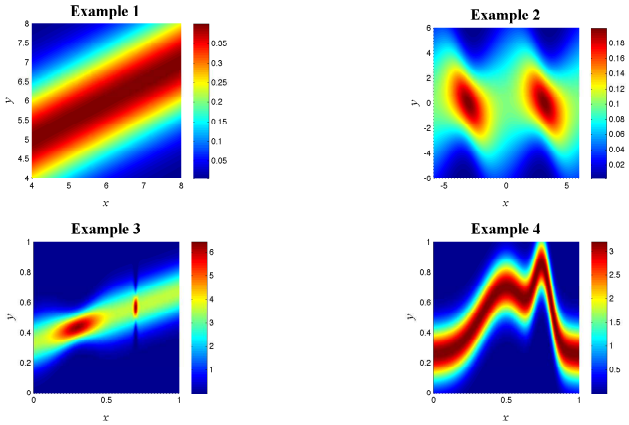

FIG 3. *Level lines of the conditional densities to estimate.*

TABLE 1
Results for $n = 250$ data and 100 simulations.

$n = 250$	$\|s\|_\infty$	$\|\hat{s}_{m\bullet}\|_\infty$	$\mathbb{E}_s\left[\|s - \tilde{s}(3)\|_n^2\right]$	$\min_c \mathbb{E}_s\left[\|s - \tilde{s}(c)\|_n^2\right]$
Example 1	0,4	0,75	0,04	0,04
Example 1 (Markov)	0,4	0,54	0,02	0,02
Example 2	0,22	0,15	0,01	0,02
Example 2 (Markov)	0,22	0,16	0,01	0,02
Example 3	6,5	3,25	0,62	0,61
Example 3 (Markov)	6,5	3,73	0,40	0,40
Example 4	3,2	2,49	0,7	0,7
Example 4 (Markov)	3,2	2,77	0,78	0,72

obtained by varying c from 0 to 4 by step 0.1. All these quantities have been estimated over 100 simulations. Besides, for Example 3, we represent in Figure 5 the selected partition for one simulation with 1000 independent data and the penalty constant $c = 3$. That partitition is both anisotropic and inhomogeneous and well adapted to the function, which illustrates the interest of allowing non-regular and non-isotropic partitions in our selection procedure. Just below, we represent two sections of that conditional density (dark line) together with the corresponding sections of $\tilde{s}(3)$.

The closeness between the minimal risk $\min_c \mathbb{E}_s\left[\|s - \tilde{s}(c)\|_n^2\right]$ and $\mathbb{E}_s\left[\|s - \tilde{s}(3)\|_n^2\right]$ indicates that a penalty constant equal to 3 seems to be a good choice. We observe that, for each example, the risks obtained for the independent and the Markovian cases are also close, which tends to confirm Theorem 4.1, otherwise said that the penalty under assumption $\boldsymbol{(D\beta_{cond})}$ is not so much affected by the dependency between the data. For Examples 1 and 2, we can compare ourselves with the results of Lacour [Lac07] in the Markovian case, obtained via regular model selection. We obtain either similar results for Example 2 or even

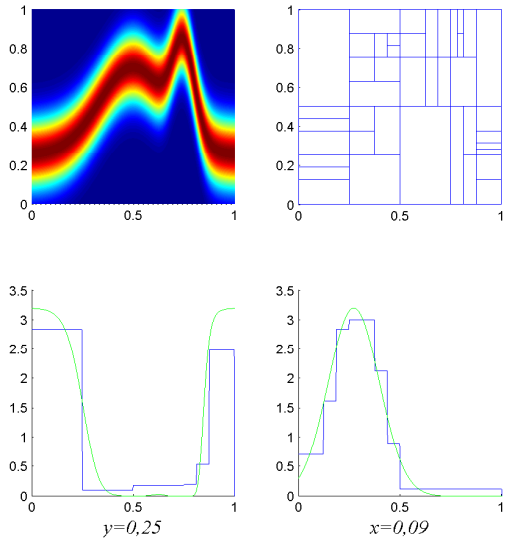

FIG 4. *Top left: Level lines of the the conditional density s for Example 4. Top right: selected partition for $c = 3$ and $n = 1000$. Bottom : two sections of s (dark line) together with the corresponding sections of $\tilde{s}(3)$ (light line).*

better results for Example 1. Last, let us mention that the performance of $\tilde{s}$, in practice, might still be improved by a data-driven choice of the penalty constant based on the slope heuristics, as described in [BMM11] for instance, but this is beyond the scope of the paper.

6. Proofs

6.1. Notation and preliminary lemmas

In all the proofs, the letter C denotes a real that may change from line to line. The notation $C(\theta)$ means that the real C may depend on θ.

For all $t \in \mathbb{L}_2([0,1]^{d_1} \times [0,1]^{d_2})$ and all $z = (x,y) \in [0,1]^{d_1} \times [0,1]^{d_2}$, let

$$\Gamma_t(z) = t(x,y) - \int_{[0,1]^{d_2}} t(x,u)s(x,u)\mathrm{d}u, \tag{6.1}$$

and let ν be the empirical process defined on $\mathbb{L}_2([0,1]^{d_1} \times [0,1]^{d_2})$ by

$$\nu(t) = \frac{1}{n}\sum_{i=1}^{n} \Gamma_t(Z_i). \tag{6.2}$$

TABLE 2
Results for $n = 1000$ data and 100 simulations.

$n = 1000$	$\|s\|_\infty$	$\|\hat{s}_{m\bullet}\|_\infty$	$\mathbb{E}_s\left[\|s - \tilde{s}(3)\|_n^2\right]$	$\min_c \mathbb{E}_s\left[\|s - \tilde{s}(c)\|_n^2\right]$
Example 1	0,4	0,62	0,02	0,02
Example 1 (Markov)	0,4	0,54	0,02	0,02
Example 2	0,22	0,20	0,01	0,01
Example 2 (Markov)	0,22	0,21	0,01	0,01
Example 3	6,5	5,18	0,38	0,38
Example 3 (Markov)	6,5	6,61	0,26	0,25
Example 4	3,2	3,36	0,39	0,39
Example 4 (Markov)	3,2	3,85	0,41	0,41

For all i, $\mathbb{E}_s[\Gamma_t(Z_i)|X_i] = \mathbb{E}_s[t(X_i, Y_i)|X_i] - \int_{[0,1]^{d_2}} t(X_i, u)s(X_i, u)\mathrm{d}u = 0$, so that $\nu(t)$ is centered.

We will use several times the following lemma to bound some variance terms.

Lemma 6.1. *Let q be a positive integer. For all $t \in \mathbb{L}_2([0,1]^{d_1} \times [0,1]^{d_2})$,*

$$\mathrm{Var}_s\left(\sum_{i=1}^{q} \Gamma_t(Z_i)\right) \leq \vartheta q^{1+\delta} \mathrm{Var}_s\left(\Gamma_t(Z_1)\right) \tag{6.3}$$

where δ and ϑ are defined in Section 4.2 for the dependent case, or $\delta = 0$ and $\vartheta = 1$ when the variables Z_i are independent. Besides,

$$\mathrm{Var}_s\left(\Gamma_t(Z_1)\right) \leq \mathbb{E}_s\left[t^2(Z_1)\right] \leq \min\{\|s\|_\infty \|t\|_f^2, \|sf\|_\infty \|t\|^2\}.$$

Proof: First we use a convexity inequality to write, without further assumption,

$$\mathrm{Var}_s\left(\sum_{i=1}^{q} \Gamma_t(Z_i)\right) \leq \mathbb{E}_s\left(q \sum_{i=1}^{q} \Gamma_t^2(Z_i)\right) \leq q^2 \mathrm{Var}_s\left(\Gamma_t(Z_1)\right)$$

whereas in the independent case, $\mathrm{Var}_s\left(\sum_{i=1}^{q} \Gamma_t(z_i)\right) = \sum_{i=1}^{q} \mathrm{Var}_s\left(\Gamma_t(Z_1)\right)$. Now, under Assumption $(\boldsymbol{D\beta\rho})$, Lemma 8.15 in [Bra07] provides

$$\mathrm{Var}_s\left(\sum_{i=1}^{q} \Gamma_t(Z_i)\right) \leq C_1 q \mathrm{Var}_s\left(\Gamma_t(Z_1)\right)$$

with $C_1 = 250 \prod_{j=0}^{\lfloor \log_2 q \rfloor} \left(1 + \rho_{\lfloor 2^{j/3} \rfloor + 1}^{\boldsymbol{Z}}\right)$. Next, by stationarity,

$$\mathrm{Var}_s\left(\sum_{i=1}^{q} \Gamma_t(z_i)\right) = q\mathrm{Var}_s\left(\Gamma_t(Z_1)\right) + 2\sum_{j=1}^{q-1}(q-j)\mathrm{Cov}_s\left(\Gamma_t(Z_1), \Gamma_t(Z_{j+1})\right).$$

Under Assumption $(\boldsymbol{D\beta 2\text{-}\rho})$, we immediately deduce from the definition of the ρ-mixing coefficients and the stationarity of $(Z_i)_{i\in\mathbb{Z}}$ that, for all $1 \leq j \leq q-1$,

$$\left|\mathrm{Cov}_s\left(\Gamma_t(Z_1), \Gamma_t(Z_{j+1})\right)\right| \leq \rho(\sigma(Z_1), \sigma(Z_{j+1}))\mathrm{Var}_s\left(\Gamma_t(Z_1)\right).$$

Thus,

$$\frac{1}{q}\mathrm{Var}_s\left(\sum_{i=1}^{q}\Gamma_t(Z_i)\right) = \mathrm{Var}_s\big(\Gamma_t(Z_1)\big) + 2\sum_{j=1}^{q-1}\left(1-\frac{j}{q}\right)\mathrm{Cov}_s\big(\Gamma_t(Z_1),\Gamma_t(Z_{j+1})\big)$$
$$\leq \left(1+2\sum_{j=1}^{q-1}\rho(\sigma(Z_1),\sigma(Z_{j+1}))\right)\mathrm{Var}_s\left(\Gamma_t(Z_1)\right).$$

Under Assumption $(\boldsymbol{D\beta_{cond}})$

$$\mathrm{Cov}_s\left(\Gamma_t(Z_1),\Gamma_t(Z_{j+1})\right) = \mathbb{E}_s\left[\Gamma_t(Z_1)\mathbb{E}_s\left[\Gamma_t(Z_{j+1})|Z_1,X_{j+1}\right]\right] = 0,$$

hence Inequality (6.3) in the last case.

Besides,

$$\mathrm{Var}_s\left[\Gamma_t(Z_1)\right] = \mathrm{Var}_s\left[t(Z_1)\right] + \mathbb{E}_s\left[\mathbb{E}_s\left[\big(\mathbb{E}_s[t(Z_1)] - \mathbb{E}_s[t(Z_1)|X_1]\big)^2|X_1\right]\right]$$
$$\leq \mathbb{E}_s[t^2(Z_1)] = \int_{[0,1]^{d_1}}\int_{[0,1]^{d_2}} t^2(x,y)s(x,y)f(x)\mathrm{d}x\mathrm{d}y.$$

□

We recall here Bernstein's Inequality for independent random variables (see [Mas07] (Section 2.2.3) for a proof).

Lemma 6.2 (Bernstein inequality)**.** *Let $(W_i)_{1\leq i\leq n}$ be an independent and identically dsitributed sequence, defined on the probability space $(\Omega,\mathcal{F},\mathbb{P})$, with values in $\mathcal{W}$. Let $n\geq 1$ and g be a real-valued and bounded function defined on $\mathcal{W}$. Let $\sigma_g^2 = \mathrm{Var}(g(W_1))$. Then, for all $x>0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}\left(g(W_i)-\mathbb{E}[g(W_i)]\right)\right| \geq \sqrt{2n\sigma_g^2 x} + \|g\|_\infty x/3\right) \leq 2\exp\left(-x\right).$$

6.2. Proof of Proposition 2.1

Since $\hat{s}_m = \underset{t\in S_m}{\mathrm{argmin}}\,\gamma(t)$, we have $\gamma(\hat{s}_m)\leq\gamma(s_m)$. The contrast γ satisfies, for all $t,u\in\mathbb{L}_2([0,1]^{d_1}\times[0,1]^{d_2})$,

$$\gamma(t)-\gamma(u) = \|t-s\|_n^2 - \|u-s\|_n^2 - 2\nu(t-u),$$

where ν is defined by (6.2), hence

$$\|s-\hat{s}_m\|_n^2 \leq \|s-s_m\|_n^2 + 2\nu(\hat{s}_m - s_m).$$

Let

$$\chi_f(m) = \sup_{\substack{t\in S_m\\ \|t\|_f=1}}\nu(t),$$

and let θ be some positive constant, to be chosen later, then

$$\begin{aligned} 2\nu(\hat{s}_m - s_m) &\leq 2\|\hat{s}_m - s_m\|_f \chi_f(m) \\ &\leq \frac{1}{\theta}\|\hat{s}_m - s_m\|_f^2 + \theta\chi_f^2(m). \end{aligned}$$

Let us fix $\eta > 1$, to be determined later, and define

$$\Omega_\eta(m) = \left\{\text{For all } t \in S_m \backslash \{0\}, \|t\|_f^2 \leq \eta\|t\|_n^2\right\}. \tag{6.4}$$

We deduce from the triangle inequality that, on $\Omega_\eta(m)$,

$$\begin{aligned} 2\nu(\hat{s}_m - s_m) &\leq \frac{\eta}{\theta}\|\hat{s}_m - s_m\|_n^2 + \theta\chi_f^2(m) \\ &\leq \frac{2\eta}{\theta}\|s - \hat{s}_m\|_n^2 + \frac{2\eta}{\theta}\|s - s_m\|_n^2 + \theta\chi_f^2(m) \end{aligned} \tag{6.5}$$

Consequently, provided $\theta > 2\eta$,

$$\left(1 - \frac{2\eta}{\theta}\right)\|s - \hat{s}_m\|_n^2 \mathbb{1}_{\Omega_\eta(m)} \leq \left(1 + \frac{2\eta}{\theta}\right)\|s - s_m\|_n^2 + \theta\chi_f^2(m),$$

so that, choosing $\eta = 7/6$ and $\theta = 7$,

$$\frac{2}{3}\mathbb{E}_s\left[\|s - \hat{s}_m\|_n^2 \mathbb{1}_{\Omega_\eta(m)}\right] \leq \frac{4}{3}\|s - s_m\|_f^2 + 7\mathbb{E}_s\left[\chi_f^2(m)\right].$$

Let $(\Phi_\lambda^f)_{\lambda\in\Lambda(m)}$ be a basis of S_m orthonormal for $\|.\|_f$. Since ν is linear, we deduce from Schwarz Inequality and its equality case that

$$\chi_f^2(m) = \sum_{\lambda\in\Lambda(m)} \nu^2\left(\Phi_\lambda^f\right),$$

so

$$n\mathbb{E}_s\left[\chi_f^2(m)\right] = \frac{1}{n}\sum_{\lambda\in\Lambda(m)} \mathrm{Var}\left(\sum_{i=1}^n \Gamma_{\Phi_\lambda^f}(Z_i)\right).$$

Since $Z_1, \ldots, Z_n$ are independent, we deduce from Lemma 6.1 that

$$n\mathbb{E}_s\left[\chi_f^2(m)\right] \leq \|s\|_\infty D_m, \tag{6.6}$$

hence

$$\mathbb{E}_s\left[\|s - \hat{s}_m\|_n^2 \mathbb{1}_{\Omega_\eta(m)}\right] \leq 2\|s - s_m\|_f^2 + 11\frac{\|s\|_\infty D_m}{n}.$$

In order to bound the risk of $\hat{s}_m$ on $\Omega_\eta^c(m)$, we use the following two lemmas, proved just below.

Lemma 6.3. *Assume that d_1 is positive and that s is bounded. Let m be a partition of $[0,1]^{d_1} \times [0,1]^{d_2}$ into rectangles built on a regular partition $m_1^\star \times m_2^\star$, where $m_1^\star$ and $m_2^\star$ are regular partitions of $[0,1]^{d_1}$ and $[0,1]^{d_2}$ into cubes. Then*

$$\|s - \hat{s}_m\|_n^2 \leq 2\|s\|_\infty^2 + 2(r+1)^{d_2}(2r+1)^{d_2}|m_2^\star|.$$

Lemma 6.4. *Let $m^\star = m_1^\star \times m_2^\star$, where $m_1^\star$ and $m_2^\star$ are regular partitions of $[0,1]^{d_1}$ and $[0,1]^{d_2}$ into cubes. Let $\eta > 1$ and $\Omega_\eta(m^\star)$ be defined by (6.4). Then there exists an absolute constant C such that*

$$\mathbb{P}_s(\Omega_\eta^c(m^\star)) \leq C(r+1)^{2d_1}|m_1^\star| \exp\left(-\frac{\iota^2(f)(1-1/\eta)^2 n}{3\|f\|_\infty C(r,d_1)|m_1^\star|}\right).$$

Since m is built on $m^\star = m_1^\star \times m_2^\star$, $\Omega_\eta(m^\star) \subset \Omega_\eta(m)$. Given the conditions on $m_1^\star$ and $m_2^\star$, we then obtain

$$\begin{aligned}\mathbb{E}_s\left[\|s-\hat{s}_m\|_n^2 \mathbb{1}_{\Omega_\eta^c(m)}\right] &\leq 2\left(\|s\|_\infty^2 + (r+1)^{d_2}(2r+1)^{d_2}|m_2^\star|\right)\mathbb{P}_s(\Omega_\eta^c(m^\star)) \\ &\leq C(r,d_1,s,f)/n,\end{aligned}$$

where $C(r,d_1,s,f)$ is a nonnegative real that only depends on r, d_1, $\|s\|_\infty$, $\iota(f)$ and $\|f\|_\infty$.

Let us end with the proofs of Lemmas 6.3 and 6.4.

Proof of Lemma 6.3: We shall use the notation

- $\|.\|_{\mathbb{R}^n}$ defined for $v = \{v_i\}_{1\leq i\leq n} \in \mathbb{R}^n$ by $\|v\|_{\mathbb{R}^n} = \sum_{i=1}^n v_i^2/n$;
- for $t \in \mathbb{L}_2([0,1]^{d_1}\times[0,1]^{d_2})$ and $y \in [0,1]^{d_2}$, $t^{\mathbf{X}}(y) = \{t(X_i,y)\}_{1\leq i\leq n} \in \mathbb{R}^n$;
- $\mathcal{V}_m^{\mathbf{X}}(y) = \{t^{\mathbf{X}}(y), t \in S_m\}$ and $\mathcal{P}_{\mathcal{V}_m^{\mathbf{X}}(y)}$ the orthogonal projection of $\mathbb{R}^n$ on $\mathcal{V}_m^{\mathbf{X}}(y)$.

For all $y \in [0,1]^{d_2}$, let us also define the $\mathbb{R}^n$-vector

$$\hat{v}_m(y) = \left\{\sum_{J\in m_2}\sum_{j\in\{0,\dots,r\}^{d_2}} \psi_{J,j}(Y_i)\psi_{J,j}(y)\right\}_{1\leq i\leq n}.$$

As [Lac07] (Proposition 2.1), we can prove that $\hat{s}_m^{\mathbf{X}}(y) = \mathcal{P}_{\mathcal{V}_m^{\mathbf{X}}(y)}\left(\hat{v}_m(y)\right)$. Using the triangle inequality and the shrinking property of $\mathcal{P}_{\mathcal{V}_m^{\mathbf{X}}(y)}$, we get

$$\begin{aligned}\|s-\hat{s}_m\|_n^2 &= \int_{[0,1]^{d_2}} \|s^{\mathbf{X}}(y) - \hat{s}_m^{\mathbf{X}}(y)\|_{\mathbb{R}^n}^2 \mathrm{d}y \\ &\leq 2\int_{[0,1]^{d_2}} \|s^{\mathbf{X}}(y)\|_{\mathbb{R}^n}^2 \mathrm{d}y + 2\int_{[0,1]^{d_2}} \|\hat{v}_m(y)\|_{\mathbb{R}^n}^2 \mathrm{d}y.\end{aligned}$$

From the orthonormality of $\{\psi_{J,j}\}_{J\in m_2, j\in\{0,\dots,r\}^{d_2}}$, we deduce that

$$\int_{[0,1]^{d_2}} \|\hat{v}_m(y)\|_{\mathbb{R}^n}^2 \mathrm{d}y = \frac{1}{n}\sum_{i=1}^n \sum_{J\in m_2}\sum_{j\in\{0,\dots,r\}^{d_2}} \psi_{J,j}^2(Y_i).$$

Now, using (5.1),

$$\|\psi_{J,j}\|_\infty^2 = \frac{\prod_{i=1}^{d_2}(2k_2(i)+1)}{\mu_{d_2}(J)} \leq \frac{(2r+1)^{d_2}}{\mu_{d_2}(J)}.$$

Then, by grouping the $\psi_{J,j}$ having the same support, we get

$$\left\| \sum_{J\in m_2} \sum_{j\in\{0,\dots,r\}^{d_2}} \psi_{J,j}^2 \right\|_\infty \leq \max_{J\in m_2} \left\| \sum_{j\in\{0,\dots,r\}^{d_2}} \psi_{J,j}^2 \right\|_\infty \leq (2r+1)^{d_2}(r+1)^{d_2} / \min_{J\in m_2} \mu_{d_2}(J),$$

hence Lemma 6.3. □

Proof of Lemma 6.4: The proof follows almost the same lines as the proof of Proposition 8 in [Lac07]. Let ν' be the centered empirical process defined for $u \in \mathbb{L}_2([0,1]^{d_1} \times [0,1]^{d_2})$ by

$$\nu'(u) = \frac{1}{n} \sum_{i=1}^{n} \left(\int_{[0,1]^{d_2}} u(X_i, y)\mathrm{d}y - \int_{[0,1]^{d_1}\times[0,1]^{d_2}} u(x,y) f(x)\mathrm{d}x\mathrm{d}y \right).$$

Since $\|t\|_n^2 = \nu'(t^2) + \|t\|_f^2$ for all $t \in \mathbb{L}_2([0,1]^{d_1} \times [0,1]^{d_2})$, ν' is linear and $\eta > 1$, we get

$$\Omega_\eta^c(m^\star) \subset \left\{ \sup_{t\in S_{m^\star}/\|t\|_f=1} |\nu'(t^2)| > 1 - 1/\eta \right\}.$$

By construction of $(\Phi_{K,k})_{K\in m^\star, k\in\{0,\dots,r\}^d}$, for all $K, L \in m^\star$ and $k, l \in \{0,\dots,r\}^d$, and all $i \in \{1,\dots,n\}$,

$$\begin{aligned} \int_{[0,1]^{d_2}} \Phi_{K,k}(X_i, y)\Phi_{L,l}(X_i, y)\mathrm{d}y &= \phi_{K_1,k_1}(X_i)\phi_{L_1,l_1}(X_i)\langle \psi_{K_2,k_2}, \psi_{L_2,l_2}\rangle \\ &= \mathbb{1}_{K_1=L_1}\mathbb{1}_{(K_2,k_2)=(L_2,l_2)}\phi_{K_1,k_1}\phi_{L_1,l_1}(X_i). \end{aligned} \tag{6.7}$$

Let $t \in S_{m^\star}\backslash\{0\}$, and for $K_1 \in m_1^\star$ and $k_1 \in \{0,\dots,r\}^{d_1}$, let

$$a_{K_1,k_1} = \sqrt{\sum_{K_2\in m_2^\star} \sum_{k_2\in\{0,\dots,r\}^{d_2}} \langle t, \Phi_{K_1\times K_2,(k_1,k_2)}\rangle^2} / \|t\|.$$

It follows from (6.7) and Schwarz inequality that

$$|\nu'(t^2)| \leq \|t\|^2 \sum_{K_1\in m_1} \sum_{k_1,l_1\in\{0,\dots,r\}^{d_1}} a_{K_1,k_1} a_{K_1,l_1} |\nu''(\phi_{K_1,k_1}\phi_{K_1,l_1})|$$

where ν'' is the centered empirical process defined on $\mathbb{L}_2([0,1]^{d_1})$ by

$$\nu''(u) = \frac{1}{n} \sum_{i=1}^{n} \left(u(X_i) - \int_{[0,1]^{d_1}} u(x) f(x)\mathrm{d}x \right).$$

Consequently

$$\sup_{t\in S_m/\|t\|_f=1} |\nu'(t^2)| \leq \iota^{-1}(f) \max_{a\in\mathcal{A}} \sum_{K_1\in m_1} \sum_{k_1,l_1\in\{0,\dots,r\}^{d_1}} a_{K_1,k_1} a_{K_1,l_1} |\nu''(\phi_{K_1,k_1}\phi_{K_1,l_1})|,$$

where $\mathcal{A} = \left\{ a = (a_{K_1,k_1})_{K_1 \in m_1, k_1 \in \{0,\dots,r\}^{d_1}} \text{ s.t. } \sum_{K_1 \in m_1} \sum_{k_1 \in \{0,\dots,r\}^{d_1}} a^2_{K_1,k_1} = 1 \right\}$.
Let us introduce $B = (B_{K_1,k_1,l_1})_{K_1 \in m_1, k_1, l_1 \in \{0,\dots,r\}^{d_1}}$ and $V = (V_{K_1,k_1,l_1})_{K_1 \in m_1, k_1, l_1 \in \{0,\dots,r\}^{d_1}}$ defined respectively by

$$B_{K_1,k_1,l_1} = \|\phi_{K_1,k_1}\phi_{K_1,l_1}\|_\infty \quad \text{and} \quad V_{K_1,k_1,l_1} = \|\phi_{K_1,k_1}\phi_{K_1,l_1}\|.$$

Let us set

$$\bar{\rho}(B) = \sup_{a \in \mathcal{A}} \sum_{K_1 \in m_1} \sum_{k_1, l_1 \in \{0,\dots,r\}^{d_1}} |a_{K_1,k_1}||a_{K_1,l_1}| B_{K_1,k_1,l_1},$$

define $\bar{\rho}(V)$ in the same way, and set $L(\phi) = \max\{\bar{\rho}^2(V), \bar{\rho}(B)\}$. Then, Schwarz Inequality and the properties of the family $(\phi_{K_1,k_1})_{K_1 \in m_1, k_1 \in \{0,\dots,r\}^{d_1}}$ recalled in Section 6.1 provide

$$L(\phi) \leq C(r, d_1)|m_1^\star|.$$

Let

$$x = \frac{\iota^2(f)(1 - 1/\eta)^2}{3\|f\|_\infty L(\phi)}$$

and

$$\Delta = \bigcap_{K_1 \in m_1^\star, k_1, l_1 \in \{0,\dots,r\}^{d_1}} \left\{ |\nu''(\phi_{K_1,k_1}\phi_{K_1,l_1})| < \sqrt{2\|f\|_\infty x} V_{K_1,k_1,l_1} + \frac{1}{3} B_{K_1,k_1,l_1} x \right\}.$$

One can easily check that, on Δ, $\sup_{t \in S_{m^\star}/\|t\|_f = 1} |\nu'(t^2)| \leq 1 - 1/\eta$, so that $\Omega_\eta^c(m) \subset \Delta^c$. Lemma 6.4 then follows from Lemma 6.2. □

6.3. *Proof of Theorem 4.1*

Let us fix $m \in \mathcal{M}$. We also fix $\eta \geq 1$ and $\theta_1 > 0$, to be determined at the end of the proof. By definition of $\hat{m}$ and $\hat{s}_m$,

$$\begin{aligned} \gamma(\tilde{s}) + \text{pen}(\hat{m}) &\leq \gamma(\hat{s}_m) + \text{pen}(m) \\ &\leq \gamma(s_m) + \text{pen}(m). \end{aligned} \tag{6.8}$$

Using the same arguments as in the proof of Proposition 2.1, we deduce from (6.8) that

$$\|s - \tilde{s}\|_n^2 \leq \|s - s_m\|_n^2 + \text{pen}(m) + 2\nu(\tilde{s} - s_m) - \text{pen}(\hat{m}).$$

As $\tilde{s} - s_m \in S_m + S_{\hat{m}} \subset S_{m^\star}$, we obtain in the same way as Inequality (6.5) that, on the set $\Omega_\eta(m^\star)$ defined as in (6.4),

$$2\nu(\tilde{s} - s_m) \leq \frac{2\eta}{\theta_1}\|s - \tilde{s}\|_n^2 + \frac{2\eta}{\theta_1}\|s - s_m\|_n^2 + \theta_1 \chi_f^2(m, \hat{m})$$

with

$$\chi_f(m, m') = \sup_{\substack{t \in S_m + S_{m'} \\ \|t\|_f = 1}} |\nu(t)|.$$

Consequently, provided $\theta_1 > 2\eta$,

$$\left(1-\frac{2\eta}{\theta_1}\right)\|s-\tilde{s}\|_n^2\mathbb{1}_{\Omega_\eta(m^\star)} \leq \left(1+\frac{2\eta}{\theta_1}\right)\|s-s_m\|_n^2\mathbb{1}_{\Omega_\eta(m^\star)} + \mathrm{pen}(m) + \theta_1\chi_f^2(m,\hat{m}) - \mathrm{pen}(\hat{m}). \quad (6.9)$$

To pursue the proof, we have to control the term $\chi_f^2(m,\hat{m})$. Since the data are β-mixing, we can introduce blockwise independent data. More precisely, let $q_n = \lceil 3b^{-1}\log(n)\rceil$ (where b is defined in Assumption $(\boldsymbol{D\beta})$) and let (d_n, r_n) be the unique couple of nonnegative integers such that $n = d_n q_n + r_n$ and $0 \leq r_n < q_n$. For the sake of simplicity, we assume in the sequel that $r_n = 0$ and $d_n = 2p_n > 0$, but the other cases can be treated in a similar way. For $l = 0, \ldots, p_n - 1$, let us set

$$A_l = \{Z_i\}_{2lq_n+1\leq i\leq(2l+1)q_n} \quad \text{and} \quad B_l = \{Z_i\}_{(2l+1)q_n+1\leq i\leq(2l+2)q_n}.$$

As recalled for instance in [Vie97] (proof of Proposition 5.1), we can build, for $l = 0, \ldots, p_n - 1$,

$$A_l^\bullet = \{Z_i^\bullet\}_{2lq_n+1\leq i\leq(2l+1)q_n} \quad \text{and} \quad B_l^\bullet = \{Z_i^\bullet\}_{(2l+1)q_n+1\leq i\leq(2l+2)q_n}$$

such that, for all $l = 0, \ldots, p_n - 1$,

- A_l, $A_l^\bullet$, B_l and $B_l^\bullet$ have the same distribution;
- $\mathbb{P}_s(A_l \neq A_l^\bullet) \leq \beta_{q_n}^{\boldsymbol{Z}}$ and $\mathbb{P}_s(B_l \neq B_l^\bullet) \leq \beta_{q_n}^{\boldsymbol{Z}}$;
- $(A_l^\bullet)_{0\leq l\leq p_n-1}$ are independent random variables, and so are $(B_l^\bullet)_{0\leq l\leq p_n-1}$.

We set

$$\Omega_\bullet = \bigcap_{i=1}^{n}\{Z_i^\bullet = Z_i\}.$$

The proof of Theorem 4.1 heavily relies on the following concentration inequality satisfied by the random variables $\chi_f^2(m,m')$, for m, m' partition built on $m^\star$. The proof of that proposition is deferred to Section 6.4.

Proposition 6.1. *Under the assumptions of Theorem 4.1, there exists a positive constant C such that*

$$\sum_{m'\in\mathcal{M}} \mathbb{E}_s\left[[\chi_f^2(m,m') - \mathrm{pen}(m) - \mathrm{pen}(m')]_+\mathbb{1}_{\Omega_\bullet}\right] \leq C\frac{\log^\delta(n)}{n},$$

where $[x]_+$ denotes the positive part of a real x and C depends on $\vartheta, \|s\|_\infty, r, d, \iota(f), b$.

We shall first bound the quadratic risk of $\tilde{s}$ on $\Omega_\eta(m^\star)\cap\Omega_\bullet$. Combining (6.9) and Proposition 6.1

$$\left(1-\frac{2\eta}{\theta_1}\right)\|s-\tilde{s}\|_n^2\mathbb{1}_{\Omega_\eta(m^\star)\cap\Omega_\bullet} \leq \left(1+\frac{2\eta}{\theta_1}\right)\|s-s_m\|_n^2\mathbb{1}_{\Omega_\eta(m^\star)} + 2\mathrm{pen}(m) + \theta_1[\chi_f^2(m,\hat{m}) - \mathrm{pen}(m) - \mathrm{pen}(\hat{m})]_+\mathbb{1}_{\Omega_\bullet}$$

hence

$$\left(1-\frac{2\eta}{\theta_1}\right)\mathbb{E}_s\|s-\tilde{s}\|_n^2\mathbb{1}_{\Omega_\eta(m^\star)\cap\Omega_\bullet} \leq \left(1+\frac{2\eta}{\theta_1}\right)\mathbb{E}_s\|s-s_m\|_n^2+2\mathrm{pen}(m) + \theta_1 C\frac{\log^\delta(n)}{n}. \tag{6.10}$$

Let us now bound the quadratic risk of $\tilde{s}$ on $\Omega_\eta^c(m^\star)\cup\Omega_\bullet^c$. A straightforward upper-bound for the $\mathbb{P}_s$-measure of $\Omega_\eta^c(m^\star)\cup\Omega_\bullet^c$ is

$$\mathbb{P}_s(\Omega_\eta^c(m^\star)\cup\Omega_\bullet^c)\leq\mathbb{P}_s(\Omega_\bullet^c)+\mathbb{P}_s(\Omega_\eta^c(m^\star)\cap\Omega_\bullet).$$

One easily deduces from one of the properties of the $A_l^\bullet$'s and $B_l^\bullet$'s that

$$\mathbb{P}_s(\Omega_\bullet^c)\leq 2p_n\beta_{q_n}^{\boldsymbol{Z}}=\frac{n}{q_n}\beta_{q_n}^{\boldsymbol{Z}}. \tag{6.11}$$

In order to bound $\mathbb{P}_s(\Omega_\eta^c(m^\star)\cap\Omega_\bullet)$, we follow the proof of Lemma 6.4. Thus there exists some constant $C(r,d_1)$ that only depends on d_1 and r such that

$$\mathbb{P}_s(\Omega_\eta^c(m^\star)\cap\Omega_\bullet)\leq 4(r+1)^{2d_1}|m_1^\star|\exp\left(-C(r,d_1)\frac{\iota^2(f)(1-1/\eta)^2}{\|f\|_\infty}\frac{n}{q_n|m_1^\star|}\right). \tag{6.12}$$

where $m_1^\star$ and $m_2^\star$ are the partition of $[0,1]^{d_1}$ and $[0,1]^{d_2}$ such that $m^\star=m_1^\star\times m_2^\star$. Combining Inequalities (6.11) and (6.12) with Lemma 6.3 then provides for $\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\mathbb{1}_{\Omega_\eta^c(m^\star)\cup\Omega_\bullet^c\cup\Omega_T^c}\right]$ the upper-bound

$$C(r,d_1,d_2)|m_2^\star|\left(\frac{n}{q_n}\beta_{q_n}^{\boldsymbol{Z}}+|m_1^\star|\exp\left(-C(\eta,r,d_1,\iota(f),\|f\|_\infty)\frac{n}{q_n|m_1^\star|}\right)\right). \tag{6.13}$$

Last, let us choose

$$\eta=7/6,\quad \theta_1=7.$$

Under Assumption **(P2)** on $m_1^\star$ and $m_2^\star$, we deduce from (6.10) and (6.13) that

$$\mathbb{E}_s\left[\|s-\tilde{s}\|_n^2\right]\leq 3\left\{\|s-s_m\|_f^2+\mathrm{pen}(m)\right\} + C(\vartheta,\|s\|_\infty,r,d_1,d_2,\iota(f),\|f\|_\infty,b,a)\frac{\log^\delta(n)}{n}.$$

Theorem 4.1 then follows by taking the minimum over $m\in\mathcal{M}$.

Notice that when the data are independent, we can take $\vartheta=1$, $q_n=1$, $\beta_{q_n}=0$ and $\delta=0$. Then Proposition 6.1 and the rest of the proof are valid under Assumption **(P1)**.

6.4. Proof of Proposition 6.1

We recall that Γ is given by (6.1) and we define on $\mathbb{L}_2([0,1]^{d_1}\times[0,1]^{d_2})$, for all $m,m'\in\mathcal{M}$ and $j=1,2$,

$$\nu^{\bullet}_{(j)}(t)=\frac{1}{n}\sum_{l=0}^{p_n-1}\sum_{i=(2l+j-1)q_n+1}^{(2l+j)q_n}\Gamma_t(Z_i^{\bullet}) \quad\text{and}\quad \chi^{\bullet}_{f,(j)}(m,m')=\sup_{\substack{t\in S_m+S_{m'}\\ \|t\|_f=1}}\nu^{\bullet}_{(j)}(t).$$

We set

$$V=\sqrt{\vartheta q_n^{\delta}\|s\|_{\infty}/2} \quad\text{and}\quad B=2q_n\sqrt{\frac{(2r+1)^d D_{m^\star}}{\iota(f)}}.$$

Since $A_0^{\bullet},\dots,A_{p_n-1}^{\bullet}$ are independent and identically distributed on $\Omega_{\bullet}$, we deduce from Lemma 6.1 that

$$\sup_{\substack{t\in S_m+S_{m'}\\ \|t\|_f=1}}\sum_{l=0}^{p_n-1}\mathrm{Var}_s\left(\frac{1}{n}\sum_{i=2lq_n+1}^{(2l+1)q_n}\Gamma_t(Z_i^{\bullet})\mathbb{1}_{\Omega_{\bullet}}\right)\le\frac{V^2}{n}$$

and also, by using the same arguments as for (6.6), that

$$\mathbb{E}_s^2\left[\chi^{\bullet}_{f,(1)}(m,m')\mathbb{1}_{\Omega_{\bullet}}\right]\le\frac{V^2(D_m+D_{m'})}{n}.$$

If $t\in S_m+S_{m'}$ and $\|t\|_f=1$, then by developing t in the basis $(\Phi_{K,k})_{K\in m^\star,k\in\{0,\dots,r\}^d}$ and using Schwarz Inequality, we get

$$\begin{aligned}\|t\|_{\infty}^2&\le\max_{K\in m^\star}\left(\sum_{k\in\{0,\dots,r\}^d}|\langle t,\Phi_{K,k}\rangle|\|\Phi_{K,k}\|_{\infty}\right)^2\\&\le\max_{K\in m^\star}\left(\sum_{k\in\{0,\dots,r\}^d}\langle t,\Phi_{K,k}\rangle^2\right)\left(\sum_{k\in\{0,\dots,r\}^d}\|\Phi_{K,k}\|_{\infty}^2\right)\\&\le\|t\|^2(r+1)^d(2r+1)^d|m^\star|\\&\le\frac{\|t\|_f^2}{\iota(f)}(r+1)^d(2r+1)^d|m^\star|=\frac{(2r+1)^dD_{m^\star}}{\iota(f)},\end{aligned}$$

hence

$$\frac{1}{n}\left|\sum_{i=2lq_n+1}^{(2l+1)q_n}\Gamma_t(Z_i^{\bullet})\mathbb{1}_{\Omega_{\bullet}}\right|\le\frac{B}{n}.$$

As $\nu=\nu^{\bullet}=\nu^{\bullet}_{(1)}+\nu^{\bullet}_{(2)}$ on $\Omega_{\bullet}$, we have

$$\chi_f(m,m')\mathbb{1}_{\Omega_{\bullet}}\le\chi^{\bullet}_{f,(1)}(m,m')\mathbb{1}_{\Omega_{\bullet}}+\chi^{\bullet}_{f,(2)}(m,m')\mathbb{1}_{\Omega_{\bullet}},$$

and, from the hypotheses on $(A_l^\bullet)_{0\leq l\leq p_n-1}$ and $(B_l^\bullet)_{0\leq l\leq p_n-1}$, $\chi^\bullet_{f,(1)}(m,m')\mathbb{1}_{\Omega_\bullet}$ and $\chi^\bullet_{f,(2)}(m,m')\mathbb{1}_{\Omega_\bullet}$ are identically distributed. Thus, denoting by ε some positive constant, applying Talagrand's inequality (as stated for instance in [Mas07], Inequality (5.50)) to each $\chi^\bullet_{f,(j)}(m,m')\mathbb{1}_{\Omega_\bullet}$, we deduce that, for all $x>0$, there exists an event $\Omega_{m,m'}(x)$ such that $\mathbb{P}_s(\Omega^c_{m,m'}(x))\leq 2\exp(-x)$ and over which

$$\sqrt{n}\chi_f(m,m')\mathbb{1}_{\Omega_\bullet}\leq 2\left(\frac{5}{2}V\sqrt{D_m+D_{m'}}+V\sqrt{2x}+B\frac{x}{\sqrt{n}}\right).$$

Let $u>0$, then on $\Omega^c_{m,m'}(|m'|L_{m'}+u)$,

$$\begin{aligned}\sqrt{n}\chi_f(m,m')\mathbb{1}_{\Omega_\bullet}\leq{}& 2\left(\frac{5}{2}V\sqrt{D_m+D_{m'}}+V\sqrt{2D_{m'}L_{m'}}+B\frac{|m'|L_{m'}}{\sqrt{n}}\right)\\&+2\left(V\sqrt{2u}+B\frac{u}{\sqrt{n}}\right)\end{aligned}$$

Since $q_n=\lceil 3b^{-1}\log(n)\rceil\leq 6b^{-1}\log(n)$ and $|m_\star|\leq\sqrt{n}/\log(n)$,

$$B|m'|\leq B\sqrt{|m_\star|}\sqrt{|m'|}\leq 12\sqrt{\frac{n(2r+1)^dD_{m'}}{b^2\iota(f)}}.$$

Therefore, still on $\Omega^c_{m,m'}(|m'|L_{m'}+u)$,

$$\begin{aligned}\sqrt{n}\chi_f(m,m')\mathbb{1}_{\Omega_\bullet}\leq{}& 2\left(\frac{5}{2}V\sqrt{D_m+D_{m'}}+V\sqrt{2D_{m'}L_{m'}}+12\sqrt{\frac{(2r+1)^dD_{m'}}{b^2\iota(f)}}L_{m'}\right)\\&+4\max\left\{V\sqrt{2u},B\frac{u}{\sqrt{n}}\right\}\end{aligned}$$

so that

$$\begin{aligned}n\chi_f^2(m,m')\mathbb{1}_{\Omega_\bullet}\leq{}& 8\left(\frac{5}{2}V\sqrt{D_m+D_{m'}}+V\sqrt{2D_{m'}L_{m'}}+12\sqrt{\frac{(2r+1)^dD_{m'}}{b^2\iota(f)}}L_{m'}\right)^2\\&+32\max\left\{2V^2u,B^2\frac{u^2}{n}\right\}.\end{aligned}$$

As $L_{m'}\geq 1$ for all $m'\in\mathcal{M}$, choosing pen such that for all $m'\in\mathcal{M}$

$$\mathrm{pen}(m')\mathbb{1}_{\Omega_\bullet}\geq 32\frac{D_{m'}L_{m'}^2}{n}\left(((5/2+\sqrt{2})^2/2)\vartheta q_n^\delta\|s\|_\infty+144^2\frac{(2r+1)^d}{b^2\iota(f)}\right),$$

we obtain

$$\mathbb{P}_s\left((\chi_f^2(m,m')-\mathrm{pen}(m)-\mathrm{pen}(m'))\mathbb{1}_{\Omega_\bullet}\geq 32\max\left\{2V^2\frac{u}{n},B^2\frac{u^2}{n^2}\right\}\right)\leq 2e^{-|m'|L_{m'}-u}.$$

Last, we recall that Fubini's Theorem yields, for any random variable Z,

$$\mathbb{E}\left([Z]_+\right) = \int_0^\infty \mathbb{P}([Z]_+ \geq z)\mathrm{d}z = \int_0^\infty \mathbb{P}(Z \geq z)\mathrm{d}z.$$

Therefore, we obtain by integrating the previous inequality

$$\mathbb{E}_s\left[[\chi_f^2(m,m') - \mathrm{pen}(m) - \mathrm{pen}(m')]_+ \mathbb{1}_{\Omega_\bullet}\right] \leq \frac{32}{n} e^{-|m'|L_{m'}} \left(2V^2 + 4B^2 n^{-1}\right)$$

and since $\sum_{m' \in \mathcal{M}} e^{-|m'|L_{m'}} \leq 1$, we conclude that

$$\sum_{m' \in \mathcal{M}} \mathbb{E}_s\left[[\chi_f^2(m,m') - \mathrm{pen}(m) - \mathrm{pen}(m')]_+ \mathbb{1}_{\Omega_\bullet}\right] \leq \frac{C(\vartheta, \|s\|_\infty, r, d, \iota(f), b)\log^\delta(n)}{n}.$$

6.5. Proof of Theorems 3.2 and 4.2

It is sufficient to use the following theorem, proved in [Aka10] (Proposition 2 and Theorem 2).

Theorem 6.1. *Let $J \in \mathbb{N}$, $R > 0$, $\boldsymbol{\sigma} \in (0, r+1)^d$ and $p > 0$ such that*

$$H(\boldsymbol{\sigma})/d > \max\{1/p - 1/2, 0\}.$$

Assume that $s \in \mathcal{B}(\boldsymbol{\sigma}, p, R)$. Then, for all $k \in \mathbb{N}$, there exists some partition m_k of $[0,1]^d$ that only contains dyadic rectangles with edge-length at least $2^{-J\underline{\boldsymbol{\sigma}}/\sigma_l}$ in the l-th direction, $l = 1, \ldots, d$, such that

$$|m_k| \leq C(d, p, \boldsymbol{\sigma}) 2^{kd}$$

and

$$\|s - s_{m_k}\|^2 \leq C(d, p, r, \boldsymbol{\sigma}) R^2 \left(2^{-2Jd(H(\boldsymbol{\sigma})/d + 1/2 - 1/p)\underline{\boldsymbol{\sigma}}/H(\boldsymbol{\sigma})} + 2^{-2kH(\boldsymbol{\sigma})}\right).$$

Under the assumptions of Theorem 3.2, we set $\delta = \mu = 0$, whereas under the assumptions of Theorem 4.2, δ is given in Section 4.2 and $\mu = 1$. Let us fix R, $\boldsymbol{\sigma}$, $p > 0$ satisfying the assumptions of Theorems 3.2 or 4.2, and $s \in \mathcal{B}(\boldsymbol{\sigma}, p, R)$. Since, $\underline{\boldsymbol{\sigma}} \leq \sigma_l$ for all $l = 1, \ldots, d$, Theorem 6.1 applied with $J = J_\star$ provides partitions $(m_k)_{k \in \mathbb{N}}$ that all belong to $\mathcal{M}^{rect}$. Thus, with $\tau = H(\boldsymbol{\sigma})/d - (1/p - 1/2)_+$, we obtain

$$\begin{aligned}
&\min_{m \in \mathcal{M}^{rect}} \left\{\|s - s_m\|^2 + \log^\delta(n)\frac{|m|}{n}\right\} \\
&\leq \inf_{k \in \mathbb{N}} \left\{\|s - s_{m_k}\|^2 + \log^\delta(n)\frac{|m_k|}{n}\right\} \\
&\leq C(d, p, r, \boldsymbol{\sigma}) \left(R^2 2^{-2J_\star d\tau\underline{\boldsymbol{\sigma}}/H(\sigma)} + \inf_{k \in \mathbb{N}} \left\{R^2 2^{-2kH(\boldsymbol{\sigma})} + \log^\delta(n)\frac{2^{kd}}{n}\right\}\right).
\end{aligned}$$

With $k_\star = \max\{k \in \mathbb{N} \text{ s.t. } 2^{kd}\log^\delta(n)/n \leq R^2 2^{-2kH(\boldsymbol{\sigma})}\}$, which is well-defined for $R^2 \geq \log^\delta(n)/n$, we obtain

$$\min_{m\in\mathcal{M}^{rect}} \left\{ \|s - s_m\|^2 + \log^\delta(n)\frac{|m|}{n} \right\}$$
$$\leq C(d,p,r,\boldsymbol{\sigma}) \left(R^2 2^{-2J_\star d\tau\underline{\boldsymbol{\sigma}}/H(\sigma)} + \left(R(n\log^{-\delta}(n))^{-H(\boldsymbol{\sigma})/d} \right)^{2d/(d+2H(\boldsymbol{\sigma}))} \right).$$

Last, since $2^{dJ_\star} \leq \sqrt{n}/\log^\mu(n)$, it is enough to choose $R^2 \leq n^{q(\boldsymbol{\sigma},d,p)-1}(\log(n))^{\delta-2\mu q(\sigma,d,p)}$ so that the first term in the upper-bound is smaller than the second. □

Acknowledgements

We are grateful to Cécile Durot for initiating this project and proofreading earlier versions of this paper. We also thank Yves Rozenhloc for his advice on the implementation of the penalized estimator and Jérôme Dedecker for his explanations on mixing.

References

[Aka09] N. Akakpo. *Estimation adaptative par sélection de partitions en rectangles dyadiques*. PhD thesis, Université Paris-Sud 11, 2009.

[Aka10] N. Akakpo. From adaptive approximation to adaptive estimation for functions with inhomogeneous and anisotropic smoothness. *ArXiv*, 2010.

[AN98] P. Ango Nze. Critères d'ergodicité géométrique ou arithmétique de modèles linéaires perturbés à représentation markovienne. *C. R. Acad. Sci. Paris Sér. I Math.*, 326(3):371–376, 1998.

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

[BCL07] E. Brunel, F. Comte, and C. Lacour. Adaptive Estimation of the Conditional Density in Presence of Censoring. *Sankhyā*, 69(4):734–763, 2007.

[Bir83] L. Birgé. Approximation dans les espaces metriques et theorie de l'estimation. *Probability Theory and Related Fields*, 65(2):181–237, 1983.

[BM97] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[BMM11] J.P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, pages 1–16, 2011.

[Bos98] D. Bosq. *Nonparametric statistics for stochastic processes*, volume 110 of *Lecture Notes in Statistics*. Springer-Verlag, New York, second edition, 1998. Estimation and prediction.

[Bra05] R. C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.*, 2:107–144 (electronic), 2005. Update of, and a supplement to, the 1986 original.

[Bra07] R. C. Bradley. *Introduction to strong mixing conditions. Vol. 1.* Kendrick Press, Heber City, UT, 2007.

[BSRM07] G. Blanchard, C. Schäfer, Y. Rozenholc, and K.R. Müller. Optimal dyadic decision trees. *Machine Learning*, 66(2):209–241, 2007.

[Clé00a] S. Clémençon. *Méthodes d'ondelettes pour la statistique non paramétrique des chaînes de Markov*. PhD thesis, Doctoral thesis, Université Paris VII, 2000.

[Clé00b] S. J. M. Clémençon. Adaptive estimation of the transition density of a regular Markov chain. *Math. Methods Statist.*, 9(4):323–357, 2000.

[CM02] F. Comte and F. Merlevède. Adaptive estimation of the stationary density of discrete and continuous time mixing processes. *ESAIM Probab. Statist.*, 6:211–238 (electronic), 2002.

[DG83] P. Doukhan and M. Ghindès. Estimation de la transition de probabilité d'une chaîne de Markov Doëblin-récurrente. Étude du cas du processus autorégressif général d'ordre 1. *Stochastic Process. Appl.*, 15(3):271–293, 1983.

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

[Don97] D. L. Donoho. CART and best-ortho-basis: a connection. *Ann. Statist.*, 25(5):1870–1911, 1997.

[Dou94] P. Doukhan. *Mixing*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1994. Properties and examples.

[DP05] J. Dedecker and C. Prieur. New dependence coefficients. Examples and applications to statistics. *Probab. Theory Related Fields*, 132(2):203–236, 2005.

[DT93] P. Doukhan and A. B. Tsybakov. Nonparametric recurrent estimation in nonlinear ARX models. *Problemy Peredachi Informatsii*, 29(4):24–34, 1993.

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

[Efr08] S. Efromovich. Oracle inequality for conditional density estimation and an actuarial example. *Annals of the Institute of Statistical Mathematics*, pages 1–27, 2008.

[Eng94] J. Engel. A simple wavelet approach to nonparametric regression from recursive partitioning schemes. *J. Multivariate Anal.*, 49(2):242–254, 1994.

[Eng97] J. Engel. The multiresolution histogram. *Metrika*, 46(1):41–57, 1997.

[Fau07] O. P. Faugeras. A product type non-parametric estimator of the conditional density by quantile transform and copula representation. www. *Arxiv preprint math.ST/0709.3192 v1*, 2007.

[FY04] J. Fan and T. H. Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834, 2004.

[GK07] L. Györfi and M. Kohler. Nonparametric estimation of conditional

distributions. *IEEE Trans. Inform. Theory*, 53(5):1872–1879, 2007.

[GW10] I. Gannaz and O. Wintenberger. Adaptive density estimation under weak dependence. *ESAIM Probab. Statist.*, 14:151–172, 2010.

[HKP98] P. Hall, G. Kerkyacharian, and D. Picard. Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.*, 26(3):922–942, 1998.

[Jon04] G. L. Jones. On the Markov chain central limit theorem. *Probab. Surv.*, 1:299–320 (electronic), 2004.

[Kle09] J. Klemelä. Multivariate histograms with data-dependent partitions. *Statist. Sinica*, 19(1):159–176, 2009.

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

[Mas07] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[Mok90] A. Mokkadem. Propriétés de mélange des processus autorégressifs polynomiaux. *Ann. Inst. H. Poincaré Probab. Statist.*, 26(2):219–260, 1990.

[MT93] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London Ltd., London, 1993.

[Rou69] G.G. Roussas. Nonparametric estimation in markov processes. *Annals of the Institute of Statistical Mathematics*, 21(1):73–87, 1969.

[RR97] G. O. Roberts and J. S. Rosenthal. Geometric ergodicity and hybrid Markov chains. *Electron. Comm. Probab.*, 2:no. 2, 13–25 (electronic), 1997.

[RR08] G. O. Roberts and J. S. Rosenthal. Variance bounding Markov chains. *Ann. Appl. Probab.*, 18(3):1201–1214, 2008.

[Sta99] Richard P. Stanley. *Enumerative combinatorics. Vol. 2*, volume 62 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999. With a foreword by Gian-Carlo Rota and appendix 1 by Sergey Fomin.

[Tri06] H. Triebel. *Theory of function spaces. III*, volume 100 of *Monographs in Mathematics*. Birkhäuser Verlag, Basel, 2006.

[Vie97] G. Viennet. Inequalities for absolutely regular sequences: application to density estimation. *Probab. Theory Related Fields*, 107(4):467–492, 1997.

[WN07] R. M. Willett and Robert D. Nowak. Multiscale Poisson intensity and density estimation. *IEEE Trans. Inform. Theory*, 53(9):3171–3187, 2007.